Doctoral theses at NTNU, 2023:222

Ashish Kumar Singh

Next Generation Sequencing based methods in genetic disease diagnostics

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

NTNU
Norwegian University of
Science and Technology

Ashish Kumar Singh

Next Generation Sequencing based methods in genetic disease diagnostics

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2023

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

ISBN 978-82-326-7144-1 (printed ver.)
ISBN 978-82-326-7143-4 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2023:222

Printed by NTNU Grafisk senter

Table of contents

Summary

Background: Identification of a germline pathogenic variant that increases risk of getting diseases in a family is important for the clinical management of the family members. DNA sequencing is an important molecular diagnostic technology that determines the order of nucleotides in an individual's genetic code. Next generation sequencing (NGS) technologies in DNA sequencing has revolutionized the research and diagnostics within the field of genetic disease, providing the opportunity to perform comprehensive genetic testing of large gene sets and to discover new causative genes. Targeted sequencing, whole exome sequencing, and whole genome sequencing are now widely employed by clinical laboratories using NGS, considering the benefits and difficulties of each technique. The project focuses on the study and use of these NGS technologies and developing bioinformatics analysis strategies for the same, suitable for usages in clinical diagnosis of hereditary diseases with focus on hereditary cancer.

Results: Studies I and III aimed to identify genetic variants that are associated with an increased risk of cancer, specifically endometrial and colorectal cancer respectively. Study I used gene panel sequencing to screen 22 genes involved in the mismatch repair pathway in 199 unselected endometrial cancer patients. Study III performed whole exome sequencing on 48 patients suspected of familial colorectal cancer. Bioinformatic pipelines were used to identify and classify variants, where use of multiple *in silico* tools improved the accuracy of predictions. Study I identified 22 potential pathogenic variants that may be associated with an increased risk of endometrial cancer, and Study III identified 26 germline variants in genes known for their association with colorectal cancer, as well as variants in other genes that may also contribute to an increased risk, hinting for a larger genetic spectrum of colorectal cancer, not

limited to just mismatch repair genes. Study II demonstrated the development of a bioinformatic pipeline for copy number variation (CNV) detection in NGS data from diagnostic gene panels. With a sensitivity of 100% and specificity of 91%, the pipeline has been successful in detect CNVs in all control samples.

Conclusions: These studies (I & III) used gene target panel and exome sequencing to find the likely genetic cause for predisposition in several patients that participated in these studies. These studies also contribute to a larger understanding of the genetic spectrum of cancer. The bioinformatic pipeline (study II) has now been incorporated into routine practices, leading to expansion of the portfolio of genes for which CNV detection can be offered and demonstrated its diagnostic value by identifying CNVs in routine tests of patient samples. This has allowed for efficient and cost-effective CNV detection, which was previously limited by wet lab methods like MLPA. The outcomes of the whole project can help identify patients with inherited increased risk for cancers and other genetic disease, allowing for lifesaving surveillance.

Sammendrag

Bakgrunn: Identifikasjon av en arvelig patogen variant som gir økt risiko for sykdom i en familie er viktig for klinisk oppfølging av familiemedlemmer. Sekvensering av DNA er en viktig teknologi for molekylær diagnostikk som identifiserer rekkefølgen av nukleotidene i den enkeltes genetiske kode. Teknologier for neste generasjon sekvensering (NGS) av DNA har revolusjonert forskning og diagnostikk innenfor genetiske sykdommer, og de gir mulighet for å utføre grundig genetisk testing av store gensett og identifikasjon av nye sykdomsrelaterte gen. Målrettet sekvensering, sekvensering av hele eksomet og sekvensering av hele genomet med NGS er nå i utstrakt bruk ved kliniske laboratorier, som må ta stilling til fordeler og ulemper ved hver metode. Dette prosjektet fokuserer på studie og bruk av disse NGS-teknologiene, og utvikling av bioinformatiske analysestrategier som egner seg for bruk i klinisk diagnose av arvelige sykdommer, med fokus på arvelig kreft.

Resultater: Studie I og III hadde som mål å identifisere genetiske varianter som kan assosieres med økt risiko for kreft, spesifikt henholdsvis endometrie- og kolorektalkreft. Studie I brukte sekvensering basert på genpanel til å undersøke 22 gener involvert i DNA *mismatch reparasjonssystemet* i 199 ikke-selekterte pasienter med endometriekreft. Studie III utførte sekvensering av hele eksomet for 48 pasienter med mistanke om familierelatert kolorektalkreft. Bioinformatiske prosedyrer ble brukt til å identifisere og klassifisere varianter, der en kombinasjon av flere *in silico* verktøy forbedret nøyaktigheten av prediksjonene. Studie I identifiserte 22 mulig patogene varianter som kan assosieres med en økt risiko for endometriekreft, og studie III identifiserte 26 arvelige varianter i gener med kjent assosiasjon til kolorektalkreft, men også varianter i andre gener som kan bidra til en økt risiko, noe som antyder et større genetisk spektrum for kolorektalkreft, ikke bare knyttet til gener involvert i *mismatch reparasjonssystemet*. Studie II viste utvikling av en bioinformatisk prosedyre for deteksjon av kopitallsvarianter (CNV) i NGS data fra diagnostiske genpanel. Med en

sensitivitet på 100% og spesifisitet på 91% lykkes denne prosedyren i å detektere CNV i alle kontrollprøvene.

Konklusjoner: Disse studiene (I & III) brukte genpanel og eksomsekvensering til å finne den sannsynlige genetiske årsaken til predisponering for kreft i flere pasienter som deltok i disse studiene. Disse studiene bidro også til en bredere forståelse av det genetiske spekteret for kreft. Den bioinformatiske prosedyren (studie II) er nå inkludert i rutineundersøkelser, og dette har utvidet utvalget av gener hvor CNV-deteksjon kan bli tilbudt, og dette har demonstrert sin diagnostiske verdi ved å identifisere CNVer i rutineundersøkelser av pasientprøver. Dette gjør det mulig med en rask og kostnadseffektiv CNV-deteksjon, noe som tidligere har vært begrenset av laboratoriemetoder som MLPA. Resultater fra hele prosjektet kan bidra til å identifisere pasienter med arvelig økt risiko for kreft og andre genetiske sykdommer, noe gjør det mulig med livsbesparende overvåking av slike pasienter.

Acknowledgements

This thesis was carried out at Department of Medical Genetics, St. Olavs Hospital and Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology.

I express my deepest gratitude to all the patients who participated in the studies included in this project.

I am immensely grateful to my PhD supervisor, **Finn Drabløs**, for your invaluable guidance, unwavering support, and immense expertise throughout my PhD journey. Your mentorship has shaped the direction of my work, and insightful feedbacks has greatly improved its quality. Finn, I feel truly fortunate and honored to have you as my supervisor, and I will always cherish your dedication and belief in me. Thank you.

I would like to extend my heartfelt gratitude to my co-supervisor, **Wenche Sjursen**, for your invaluable support, guidance, and contributions to my PhD journey. Your profound knowledge in the field of medical genetics has been instrumental in shaping the direction of my research. Your availability and dedication have been greatly appreciated. Thank you.

I would like to thank to my collaborator, Bente Talseth-Palmer, for providing me with the opportunity to collaborate on her two remarkable projects, which led to the creation of Study I and Study III.

I would like to acknowledge and credit my co-authors for their scientific input, as well as their supportive comments and suggestions during the writing process, which greatly contributed to the quality and depth of this work.

I would like to thanks to all my incredible colleagues at LAB-CLG-FoU-AMG-STO. Your cheers, support, and vibrant presence have made the office a lively and enjoyable place. Bodil, thanks for Yoga sessions and for being tante Bodil to my family. Kristine, I appreciate our engaging scientific discussions and your undocumented mentorship. Liss-Anne, your unwavering support and care for my well-being make you a true rock star. Maren, thank you for being project partner (back in 2014), a good friend, and a supportive manager.

I want to thank my amazing colleague and friend, Per. You not only has been my go-to person for scientific wisdom and guidance, but has also been an expert listener during those "losing my marbles" moments. Your support and witty banter have been a lifesaver.

I want to give a big shout-out to my best buddy, Bjørner, for surviving a decade of friendship with me. Not only have you been a constant pillar of support and a fantastic listener, but you've also willingly suffered through countless hours of gym torture with me. Our friendship is the perfect blend of laughter, inside jokes, and shared memories that will always hold a special place in my heart. Thank you for being my family away from family.

I am deeply grateful to my siblings and parents for their unwavering support in my journey. Didi and Bade-bhai thanks for all the love and fights. Papa, your expertise in Chemistry and Enfield-bullet charisma continue to inspire and guide me. Mumma, your unconditional love has been anchoring during the ups and downs. I am forever thankful for your presence in my life. Thanks to my father (in-law), for motivations towards scientific quest.

My beautiful little monkeys Arna and Arnay, you are my greatest motivation and source of inspiration. Your infectious laughter and boundless energy have brought light and warmth to even the most challenging days. Seeing your innocent curiosity and eagerness to learn has reminded me of the importance of perseverance and the joy of discovery.

I would like to take a moment to express my deepest appreciation and admiration for my ultimate cheerleader and partner-in-crime, my extraordinary wife, **Neha**, whose unwavering support and love have been the driving force behind my PhD journey. You are the rock that keeps me grounded, the source of endless inspiration, and my partner-in-everything. Your brilliance, wit, and unwavering belief in me have pushed me to reach new heights and overcome any obstacle. From late nights filled with research to moments of self-doubt, you have been my guiding light and confidante. Your endless patience, understanding, and the ability to put up with my thesis-induced quirks are nothing short of remarkable. Thank you for being the wind beneath my wings, my biggest cheerleader, and the love of my life. I am forever grateful to have you by my side. Together, we make an unbeatable team, and this thesis is a testament to our shared journey.

Lastly, thank you Gurudev for enlightening the path. **Jai Gurudev!**

List of included articles

The papers included in this PhD thesis are listed below and will be referred to in the text by their Roman numerals. Full format papers are attached at the end of this thesis.

I. **Ashish Kumar Singh**, Bente Talseth-Palmer, Mary McPhillips, Liss Anne Solberg Lavik, Alexandre Xavier, Finn Drabløs, Wenche Sjursen . **Targeted sequencing of genes associated with the mismatch repair pathway in patients with endometrial cancer.** *PLOS ONE 15(7): e0235613 (2020).* *https://doi.org/10.1371/journal.pone.0235613*

II. **Ashish Kumar Singh**, Maren Fridtjofsen Olsen, Liss Anne Solberg Lavik, Trine Vold, Finn Drabløs, Wenche Sjursen. **Detecting copy number variation in next generation sequencing data from diagnostic gene panels.** *BMC Medical Genomics 14, 214 (2021).* *https://doi.org/10.1186/s12920-021-01059-x*

III. **Ashish Kumar Singh**, Bente Talseth-Palmer, Alexandre Xavier, Rodney J. Scott Finn Drabløs, Wenche Sjursen. **Detection of germline variants with pathogenic potential in 48 patients with familial colorectal cancer by using whole exome sequencing.** (~~Under review~~ Accepted after revision) *BMC Medical Genomics (2023).* *https://doi.org/10.1186/s12920-023-01562-3*

Other publications where the author contributed:

1. Alexandre Xavier, Maren Fridtjofsen Olsen, Liss Anne Lavik, Jostein Johansen, **Ashish Kumar Singh**, Wenche Sjursen, Rodney J. Scott, Bente A. Talseth-Palmer . **Comprehensive mismatch repair gene panel identifies variants in patients with Lynch-like syndrome.** *Mol Genet Genomic Med. 2019; 7:e850.* *https://doi.org/10.1002/mgg3.850*
2. Marieve J Rocque, Vilde Leipart, **Ashish Kumar Singh**, Pilar Mur, Maren F. Olsen, Lars F. Engebretsen, Edgar Martin-Ramos, Rosa Aligué, Pål Sætrom, Laura Valle, Finn Drabløs, Marit Otterlei, Wenche Sjursen. **Characterization of POLE c.1373A > T p.(Tyr458Phe), causing high cancer risk.** Mol Genet Genomics (2023). *https://doi.org/10.1007/s00438-023-02000-w*

Abbreviations

ACMG	American College of Medical Genetics and Genomics
BAM	Binary Alignment Map
ChIP	Chromatin Immunoprecipitation
CNV	Copy Number Variation
CRC	Colorectal Cancer
DHPLC	Denaturing high performance liquid chromatography
EC	Endometrial Cancer
FAP	Familial Adenomatous Polyposis
FISH	Fluorescence in situ hybridization
GWAS	Genome-wide association study
HNPCC	Hereditary Non-Polyposis Colorectal Cancer
HTS	High Throughput Sequencing
IHC	Immunohistochemistry
Indel	Insertion/Deletion
LS	Lynch Syndrome
MLPA	Multiplex Ligation-dependent Probe Amplification
MMR	Mismatch Repair
MSI	Microsatellite Instability
NGS	Next Generation Sequencing
OG	Oncogene
PCR	Polymerase Chain Reaction
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variation
SV	Structural Variation
TFBS	Transcription Factor Binding Site
TSG	Tumor Suppressor Gene
uORF	upstream open reading frame
UTR	Untranslated Region
VUS	Variant of Unknown Significance
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

Glossary

Big data	Large and complex data sets that are difficult to process using traditional data processing applications.
Binary Alignment Map (BAM)	A format for storing DNA sequencing data in a binary form, used for storing the alignment of the reads from a high-throughput sequencing experiment.
Coverage	The number of times a particular region of a genome is covered by sequencing reads, providing information about the quality of the sequencing data.
Denaturing high performance liquid chromatography (DHPLC)	A method for detecting variations in DNA sequences, used for genotyping and mutation detection.
Exome	The portion of the genome that contains the exons, or protein-coding regions, of the genes.
Exon	A coding region of a gene that is transcribed into messenger RNA (mRNA) and translated into protein.
Familial Adenomatous Polyposis (FAP)	A hereditary condition characterized by the development of many polyps in the colon, leading to an increased risk of colorectal cancer.
FASTQ	A file format used to store high-throughput sequencing data, containing the base call quality scores and the DNA sequence.
Fluorescence in situ hybridization (FISH)	A technique used to study the location of DNA sequences within cells or tissues.
Genome-wide association study (GWAS)	A study that identifies genetic variations associated with a particular trait or disease by comparing the genomes of individuals with and without the trait or disease.
Germline DNA	DNA that is present in the eggs or sperm and is passed from one generation to the next.
High Throughput Sequencing (HTS)	A method for sequencing DNA that allows for the rapid determination of the order of the four nitrogenous bases (A, C, G, and T) in a DNA molecule.
In silico	A term used to describe computational or computer-based analysis or simulations.
Intron	A non-coding region of a gene that is transcribed into RNA but not translated into protein.

Microarray	A technology used to measure the expression of many genes simultaneously by hybridizing labeled RNA or DNA to a solid surface.
Multiplex Ligation-dependent Probe Amplification (MLPA)	A method for detecting changes in the number of copies (copy number variation) of specific genes or regions of DNA, used for genetic testing and diagnosis.
Polymerase chain reaction (PCR)	A method for amplifying specific DNA sequences in vitro, used for a variety of applications in molecular biology and genetic research.
Pseudogene	A non-coding region of DNA that resembles a gene but does not produce a functional protein.
RNA sequencing (RNA-seq)	A method for sequencing the transcriptome, or the collection of all the RNA molecules in a cell, used for the study of gene expression and alternative splicing.
Sanger sequencing	A method for sequencing DNA, also known as dideoxy sequencing, that uses a chain termination approach to determine the order of the four nitrogenous bases (A, C, G, and T) in a DNA molecule.
Short read sequencing	A method for sequencing DNA that generates short reads, typically around 100-300 base pairs in length.
Untranslated region (UTR)	A region of a gene that is transcribed into RNA but not translated into protein, located at the 5' or 3' end of the coding region.
Upstream open reading frames (uORFs)	Short open reading frames located upstream of the main coding sequence in a gene, which can regulate gene expression by disrupting the initiation of translation of the main coding sequence

1. Introduction

1.1 Cancer overview: caseload

Cancer, often referred to as a single disease, actually consists of over 100 different conditions characterized by the uncontrolled growth and spread of abnormal cells. It can arise in many sites and behave differently depending on its organ of origin and tissues type. It starts as a pre-cancerous lesion and progresses to a full-blown malignant tumor, led by uncontrolled cell division and transformation of normal cells into malignant tumor cells forming tumors, invading surrounding tissues, and spreading throughout the body, leading to severe illness and death. In 2020, there were 19.3 million new cases and 10 million deaths globally, making it one of the leading causes of premature mortality worldwide (1). The continuously rising number of cases makes it a significant burden to society. The most frequently diagnosed types of cancer include breast, lung, colon, rectum, prostate, and stomach cancers (2). Due to an aging population and shifts in lifestyle, it is projected that the incidence of cancer will increase by approximately 50% over the next two decades, posing a significant burden on society (2).

1.1.1 Colorectal cancer

Colorectal cancer (CRC) is also referred to as cancer of the colon or rectum. Colon is critical to the growth and development of the organism due to its high rate of cellular renewal. However, this makes the colon susceptible to a range of physical, chemical, and biological agents, increasing the risk of developing diseases such as CRC. CRC is a prevalent form of cancer globally. In 2020, it was estimated that there were over 1.9 million new cases and 935,000 deaths due to CRC, accounting for approximately 10% of all cancer cases and deaths,

and ranking third in terms of incidence and second in terms of cancer-related mortality (2,3). There is a significant discrepancy in CRC incidence rates between developed and developing countries, with rates in developed countries being about four times higher. Furthermore, there is a nine-fold difference in CRC incidence rates among world regions, with the highest rates found in European regions, Australia/New Zealand, and North America, with particularly highest rates in Hungary for men and in Norway for women (2). As such, CRC incidence also serves as an indicator of socioeconomic development (4,5). The etiology of CRC can be attributed to a combination of both elastic and inelastic factors (6). Elastic factors, such as environmental and lifestyle factors, are estimated to account for 70% of all CRC cases, while inelastic factors, including age and hereditary predisposition, are estimated to account for 30% of cases.

1.1.2 Endometrial cancer

Endometrial cancer (EC) is a prevalent gynaecological cancer in developed countries (7), ranking as the sixth most frequently diagnosed cancer among women, with 417,000 new cases and 97,000 deaths reported in 2020. The number of annual incidences of EC has been steadily increasing in recent years (2,8). EC arises from the tissue lining the uterus, known as the endometrium, and its abnormal growth can lead to invasiveness or spread to other parts of the body (8). Risk factors for EC include environmental factors, lifestyle changes, high body mass index, hypertension, menstrual irregularities, and hormonal imbalances (9). However, likelihood of developing EC is also influenced by hereditary factors, as EC has been found to have a higher occurrence among close relatives of EC patients (10).

1.2 Factors behind cancer causes

Cancer is a complex genetic disorder characterized by the progressive accumulation of genetic and epigenetic alterations in genes that control cell growth and division. These genetic alterations are either spontaneous mutations in an individual's DNA during a person's lifetime acquired by exogenous factors such as exposure to environmental carcinogens or endogenous factors such as inherited (passed down within families) genetic predisposition . These genetic changes can increase the risk of developing cancer or influence the progression of the disease.

1.2.1 External factors behind cancer

External factors that contribute to carcinogenesis can be grouped into three categories: physical, chemical, and biological. Physical carcinogens include ultraviolet and electromagnetic radiations, chemical carcinogens encompass substances such as tobacco smoke, alcohol, asbestos, aflatoxin, and arsenic, while biological carcinogens involve infections from viruses, bacteria, or parasites (11). The development of cancer can also be influenced by interactions between an individual's genetic factors and environmental exposures, such as elevated body mass index, hypertension, hormonal imbalances, and menstrual irregularities (9,12).

1.2.2 Hereditary factors behind cancer

The significance of genetics in the occurrence of cancer and other diseases is of utmost importance (13). The molecular mechanisms underlying tumorigenesis are a result of the accumulation of genetic alterations that play a role in regulating epithelial development and cellular differentiation. Changes in DNA structure and function, either spontaneous or facilitated by environmental factors, can lead to the development of diseases, including single gene disorders, chromosomal imbalances, epigenetics, complex disorders, and cancer.

Although diseases caused by a single genetic factor are rare, they account for approximately 80% of all rare diseases, which number in the thousands (13).

The hereditary nature of cancer has been understood for nearly a century. However, the specifics of inherited cancer susceptibility were unclear until recent advancements in diagnostic techniques. These advancements have improved our ability to accurately identify genetic predispositions to cancer (14). Hereditary predisposition to cancers may result from either rare, single gene germline mutations, or more commonly from multiple less-penetrant genes interacting with environmental factors (15).

1.2.3 Cancer genes

Cancer predisposing genes can be divided into two categories: tumor suppressor genes (TSGs) and oncogenes (OGs). Loss of function in TSGs is caused by biallelic mutations, while monoallelic mutations in OGs result in a gain of function. Both of these increase the likelihood of developing cancer.

TSGs play a crucial role in controlling the onset of neoplastic processes by functioning as gatekeepers. This includes regulating cell growth by managing basic cell functions such as cell cycling, proliferation, differentiation, and apoptosis. TSGs also act as caretaker genes, repairing DNA errors and correcting DNA damage, and as landscaper genes, maintaining the stability of the cellular microenvironment. Loss of function mutations in these TSGs can disrupt their ability to control cell division, leading to uncontrolled cell growth and the formation of tumors. These mutations can occur in a variety of ways, such as deletion of genetic material, insertion of foreign genetic material, or changes to the DNA sequence that result in a non-functional protein product. In some cases, loss of function mutations can be inherited, while in others they may develop as a result of exposure to environmental factors such as

radiation or chemicals. Ultimately, loss of function mutations in TSGs can result in the development of cancer by allowing cells to grow and divide in an uncontrolled manner. One common example of TGSs is mismatch repair (MMR) genes, commonly associated with CRC include *MLH1, MSH2, MSH6*, and *PMS2*. In endometrial cancer, the tumor suppressing MMR gene most frequently altered is MSH2. Some other examples of TSGs in CRC include *APC* (16), *TP53* (17), and *SMAD4* (18), and in EC *PTEN* and *P53* (19) .

OGs, on other hand, arise from proto-oncogenes, which regulate normal cell division by encoding growth factor proteins, receptors, membrane-associated signalling proteins, or transcription factors. Gain-of-function mutations in proto-oncogenes result in the activation of OGs, and/or increased protein production leading to uncontrolled signalling and cell proliferation (15). Some frequently observed examples of OGs include *KRAS*, *BRAF* and *PIK3CA*. Mutations in *KRAS* are associated with a high risk of developing CRC and EC, and are often seen in advanced stages of the CRC (20), whereas, mutations in *BRAF* are relatively rare (10%) in CRC but are associated with a more aggressive form of the disease (21). (21)Mutations in *PIK3CA* have been associated with increased activation of the PI3K pathway and promotion of tumor growth in EC (22).

1.2.4 DNA mismatch repair mechanism

The DNA repair genes, a type of TSGs, code for proteins that play a crucial role in maintaining DNA integrity and error correction. This correction process is referred to as the DNA Mismatch Repair (MMR) mechanism. It is a crucial post-replication process that ensures the stability of the genome and promotes genetic stability (23). The main function of the MMR system is to correct DNA replication errors, prevent non-identical DNA recombination, and repair spontaneous base-base mismatches and small insertions/deletions (indels) loops that occur

during DNA replication while ensuring that the genetic material is correctly copied from one generation of cells to the next. MMR mechanism primarily involves 8 TSGs that code for its components, including *MutS* homologs (*MSH2*, *MSH3*, *MSH5*, *MSH6*) and *MutL* homologs (*MLH1*, *PMS1* [*MLH2*], *MLH3*, *PMS2* [*MLH4*]) (24–26). When the MMR mechanism is malfunctioned, it leads to an increased mutation rate in TSGs and loss of function in proteins coded by these genes, which are critical mediators between DNA damage repair and cell survival (24). The DNA damage repair includes correcting alterations in the lengths of microsatellites, which are short repetitive regions in human DNA and are polymorphic in their nature. The development of microsatellite instability (MSI) – a type of genomic instability, is commonly seen in MMR deficient tumor cells (27,28).

1.3 Classification of cancer

Cancer can be classified into three categories based on the genetic basis of its occurrence: sporadic, familial, and hereditary. Approximately 90-95% of all cancers are sporadic, with 5-10% being hereditary or familial (29,30). However, exact distribution may vary depending on the type of cancer and population studied.

1.3.1 Sporadic cancer

Sporadic cancer is the most common type of cancer and occurs spontaneously in an individual without any family history of the disease. The cause of sporadic cancer is often due to acquired mutations in cells that develop during lifetime of an individual due to environmental and lifestyle factors, such as exposure to carcinogens, aging, or lifestyle choices. These accumulated mutations can eventually lead to the development of cancer over the time. Sporadic cancers do not show a pattern of inheritance and are not associated with an inherited

predisposition to cancer. However, some studies have suggested for significant inherited polygenic-risk component in this class of cancer (31).

1.3.2 Familial and hereditary cancer

Familial and hereditary cancers refers to form of cancers that are inherited or run in families. Both caused by the presence of genetic mutations that are passed down from generation to generation.

Familial cancers are more common and often caused by low-penetrant multigenic mutations or sometimes without any clear pattern of inheritance or specific gene mutations that can be identified.

Hereditary cancer are rarer, caused by the inheritance of genetic mutations from one or both parents that predispose individuals to develop cancer. These rare mutations are usually found in a single, highly penetrant gene, which are mostly TSGs. These genetic mutations can significantly increase the risk of developing cancer, especially at a young age, and they can be inherited from either parent. Some common examples of hereditary cancers are *BRCA1/BRCA2* associated hereditary breast and ovarian cancer, and DNA MMR genes associated Lynch syndrome.

1.3.3 Hereditary cancer syndromes

About 5% of all cancers are part of a hereditary cancer syndrome (32). Examples of hereditary cancer syndromes caused by mutated TSGs include Lynch syndrome, Familial Adenomatous Polyposis (FAP), Retinoblastoma, Li-Fraumeni syndrome, and von Hippel-Lindau syndrome. FAP is caused by mutations in the *APC* gene and increases the risk of developing CRC (33). Retinoblastoma is caused by mutations in the *RB1* gene and increases the risk of developing a

rare form of eye cancer (34,35). Li-Fraumeni syndrome is caused by mutations in the *TP53* gene and increases the risk of developing a range of cancers, including sarcoma, breast, and brain cancers (36). Von Hippel-Lindau syndrome is caused by mutations in the *VHL* gene and increases the risk of developing a range of cancers, including kidney, pancreas, and brain cancers (37).

Diagnosis of these hereditary cancer syndromes can help in early prediction, detection and management of potentially associated cancers.

1.3.4 Lynch syndrome

Lynch Syndrome (LS) is an inherited autosomal dominant cancer susceptibility syndrome. It is also known as hereditary non-polyposis colon cancer (HNPCC), and it is characterized by early-onset epithelial cancers (38–41). It is estimated to account for 2-5% of all CRC cases and is more prevalent in individuals with a family history of the disease. While *de novo* mutations are a rare cause of LS, they may account for about 4-8% of all the LS cases (42–44).

LS is caused by mutations in DNA MMR genes, leading to the accumulation of genetic mutations. Individuals with LS are at high risk of developing CRC and EC, as well as other epithelial malignancies like bowel, stomach, ovary, bladder, or pancreas cancer (38,39). LS causing MMR gene alterations, are responsible for one third of all sporadic CRC cases associated with a hereditary condition (45). Moreover, approximately 30% of genetic EC cases have a hyper-mutable phenotype and MSI as a result of MMR dysfunction (46–48).The specific type of cancer that is most likely to develop depends on the MMR gene that is affected. For instance, mutations in *MLH1* and *MSH2* genes increase the risk of CRC and EC, while mutations in *MSH6* and *PMS2* genes are associated with an increased risk of CRC and gastric cancer (42,49).

The lifetime risk of any cancer for LS-affected individuals is to be 64% by the age of 70 years (50), with a risk of 33-61% for EC and 40-80% for CRC (51,52).

1.4 Importance of genetic diagnostics

Identifying genetic factors causing cancer is important as it can lead to a better understanding of the underlying mechanisms of the disease, and further help in improved diagnosis and prognosis, and the development of targeted therapies. Understanding the specific genetic alterations that are responsible for a particular type of cancer can help in determining the most effective targeted treatment options that are more effective and have fewer side effects. For example, some cancers may be more responsive to chemotherapy or radiation, while others may respond better to targeted therapies that specifically target the genetic mutations that drive the cancer. Additionally, identifying genetic factors can help identify people who are at higher risk of developing cancer, enabling earlier detection and intervention. This can improve outcomes and quality of life for affected individuals. Over all, identifying genetic factors in cancer is crucial for improving patient care and outcomes, and for advancing the field of cancer research.

1.5 Genetic test

A genetic test is a laboratory analysis of an individual's genetic material, such as DNA , RNA or protein, to identify certain genetic variations (53–55). There are several different types of genetic tests, including:

- Diagnostic testing to confirm a suspected diagnosis of a genetic disorder.
- Predictive testing to determine an individual's risk of developing a certain disease.

- Carrier testing to identify individuals who carry a gene for a certain genetic disorder.
- Prenatal testing to identify certain genetic disorders in the fetus.
- Newborn screening to identify certain genetic disorders that can be treated early in life.

1.5.1 Cancer diagnostics

Streamlined models of genetic counseling and testing have made genetic diagnosis for cancer susceptibility a key aspect of clinical management for individuals or families with well-defined inherited cancer syndromes (56). In cases of genetic predisposition to cancer, the genetic diagnosis is often made alongside or after the detection of cancer. The use of advanced diagnostic methods to identify predispositions, in conjunction with appropriate surgical interventions and follow-up surveillance, contributes to improved patient survival."

1.6 Disease-causing genetic variations

Genetic variants can impact gene function in various ways. For example, in coding regions, there can be premature stop codons introduced by nonsense mutations, alterations to key residues in the protein by missense mutations, or changes in splicing or exon skipping by synonymous or splice site mutations at exon-intron boundaries (57). Variants in 5’ untranslated regions (UTRs) specifically in transcription factor binding sites (TFBSs), can affect binding affinity with transcription factors, leading to changes in gene expression (58,59). Variants in upstream open reading frames (uORFs) can impact translation initiation rates by sequestering ribosomes, potentially leading to functional activation or the creation of novel uORFs (60–63). Variants in 3' UTRs, which are target sites for microRNAs, can regulate gene expression by

controlling mRNA localization, stability, or translational efficiency, and such variants can affect these processes (64–66).

1.6.1 Types of genetic variations

Disease-causing DNA mutations can vary in length, from a single nucleotide to entire chromosomes. A change of 1 nucleotide is called a single nucleotide variation (SNV), and changes up to 50 nucleotides are called short insertion-deletion variations (indels). Alterations larger than 50 nucleotides are referred to as structural variants (SVs), which can include insertions, deletions, duplications, inversions, translocations, or a combination of these different types, co-occurring in a single genome (67). Deletions and duplications larger than 50 nucleotides, specifically, are referred to as copy number variations (CNVs).

1.7 Next generation sequencing

DNA structure was fundamentally discovered in 1953 by Franklin, Wilkins, Watson and Crick (68). This led to pioneering development of methods to sequence the DNA, starting with development of foundational sequencing methods such as sanger sequencing (69) and Maxam-Gilbert method (70), commonly known as first generation of sequencing technologies. These methods had high accuracy and long read lengths but were slow, labor-intensive, and expensive, making them unsuitable for large-scale projects. Rapid evolution of more methods and newer technologies, i.e. polymerase chain reaction (PCR) method (71,72), widespread availability of modification enzymes and the development of fluorescent DNA sequencing (73), scaled the sequencing yield and led to completion of first human genome project (74–76). However, the need for more advanced sequencing methods became evident. This led to the development of next-generation sequencing (NGS) in the mid-2000s with the advent of high-

throughput sequencing (HTS) machines (77). While other technologies such as DNA microarrays (78), NanoString (79), qPCR (80) and Optical mapping (81) exist, NGS remains the popular term for high-throughput sequencing methods producing millions or trillions of sequencing reads in a single run.

NGS technology has been utilized in different ways, such as targeted gene panels, whole exomes, or whole genomes, depending on the level of genomic coverage desired and the focus of the investigation on either coding or both coding and non-coding regions of DNA. This results in a more comprehensive understanding of genomic information and its biological implications. The high data production capability of NGS has contributed to numerous large-scale studies, such as the NHLBI exome project (ESP) (82), 1000 genome project (1000G) (83), Iceland's deCODE project (84), Genomic England's 100,000 genome (100 K) project (85), giving researchers a comprehensive view of genomic information and its biological implications.

The second generation NGS methods, also known as short-read sequencing methods, produce reads with length of approximately 300-350 bases per read. The genome, with its complex and long repetitive regions (86), presents challenges for these short-read methods to sequence. Additionally, large, complex alterations such as copy number variations and structural variations which play significant roles in evolution, adaptation, and disease (87,88), are longer and cannot be detected using short-read technologies. The more advanced third-generation NGS methods are long-read sequencing techniques, capable of producing read lengths of several kilobases. This enables the spanning of complex or repetitive regions in a single continuous read, reducing ambiguity in the position or size of genomic elements and detecting complex structural variants (89). Furthermore, long read lengths are useful for *de novo* genome assembly and full-length isoform sequencing (90,91). However, high costs, error rates (92) and

limited available analysis tools and pipelines, currently making it challenging for routine use. These third-generation long-read sequencing technologies are still under active development, and future advancements in this field hold the promise of more robust, reproducible, and accurate sequencing of long fragments at higher throughput and lower cost. It will likely be several years before these platforms rival second-generation instruments. However, the second generation short-read NGS technologies have been evolving for over a decade, and have achieved high sequencing yield, cost-effectiveness, accuracy, with support from a wide range of analysis tools and pipelines to detect genetic alterations. These technologies have met the current goals of the diagnostic world, which heavily relies on them. In addition to second-generation short-read DNA sequencing, short-read NGS based RNA sequencing (93,94), DNA methylation-based epigenetic profiling (95), chromatin immunoprecipitation (ChIP) sequencing (96) and microbiome sequencing (97) are also being performed using NGS in routine diagnostics.

Studies have extensively explored the development of NGS technologies, delving into their procedures and highlighting both their benefits and limitations (98–100). Considering the currently available short and long-read sequencing technologies and their respective advantages and disadvantages, choice between short and long sequencing technologies ultimately depends on the purpose of research or diagnosis. Despite current limitations, future advancements in sequencing technology will aim to overcome these hurdles to facilitate further scientific discoveries and clinical applications.

1.8 Bioinformatics

With the recent advancements of NGS technology, the size and complexity of omics datasets, particularly genomic datasets, have increased dramatically, making genomics one of the four biggest big data contributors, alongside astronomy, YouTube, and Twitter. The typical lifecycle of these datasets involves acquisition, storage, distribution and analysis. For the other three big data contributors, only one of these four domains have been cumbersome to handle, i.e., for astronomy it is data acquisition, for YouTube it is data storage, for Twitter it is data mining. Whereas genomics is a "four-headed beast" and is demanding in all the four data domains: acquisition, storage, distribution, and analysis (101). These large-scale genomic and postgenomic datasets, coupled with high-throughput technologies, have paved the way for data-driven biological research that was previously based on hypothesis-testing systems. Computational approaches are revolutionizing biology by integrating large datasets with disease-specific mutation databases, genotype-phenotype analyses, statistics, ontologies, and more. As biological knowledge is defined, organized, and accessed through computation, this is making biological concepts more rigorous and testable, providing a new reference map for biology. In the future, biological research will be driven by mathematical, statistical, and computational methods, turning biology into a quantitative science (102).

Bioinformatics is a branch of data science aimed at solving biological questions. It integrates unique blend of methodologies and scientific cultures, from computer science, statistics, information science, and applied mathematics to decipher the digital codes of life using data-driven approaches (103,104). It involves the development of algorithms, software tools, and databases to process and interpret large amounts of genetic and genomic information. The field plays a crucial role in various areas of biological research including gene expression analysis, comparative genomics, molecular evolution, drug discovery, and personalized medicine. More specifically, tasks such as DNA/protein sequence alignments, genetic variation detection and

functional prediction, macromolecular structure and function prediction, gene/biomolecular interaction network simulation, integration of diverse biological databases, drug discovery, and large-scale genome-wise association studies utilize bioinformatics approaches. Additionally, Bioinformatics tools and platforms enable integration and analysis of multi-scale and multi-omics data including genomics, transcriptomics, proteomics, metabolomics, and microbiomics, providing a more comprehensive understanding and facilitating the develop of new treatments and therapies for human diseases.

1.8.1 Bioinformatics in medical genetics and genomics

Clinical laboratories have adopted NGS technology to enhance their molecular genetics testing, resulting in a surge of data size and complexity. To address this, bioinformatics involves the use of advanced computational methods, algorithms, and software to store, manage, process, analyze, integrate and interpret these large datasets, to make informed decisions in field of genetic disease diagnostics and personalized medicine. As a critical component of medical genetics, it plays a vital role in transforming genetic data into actionable knowledge to improve patient care. In this context, bioinformatics approaches are used in analyzing and interpreting large amounts of genetic data generated by various technologies such as NGS, microarrays, and others. Bioinformatic tools and algorithms are used to process genetic data to identify variants, predict their functional impact, and associate them with disease phenotypes. This information is used to improve the diagnosis, prognosis, and treatment of various genetic conditions. For example, in the case of cancer genetics, bioinformatics is used to analyze the genomic changes in tumor cells to identify potential therapeutic targets. In clinical genetics, bioinformatics is used to analyze patient genetic data to diagnose hereditary diseases and to develop personalized treatment plans. This leads to molecular genetic testing that is faster, more efficient, and cost-effective. With continuously evolving and improving bioinformatics

tools and approaches, the integration of diverse and large amounts of genomic and phenotype information is possible, leading to improved accuracy in results and better patient care.

