Identification of new causative genes in inherited colorectal cancer

Alexandre Xavier MA (Rennes)

Thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Medical Genetics

November 2019

Statement of originality

I hereby certify that the work embodied in the thesis is my own work, conducted under normal supervision. The thesis contains no material which has been accepted, or is being examined, for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository, subject to the provisions of the Copyright Act 1968 and any approved embargo.

16/10/2019

…….. ……..

Date Alexandre Xavier

Statement of Collaboration:
I hereby certify that the work embodied in this thesis has been done in collaboration with other researchers. I have included as part of the thesis a statement clearly outlining the extent of collaboration, with whom and under what auspices.

Statement of Authorship:

I hereby certify that the work embodied in this thesis contains a published papers/scholarly work of which I am a joint author. I have included as part of the thesis a written statement, endorsed by my supervisor, attesting to my contribution to the joint publications/scholarly work.

Thesis by Publication:

I hereby certify that this thesis is in the form of a series of published papers of which I am a co-author. I have included as part of the thesis a written statement from each co- author, endorsed by the Faculty Assistant Dean (Research Training), attesting to contribution to the joint publications.

16/10/2019

…….. ……..

Date Alexandre Xavier

Acknowledgements:

I would like, first and foremost, to thank Laureate Professor Rodney J. Scott and Dr Bente Talseth-Palmer for giving me the opportunity to fulfil my PhD project with them. I will always be thankful for Rodney's approachability as his doors remained open even at the busiest of time. Rodney also taught me how to distinguish important things from others. I am also infinitely grateful for Bente's unyielding support during my project as well as her encouragement when I wanted to explore new fields.

I would also give special thanks to Dr Kelly Kiejda. Kelly, I consider you as my 3rd unofficial supervisor and I will always be grateful that you took me under your wing when I was alone and with no one else to turn to.

I want to express my gratitude to Dr Katherine Bolton. You were the one who welcomed me to this lab and you took it upon yourself to teach me everything. Nothing I have done would have been possible without your help.

I would also like to thank Xiajie Zhang, my wonderful stable partner for her support and love through nearly three years of PhD together. You are the reason I have been able to go through all the hardships that PhD life has thrown at me and I would like to share all my achievements with you, as they are all yours.

I want to express all my love to my parents. Thank you, it looks like you did not do a bad job. Thanks you for everything.

The Hunter Cancer Research Alliance played a key role during my PhD. On top of partially funding my PhD scholarship and funding my project through the RHD BTT award, they provided support for every aspect of my project. They were instrumental in the completion of this thesis. In addition to being an amazing structure, they are amazing human beings and I would like to personally thank Dr Stephen Ackland, Ann Thomas, Gemma Hulsing and Sarah Nielsen for all the help they provided.

Many thanks to Joshua Atkins, the Intersect people and Carlos Riveros. Thanks to you, I discovered genomic analysis, coding and bioinformatics.

My gratitude also goes to every single individuals in the IBM lab (and more generally level 3 west at HMRI). Thanks you for being so helpful and for all the smiles in the corridors, they help everyone to go through their day.

I also thank the University of Newcastle for funding a large portion of my PhD scholarship. UoN and HMRI are one of the best working environment I have ever seen.

I would also like to thank, in no particular order and for various reasons:

- Alexandre Gangnant and Paul Noël, you are the “ME” and “TE” to my “OR”.
- William Laporte. Before you, I was a pack of one wolf and now we are a pack of two wolfs.
- Chloe Sanini. For not giving up when I told her to.
- Ophélie Laporte and Romain Nugier. You are at the same time the chillest and most hardcore people I know.
- Ornella Tilly and Jérôme Xavier. For actually reading this thesis.
- Sofia Sanini. For being herself.
- Élodie Gourreau. For simultaneously making the best hummus and having the coolest hair.
- Marie-Caroline Grosset. For all the art.
- Debbie and John Amas for always being there for us.
- Sean Burnard. For still trying to keep me fit and being so wholesome.
- Kumar Uddipto. For Cookies!
- Mamta Pariyar. For being a little bundle of joy.
- Brianna Morten. For being the most helpful person I have ever met.
- Shankar.
- David Mossman and Michael Hipwell for teaching me the same things again and again

ABSTRACT

Colorectal cancer (CRC) remains a heavy burden for all national health systems. It is the third most frequently diagnosed cancer and the second leading cause of death in Australia and worldwide. Around 80% of CRC diagnosed each year are sporadic and somewhere between 7% and 8% have a clearly identified genetic predisposition (inherited CRC cancer; 5% for Lynch Syndrome (LS), 1% for Familial Adenomatous Polyposis (FAP) and 1-2% inclusive for various syndromes with very low incidences), with the remaining ~ 12%-13% being described as "familial". For many patients with a clinical diagnosis of LS and FAP, no causative mutation has been identified in *MSH6*, *MLH1*, *MSH2* or *PMS2* (for LS patients) and in *APC* or *MUTYH (*for FAP patients) as a result of genetic testing.