1.8.2 Bioinformatics analysis of NGS data in diagnostic routines

Bioinformatics operations on NGS datasets generated in clinical settings are typically divided into three stages (105,106). The first stage focuses on sequence generation, the second stage focuses on sequence processing (read alignment and variant detection), and the third stage focuses on results interpretation, which includes computational and evidence-based annotation of the variants. Figure 1 illustrates these bioinformatics analysis stages.

The first stage of bioinformatics operations involves converting raw signals detected by the sequencer into nucleotide bases and assigning a quality score to each position. These bases are then combined to form complete sequences, also known as raw reads. The raw reads are then demultiplexed, which involves distributing the sequences to multiple samples that were indexed and pooled together during a single sequencing run. The bioinformatics steps in this stage can either be done on the sequencing instrument itself or on high-performance computing clusters or cloud-based architectures for larger datasets.

In the second stage of the bioinformatics process, sequence processing takes place (107). This involves mapping raw reads to a reference genome (for short read sequencing) or performing a *de novo* assembly (for long read sequencing) to recreate the full length of the sequenced DNA fragment. Additionally, post-alignment error correction procedures are performed, such as de-duplication, indel realignment, and base quality score recalibration. After alignment, variant calling is carried out on the final aligned reads against the chosen reference genome, such as GRCh37 (108) or GRCh38 (109). These variant calls are then post-processed for error

correction and filtered based on call quality, considering the type of variant (e.g., SNV, Indel, CNV, or SV).

The final step of bioinformatics operations in NGS datasets from clinical settings is variant interpretation and pathogenicity classification. This involves annotating variants using annotation toolkits which collectively assigns all the annotation information from various tools and databases, such as prediction-based tools, evidence-based databases, and frequency-based databases to the variant (110). Further, variants are assigned pathogenicity classes based on rule-based classifiers, such as ACMG guidelines (111). Variants with high class of pathogenicity may then be validated with a secondary sequencing methods, e.g., for SNP/Indel sanger sequencing based validation is used (112); and for validation of CNVs, Multiplex Ligation-dependent Probe Amplification (MLPA) (113), SNP microarrays (114), RNA sequencing (115), fluorescence in situ hybridization (FISH) (116) or PCR based methods (117) are commonly used. Validated variants of pathogenic significance are then reported back to genetic counselors.

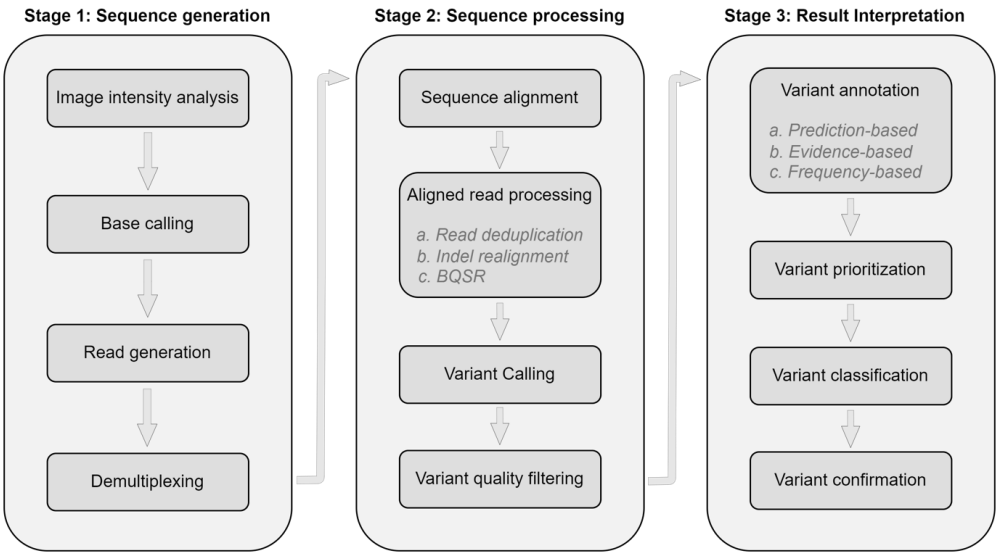


Figure 1: Bioinformatics analysis stages of NGS data

2. Aim of the study

The overall aim of this project has been to develop and implement strategies and tools for the analysis of data from NGS in clinical diagnostics, with focus on improved throughput and performance of a workflow for routine diagnostics of genetic disease.

Current approaches to molecular diagnostics are using NGS as the primary method for producing sequencing data, where the data is utilized for detecting pathogenic variants in genes known to be associated with the disease, but also to search for new genes that also may be associated with the disease. The ever-evolving NGS technology is producing enormous and complex dataset which are still not fully explored and utilized in routine diagnostics, and this poses a challenge to the laboratories towards a more comprehensive and productive utilization of these data.

The overall aim of the project was realized through a few specific goals:

1. *Use real data from clinical diagnostics by targeted gene panels and whole exome sequencing methods to identify challenges and benefits of relevant methods for data analysis, in order to identify good strategies for data analysis in clinical diagnostics.* **Study I and III** were performed towards this goal, where NGS data from EC and CRC patient samples produced by these methods was analyzed.
2. *Analyze both common and novel genes and genomic regions within cancer datasets to identify new genetic variants (including recessive and de novo) that may play a role in cancer.* **Study I and III** were performed to identify variants with possible pathogenic potential towards cancer in both commonly known genes and novel genes.
3. *Use existing bioinformatic tools to develop and implement new analysis strategies for large datasets in diagnostic settings, including variant identification, annotation, prioritization and pathogenicity classification.* To achieve this goal, **Study III**

developed a custom analysis approach for analyzing the large dataset generated by exome sequencing of germline DNA obtained from CRC patients.

4. *Develop a bioinformatic tool to detect larger genetic variants (e.g., CNVs) by using NGS datasets with high precision and sensitivity according to diagnostic standards.* **Study II** was conducted to develop of a bioinformatic pipeline aimed at achieving this goal.

3. Materials and Methods

3.1 Samples

In **Study I**, germline DNA was extracted from blood samples of 199 women diagnosed with sporadic EC and treated at the Hunter Centre for Gynaecological Cancer, John Hunter Hospital in Newcastle, New South Wales, Australia between 1992 and 2005. The samples were collected consecutively from the recruited patients.

The bioinformatics pipeline described in **Study II** was validated using diagnostic routine samples, including 36 positive control samples with previously known CNVs and 11 routine samples without CNVs. The positive control samples were collected from various genetic diagnostic laboratories across Norway, including Haukeland University Hospital (Bergen), University Hospital of North Norway (Tromsø), and St. Olavs Hospital (Trondheim), based on the availability of known CNV positive samples. The 11 routine samples were collected from the Department of Medical Genetics at St. Olavs Hospital. All samples were germline DNA extracted from blood. The validation aimed to assess the sensitivity and specificity of the bioinformatic pipeline by checking only genes with known CNVs in the positive control samples and all genes in the panel in the routine samples.

Study III was conducted to determine the genetic causes of CRC in 48 patients from Australia, who had been diagnosed with CRC or other LS associated cancers (n=6), fulfilling the Amsterdam-II Criteria for Lynch syndrome. The study included 16 related individuals from 8 families and 32 unrelated individuals, and was performed using germline DNA extracted from blood samples. DNA sequencing was performed to identify the underlying genetic factors

contributing to the development of cancer in these patients. Some of the MMR genes were previously tested with Sanger sequencing.

3.2 Next Generation Sequencing

In **Study I**, targeted sequencing was carried out on 199 EC patient samples using Illumina MiSeq (118) at the Medical Genetics Laboratory, Hunter Medical Research Institute, University of Newcastle, Australia. The gene panel consisted of 22 genes associated with the DNA MMR pathway, where all introns, exons, 5' and 3' UTRs were included in the sequencing. A total of 1.213 Mb of nucleotides were captured by 6961 probes utilizing the Illumina Nextera Rapid Capture Custom Enrichment Kit.

In **Study II**, targeted sequencing was performed on 36 patient germline DNA samples extracted from blood, using Illumina MiSeq and Illumina NextSeq 500 (119) sequencer at the Department of Medical Genetics, St Olavs Hospital, Trondheim, Norway. The gene panel comprised 126 genes known to be associated with cancer, with only exons, 5' and 3' UTR regions, and a vicinity of ±25 nucleotides in intronic regions being captured. The total size of nucleotides captured was 0.71 Kb, achieved through the utilization of the Illumina Nextera Rapid Capture Custom Enrichment Kit.

In **Study III**, Whole Exome Sequencing (WES) was performed on 48 germline DNA samples extracted from blood from suspected LS patients. The paired-end library preparation utilized the Illumina Truseq Exome Capturing Kit to target the exons of all protein-coding genes. DNA fragmentation was performed to achieve a size of approximately 150 base pairs. Sequencing was carried out at the Beijing Genomic Institute, China using the Illumina Nextseq 500 kit, with a paired-end read configuration of 150 cycles.

3.3 Bioinformatic data analysis

The data analysis for **Studies I and III** was a multi-step process that included pre-processing of raw data (FASTQ files), variant calling, annotation, and prioritization. Pre-processing and variant calling for both studies were carried out using a standardized BWA-Picard-GATK pipeline (120), which is a widely accepted best practice for NGS data processing. The annotation of variants in both studies was conducted using multiple databases and prediction tools. For annotation, Study I utilized the Alamut-batch software toolkit (121), while Study III utilized the Ensembl variant effector prediction (VEP) toolkit (122). Subsequently, the annotated variants were prioritized based on significance. Study I used the Filtus (123) tool for this step, while Study III used the filter_vep tool from the VEP toolkit.

The data analysis for **Study II** consisted of pre-processing of raw data and identifying copy number variations (CNVs). The raw data was preprocessed using the standardized BWA-Picard-GATK pipeline, similar to the other two studies. The pre-processed data, including the aligned reads (Binary Alignment Map files), were used to determine the per-locus coverage (nucleotide level coverage) of the samples through the use of the GATK toolkit DepthOfCoverage tool. The procedure for identifying CNVs was implemented as pipeline (see article II for details). The pipeline is intended to function within a Unix based operating system and was constructed using a combination of bash scripting and the R programming language. During the development and testing phase, the pipeline utilized R version 3.4 and relied on default R libraries. Furthermore, for improved visualizations, the pipeline can utilize the R library ggplot2.

In all three studies, the human reference genome version GRCh37 (124) was used as the reference genome for bioinformatics analysis.

3.4 Result validation

In **Study I**, Sanger sequencing was utilized for validation of the results. The protocol involved fragment amplification, cycle sequencing, electrophoresis by capillary, and analysis of data using SeqScape software. Some significant variants were not validated through Sanger sequencing, which was partly due to the unavailability of primers for specific genes, as well as logistical challenges. Nevertheless, these variants were carefully scrutinized in Binary Alignment Map (BAM) files to ensure their likely authenticity as true positive variants.

In **Study II**, validation of the bioinformatic pipeline was performed using control samples with known true Copy Number Variants (CNVs) previously detected via methods such as MLPA and/or RNA sequencing and routine diagnostic samples without any CNVs. The pipeline demonstrated 100% sensitivity in detecting all CNVs in the control samples. Furthermore, evaluation against the routine samples resulted in a specificity of 90.9% and total accuracy of 91.14% for the pipeline.

The findings from **Study III** were not validated using alternative techniques such as Sanger sequencing or MLPA due to sample material exhaustion. However, given the high accuracy of current next-generation sequencing (NGS)-based detection of SNVs and indels variants, supplementary validation is often not deemed necessary (125). As was performed in Study I, the variants were thoroughly scrutinized in BAM files to verify their likelihood of being true positive.

3.5 Ethics and consent

Study I was approved by Hunter New England (HNE) Human Research Ethics Committee (HNE HREC: 05/03/09/3.14). Written informed consent was obtained from all participants for the study.

Study II was designated as a quality assurance audit in accordance with the guidelines outlined in the "Guide for Research Ethics Committee Members" by the Steering Committee on Bioethics (Council of Europe, April 2012). These guidelines are also adopted by the Regional Ethical Committee (REK) in Norway (https://rekportalen.no). Consequently, evaluation by the local ethics committee was deemed unnecessary. Written consent was obtained from all patients for conducting the diagnostic genetic testing. The Department of Medical Genetics at St. Olavs Hospital evaluated the utilization of the genetic testing data with regards to anonymity and determined that the study results are anonymous and cannot be linked back to individual patients.

Study III was conducted in accordance with the principles of the Helsinki Declaration, and received approval from both the Hunter New England Human Research Ethics Committee (HNE HREC: 04/03/10/3.11), Australia and the Regional Ethics Committee (REK), Norway (2015/838). Written informed consent was obtained from all participants in the study.

4. Results and summary of studies

4.1 Study I: Targeted sequencing of genes associated with the mismatch repair pathway in patients with endometrial cancer

The genetic causes of familial EC are commonly linked to Lynch syndrome, which is caused by germline variants that inactivate the MMR genes *MLH1*, *MSH2*, *MSH6* and *PMS2*. The MMR pathway comprises a total of 22 genes, including these four. To identify causative pathogenic variants that increase the risk of EC or other related cancers, and to provide life-saving surveillance to patients with known diagnosis, it is crucial to examine additional genes that may also contribute to cancer risk. This can be accomplished by considering all known genes involved in the MMR pathway.

For this study, next-generation sequencing was performed on constitutional DNA extracted from full blood from 199 unselected EC patients to screen all 22 genes involved in the MMR pathway. The sequencing covered the entire gene regions including coding (exonic), noncoding (intronic) and regulatory (UTR) regions to detect all potential variant types, such as those altering protein coding, gene regulation, or splicing, that may contribute to EC causes. The sequencing was done in 12 different batches (runs), and the quality (mean read coverage depth) of the samples varied across different runs, from a minimum 1X up to a maximum 169X of coverage depth. To check the potential of low coverage NGS data, we used low-quality samples as part of our study, which led to identification of true-positive variants in low coverage regions (findings were verified by Sanger sequencing). After annotating and filtrating all 10,680 uniquely detected SNV/Indel variants, we shortlisted 35 significant variants (22 exonic, 4 UTR, 9 intronic) among 34 patients, with three patients having class 5 variants (in the *MSH6* gene)

and two patients having the same class 4 variant (in the *PMS2* gene). These five patients, approximately 2.5% of this patient cohort size, have pathogenic variants that are very likely the cause of their cancer. For the other 29 of 34 patients, we identified class 3 variants with high suspicion of being pathogenic. Of the significant 35 variants, 15 are associated to 4 MMR genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*) while 20 variants are associated to additional genes from MMR pathway including *MSH3, POLD1, RFC1, RFC3, RFC4, LIG1, EXO1, RPA1* and *RPA3*, which are not commonly studies for EC causes. For the remaining 165 patients we could not identify variants with significance for EC or other disease causes, probably due to limitations with the target panel size, data quality and limitation in annotation (especially around intronic regions) or due to sporadic nature of cancer in these patients.

4.2 Study II: Detecting copy number variation in next generation sequencing data from diagnostic gene panels

The size of genetic variants with potential for causing diseases can range from single nucleotides to whole chromosome alterations. To detect small alterations, such as SNV/Indels up to 50 nucleotides, clinical diagnostics labs typically use NGS technologies. However, for detection of larger variations, such as CNVs, traditional methods like MLPA or microarray/arrayCGH are still used by diagnostic labs. These methods are lab-intensive, cost inefficient and have limitation in terms of the availability of gene-specific testing kits. Currently, NGS technologies produce high-quality data, providing the possibility to detect CNVs form NGS data, with resolution down to single exons and with accurate detection of the exact breakpoints of those CNVs.

The study developed a CNV detection pipeline using NGS data from a diagnostic gene panel, which was validated and integrated into routine practices at the Department of Medical Genetics at St. Olavs Hospital.

Pipeline uses coverage depth information of captured regions in a sample to calculate the copy number ratio scores. To increase the resolution of results, each region is divided into fixed sliding windows, allowing detection of small or partial exons.

We validated the pipeline using 36 CNV positive control samples consisting of 36 known CNVs in 12 different genes. These included 4 whole gene deletions, 6 single exon deletions, 17 multi-exon deletions, 2 single + partial exon deletions (break point inside the second exon), 3 single exon duplications and 4 multi-exon duplications, which were previously detected by MLPA or/and RNA sequencing. The pipeline detected all variants, resulting in a 100% sensitivity for this set of samples. Additionally, we used 11 pre-tested diagnostic samples to determine specificity and accuracy (90.9% and of 91.14% respectively). With high sensitivity, specificity and accuracy, this pipeline was approved and implemented in our diagnostic lab in August 2018. Since the implementation in routine, it has proven its diagnostic value by identifying more then 50 CNVs (until year 2021), including whole gene and exonic-level deletions/duplications, partial exonic deletions, and mosaic deletions. Some of these CNVs findings were in genes, never been tested for CNVs (with MLPA) in our lab. It shows the usefulness of this pipeline in routine practices. This has helped in expanding our lab's diagnostic capacity to offer CNV detection on whole gene panel, improving the quality of our diagnostic work.

4.3 Study III: Detection of germline variants with pathogenic potential in 48 patients with familial colorectal cancer by using whole exome sequencing

CRC is one of the most frequently occurring types of cancer, leading to premature mortality. Almost 30% of all CRC cases have a familial component, though only one third of this are caused by high penetrant pathogenic variants leading to disease predisposition including Lynch syndrome caused by defect in MMR genes (45). To discover the etiology of remaining 20% cases, a wider genetic quest is required aimed at identifying novel candidate genes and causal variants, which have not yet been linked to familial CRC.

The 48 patients in this study cohort were diagnosed with CRC or another LS associated cancer, and initial testing for germline mutations in MMR genes via denaturing high performance liquid chromatography (DHPLC) and Sanger sequencing did not detect any mutations. Of these 48 patients, 16 were related individuals from 8 families and 32 were unrelated individuals. The aim was to discover the etiology of familial CRC, which was not caused by high penetrant pathogenic variants leading to disease predisposition as seen in Lynch syndrome.

With aim to identify germline variants with pathogenic potential in genes additional to MMR genes, we performed whole exome sequencing on constitutional DNA extracted from blood from these 48 patients. Variant calling and further three-stage filtration on the complete dataset, led to identification of 346 variants in 302 gene. Of these 302 genes, 38 have known or expected roles in cancer as OGs, TSGs or fusion genes. These 38 genes are associated with 46 variants, which occur in 33 samples. This includes 14 pathogenic or likely pathogenic, 6 variants of uncertain significance (VUS), 4 variants with conflicting interpretation (between pathogenic and VUS in ClinVar) and 22 unknown variants (not reported in ClinVar). These 38 genes include seven well-known cancer genes with high impact towards cancer. These included

BRCA2, MLH1, MSH2, MSH6, PMS2, PTCH1 and *SDHA*. These seven genes are associated with 9 variants with pathogenicity classes 5, 4 or 3, which occur in 10 patients. Identification of high penetrant variants in MMR genes shows the limitations on preliminary testing methods. The identification of variants in 31 genes that have not previously been associated with familial CRC suggests a larger spectrum of genetic variants associated with this disease that is not limited to DNA MMR genes or other known cancer-associated genes.

5. Discussion

The overall discussion highlights the major outcomes of the three studies presented in the thesis. This includes how these studies demonstrate the effectiveness of the analysis strategies and methods developed, and their impact on advancing medical genetics and improving diagnostic practices. For a more in-depth examination of the results from each study, readers are referred to the corresponding articles I, II, and III.

5.1 Highlights of this thesis

This thesis explores the benefits of using NGS in diagnostic practices, focusing on its bioinformatics applications for target panel and WES data in diagnosing hereditary cancers and genetic diseases. The results show the operational advantages of NGS-based diagnostics, including reduced costs, increased diagnostic yield, and improved efficiency.

Study I emphasizes the potential significance of analyzing all 22 genes in the MMR pathway for accurate diagnosis and risk assessment of EC. This can lead to the identification of variants in genes that may contribute to the development of EC, and further research can be done for a better understanding of EC development. The study also highlights the benefits and limitations of using computational tools for predicting the pathogenicity of genetic variants.

Study II describes an in-house developed bioinformatic pipeline that detects CNVs using NGS sequencing data from targeted gene panels. The study highlights the benefits of implementing this tool in routine genetic diagnostics at AMG, St. Olavs hospital, Trondheim, including the ability to detect partial, single or multi-exonic, and intragenic CNVs in all genes in target panel, which expands the diagnostic portfolio to include CNV detection in all the genes in panel. This operational capability was previously limited to only gene with available MLPA kits.

Study III explores the broader genetic landscape of familial CRC by discovering variants in previously underexplored genes, beyond the traditional MMR genes. It also highlights the challenges of handling large list of genetic variants and proposes filtration strategies to streamline this process, a crucial step in WES data analysis. The results of this study contribute to our understanding of the genetic factors that may play a in the development of CRC.

5.2 Impact of these studies in medical genetics and diagnostics

5.2.1 Usages of larger panels for cancer diagnostics

The genetic landscape of hereditary cancers remains largely unknown, with the majority of cancer-causing factors still undiscovered. To effectively diagnose hereditary cancers, it is crucial to take a comprehensive approach in genetic testing that goes beyond commonly known genes. Familial cancers such as CRC and EC are often caused by a combination of low penetrance genes, which are not well defined. Testing for only a small number of genes may not be effective, particularly in complex cases with overlapping phenotypes.

Pathway-focused panels, such as targeted MMR gene panel or even larger ones such as the exome panel, provide a more comprehensive approach to identify disease-causing variants, even in less commonly studied genes. This increases the likelihood of identifying potential causes of disease, particularly in cases where the spectrum of genes involved is not well defined. As shown in Study I, the presence of germline variants in the MMR pathway genes that have not been commonly tested before may contribute to an elevated risk of EC. Study III expanded the analysis scope to the entire exome, resulting in the identification of variants in

cancer-associated genes that may play a role in the development of familial CRC, a disease with a more extensive genetic landscape beyond the MMR genes.

The implementation of expanded NGS-based gene panel analysis is crucial in the diagnosis of hereditary cancers, as it not only increases the chances of a diagnosis but also offers operational cost savings and improved efficiency. However, larger gene panels also presents the challenge of handling a large list of variants and incidental findings. This can be addressed through careful filtration using appropriate bioinformatics approaches.

5.2.2 Usages of bioinformatic tools and approaches in cancer diagnostics

The advancements in sequencing technology have led to a rapid increase in the size, depth, and complexity of sequencing data, requiring the use of bioinformatics approaches to effectively analyze this data. In studies I and III, bioinformatics pipelines were developed to tackle this challenge, utilizing existing tools and databases for variant detection, annotation, filtration, and pathogenicity prediction. These pipelines are suitable for routine diagnostic purposes that often involve a large volume of sequencing data. Study II focused on developing a bioinformatic pipeline for detecting CNVs using NGS data from targeted gene panels. The successful implementation of this pipeline at St. Olavs Hospital has expended the diagnostic lab's capabilities, allowing for the detection of partial, single, multi-exonic, and intragenic CNVs in all genes within the target panel, increasing the number of genes for which CNV detection is possible. This improvement was made possible by overcoming previous limitations imposed by the availability of MLPA kits.

5.3 Contribution of these studies towards new knowledge and resources

Outcomes of these studies contribute towards available knowledge base of hereditary EC and CRC diagnosis.

In study I, a comprehensive analysis was conducted on the genetic regions within all genes in the MMR pathway in order to identify variants that may contribute to EC. The analysis found variants in various regions, including intronic, UTR, and exonic. Including non-coding regions in the analysis enhances the possibility of discovering variants that alter gene regulation or splicing sites. The study emphasizes the significance of analyzing variants in non-coding regions and taking a comprehensive approach that considers the entire MMR pathway to increase the chances for getting a diagnosis in EC.

Study II resulted in the development of an efficient *in silico* method for CNV detection, meeting the high diagnostic accuracy, sensitivity, and specificity standards. Since August 2018, this method has been applied in routine diagnostic procedures at St. Olavs Hospital and has proven its diagnostic value with the discovery of numerous CNVs, including those in genes that were previously not tested using MLPA methods or were limited by MLPA kits availability.

In Study III, the entire protein-coding genome (i.e., the whole exome) was analyzed to identify variants in cancer-associated genes and their potential link to hereditary CRC. The analysis resulted in the identification of significant variants in genes linked to various types of cancer, including TSGs, OGs, and fusion genes. The discovery of these variants in genes not previously linked to familial CRC indicates a wider range of genetic variants that could contribute to the disease, beyond just DNA MMR genes or previously known cancer-associated genes. The study relied heavily on bioinformatics and offers a reliable diagnostic approach, particularly when analyzing large datasets like whole exome and whole genome sequencing data.

5.4 Limitation of the studies

Study I highlight the importance of including non-coding regions such as UTRs and intronic regions in the identification of potential hereditary mutations, which contribute to nearly 10% of all disease-causing mutations (126–128). However, annotating variants in these regions can be challenging due to limitations in annotation databases and tools. In a clinical setting, these variants can easily be overlooked unless RNA studies are performed to examine exon skipping, the generation of new donor sites or cryptic site activation. These limitations in annotation tools and databases make it difficult to predict the effects of most mutations in UTR and intronic regions, leading to a more stringent variant filtering compared to standard diagnostics, to reduce the number of variants to a manageable size. Despite these limitations, Study I was still able to identify ten significant intronic variants, including four in the splice site vicinity and six in deep intronic regions.

In Study I, all known genes in the MMR pathway were included in the analysis. However, the possibility of undiscovered genes and variants with similar disease effects cannot be ruled out, and could only be addressed through expanding the panel to cover the entire genome. However, this would increase the potential for noise and complexity in the analysis.

The similarity between the six of the exons in PMS2 gene in panel and its pseudogene PMS2CL created difficulties in accurate read alignment, resulting in the possibility of artifacts during variant calling. To overcome this limitation, manual checking of reads and coverage in a genomic viewer, and Sanger verification of variants, was conducted for the PMS2 gene associated variants.

Study I aimed to identify CNVs via NGS data, however, it was not possible due to limitations in data quality, such as non-uniform and low coverage depth. MLPA was also not performed

to overcome this limitation, partly due to the unavailability of MLPA kits for many genes in the panel.

In study I, 29 patients were identified to have class three variants, some of which are considered to be highly likely pathogenic in nature. However, the actual significance of these variants as causes of the disease requires additional confirmation through further studies. A key limitation in the interpretation of these class three variants is the absence of information regarding the patients' debut age of cancer and results from related tumor analyses, such as MSI status and immunohistochemical analysis of MMR genes.

Bioinformatic pipeline described by study II was validated using only 36 CNV positive control samples consisting of different types of whole gene and intrageneric CNVs in 12 different genes. The use of a larger number of positive control samples is often recommended for validation, but this was limited by the availability of known positive controls. Additional limitation, previously mentioned, was presence of pseudogene of *PMS2* gene, causing difficulties in accurate read alignment, leading to false estimation of coverage depth in genomic region of this gene. This may lead to false positive signals of CNVs generated by pipeline. Hence, all the samples are recommended to be tested though MLPA for CNV detection in *PMS2* gene. However, despite this challenge pipeline not only could detect the true CNVs in this gene in all the control sample, but also have managed to detect true signals in routine diagnostic samples tested at department of Medical Genetics at St. Olavs hospital, Trondheim. An additional limitation of this NGS-based CNV detection approach is due to the potential for artifacts to occur while sequencing of genes with smaller exons that are less than half the length of the capturing probe. This can lead to a bias towards larger intronic regions, resulting in false CNV signals, such as those frequently observed in exons (exon 2 and 5) of the *SMAD4* gene.

To address this, alternative CNV detection methods are recommended for genes with small exons.

The limitations of study III included the lack of validation of identified variants using supplementary techniques, such as Sanger sequencing or MLPA, due to sample material exhaustion. However, this was compensated for by a thorough scrutiny of the variants in Binary Alignment Map (BAM) files to confirm their validity as true positives, as NGS-based detection of SNVs and indels is known for its high accuracy.

A strict filtering criterion was applied on the variants called from whole exome regions in study III to reduce the substantial number of these variants and focus on those with potential impact on gene function. This resulted in a much smaller set of variants that passed the filters. Despite the increased chance of identifying variants with negative effects on gene function by using chosen stringent filtering criteria, it also raises the risk of missing significant variants, which may cause biases in study outcome. However, any slight adjustment towards less stringent filtering would have resulted in a minimum two-fold increase in the number of variants passing the filtering process. In this study, no significant variants were detected in known or candidate cancer-associated genes in 15 patients. This could be due to the strict filtering criteria, but there are other factors that can also contribute to a missed molecular diagnosis, such as somatic mosaicism, epigenetic inheritance, technological limitations, non-genetic risk factors, and insufficient information leading to an incomplete clinical diagnosis.

In Study III, variants in regulatory regions, such as in uORFs, were not analyzed due to limited annotation data, mainly due to the scarcity of sequencing in these regions in targeted sequencing efforts. This lack of annotation data restricts the inclusion of these variants in the analysis.

The project initially was also aimed to conduct whole genome sequencing for unsolved patient cases and set up bioinformatics workflows to analyze WGS data for disease-causing structural variants and CNVs in clinical diagnostics. However, due to logistical and time limitations, this study could not be carried out.

6. Conclusions

- Inclusion of all genes of the MMR pathway in a gene test panel provides opportunity of discovering variants in additional genes that could be associated with EC risk.
- The inclusion of non-coding parts in the sequencing target increases the chances of identifying gene regulation or splice site alteration variants, but also leads to a larger number of variants with unknown clinical significance, which are difficult to annotate.
- Low-quality data can help identify informative variants, but it should be avoided as it leads to increased noise in the analysis.
- NGS data from targeted gene panels can be used to detect CNVs *in silico* with high sensitivity, specificity, and accuracy that meets diagnostic standards.
- The CNV detection pipeline documented in Study II has shown diagnostic value and provided significant findings in routine diagnostics since its implementation.
- The implementation of the CNV detection pipeline has reduced operational costs and expanded the gene portfolio for the diagnostic lab.
- WES offers the opportunity to identify important variants across a full set of genes, but can also result in a large list of variants of uncertain significance.
- Using consensus predictions for pathogenicity by combining multiple *in silico* tools and accurate filtration strategy narrows down the large list of variants to those most likely to affect gene function.
- The outcomes of Study III suggest a wider spectrum of genes and genetic variants with an association to familial CRC, beyond the usual DNA MMR genes.

7. Future perspectives

The results of Study I provide potential new insight into the genetic basis of EC through the identification of new genes and variants that warrant further investigation. These discoveries could lead to a deeper understanding of the underlying causes of EC in the future. One avenue of future study could be the more comprehensive examination of variants in non-coding regions, which have been historically difficult to study and annotate. This line of inquiry could yield further discoveries that contribute to our knowledge of EC.

Study II demonstrated the feasibility of detecting copy number variations (CNVs) in diagnostic gene panels through the use of next-generation sequencing data. This proof-of-concept could be expanded in the future to detect CNVs in larger sequencing panels, such as whole exomes or whole genomes. Additionally, future developments could involve the integration of advanced technologies such as machine learning algorithms to improve the accuracy and efficiency of CNV detection.

The use of whole exomes in Study III resulted in the identification of a large set of variants, many of which were unknown. As prediction tools and variant databases continue to advance, it will become increasingly feasible to handle and analyze such large sets of unknown variants. The study also uncovered potentially pathogenic variants in genes that have not previously been associated with familial CRC, making it necessary to further investigate the potential involvement of these genes in CRC development. The bioinformatics approaches used in this study for variant annotation and filtration were effective in identifying important variants. Future improvements to these approaches could involve the integration of AI-based automation techniques, making the approach even more effective and potentially more suitable for diagnostic purposes.

References

1. Bray F, Laversanne M, Weiderpass E, Soerjomataram I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer. 2021 Aug;127(16):3029–30.
2. H S, J F, RL S, M L, I S, A J, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209–49.
3. Dekker E, Tanis PJ, Vleugels JLA, Kasi PM, Wallace MB. Colorectal cancer. Lancet. 2019 Oct;394(10207):1467–80.
4. Bray F. Transition in human development and the global cancer burden. World cancer Rep 2014. 2014 Jan;54–68.
5. Fidler MM, Soerjomataram I, Bray F. A global view on cancer incidence and national levels of the human development index. Int J Cancer. 2016 Dec;139(11):2436–46.
6. Lewandowska A, Rudzki G, Lewandowski T, Stryjkowska-Góra A, Rudzki S. Title: Risk Factors for the Diagnosis of Colorectal Cancer. Cancer Control. 2022 Jan;29:10732748211056692.
7. Jemal A, Siegel R, Xu J, Ward E. Cancer Statistics, 2010. CA Cancer J Clin. 2010 Sep;60(5):277–300.
8. Morice P, Leary A, Creutzberg C, Abu-Rustum N, Darai E. Endometrial cancer. Lancet (London, England). 2016 Mar;387(10023):1094–108.
9. Jenabi E, Poorolajal J. The effect of body mass index on endometrial cancer: a meta-analysis. Public Health. 2015 Jul;129(7):872–80.
10. Banno K, Yanokura M, Kobayashi Y, Kawaguchi M, Nomura H, Hirasawa A, et al. Endometrial cancer as a familial tumor: pathology and molecular carcinogenesis

(review). Curr Genomics. 2009 Apr;10(2):127–32.

11. Das S, Kundu M, Jena BC, Mandal M. Chapter 25 - Causes of cancer: physical, chemical, biological carcinogens, and viruses. In: Kundu SC, Reis RLBT-B for 3D TM, editors. Materials Today. Elsevier; 2020. p. 607–41.
12. Carbone M, Arron ST, Beutler B, Bononi A, Cavenee W, Cleaver JE, et al. Tumour predisposition and cancer syndromes as models to study gene–environment interactions. Nat Rev Cancer. 2020;20(9):533–49.
13. Jackson M, Marks L, May GHW, Wilson JB. The genetic basis of disease. Essays Biochem. 2018 Dec;62(5):643–723.
14. Sokolenko AP, Imyanitov EN. Molecular diagnostics in clinical oncology. Front Mol Biosci. 2018;5(AUG):1–15.
15. Turnbull C, Hodgson S. Genetic predisposition to cancer. Clin Med. 2005;5(5):491–8.
16. Hankey W, Frankel WL, Groden J. Functions of the APC tumor suppressor protein dependent and independent of canonical WNT signaling: implications for therapeutic targeting. Cancer Metastasis Rev. 2018;37(1):159–72.
17. Liebl MC, Hofmann TG. The Role of p53 Signaling in Colorectal Cancer. Cancers (Basel). 2021;13(9):2125.
18. Fang T, Liang T, Wang Y, Wu H, Liu S, Xie L, et al. Prognostic role and clinicopathological features of SMAD4 gene mutation in colorectal cancer: a systematic review and meta-analysis. BMC Gastroenterol. 2021;21(1):297.
19. Okuda T, Sekizawa A, Purwosunu Y, Nagatsuka M, Morioka M, Hayashi M, et al. Genetics of Endometrial Cancers. Hernández EA, editor. Obstet Gynecol Int. 2010;2010:984013.
20. Zhu G, Pei L, Xia H, Tang Q, Bi F. Role of oncogenic KRAS in the prognosis, diagnosis and treatment of colorectal cancer. Mol Cancer. 2021;20(1):143.

21. Caputo F, Santini C, Bardasi C, Cerma K, Casadei-Gardini A, Spallanzani A, et al. BRAF-Mutated Colorectal Cancer: Clinical and Molecular Insights. Int J Mol Sci. 2019;20(21):5369.

22. Dedes KJ, Wetterskog D, Ashworth A, Kaye SB, Reis-Filho JS. Emerging therapeutic targets in endometrial cancer. Nat Rev Clin Oncol. 2011;8(5):261–71.

23. Kunkel TA. Evolving Views of DNA Replication (In)Fidelity. Cold Spring Harb Symp Quant Biol . 2009 Jan;74:91–101.

24. Clark N, Wu X, Her C. MutS Homologues hMSH4 and hMSH5: Genetic Variations, Functions, and Implications in Human Diseases. Curr Genomics. 2013;14(2):81–90.

25. Amaral-Silva GK do, Martins MD, Pontes HAR, Fregnani ER, Lopes MA, Fonseca FP, et al. Mismatch repair system proteins in oral benign and malignant lesions. J Oral Pathol Med. 2017 Apr;46(4):241–5.

26. Lipkin SM, Wang V, Jacoby R, Banerjee-Basu S, Baxevanis AD, Lynch HT, et al. MLH3: a DNA mismatch repair gene associated with mammalian microsatellite instability. Nat Genet. 2000;24(1):27–35.

27. Schmidt MHM, Pearson CE. Disease-associated repeat instability and mismatch repair. DNA Repair (Amst). 2016;38:117–26.

28. Pećina-Šlaus N, Kafka A, Salamon I, Bukovac A. Mismatch Repair Pathway, Genome Stability and Cancer. Front Mol Biosci. 2020;7:122.

29. Nagy R, Sweet K, Eng C. Highly penetrant hereditary cancer syndromes. Oncogene. 2004;23(38):6445–70.

30. Garber JE, Offit K. Hereditary Cancer Predisposition Syndromes. J Clin Oncol. 2005 Jan;23(2):276–92.

31. Lu Y, Ek WE, Whiteman D, Vaughan TL, Spurdle AB, Easton DF, et al. Most common ‘sporadic’ cancers have a significant germline genetic component. Hum Mol

Genet. 2014 Nov;23(22):6112–8.