For those patients and their families, it is critical to identify the genetic cause underlying their increased CRC risk to offer early detection, tightened monitoring and, if required, suitable surgical management.

Establishing an exhaustive list of known genetic risk factors for inherited CRC is essential for families burdened with a high incidence of CRC. Patients with a strong family history of CRC will usually undergo a tighten monitoring. Removing this psychological burden in individuals proven to be non-carriers of pathogenic germline variants is critical.

Initial investigations focused on the Mismatch Repair (MMR) pathway in patients with LS and those with Lynch-Like Syndromes (LLS). 274 DNA samples from LLS patients were sequenced for the 22 genes involved in the MMR pathway to determine the presence of pathogenic variants. The results confirmed that LLS patients harbour pathogenic variants in genes that are not part of routine clinical screening: *POLD1*, *EXO1*, *MLH3*, *RFC1* and *RPA1*. The results indicate that additional MMR genes are involved in the increased risk of CRC in LLS patients.

As the technology evolved and became more cost-effective, whole exome sequencing (WES) was employed. Forty-eight patients with a clinical diagnosis of FAP were recruited based on their family history of CRC, their polyp status and their negative mutational status of *APC* and/or *MUTYH*. WES was used to interrogate all coding regions of the genome. Analysis of pathogenic variants showed that genes involved in DNA repair were frequently associated with a pathogenic variant. In addition, CNV analysis revealed the deletion of large portions of *CFHR3,* known to cause Atypical Haemolytic Uremic Syndrome, leading to ulcerative colitis, a known risk factor in CRC. Analysing the Polygenic Risk Score (PRS) for CRC risk-factors show an enrichment in inflammatory bowel syndrome-related markers.

During the WES analysis of FAP-like patients, an absence of a precise and automated method to predict pathogenicity in cohorts sharing the same phenotype was apparent. To overcome this, we

developed TAPES, a bioinformatics tool that can predict pathogenicity more precisely that can also calculate variant enrichment using only publicly available control sequences. TAPES also integrate powerful variant filtering and can generate useful reports (such as pathway analysis or calculating the total gene burden in a cohort).

In conclusion, the research presented herein helps strengthen the knowledge of familial CRC. The involvement of novel MMR genes in LLS was also revealed thereby expanding the known number of genes associated with this disorder. DNA-repair related genes as well as those involved in inflammation were shown to play an important role in FPS. Finally, a refined analytical pipeline for WES sequencing interpretation was developed providing new bioinformatics tools for the rapid delivery of results.

List of publications included as part of this thesis:

1) 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;7(8):e850.
2) Xavier A, Scott RJ, Talseth-Palmer BA. TAPES: A tool for assessment and prioritisation in exome studies. PLOS Computational Biology. 2019;15(10):e1007453.
3) Xavier A, Scott RJ, Talseth-Palmer BA. Exome sequencing of unexplained familial polyposis identifies both known and novel causative genes *To be submitted to Clinical Genetics (August 2020)*
4) Xavier A, Scott RJ, Talseth-Palmer BA. IBD-related markers associate with the age of onset for unexplained familial polyposis patients *To be submitted to Clinical Genetics (August 2020)*

List of additional publications:

1) Hansen MF, Johansen J, Sylvander AE, Bjornevoll I, Talseth-Palmer BA, Lavik LA, Xavier A, Engebretsen L. F, Scott R. J, Drablos F, Sjursen W. Use of multigene-panel identifies pathogenic variants in several CRC-predisposing genes in patients previously tested for Lynch Syndrome. Clin Genet. 2017. In Appendices section

List of oral/poster conference presentations:

1) Xavier A, Scott RJ, Talseth-Palmer BA. TAPES: a Tool for Assessment and Prioritisation in Exome Studies. *Australian Society for Medical Research satellite conference 2019*, Newcastle, NSW, Australia. Poster
2) Alexandre Xavier, Maren Fridtjofsen Hansen, Liss A. Lavik, Ashish Kumar Singh, Rodney J. Scott1,4, Wenche Sjursen and BenteA. Talseth-Palmer: New causative genes in inherited colorectal cancer: A new landscape of mutation for Hereditary Non-Polyposis Colorectal Cancer, *Australian Society for Medical Research conference* 2017*,* Sydney, Australia. Poster
3) Xavier A, Scott RJ, Talseth-Palmer BA. Colorectal Polyposis syndromes. Friday Seminar Series, 19 October 2018, Newcastle, NSW, Australia. Oral