32. Rahner N, Steinke V. Hereditary Cancer Syndromes. Dtsch Arztebl Int. 2008 Oct;105(41):706–13.

33. Zhang L, Shay JW. Multiple Roles of APC and its Therapeutic Implications in Colorectal Cancer. JNCI J Natl Cancer Inst. 2017 Aug;109(8):djw332.

34. Aerts I, Lumbroso-Le Rouic L, Gauthier-Villars M, Brisse H, Doz F, Desjardins L. Retinoblastoma. Orphanet J Rare Dis. 2006;1(1):31.

35. Collard TJ, Urban BC, Patsos HA, Hague A, Townsend PA, Paraskeva C, et al. The retinoblastoma protein (Rb) as an anti-apoptotic factor: expression of Rb is required for the anti-apoptotic function of BAG-1 protein in colorectal tumour cells. Cell Death Dis. 2012;3(10):e408–e408.

36. Guha T, Malkin D. Inherited TP53 Mutations and the Li–Fraumeni Syndrome. Cold Spring Harb Perspect Med . 2017 Apr;7(4).

37. Kim WY, Kaelin WG. Role of VHL Gene Mutation in Human Cancer. J Clin Oncol. 2004 Dec;22(24):4991–5004.

38. Li X, Liu G, Wu W. Recent advances in Lynch syndrome. Exp Hematol Oncol. 2021;10(1):37.

39. Bansidhar BJ. Extracolonic Manifestations of Lynch Syndrome. Clin Colon Rectal Surg. 2012;25:103–10.

40. Lynch HT, de la Chapelle A. Hereditary Colorectal Cancer. N Engl J Med. 2003 Mar;348(10):919–32.

41. Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. Milestones of Lynch syndrome: 1895–2015. Nat Rev Cancer. 2015;15(3):181–94.

42. Bonadona V, Bonaïti B, Olschwang S, Grandjouan S, Huiart L, Longy M, et al. Cancer Risks Associated With Germline Mutations in MLH1, MSH2, and MSH6 Genes in

Lynch Syndrome. JAMA. 2011 Jun;305(22):2304–10.

43. Gazzoli I, Loda M, Garber J, Syngal S, Kolodner RD. A Hereditary Nonpolyposis Colorectal Carcinoma Case Associated with Hypermethylation of the MLH1 Gene in Normal Tissue and Loss of Heterozygosity of the Unmethylated Allele in the Resulting Microsatellite Instability-High Tumor1. Cancer Res. 2002 Jul;62(14):3925–8.

44. Win AK, Buchanan DD, Rosty C, MacInnis RJ, Dowty JG, Dite GS, et al. Role of tumour molecular and pathology features to estimate colorectal cancer risk for first-degree relatives. Gut. 2015 Jan;64(1):101 LP – 110.

45. Mao R, Krautscheid P, Graham RP, Ganguly A, Shankar S, Ferber M, et al. Genetic testing for inherited colorectal cancer and polyposis, 2021 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG). Genet Med 2021 2310. 2021 Jun;23(10):1807–17.

46. Kunitomi H, Banno K, Yanokura M, Takeda T, Iijima M, Nakamura K, et al. New use of microsatellite instability analysis in endometrial cancer. Oncol Lett. 2017;14(3):3297.

47. Resnick KE, Frankel WL, Morrison CD, Fowler JM, Copeland LJ, Stephens J, et al. Mismatch repair status and outcomes after adjuvant therapy in patients with surgically staged endometrial cancer ☆. Gynecol Oncol. 2010;117:234–8.

48. Kawaguchi M, Banno K, Yanokura M, Kobayashi Y, Kishimi A, Ogawa S, et al. Analysis of candidate target genes for mononucleotide repeat mutation in microsatellite instability-high (MSI-H) endometrial cancer. Int J Oncol. 2009 Sep;35(05):977–82.

49. Moreira L, Balaguer F, Lindor N, de la Chapelle A, Hampel H, Aaltonen LA, et al. Identification of Lynch Syndrome Among Patients With Colorectal Cancer. JAMA. 2012 Oct;308(15):1555–65.

50. Bucksch K, Zachariae S, Aretz S, Büttner R, Holinski-Feder E, Holzapfel S, et al. Cancer risks in Lynch syndrome, Lynch-like syndrome, and familial colorectal cancer type X: a prospective cohort study. BMC Cancer. 2020;20(1):460.

51. Barrow E, Hill J, Evans DG. Cancer risk in Lynch Syndrome. Fam Cancer. 2013 Jun;12(2):229–40.

52. Ferguson SE, Aronson M, Pollett A, Eiriksson LR, Oza AM, Gallinger S, et al. Performance characteristics of screening strategies for Lynch syndrome in unselected women with newly diagnosed endometrial cancer who have undergone universal germline mutation testing. Cancer. 2014 Dec;120(24):3932–9.

53. Holtzman NA. Promoting Safe and Effective Genetic Tests in the United States: Work of the Task Force on Genetic Testing. Clin Chem. 1999 May;45(5):732–8.

54. Burke W. Genetic Testing. N Engl J Med. 2002 Dec;347(23):1867–75.

55. McPherson E. Genetic Diagnosis and Testing in Clinical Practice. Clin Med Res . 2006 Jun;4(2):123–9.

56. Ponder B. Genetic Testing for Cancer Risk. Science (80-). 1997 Nov;278(5340):1050–4.

57. Mueller WF, Larsen LSZ, Garibaldi A, Hatfield GW, Hertel KJ. The Silent Sway of Splicing by Synonymous Substitutions *. J Biol Chem. 2015 Nov;290(46):27700–11.

58. Carrasco Pro S, Bulekova K, Gregor B, Labadorf A, Fuxman Bass JI. Prediction of genome-wide effects of single nucleotide variants on transcription factor binding. Sci Reports 2020 101. 2020 Oct;10(1):1–11.

59. Tseng C-C, Wong M-C, Liao W-T, Chen C-J, Lee S-C, Yen J-H, et al. Genetic Variants in Transcription Factor Binding Sites in Humans: Triggered by Natural Selection and Triggers of Diseases. Int J Mol Sci. 2021 Apr;22(8):4187.

60. Hood HM, Neafsey DE, Galagan J, Sachs MS. Evolutionary Roles of Upstream Open

Reading Frames in Mediating Gene Regulation in Fungi. Annu Rev Microbiol. 2009 Sep;63(1):385–409.

61. Whiffin N, Karczewski KJ, Zhang X, Chothani S, Smith MJ, Evans DG, et al. Characterising the loss-of-function impact of 5' untranslated region variants in 15,708 individuals. Nat Commun 2020 111. 2020 May;11(1):1–12.

62. Zhang H, Wang Y, Lu J. Function and Evolution of Upstream ORFs in Eukaryotes. Trends Biochem Sci. 2019 Sep;44(9):782–94.

63. Takahashi H, Miyaki S, Onouchi H, Motomura T, Idesako N, Takahashi A, et al. Exhaustive identification of conserved upstream open reading frames with potential translational regulatory functions from animal genomes. Sci Reports 2020 101. 2020 Oct;10(1):1–10.

64. Hughes TA. Regulation of gene expression by alternative untranslated regions. Trends Genet. 2006 Mar;22(3):119–22.

65. Pamuła-Piłat J, Tęcza K, Kalinowska-Herok M, Grzybowska E. Genetic 3′UTR variations and clinical factors significantly contribute to survival prediction and clinical response in breast cancer patients. Sci Reports 2020 101. 2020 Mar;10(1):1–15.

66. Grimson A, Farh KKH, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. Mol Cell. 2007 Jul;27(1):91–105.

67. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17(4):224–38.

68. WATSON JD, CRICK FHC. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature. 1953;171(4356):737–8.

69. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed

synthesis with DNA polymerase. J Mol Biol. 1975;94(3):441–8.

70. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci. 1977 Feb;74(2):560–4.

71. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, et al. Enzymatic Amplification of β-Globin Genomic Sequences and Restriction Site Analysis for Diagnosis of Sickle Cell Anemia. Science (80-). 1985 Dec;230(4732):1350–4.

72. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase. Science (80-). 1988 Jan;239(4839):487–91.

73. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, et al. Fluorescence detection in automated DNA sequence analysis. Nature. 1986;321(6071):674–9.

74. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921.

75. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The Sequence of the Human Genome. Science (80-). 2001 Feb;291(5507):1304–51.

76. Consortium IHGS. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931–45.

77. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376–80.

78. Augenlicht LH, Kobrin D. Cloning and screening of sequences expressed in a mouse colon tumor. Cancer Res. 1982 Mar;42(3):1088–93.

79. Malkov VA, Serikawa KA, Balantac N, Watters J, Geiss G, Mashadi-Hossein A, et al. Multiplexed measurements of gene signatures in different analytes using the Nanostring nCounter^{TM} Assay System. BMC Res Notes. 2009;2(1):80.

80. MORIN PA, MCCARTHY M. Highly accurate SNP genotyping from historical and low-quality samples. Mol Ecol Notes. 2007 Nov;7(6):937–46.

81. Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang Y-K. Ordered Restriction Maps of Saccharomyces cerevisiae Chromosomes Constructed by Optical Mapping . Science (80-). 1993 Oct;262(5130):110–4.

82. Fu W, O'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature. 2013;493(7431):216–20.

83. McVean GA, Altshuler (Co-Chair) DM, Durbin (Co-Chair) RM, Abecasis GR, Bentley DR, Chakravarti A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Oct;491(7422):56–65.

84. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. Large-scale whole-genome sequencing of the Icelandic population. Nat Genet. 2015;47(5):435–44.

85. Trotman J, Armstrong R, Firth H, Trayers C, Watkins J, Allinson K, et al. The NHS England 100,000 Genomes Project: feasibility and utility of centralised genome sequencing for children with cancer. Br J Cancer. 2022;127(1):137–44.

86. Mirkin SM. Expandable DNA repeats and human disease. Nature. 2007;447(7147):932–40.

87. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet. 2007;39(7):S37–42.

88. Stankiewicz P, Lupski JR. Structural Variation in the Human Genome and its Role in Disease. Annu Rev Med. 2010 Feb;61(1):437–55.

89. Ritz A, Bashir A, Sindi S, Hsu D, Hajirasouliha I, Raphael BJ. Characterization of structural variants with single molecule and hybrid sequencing approaches.

Bioinformatics. 2014 Dec;30(24):3458–66.

90. Chaisson MJP, Wilson RK, Eichler EE. Genetic variation and the de novo assembly of human genomes. Nat Rev Genet. 2015;16(11):627–40.

91. Chaisson MJP, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015;517(7536):608–11.

92. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biol. 2020 Feb;21(1):30.

93. Curry PDK, Broda KL, Carroll CJ. The Role of RNA-Sequencing as a New Genetic Diagnosis Tool. Curr Genet Med Rep. 2021;9(2):13–21.

94. Yépez VA, Gusic M, Kopajtich R, Mertes C, Smith NH, Alston CL, et al. Clinical implementation of RNA sequencing for Mendelian disease diagnostics. Genome Med. 2022;14(1):38.

95. Barros-Silva D, Marques CJ, Henrique R, Jerónimo C. Profiling DNA Methylation Based on Next-Generation Sequencing Approaches: New Insights and Clinical Applications. Genes (Basel). 2018;9(9):429.

96. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science (80-). 2007 Jun;316(5830):1497–502.

97. Schlaberg R. Microbiome Diagnostics. Clin Chem. 2020 Jan;66(1):68–76.

98. Goodwin S, McPherson JD, Richard McCombie W. Coming of age: ten years of next-generation sequencing technologies. Nat Publ Gr. 2016;17.

99. Levy SE, Boone BE. Next-Generation Sequencing Strategies. Cold Spring Harb Perspect Med. 2019 Jul;9(7).

100. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Hum Immunol. 2021;82(11):801–11.

101. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015 Jul;13(7):e1002195.

102. Markowetz F. All biology is computational biology. PLOS Biol. 2017 Mar;15(3):e2002050.

103. Hogeweg P. The Roots of Bioinformatics in Theoretical Biology. PLOS Comput Biol. 2011 Mar;7(3):e1002021.

104. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. Brief Bioinform. 2019 Nov;20(6):1981–96.

105. Oliver GR, Hart SN, Klee EW. Bioinformatics for Clinical Next Generation Sequencing. Clin Chem. 2015 Jan;61(1):124–35.

106. Kanzi AM, San JE, Chimukangara B, Wilkinson E, Fish M, Ramsuran V, et al. Next Generation Sequencing and Bioinformatics Analysis of Family Genetic Inheritance. Front Genet. 2020;11:544162.

107. Koboldt DC. Best practices for variant calling in clinical sequencing. Genome Med. 2020;12(1):91.

108. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing Reference Genome Assemblies. PLOS Biol. 2011 Jul;9(7):e1001091.

109. Schneider VA, Graves-Lindsay T, Howe K, Bouk N, Chen H-C, Kitts PA, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. Genome Res . 2017 May;27(5):849–64.

110. Tuteja S, Kadri S, Yap KL. A performance evaluation study: Variant annotation tools - the enigma of clinical next generation sequencing (NGS) based genetic testing. J Pathol Inform. 2022;13:100130.

111. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, et al. ACMG recommendations for standards for interpretation and reporting of sequence variations:

Revisions 2007. Genet Med. 2008 Apr;10(4):294–300.

112. Baudhuin LM, Lagerstedt SA, Klee EW, Fadra N, Oglesbee D, Ferber MJ. Confirming Variants in Next-Generation Sequencing Panel Testing by Sanger Sequencing. J Mol Diagnostics. 2015 Jul;17(4):456–61.

113. Shen Y, Wu BL. Designing a simple multiplex ligation-dependent probe amplification (MLPA) assay for rapid detection of copy number variants in the genome. J Genet Genomics. 2009 Apr;36(4):257–65.

114. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet. 2007 Jun;39(7S):S16–21.

115. Serin Harmanci A, Harmanci AO, Zhou X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. Nat Commun. 2020;11(1):89.

116. Buysse K, Delle Chiaie B, Van Coster R, Loeys B, De Paepe A, Mortier G, et al. Challenges for CNV interpretation in clinical molecular karyotyping: Lessons learned from a 1001 sample experience. Eur J Med Genet. 2009 Nov;52(6):398–403.

117. Ito T, Kawashima Y, Fujikawa T, Honda K, Makabe A, Kitamura K, et al. Rapid screening of copy number variations in STRC by droplet digital PCR in patients with mild-to-moderate hearing loss. Hum Genome Var. 2019;6(1):41.

118. Illumina. Illumina MiSeq; https://www.illumina.com/systems/sequencing-platforms/miseq.html [Internet].

119. Illumina. Illumina NextSeq; https://www.illumina.com/systems/sequencing-platforms/nextseq.html [Internet].

120. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;43(1110):11.10.1-11.10.33.

121. Alamut. Alamut-batch; https://www.interactive-biosoftware.com/alamut-batch/ [Internet]. Interactive Biosoftware, Rouen, France;

122. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122.

123. Vigeland MD, Gjøtterud KS, Selmer KK. FILTUS: a desktop GUI for fast and efficient detection of disease-causing variants, including a novel autozygosity detector. Bioinformatics. 2016 Jan;32(10):1592–4.

124. Genome-Reference-Consortium. GRCh37 - hg19 - Genome - Assembly - NCBI; https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/ [Internet]. 2009.

125. Arteche-López A, Ávila-Fernández A, Romero R, Riveiro-Álvarez R, López-Martínez MA, Giménez-Pardo A, et al. Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes. Sci Rep. 2021;11(1):5697.

126. Krawczak M, Thomas NST, Hundrieser B, Mort M, Wittig M, Hampe J, et al. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum Mutat. 2007 Feb;28(2):150–8.

127. Stenson PD, Mort M, Ball E V, Howells K, Phillips AD, Thomas NS, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009 Jan;1(1):13.

128. Cooper DN. Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes. Hum Genomics. 2010 Jun;4(5):284–8.

Articles

Article 1

PLOS ONE

RESEARCH ARTICLE

Targeted sequencing of genes associated with the mismatch repair pathway in patients with endometrial cancer

Ashish Kumar Singh[1,2]*, Bente Talseth-Palmer[1,3,4], Mary McPhillips[5], Liss Anne Solberg Lavik[1], Alexandre Xavier[3], Finn Drabløs[2], Wenche Sjursen[1,2]

1 Department of Medical Genetics, St. Olavs Hospital, Trondheim, Norway, **2** Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, NTNU—Norwegian University of Science and Technology, Trondheim, Norway, **3** School of Biomedical Science and Pharmacy, Faculty of Health and Medicine, University of Newcastle and Hunter Medical Research Institute, Newcastle, Australia, **4** Department of Research and Development, Møre og Romsdal Hospital Trust, Molde, Norway, **5** NSW Health Pathology, Molecular Medicine, John Hunter Hospital, Newcastle, NSW, Australia

* ashish.kumar.singh3@stolav.no

Citation: Singh AK, Talseth-Palmer B, McPhillips M, Lavik LAS, Xavier A, Drabløs F, et al. (2020) Targeted sequencing of genes associated with the mismatch repair pathway in patients with endometrial cancer. PLoS ONE 15(7): e0235613. https://doi.org/10.1371/journal.pone.0235613

Editor: Noel F. C. C. de Miranda, Leiden University Medical Centre, NETHERLANDS

Received: December 31, 2019

Accepted: June 19, 2020

Published: July 7, 2020

Data Availability Statement: Data contain confidential information and cannot be shared publicly as per guidelines from Hunter New England Human Research Ethics Committee. Data may be made available by request to Rodney Scott (rodney.scott@health.nsw.gov.au) for researchers who meet the criteria for access to this confidential data.

Funding: This research grant was awarded to BTP by Cancer Institute NSW (https://www.cancer.nsw.gov.au/), (Grant number: 12/ECF/2-34). The

Abstract

Germline variants inactivating the mismatch repair (MMR) genes *MLH1*, *MSH2*, *MSH6* and *PMS2* cause Lynch syndrome that implies an increased cancer risk, where colon and endometrial cancer are the most frequent. Identification of these pathogenic variants is important to identify endometrial cancer patients with inherited increased risk of new cancers, in order to offer them lifesaving surveillance. However, several other genes are also part of the MMR pathway. It is therefore relevant to search for variants in additional genes that may be associated with cancer risk by including all known genes involved in the MMR pathway. Next-generation sequencing was used to screen 22 genes involved in the MMR pathway in constitutional DNA extracted from full blood from 199 unselected endometrial cancer patients. Bioinformatic pipelines were developed for identification and functional annotation of variants, using several different software tools and custom programs. This facilitated identification of 22 exonic, 4 UTR and 9 intronic variants that could be classified according to pathogenicity. This study has identified several germline variants in genes of the MMR pathway that potentially may be associated with an increased risk for cancer, in particular endometrial cancer, and therefore are relevant for further investigation. We have also developed bioinformatics strategies to analyse targeted sequencing data, including low quality data and genomic regions outside of the protein coding exons of the relevant genes.

Introduction

Cancer is a life-threatening disease, with 18.1 million new cancer cases and 9.6 million cancer deaths worldwide in 2018 [1]. There is an increasing number of cases every year, and it has become an enormous burden to society. With longer life span, increased population and changed lifestyle, we can expect to have even more cases of cancer in the future. Among many

funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

types of cancers, incidences of endometrial cancer (EC) have increased worldwide in recent years [2], and it is currently the most common gynecological disease in western world [3]. This is the sixth most commonly diagnosed cancer and the fourteenth leading cause of death for women worldwide, with 380,000 estimated new cases in 2018 [1]. In Europe around 88,000 women get affected with EC every year, making EC the fourth most common cancer in women and tenth most common cancer among cancer related deaths [4]. With these high rates, it is important to diagnose EC at early and treatable stages. Environmental factors, changed lifestyle, high BMI, hypertension, menstrual irregularities and hormonal imbalances can play important roles towards carcinogenesis [5].

Hereditary factors also contribute towards EC. Higher incidences of EC are common among close relatives of EC patients [6]. Micro-satellite instability (MSI), due to dysfunction of the DNA mismatch repair (MMR) pathway, has frequently been reported as an oncogenic mechanism in EC [7]. The MMR system corrects replication errors, in particular single nucleotide variants and insertion-deletion (INDEL) loops, and failure in this system can result in MSI. Around ~30% of EC patients have been found with hyper-mutable phenotype and MSI [7–9] induced by dysfunctional MMR. MMR dysfunction is the cause of Lynch syndrome (LS), an autosomal dominant inherited cancer susceptibility syndrome, also known as hereditary non-polyposis colon cancer (HNPCC). LS is characterized by early-onset epithelial cancers. Individuals affected with LS have high risk of colorectal cancer (CRC) and EC, in addition to an increased risk of other epithelial malignancies like bowel, stomach, ovary, bladder, or pancreas cancer to mention a few [10]. Life-time risk of LS-affected individuals for EC is 33–61% and for CRC 40–80% [11, 12]. Not all CRC and EC with MMR deficiency are due to germline mutation, rather, most of the cases are sporadic cancers occurring due to epigenetic silencing of the MMR gene *MLH1* by DNA methylation [13–15]. It is important to identify EC cases with LS as they require regular surveillance, like colonoscopy. Given the high risk for developing new primary cancers, including CRC, this has been proven to reduce the overall mortality of the disease. If mutations in MMR genes are identified it will give the patient a diagnosis of LS and also enable at-risk relatives to be informed about their cancer risks. In addition, if pathogenic variants are identified in novel genes it could possibly explain why pathogenic variants are identified only in approximately 50% of families with a clinical diagnosis of LS (i.e. they fulfil the Amsterdam criteria) [16].

Since the rate of MSI tumours reported in EC cases is higher (30%) compared to other cancers (ie 15% in CRC), illustrating that an abnormal DNA MMR pathway plays a role in EC tumorigenesis, we decided to look into a more extended set of genes than those known to be involved in LS (*MLH1*, *MSH2*, *MSH6*, *PMS2* and deletions in *EPCAM1*). In the present study, 22 genes (both coding and noncoding parts) involved in the MMR pathway were sequenced in DNA from 199 sporadic EC patients. Targeted next generation sequencing (NGS) was used, aiming to identify novel genetic variation like substitutions, insertions/deletions (indels) and structural alterations (e.g. copy number variations) that may lead to the multi-step process of carcinogenesis.

Materials and methods

The study was performed on DNA extracted from full blood from 199 patient samples from a study which included consecutively recruited women with histologically confirmed EC (sporadic cases) who presented for treatment at the Hunter Centre for Gynaecological Cancer, John Hunter Hospital, Newcastle, New South Wales, Australia between the years 1992 and 2005 [17]. Blood samples were taken in year 2005 for the present study. The study has been approved by Hunter New England (HNE) Human Research Ethics Committee (HNE HREC: 05/03/09/3.14). Written informed consent was obtained from all participants.

Targeted next generation sequencing (NGS)

Targeted NGS sequencing was performed on the 199 patient samples, using an Illumina MiSeq [18] instrument. Initially 12 runs were performed to sequence the samples; later 15 samples were re-sequenced due to low quality of the initial sequencing. The target regions (all introns, exons, 5' and 3' UTRs) of 22 MMR genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*, *MSH3*, *PMS1*, *MLH3*, *EXO1*, *RFC1*, *RFC2*, *RFC3*, *RFC4*, *RFC5*, *PCNA*, *LIG1*, *RPA1*, *RPA2*, *RPA3*, *POLD1*, *POLD2*, *POLD3* and *POLD4*) with a total size of 1.213 Mb were captured using 6961 probes and the Illumina Nextera Rapid Capture Enrichment Kit (custom, 96 samples). An overview of these 22 genes, their function and associated phenotypes are shown in Table 1. Sequencing was performed at the Medical Genetics Laboratory at Hunter Medical Research Institute (HMRI), University of Newcastle, Australia.

Bioinformatic analysis

Raw reads (.fastq files) generated by the sequencer were processed by the following three major steps:

1. *Data pre-processing*: Raw reads were aligned to the reference genome (version hg19), and sequence alignment maps were generated. These alignment maps were used for read visualization and to call variants.
2. *Variants discovery*: The alignment maps generated from previous steps were compared against the reference genome to generate a list of nucleotide variants.
3. *Variants annotation*: Variants were annotated using different databases and tools.

A pipeline was constructed to perform the above-mentioned steps of analysis. Detailed overview of pipeline and tools used can be found as S1 File. Schematic overview of the pipeline is shown in Fig 1.

Filtration of variants

All called variants were annotated by using Alamut-batch [19] before filtering. Filtus [20] was used for filtering variants. All variants were classified into 4 region-wise categories; exons, UTRs, introns, and splice sites (variant distance $\leq$ 10 nucleotides from nearest splice site). In the first stage of filtering, variants from all these four regions were filtered based on frequencies of variants in the gnomAD database [21]. Exonic variants, intronic variants, and variants near splice sites were filtered-in for frequencies less than 0.1% (or no frequency). UTR variants were filtered-in for frequencies less than 0.01% (or no frequency). In further stages of filtering, different strategies were adopted for every region. See Fig 2 for the workflow. Detailed filtering steps can be found in S2 File.

Validation of variants

Sanger sequencing was performed for validation of selected variants. The fragments were amplified using AmpliTaq Gold® 360 MasterMix and 360 GC Enhancer (Life Technologies). Cycle sequencing reaction was performed with BigDye® Terminator v3.1 (Life Technologies) and subsequent capillary electrophoresis was performed on the ABI 3130xl or ABI 3730 (Life Technologies). List of primer sequences can be provided upon request. Sanger sequencing data was analysed using SeqScape Software v3.0 (Life Technologies). Some variants have not been verified by Sanger sequencing, partly due to unavailability of primers for some of these gene, but also due to logistic issues. But variants were thoroughly inspected in BAM files to

Table 1. List of genes in target panel.

Gene		Gene function (Info source: NCBI-gene)	Phenotype (Info source: OMIM)	OMIM ID
MLH	*MLH1*	Encodes a protein which heterodimerizes with MMR endonuclease PMS2 to form MutL alpha.	Colorectal cancer, hereditary nonpolyposis, type 2; Muir-Torre syndrome; Mismatch repair cancer syndrome	120436
	MLH3	Member of the MutL-homolog (MLH) family, maintains genomic integrity during DNA replication and after meiotic recombination.	Colorectal cancer, hereditary nonpolyposis, type 7; Colorectal cancer, somatic; Susceptibility to Endometrial cancer	604395
MSH	*MSH2*	Forms 2 different heterodimers: MutS alpha (MSH2-MSH6 heterodimer) and MutS beta (MSH2-MSH3 heterodimer) which binds to DNA mismatches to initiate DNA repair	Colorectal cancer, hereditary nonpolyposis, type 1; Muir-Torre syndrome; Mismatch repair cancer syndrome	609309
	MSH3	Forms a hetero-dimer with MSH2 to form MutS beta which forms a complex with MutL alpha heterodimer and initiates mismatch repair by binding to a mismatch.	Endometrial carcinoma, somatic; Familial adenomatous polyposis 4	600887
	MSH6	A component of the post-replicative DNA MMR system. Heterodimerizes with MSH2 to form MutS alpha, which binds to DNA mismatches to initiate DNA repair.	Colorectal cancer, hereditary nonpolyposis, type 5; Endometrial cancer, familial; Mismatch repair cancer syndrome	600678
PMS	*PMS1*	Forms heterodimers with MLH1. Encoded protein belongs to the DNA MMR mutL/hexB family.	Hereditary nonpolyposis colorectal cancer type 3 (HNPCC3); Lynch syndrome	600258
	PMS2	Forms MutL-alpha heterodimer (MLH1-PMS2 hetetrodimer) which activates endonucleolytic activity following recognition of mismatches and insertion/deletion loops by the MutS-alpha and MutS-beta heterodimers.	Colorectal cancer, hereditary nonpolyposis, type 4; Mismatch repair cancer syndrome	600259
	EXO1	Encodes a protein with 5' to 3' exonuclease and RNase H activities. Similar to the Saccharomyces cerevisiae protein Exo1 which interacts with Msh2 for MMR.		606063
RFC	*RFC1*	Encodes large subunit of replication factor C, a 5 subunit DNA polymerase accessory protein (DNA-dependent ATPase required for eukaryotic DNA replication and repair).		102579
	RFC2	Encodes 40-kD subunit, responsible for binding ATP and may help promote cell survival	Disruption of this gene is associated with Williams syndrome	600404
	RFC3	Encodes 38-kD subunit, responsible for binding ATP and may help promote cell survival		600405
	RFC4	Encodes 37-kD subunit, responsible for binding ATP and may help promote cell survival		102577
	RFC5	Encodes 36.5-kD subunit, responsible for binding ATP and may help promote cell survival		600407
	PCNA	Encodes a protein which acts as a homotrimer and helps increase the process of leading strand synthesis during DNA replication, also involved in the RAD6-dependent DNA repair pathway	ataxia-telangiectasia-like disorder-2 (ATLD2)	176740
	LIG1	Encodes a member of the ATP-dependent DNA ligase protein family, which functions in DNA replication, recombination, and the base excision repair process.	Mutations in gene leads to ligase-I deficiency resulting in immunodeficiency and increased sensitivity to DNA-damaging agents associated with variety of cancers	126391
RPA	*RPA1*	Encodes the subunit of heterotrimeric Replication Protein A (RPA) complex, which binds to single-stranded DNA, forming a nucleoprotein complex. Complex is involved in DNA metabolism, replication, repair, recombination, telomere maintenance.	knockdown of RPA1 in HeLa cells caused accumulation of cells in S and G2/M phases, followed by cell death	179835
	RPA2	Same as above		179836
	RPA3	Same as above		179837
POLD	*POLD1*	Encodes a catalytic subunit of DNA polymerase delta, which possesses both polymerase and 3' to 5' exonuclease activity, important for DNA replication and repair.	Colorectal cancer, Susceptibility to, CRC-10; CRCS10; Mandibular hypoplasia, deafness, progeroid features, and lipodystrophy syndrome	174761
	POLD2	Encodes 50-kDa catalytic subunit of DNA polymerase delta which possesses both polymerase and 3' to 5' exonuclease activity and plays a critical role in DNA replication and repair.	Expression of this gene may be a marker for ovarian carcinomas	600815
	POLD3	Encodes the 66-kDa subunit of DNA polymerase delta.		611415
	POLD4	Encodes the smallest subunit of DNA polymerase delta POLDS-P12.		611525

https://doi.org/10.1371/journal.pone.0235613.t001

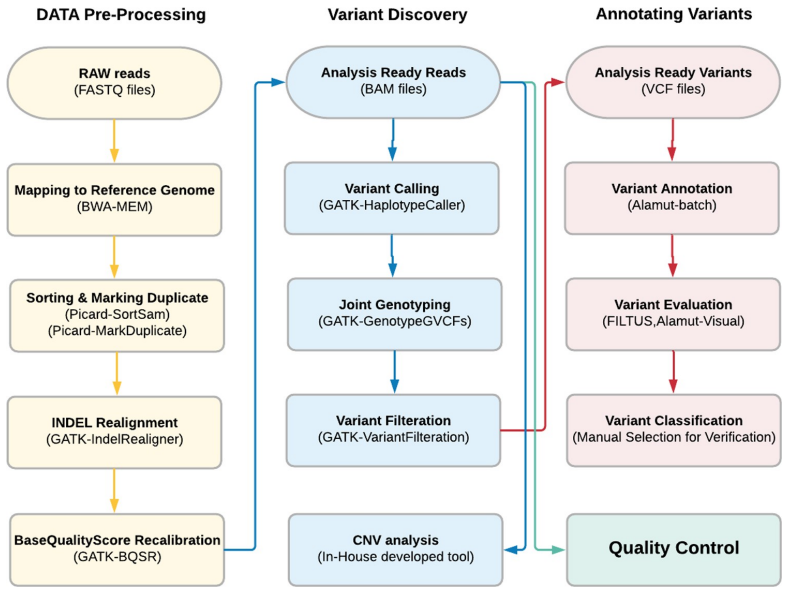


Fig 1. Schematic overview of the bioinformatics pipeline.

https://doi.org/10.1371/journal.pone.0235613.g001

assure they were likely to be true positive variants (enough coverage and an allele fraction of about 50%, between 30 and 75%).

Interpretation and classification of DNA variants

The remaining variants after filtering were classified into 5 classes according to the American College of Medical genetics (ACMG) guidelines [22]. To determine whether these variants had

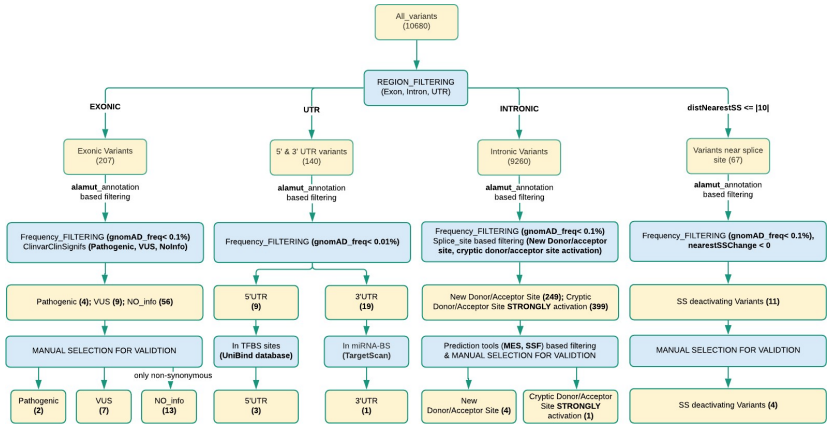


Fig 2. Filtering workflow and number of genetic variants detected.

https://doi.org/10.1371/journal.pone.0235613.g002

been detected before, literature and databases including LOVD/InSIGHT (https://www.insight-group.org/variants/databases/) and ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/) were searched. Potential pathogenicity of missense variants was interpreted using Alamut batch (annotation) [19] and Alamut Visual (interpretation) [23].

Results and discussion

From all 199 samples, on average 99.8% of reads (per run) could be aligned to the reference genome (hg19) using BWA for alignment (see S1 File). Coverage depth of reads for samples and mean coverage depth for runs varied a lot among the 12 runs. Only 23 samples had a coverage of more than 100X (maximum 169X), and 50 samples had coverage of less than 30X (minimum 1X) (see S1 Table). Despite having multiple samples with low quality, the strategy for variant calling was uniformly applied to all samples. This was done to investigate the potential for identifying true variants even from target regions with low coverage depth. However, these low-quality data were not suitable for identification of copy number variants, and therefore CNV calling was not included in the final analysis.

In total 10,680 unique variants (substitutions and INDELs) were called using the GATK toolkit. These variants could be classified into four categories according to genomic region; exonic, intronic, UTRs and splice-site neighbourhood ($\leq |10|$bp). See Fig 2 for the workflow. After filtering and annotation, **22** exonic, **9** intronic (4 variants in splice-site neighbourhood) and **4** UTR variants (Fig 3) were selected for further investigation for pathogenicity as potential cancer risk variants, and these variants are described below. See Table 2 for an aggregate list. Sanger verification was performed for 21 of these 35 variants. Remaining 14 variants are not Sanger validated. These 14 variants were designated as true variants by observing BAM files.

Exonic variants

A total of 207 variants were called in exonic regions of the target panel, over all samples. The variants were filtered by removing cases according to their frequency in gnomAD ($> 0.1\%$) and annotation in ClinVar (benign/likely-benign) [49]. Of the 22 exonic variants (S3 File) that remained after filtering, there were 2 putative pathogenic variants, 7 variants of unknown significance (VUS) and 13 variants without any information (NO_Info) according to ClinVar (only non-synonymous variants).

Among these 22 variants there were 2 variants in *MLH1* (NM_000249.3). Both *MLH1* variants were classified as class 3 in pathogenicity, according to ACMG guidelines [22]. The variant c.453G>A p.(Thr151 =) is found in the last nucleotide of exon 5. It may alter the ligation of adjacent exons 5 and 6 and is predicted to be splice site deactivating by prediction tools (SSF [50], MES [51]) (nearest-SS-change score: -0.29). The first and the last three positions of the exon are an integral part of the 3' and 5'splice site consensus sequences [52], the variant position is highly conserved (PhastCons score: 0.99), and predicted as pathogenic by UMD-predictor [53]. According to ClinVar it is classified as a likely pathogenic / VUS variant, with multiple submissions in ClinVar where many of them has a HNPCC/Lynch syndrome phenotype. With strong evidences for being a pathogenic variant, it is a candidate for further RNA/functional studies. The variant c.2009A>G p.(Lys670Arg) has no frequency in the gnomAD database, but has recently been reported in ClinVar (as VUS) and in other databases. This variant has been associated with a HNPCC phenotype and hereditary cancer-predisposing syndrome, according to ClinVar. The variant position is highly conserved (PhastCons: 1, phyloP: 4.6) and lies in a helix secondary structure of the protein. It has been predicted as pathogenic (UMD-prediction, MutationTaster).

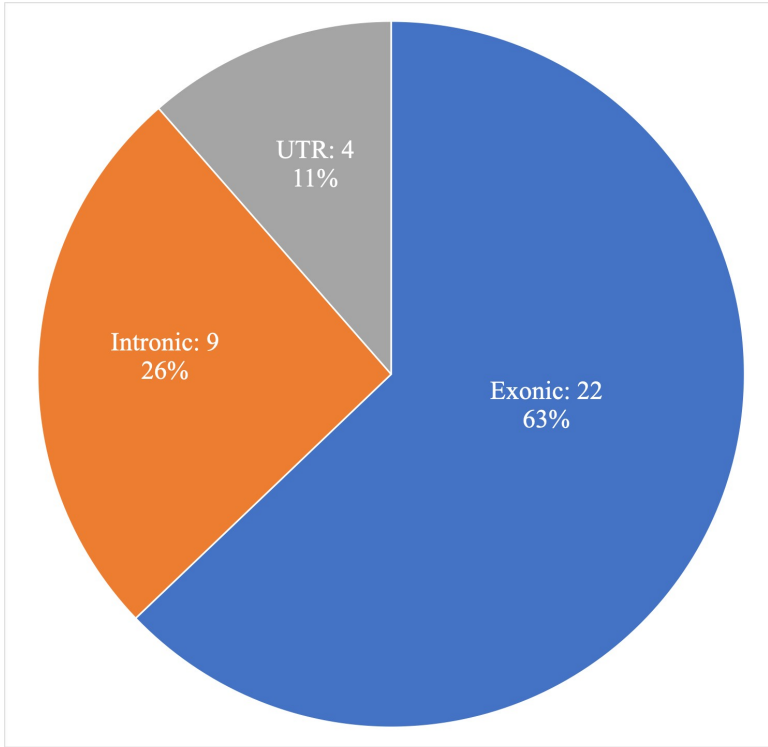


Fig 3. Investigated variants in different genomic regions.

https://doi.org/10.1371/journal.pone.0235613.g003

There were also four exonic variants in *MSH6* (NM_000179.2). Two of the variants, c.335A>G p.(Asn112Ser) and c.2203C>A (p.Leu735Ile), have been classified as VUS by ClinVar. These two variants have previously been associated with Lynch syndrome, HNPCC and hereditary cancer-predisposing syndrome-like phenotype according to ClinVar and other databases. Both variants are at highly conserved positions and both have been predicted as damaging by prediction tools. We classified these two variants as class 3. Variants c.1409C>G p.(Ser470*) and c.3802-4_3825dup p.(Glu1276*) both code for "STOP gain" and are disease causing. None of these variants have entries in ClinVar or frequency in gnomAD. We classify these as class 5 variants.

Three exonic variants in *MSH2* was identified (NM_000251.2), c.97A>C p.(Thr33Pro), c.1228G>T p.(Gly410Cys) and c.2732T>G p.(Leu911Arg), with all three classified as VUS by ClinVar. All three have phenotypic association to Lynch syndrome/HNPCC and hereditary cancer-predisposing syndrome and have been predicted as pathogenic/disease-causing by many prediction tools (UMD-prediction, PolyPhen, SIFT and MutationTaster). All three variant positions are highly conserved (with high scores in PhastCons and phyloP). Variant c.97A>C p.(Thr33Pro) was identified from a low quality sample (coverage depth at variant position 4X and sample coverage 7X), but was verified as a true variant by Sanger sequencing. It has been scored with a high value for decreasing protein stability (SNPs3D [54] score: -1.08) and has been suggested as a cause of reduced mismatch binding/release efficiency compared

Table 2. Aggregate list of variants and their classification according to the ACMG system.

Sample ID	Other Cancers[c]	Family history of Cancer[d]	Gene[a]	gNomen, cNomen, pNomen, rsID	Variant allele fraction	ClinVar / gnomAD	Class	Comments[b]	References
051783	BC	2nd* BC	*MLH1*	Chr3:g.37042548A>G, NM_000249.3: c.306+4A>G, p.(?), rs267607733	0.35	6 x VUS /0.0012%	Class 3	Activation of a cryptic donor site and the skipping of exon 3 in an ex vivo splicing minigene assay	[24]
060337	BC	NO	*MLH1*	Chr3:g.37090414A>G, NM_000249.3: c.2009A>G, p.(Lys670Arg), rs905983196	0.49	3 x VUS / NIL	Class 3	SS: Helix	
051456	NO	NO	*MSH6**	Chr2:g.48010297:G>T, NM_000179.2: c.-76G>T, p.(?)	0.64	NIL/ NIL	Class 3	No frequency, highly conserved	
051026	SC	NO	*MSH6*	Chr2:g.48018140A>G, NM_000179.2: c.335A>G, p.(Asn112Ser), rs587779934	0.44	6 X VUS /0.0025%	Class 3	New acceptor site predicted SS: Turn	
051408	NO	2nd* CRC	*MSH6*	Chr2:g.48025743C>A, NM_000179.2: c.628-7C>A, p.(?), rs373129248	0.41	6 x VUS /0.0093%	Class 3		
051476	NO	1st* EC	*MSH6*	Chr2:g.48026531C>G, NM_000179.2: c.1409C>G, p.(Ser470*)	0.48	NIL/ NIL	**Class 5**	SS: Helix	
051791	NO	1st* EC	*MSH6*	Chr2:g.48027325C>A, NM_000179.2: c.2203C>A, p.(Leu735Ile), rs786204071	0.4	6 X VUS / NIL	Class 3		
051280	NO	NO	*MSH6*	Chr2:g.48032048G>T, NM_000179.2: c.3439-1G>T, p.(?), rs587779263	0.5	8 X Pathogenic / NIL	**Class 5**		
060162	NO	2nd* OC	*MSH6*	Chr2:g.48033587_48033614dup, NM_000179.2:c.3802-4_3825dup, p.(Glu1276*)	0.46	NIL/ NIL	**Class 5**		[25]
051233	NO	1st* BC	*MSH2*	Chr2:g.47630427A>C, NM_000251.2: c.97A>C, p.(Thr33Pro), rs63751107	0.75	6 X VUS / 0.0056%	Class 3	SS: Beta strand	[26, 27, 28–35, 36]
			*MSH3**	Chr5:g.79968115C>T, NM_002439.4: c.845C>T, p.(Thr282Ile), rs202184623	0.67	NIL// 0.0053%	Class 3	SS: Beta strand	
051872	3 Melanomas	NO	*MSH2*	Chr2:g.47657032G>T, NM_000251.2: c.1228G>T, p.(Gly410Cys), rs587782242	0.47	1 X VUS/ NIL	Class 3	SS: Helix	
051107	Melanoma	NO	*MSH2*	Chr2:g.47672680C>A, NM_000251.2: c.1277-7C>A, p.(?), rs375437307	0.57	3 X VUS/ 0.0037%	Class 3		
051271	BC	1st* BC & PC, 2nd* CRC	*MSH2*	Chr2:g.47710015T>G, NM_000251.2: c.2732T>G, p.(Leu911Arg), rs41295182	1	1 X VUS gnomAD: 0.0062%	Class 3	SS: Helix	[34, 37–39]
051179	SKIN SPOTS	1st* CRC	*PMS2*	Chr7:g.6045549C>A, NM_001322014.1:c.137G>T, p.(Ser46Ile), rs121434629	0.44	12 X VUS:/ 0.0169%	**Class 4**	Associated with diagnosis of CMMRD syndrome, SS: Helix	[40–47]
051300	NO	1st* CRC & PC			0.47				
051657	NO	1st* OC	*MSH3*	Chr5:g.80021325A>G, NM_002439.4: c.1394A>G, p.(Tyr465Cys), rs35009542	0.58	NIL/ 0.0202%	Class 3	SS: Helix	
051172	NO	NO	*MSH3**	Chr5:g.79974804G>A, NM_002439.4: c.1232G>A, p.(Arg411His), rs764885728	0.49	NIL/ 0.0012%	Class 3		
051469	PCOS	2nd* UC	*MSH3**	Chr5:g.80021327A>G, NM_002439.4: c.1396A>G, p.(Ser466Gly), rs766948921	0.51	NIL/ 0.0025%	Class 3	SS: Helix	
060161	NO	NO	*MSH3**	Chr5:g.80063896C>T, NM_002439.4: c.2041C>T, p.(Pro681Ser), rs115198722	0.48	NIL/ 0.0787%	Class 3	SS: Helix	
051330	BC, SpC, LC, KC AND LiC	NO			0.55				
051610	BrT	NO	*POLD1*	Chr19:g.50905980G>A, NM_001308632.1:c.952G>A, p.(Glu318Lys), rs775232133	0.63	1 X VUS / NIL	Class 3	highly conserved DI**E**	

(*Continued*)

Table 2. (Continued)

Sample ID	Other Cancers[c]	Family history of Cancer[d]	Gene[a]	gNomen, cNomen, pNomen, rsID	Variant allele fraction	ClinVar / gnomAD	Class	Comments[b]	References
051406	BC	1st* BC	*RFC1*	Chr4:g.39290383A>T, NM_001204747.1:c.3445T>A, p.(*1149Argext*15), rs149767968	0.59	NIL/ 0.0065%	Class 3	Altered stop codon, extension of protein with 15 aa	
			RPA3	Chr7:g.7676702A>G, NM_002947.4: c.295T>C, p.(Tyr99His)	0.62	NIL/ 0.0004%	Class 3	SS: Helix	
051663	NO	1st* BC	*RFC1**	Chr4:g.39306530C>A, NM_001204747.1:c.2017G>T, p.(Val673Leu), rs28903096	0.33	NIL/ 0.057%	Class 3		
051400	NO	1st* Unknown CANCER	*RFC1**	Chr4: g.39346049A>CNM_001204747.1:c.208 +972T>G, p.(?)	0.63	NIL/ NIL	Class 3		
051471	NO	NO	*RFC3*	Chr13:g.34392210:A>G, NM_002915.3: c.-106A>G, p.(?), rs554574193	0.57	NIL/ 0.0064%	Class 3	highly conserved	
051640	NO	2nd* CRC	*RFC4*	Chr3:g.186524157:G>A, NM_002916.3: c.-90C>T, p.(?)	0.52	NIL/ NIL	Class 3	**not** conserved	
051439	NO	NO	*RFC4**	Chr3:g.186518351T>C,NM_002916.3: c.210+555A>G, p.(?), rs781729102	0.55	NIL/ 0.0387%	Class 3		
051802	UC	1st* EC	*LIG1*	Chr19:g.48640874G>A, NM_000234.2: c.1159C>T, p.(Arg387Cys), rs749929415	0.48	NIL/ 0.0018%	Class 3	SS: Beta strand	
051133	BC & OC	1st* BC	*LIG1**	Chr19:g.48653350A>C, NM_002439.4: c.692T>G, p.(Phe231Cys), rs767343361	0.37	NIL/ 0.0079%	Class 3	SS: Turn	
051166	BC & OC	1st* BC	*MLH1*	Chr3:g.37048554G>A, NM_000249.3: c.453G>A, p.(Thr151 =), rs369521379	0.51	9 x VUS /0.0011%	Class 3	Last nucleotide of exon 5. ClinVar Miner: damage the nearby splice donor site (at -1 distance) and cause abnormal splicing. SS: Beta strand	
			*EXO1**	Chr1:g.242020650G>T, NM_006027.4: c.409G>T, p.(Ala137Ser), rs147663824	0.55	NIL/ 0.0094%	Class 3	SS: Helix	[48]
			*RPA3**	Chr7:g.7753847G>T, NM_002947.4:c.-1028+959C>A, p.(?)	0.51	NIL/ NIL	Class 3		
051267	NO	1st* PCOS	*EXO1**	Chr1:g.242052986T>G, NM_130398.3: c.*84T>G, p.(?)	0.5	NIL/ NIL	Class3	Can affect miRNA binding, for miR-370-3p and miR-93-3p	
051007	NO	NO	*RPA1**	Chr17:g.1785509A>G,NM_002945.4: c.1241+1524A>G, p.(?), rs536796524	0.43	NIL/ 0.0323%	Class 3	It is within 1000 bp of a region that may be important for chromatin folding (Insulator / CTCF / SMC3 / RAD21)	
051291	NO	NO	*RPA3**	Chr7:g.7695875T>C,NM_002947.4:c.-757-15069A>G, p.(?), rs946965390	0.54	NIL/ NIL	Class 3	It is within 2500 bp of a region that may be important for chromatin folding (Insulator / CTCF / SMC3 / RAD21)	

[a]Variants with

* not yet verified by Sanger sequencing;

[b]SS: Variant lies in Secondary Structure (UniProt)

[c]Other cancers: BC: Breast Cancer, CRC: Colorectal Cancer, PCOS: Polycystic Ovary Syndrome, PC: Prostate Cancer, UC: Uterine Cancer, EC: Endometerial Cancer, OC: Ovarian Cancer, SC: Skin Cancer, SpC: Spine Cancer, LC: Lung Cancer, KC: Kidney Cancer, LiC:Liver Cancer, BrT: Brain Tumor

[d]Family history of cancer: 1st* & 2nd*: 1st & 2nd Degree relatives with cancer-type

https://doi.org/10.1371/journal.pone.0235613.t002

to wild-type protein in previous studies by Ollila *et al*. [26, 27]. Variant c.1228G>T p.(Gly410Cys) has no frequency in gnomAD, but has been reported to ClinVar and is located in a helix secondary structure of the protein. It has a high score for structural change (Grantham-Distance: 159), but it is predicted not to alter protein stability (SNPs3D: +3.43). Variant c.2732T>G p.(Leu911Arg) was also identified in a low quality sample (coverage depth at variant position 5X, sample 10X). It lies in a helix secondary structure, has high score for structural change (Grantham-Distance: 102) and for decreased protein stability (SNPs3D: -1.08). These three *MSH2* variants have been classified as class 3.

A missense exonic variant in *PMS2* was also detected (NM_001322014.), c.137G>T p.(Ser46Ile), which was found in two samples. It was classified as likely pathogenic according to ClinVar, and is reported to be a founder mutation [55]. The protein region has helix-like secondary structure (UniProt [56]), and the position is highly conserved (PhastCons [57] score:1; phyloP [58] score: 6.178). It has been classified as pathogenic by several prediction tools (UMD-predictor, PolyPhen [59], SIFT [60], and MutationTaster [61], the variant has been referred to in many previous studies, and it has been considered for strongly decreased DNA mismatch repair activity. This variant was classified as class 4.

One exonic variant in *POLD1* was identified (NM_001308632.1), c.952G>A p.(Glu318Lys), was classified as VUS according to ClinVar. It was called in a low-quality sample (coverage depth at variant position 11X, sample's mean coverage depth 22X). The position is highly conserved (PhastCons: 1, phyloP:3.9). The variant is in the DNA binding cleft of the exonuclease active domain of *POLD1*, it has a high score for decreased protein stability (SNPs3D: -2.68), and is predicted as damaging by prediction tools. A previous study has predicted it to be disease causing [62]. However, functional studies are needed to confirm pathogenicity, and therefore it was classified as class 3.

Exonic variants were also found in five other genes; *MSH3*, *LIG1*, *RFC1*, *EXO1* and *RPA3*. All these variants were classified as class 3. In *MSH3* (NM_002439.4) variants were c.1394A>G p.(Tyr465Cys), c.845C>T p.(Thr282Ile), c.1232G>A p.(Arg411His), c.1396A>G p.(Ser466Gly), c.2041C>T p.(Pro681Ser). In *LIG1* (NM_000234.2) variants were c.1159C>T p.(Arg387Cys) and c.692T>G p.(Phe231Cys). In *RFC1* (NM_001204747.1) this was c.3445T>A p.(*1149Argext*15), which introduces a "STOP loss" and extension of 15 amino acids in the product protein and c.2017G>T p.(Val673Leu); in *EXO1* (NM_006027.4) c.409G>T p.(Ala137Ser); In *RPA3* (NM_002947.4) it was c.295T>C p.(Tyr99His).

Intronic variants

Among all detected variants, 9,260 were identified as intronic, which was ~97% of all variants. Intronic regions of human DNA, being extraordinarily larger in comparison to other regions, it is expected to find most of the variants in these non-coding regions. After frequency-based filtering (< 0.1%), this list was reduced to 4,197 variants, which was further reduced by splice site related filtering, using strict filtering criteria to reduce the large number of variants. These variants were filtered for two categories, first for "New Donor/Acceptor site" and then for "Cryptic Donor/Acceptor Site STRONG activation" (see S2 File for filtering details). We found in total five variants, with four variants in the first category and one in the second (see S3 File). Two of these variants were in *RPA1* and *RPA3*, and have been predicted as new acceptor site, two were in *RFC1* and *RFC4* and have been predicted as new donor sites, and one was in *RPA3* and has been predicted to give strong activation of a cryptic donor site.

According to the Human Gene Mutation Database (HGMD) more than 10% of all disease-causing hereditary mutations are splice site altering [63–65]. Variants in vicinity of exon-intron junctions were therefore studied. After filtering (see Supporting Material), we found

four variants of interest in the vicinity of splice sites (see S3 File). Two of these were in *MSH6* (NM_000179.2). The variant c.628-7C>A has been classified as VUS by us and ClinVar, The variant c.3439-1G>T, at the last nucleotide of 5th intron, has been classified as pathogenic by ClinVar, it has been linked to LS/HNPCC phenotype, and has a maximum score (-1) for splice-site deactivation. We classified it as class 5 and hence disease causing.

An intronic variant in *MLH1* (NM_000249.3: c.306+4A>G) is found close to a splicing junction and was predicted for splice site deactivation. It is in a highly conserved position (PhastCons:1, phyloP:4.2), and has been classified as VUS in ClinVar. Experimental studies have shown that this variant results in the activation of a cryptic donor site and skipping of exon 3 in an ex-vivo splicing minigene assay [24], but as no studies have verified this in patient samples, we classified it as class 3 variant. An intronic variant in *MSH2* (NM_000251.2:c.1277-7C>A), previously classified as likely benign, we classified as a class 3 variant.

UTR variants

There were 140 variants identified in UTR regions. Due to limitations of annotation tools and databases, any effects of most mutations in these regions are hard to predict. Hence, a relatively strict filtering compared to standard (for diagnostics) [66] was used for variants in these regions, to reduce the number of variants to a manageable size. After frequency-based filtering (< 0.01%) this list reduced to 28 variants, of which 9 variants were in 5' and 19 variants were in 3' UTR.

Variants in 5' UTR were annotated for transcription factor binding sites (TFBSs), using the UniBind database [67]. Among the 9 variants in 5' UTR, three had significant hits in the database, where each of these three variants was found inside a potential binding site for at least one transcription factor (TF) according to UniBind data (see S2 Table). One variant in *MSH6* (NM_000179.2: c.-76G>T) had overlap with potential binding sites for the TFs CTCF, STAT3, E2F7 and E2F1. For CTCF there is a high frequency of the reference allele (G) compared to the alternate allele (T) at the variant position, which can indicate a strong preference for the reference variant, and possibly a significant effect of the alternate variant on TFBS specificity (frequency matrices from the JASPAR database [68, 69] were used for this analysis). According to ChIP-seq data visualized with the UCSC genome browser [70] there are relatively strong signals for CTCF at this position (see S1 Fig) compared to other potential TFs. Mutations in CTCF binding sites have for example been associated with chromosomal instability and aberration and have been found in gastric and colorectal cancer [71], which strengthens the possibility that this variant may have an effect through altered binding of CTCF. A variant in *RFC3* (NM_002915.3: c.-106A>G) had hits for the TFs GABPA, JUN, CREM, JUND, ATF1, MITF, NR3C1, ATF7 and CREB1. Among these hits, 6 TFs (JUN, CREM, JUND, R3C1, ATF7 and CREB1) had a very high frequency of the reference allele (A) compared to the alternate allele (G) at the variant position. ChIP-seq data shows strong signals for CREB1 (see S2 Fig.), which may indicate a potential for significant effects due to alteration in the binding site. A variant in *RFC4* (NM_002916.3: c.-90C>T) had a hit for the TF AR.

Nineteen variants in 3'UTR were annotated using TargetScan v6.2 [72] and a two-step SVM prediction of micro-RNA (miRNA) target sites [73]. A SVM score normalization method [74] was used to normalize the score and miRNA data were taken from MirBase v22 [75]. Only a variant in gene *EXO1* (NM_130398.3:c.*84T>G) was predicted as a likely true candidate for affecting miRNA binding, for miR-370-3p and miR-93-3p (see S3 Table). Several studies have shown the importance of *EXO1* in replication, DNA repair pathways, cell cycle checkpoints and its association to cancer [76], and GWAS studies have identified specific mutations in *EXO1* gene as risk alleles for different types of cancer [77, 78]. SNPs in miRNA

binding sites have been associated with CRC [79]. For the two miRNAs predicted to be affected by variation in their binding site, miR-370-3p has been identified as a tumour suppressor in EC via endoglin regulation [80]. The miR-93-3p can be considered as an important factor for CRC suppression and inhibition of tumorigenesis [81], as a previous study has associated the down-regulation of miR-93 with unfavourable clinicopathologic features and short overall survival of CRC patients [82].

Implications of the study

In this study we found 35 significant variants (22 exonic, 4 UTR, 9 intronic), with 15 variants in the 4 MMR genes known to cause LS (*MLH1*, *MSH2*, *MSH6*, *PMS2*) and 20 in the additional MMR genes included in this study (*MSH3*, *POLD1*, *RFC1*, *RFC3*, *RFC4*, *LIG1*, *EXO1*, *RPA1*, *RPA3*). This helped in identification of variants in less studied genes, as well as polygenic variations (although none of the 199 samples in this particular study showed polygenic variants of interest for further investigations). This study also used the complete genomic regions of the genes, which very few previous studies have done [83, 84].

Though all known genes in the MMR pathway were studied, there will always be a possibility of additional genes and associated variants with similar disease effects, e.g., *POLE* mutations in EC cases contributing towards Polymerase Proofreading Associated Polyposis (PPAP) [85, 86], or germline deletions in another gene (*EPCAM1)* leading to silencing of the MSH2 gene, causing Lynch syndrome [87]. These limitations can only be removed by expanding the panel by including more genes, up to the extent of the whole genome. However, this will also increase the potential of noise and complexity of the analysis, by including more genes and variants that are less likely to be relevant in a given study. Another limitation is associated with the *PMS2* gene in this panel, which has a pseudo-gene (*PMS2CL*), and where 6 exons (exon 9, 11, 12, 13, 14 & 15) are highly similar to *PMS2CL*. This creates challenges in alignment of correct reads at these exons and creates artefacts during variant calling. This limitation has also been mentioned in a pilot study [83]. This makes it important to manually check reads and coverage in a genomic viewer, and to do Sanger verification of variants, as we did for the *PMS2* gene.

The current study emphasises the importance of including non-coding intronic regions. These regions will often have splice site variants, which may contribute to 10% of all disease-causing hereditary mutations according to HGMD [63–65], and deep intronic variants (e.g., in branch-point sequences, U2 type introns) which also contribute towards disease, most frequently by creating new pseudo-exons by activating non-trivial splice sites or by changing splicing regulatory elements. Intronic variants can also disrupt transcription regulatory motifs and non-coding RNA genes [88]. However, it is challenging to annotate these intronic variants due to limitations of annotation databases and tools. In a clinical setting, these variants can easily be missed unless RNA studies are performed to check for exon skipping, generation of new donor sites or cryptic site activation. Considering the potential importance of such variants, the current study included all intronic regions in order to search for this type of variant. Among 10 significant intronic variants we found four in the splice site vicinity and six in deep intronic regions.

NGS was performed, aiming at a data quality greater than 100X (average read coverage depth) for all samples. However, only 23 samples achieved this coverage (highest among them 169X), whereas 50 samples had coverage of less than 30X. These low-quality samples were included in the study, with the aim of exploring the value of low-quality data when searching for true positive (TP) variants. Using low quality data (i.e., with low coverage) led to a higher fraction of false positive (FP) variants, as 16 variants identified from the data analysis were

subsequently identified as false positive variants by Sanger sequencing. Most of them had low coverage at the variant position (between 14X to 6X coverage), whereas others were in repeat regions. FP variants in *MLH1*, *MSH6* & *PMS2* genes were in repeat regions, and had low coverage (except a *PMS2* variant with 84X coverage), which possibly led to their false SNV call. On the other hand, we cannot rule out that some were not verified due to SNPs in the primer binding site (allelic dropout). However, we also found many true positive variants in low coverage regions, as we found and confirmed 6 true positive variants in regions with low coverage (between 16X and 4X). Among these were two class 3 variants (*MSH2*), one class 3+ variant (*POLD1*) and a class 5 variant (*MSH6*). This shows the potential for finding true variants of significance even in low-quality samples, given that the variants can be verified.

Our initial aim also included identification of CNVs. CNVs can occur in both exonic and intronic regions of protein coding genes, with intronic CNVs being more frequent [89], and both types can contribute towards disease. However, due to the limitations of data quality (non-uniform and low coverage depth), it was not possible to do reliable CNV calling. Also, there is no availability of MLPA kits (MRC-Holland) for detecting CNVs for many genes in this panel.

To associate variants with possible effects we utilized *in silico* resources and tools, in addition to published literature. Effect prediction and annotation of all variants was done using multiple tools as mentioned in the methods section. Also, multiple potential factors and effects, like conservation in variant position or structural changes at protein level, were checked for each variant. This consensus-like approach (multiple tools, multiple potential effects) increases the robustness of predictions and annotations of the variants, although we also had cases of contradictory predictions, which illustrates the challenge of using *in silico* prediction tools.

Among the 199 EC patient samples, we identified variants of interest (for further investigations) in 34 patients. Among these, we found 3 patients with class 5 variants (in *MSH6* gene) and two patients with the same class 4 variant (in *PMS2* gene); ~2,5% of patients had pathogenic variants representing a very likely cause of cancer in these five patients. This is in accordance with other studies. One meta-analysis of 53 studies concluded the prevalence of LS in EC patients to be approximately 3% [90]. These studies have only looked into the coding part of the four MMR genes MLH1, MSH2, MSH6 and PMS2. We found class 3 variants in another 29 patients, some of which are highly suspicious of being pathogenic variants. This indicates potential causes of their disease, although further studies are required to confirm their actual significance. It is an important limitation for further interpretation of these class 3 variants that we lack information about the patients debut age of cancer, and results from tumour analyses (MSI status and immunohistochemistry of MMR genes). For the remaining 164 patients, we did not find any significant variant to explain their disease. Expansion of panel size with more genes, improved annotation (particularly of variants in non-protein-coding regions), and improved data quality may help in explanation of some of these cases. However, since the study cohort consist of consecutive EC patients, most of the cancers will be sporadic with no underlying high penetrant genetic cause.

Conclusions

Including all genes of the MMR pathway in a gene panel provides opportunity to discover variants in additional genes that potentially can be associated with a risk for EC, and hence are relevant for further investigation towards a better understanding of the development of EC. Including non-coding parts provides chances of identifying gene regulation or splice site alteration variants, although this will lead to a larger number of unknown variants which is challenging to study and annotate. In silico tools can be useful to find some leads in this situation,

although their predictions can be ambiguous and noisy. Hence in silico tools should not be used in identifying pathogenicity by themselves. In addition, although low-quality data should be avoided, such data can still support identification of informative variants. But such data will also lead to increased noise in the analysis, and experimental verification of such variants is essential. We identified pathogenic MMR variants in the same order of magnitude as earlier reported. In addition, we identified 31 class-3 (VUS) variants some of which may be disease causing. This supports that screening for LS among EC patients should be recommended. However, to determine whether the use of an extended panel of MMR genes (beyond *MLH1*, *MSH2*, *MSH6* and *PMS2*) has clinical value needs further investigation.

Supporting information

S1 Fig. Transcription factor ChIP-seq cluster for MSH6 5'UTR variant.
(TIF)

S2 Fig. Transcription factor ChIP-seq cluster for RFC3 5'UTR variant.
(TIF)

S1 Table. Read coverage depth of samples across 12 sequencing runs.
(PDF)

S2 Table. List of 5'UTR variants.
(PDF)

S3 Table. List of 3'UTR variants.
(PDF)

S1 File. Bioinformatics analysis steps.
(PDF)

S2 File. Variants filtering steps.
(PDF)

S3 File. All significant variants with annotation details.
(XLSX)

Author Contributions

Conceptualization: Ashish Kumar Singh, Bente Talseth-Palmer, Finn Drabløs, Wenche Sjursen.

Data curation: Ashish Kumar Singh.

Formal analysis: Ashish Kumar Singh.

Funding acquisition: Bente Talseth-Palmer.

Investigation: Ashish Kumar Singh, Finn Drabløs, Wenche Sjursen.

Methodology: Ashish Kumar Singh, Finn Drabløs, Wenche Sjursen.

Project administration: Bente Talseth-Palmer, Finn Drabløs, Wenche Sjursen.

Resources: Bente Talseth-Palmer.

Software: Ashish Kumar Singh.

Supervision: Finn Drabløs, Wenche Sjursen.

Validation: Mary McPhillips, Liss Anne Solberg Lavik, Alexandre Xavier.

Visualization: Ashish Kumar Singh.

Writing – original draft: Ashish Kumar Singh, Bente Talseth-Palmer, Finn Drabløs, Wenche Sjursen.

Writing – review & editing: Ashish Kumar Singh, Bente Talseth-Palmer, Finn Drabløs, Wenche Sjursen.

References

1. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov 1; 68(6):394–424. https://doi.org/10.3322/caac.21492 PMID: 30207593
2. Lortet-Tieulent J, Ferlay J, Bray F, Jemal A. International Patterns and Trends in Endometrial Cancer Incidence, 1978–2013. JNCI J Natl Cancer Inst. 2017 Oct 16; 110(4):354–61.
3. Jemal A, Siegel R, Xu J, Ward E. Cancer Statistics, 2010. CA Cancer J Clin. 2010 Sep 1; 60(5):277–300. https://doi.org/10.3322/caac.20073 PMID: 20610543
4. Colombo N, Preti E, Landoni F, Carinelli S, Colombo A, Marini C, et al. Endometrial cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol. 2013 Oct 1; 24(suppl 6): vi33–8.
5. Jenabi E, Poorolajal J. The effect of body mass index on endometrial cancer: a meta-analysis. Public Health. 2015 Jul 1; 129(7):872–80. https://doi.org/10.1016/j.puhe.2015.04.017 PMID: 26026348
6. Banno K, Yanokura M, Kobayashi Y, Kawaguchi M, Nomura H, Hirasawa A, et al. Endometrial cancer as a familial tumor: pathology and molecular carcinogenesis (review). Curr Genomics. 2009 Apr; 10 (2):127–32. https://doi.org/10.2174/138920209787847069 PMID: 19794885
7. Kunitomi H, Banno K, Yanokura M, Takeda T, Iijima M, Nakamura K, et al. New use of microsatellite instability analysis in endometrial cancer. Oncol Lett. 2017; 14(3):3297. https://doi.org/10.3892/ol.2017.6640 PMID: 28927079
8. Resnick KE, Frankel WL, Morrison CD, Fowler JM, Copeland LJ, Stephens J, et al. Mismatch repair status and outcomes after adjuvant therapy in patients with surgically staged endometrial cancer ☆. Gynecol Oncol. 2010; 117:234–8. https://doi.org/10.1016/j.ygyno.2009.12.028 PMID: 20153885
9. Kawaguchi M, Banno K, Yanokura M, Kobayashi Y, Kishimi A, Ogawa S, et al. Analysis of candidate target genes for mononucleotide repeat mutation in microsatellite instability-high (MSI-H) endometrial cancer. Int J Oncol. 2009 Sep 15; 35(05):977–82.
10. Bansidhar BJ. Extracolonic Manifestations of Lynch Syndrome. Clin Colon Rectal Surg. 2012; 25:103–10. https://doi.org/10.1055/s-0032-1313781 PMID: 23730225
11. Barrow E, Hill J, Evans DG. Cancer risk in Lynch Syndrome. Fam Cancer. 2013 Jun 21; 12(2):229–40. https://doi.org/10.1007/s10689-013-9615-1 PMID: 23604856
12. Ferguson SE, Aronson M, Pollett A, Eiriksson LR, Oza AM, Gallinger S, et al. Performance characteristics of screening strategies for Lynch syndrome in unselected women with newly diagnosed endometrial cancer who have undergone universal germline mutation testing. Cancer. 2014 Dec 15; 120(24):3932–9. https://doi.org/10.1002/cncr.28933 PMID: 25081409
13. Cunningham JM, Christensen ER, Tester DJ, Kim CY, Roche PC, Burgart LJ, et al. Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. Cancer Res. 1998 Aug 1; 58 (15):3455–60. PMID: 9699680
14. Herman JG, Umar A, Polyak K, Graff JR, Ahuja N, Issa JP, et al. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. Proc Natl Acad Sci U S A. 1998 Jun 9; 95(12):6870–5. https://doi.org/10.1073/pnas.95.12.6870 PMID: 9618505
15. Haraldsdottir S, Hampel H, Tomsic J, Frankel WL, Pearlman R, de la Chapelle A, et al. Colon and Endometrial Cancers With Mismatch Repair Deficiency Can Arise From Somatic, Rather Than Germline, Mutations. Gastroenterology. 2014 Dec 1; 147(6):1308-1316.e1.
16. Bonis PA, Trikalinos TA, Chung M, Chew P, Ip S, Devine DA, et al. Hereditary Nonpolyposis Colorectal Cancer: Diagnostic Strategies and Their Implications [Internet]. 2007 [cited 2018 Sep 3].
17. Ashton KA, Proietto A, Otton G, Symonds I, McEvoy M, Attia J, et al. The influence of the Cyclin D1 870 G>A polymorphism as an endometrial cancer risk factor. BMC Cancer. 2008 Sep 29; 8:272. https://doi.org/10.1186/1471-2407-8-272 PMID: 18822177
18. Illumina. Illimina MiSeq.

19. Alamut. Alamut-batch [Internet]. Interactive Biosoftware, Rouen, France;

20. Vigeland MD, Gjøtterud KS, Selmer KK. FILTUS: a desktop GUI for fast and efficient detection of disease-causing variants, including a novel autozygosity detector. Bioinformatics. 2016 Jan 27; 32 (10):1592–4. https://doi.org/10.1093/bioinformatics/btw046 PMID: 26819469

21. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRxiv. 2019 Jan 30;531210.

22. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med. 2015; 17(5):405–23. https://doi.org/10.1038/gim.2015.30 PMID: 25741868

23. Alamut. Alamut-visual [Internet]. Interactive Biosoftware, Rouen, France;

24. Tournier I, Vezain M, Martins A, Charbonnier F, Baert-Desurmont S, Olschwang S, et al. A large fraction of unclassified variants of the mismatch repair genes *MLH1* and *MSH2* is associated with splicing defects. Hum Mutat. 2008 Dec 1; 29(12):1412–24. https://doi.org/10.1002/humu.20796 PMID: 18561205

25. Lo SM, Choi M, Liu J, Jain D, Boot RG, Kallemeijn WW, et al. Phenotype diversity in type 1 Gaucher disease: discovering the genetic basis of Gaucher disease/hematologic malignancy phenotype by individual genome analysis. Blood. 2012 May 17; 119(20):4731. https://doi.org/10.1182/blood-2011-10-386862 PMID: 22493294

26. Ollila S, Sarantaus L, Kariola R, Chan P, Hampel H, Holinski–Feder E, et al. Pathogenicity of MSH2 Missense Mutations Is Typically Associated With Impaired Repair Capability of the Mutated Protein. Gastroenterology. 2006 Nov 1; 131(5):1408–17. https://doi.org/10.1053/j.gastro.2006.08.044 PMID: 17101317

27. Ollila S, Bebek DD, Jiricny J, Nyström M. Mechanisms of pathogenicity in human MSH2 missense mutants. Hum Mutat. 2008; 29(11):1355–63. https://doi.org/10.1002/humu.20893 PMID: 18951462

28. Hegde M, Blazo M, Chong B, Prior T, Richards C. Assay validation for identification of hereditary nonpolyposis colon cancer-causing mutations in mismatch repair genes MLH1, MSH2, and MSH6. J Mol Diagnostics. 2005; 7(4):525–34.

29. Hampel H, Frankel W, Panescu J, Lockman J, Sotamaa K, Fix D, et al. Screening for Lynch syndrome (hereditary nonpolyposis colorectal cancer) among endometrial cancer patients. Cancer Res. 2006 Aug 1; 66(15):7810–7. https://doi.org/10.1158/0008-5472.CAN-06-1114 PMID: 16885385

30. Ou J, Niessen RC, Lützen A, Sijmons RH, Kleibeuker JH, de Wind N, et al. Functional analysis helps to clarify the clinical importance of unclassified variants in DNA mismatch repair genes. Hum Mutat. 2007 Nov; 28(11):1047–54. https://doi.org/10.1002/humu.20580 PMID: 17594722

31. Chao EC, Velasquez JL, Witherspoon MSL, Rozek LS, Peel D, Ng P, et al. Accurate classification of MLH1/MSH2 missense variants with Multivariate Analysis of Protein Polymorphisms-Mismatch Repair (MAPP-MMR). Hum Mutat. 2008 Jun; 29(6):852–60. https://doi.org/10.1002/humu.20735 PMID: 18383312

32. Martinez SL, Kolodner RD. Functional analysis of human mismatch repair gene mutations identifies weak alleles and polymorphisms capable of polygenic interactions. Proc Natl Acad Sci U S A. 2010 Mar 16; 107(11):5070–5. https://doi.org/10.1073/pnas.1000798107 PMID: 20176959

33. Kansikas M, Kariola R, Nyström M. Verification of the three-step model in assessing the pathogenicity of mismatch repair gene variants. Hum Mutat. 2011 Jan; 32(1):107–15. https://doi.org/10.1002/humu.21409 PMID: 21120944

34. Thompson BA, Greenblatt MS, Vallee MP, Herkert JC, Tessereau C, Young EL, et al. Calibration of Multiple In Silico Tools for Predicting Pathogenicity of Mismatch Repair Gene Missense Substitutions. Hum Mutat. 2013 Jan; 34(1):255–65. https://doi.org/10.1002/humu.22214 PMID: 22949387

35. Chubb D, Broderick P, Frampton M, Kinnersley B, Sherborne A, Penegar S, et al. Genetic diagnosis of high-penetrance susceptibility for colorectal cancer (CRC) is achievable for a high proportion of familial CRC by exome sequencing. J Clin Oncol. 2015 Feb 10; 33(5):426–32. https://doi.org/10.1200/JCO.2014.56.5689 PMID: 25559809

36. Houlleberghs H, Dekker M, Lantermans H, Kleinendorst R, Dubbink HJ, Hofstra RMW, et al. Oligonucleotide-directed mutagenesis screen to identify pathogenic Lynch syndrome-associated MSH2 DNA mismatch repair gene variants. Proc Natl Acad Sci U S A. 2016 Apr 12; 113(15):4128–33. https://doi.org/10.1073/pnas.1520813113 PMID: 26951660

37. Pal T, Akbari MR, Sun P, Lee J-H, Fulp J, Thompson Z, et al. Frequency of mutations in mismatch repair genes in a population-based study of women with ovarian cancer. Br J Cancer. 2012 Nov 6; 107 (10):1783–90. https://doi.org/10.1038/bjc.2012.452 PMID: 23047549

38. Barnetson RA, Cartwright N, van Vliet A, Haq N, Drew K, Farrington S, et al. Classification of ambiguous mutations in DNA mismatch repair genes identified in a population-based study of colorectal cancer. Hum Mutat. 2008 Mar; 29(3):367–74. https://doi.org/10.1002/humu.20635 PMID: 18033691

39. Karageorgos I, Mizzi C, Giannopoulou E, Pavlidis C, Peters BA, Zagoriti Z, et al. Identification of cancer predisposition variants in apparently healthy individuals using a next-generation sequencing-based family genomics approach. Hum Genomics. 2015 Dec 20; 9(1):12.