4) Xavier A, Scott RJ, Talseth-Palmer BA. New causative genes in inherited colorectal cancer: Colorectal Polyposis syndromes. *HEaPS seminar 29 March 2017,* Newcastle, NSW, Australia. Oral

5) Hansen MF, Johansen J, Sylvander AE, Bjornevoll I, Talseth-Palmer BA, Lavik LA, Xavier A, Engebretsen L. F, Scott R. J, Drablos F, Sjursen W. A new landscape of mutation for Hereditary Non-Polyposis Colorectal Cancer. *HCRA conference 2018, Rapid Fire Session.* Newcastle, NSW, Australia. Oral

6) Xavier A, Scott RJ, Talseth-Palmer BA.: Genetics in Inherited Colorectal Cancer, HCRA community showcase, 2017, Newcastle NSW, Australia

List of Awards:

1) *HCRA RHD* Award from the Biomarkers and Targeted Therapies Flagship, Hunter Cancer Research Alliance, 2017, AUD $5000 for research purpose
2) *Highly commended Rapid Fire* for A new landscape of mutation for Hereditary Non-Polyposis Colorectal Cancer. *HCRA conference 2018, Rapid Fire Session.*
3) HCRA Future Leaders Group publication award for Xavier A, Scott RJ, Talseth-Palmer BA. TAPES: A tool for assessment and prioritisation in exome studies. PLOS Computational Biology. 2019;15(10):e1007453. AUD $2500 toward open access publication fees

List of Abbreviations

Abbreviation	Expanded term
AC	Amsterdam Criteria
ACMG	American College of Medical Genetics
AMP	Association for Molecular Pathology
BER	Base Excision Repair (pathway)
BG	Bethesda Guidelines
bp	Base Pair
CRC	Colorectal Cancer
CS	Cowden Syndrome
DNA	DeoxyriboNucleic Acid
FAP	Familial Adenomatous Polyposis
FCCTX	Familial Colorectal Cancer Type X
FPS	Familial Polyposis Syndrome
GO	Gene Ontology
HNPCC	Hereditary Non-Polyposis Colorectal Cancer
IBD	Inflammatory Bowel Disease
IHC	Immunohistochemistry
Indel	Insertion/Deletion
JPS	Juvenile Polyposis Syndrome
LLS	Lynch-Like Syndrome
LS	Lynch Syndrome
MAP	MUTYH-Associated Polyposis
MLPA	Multiplex Ligation-dependent Probe Amplification
MMR	MisMatch Repair (pathway)
MSI	Micro-Sattelite Instability
NAP	NTHL1-Associated Polyposis
NGS	Next-Generation Sequencing
NSAID	Non Steroidal Anti-Inflammatory Drug
OMIM	Online Mendelian Inheritance in Man
PHTS	PTEN Hamartoma Tumour Syndrome
PJS	Peutz–Jeghers Syndrome
PPAP	Polymerase Proofreading-Associated Polyposis
PRS	Polygenic Risk Score
SAM / BAM / CRAM	Sequence Alignment Map / Binary / Compressed
SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SPS	Serrated Polyposis Syndromes
VCF	Variant Calling Format (file)
VUS	Variant of Unknown Significance
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

List of Tables

List of Figures

TABLE OF CONTENTS:

CHAPTER 1

Colorectal cancer – General Background

CHAPTER 1: Colorectal cancer – General Background

1.1)General Colorectal Cancer Background

Worldwide cancer incidence

Cancer (malignant neoplasms) is the second leading cause of non-communicable death worldwide (after cardiovascular diseases) with over 8,966,000 deaths each year (1). Both incidence and mortality of cancer are increasing worldwide. This is due to a combination of an aging population and both lifestyle and environmental factors. Recently, cancer was found to be the leading cause of death in high-income countries (2).

Colorectal cancer (CRC) is a worldwide burden on the health systems. CRC is the third most frequently diagnosed cancer (1) with more than 1.8 million newly diagnosed cases each year. It is also the second leading cause of death from cancer with more than 794,000 deaths each year (1).

Most CRCs (around 80% (3)) are sporadic, which suggests that they are caused by lifestyle or environmental factors.

Risk factors

CRC risk factors include, inflammatory bowel disease, smoking, exercise, alcohol, obesity (4), diet and diabetes to mention but a few (see Table 1). The most significant risk remains age with a cumulative risk of developing CRC of 0.35% under 49 to 3.15% above 70 (5). However, there is no single individual risk factor that can explain CRC risk.

Diet is a very important risk factor with recent research showing high levels of ultra-processed food and high consumption of red meat increased CRC risk (6, 7). This has now been coupled with metagenomics data that indicates a new potential causal relationship (8).

Table 1 List of risk factors for CRC. HR = Hazard Ratio, CI = Confidence Interval from "*A Comprehensive Model of Colorectal Cancer by Risk Factor Status and Subsite Using Data From the Nurses' Health Study*" (9)

Risk Factor	**Colorectal Cancer (n = 1,759)**		
	HR	95% CI	P Value
Age (years; 60 vs. 50)	1.81	1.70 - 1.92	<0.0001
Family history of colon or rectal cancer (yes vs. no)	1.45	1.29 - 1.63	<0.0001
Red meat intake (servings/day per year; 1 vs. 0)	1.01	0.87 - 1.18	0.87
Processed meat intake (servings/day per year; 1 vs. 0)	1.11	0.92 - 1.33	0.29
Folate intake (µg/day per year; 600 vs. 200)	0.83	0.74 - 0.95	0.004
Smoking history (total pack-years; 40 vs. 0)	1.2	1.10 - 1.31	<0.0001
BMI (units per year; 30 vs. 20)	1.37	1.19 - 1.57	<0.0001
Physical activity level (MET-hours/week per year; 21 vs. 2)	0.61	0.48 - 0.76	<0.0001
Height (inches per year; 67 vs. 61)	1.24	1.09 - 1.41	0.001
Alcohol (g/day per year; 30 vs. 0)	1.15	0.98 - 1.34	0.082
Aspirin use (tablets per week per year; 7 vs. 0)	0.78	0.70 - 0.86	<0.0001
Endoscopic screening (yes vs. no; 20 years vs. 0)	0.74	0.67 - 0.83	<0.0001
Calcium intake (mg/day per year; 1,000 vs. 500)	0.82	0.73 - 0.91	0.0002

Evolution of colorectal cancer incidence

The number of CRC cases tends to decrease each year in developed countries. However, with a rapidly changing lifestyle in developing countries, some countries have seen the incidence of CRC increasing recently (10).

Colorectal cancer management in Australia

In Australia, there have been campaigns for the early detection of CRC. Due to the increased risk of disease after the age of 50 years, a detection kit is provided to all citizens over 50 years of age within 6 months of their birthday. Thereafter, a screening kit is sent every 2 years, until the age of 74 years.

The kits detect trace quantities of blood contained in faeces. Because of the nature of this test, acceptance rates of the test remain rather low (reaching 40.9% of the eligible population during the 2015-2016 campaign), but is improving (from 36.1% during the 2012-2013 campaign) (11).

Systematic screening has proven to be effective in reducing mortality: individuals with CRC detected as part of the screening program had a significantly lower mortality rate than those with CRC detected outside of the screening program (Hazard Ratio = 0.39, CI = 0.35-0.43) (11).

Colon and rectum anatomy

The human gastrointestinal tract starts from the mouth followed by the oesophagus. The stomach is the first organ processing food intake, which is then driven to the small intestine and then the colon (or large intestine). The colon then leads to the rectum. On a more functional level, the stomach is responsible for pre-processing food. The liver, gallbladder and pancreas secrete enzymes and chemicals that help breaking down fat, proteins and carbohydrates. The small intestine is where most of the nutrients are absorbed while water and electrolytes are mostly absorbed through the colon. The rectum and anus are then responsible for the excretion (see Figure 1).

The combination of the descending and the sigmoid colon is referred to as distal colon (or left side) while the combination of the ascending colon and the transverse colon is referred to as proximal colon (or right side).

There is evidence showing that CRC originating from distal colon and proximal colon exhibit differences in protein expression and overall molecular profile (12). Similarly, different CRC syndromes will result in more distal or proximal tumours (13).