40. Drost M, Koppejan H, de Wind N. Inactivation of DNA mismatch repair by variants of uncertain significance in the PMS2 gene. Hum Mutat. 2013 Nov; 34(11):1477–80. https://doi.org/10.1002/humu.22426 PMID: 24027009

41. Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, et al. The Clinical Phenotype of Lynch Syndrome Due to Germ-Line PMS2 Mutations. Gastroenterology. 2008 Aug 1; 135(2):419–428. e1. https://doi.org/10.1053/j.gastro.2008.04.026 PMID: 18602922

42. van der Klift HM, Mensenkamp AR, Drost M, Bik EC, Vos YJ, Gille HJJP, et al. Comprehensive Mutation Analysis of *PMS2* in a Large Cohort of Probands Suspected of Lynch Syndrome or Constitutional Mismatch Repair Deficiency Syndrome. Hum Mutat. 2016 Nov 1; 37(11):1162–79. https://doi.org/10.1002/humu.23052 PMID: 27435373

43. Borràs E, Pineda M, Cadiñanos J, Del Valle J, Brieger A, Hinrichsen I, et al. Refining the role of PMS2 in Lynch syndrome: germline mutational analysis improved by comprehensive assessment of variants. J Med Genet. 2013 Aug 1; 50(8):552–63. https://doi.org/10.1136/jmedgenet-2012-101511 PMID: 23709753

44. Auclair J, Leroux D, Desseigne F, Lasset C, Saurin JC, Joly MO, et al. Novel biallelic mutations inMSH6 andPMS2 genes: gene conversion as a likely cause ofPMS2 gene inactivation. Hum Mutat. 2007 Nov 1; 28(11):1084–90. https://doi.org/10.1002/humu.20569 PMID: 17557300

45. Brohl AS, Patidar R, Turner CE, Wen X, Song YK, Wei JS, et al. Frequent inactivating germline mutations in DNA repair genes in patients with Ewing sarcoma. Genet Med. 2017 Aug 1; 19(8):955–8. https://doi.org/10.1038/gim.2016.206 PMID: 28125078

46. Nowak JA, Yurgelun MB, Bruce JL, Rojas-Rudilla V, Hall DL, Shivdasani P, et al. Detection of Mismatch Repair Deficiency and Microsatellite Instability in Colorectal Adenocarcinoma by Targeted Next-Generation Sequencing. J Mol Diagnostics. 2017 Jan 1; 19(1):84–91.

47. Rengifo-Cam W, Jasperson K, Garrido-Laguna I, Colman H, Scaife C, Samowitz W, et al. A 30-Year-Old Man with Three Primary Malignancies: A Case of Constitutional Mismatch Repair Deficiency. ACG Case Reports J. 2017; 4(1):e34.

48. Jagmohan-Changur S, Poikonen T, Vilkki S, Launonen V, Wikman F, Orntoft TF, et al. EXO1 variants occur commonly in normal population: evidence against a role in hereditary nonpolyposis colorectal cancer. Cancer Res. 2003 Jan 1; 63(1):154–8. PMID: 12517792

49. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2013 Nov 14; 42 (D1):D980–5.

50. Shapiro MB, Senapathy P. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Res. 1987 Sep 11; 15(17):7155–74. https://doi.org/10.1093/nar/15.17.7155 PMID: 3658675

51. Yeo G, Burge CB. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. J Comput Biol. 2004 Mar; 11(2–3):377–94. https://doi.org/10.1089/1066527041410418 PMID: 15285897

52. Soukarieh O, Gaildrat P, Hamieh M, Drouet A, Baert-Desurmont S, Frébourg T, et al. Exonic Splicing Mutations Are More Prevalent than Currently Estimated and Can Be Predicted by Using In Silico Tools. Aretz S, editor. PLOS Genet. 2016 Jan 13; 12(1):e1005756. https://doi.org/10.1371/journal.pgen.1005756 PMID: 26761715

53. Salgado D, Desvignes J-P, Rai G, Blanchard A, Miltgen M, Pinard A, et al. UMD-Predictor: A High-Throughput Sequencing Compliant System for Pathogenicity Prediction of any Human cDNA Substitution. Hum Mutat. 2016 May; 37(5):439–46. https://doi.org/10.1002/humu.22965 PMID: 26842889

54. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics. 2006 Mar 22; 7(1):166.

55. Tomsic J, Senter L, Liyanarachchi S, Clendenning M, Vaughn CP, Jenkins MA, et al. Recurrent and founder mutations in the *PMS2* gene. Clin Genet. 2013 Mar 1; 83(3):238–43. https://doi.org/10.1111/j.1399-0004.2012.01898.x PMID: 22577899

56. Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017 Jan 4; 45(D1):D158–69. https://doi.org/10.1093/nar/gkw1099 PMID: 27899622

57. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 2005 Aug 1; 15(8):1034–50. https://doi.org/10.1101/gr.3715005 PMID: 16024819

58. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010 Jan 1; 20(1):110–21. https://doi.org/10.1101/gr.097857.109 PMID: 19858363

59. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet. 2013 Jan;Chapter 7:Unit7.20.

60. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003 Jul 1; 31(13):3812–4. https://doi.org/10.1093/nar/gkg509 PMID: 12824425

61. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014 Apr 1; 11(4):361–2. https://doi.org/10.1038/nmeth.2890 PMID: 24681721

62. Haradhvala NJ, Kim J, Maruvka YE, Polak P, Rosebrock D, Livitz D, et al. Distinct mutational signatures characterize concurrent loss of polymerase proofreading and mismatch repair. Nat Commun. 2018 Dec 1; 9(1):1746. https://doi.org/10.1038/s41467-018-04002-4 PMID: 29717118

63. Krawczak M, Thomas NST, Hundrieser B, Mort M, Wittig M, Hampe J, et al. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum Mutat. 2007 Feb 1; 28(2):150–8. https://doi.org/10.1002/humu.20400 PMID: 17001642

64. Stenson PD, Mort M, Ball E V, Howells K, Phillips AD, Thomas NS, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009 Jan 22; 1(1):13. https://doi.org/10.1186/gm13 PMID: 19348700

65. Cooper DN. Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes. Hum Genomics. 2010 Jun; 4(5):284–8. https://doi.org/10.1186/1479-7364-4-5-284 PMID: 20650817

66. InSiGHT Variant Interpretation Committee: Mismatch Repair Gene Variant Classification Criteria Rules for Variant Classification [Internet]. 2018 [cited 2019 Nov 15].

67. Gheorghe M, Sandve GK, Khan A, Cheneby J, Ballester B, Mathelier A. A map of direct TF-DNA interactions in the human genome. bioRxiv. 2018 Aug 17;394205.

68. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004 Jan 1; 32(90001):91D – 94.

69. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. Nucleic Acids Res. 2018 Jan 4; 46(D1):D260–6. https://doi.org/10.1093/nar/gkx1126 PMID: 29140473

70. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res. 2002 Jun; 12(6):996–1006. https://doi.org/10.1101/gr.229102 PMID: 12045153

71. Guo YA, Chang MM, Huang W, Ooi WF, Xing M, Tan P, et al. Mutation hotspots at CTCF binding sites coupled to chromosomal instability in gastrointestinal cancers. Nat Commun. 2018 Dec 18; 9(1):1520. https://doi.org/10.1038/s41467-018-03828-2 PMID: 29670109

72. Garcia DM, Baek D, Shin C, Bell GW, Grimson A, Bartel DP. Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. Nat Struct Mol Biol. 2011 Oct 11; 18(10):1139–46. https://doi.org/10.1038/nsmb.2115 PMID: 21909094

73. Saito T, Sætrom P. A two-step site and mRNA-level model for predicting microRNA targets. BMC Bioinformatics. 2010 Dec 31; 11(1):612.

74. Thomas LF, Saito T, Sætrom P. Inferring causative variants in microRNA target sites. Nucleic Acids Res. 2011 Sep 1; 39(16):e109–e109. https://doi.org/10.1093/nar/gkr414 PMID: 21693556

75. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011 Jan 1; 39(Database):D152–7.

76. Keijzers G, Bakula D, Petr M, Madsen N, Teklu A, Mkrtchyan G, et al. Human Exonuclease 1 (EXO1) Regulatory Functions in DNA Replication with Putative Roles in Cancer. Int J Mol Sci. 2018 Dec 25; 20 (1):74.

77. Zhang M, Zhao D, Yan C, Zhang L, Liang C. Associations between Nine Polymorphisms in EXO1 and Cancer Susceptibility: A Systematic Review and Meta-Analysis of 39 Case-control Studies. Sci Rep. 2016 Sep 8; 6(1):29270.

78. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet. 2015 Apr 9; 47(4):373–80. https://doi.org/10.1038/ng.3242 PMID: 25751625

79. Kang BW, Jeon H-S, Chae YS, Lee SJ, Park JS, Choi GS, et al. Impact of Genetic Variation in MicroRNA-binding Site on Susceptibility to Colorectal Cancer. Anticancer Res. 2016 Jul 1; 36(7):3353–61. PMID: 27354594

80. Chen X-P, Chen Y-G, Lan J-Y, Shen Z-J. MicroRNA-370 suppresses proliferation and promotes endometrioid ovarian cancer chemosensitivity to cDDP by negatively regulating ENG. Cancer Lett. 2014 Oct 28; 353(2):201–10. https://doi.org/10.1016/j.canlet.2014.07.026 PMID: 25063739

81. Yang I-P, Tsai H-L, Hou M-F, Chen K-C, Tsai P-C, Huang S-W, et al. MicroRNA-93 inhibits tumor growth and early relapse of human colorectal cancer by affecting genes involved in the cell cycle. Carcinogenesis. 2012 Aug 1; 33(8):1522–30. https://doi.org/10.1093/carcin/bgs166 PMID: 22581829

82. Xiao Z-G, Deng Z-S, Zhang Y-D, Zhang Y, Huang Z-C. Clinical significance of microRNA-93 downregulation in human colon cancer. Eur J Gastroenterol Hepatol. 2013 Mar; 25(3):296–301. https://doi.org/10.1097/MEG.0b013e32835c077a PMID: 23354160

83. Talseth-Palmer BA, Bauer DC, Sjursen W, Evans TJ, Mcphillips M, Proietto A, et al. Targeted next-generation sequencing of 22 mismatch repair genes identifies Lynch syndrome families. Cancer Med. 2016; 5(5):929–41. https://doi.org/10.1002/cam4.628 PMID: 26811195

84. Xavier A, Olsen MF, Lavik LA, Johansen J, Singh AK, Sjursen W, et al. Comprehensive mismatch repair gene panel identifies variants in patients with Lynch-like syndrome. Mol Genet Genomic Med. 2019 Aug 1; 7(8):e850. https://doi.org/10.1002/mgg3.850 PMID: 31297992

85. Billingsley CC, Cohn DE, Mutch DG, Stephens JA, Suarez AA, Goodfellow PJ. Polymerase ε (POLE) mutations in endometrial cancer: clinical outcomes and implications for Lynch syndrome testing. Cancer. 2015 Feb 1; 121(3):386–94. https://doi.org/10.1002/cncr.29046 PMID: 25224212

86. Konstantinopoulos PA, Matulonis UA. *POLE* mutations as an alternative pathway for microsatellite instability in endometrial cancer: Implications for Lynch syndrome testing. Cancer. 2015 Feb 1; 121(3):331–4. https://doi.org/10.1002/cncr.29057 PMID: 25224324

87. Huth C, Kloor M, Voigt AY, Bozukova G, Evers C, Gaspar H, et al. The molecular basis of EPCAM expression loss in Lynch syndrome-associated tumors. Mod Pathol. 2012; 25(6):911–6. https://doi.org/10.1038/modpathol.2012.30 PMID: 22388758

88. Vaz-Drago R, Custódio N, Carmo-Fonseca M. Deep intronic mutations and human disease. Hum Genet. 2017 Sep 12; 136(9):1093–111. https://doi.org/10.1007/s00439-017-1809-4 PMID: 28497172

89. Rigau M, Juan D, Valencia A, Rico D. Intronic CNVs and gene expression variation in human populations. Semple C, editor. PLOS Genet. 2019 Jan 24; 15(1):e1007902. https://doi.org/10.1371/journal.pgen.1007902 PMID: 30677042

90. Ryan NAJ, Glaire MA, Blake D, Cabrera-Dandy M, Evans DG, Crosbie EJ. The proportion of endometrial cancers associated with Lynch syndrome: a systematic review of the literature and meta-analysis. Genet Med. 2019; 21(10):2167–80. https://doi.org/10.1038/s41436-019-0536-8 PMID: 31086306

Article 2

BMC Medical Genomics

SOFTWARE Open Access

Detecting copy number variation in next generation sequencing data from diagnostic gene panels

Ashish Kumar Singh[1,2*], Maren Fridtjofsen Olsen[1], Liss Anne Solberg Lavik[1], Trine Vold[1], Finn Drabløs[2] and Wenche Sjursen[1,2]

Abstract

Background: Detection of copy number variation (CNV) in genes associated with disease is important in genetic diagnostics, and next generation sequencing (NGS) technology provides data that can be used for CNV detection. However, CNV detection based on NGS data is in general not often used in diagnostic labs as the data analysis is challenging, especially with data from targeted gene panels. Wet lab methods like MLPA (MRC Holland) are widely used, but are expensive, time consuming and have gene-specific limitations. Our aim has been to develop a bioinformatic tool for CNV detection from NGS data in medical genetic diagnostic samples.

Results: Our computational pipeline for detection of CNVs in NGS data from targeted gene panels utilizes coverage depth of the captured regions and calculates a copy number ratio score for each region. This is computed by comparing the mean coverage of the sample with the mean coverage of the same region in other samples, defined as a pool. The pipeline selects pools for comparison dynamically from previously sequenced samples, using the pool with an average coverage depth that is nearest to the one of the samples. A sliding window-based approach is used to analyze each region, where length of sliding window and sliding distance can be chosen dynamically to increase or decrease the resolution. This helps in detecting CNVs in small or partial exons. With this pipeline we have correctly identified the CNVs in 36 positive control samples, with sensitivity of 100% and specificity of 91%. We have detected whole gene level deletion/duplication, single/multi exonic level deletion/duplication, partial exonic deletion and mosaic deletion. Since its implementation in mid-2018 it has proven its diagnostic value with more than 45 CNV findings in routine tests.

Conclusions: With this pipeline as part of our diagnostic practices it is now possible to detect partial, single or multi-exonic, and intragenic CNVs in all genes in our target panel. This has helped our diagnostic lab to expand the portfolio of genes where we offer CNV detection, which previously was limited by the availability of MLPA kits.

Keywords: Next generation sequencing (NGS), Copy number variation (CNV), Structural variant, Multiplex ligation-dependent probe amplification (MLPA), Sliding window

Background

Potentially disease-causing DNA mutations include alterations of single nucleotides up to whole chromosomes. Small changes of 1 nucleotide (nt) are called single nucleotide variation and changes up to 50 nt at single locus are called short insertion-deletion variation (indel). Whereas alterations larger than 50 nt are called

*Correspondence: ashish.kumar.singh3@stolav.no; ashish.k.singh@ntnu.no
[1] Department of Medical Genetics, St. Olavs Hospital, Trondheim, Norway
Full list of author information is available at the end of the article

structural variants (SVs) [1], which previously has been defined as alterations larger than 1000 nt [2, 3]. Such SVs include insertions, deletions, duplications, inversions, and translocations. Combinations of these SVs are also possible in a single genome [4]. Deletions and duplications, commonly called copy number variations (CNVs), contribute to a large fraction of all genetic alterations and are of diagnostic relevance as they can play important roles in causing genetic diseases [5].

Several laboratory-based approaches have been developed and can be used for detecting CNVs, including multiplex ligation-dependent probe amplification (MLPA) [6], microarray based comparative genomic hybridization (aCGH) and SNP microarrays [7], RNA sequencing [8], fluorescence in situ hybridization (FISH) [9] and PCR based methods [10]. All these methods are laboratory intensive, have low throughput and are expensive. Among these, diagnostics labs most commonly use aCGH/SNP microarray and MLPA. The aCGH method is sensitive, but it is limited to detect only CNVs of sequences present in the reference assembly used to design the array probes [11]. Limitation in MLPA-based testing is the number of probes included in the kit. It is designed to multiplex up to approximately 50 probes, hence most suitable for one or a few smaller genes.

With the evolution of next generation sequencing (NGS) technologies, diagnostics laboratories are heavily utilizing NGS data in detection of SNPs and indels. With the current quality of NGS data it is also possible to detect CNVs [12]. In addition, NGS provides the benefit of detecting exact CNV breakpoint positions in the genome. Hence using NGS for CNV detection will help diagnostic labs in testing larger number of genes for CNVs. In traditional routine diagnostic practices, samples are analyzed by MLPA testing of genes according to requests. As CNVs do not occur that often, MLPA results are often negative. It has been shown that using NGS in diagnostics provides better throughput at a lower cost compared to using MLPA-based testing for CNVs [13], and this is also consistent with the experience of our in-house diagnostic lab. MLPA is then used mainly for verification on those genes where analysis of the NGS data has indicated a CNV.

Four different approaches are currently used for detecting CNVs from NGS data [14, 15]; paired-end mapping based detection (PE), split read based detection (SR), de novo assembly based detection (DA) and read depth based detection (RD). Additionally, mixed approaches are used. All these approaches use NGS generated reads to create consensus sequences by mapping to a reference genome or by de novo assembly and looking for anomalies occurring due to SVs. Among these approaches, PE, SR and DA can be used to discover all types of SVs, but application of these approaches requires high data quality and data consistency across regions [14], which often limits their applicability to whole genome sequencing data. On the other hand, the RD approach can only detect CNVs (deletions and duplications), but it predicts exact copy numbers, including mosaicism [16, 17], and can also detect small or very large CNVs in all types of regions in a genome. Depending on data quality, coverage depth, read length, and captured regions, RD can also detect exact breakpoints with high accuracy. The best approach for CNV detection will depend upon the available sequencing data. Data from targeted gene panels represent selected genetic regions of the genome, like specific exons, which means that it does not represent continuous regions of the genome. However, as the RD approach uses region-specific information (coverage depth) to detect CNVs, this is a good approach for targeted gene panels. Due to being deep-sequenced the panel data often have high coverage depth, which increases accuracy of CNV detection via the RD approach, although the fact that intronic regions are not included in the analysis may give a somewhat lower sensitivity to certain CNVs compared to using whole genome data [18].

There are several bioinformatic tools that have been developed to detect CNVs in NGS data [13, 19]. The majority of these tools have been developed for detecting large CNVs (in the size of megabases) and hence suitable only for whole genome or whole exome sequencing data [13]. In diagnostics labs where sequencing of targeted gene panels is common practice, the main goal is often to detect small (intragenic) disease-associated CNVs in partial, single or a few small exons [20]. There are a few available tools that claim to be suitable for data from targeted gene panels [21–25], but it is always challenging to detect smaller CNVs, especially partial or single exons or mosaic CNVs, with high sensitivity and specificity consistent with diagnostic standards.

We have developed a computational pipeline to detect CNVs in NGS data from targeted gene panels, which enables us to detect small CNVs in all targets included in our panel. Since implementation of the pipeline for routine diagnostics in our lab in August 2018 it has proved its diagnostic value by detecting 45 CNVs in 16 different genes, which includes partial exonic, single exonic, multi exonic, whole gene and mosaic CNVs. By implementing this method in our routine, we have reduced cost and lab-work overhead and improved diagnostic throughput.

Implementation

Here we describe our CNV detection pipeline which has been developed to work on NGS data generated from targeted gene panels. To identify potential CNVs the pipeline utilizes coverage depth information of reads in target

regions defined by the gene panel. If a target region has CNV, the coverage depth in this region will differ from the expected coverage depth. When duplicated, the target region will have 1.5 times more coverage depth than the expected coverage depth. On the other hand, in case of deletion the target region will have half the expected coverage depth. Figure 1 illustrates this approach of CNV detection.

To detect CNVs in a target region of a query sample, our pipeline (Fig. 2) utilized this principle by comparing coverage depth in this region of the query sample with average depth in same region for normal samples with similar coverage depth as the query sample. The normal samples are provided to the analysis, and the pipeline creates pools of normal samples, where each pool contains normal samples with similar coverage depth. These pools are called static pools and can be repeatedly used for CNV detection of any query sample where the coverage depth is similar to the average coverage depth of the pool. The pipeline is illustrated in Fig. 2.

Target region based sliding windows

To increase resolution each target region is divided into overlapping sub-regions in a sliding window approach as shown in Fig. 3, forming the template for a window-based representation of each target region. This approach is called the Target Region based Sliding Windows (TRSW) approach, or just sliding windows. This also helps in detecting CNVs occurring in smaller sub-regions, e.g., part of an exon. Selection of window size is based on length of sequencing reads and the required resolution of CNV predictions. Sliding length for two adjacent overlapping sliding windows remains the same across all regions and is kept relatively small compared to window size. This helps in detecting the start- and end-points of CNVs more accurately, up to the resolution of the sliding length. At our diagnostic lab standard sequencing read length is 150 nt (X2 paired-end reads). Hence a window size of 75 nt, i.e., half of the read length, along with a sliding length of 10 nt has been chosen for validation samples and for standard routine CNV detection in NGS runs. This gives an overlap

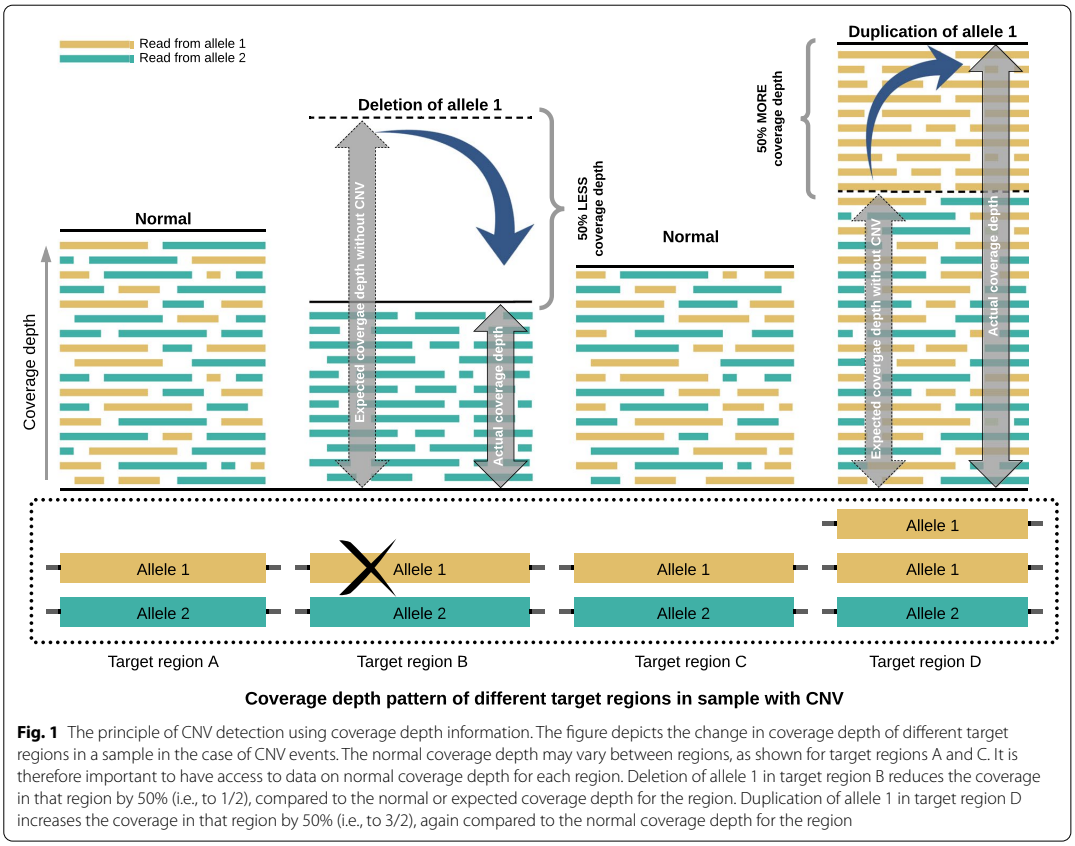


Fig. 1 The principle of CNV detection using coverage depth information. The figure depicts the change in coverage depth of different target regions in a sample in the case of CNV events. The normal coverage depth may vary between regions, as shown for target regions A and C. It is therefore important to have access to data on normal coverage depth for each region. Deletion of allele 1 in target region B reduces the coverage in that region by 50% (i.e., to 1/2), compared to the normal or expected coverage depth for the region. Duplication of allele 1 in target region D increases the coverage in that region by 50% (i.e., to 3/2), again compared to the normal coverage depth for the region

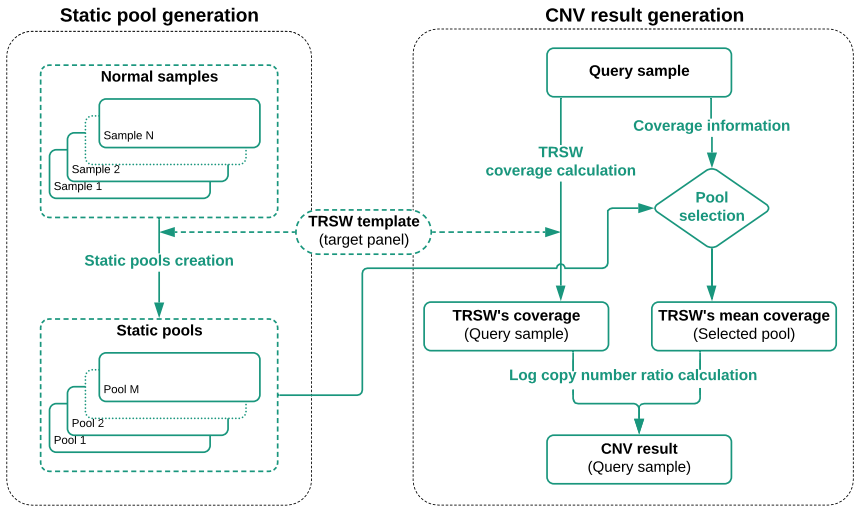


Fig. 2 The general workflow for CNV detection. In "static pool generation" a series pools or collections of normal samples are defined to be used as reference data during CNV detection, representing the expected coverage depth of each region. In "CNV result generation" the coverage depth of the query sample is used to select a suitable pool, and a region-wise comparison between the query sample and the selected pool is used to identify regions with potential CNVs. The resolution of the comparison is improved by using a template of overlapping windows across each region of the target panel, defined as target region-based sliding windows (TRSW), see the main text for details

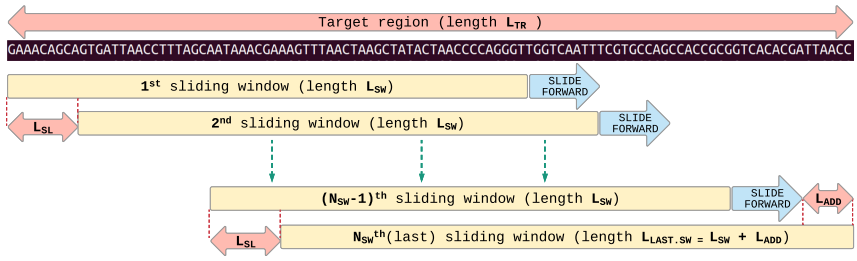


Fig. 3 Defining the sliding window template for a target region. The selected region is divided into smaller sub-regions by a sliding window approach where each window is of a fixed size and slides forward with a fixed sliding length. The last window can be larger than the chosen window size if the length of the remaining region is smaller than sliding length. Then the remaining region is just added to the last sliding window

of 65 nt between two consecutive windows. This selection of window size and sliding length gives a good tradeoff between computational complexity and resolution.

Equation 1a is used for calculating N_{SW}, the number of sliding windows for a target region of length L_{TR}, where sliding window length is L_{SW} and sliding length is L_{SL}.

$$N_{SW} = \frac{L_{TR} - L_{SW}}{L_{SL}} + 1 \tag{1a}$$

Window traversal for a region starts by aligning the first window at start of the region and sliding forward (with sliding length) until end of region. If for the last slide the remaining length of the region is less than sliding length, then the remaining length is added as an additional length to the last window. Hence the size of the last window in a region can be bigger than the chosen window size. Equations 1b and 1c are used for calculation of this additional length L_{ADD} and length of the last sliding window $L_{LAST.SW}$, respectively.

$$L_{ADD} = (L_{TR} - L_{SW})\%L_{SL} \quad (1b)$$

$$L_{LAST.SW} = L_{SW} + L_{ADD} \quad (1c)$$

Once window traversal ends for a target region, the next window starts at the beginning of the next target region. If the length of a target region is smaller than the chosen window size, then there will not be any splitting of that region into windows and there will only be one window for that region, of the same size as the region.

Static pools from normal samples

In first part of the pipeline static pools are created from normal samples with no CNVs, sorted according to coverage depth. The pipeline can then select a pool of samples that matches the coverage depth of the query sample and use this to estimate expected coverage depth (without any CNVs) for a region of interest. Figure 4 shows the workflow of static pool creation.

Targeted capturing kits always have batch effects in capturing quality due to differences in batches or lots of kits as provided from vendor [26]. This is a common issue with sequencing of targeted panels. Using samples from the same sequencing batch or lot reduces the level of noise by reducing batch effects in the CNV analysis. Therefore, normal samples used in creation of static pools for a CNV analysis should be sequenced using

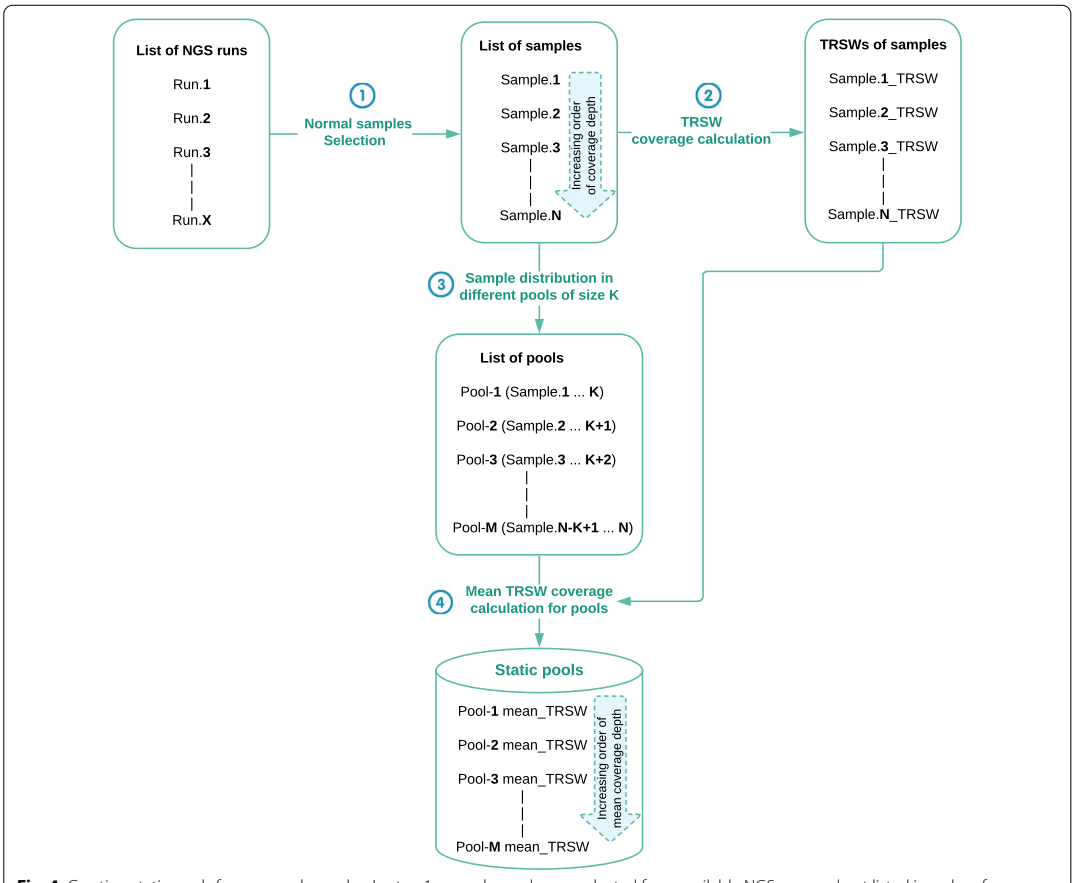


Fig. 4 Creating static pools from normal samples. In step 1 normal samples are selected from available NGS runs and get listed in order of increasing coverage depth. In step 2 the coverage depth is calculated for each window across each sample. In step 3 the list of selected normal samples is divided into different pools of size K, where Pool-1 consists of the first K samples, followed by the next pool consisting of the next K samples after skipping the first sample of the previous pool. In step 4 the mean TRSW of each pool is calculated

the same batch of target capturing kit as was used for the query samples.

Results from several NGS runs are used as input data in pool creation. The pipeline extracts normal samples (with depth of coverage higher than the assigned cut-off) from the provided runs and lists them in increasing order of coverage depth (Step 1 in Fig. 4).

To increase the resolution of CNV results the sliding windows approach (TRSWs, see above) is used. For each normal sample, coverage for all sliding windows is calculated (Step 2 in Fig. 4).

This list of samples is used for creating the static pools. Equation 2 is used for calculating M, the total number of pools generated from these samples given N, the number of normal samples, and K, the pool size.

$$M = N - K + 1 \tag{2}$$

Provided the size for each pool is K, the first K samples of the list are used to create the 1st static pool of normal samples, the 2nd pool skips first sample and uses the next K samples (2nd till K + 1th sample), and the same follows for next remaining pools. The Mth (last) pool uses last K samples (N − K + 1th till Nth sample) from the list (Step 3 in Fig. 4).

For each sliding window in the panel the mean coverage depth over all samples in each pool is calculated (Step 4 in Fig. 4). This list of mean coverage depth of each sliding window (mean_TRSW) of a pool is stored and used for CNV score calculations.

CNV calculation

As all regions in the target panel are split into smaller sliding windows (TRSWs) to increase the resolution of results, CNV score is calculated for each window. Figure 2 illustrates the CNV calculation workflow.

For a given query sample the coverage depth is first calculated for each sliding window. A static pool is then chosen from the set of static pools where mean coverage depth of the selected pool is closest to coverage depth of the sample. The coverage depth for each window of the query sample is compared against mean coverage depth of each corresponding window of the selected pool. This ratio is converted to $\log_2$ scale to calculate the final CNV score, i.e., log copy number ratio score (logCNR score) for that window. Equation 3 is used for calculating the $\text{logCNR}_{\text{score}}$ for a window, where L_{SW} is sliding window length, ND_i is nucleotide depth at ith position of query sample, ND_{ij} is nucleotide depth at ith position of jth sample in the static pool, and n is the number of samples in the selected static pool.

$$logCNR_{score} = log_2 \frac{1/L_{SW} \sum_{i}^{i+L_{SW}-1} ND_i}{1/n \sum_{j=1}^{n} \left(1/L_{SW} \sum_{i}^{i+L_{SW}-1} ND_{ij}\right)} \tag{3}$$

Theoretical values of $\text{logCNR}_{\text{score}}$ are 0.0 for 2 alleles (normal), −1.0 for 1 allele (deletion), and +0.58 for 3 alleles (duplication). The $\text{logCNR}_{\text{score}}$ for each sliding window is stored as CNV results of the query sample.

Quality control

The quality of the pools relatively to the query sample is important for the performance of our approach, and quality control of query and pools is therefore an important step for reducing noise in the analysis. Three quality checks are used. First, comparing the coverage depth of the query sample to average depth of the selected pool. Second, checking the uniformity in coverage depth among samples in the selected static pool. And third, comparing CNV results generated using static pools to results generated with run-wise pools (see below).