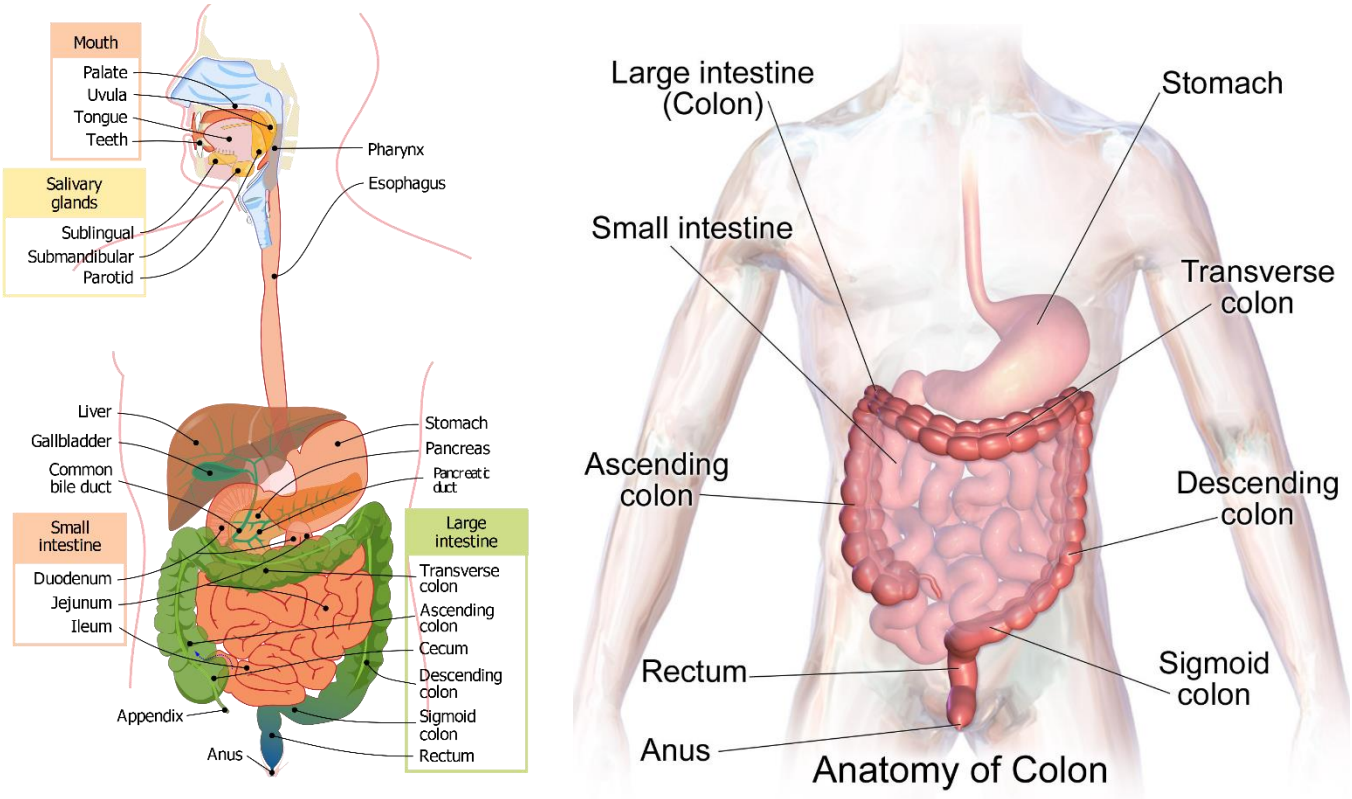

Figure 1. Anatomy of the gastrointestinal tract, colon and rectum Left: Anatomy of the gastrointestinal tract. Right: Anatomy of the colon and rectum (Illustrations credit: Wikimedia, Wikibooks)

1.2) Hereditary Colorectal Cancer syndromes

Approximately 20% of all CRCs can be explained by genetic factors. Two different cases need to be distinguished. First, hereditary CRC syndromes can be defined as all the predisposition syndromes causing CRC with a known genetic cause. They need to be differentiated from the so-called familial CRC, where cancers can be theoretically linked to genetic causes, but no known causative gene has been identified.

Lynch Syndrome/Hereditary Non-Polyposis Colorectal Cancer

Disambiguation

Lynch Syndrome (LS) and Hereditary Non-Polyposis Colorectal Cancer (HNPCC) can be confusing terms when going through the literature (14). They are often used as synonyms, but their respective definition is not identical.

> *"HNPCC is defined clinically, usually as families satisfying Amsterdam I or II criteria. Lynch syndrome is defined genetically, by the presence of a germline mutation in DNA mismatch repair (MMR) or EPCAM genes." Quote From Kravochuck et al. 2014 (14)*

An individual is diagnosed with LS if they have an identified pathogenic variant in *MLH1*, *MSH2*, *MSH6, PMS2* or a deletion of the 3' end of *EPCAM*. The first four genes are involved in the DNA Mismatch Repair (MMR) pathway,

while *EPCAM* deletions leads to epigenetic silencing of *MSH2* as a result of *EPCAM* transcriptional start site read through (15). Individuals diagnosed with LS are not required to have cancer to be considered LS as they have a molecular diagnosis of the disease. Indeed, LS individuals have a higher risk of cancer but not every LS individual develops cancer (see Cancer Institute NSW guidelines (16)).

HNPCC is, by definition, referring to all CRC inherited genetically without polyps (as opposed to FAP which exhibits a polyp phenotype). While Henry T. Lynch referred to the initial cases as "Cancer Family Syndrome" in 1966 (17), he used HNPCC in 1985 (18) to distinguish between families affected only by CRC and those with increased risk of other cancers. The Amsterdam Criteria (AC) (19) and the refined Amsterdam criteria II (ACII) (20) (see Table 2) were tools used to identify HNPCC. HNPCC was the term generally used before the molecular background of LS was well established. As such, HNPCC also encompasses Familial Colorectal Cancer Type X (FCCTX, discussed later in this thesis), not presenting pathogenic MMR variants.

Table 2. The Amsterdam Criteria I and II

The Amsterdam Criteria (AC I)
• At least 3 relatives have been diagnosed with **CRC**, with at least one being a first degree relative to the other, polyposis are excluded. • At least 2 successive generations should be involved • At least 1 CRC patient should be diagnosed before the age of 50 years
The Amsterdam Criteria II (AC II)
• At least 3 relatives have been diagnosed with **HNPCC-related cancers**, with at least one being a first degree relative to the other, polyposis are excluded. • At least 2 successive generations should be involved • At least 1 CRC patient should be diagnosed before the age of 50 years

The ACI and ACII are empirical criteria that pinpoints the hereditary basis of cancers (using early age of onset as well as several affected individuals).

In addition to the AC, the Bethesda Guidelines (BG) (21) and the Revised Bethesda Guidelines (RBG) (22), are guidelines to follow to decide whether to test patients for Microsatellite instability (MSI).

Table 3. The Revised Bethesda Guidelines

Tumours from individuals should be tested for MSI in the following situations:

1. Colorectal cancer diagnosed in a patient who is less than 50 years of age.
2. Presence of synchronous, metachronous colorectal, or other HNPCC-associated tumors,[*] regardless of age.
3. Colorectal cancer with the MSI-H[†] histology[‡] diagnosed in a patient who is less than 60 years of age.[§]
4. Colorectal cancer diagnosed in one or more first-degree relatives with an HNPCC-related tumor, with one of the cancers being diagnosed under age 50 years.
5. Colorectal cancer diagnosed in two or more first- or second-degree relatives with HNPCC-related tumors, regardless of age.

[*]Hereditary nonpolyposis colorectal cancer (HNPCC)-related tumors include colorectal, endometrial, stomach, ovarian, pancreas, ureter and renal pelvis, biliary tract, and brain (usually glioblastoma as seen in Turcot syndrome) tumors, sebaceous gland adenomas and keratoacanthomas in Muir–Torre syndrome, and carcinoma of the small bowel

[†]MSI-H = microsatellite instability–high in tumors refers to changes in two or more of the five National Cancer Institute-recommended panels of microsatellite markers.

[‡]Presence of tumor infiltrating lymphocytes, Crohn's-like lymphocytic reaction, mucinous/signet-ring differentiation, or medullary growth pattern.

[§]There was no consensus among the Workshop participants on whether to include the age criteria in guideline 3 above; participants voted to keep less than 60 years of age in the guidelines.

The BG were kept much less restrictive than the AC to identify most of the LS cases. But, when screening for *MLH1* and *MSH2* pathogenic variants, the AC reached a sensitivity of 61% and a specificity of 67% while the BG achieved a sensitivity of 94% and a specificity of only 25% (23).

In practice, LS and HNPCC are often used interchangeably but in this thesis the term Lynch Syndrome will be used.

LS Description

Each year, 5% of all diagnosed CRC are caused by LS, making it the most diagnosed hereditary type of CRC. As described above, LS is diagnosed by screening MMR genes for pathogenic variants. These genes are all involved in the DNA MMR pathway and pathogenic variants will affect the capacity to recognise DNA mismatches.