Query sample versus pool quality

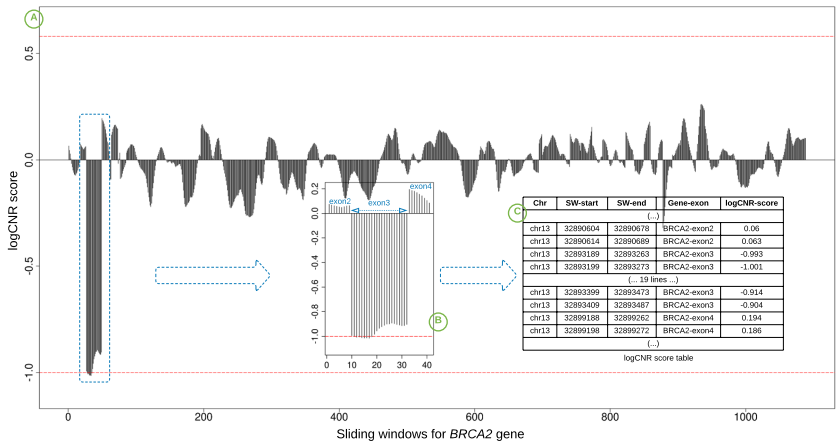
Quality of CNV results depends on a similar coverage depth of query sample and selected static pool. Hence for all query samples, percentage deviation of mean depth of the query sample relative to mean depth of the selected pool is checked. If this percentage deviation is larger than a cutoff (set by lab, for example 5%), then the query sample is re-analyzed with a larger (updated) list of static pools. If the deviation is still too large, then re-sequencing or a MLPA test is used, depending on the number of genes requested for analysis.

Static pool quality

The quality of the selected static pool can also affect the CNV results. Even when the percentage deviation of the coverage depth of the query sample compared to mean depth of the selected pool is lower than cutoff, differences in depth of normal samples used in making of selected pools can introduce noise. Hence only good quality pools (i.e., samples with uniform coverage depth) should be used for CNV detection. Additionally, run-wise pools (created by using all samples from the same NGS run of the query sample) can also be used to check quality of the static pool in case of noisy results.

Interpretation of output

For each gene in the target panel, logCNR score of windows belonging to that gene are plotted. These plots are checked for initial assessment. Once potential signals are identified, gene specific regions are looked up in the table of logCNR scores. As example of a deletion event, Fig. 5 shows plots of logCNR score of all sliding windows of



Chr	SW-start	SW-end	Gene-exon	logCNR-score
(...)				
chr13	32890604	32890678	BRCA2-exon2	0.06
chr13	32890614	32890689	BRCA2-exon2	0.063
chr13	32893189	32893263	BRCA2-exon3	-0.993
chr13	32893199	32893273	BRCA2-exon3	-1.001
(... 19 lines ...)				
chr13	32893399	32893473	BRCA2-exon3	-0.914
chr13	32893409	32893487	BRCA2-exon3	-0.904
chr13	32899188	32899262	BRCA2-exon4	0.194
chr13	32899198	32899272	BRCA2-exon4	0.186
(...)				

Fig. 5 Example of output from the CNV tool. The example shows identification of a deletion of exon 3 in the BRCA2 gene in one of the control samples (CS_12). Part A shows a plot of the logCNR score for all sliding windows across the exons of the BRCA2 gene. The horizontal lines at −1.00 and +0.58 represent expected score values for single-allele deletions and duplications, respectively. The score values in the region of exon 3 show clear signals of a deletion. Part B shows a zoomed in representation of the plot for exon 3 with the CNV deletion, whereas its neighbouring exons (2 and 4) have normal coverage depth. Part C shows the logCNR table listing score values for selected sliding windows covering the exon2–exon3 and exon3–exon4 junctions

BRCA2 gene in a control sample (CS_12) depicting signals of deletion of exon3, and the table in Fig. 5 enlists the logCNR scores of all sliding windows of same exon3 and its adjacent exon2 and exon4. In some cases, to get the best possible resolution (i.e., to locate exact break point) nucleotide-level coverage files are also checked. In our lab's diagnostic practices, we also generate merged plots for the same gene across all the samples sequenced in same run (without naming the samples to avoid incidental findings), which helps in detecting or rectifying any noise or signal. We also generate merged plots for run-wise versus static pooling results for all genes over all samples, which helps us in predicting or identifying any noise associated with static pools (see Quality control).

Once CNV signals have been confirmed in the logCNR score table, MLPA-based validation in performed on the sample. In cases of specific genes where MLPA test is not available, RNA sequencing or long-range PCR is performed for CNV verification.

Control samples

Selection of control samples for validation has been based on availability of known CNV positive samples, previously detected through MLPA. These samples were collected from the genetic diagnostic laboratories at Haukeland University Hospital (Bergen, Norway), University Hospital of North Norway (Tromsø, Norway), and St. Olavs Hospital (Trondheim, Norway). In total 36 positive control samples were used for validation of the CNV detection pipeline, where only genes with known CNVs were checked to reduce the risk of incidental findings. Additionally, 11 routine samples were chosen for calculating the specificity of the pipeline, where all the genes in the panel were checked for CNVs. These samples were collected at Department of Medical Genetics, St. Olavs Hospital, Trondheim, Norway. Both the 36 positive control samples and the 11 routine samples were germline samples where DNA had been extracted from blood.

The target gene panel consisted of 126 genes. For all genes, only exons, UTR regions and approximately ± 25 nucleotides in intronic regions were captured. These 126 genes are mainly cancer associated genes. Additional file 1 lists target regions and capturing probes.

Illumina's Nextera Rapid Capture Custom Enrichment kit was used for capturing the target sequences. Illumina MiSeq and Illumina NextSeq 500 sequencers were used for sequencing the samples.

Among 36 positive control samples, 22 samples were sequenced once (12 on MiSeq and 10 on NextSeq sequencer), 14 samples were sequenced twice (once on MiSeq and once on NextSeq sequencer). The 11 routine samples were sequenced on MiSeq. Repetition of sequencing was performed to replicate the results, and the use of different sequencers was done to test the

robustness of pipeline for differences in data quality due to different sequencing platforms.

Data pre-processing

Sequencing data (as FASTQ files) was preprocessed to generate suitable input for the CNV pipeline, using human genome version GRCh37 [27] as the reference genome. GATK best practices guidelines [28] were used for the preprocessing, which included alignment of raw pair-end reads (FASTQ files) to the reference genome using the BWA tool [29], further sorting, marking of duplicates, INDEL realignment and base quality score recalibration steps using the GATK toolkit to generate analysis-ready aligned reads (BAM files). These aligned reads (BAM files) were used for calculating average coverage and per-locus coverage (nucleotide level coverage) of samples by using the GATK toolkit tool DepthOfCoverage. Both average coverage and per-locus coverage of samples were used as input data to the CNV pipeline for both static pool creation and CNV calculation steps.

Results

Validation of the pipeline

The pipeline was validated with the 36 CNV positive control samples. Only the 12 genes with known CNVs detected with MLPA or RNA sequencing were looked at. All the previously detected CNV were found with the pipeline, and comprising 4 whole gene deletions, 6 single exon deletions, 17 multi-exon deletions, 2 single + partial exon deletions (break point inside the second exon), 3 single exon duplications and 4 multi-exon duplications. Table 1 lists these 12 genes and number of findings for each of them. Additional file 2 lists the CNV findings with genomic positions.

Sensitivity, specificity and accuracy

Calculation of sensitivity for this method was based on the 36 known CNVs in the 36 positive control samples. Since all the variants were detected by the pipeline, the measured sensitivity is 100%, at least for this set of samples.

Calculation of specificity was based on the results from 11 diagnostic routine samples where all 126 genes in the target panel were checked for CNVs. In total we analyzed 1386 (11×126) individual genetic regions for CNVs. Of these 1386 regions, we detected 126 false positive results. This provides specificity of 90.9% and total accuracy of 91.14% for the pipeline. Additional file 3 shows details of the false positives and regions of systematic error and their respective genes in these samples.

Using the pipeline in routine diagnostics

Since implementation of the CNV detection pipeline in routine work of our diagnostic lab in August 2018, we have detected 45 germline samples with CNVs. These CNVs were found in 16 different genes and include 5 whole gene deletions, 6 single exon deletions, 18 multi exon deletions, 1 multi + partial exon deletion, 1 multi-exon mosaic deletion, 1 whole gene duplication and 13 multi-exon duplications. Table 2 lists these 16 genes and the number of findings in each. Some of the diagnostic samples show similar CNV events, e.g., all 7 samples with CNVs in *ATM* genes, 4 out of 5 samples with CNVs in the *PMS2* gene and 7 out of 8 samples with CNVs in the *RAD51C* gene show the same duplication events. Some of these samples are from related family members. Additional file 4 lists these CNV findings with genomic positions. All these findings were verified by MLPA and/or RNA sequencing.

Table 1 Genes with CNVs identified in positive control samples

Gene name	Type of CNV: number of findings
APC	Whole gene deletion:1
BRCA1	Single exon deletion: 2; Multi exon deletion: 6; Single exon duplication: 1
BRCA2	Single exon deletion: 1; Multi exon deletion: 1; Single exon duplication: 2; Whole gene deletion: 2
CDH1	Multi exon deletion: 1
CDKN2A	Single + partial* exon deletion: 2
MLH1	Multi exon deletion: 3
MSH2	Single exon deletion: 1; Multi exon deletion: 4; Multi exon duplication: 1
NF1	Multi exon deletion: 1; Multi exon duplication: 1; Whole gene deletion: 1
PMS2	Multi exon duplication: 2
PTEN	Single exon deletion: 1
STK11	Single exon deletion: 1
VHL	Multi exon deletion: 1

* Here partial exon means that CNV breakpoint is inside exon

Table 2 Genes with CNVs identified in routine diagnostic samples

Gene name	Type of CNV: number of findings
ATM	Multi exon duplication: 7
BRCA1	Single exon deletion: 3; Multi exon deletion: 1; Multi + partial* exon deletion: 1; Multi exon duplication: 1
BRCA2	Single exon deletion: 1; Multi exon deletion: 1
CDC73	Multi exon deletion: 1
CDKN2A	Whole gene deletion (homozygote): 1
DICER1	Single exon deletion: 2
MLH1	Multi exon deletion: 1
MSH2	Multi exon deletion: 5; Whole gene deletion: 1
MSH6	Whole gene duplication: 1
NF1	Multi exon mosaic deletion: 1 (30% mosaicism)
NF2	Multi exon deletion: 1
PMS2	Multi exon duplication: 5
PTCH1	Whole gene deletion: 1
RAD51C	Multi exon deletion: 8
RB1	Whole gene deletion: 1
PTKAR1A	Whole gene deletion: 1

* Here partial exon means that CNV breakpoint is inside exon

Discussion

While keeping the needs of diagnostic labs as our central aim we have developed a CNV detection pipeline that works on NGS data from target panels. We have validated the pipeline and implemented it in routine diagnostics, and it has been used in diagnostic practices in our lab since mid-2018. Based on the experience from routine diagnostics of more than 3000 samples it has proven its diagnostic value. By using a sliding window approach to increase resolution and static pooling to reduce noise this pipeline generates high quality CNV results. With this pipeline we have detected different types of CNVs, including whole gene CNVs and CNVs occurring at exonic level, e.g., multi exonic (intra-genic), single exonic, partial exonic and mosaic CNVs. Detecting partial exonic CNVs with exact breakpoints as well as mosaic CNVs with relatively weak signals from target panel data can be challenging with available in silico methods. By being able to handle also such data this pipeline has shown its value in diagnostic use.

Validation of the pipeline was done using 36 CNV positive control samples consisting of different types of whole gene and intrageneric CNVs in 12 different genes (Table 1). The use of a larger number of positive control samples is often recommended for validation, but this was limited by the availability of known positive controls. However, by detecting all control sample CNVs, and hence giving a measured sensitivity of 100%, this pipeline meets the diagnostics requirement of no false negative results during the validation. Although we have to consider the fact that sensitivity calculation on a certain number of already known CNV positive genes may not be entirely representative of the actual performance during normal use.

The high sensitivity, specificity and accuracy of the pipeline shows that it is well suited for clinical practice. All the 126 false positive CNV detected in the 11 validation samples are in regions with very low coverage depth, which occurs due to non-optimal capturing by the capturing kit in these regions. We found 54 of these false positives to be systematic errors as these were observed consistently in the same regions in same genes across all samples. Most of these regions are homologous or repetitive regions and high GC content regions that are challenging to sequence and map. In routine practices some of these regions (often described as systematic gap regions) are tested by other methods, such as Sanger sequencing or long-range PCR. Updating the capturing kit by adding more capturing probes and modifying the target panel by removing some of the most challenging genes has over time helped our lab to improve the sequencing quality of these regions. In addition, several of the areas with systematic errors are in UTRs that are outside of the relevant analysis area, and therefore not reported to requisitioners. The analysis is therefore in practice even more specific than shown here. The calculation nevertheless provides a rough estimate of the specificity of the analysis. The 11 validation samples were chosen because no CNVs had been detected during previous analyses (MLPA) of these samples. However, not all the 126 genes were checked with MLPA in each case, which in principle can give some false negative tests, but we believe that the probability of this is very small. The number of false negatives would in any case be small, and therefore have only minor impact on the estimated specificity.

The level of systematic sequencing errors may also change when changing to a different lot of the capturing kit [26]. This can change the capturing efficiency, and hence change the quality of sequencing data. That is, a region showing systematic errors in the analysis may not have the same systematic errors when moving to a new lot. Conversely, new regions with systematic errors may also arise with the introduction of a new lot, in genes that have not previously shown such errors. To avoid this kind of batch effects, the lot number of capturing kits should therefore be changed as infrequently as possible, and a verification must always be made when introducing a new lot.

The CNV analysis may also be affected by sample properties. In in some rare cases SNPs occurring in the binding site of a probe may affect capturing

of this region, and hence reduce depth. For example, in one of our routine diagnostic samples a mutation (Chr2(GRCh37):g.47643457G > A) in the middle of exon 6 (of length 134 nt) in the *MSH2* gene led to a false signal of deletion of this exon by the pipeline. This type of noise is hard to avoid but important to be aware of and consider by checking for SNPs that can affect probe binding.

The CNV analysis may also be affected by various genomic properties. Genes with repeats or with almost identical pseudogenes are always challenging for short read alignment algorithms in assigning reads to their correct genomic position, due to the ambiguity in placing a read which matches two or more identical regions. Therefore, it is challenging to estimate the correct coverage depth for such genes or regions. For example, exon 11–15 in the *PMS2* gene have duplicated sequences in the *PMS2CL* pseudogene. This can interfere with correct identification of CNVs in these regions, in most cases affecting exons 13–15 of the gene. However, we have correctly detected CNVs for this gene in all our control samples, and also detected and verified it in 5 diagnostic samples. To avoid the risk of false negatives in this gene, it always goes through MLPA test (for the whole gene) and long-range PCR test (for only exons 11–15) for CNV detection. Similarly a *SMAD4* processed pseudogene which consists of only the exonic regions of exons 2–12 of the *SMAD4* gene introduces false signals for CNVs in exons 2–12 for this gene, and not in the introns [30]. These false signals are found not only by the pipeline, but also by MLPA. However, as deletions and duplications are not restricted to exonic sequences, but should also be found in intronic regions, we can identify these CNVs as false signals introduced due to processed pseudogene.

This pipeline has now been used in our routine diagnostic practice for more than two years. Since its implementation in our diagnostics the pipeline has detected different types of challenging CNVs in 16 different genes in 45 diagnostic germline samples, as listed above (Table 2), and several of these genes were previously not tested for CNVs (with MLPA) in our diagnostic practice. This shows that the use of this pipeline has been an important expansion of our capacity for clinical diagnosis. Although most of our use so far has been on DNA extracted from blood, in a few cases the pipeline has also been used on sequencing data generated with DNA extracted from fresh frozen tissue samples. In principle the pipeline can also be used on somatic samples, and as part of our work towards future versions of the pipeline it will be tested and further developed also for the analysis of somatic samples.

Compared to some other tools our pipeline is specially designed to detect smaller CNVs in target panel-based data, e.g., single exonic and partial exonic CNVs. Splitting of larger regions into overlapping sliding windows and the possibility to choose smaller sliding length with respect to window length provides high resolution of CNV results. This improves the detection of small CNV events and predicts the variant boundaries (breakpoints) more accurately. Also, the availability of nucleotide level coverage information has facilitated prediction of exact breakpoints, especially for partial exonic CNVs. Some tools [22, 25] claim to detect CNVs at single exonic level, but it is still challenging to detect partial exonic and mosaic CNVs. Our pipeline has successfully managed to detect such CNVs in routine diagnostics, in addition to exonic CNVs.

Presently the pipeline uses a fixed window size for sliding windows across all regions in the target panel (except for last window of a region and for regions smaller than window size). As a future improvement we are considering whether the sliding window size should be chosen based on the length of each region, and the pattern of sliding windows created accordingly. This will make it possible to use larger sliding windows for larger regions, but also smaller sliding windows for smaller regions. Sliding length may also be selected according to size of the window length. This more dynamic approach can speed up the computation for larger regions, while at the same time giving sufficient resolution of CNV scores for smaller regions.

The CNV score ($\log CNR_{score}$) in our approach has a theoretical value of $+0.58$ for duplications and -1.0 for deletions. As the numeric value of the duplication score is less than the deletion score ($|+0.58| < |-1.0|$), signals for duplications are weaker than for deletions. Interpretation of the pipeline output is based on logCNR scores and their plots, rather than a list of CNV calls. This means that no strict numerical cutoff on logCNR scores is used by our diagnostic lab. This reduces the risk of false negatives due to weak or somewhat noisy signals, and any false positives from this approach will be found by the subsequent experimental verification by sequencing or MLPA. This manual approach to output analysis is doable because most often we are asked to analyze only some of the genes included in the panel (1–15 genes), hence interpretation for this small numbers of genes can easily be managed without using strict cutoffs on CNV score. But for investigating larger sets of queries, like larger target panels with hundreds of genes, or exome panels, certain cutoffs based on statistical analysis will be necessary in order to remove most of the false positive signals caused by noise, to reduce workload and to narrow down investigation towards the most reliable CNV signals. This will be considered for future versions of our pipeline, adapted to large query sets.

To further improve the pipeline, we will in the future also update our approach to sample selection for pool creations. Presently this is based on similarity of coverage depth across samples for creating pools, for selecting pool size, and for selecting the optimal pool for a given query sample. The pattern of coverage depth in target regions remains the same across different samples sequenced from the same lot of a capturing kit. Normalization of the coverage depth of normal samples in pools and of query samples will help in creating a single pool with all available normal samples, which can be used with all query samples. This will also reduce the overhead in the pipeline in creating different pools, and in pool selection for each query sample. However, this also requires a good understanding of optimal approaches for normalization of samples and will therefore be considered mainly for future versions of our pipeline.

Conclusions

We have here described a pipeline for detection of CNVs in NGS sequencing data from targeted gene panels. This pipeline has high sensitivity, specificity, and accuracy, and has already proven its diagnostic value with more than 45 CNV findings in routine diagnostics in our laboratory since August 2018. These findings include partial exonic, single exonic, multi exonic, whole gene and mosaic CNVs, often in genes that previously were not tested, for example because MLPA tests were not available. By using this pipeline our lab has expanded the portfolio of genes up to whole gene panels where we can offer CNV detection, which is important for the quality of our diagnostic work.

Abbreviations

CNV: Copy number variation; CNR: Copy number ratio; TRSW: Target region based sliding windows; MLPA: Multiplex ligation-dependent probe amplification; NGS: Next generation sequencing; PE: Paired-end mapping based detection; SR: Split read based detection; DA: De novo assembly based detection; RD: Read depth based detection; SV: Structural variant; aCGH: Array comparative genomic hybridization; SNP: Single nucleotide polymorphism; PCR: Polymerase chain reaction; FISH: Fluorescence in situ hybridization; UTR : Untranslated region.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12920-021-01059-x.

Additional file 1. Target gene panel, consisting of 126 genes.

Additional file 2. CNV findings (with genomic positions) in 36 control samples.

Additional file 3. Details of false positives and regions of systematic errors and their respective genes in 11 routine samples used for quality control.

Additional file 4. CNV findings (with genomic positions) in 45 diagnostic routine samples.

Acknowledgements

We thank the Department of Medical Genetics, St. Olavs Hospital, Trondheim, Norway for providing access to data for analysis. We also thank Kristine Rosnes for reviewing a draft of our manuscript and providing useful suggestions.

Authors' contributions

AKS wrote the code, implemented the pipeline, performed the analysis, interpreted the results and wrote the manuscript. AKS and MFO conceptualised, designed and administrated the project. TV, LASL and MFO interpreted and validated the results. MFO and LASL participated in manuscript writing. MFO participated in supervision of the project. WS and FD supervised the project and wrote the manuscript. All authors have read and approved the final manuscript.

Funding

This work was done as part of establishing diagnostic methods at Department of Medical Genetics, St. Olavs Hospital, Trondheim, Norway. The department has supported the work. Open Access funding was provided by St. Olavs Hospital. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

Project name: CNV detection in diagnostic gene panel. Project home page: https://github.com/ash9nov/Target-panel-based-CNV-detection. Operating system(s): Unix. Programming language: Shell Scripting, R. Other requirements: None. Licence: GNU GPLv3. Any restrictions to use by non-academics: none. Due to confidentiality and ethical concerns, data cannot be made publicly available. Further information about the data and conditions for access can be provided by the corresponding author AKS (Ashish.Kumar.Singh3@stolav.no; ashish.k.singh@ntnu.no) and by Department of Medical Genetics, St. Olavs Hospital, Trondheim (genetikk@stolav.no).

Declarations

Ethics approval and consent to participate

The project has been classified as a quality assurance audit according to the recommendations given in "Guide for Research Ethics Committee Members" by the Steering Committee on Bioethics (Council of Europe, April 2012). These recommendations are also used by REK, the Regional Ethical Committee (https://rekportalen.no). Evaluation by the local ethics committee is therefore not required. The patients have given written consent to do the diagnostic genetic testing. The Department of Medical Genetics, St. Olavs Hospital has evaluated the use of the data from the genetic testing with respect to anonymity and concluded that the CNV-results listed in the manuscript are anonymous and cannot be traced back to individual patients.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Medical Genetics, St. Olavs Hospital, Trondheim, Norway. [2]Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, NTNU - Norwegian University of Science and Technology, Trondheim, Norway.

Received: 14 May 2021 Accepted: 16 August 2021
Published online: 31 August 2021

References

1. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019;20(1):246.
2. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: new insights in genome diversity. Genome Res. 2006;16:949–61.

3. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12(5):363–76.
4. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17(4):224–38.
5. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. Annu Rev Med. 2010;61(1):437–55.
6. Shen Y, Wu BL. Designing a simple multiplex ligation-dependent probe amplification (MLPA) assay for rapid detection of copy number variants in the genome. J Genet Genom. 2009;36(4):257–65.
7. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet. 2007;39(7S):S16-21.
8. Serin Harmanci A, Harmanci AO, Zhou X. CaSpER identifies and visualizes CNV events by integrative analysis of single-cell or bulk RNA-sequencing data. Nat Commun. 2020;11(1):89.
9. Buysse K, Delle Chiaie B, Van Coster R, Loeys B, De Paepe A, Mortier G, et al. Challenges for CNV interpretation in clinical molecular karyotyping: lessons learned from a 1001 sample experience. Eur J Med Genet. 2009;52(6):398–403.
10. Ito T, Kawashima Y, Fujikawa T, Honda K, Makabe A, Kitamura K, et al. Rapid screening of copy number variations in STRC by droplet digital PCR in patients with mild-to-moderate hearing loss. Hum Genome Var. 2019;6(1):41.
11. Kidd JM, Sampas N, Antonacci F, Graves T, Fulton R, Hayden HS, et al. Characterization of missing human genome sequences and copy-number polymorphic insertions. Nat Methods. 2010;7(5):365–71.
12. Mu W, Li B, Wu S, Chen J, Sain D, Xu D, et al. Detection of structural variation using target captured next-generation sequencing data for genetic diagnostic testing. Genet Med. 2019;21(7):1603–10.
13. Moreno-Cabrera JM, del Valle J, Castellanos E, Feliubadaló L, Pineda M, Brunet J, et al. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. Eur J Hum Genet. 2020;28(12):1645–55.
14. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 2009;6(11S):S13.
15. Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, et al. Mapping copy number variation by population-scale genome sequencing. Nature. 2011;470(7332):59–65.
16. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061–7.
17. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009;19(9):1586–92.
18. van den Akker J, Hon L, Ondov A, Mahkovec Z, O'Connor R, Chan RC, et al. Intronic breakpoint signatures enhance detection and characterization of clinically relevant germline structural variants. J Mol Diagn. 2021;23(5):612–29.
19. Mason-Suares H, Landry LS, Lebo M. Detecting copy number variation via next generation technology. Curr Genet Med Rep. 2016;4(3):74–85.
20. Truty R, Paul J, Kennemer M, Lincoln SE, Olivares E, Nussbaum RL, et al. Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. Genet Med. 2019;21(1):114–23.
21. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. PLOS Comput Biol. 2016;12(4):e1004873.
22. Johansson LF, van Dijk F, de Boer EN, van Dijk-Bos KK, Jongbloed JDH, van der Hout AH, et al. CoNVaDING: single exon variation detection in targeted NGS data. Hum Mutat. 2016;37(5):457–64.
23. Fowler A, Mahamdallie S, Ruark E, Seal S, Ramsay E, Clarke M, et al. Accurate clinical detection of exon copy number variants in a targeted NGS panel using DECoN. Wellcome Open Res. 2016;1:20.
24. Povysil G, Tzika A, Vogt J, Haunschmid V, Messiaen L, Zschocke J, et al. panelcn.MOPS: Copy-number detection in targeted NGS panel data for clinical diagnostics. Hum Mutat. 2017;38(7):889–97.
25. Chiang T, Liu X, Wu T-J, Hu J, Sedlazeck FJ, White S, et al. Atlas-CNV: a validated approach to call single-exon CNVs in the eMERGESeq gene panel. Genet Med. 2019;21(9):2135–44.
26. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11(10):733–9.
27. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. PLoS Biol. 2011;9(7):e1001091.
28. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinform. 2013. https://doi.org/10.1002/0471250953.bi1110s43.
29. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
30. Millson A, Lewis T, Pesaran T, Salvador D, Gillespie K, Gau C-L, et al. Processed pseudogene confounding deletion/duplication assays for SMAD4. J Mol Diagn. 2015;17:576–82.

Publisher's Note

Article 3

Detection of germline variants with pathogenic potential in 48 patients with familial colorectal cancer by using whole exome sequencing

Ashish Kumar Singh[1,2*], Bente Talseth-Palmer[3,4,5], Alexandre Xavier[3], Rodney J Scott[3,5], Finn Drabløs[2], Wenche Sjursen[1,2]

[1]Department of Medical Genetics, St. Olavs Hospital, Trondheim, Norway

[2]Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, NTNU - Norwegian University of Science and Technology, Trondheim, Norway

[3]School of Biomedical Science and Pharmacy, Faculty of Health and Medicine, University of Newcastle and Hunter Medical Research Institute, Newcastle, Australia

[4]Møre and Romsdal Hospital Trust, Research Unit, Ålesund, Norway

[5]NSW Health Pathology, Newcastle, Australia

[*]Corresponding author: Ashish Kumar Singh; E-mail: ashish.kumar.singh3@stolav.no; ashish.k.singh@ntnu.no

Abstract

Background: Hereditary genetic mutations causing predisposition to colorectal cancer are accountable for approximately 30% of all colorectal cancer cases. However, only a small fraction of these are high penetrant mutations occurring in DNA mismatch repair genes, causing Lynch Syndrome. Most of the mutations are low-penetrant variants, contributing to an increased risk of familial colorectal cancer, and they are often found in additional genes and pathways. The aim of this study was to identify such variants.

Methods: We performed whole exome sequencing on constitutional DNA extracted from blood of 48 patients suspected of familial colorectal cancer and used multiple *in silico* prediction tools and available literature-based evidence to detect and investigate genetic variants.

Results: We identified several causative and some possibly causative germline variants in genes known for their association with colorectal cancer. Additionally, several variants have been identified in additional genes, normally not included in relevant gene panels for colorectal cancer, that potentially can be associated with an increased risk for cancer.

Conclusions: Identification of variants in additional genes that potentially can be associated with familial colorectal cancer indicates a larger genetic spectrum of this disease, not limited only to mismatch repair genes. Usage of multiple *in silico* tools based on different methods and combined through a consensus approach increases the sensitivity of predictions and narrows down a large list of variants to the ones that are most likely to be significant.

Keywords: Whole exome sequencing (WES), Colorectal cancer (CRC), Lynch syndrome (LS), Mismatch repair (MMR), Copy number variation (CNV), Variant annotation, Variant filtration.

Background

Cancer is a leading cause of premature mortality in the population (1) with 19.3 million newly diagnosed cases and 10 million deaths worldwide in 2020 (2). Colorectal cancer (CRC) ranks third in cancer incidence, but second with respect to cancer-related mortality (2). Of all CRC cases, 30% are thought to have a familial component but only one third of these are associated with a hereditary condition (3) where high-penetrance pathogenic variants account for their genetic predisposition to disease, e.g., defects in DNA mismatch repair (MMR) genes leading to Lynch syndrome. In addition to the MMR genes (*MLH1, MSH2, MSH6* & *PMS2*), *APC, MUTYH* (biallelic), *NTHL1* (biallelic), and the exonuclease domains of *POLE* and *POLD1* are known high penetrant CRC predisposing genes (4–7). For the remaining 20% of familial CRC causal genetic factors for CRC predisposition remain to be revealed. Next generation sequencing (NGS) and genome-wide association studies (GWASs) have been used to discover the etiology of familial CRC by identifying novel candidate genes and causal variants which have not yet been linked to CRC (4,8). Additionally, whole exome sequencing (WES) has been used to identify bi-allelic and polygenic mutations in Lynch-like cases (5,9–11). Polygenic variation is also recognised as a potential cause of increased disease penetrance in Lynch syndrome (12).

DNA sequencing of protein coding regions enables the study of novel candidate genes and their potential role in cancer risk. The selection of candidate genes can be based on prioritization scores (13). With WES it is possible to expand genomic sequencing beyond just exons, towards 5' untranslated regions (5'UTRs) to capture transcription factor binding sites (TFBSs) and upstream open reading frames (uORFs), and towards 3'UTRs to reveal microRNA binding sites associated with gene regulation. According to GWAS analysis these regions may account for up to 93% of functional variants (14) that are linked to gene regulation.

Genetic variants are often classified as single nucleotide (nt) variation (SNV) (1 nt), short insertion-deletion variation (indel) (up to 50 nt), and structural variation (SV) (larger than 50 nt) (15). Here SVs include insertions, deletions, duplications, inversions, translocations or a combination of these, co-occurring in a single genome (16). Deletions and duplications of SVs (17,18), known as copy number variants (CNVs), have been associated with disease and can contribute to a large fraction of the disease-causing variation (19). WES has mainly been used to detect disease-causing SNV/indel variants (20). However, with the use of recently developed *in silico* methods it is possible to identify also CNVs from WES data (21).

An essential step in NGS data analysis is the assessment of a variant's effect on gene function and any causative association with disease. This is achieved by assigning annotations consisting of both theoretical pathogenicity scores calculated by prediction tools and experimental data extracted from various databases. Annotation tools can provide a diverse set of annotations in one place (22,23), and these annotations can then be used to filter down large lists of variants to the most significant ones.

In the present study, WES was performed on constitutional DNA extracted from blood of 48 patients with suspected familial colorectal cancer. Variant calling was undertaken to detect SNVs, indels and CNVs in all target regions of the exome. Consensus prediction based on multiple *in silico* tools and literature-based evidence was used to search for disease association of detected variants. We identified several potentially causative germline pathogenic variants in genes known to be associated with colorectal cancer. Additionally, several variants were identified in genes normally not included in gene panels for colorectal cancer, and these may be associated with an increased cancer risk.

Methods

Samples and study design

Germline DNA were extracted from blood samples from 48 Australian patients diagnosed with colorectal cancer fulfilling the Amsterdam-II Criteria for Lynch syndrome, including 16 related individuals from 8 families while 32 individuals were unrelated individuals. Sanger sequencing performed previously detected no germline MMR mutations in the samples. Therefore, these patients are defined as Lynch-like syndrome (LlS) patients.

Whole exome sequencing (WES)

WES was performed on germline DNA from these 48 samples. Paired-end library preparation was performed using the Illumina Truseq exome capturing kit. DNA was sheared to ~150bp using the Bioruptor Pico (Diagenode) followed by the recommended protocol using a single index. The final libraries were sequenced using an Illumina Nextseq 500 kit (Illumina), 150 cycles pair ended. Libraries were quantified using Qbit High Sensitivity D100 (Agilent) and were checked using either TapeStation on Bioanalyzer (Agilent) for quality and size.

Variant calling and annotation

SNV/indel variant calling was performed on the dataset using a standardized BWA-Picard-GATK pipeline (20). Joint annotation of variants was performed using the command-line based batch annotation software tool Ensembl variant effector prediction (VEP) (22) complimented with additional annotations from database dbNSFP (24,25) used as plugin with VEP.

Detection of CNVs was performed using an in-house developed method (26) for detecting CNVs in targeted sequencing data.

Variant prioritization

Prioritization steps were performed on the initial set of 125.686 variants detected from variant calling on 48 samples, using the command-line based tool filter_vep from the VEP toolkit. This was performed

in three stages. In stage one, variants were selected based on their occurrence in the population database gnomAD (13). In the second stage, variants were classified and prioritized based on their clinical significance assigned in the ClinVar database (27). In the third and final stage, variants passing through the previous two stages were filtered based on pathogenicity estimation scores of selected tools. This included REVEL (28), CADD (29), ClinPred (30), M-CAP (31), VEST4 (32), MetaSVM (33), BayesDel (34) for missense, nonsense and start-loss prediction; SpliceAI (35) for splicing alteration prediction; and Loftee (13) for loss of function prediction.

Selection of *in silico* prediction tools was based on ranking generated by our benchmarking study comparing the performance of 45 different pathogenicity prediction tools (see Supplementary file 1). We also took into consideration other benchmarking studies with similar goals (36–38). Fig. 1 shows workflow for these filtering steps and the outcome of each step. For detailed information about the various filtering steps, please see Supplementary file 2.

Results

In all 48 samples, on average more than 99% of reads aligned to the reference genome GRCh37, with an average coverage depth of 92X. A total of 125.686 SNP/indel variants (for 25.664 genes) were identified in the 48 samples after the variant calling step. The three-stage filtering strategy detailed above (also displayed in Fig. 1) was applied to these variants, resulting in 346 variants (for 302 genes) with variants in different filtering categories. These variants were assigned to different pathogenicity classes according to the ClinVar database. Table 1 displays a breakdown of these 346 variants into the number of variants for the different filtering steps. The full list of these variants is listed in Supplementary file 3.

Table 1: Number of variants as outputs from different filtering stages.

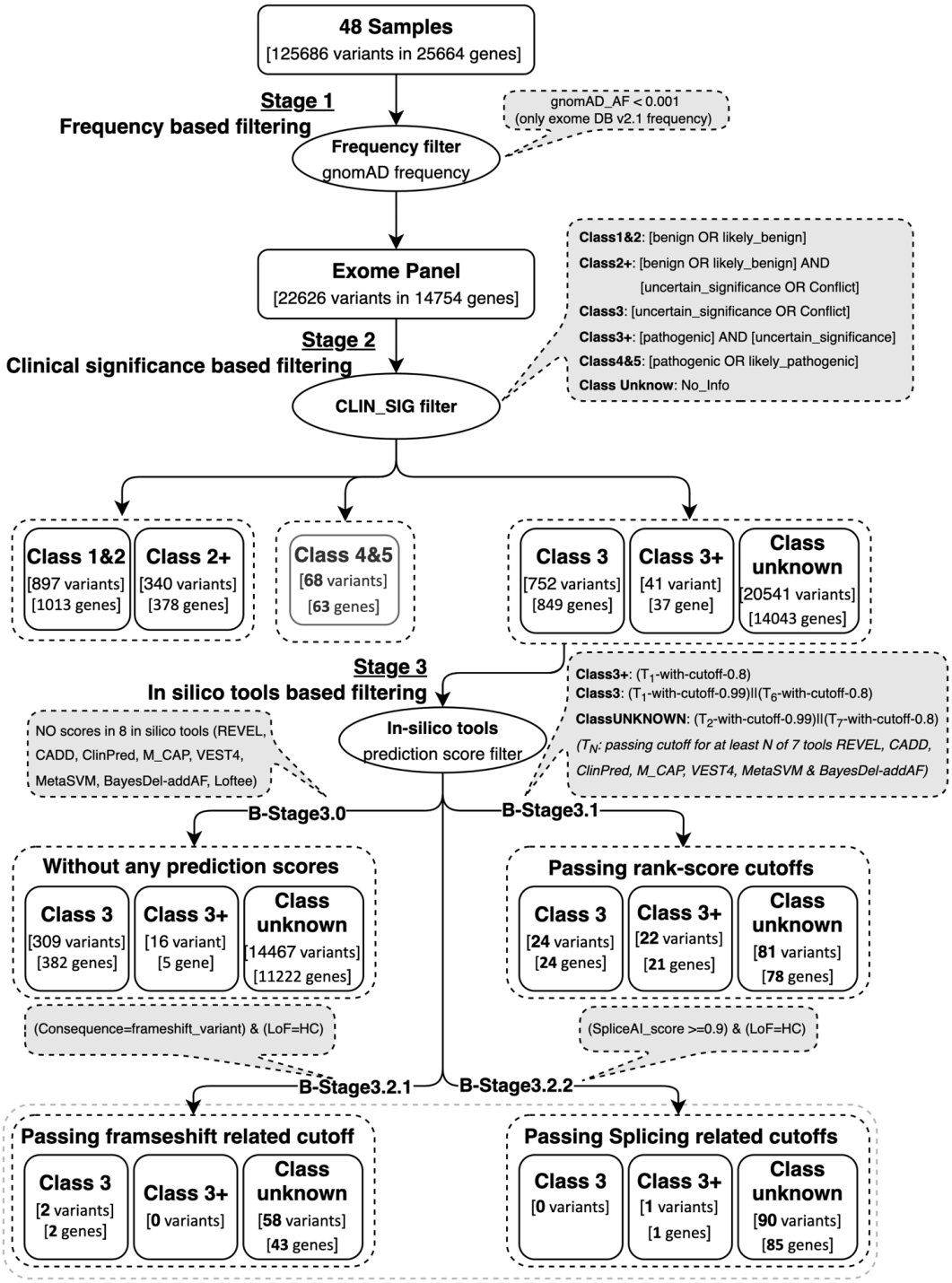
Variant calling	125686 variants (in 25664 genes)					
Stage 1	22626 variants (in 14754 genes)					
	Class Unknown	**Class 3**	**Class 3+**	**Class 4&5**	**Class1&2**	**Class 2+**

Stage 2			20541 variants (14043 genes)	752 variants (849* genes)	41 variants (37 genes)	**68** variants (91* genes)	897 variants (1013* genes)	340 variants (378* genes)
Stage 3	**3.0**		14467 variants (11222 genes)	309 variants (382* genes)	16 variant (5 genes)	No Stage 3 filtering in these classes (4&5, 1&2, 2+)		
	3.1		**81** variants (78 genes)	**24** variants (24 genes)	**22** variants (21 genes)			
	3.2	**3.2.1**	**90** variants (85 genes)	0	**1** variant (1 gene)			
		3.2.2	**58** variants (43 genes)	**2** variants (2 genes)	0			

*Number of genes is higher than number of associated variants, due to multiple naming of some genes.