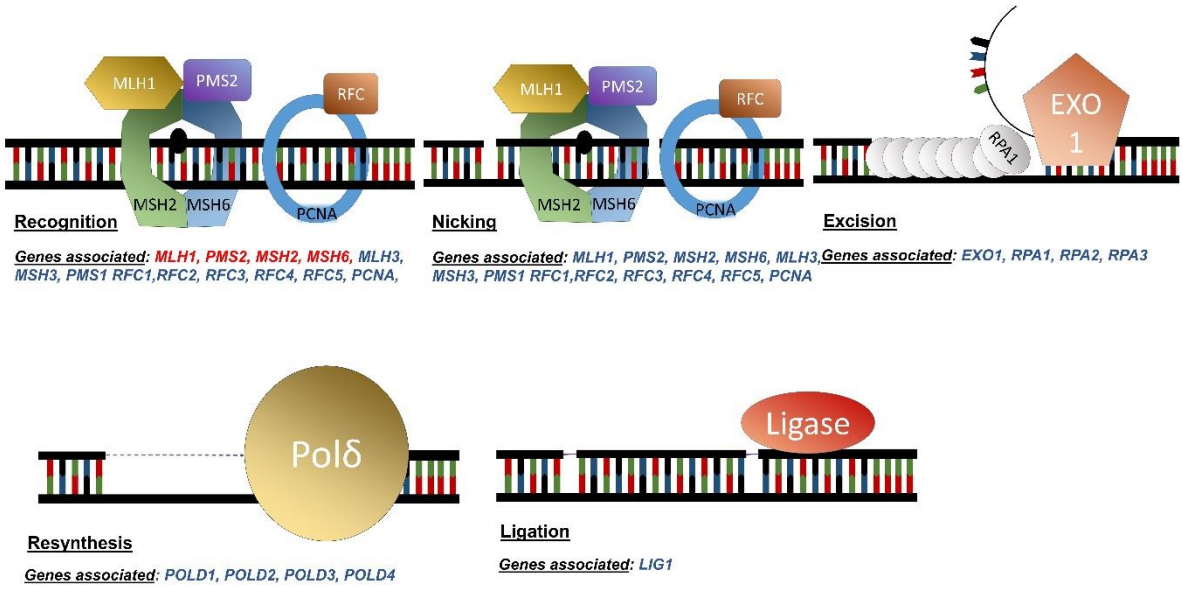

Figure 2. Genes involved in the DNA MMR pathway by step (from Xavier et al. (24))

The DNA MMR process corrects both mismatched bases and small insertions/deletions (indels) that occurred during DNA replication. An observable phenotype of the alteration of the MMR efficiency as a result of inherited pathogenic variants is MicroSatellite Instability (MSI).

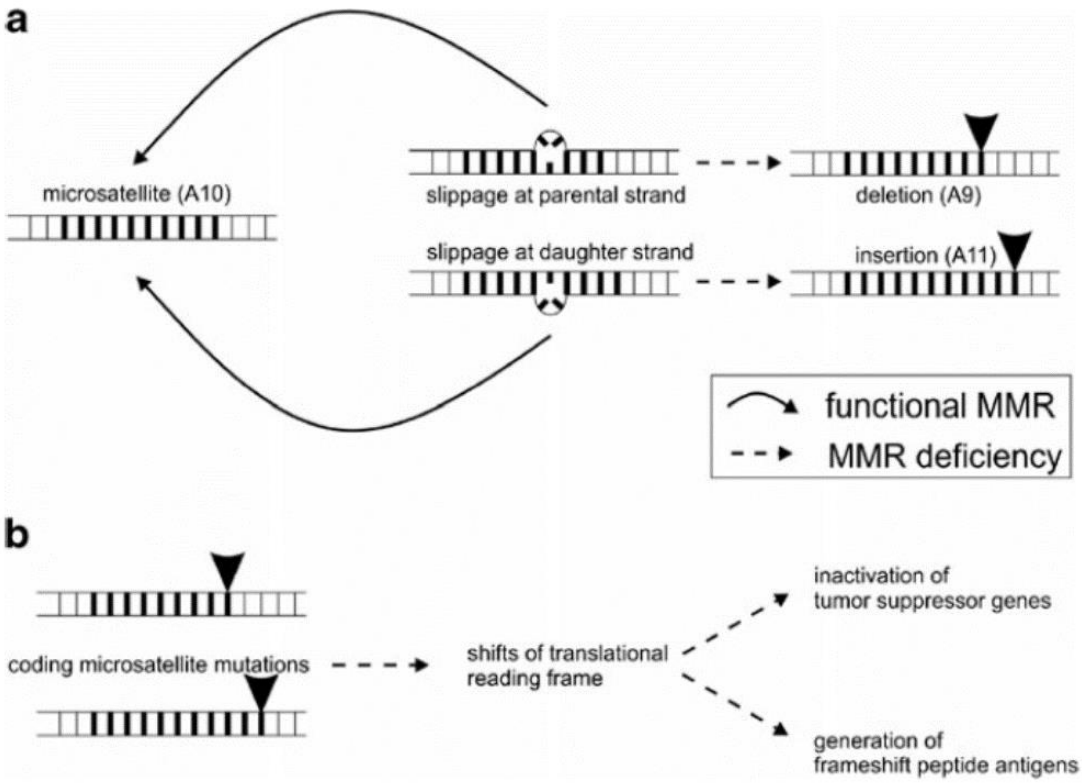

Figure 3. Microsatellite instability process. a) Process b) Consequences from Kloor et al (25)

Microsatellites are repetitive regions of the genome where one or more (usually 1 to 6) nucleotides will be repeated n-times (typically from 5-50 times (26)). They represent roughly 3% of the whole human genome (27). MSI is mainly due to slippage of polymerase during DNA replication, happening both in prokaryotes and eukaryotes (28) . High MSI is a recognised marker of LS derived tumours and testing for MSI can be guided using the Revised Bethesda Guidelines.

A portion of sporadic cancer are, however, also characterised as MMR-deficient (MMR-D) and can exhibit high MSI, underlining the importance of both germline and tumour testing for CRC (29).

Penetrance and risks of cancer are different between the genes causing LS. While pathogenic variant carriers in *MLH1*, *MSH2* and *MSH6* have a respective risk of any LS-related cancer of 77.1%, 81% and 52% by the age of 75 (30), carriers of *PMS2* mutations have a risk of any LS cancer of 34% (with a risk of CRC of 10.4%). Recently, there is growing evidence for PMS2 being a recessive disease rather than dominant (31) with a low risk of CRC in heterozygous pathogenic variant carriers.

These findings (see Table 4) were obtained through the Prospective Lynch Syndrome Database (PLSD) which monitors pathogenic variant carriers over time to evaluate cancer risks associated with each MMR gene, gender and age.

In addition to heterogeneity of penetrance between MMR genes, there is also significant heterogeneity of penetrance between individuals. As illustrated in figure 4, the curve representing cumulative risk of CRC has a distinctive U shape (32). For example, while around, 25% of MLH1 carriers have under 10% cumulative risk of CRC, only 7% have a 50-60% risk. At the other end of the spectrum, around 10% or MLH1 carriers have more than 90% of CRC risk at 70 years. This highlights the need to investigate modifier genes and cofactors modulating the risk of CRC in MMR mutation carriers.

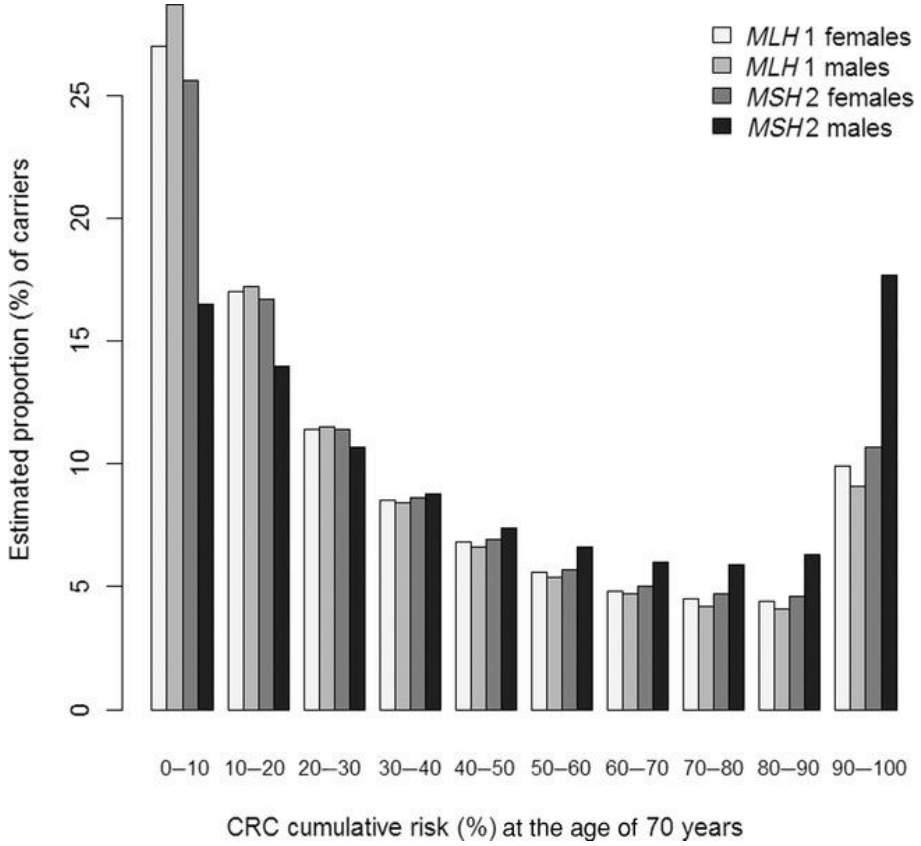

Figure 4. Estimated distribution of CRC cumulative risk at the age of 70 years for each combination of gene and sex (adapted from Moller et al 2017)

LS-associated cancer

Being caused by DNA repair deficiency, LS does not only confer a higher risk of colorectal cancer. All tissues with epithelial cells have an increased risk of high MSI (see Table 4).

LS has been proven to increase the risk of several other cancer, especially endometrial cancer in women with a lifetime cumulative risk of 48.9% for MMR pathogenic variant carriers (30).

In addition to non-colorectal LS-related cancers, there has been reports of MMR mutations being linked to a