Filtering stages: Stage 1: Frequency (gnomAD) based filtering; Stage 2: Clinical significance (ClinVar) based filtering; Stage 3: Chosen *In silico* tools-based filtering, 3.0: Variants without any scores in chosen *in silico* tools, 3.1: Variants passing tool rank-scores cut-offs, 3.2.1: Variants passing splicing related filters, 3.2.2: Variants passing frameshift related filters. See Fig. 1 for further details and explanation of class 2+ and 3+.

Fig. 1: Variants filtering workflow, illustrating all filtering stages and their outcomes.



To identify any associations of these variants with cancer all 302 genes were checked against three cancer-associated databases; COSMIC (39), OncoKB (40) and TSGene (41). Only 38 of the 302 genes were listed in at least one of these databases. All these 38 genes are either known or have expected roles in cancer as oncogenes, tumor suppressor genes or fusion genes according to database classification. The list of the 38 genes with their respective roles in cancer are given in Supplementary Table S1. Of the 346 variants passing the filtering stages, only 46 are associated with these 38 genes. Among these 46 variants, 14 are pathogenic/likely pathogenic, 6 are VUS, and 4 have conflicting interpretation between pathogenic/likely pathogenic and VUS according to ClinVar. The remaining 22 variants are not reported in ClinVar. Thirty-three of the 48 samples carried one or more of these 46 variants, the remaining 15 samples did not harbor any variant with a known cancer association, and hence lacked a clear link to an established cancer-associated variant. Table 2 lists these 33 samples and the associated 46 variants.

Table 2: List of 33 samples and associated 46 variants.

ID (Fid)	CRC diagnosis age; OC (age)	gNomen, cNomen (pNomen), Existing variation	ClinVar	No. of samples
S.02 (F.8)	30s; BrC (30)	NM_022552.5(***DNMT3A***):c.2210T>C (p.Leu737Pro)	NR	1
S.03 (F.7)	*; OvC (78)	NM_000059.4(***BRCA2***):c.2808_2811del (p.Ala938Profs*21), rs80359351	P	2 (F.7)
		NM_006293.3(***TYRO3***):c.1660+1G>C (p.?), rs757748573	NR	3
		NM_001009944.3(***PKD1***):c.6605C>T (p.Ala2202Val), rs764264106	VUS	1
		NM_002894.3(***RBBP8***):c.298C>T (p.Arg100Trp), rs373804633	P	1
S.04 (F.7)	31;	NM_000059.4(***BRCA2***):c.2808_2811del (p.Ala938Profs*21), rs80359351	P	2 (F.7)
		NM_003331.5(***TYK2***):c.1011+2T>G (p.?), rs1463636749	NR	1
		NM_002568.4(***PABPC1***):c.739-1G>A (p.?), rs759516741	NR	7
S.08	48;	NM_000492.4(***CFTR***):c.1392G>T (p.Lys464Asn), rs397508198	P	5
S.09	65; UtC (50)	NM_007289.4(***MME***):c.467del (p.Pro156Leufs*14), rs749320057	P	1
		NM_001128840.3(***CACNA1D***):c.1750G>A (p.Val584Ile), rs773365038	VUS	1
S.11	60; BoC (32)	NM_001349338.3(***FOXP1***):c.179A>G (p.Gln60Arg), rs374060287	LP	1
S.13 (F.1)	50;	NM_006343.3(***MERTK***):c.1450G>A (p.Gly484Ser), rs527236084	VUS	1
		NM_002568.4(***PABPC1***):c.739-1G>A (p.?), rs759516741	NR	7
S.14	52; BrC (52)	NM_005168.5(***RND3***):c.349-2A>T (p.?), rs1222374113	NR	1
S.15	42;	NM_000088.4(***COL1A1***):c.4066C>A (p.Arg1356Ser), rs1341595487	VUS	1

		NM_002568.4(***PABPC1***):c.739-1G>A (p.?), rs759516741	NR	7
		NM_002568.4(***PABPC1***):c.388-1G>A (p.?), rs771446357	NR	1
S.16	48; CRC (67)	NM_004431.5(***EPHA2***):c.2162G>A (p.Arg721Gln), rs116506614	CI	1
		NM_000492.4(***CFTR***):c.1392G>T (p.Lys464Asn), rs397508198	P	5
		NM_002568.4(***PABPC1***):c.739-1G>A (p.?), rs759516741	NR	7
S.17	72; KC (71)	NM_002911.4(***UPF1***):c.2474G>T (p.Ser825Ile)	NR	1
		NM_006941.4(***SOX10***):c.718A>C (p.Thr240Pro), rs1332625359	VUS	1
S.19	21; CRC (40)	NM_006293.3(***TYRO3***):c.1660+1G>C (p.?), rs757748573	NR	3
		NM_000535.7(***PMS2***):c.614A>C (p.Gln205Pro), rs587779342	CI	2
		NM_000535.7(***PMS2***):c.1A>G (p.Met1Val), rs587779333	P/LP	2
		NM_002568.4(***PABPC1***):c.739-1G>A (p.?), rs759516741	NR	7
S.20	55; EC (41)	NM_002693.3(***POLG***):c.2209G>C (p.Gly737Arg), rs121918054	P/LP	1
		NM_022552.5(***DNMT3A***):c.1122+2T>G (p.?), COSV53057339	NR	1
S.21	*; PC (62)	NM_000059.4(***BRCA2***):c.7977-1G>C (p.?), rs81002874	P	1
		NM_052839.4(***PANX2***):c.1479dup (p.Gly494Argfs*13)	NR	1
S.23 (F.5)	51;	NM_000400.4(***ERCC2***):c.1480-1G>C (p.?), rs375284572	NR	1
		NM_000249.4(***MLH1***):c.514G>A (p.Glu172Lys), COSV51617106	NR	2 (F.5)
S.24	50;	NM_033084.5(***FANCD2***):c.1588C>T (p.Arg530*), rs962867926	NR	1
		NM_004625.4(***WNT7A***):c.874C>T (p.Arg292Cys), rs104893835	P	1
		NM_006424.3(***SLC34A2***):c.1267G>A (p.Gly423Arg), rs769110830	NR	1
S.25 (F.5)	52;	NM_000249.4(***MLH1***):c.514G>A (p.Glu172Lys), COSV51617106	NR	2 (F.5)
		NM_004168.4(***SDHA***):c.762_770+17del (p.Ala255_Gly257del), rs1041809852	P	1
S.27	*; UrC (60)	NM_000492.4(***CFTR***):c.2723C>A (p.Thr908Asn), rs369521395	P	1
		NM_002568.4(***PABPC1***):c.739-1G>A (p.?), rs759516741	NR	7
S.29	48; UtC (56)	NM_002568.4(***PABPC1***):c.739-1G>A (p.?), rs759516741	NR	7
S.30	68; UrC (79)	NM_004963.4(***GUCY2C***):c.612-1G>A (p.?), rs763904634	NR	1
		NM_006293.3(***TYRO3***):c.308+1G>C (p.?), rs764446020	NR	1
		NM_000179.3(***MSH6***):c.3724_3726del (p.Arg1242del), rs63749942	P/LP	1
		NM_000492.4(***CFTR***):c.1392G>T (p.Lys464Asn), rs397508198	P	5
		NM_001001548.3(***CD36***):c.1202_1205del (p.Val401Glufs*4), rs769354931	CI	1
S.31	57;	NM_006092.4(***NOD1***):c.689T>G (p.Phe230Cys), CM1612670	NR	1
S.32	58; PC (60)	NM_007371.4(***BRD3***):c.71dup (p.Glu25Glyfs*51), rs768970491	NR	1
S.33	63; EC (29)	NM_002568.4(***PABPC1***):c.367G>T (p.Gly123Cys), rs755674364	NR	1
S.34 (F.2)	48;	NM_001371290.1(***ZBTB7C***):c.402_403insC (p.Glu135Argfs*4)	NR	1
		NM_000492.4(***CFTR***):c.1392G>T (p.Lys464Asn), rs397508198	P	5
S.36 (F.2)	*; UtC (41)	NM_145728.3(***SYNM***):c.2523del (p.His842Thrfs*47), COSV60376961	NR	1

		NM_000535.7(***PMS2***):c.614A>C (p.Gln205Pro), rs587779342	CI	2
		NM_000535.7(***PMS2***):c.1A>G (p.Met1Val), rs587779333	P/LP	2
S.37 (F.4)	38;	NM_006293.3(***TYRO3***):c.1660+1G>C (p.?), rs757748573	NR	3
S.38	45;	NM_024415.3(***DDX4***):c.673+2T>C (p.?), rs201596382	NR	1
S.39	77; RC (51)	NM_000251.3(***MSH2***):c.2228C>G (p.Ser743*), rs63751155	P	1
S.40	66; SC(-)	NM_006293.3(***TYRO3***):c.1483+2T>C (p.?), rs138345868	NR	1
S.43	51; CRC (64)	NM_002335.4(***LRP5***):c.3562C>T (p.Arg1188Trp), rs141178995	P	1
S.44 (F.6)	33; UtC (-)	NM_000264.5(***PTCH1***):c.104G>A (p.Arg35Gln), rs587778627	VUS	1
S.47 (F.3)	34; Mel (32)	NM_000492.4(***CFTR***):c.1392G>T (p.Lys464Asn), rs397508198	P	5
S.48 (F.3)	*; RC (53)	NM_017563.5(***IL17RD***):c.392A>C (p.Lys131Thr), rs184758350	CI	1

*No colorectal cancer.

Abbreviations: ID: patient ID; **Fid**: family ID; **OC**: other cancers. **LP**: likely pathogenic; **P**: pathogenic; **VUS**: uncertain significance; **LB**: likely benign; **NR**: not reported; **CI**: conflicting interpretations (P/LP; VUS). **LS**: lynch syndrome; **LlS**: lynch-like syndrome; **CRC**: colorectal cancer; **LC**: lung cancer; **UtC**: uterine cancer; **Mel**: melanoma; **RC**: renal cancer; **OvC**: ovarian cancer; **BrC**: breast cancer; **UrC**: ureteral cancer; **KC**: kidney cancer; **PC**: pancreatic cancer; **SC**: stomach cancer; **BoC**: bone cancer; **BlC**: bladder cancer; **EC**: endometrial cancer.

Among these 38 genes, 7 are well known cancer genes with high impact towards cancer. These included *BRCA2, MLH1, MSH2, MSH6, PMS2, PTCH1* & *SDHA*. These 7 genes have 9 variants with pathogenicity classes 5, 4 or 3, occurring in 10 patients. Table 3 lists these genes and the associated variants.

Table 3: List of known cancer genes and associated variants.

Gene	Linked to cancer	Variant	ACMG-AMP	ID (Fid)	CRC diagnosis age; OC (age)
BRCA2 (NM_000059.4)	BrC	c.2808_2811del (p.Ala938Profs*21), rs80359351	Class 5	S.03 (F.7)	*; OvC (78)
				S.04 (F.7)	31;
		c.7977-1G>C (p.?), rs81002874	Class 5	S.21	62*; PC (62)
MLH1 (NM_000249.4)	CRC	c.514G>A (p.Glu172Lys), COSV51617106	Class 3	S.23 (F.5)	51;
				S.25 (F.5)	52;
MSH2 (NM_000251.3)	CRC	c.2228C>G (p.Ser743*), rs63751155	Class 5	S.39	77; RC (51)

MSH6 (NM_000179.3)	CRC	c.3724_3726del (p.Arg1242del), rs63749942	Class 4	S.30	68; UrC (79)
PMS2 (NM_000535.7)	CRC	c.614A>C (p.Gln205Pro), rs587779342	Class 4	S.19	21; CRC (40)
				S.36 (F.2)	*; UtC (41)
		c.1A>G (p.Met1Val), rs587779333	Class 5	S.19	21; CRC (40)
				S.36 (F.2)	*; UtC (41)
PTCH1 (NM_000264.5)	GS	c.104G>A (p.Arg35Gln), rs587778627	Class 3	S.44 (F.6)	33; UtC (-)
SDHA (NM_004168.4)	PG	c.762_770+17del (p.Ala255_Gly257del), rs1041809852	Class 3	S.25 (F.5)	52;

*No colorectal cancer

Abbreviations: ID: patient ID; **Fid**: family ID; **OC**: other cancers**. LP**: likely pathogenic; **P**: pathogenic; **VUS**: uncertain significance; **LB**: likely benign; **CI**: conflicting interpretations (P/LP; VUS). **LS**: lynch syndrome; **LlS**: lynch-like syndrome; **CRC**: colorectal cancer; **BrC**: breast cancer; **GS**: gorlin syndrome; **PG**: paraganglioma

Copy number variant calling

We performed CNVs calling step only for 88 known cancer genes. List of these genes is provided as Supplementary file 4. We detected 5 CNVs associated to 5 genes including 1 deletion and 4 duplications in 5 patients. More details are available as Supplementary Table S2.

Discussion

This study uses a WES-based approach to identify the genetic causes of disease in LlS patients, where the majority of the patients had been diagnosed with CRC, and all fulfilling the AMS criteria. Most of the patients in this cohort were at the time of genetic testing pre-screened using denaturing high performance liquid chromatography (DHPLC) prior to Sanger sequencing. DHPLC is an inferior screening method that does miss some genetic variants and hence some samples were not further processed for Sanger Sequencing. The use of WES was therefore performed to identify relevant variants in additional cancer-associated genes, as well as the MMR genes. We identified significant variants in 38 genes known for cancer associations; this included 7 well established cancer genes with high cancer penetrance. Because of deficiencies in DHPLC pre-screening some of the patients were shown by WES to carry pathogenic MMR variants.

Four patients harbored pathogenic MMR variants, and therefore have a molecular diagnosis of Lynch syndrome. In addition, one suspicious VUS (*MLH1* c.514G>A) was detected in two members of family 5 and may also represent Lynch syndrome.

The *MLH1* variant c.514G>A (p.Glu172Lys) found in two patients from the same family (S.23 and S.25, family F.5), is a VUS with the potential to be pathogenic. Immunohistochemistry showed missing protein staining for *MLH1* and *PMS2* in tumors from both family members. The variant is not reported in gnomAD, and the REVEL score (0.876) indicate pathogenicity. Residue Glu172 is highly conserved and located in the ATPase domain of *MLH1*, within an α-helix structure. The switch from Glu to Lys results in a change from acidic to basic residue, which may disrupt the α-helix. In addition, this variant has been observed as a somatic change in three carcinomas (COSMIC database) of the breast, endometrium and large intestine.

PMS2 has two mutations c.614A>C (p.Gln205Pro) and c.1A>G (p.Met1Val) classified as class 4 and class 5 respectively. Both were found in two unrelated patients, S.19 and S.36. These two variants have also been detected in one patient by a previous study (42). For variant c.614A>C functional studies have demonstrated significantly higher repair efficiency than that of a pathogenic control, but 50% compromised when compared to wild type (43). Biallelic defects in MMR genes are known as constitutive mismatch repair defect (CMMRD), and CMMRD patients often have more severe phenotypes than Lynch syndrome patients have. Previous studies have identified biallelic pathogenic *PMS2* mutations driven CMMRD leading to cancers in younger patients (44–46). Patient S.19 was diagnosed with CRC at early age of 21 years and a second CRC at age of 40 years, whereas patient S.36 was diagnosed with uterine cancer at 41, and she did not develop CRC. We are not able to distinguish whether the two *PMS2* variants are biallelic or in cis (same allele) in these two patients. Gene *SDHA* has variant c.762_770+17del (p.Ala255_Gly257del) in patient S.25 (F.5), a deletion of three amino acids predicted to cause loss of a splice donor site (SpliceAI score:1). Loss of donor splice site is predicted to disrupt RNA splicing and culminate in either the absence or disruption of the protein product. As a tumor suppressor gene, *SDHA* is more likely to be associated with neuroendocrine related cancers, more commonly paragangliomas, with germline mutations accounting for 7.6% of patients with this cancer type (47).

Family 7 harbor a pathogenic *BRCA2* mutation c.2808_2811del causing hereditary breast and ovarian cancer. One of the two included family members had been diagnosed with ovarian cancer, while extended family members had breast and ovarian cancer, in addition to CRC. There has been a discussion whether pathogenic *BRCA1/2* variants are associated with CRC. However, a recent meta-analysis concluded that *BRCA1* and/or *BRCA2* mutation carriers are not at a higher risk of colorectal cancer (68).

The remaining 31 genes are associated with a variety of different roles including tumor suppressor genes, oncogenes, and fusion genes in various types of cancers (Supplementary Table S1). The identification of variants in these genes that have not previously been associated with familial CRC suggests a larger spectrum of genetic variants associated with this disease that is not limited to DNA mismatch repair genes or other known cancer-associated genes. Among these 31 candidate genes *CFTR*, *PABPC1*, and *TYRO3* have variants over-represented in this patient cohort.

We identified two pathogenic variants in gene *CFTR* (NM_000492.4); c.2723C>A (p.Thr908Asn) and c.1392G>T (p.Lys464Asn). Variant c.2723C>A is occurring in one patient whereas c.1392G>T is occurring in five patients, all five have CRC. That the pathogenic variant c.1392G>T (p.Lys464Asn) is over-represented in this cohort of cancer patients indicates that it could contribute to CRC development, but this needs further investigation. Previously, *CFTR* has primarily been associated with cystic fibrosis (CF) (a recessive disease), but has recently been categorized as a CRC risk gene (48). This is seen as a result of CF patients surviving long enough to develop CRC. In addition, recent evidence indicates that low expression levels of *CFTR* is associated with a significant risk towards CRC (49).

Variants in the gene *PABPC1* (NM_002568.4) were found in nine patients, and the most frequently occurring variant (c.739-1G>A) was found in seven of these. The gene product of *PABPC1* is PABP-1, which is a poly(A) binding protein involved in several aspects of mRNA metabolism, including splicing of pre-mRNA, initiation of translation of mRNA, and mRNA decay (50). *PABPC1* has been shown to be an oncogene that is upregulated in gastric carcinoma, where high expression predicts poor survival (51). However, for esophageal cancer it has been shown that reduced expression of *PABPC1* correlates with tumor progression and poor prognosis after surgery (52), indicating a complex relationship between *PABPC1* expression levels and cancer. Recently *PABPC1* has been identified as

a putative CRC driver gene in some patients (53). Several domains have been identified in the protein, including four RNA recognition motif (RRM) domains, and a PAB C-terminal (PABC) domain that can bind interacting proteins(50). One of the variants revealed in this study (rs759516741) has been classified as a splice acceptor, and the variant is located at position -1 relative to the start of exon 6 (NM_002568.4:c.739-1G>A). This exon is coding for residues 247 to 292 of the protein sequence, which overlaps partly with the third RRM domain (191-268). The variant may therefore affect the RRM 3 domain as well as domains further downstream, RRM 4 (294-370) and PABC (542-619). However, it is difficult to estimate how this may affect the function of PABP-1 and any processes where it is involved.

For *TYRO3* (NM_006293.3) three different variants in five patients were identified. The most frequent (c.1660+1G>C) was found in three patients. TYRO3 is a receptor tyrosine kinase of 890 residues, and signals are transduced into the cytoplasm when extracellular ligand binding induces dimerization and autophosphorylation of its intracellular domain (54). TYRO3 acts as an oncogenic protein (55). Overexpression has been observed in several cancers and is associated with a poor prognosis. Somatic mutations have also been observed, but without validation of their effect (54). Regulation of *TYRO3* in CRC by specific non-coding RNA molecules has recently been documented (56,57), and these studies also highlight the clear relationship between *TYRO3* overexpression and cancer. The most frequent variant in this dataset (rs757748573) has been classified as a splice donor variant. It is found at position +1 relative to the end of exon 13 (NM_006293.3:c.1660+1G>C). The start of exon 14 corresponds to position 553 of the protein, which is in the intracellular domain (451-890). This means that the variant may affect signal transduction. However, whether that can give a similar metabolic effect as a general overexpression of *TYRO3* and activation of the protein is difficult to predict.

One among five detected CNVs, *RB1* (NM_000321.2) ex6.del, will most probably cause frameshift and affect the function of the gene. However, *RB1*, a tumor suppressor gene, often retains higher expression levels compared with adjacent normal tissue in CRC cells (58). It is less likely that this CNV is associated to CRC and hence not significant. For the other four detected CNVs we could not establish any functional significance towards CRC.

Variant calling in whole exome regions identified 125.686 SNPs/indels in 25.664 genes. In order to focus on the variants that were most likely to have an effect on gene function, a set of strict filtering criteria were used, see Methods for details. This led to a relatively much smaller set of 346 variants passing these filters, which uses both very high cut-off values of individual tools as well as consensus predictions from multiple tools based on different prediction approaches. Although the use of very strict filtering criteria increases the chance of detecting variants which are more likely to have a negative effect on gene function, it will also increase the risk of missing significant variants, causing a bias in the study. For example, a known (likely pathogenic) variant NM_000251.3(*MSH2*): (p.Ala689Asp) was c.2066C>A identified in sample S.20. Even though it had very high pathogenicity scores by all seven prediction tools, it was filtered out by class unknown filters (*[T_2-cutoff-0.99]* || *[T_7-cutoff-0.8]*) by a very small margin (scored 0.78 rank score by CADD-raw). However, even a small adjustment towards less stringent filtering allowing this variant to pass, would also have increased the number of unknown variants after filtering by a minimum two fold. The number of variants passing different combination filters based on rank-scores of 7 *in silico* tools is provided as Supplementary Table S3. Additionally, the *MSH2* c.2066C>A variant assigned as unknown (not reported) by VEP-based offline ClinVar annotation is a VUS according to the most recent online ClinVar records. Using this more recent ClinVar classification as VUS, i.e., a class 3 variant, it passes the class 3 filters (*[T_6-cutoff-0.8]* || *[T_1-cutoff-0.99]*). Hence, it is not only strict filtering but also the discrepancy between offline and online annotation records which may lead to a loss of significant variants during filtering. Another variant, *PTCH1* c.104G>A (p.Arg35Gln) a VUS passed the class 3 filters. Mutations in *PTCH1* can cause nevoid basal cell carcinoma syndrome (NBCCS) an autosomal dominant disorder commonly known as Gorlin syndrome (59). This variant (c.104G>A) has very low rank-scores in all selected *in silico* tools except in M-CAP (rank-score of 0.99639) which let it pass the filtering. A stricter filtering may have removed this variant from the final list. These examples demonstrate the challenges of setting up filters for large datasets, e.g., from whole genome or exome sequencing. This may be a problem mainly in more explorative analysis where exome or genome wide data are analysed and relatively stricter filtering is required in order to keep the number of variants at a manageable level. This will normally not be the case in diagnostic settings, where fewer and mainly well-characterized genes are screened.

Hence a relatively smaller number of variants are used as input for filtering, and less strict filtering criteria may be applied. Variants passing filtering stage 3.0, i.e., variants without any prediction scores, were not included in final list. Class 3 & 3+ variants of this stage were briefly checked for any significance, but none were found. There is a large number of variants in class unknown passing filtering in this stage, which can be used in future studies.

Mean coverage of these 48 samples was 92X where 84.3% of all variants in these samples had coverage depth higher than 30X. But one of these samples (S.36) had low coverage depth of 9X. However, detected variants in this sample were known to variant databases ClinVar (27) and dbSNP (60) with enlisted phenotypic effects matching the patient's phenotype. This supports our findings in this patient sample.

Validation of variants with an alternative technique (i.e., Sanger sequencing) could not be performed because most of the sample material has been exhausted. However, given the high accuracy of present day NGS-based detection of SNV/indel variants, additional validations are often not necessary (61).

In this cohort of 48 patients, 33 patients have variants in genes with known associations to cancer. Only 7 of these 33 patients have pathogenic or likely pathogenic variants (classified according to ACMG-AMP guidelines) in known cancer genes and hence have a confirmed causative variant associated with LS/LlS (Table 3). For 26 patients we detected significant variants in candidate genes with a potential to be associated with familial CRC. In the remaining 15 patients, we have not detected any significant variants passing our filtering criteria in known or candidate cancer-association genes. A possible explanation for missing variants in these 15 samples is the strict filtering criteria. Less strict filtering is one possible approach to identify significant variants in these samples. It is also possible to incorporate the combined effect of multiple variants as causative factor for disease susceptibility. The co-occurrence of multiple rare low-to-moderate risk alleles are likely to be associated with a complex genetic predisposition (62), as the combined effect of common low-risk loci is currently estimated to be up to 15% of the familial risk for cancer (63). Polygenic risk score based models are one of the latest methods utilizing this approach (64). Additionally, with exome sequencing deep intronic mis-splicing variants may be missed, and such variants also contribute towards cancer (65). We also have not included variants in regulatory regions, e.g., variants in uORF (up-stream open reading frames) in our analysis,

mainly because of very sparse annotation for such variants. This is mainly due to the fact that these regions are not commonly sequenced in targeted sequencing, hence annotation data for relevant tools (e.g., UTRannotator (66)) is very sparse. In addition to these factors, there are many more that can also lead to a missed molecular diagnosis (67) for these 15 samples, e.g., somatic mosaicism, epigenetic inheritance, technological limitations, non-genetic risk factors and the fact that the clinical diagnosis may be incorrect due to insufficient information.

Conclusions

In this study we have used whole exome sequencing (WES) to identify germline variants with a pathogenic potential in patients with familial CRC and Lynch-like syndrome. This provides an opportunity to identify important variants in the full set of genes, not limited to a predefined subset of genes from a gene panel. However, it also gives very large lists of variants where most are of uncertain significance. The use of consensus predictions for pathogenicity by combining multiple *in silico* tools based on different approaches helps in narrowing down the list to the variants that are most likely to affect gene function. Although a strict approach means that important variants may be missed out from detection, such filtering is still an essential step in the analysis of WES data. Our analysis identified possibly pathogenic variants in genes that have not previously been associated with familial CRC that warrants further investigation to establish any potential role of these genes with respect to CRC, the results indicate that a larger spectrum of genes and genetic variants may be associated with this disease, not limited to the usual suspects like the DNA MMR genes.

List of abbreviations

CRC: Colorectal cancer

MMR: Mismatch repair

NGS: Next generation sequencing

WES: Whole exome sequencing

TFBS: transcription factor binding sites

SNV: Single nucleotide variation

Indel: Insertion-deletion variation

CNV: Copy number variant

LIS: Lynch-like syndrome

DHPLC: high performance liquid chromatography

CMMRD: constitutive mismatch repair defect

Declarations

Ethics approval and consent to participate

The study was performed according to the Helsinki Declaration and approved by the Hunter New England Human Research Ethics Committee, Australia (04/03/10/3.11) and Regional Ethics Committee (REK), Norway (2015/838). Written informed consent was obtained from all participants.

Consent for publication

Not applicable.

Availability of data and materials

The raw data of whole-exome sequencing of the patients in this study and the full list of variants called are not publicly available to protect participant confidentiality. All variants after filtering are given in the paper. Further information about the data and conditions for access can be provided by the corresponding author AKS (Ashish.Kumar.Singh3@stolav.no; ashish.k.singh@ntnu.no) and by NSW Health Pathology, Newcastle, Australia (https://www.pathology.health.nsw.gov.au/).

Competing interests

The authors declare that they have no competing interests.

Funding

This work has been funded by Cancer Institute NSW (Grant-number:12/ECF/2-34); Liaison committee between Helse Midt-Norge RHF and NTNU. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

AKS performed the analysis, interpreted the results and wrote the manuscript. AKS, WS and FD conceptualised and designed the project. RJS and BTP administrated the project and participated in manuscript writing. AX performed the experiments. WS and FD supervised the project and participated in manuscript writing. All authors have read and approved the final manuscript.

Acknowledgements

We thank all the patients for their consent to donate biological samples for the study.

Authors' information

AKS, WS and FD are affiliated to **Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, NTNU - Norwegian University of Science and Technology, Trondheim, Norway**

AKS and WS are affiliated to **Department of Medical Genetics, St. Olavs Hospital, Trondheim, Norway**

BTP, AX and RJS are affiliated to **School of Biomedical Science and Pharmacy, Faculty of Health and Medicine, University of Newcastle and Hunter Medical Research Institute, Newcastle, Australia**

BTP and RJS are affiliated to **NSW Health Pathology, Newcastle, Australia**

BTP is associated to **Møre and Romsdal Hospital Trust, Research Unit, Ålesund, Norway**

Corresponding author

Correspondence to AKS (Ashish.Kumar.Singh3@stolav.no; ashish.k.singh@ntnu.no).

References

1. Bray F, Laversanne M, Weiderpass E, Soerjomataram I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer. 2021 Aug;127(16):3029–30.
2. H S, J F, RL S, M L, I S, A J, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021

May;71(3):209–49.

3. Mao R, Krautscheid P, Graham RP, Ganguly A, Shankar S, Ferber M, et al. Genetic testing for inherited colorectal cancer and polyposis, 2021 revision: a technical standard of the American College of Medical Genetics and Genomics (ACMG). Genet Med 2021 2310. 2021 Jun;23(10):1807–17.
4. Valle L, de Voer RM, Goldberg Y, Sjursen W, Försti A, Ruiz-Ponte C, et al. Update on genetic predisposition to colorectal cancer and polyposis. Mol Aspects Med. 2019 Oct;69:10–26.
5. Fernández-Rozadilla C, Álvarez-Barona M, Quintana I, López-Novo A, Amigo J, Cameselle-Teijeiro JM, et al. Exome sequencing of early-onset patients supports genetic heterogeneity in colorectal cancer. Sci Reports 2021 111. 2021 May;11(1):1–9.
6. Valle L, Vilar E, Tavtigian S V, Stoffel EM. Genetic predisposition to colorectal cancer: syndromes, genes, classification of genetic variants and implications for precision medicine. J Pathol. 2019 Apr;247(5):574–88.
7. Hahn MM, de Voer RM, Hoogerbrugge N, Ligtenberg MJL, Kuiper RP, van Kessel AG. The genetic heterogeneity of colorectal cancer predisposition - guidelines for gene discovery. Cell Oncol 2016 396. 2016 Jun;39(6):491–510.
8. Paske IBAW te, Ligtenberg MJL, Hoogerbrugge N, Voer RM de. Candidate Gene Discovery in Hereditary Colorectal Cancer and Polyposis Syndromes–Considerations for Future Studies. Int J Mol Sci 2020, Vol 21, Page 8757. 2020 Nov;21(22):8757.
9. Adam R, Spier I, Zhao B, Kloth M, Marquez J, Hinrichsen I, et al. Exome Sequencing Identifies Biallelic MSH3 Germline Mutations as a Recessive Subtype of Colorectal Adenomatous Polyposis. 2016;
10. Q. Rana H, Syngal S. Biallelic Mismatch Repair Deficiency: Management and Prevention of a Devastating Manifestation of the Lynch Syndrome. Gastroenterology. 2017 May;152(6):1254–7.
11. Fabišíková K, Hamidová O, Behulová RL, Závodná K, Priščáková P, Repiská V. Case Report: The Role of Molecular Analysis of the MUTYH Gene in Asymptomatic Individuals. Front

Genet. 2020 Dec;0:1567.

12. Fahed AC, Wang M, Homburger JR, Patel AP, Bick AG, Neben CL, et al. Polygenic background modifies penetrance of monogenic variants for tier 1 genomic conditions. Nat Commun. 2020 Dec;11(1):1–9.
13. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nat 2020 5817809. 2020 May;581(7809):434–43.
14. Tak YG, Farnham PJ. Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome. Epigenetics Chromatin 2015 81. 2015 Dec;8(1):1–18.
15. Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural variant calling: the long and the short of it. Genome Biol. 2019 Dec;20(1):246.
16. Carvalho CMB, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17(4):224–38.
17. Moreno-Cabrera JM, del Valle J, Castellanos E, Feliubadaló L, Pineda M, Brunet J, et al. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. Eur J Hum Genet. 2020 Jun;1–11.
18. Gabrielaite M, Torp MH, Rasmussen MS, Andreu-Sánchez S, Vieira FG, Pedersen CB, et al. A Comparison of Tools for Copy-Number Variation Detection in Germline Whole Exome and Whole Genome Sequencing Data. Vol. 13, Cancers . 2021.
19. Ried T, Meijer GA, Harrison DJ, Grech G, Franch-Expósito S, Briffa R, et al. The landscape of genomic copy number alterations in colorectal cancer and their consequences on gene expression levels and disease outcome. Mol Aspects Med. 2019 Oct;69:48–61.
20. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. Curr Protoc Bioinforma. 2013;(SUPL.43).
21. Välipakka S, Savarese M, Sagath L, Arumilli M, Giugliano T, Udd B, et al. Improving Copy Number Variant Detection from Sequencing Data with a Combination of Programs and a

Predictive Model. J Mol Diagnostics. 2020 Jan;22(1):40–9.

22. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17(1):122.

23. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010 Sep;38(16):e164–e164.

24. Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat. 2011 Aug;32(8):894–9.

25. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. Genome Med. 2020;12(1):103.

26. Singh AK, Olsen MF, Lavik LAS, Vold T, Drabløs F, Sjursen W. Detecting copy number variation in next generation sequencing data from diagnostic gene panels. BMC Med Genomics. 2021;14(1):214.

27. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan;46(D1):D1062–7.

28. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016 Oct;99(4):877–85.

29. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan;47(D1):D886–94.

30. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet. 2018 Oct;103(4):474–83.

31. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48(12):1581–6.

32. Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, et al. MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Res. 2014 Jan;42(D1):D297–303.

33. Kim S, Jhong J-H, Lee J, Koo J-Y. Meta-analytic support vector machine for integrating multiple omics data. BioData Min. 2017;10(1):2.

34. Feng B-J. PERCH: A Unified Framework for Disease Gene Prioritization. Hum Mutat. 2017 Mar;38(3):243–51.

35. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019 Jan;176(3):535-548.e24.

36. Anderson D, Lassmann T. An expanded phenotype centric benchmark of variant prioritisation tools. Hum Mutat. 2022 Feb;n/a(n/a).

37. Borges P, Pasqualim G, Matte U. Which Is the Best In Silico Program for the Missense Variations in IDUA Gene? A Comparison of 33 Programs Plus a Conservation Score and Evaluation of 586 Missense Variants [Internet]. Vol. 8, Frontiers in Molecular Biosciences . 2021.

38. Gunning AC, Fryer V, Fasham J, Crosby AH, Ellard S, Baple EL, et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. J Med Genet. 2021 Aug;58(8):547 LP – 555.

39. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Res. 2019 Jan;47(D1):D941–7.

40. Chakravarty D, Gao J, Phillips S, Kundra R, Zhang H, Wang J, et al. OncoKB: A Precision Oncology Knowledge Base. JCO Precis Oncol. 2017 May;(1):1–16.

41. Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res. 2016 Jan;44(D1):D1023–31.

42. Senter L, Clendenning M, Sotamaa K, Hampel H, Green J, Potter JD, et al. The Clinical Phenotype of Lynch Syndrome Due to Germ-Line PMS2 Mutations. Gastroenterology. 2008 Aug;135(2):419-428.e1.

43. Drost M, Koppejan H, de Wind N. Inactivation of DNA mismatch repair by variants of uncertain significance in the PMS2 gene. Hum Mutat. 2013 Nov;34(11):1477–80.

44. Hildreth A, Valasek MA, Thung I, Savides T, Sivagnanam M, Ramamoorthy S, et al. Biallelic Mismatch Repair Deficiency in an Adolescent Female. Yapijakis C, editor. Case Rep Genet. 2018;2018:8657823.

45. Ramos D, Brandão C, Sousa C, Dinis-Ribeiro M. Biallelic mismatch repair deficiency: A rare and trouble genetic syndrome. J Neurol Disord. 2022 May;10(4):491.

46. Sjursen W, Bjørnevoll I, Engebretsen LF, Fjelland K, Halvorsen T, Myrvold HE. A homozygote splice site PMS2 mutation as cause of Turcot syndrome gives rise to two different abnormal transcripts. Fam Cancer. 2009;8(3):179–86.

47. van der Tuin K, Mensenkamp AR, Tops CMJ, Corssmit EPM, Dinjens WN, van de Horst-Schrivers AN, et al. Clinical Aspects of SDHA-Related Pheochromocytoma and Paraganglioma: A Nationwide Study. J Clin Endocrinol Metab. 2018 Feb;103(2):438–45.

48. Scott P, Anderson K, Singhania M, Cormier R. Cystic Fibrosis, CFTR, and Colorectal Cancer. Vol. 21, International Journal of Molecular Sciences . 2020.

49. Anderson KJ, Cormier RT, Scott PM. Role of ion channels in gastrointestinal cancer. World J Gastroenterol. 2019 Oct;25(38):5732–72.

50. Kühn U, Wahle E. Structure and function of poly(A) binding proteins. Biochim Biophys Acta - Gene Struct Expr. 2004;1678(2):67–84.

51. Zhu J, Ding H, Wang X, Lu Q. PABPC1 exerts carcinogenesis in gastric carcinoma by targeting miR-34c. Int J Clin Exp Pathol. 2015;8(4):3794–802.

52. Takashima N, Ishiguro H, Kuwabara Y, Kimura M, Haruki N, Ando T, et al. Expression and prognostic roles of PABPC1 in esophageal cancer: Correlation with tumor progression and postoperative survival. Oncol Rep. 2006;15(3):667–71.

53. Jeon SA, Ha YJ, Kim J-H, Kim J-H, Kim S-K, Kim YS, et al. Genomic and transcriptomic analysis of Korean colorectal cancer patients. Genes Genomics. 2022;44(8):967–79.

54. Smart SK, Vasileiadi E, Wang X, DeRyckere D, Graham DK. The Emerging Role of TYRO3 as a Therapeutic Target in Cancer. Vol. 10, Cancers . 2018.

55. Al Kafri N, Hafizi S. Identification of signalling pathways activated by Tyro3 that promote cell survival, proliferation and invasiveness in human cancer cells. Biochem Biophys Reports. 2021;28:101111.

56. Du J, Xu J, Chen J, Liu W, Wang P, Ye K. circRAE1 promotes colorectal cancer cell migration and invasion by modulating miR-338-3p/TYRO3 axis. Cancer Cell Int. 2020;20(1):430.

57. Huang Y, Chen Z, Zhou X, Huang H. Circ_0000467 Exerts an Oncogenic Role in Colorectal Cancer via miR-330-5p-Dependent Regulation of TYRO3. Biochem Genet. 2022;

58. Collard TJ, Urban BC, Patsos HA, Hague A, Townsend PA, Paraskeva C, et al. The retinoblastoma protein (Rb) as an anti-apoptotic factor: expression of Rb is required for the anti-apoptotic function of BAG-1 protein in colorectal tumour cells. Cell Death Dis. 2012;3(10):e408–e408.

59. Gianferante DM, Rotunno M, Dean M, Zhou W, Hicks BD, Wyatt K, et al. Whole-exome sequencing of nevoid basal cell carcinoma syndrome families and review of Human Gene Mutation Database PTCH1 mutation data. Mol Genet Genomic Med. 2018 Nov;6(6):1168–80.

60. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan;29(1):308–11.

61. Arteche-López A, Ávila-Fernández A, Romero R, Riveiro-Álvarez R, López-Martínez MA, Giménez-Pardo A, et al. Sanger sequencing is no longer always necessary based on a single-center validation of 1109 NGS variants in 825 clinical exomes. Sci Rep. 2021;11(1):5697.

62. Fletcher O, Houlston RS. Architecture of inherited susceptibility to common cancer. Nat Rev Cancer. 2010;10(5):353–61.

63. Schmit SL, Edlund CK, Schumacher FR, Gong J, Harrison TA, Huyghe JR, et al. Novel Common Genetic Susceptibility Loci for Colorectal Cancer. JNCI J Natl Cancer Inst. 2019 Feb;111(2):146–57.

64. Sassano M, Mariani M, Quaranta G, Pastorino R, Boccia S. Polygenic risk prediction models for colorectal cancer: a systematic review. BMC Cancer. 2022;22(1):65.

65. Jung H, Lee KS, Choi JK. Comprehensive characterisation of intronic mis-splicing mutations

in human cancers. Oncogene. 2021;40(7):1347–61.

66. Zhang X, Wakeling M, Ware J, Whiffin N. Annotating high-impact 5′untranslated region variants with the UTRannotator. Bioinformatics. 2021 Apr;37(8):1171–3.

67. Schubert SA, Morreau H, de Miranda NFCC, van Wezel T. The missing heritability of familial colorectal cancer. Mutagenesis. 2020 Jul;35(3):221–31.

68. Cullinane CM, Creavin B, O'Connell EP, Kelly L, O'Sullivan MJ, Corrigan MA, et al. Risk of colorectal cancer associated with BRCA1 and/or BRCA2 mutation carriers: systematic review and meta-analysis. Br J Surg. 2020 Jul;107(8):951–9.

Supplementary material:

Supplementary file 1 (attached below)

Supplementary_file_1.pdf: Benchmarking study comparing the performance of 45 different pathogenicity prediction tools.

Supplementary file 2 (attached below)

Supplementary_file_2.pdf: Detailed information about the various filtering steps.

Supplementary file 3 (available upon request)

Supplementary_file_3.xlsm: The full list of 346 variants passing filtering steps.

Supplementary file 4 (available upon request)

Supplementary_file_4.bed: List of 88 known cancer genes used as targets for CNV calling.

Supplementary file 5 (attached below)

Supplementary_file_5.pdf: enlists supporting tables S1, S2 & S3, with details below.

- Table S1: List of 38 genes and their roles in Cancer (as per COSMIC, OncoKB, and TSGene databases).
- Table S2: The list of detected CNVs.

- Table S3: Number of variants passing different combinations of filters based on rank-scores of 7 *in silico* tools (CADD, ClinPred, M-CAP, BayesDel-addAF, MetaSVM, REVEL, VEST4).

This document is included as Supplementary Material for the paper "Detection of germline variants with pathogenic potential in 48 patients with familial colorectal cancer by using whole exome sequencing" by Ashish Kumar Singh, Bente Talseth-Palmer, Alexandre Xavier, Rodney J Scott, Finn Drabløs and Wenche Sjursen

Evaluation of methods for computational prediction of pathogenicity of missense variants

Ashish Kumar Singh[1,2,*], *Finn Drabløs*[1,*,+]

1. Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, NTNU - Norwegian University of Science and Technology, Trondheim, Norway
2. Department of Medical Genetics, St. Olavs Hospital, Trondheim, Norway
* The authors contributed equally.
+ Corresponding author <finn.drablos@ntnu.no>

Abstract

Background: Several methods have been developed for predicting pathogenicity of missense variants in protein-coding regions of genes. Here we try to evaluate the performance of several such methods on a relevant exome-wide dataset of variants associated with colorectal cancer, to identify well-performing methods.

Results: Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) curves for the prediction of pathogenicity of the missense variants found in the exome data compared to the ClinVar classification of the same variants (pathogenic or benign) was estimated. This was used to rank the performance of 45 different tools or pathogenicity scores. This was compared to the performance of these tools according to several previous benchmarking studies, and the overall performance was used to select the best-performing tools.

Conclusions: The seven tools with best overall performance were selected. These tools were ClinPred, VEST4, BayesDel-addAF, REVEL, CADD, M-CAP, and MetaSVM.

Keywords: Pathogenicity, Missense variants, Prediction, AUC, ROC.

Background

This project is part of a larger project on identifying possibly causative variants in whole-exome sequencing (WES) data from cancer patients. In general, the goal of this part of the project was to find a suitable set of computational tools for identifying pathogenic variants in exomes. This should be more than one tool, as no single tool has perfect sensitivity (*i.e.*, being able to find all relevant pathogenic variants) and selectivity (*i.e.*, being able to distinguish perfectly between pathogenic and benign variants). Not only because this is a challenging computational problem, but also because the distinction between pathogenic and benign can be unclear and may depend upon cancer type. It has been shown that using a consensus prediction based on a combination of tools (for example in the form of meta-predictors) may improve the overall performance [1], although also the opposite effect

has been observed [2]. But the performance of each individual tool should in any case be as good as possible. The goal has therefore been to identify suitable prediction methods for pathogenicity of variants for a specific dataset on colorectal cancer. However, although this was done with a specific dataset in mind, it is likely that the analysis may have a more general relevance.

A standard approach based on receiver operating characteristics was used for assessing the performance of relevant methods. Score values for several different tools were computed on a set of WES samples. ClinVar [3] classification of the variants was added, and the sample set was split into pathogenic and benign variants, based on the ClinVar classification, leaving out variants without a clear classification. The performance of a given tool at a given cutoff for the computed score value can then be estimated as the number of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) classifications, which can be used to compute the true positive rate (sensitivity) and the false positive rate (1-specificity) at that cutoff. This can be computed across all score values as a Receiver Operating Characteristic (ROC) curve, and the overall performance of the method can be estimated by the Area Under Curve (AUC) for the ROC. An AUC close to 1.0 will indicate a near perfect classification performance of that tool, whereas an AUC close to 0.5 indicates that the performance of the classifier is similar to random classification.

Methods

Please see the main paper for details on the sequencing data that were used. Briefly, germline DNA was extracted from blood samples of 48 patients diagnosed with colorectal cancer fulfilling the Amsterdam-II Criteria for Lynch syndrome, consisting of 32 unrelated individuals, and 16 related individuals from 8 different families. Sanger sequencing detected no germline MMR mutations in these samples. Whole exome sequencing was performed on the 48 samples. SNP/indel variant calling was performed using a standardized BWA-Picard-GATK pipeline [4]. The variants were annotated with VEP, the Ensembl Variant Effect Predictor [5], using data from dbNSFP [6, 7]. Filtering of variants for further analysis was done with local scripts. ROC and AUC values were computed with the ROCR tool [8] in R (https://www.r-project.org/) version 3.6.3 (2020), and also correlations were computed in R.

Results

The initial input file consisted of 350.126 variants, including variants assigned to alternative transcripts of the same gene. This list was annotated with VEP and then filtered according to two main criteria.

Following the VEP annotation all prediction tools giving score values as a rank score were selected, in total 45 different methods or score sets, although several of these represent variants of the same basic prediction method (like BayesDel_addAF and BayesDel_noAF, representing the same method with and without allele frequencies). Then the number of annotated variants for each set of scores were counted, and two methods were excluded for further analysis (LINSIGHT and MutPred) because of low coverage. The coverage of the remaining methods varied from 34.165 (LRT_converted) to 46.240 (CADD_raw, DANN, GenoCanyon), although most methods had coverage >40.000. Then, only fully annotated variants (*i.e.*, variants annotated by all selected methods) were used.

The variants were also annotated according to ClinVar classification (CLIN_SIG in VEP output), and only variants that could be identified as either pathogenic or benign were used. All variants classified

as *pathogenic* or *likely_pathogenic* were counted as pathogenic. All variants classified as *benign* or *likely_benign* were counted as benign. All variants classified for example only as uncertain (*uncertain_significance*) or with conflicting classifications (*benign, pathogenic*) were excluded.

This gave a final dataset consisting of 961 variants, with 161 variants classified as pathogenic and 800 variants classified as benign. This dataset was then analyzed with ROCR in R.

The performance (given as AUC) for the 15 tools with best performance are given in Table 1. The full table is given as Supplementary Table S1. The tools indicated in red were later selected to be used in the main project, please see Discussion for details.

Table 1 – AUC for the 15 best-scoring methods

Method	AUC
ClinPred_rankscore	0.9312
VEST4_rankscore	0.8858
BayesDel_addAF_rankscore	0.8844
REVEL_rankscore	0.8799
Eigen-raw_coding_rankscore	0.8666
BayesDel_noAF_rankscore	0.8653
Eigen-PC-raw_coding_rankscore	0.8600
CADD_raw_rankscore	0.8547
M-CAP_rankscore	0.8540
CADD_raw_rankscore_hg19	0.8415
Polyphen2_HDIV_rankscore	0.8325
MutationAssessor_rankscore	0.8319
MetaSVM_rankscore	0.8243
Polyphen2_HVAR_rankscore	0.8227

Discussion

The results in Table 1 show that in particular ClinPred [9] has a very good performance on our dataset. This is hardly surprising, since ClinPred was developed by using both data and a strategy that is similar to what we have used in this benchmarking. However, to have a good basis for selecting a set of tools, data from several benchmarking studies were used.

Three quite recent benchmarking studies have included the ClinPred method. The study by Anderson & Lassman [10] was an extension of a previous benchmarking [11] and compared 37 different tools. Briefly, they made a data set of pathogenic variants from ClinVar, and terms on human phenotypic abnormalities from the Human Phenotype Ontology (HPO) resource [12] were associated with the relevant genes. They then used area under the precision-recall curve to estimate tool performance for discriminating between pathogenic and benign variants. Based on this they recommended mainly three methods: BayesDel_addAF (*i.e.*, with allele frequencies), CADD, and ClinPred.

The study by Borges *et al.* [13] compared 33 different methods. They used a large set of disease-causing variants of a specific gene (alpha-L-iduronidase, IDUA, involved in mucopolysaccharidosis type

I, MPS I) and used several statistical tests to evaluate how well each method could distinguish between deleterious and neutral variants. Based on this analysis the authors recommended in particular BayesDel (addAF and noAF), PON-P2 (genome and protein), and ClinPred.

Finally, the study by Gunning *et al.* [2] used two datasets, one 'open' set with data from ClinVar and gnomAD [14], and one 'clinically representative' dataset with variants identified through exome sequencing from diagnostics and research. They used this to evaluate a small number of methods using AUC for ROC curves, comparing three different meta-predictors (REVEL, GAVIN and ClinPred) against two commonly used *in silico* tools (SIFT and PolyPhen-2). They confirmed a good performance of in particular REVEL and ClinPred.

These studies confirm that ClinPred has a very good performance. However, there are also several other recent studies that can be relevant, with comparison or benchmarking of several methods. Suybeng *et al.* [15] used AUC scoring of more than 20 tools on a gold standard set of somatic single-nucleotide variants classified as oncogenic or neutral, and found the best performing tools to be CADD, Eigen, PolyPhen-2, PROVEAN, UMD-Predictor and REVEL.

Tian *et al.* [16] compared 7 predictors, using a high-quality consensus set of missense variants in clinically relevant genes which had been classified and reviewed by experts. Here REVEL and BayesDel showed the best performance.

Jaravine *et al.* [17] used machine-learning approaches to build ensemble or meta predictors consisting of several basic prediction methods, and by building these models in a stepwise manner (stacked ensembles) they could estimate how much each prediction method contributed to the overall performance of each ensemble. The best total performance was achieved with a distributed random forest (DRF), which showed that out of 39 different annotation scores (29 predictions, 9 conservation score and 1 indispensability score) tested on ClinVar-annotated variants from gnomAD [14], the best performance could be achieved with VEST4, M-CAP, CADD, MutPred, MVP and MetaLR, in that order.

Li *et al.* [18] used three different datasets for benchmarking, including ClinVar data, to compare 23 methods by using 12 different performance measures, including AUC. The methods included both function prediction methods, conservation methods and ensemble methods. It is difficult to extract overall recommendations from the results, but the AUC results for the ClinVar set showed particularly good performance for VEST3 and REVEL (AUC >0.9), but also very good performance for ensemble methods like CADD, Eigen, and MetaLR (AUC >0.87).

Chen *et al.* [19] used a slightly different approach where they focused on prediction of cancer driver mutations. They used five different datasets to compare 33 different methods. Based on AUC scores the methods like MetaSVM, MetaLR, M-CAP, and REVEL showed good performance on a dataset based on mutation clustering patterns, whereas PolyPhen2, PROVEAN, MetaLR, MutPred, REVEL, and VEST4 showed good performance on a dataset based on TP53 mutations.

Interestingly, our data and the benchmarking studies mentioned above show a quite consistent pattern with respect to identifying prediction methods with good performance. Based both on our own results and on benchmarking studies mentioned here several score methods could be relevant

for inclusion. However, to limit the number of methods it was also decided to leave out some methods that were very similar or highly correlated with already included methods, as it was assumed that such methods would provide limited additional information. Relevant examples are Eigen (correlation to CADD >0.92), MetaLR (similar to MetaSVM with correlation >0.94), and Polyphen2 (already included through most meta-methods). Also, MutationAssessor was left out, even though it had a quite good AUC score, because it was one of the methods with relatively low coverage with respect to the number of variants on which it had data (approximately 75% coverage compared to methods that were selected).

Based on the overall evaluation, 7 different methods were then selected to be used in a consensus-like approach in the main project; ClinPred [9], VEST4 [20], BayesDel-addAF [21], REVEL [22], CADD [23], M-CAP [24], and MetaSVM [25]. These methods are highlighted in Table 1, and the ROC curve for each of these methods is shown in Figure 1.

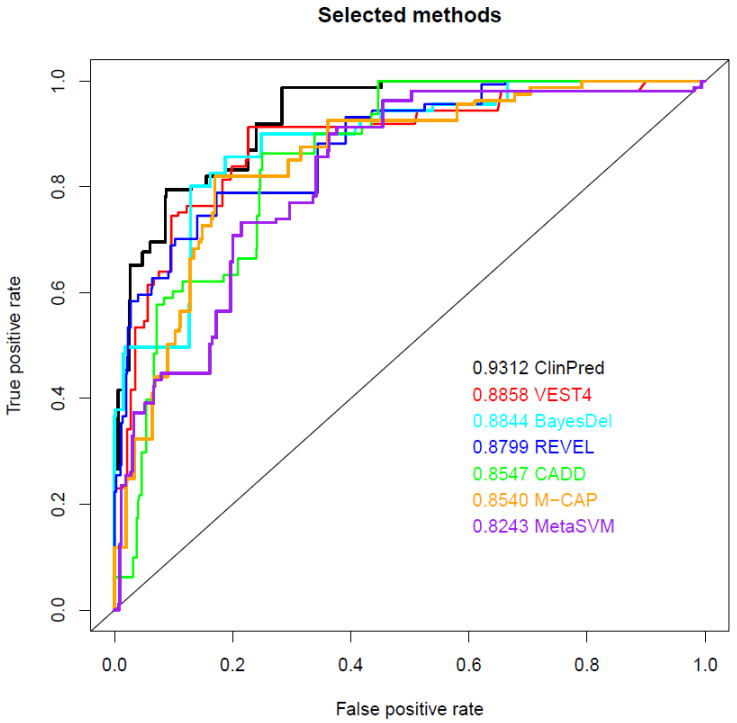


Figure 1 – ROC curve for selected methods

This analysis is included as supplementary material to the main paper, rather than a full, separate publication. It was mainly meant as a quick analysis to support the literature survey that already had been done, and to check whether this specific dataset behaved similarly to the more general benchmarking datasets used in published studies. There are several aspects of this study that certainly can be (and maybe should be) improved for this to be published as a proper benchmarking. It has not been attempted to do a selection of relevant transcripts, for example by using only

canonical transcripts. It can be argued that alternative transcripts in some cases may show different effects of specific variants, but the inclusion of several alternative transcripts can also lead to a bias towards genes with many know transcripts. It has not been attempted to make a more balanced dataset, with similar numbers of pathogenic and benign cases. This can also affect the analysis, although the effect here probably is small [9], at least partly because the AUC measure is supposed to be relatively robust for unbalanced datasets (see for example [10]). Finally, the interpretation of ClinVar classifications may be simplistic in cases with several alternative classifications. For example, a ClinVar classification as (*uncertain_significance, likely_pathogenic*) was selected to be counted as pathogenic, but this may not be a sufficiently reliable classification. However, the results from this analysis are similar to most of the benchmarking studies that we have compared it against, which seems to indicate that although the study was carried out in a relatively simplistic way (one might even say "quick-and-dirty"), the results seem to be quite robust and comparable to previous studies. Therefore, they provide a good basis for the selection of tools for the main project.

Conclusion

The study identified 7 different well-performing prediction methods for pathogenicity prediction of genetic variants. These 7 methods (ClinPred, VEST4, BayesDel-addAF, REVEL, CADD, M-CAP, and MetaSVM) were used in the main project, based on a consensus approach.

Author contributions

AKS processed sequencing data, identified variants, and annotated data with VEP. FD did the statistical analysis and wrote the first version of the manuscript. Both authors have approved the final version.

References

1. Gonzalez-Perez A, Lopez-Bigas N: **Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel**. *Am J Hum Genet* 2011, **88**(4):440-449.
2. Gunning AC, Fryer V, Fasham J, Crosby AH, Ellard S, Baple EL, Wright CF: **Assessing performance of pathogenicity predictors using clinically relevant variant datasets**. *J Med Genet* 2021, **58**(8):547-555.
3. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W *et al*: **ClinVar: improving access to variant interpretations and supporting evidence**. *Nucleic Acids Res* 2018, **46**(D1):D1062-D1067.
4. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J *et al*: **From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline**. *Curr Protoc Bioinformatics* 2013, **43**:11 10 11-11 10 33.
5. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F: **The Ensembl Variant Effect Predictor**. *Genome Biol* 2016, **17**(1):122.
6. Liu X, Jian X, Boerwinkle E: **dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions**. *Hum Mutat* 2011, **32**(8):894-899.
7. Liu X, Li C, Mou C, Dong Y, Tu Y: **dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs**. *Genome Med* 2020, **12**(1):103.
8. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R**. *Bioinformatics* 2005, **21**(20):3940-3941.
9. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD: **ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants**. *Am J Hum Genet* 2018, **103**(4):474-483.

10. Anderson D, Lassmann T: **An expanded phenotype centric benchmark of variant prioritisation tools**. *Hum Mutat* 2022.
11. Anderson D, Lassmann T: **A phenotype centric benchmark of variant prioritisation tools**. *NPJ Genom Med* 2018, **3**:5.
12. Kohler S, Doelken SC, Mungall CJ, Bauer S, Firth HV, Bailleul-Forestier I, Black GC, Brown DL, Brudno M, Campbell J *et al*: **The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data**. *Nucleic Acids Res* 2014, **42**(Database issue):D966-974.
13. Borges P, Pasqualim G, Matte U: **Which Is the Best In Silico Program for the Missense Variations in IDUA Gene? A Comparison of 33 Programs Plus a Conservation Score and Evaluation of 586 Missense Variants**. *Front Mol Biosci* 2021, **8**:752797.
14. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP *et al*: **The mutational constraint spectrum quantified from variation in 141,456 humans**. *Nature* 2020, **581**(7809):434-443.
15. Suybeng V, Koeppel F, Harle A, Rouleau E: **Comparison of Pathogenicity Prediction Tools on Somatic Variants**. *J Mol Diagn* 2020, **22**(12):1383-1392.
16. Tian Y, Pesaran T, Chamberlin A, Fenwick RB, Li S, Gau CL, Chao EC, Lu HM, Black MH, Qian D: **REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification**. *Sci Rep* 2019, **9**(1):12752.
17. Jaravine V, Balmford J, Metzger P, Boerries M, Binder H, Boeker M: **Annotation of Human Exome Gene Variants with Consensus Pathogenicity**. *Genes (Basel)* 2020, **11**(9).
18. Li J, Zhao T, Zhang Y, Zhang K, Shi L, Chen Y, Wang X, Sun Z: **Performance evaluation of pathogenicity-computation methods for missense variants**. *Nucleic Acids Res* 2018, **46**(15):7793-7804.
19. Chen H, Li J, Wang Y, Ng PK, Tsang YH, Shaw KR, Mills GB, Liang H: **Comprehensive assessment of computational algorithms in predicting cancer driver mutations**. *Genome Biol* 2020, **21**(1):43.
20. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R: **Identifying Mendelian disease genes with the variant effect scoring tool**. *BMC Genomics* 2013, **14 Suppl 3**:S3.
21. Feng BJ: **PERCH: A Unified Framework for Disease Gene Prioritization**. *Hum Mutat* 2017, **38**(3):243-251.
22. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D *et al*: **REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants**. *Am J Hum Genet* 2016, **99**(4):877-885.
23. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J: **A general framework for estimating the relative pathogenicity of human genetic variants**. *Nat Genet* 2014, **46**(3):310-315.
24. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, Bernstein JA, Bejerano G: **M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity**. *Nat Genet* 2016, **48**(12):1581-1586.
25. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X: **Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies**. *Hum Mol Genet* 2015, **24**(8):2125-2137.

Supplementary

Supplementary Table S1 – AUC for all score methods

Method	AUC
ClinPred_rankscore	0.9312
VEST4_rankscore	0.8858
BayesDel_addAF_rankscore	0.8844
REVEL_rankscore	0.8799
Eigen-raw_coding_rankscore	0.8666
BayesDel_noAF_rankscore	0.8653
Eigen-PC-raw_coding_rankscore	0.8600
CADD_raw_rankscore	0.8547
M-CAP_rankscore	0.8540
CADD_raw_rankscore_hg19	0.8415
Polyphen2_HDIV_rankscore	0.8325
MutationAssessor_rankscore	0.8319
MetaSVM_rankscore	0.8243
Polyphen2_HVAR_rankscore	0.8227
SIFT4G_converted_rankscore	0.8131
MetaLR_rankscore	0.7890
fathmm-XF_coding_rankscore	0.7863
SIFT_converted_rankscore	0.7851
fathmm-MKL_coding_rankscore	0.7807
PROVEAN_converted_rankscore	0.7723
phyloP100way_vertebrate_rankscore	0.7644
phastCons100way_vertebrate_rankscore	0.7525
MVP_rankscore	0.7522
DANN_rankscore	0.7443
DEOGEN2_rankscore	0.7411
LIST-S2_rankscore	0.7312
GERP++_RS_rankscore	0.7213
LRT_converted_rankscore	0.7174
MutationTaster_converted_rankscore	0.7125
SiPhy_29way_logOdds_rankscore	0.6858
phastCons30way_mammalian_rankscore	0.6837
PrimateAI_rankscore	0.6727
FATHMM_converted_rankscore	0.6575
MPC_rankscore	0.6255
phyloP30way_mammalian_rankscore	0.6229
GenoCanyon_rankscore	0.6214
phastCons17way_primate_rankscore	0.6085
phyloP17way_primate_rankscore	0.5850
H1-hESC_fitCons_rankscore	0.4922
bStatistic_converted_rankscore	0.4907
integrated_fitCons_rankscore	0.4639
HUVEC_fitCons_rankscore	0.4545
GM12878_fitCons_rankscore	0.4239

Variant prioritization

Prioritization steps were performed on initial set of 125686 variants detected via variant calling on 48 samples. Variant occurrences among global population, clinical significance based on already assigned predictions and pathogenicity estimation scores of selective tools (among full list of annotation tools) were taken in account. The selected tools were as follows:

- Database for calculation of population frequency: gnomAD version r2.1 (1)
- Database for estimation of clinical significance: ClinVar (2)
- *In silico* tools for pathogenicity prediction:
    - Missence, Nonsence and Start-loss prediction: REVEL (3), CADD (4), ClinPred (5), M-CAP (6), VEST4 (7), MetaSVM (8), BayesDel (9)
    - Splicing alteration prediction: SpliceAI (10)
    - Loss of function prediction: Loftee (1)

Selection of *in silico* prediction tools was based on ranking generated by our benchmarking study comparing performance of presently available 45 pathogenicity prediction tools (See supplementary document Sup1). Additionally we took inspiration from other benchmark studies with similar goals (11–13).

Variant filtration was performed using command-line based tool filter_vep form VEP toolkit. Filtering criteria have been explained in following steps:

1. First stage filtering was based on variant frequencies, where frequency database gnomAD was used to filter out variants with frequency higher then 0.001.
2. Second stage filtering was based on estimated clinical significance of variants provided by ClinVar database. Variants passing stage1 filtering were firstly categorized in below mentioned 6 different categories based on their clinical significance. Variants in Class1&2 and Class2+ where considered insignificant and were discarded after this stage. Variants in Class4&5 were manually curated after this stage. Remaining variants in categories Class3, Class3+ and ClassUnknown were further filtered in next stages based on different criteria.
    2.1. Class1&2: [benign OR likely benign]
    2.2. Class2+: [benign OR likely benign] AND [uncertain significance OR Conflicting interpretation]
    2.3. Class3: [uncertain significance OR Conflicting interpretation]
    2.4. Class3+: [pathogenic] AND [uncertain significance]

2.5. Class4&5: [pathogenic OR likely pathogenic]

2.6. Class Unknown: No Information in clinvar

3. Third stage filtering was based on predictions from 9 selected *in silico* tools for pathogenicity prediction. These tools were used in following three different ways to make filters.

3.1. In step3.1, filters were made based on predictions from seven of these nine tools i.e., REVEL, CADD, BayesDel, ClinPred, M-CAP, MetaSVM and VEST4. These tools were used for predicting effects of missense, nonsense and start-loss related variants. Combinatorial approach was used to test the filter effects ahead of applying the filters, where all possible combinations of these seven tools with different rank-score cut-offs were tested on this filtration step. Output of this analysis is provided as supplementary document (See supplementary table S4). Filters selected among those are described below.

3.1.1. Filter for Class3+: Variants with rank-score higher than or equal to 0.8 in at least one of these seven tools.

3.1.2. Filter for Class3: (Variants with rank-score higher than or equal to 0.99 in at least one of these seven tools) AND (Variants with rank-score higher than or equal to 0.8 in at least six on these seven tools).

3.1.3. Filter for ClassUnknown: (Variants with rank-score higher than or equal to 0.99 in at least two of these seven tools) AND (Variants with rank-score higher than or equal to 0.8 in all of these seven tools).

3.2. In step3.2, filtering was done to detect variants without any score or prediction in any of these following eight of these nine tools i.e., REVEL, CADD, BayesDel, ClinPred, M-CAP, MetaSVM, VEST4 and Loftee.

3.3. In step3.3, filtering was done to detect splicing-alteration and frameshift related variants.

3.3.1. Filter for splicing-alteration: Variants with delta-score higher than or equal to 0.9 for any of four changes (i.e., acceptor-gain, acceptor-loss, donor-gain and donor-loss) and high confidence for loss-of-function from Loftee tool based annotation.

3.3.2. Filter for frameshift variant: variants with consequence for frameshift and high confidence for loss-of-function from Loftee tool based annotation.

References:

1. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nat 2020 5817809. 2020 May;581(7809):434–43.
2. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan;46(D1):D1062–7.
3. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016 Oct;99(4):877–85.
4. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 2019 Jan;47(D1):D886–94.
5. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum Genet. 2018 Oct;103(4):474–83.
6. Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet. 2016;48(12):1581–6.
7. Madej T, Lanczycki CJ, Zhang D, Thiessen PA, Geer RC, Marchler-Bauer A, et al. MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Res. 2014 Jan;42(D1):D297–303.
8. Kim S, Jhong J-H, Lee J, Koo J-Y. Meta-analytic support vector machine for integrating multiple omics data. BioData Min. 2017;10(1):2.
9. Feng B-J. PERCH: A Unified Framework for Disease Gene Prioritization. Hum Mutat. 2017 Mar;38(3):243–51.
10. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019 Jan;176(3):535-548.e24.
11. Anderson D, Lassmann T. An expanded phenotype centric benchmark of variant prioritisation tools. Hum Mutat. 2022 Feb;n/a(n/a).

12. Borges P, Pasqualim G, Matte U. Which Is the Best In Silico Program for the Missense Variations in IDUA Gene? A Comparison of 33 Programs Plus a Conservation Score and Evaluation of 586 Missense Variants [Internet]. Vol. 8, Frontiers in Molecular Biosciences . 2021.
13. Gunning AC, Fryer V, Fasham J, Crosby AH, Ellard S, Baple EL, et al. Assessing performance of pathogenicity predictors using clinically relevant variant datasets. J Med Genet. 2021 Aug;58(8):547 LP – 555.

Table S1: List of 38 genes and their roles in Cancer (as per COSMIC, OncoKB, and TSGene databases)

Gene	Role in Cancer	Database
BRCA2	TSG	cosmic; OncoKB; TSG-db
BRD3	oncogene; fusion; In Sanger Cancer Gene Census	cosmic; OncoKB
CACNA1D	oncogene; In Sanger Cancer Gene Census	cosmic; OncoKB
CD36	In cancer panels (FOUNDATION ONE HEME)	OncoKB
CFTR	TSG	TSG-db
COL1A1	fusion; In cancer panels (FOUNDATION ONE HEME)	cosmic; OncoKB
DDX4	Oncogene	OncoKB
DNMT3A	TSG	cosmic; OncoKB; TSG-db
EPHA2	TSG	TSG-db
ERCC2	TSG	cosmic; OncoKB
FANCD2	TSG	cosmic; OncoKB
FOXP1	oncogene; fusion; TSG	cosmic; OncoKB; TSG-db
GUCY2C	TSG	TSG-db
IL17RD	TSG	TSG-db
LRP5	oncogene; TSG	OncoKB
MERTK	In cancer panels (FOUNDATION ONE)	OncoKB
MLH1	TSG	cosmic; OncoKB; TSG-db
MME	TSG	TSG-db
MSH2	TSG	cosmic; OncoKB; TSG-db
MSH6	TSG	cosmic; OncoKB
NOD1	In cancer panels (FOUNDATION ONE HEME)	OncoKB
PABPC1	oncogene; TSG	cosmic
PANX2	TSG	TSG-db
PKD1	TSG	TSG-db
PMS2	TSG	cosmic; OncoKB
POLG	TSG; Cancer gene (OncoKB assigned)	cosmic; OncoKB
PTCH1	TSG	cosmic; OncoKB; TSG-db
RBBP8	TSG	TSG-db
RND3	TSG	TSG-db
SDHA	TSG	cosmic; OncoKB; TSG-db
SLC34A2	TSG; fusion; In cancer panels (FOUNDATION ONE)	cosmic; OncoKB
SOX10	In cancer panels (FOUNDATION ONE HEME)	OncoKB
SYNM	TSG	TSG-db
TYK2	Oncogene	OncoKB
TYRO3	In cancer panels (FOUNDATION ONE)	OncoKB
UPF1	In cancer panels (MSK-IMPACT;MSK-HEME)	OncoKB
WNT7A	TSG	TSG-db
ZBTB7C	TSG	TSG-db

Table S2: The list of detected CNVs.

Sample ID	Gene: exon	Genomic position	CNV Type
S.01	RB1: exon6	chr13:48923092-48923159	Deletion
S.03	BARD1: exon5	chr2:215633956-215634036	Duplication
S.14	BLM: whole gene	chr15:91290623-91358509	Duplication
S.19	POLE: exon12	chr12:133251984-133252103	Duplication
S.34	PTCH2: exon22	chr1:45288087-45288341	Duplication

Table S3: Number of variants passing different combinations filters based on rank-scores of 7 in silico tools (CADD, ClinPred, M-CAP, BayesDel-addAF, MetaSVM, REVEL, VEST4)

Class4-5	Number of variants (and associated genes) passing filter: [rank-score>=Cutoff in at least **N** of **7** meta-tools]						
Total: 68 (91)	**N=1**	**N=2**	**N=3**	**N=4**	**N=5**	**N=6**	**N=7**
Cutoff>=0.8	49 (46)	45 (43)	32 (31)	24 (23)	14 (14)	8 (8)	4 (4)
Cutoff>=0.85	46 (45)	39 (38)	28 (28)	14 (14)	9 (9)	4 (4)	3 (3)
Cutoff>=0.9	41 (40)	33 (32)	18 (17)	10 (10)	5 (5)	3 (3)	0
Cutoff>=0.95	34 (33)	20 (20)	12 (12)	7 (7)	2 (2)	1 (1)	0
Cutoff>=0.99	12 (11)	3 (3)	2 (2)	0	0	0	0
Class3+	Number of variants (and associated genes) passing filter: [rank-score>=Cutoff in at least **N** of **7** meta-tools]						
Total: 41 (37)	**N=1**	**N=2**	**N=3**	**N=4**	**N=5**	**N=6**	**N=7**
Cutoff>=0.8	**22 (21)**	20 (19)	15 (14)	14 (13)	9 (8)	4 (4)	1 (1)
Cutoff>=0.85	21 (20)	18 (17)	14 (13)	9 (8)	4 (4)	2 (2)	0
Cutoff>=0.9	18 (17)	13 (12)	9 (8)	6 (6)	1 (1)	0	0
Cutoff>=0.95	14 (13)	10 (9)	4 (3)	1 (1)	0	0	0
Cutoff>=0.99	10 (9)	4 (4)	0	0	0	0	0
Class3	Number of variants (and associated genes) passing filter: [rank-score>=Cutoff in at least **N** of **7** meta-tools]						
Total: 752 (849)	**N=1**	**N=2**	**N=3**	**N=4**	**N=5**	**N=6**	**N=7**
Cutoff>=0.8	258 (224)	162 (143)	104 (98)	64 (61)	31 (31)	**13 (13)**	3 (3)
Cutoff>=0.85	210 (186)	111 (101)	63 (61)	34 (34)	13 (13)	4 (4)	1 (1)
Cutoff>=0.9	143 (133)	68 (67)	35 (36)	8 (8)	2 (2)	0	0
Cutoff>=0.95	85 (82)	29 (30)	11 (12)	1 (1)	0	0	0
Cutoff>=0.99	**14 (14)**	1 (1)	0	0	0	0	0
ClassUNKNOWN	Number of variants (and associated genes) passing filter: [rank-score>=Cutoff in at least **N** of **7** meta-tools]						
Total: 20541 (14043)	**N=1**	**N=2**	**N=3**	**N=4**	**N=5**	**N=6**	**N=7**
Cutoff>=0.8	2547 (2120)	1503 (1289)	886 (773)	520 (480)	305 (291)	151 (151)	**72 (70)**
Cutoff>=0.85	2037 (1707)	1057 (917)	561 (510)	307 (294)	158 (158)	85 (87)	26 (26)
Cutoff>=0.9	1447 (1235)	636 (570)	293 (276)	142 (141)	67 (69)	34 (35)	6 (6)
Cutoff>=0.95	769 (688)	240 (225)	89 (89)	27 (27)	15 (15)	1 (1)	0
Cutoff>=0.99	165 (155)	**20 (20)**	1 (1)	0	0	0	0

ISBN 978-82-326-7144-1 (printed ver.)
ISBN 978-82-326-7143-4 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology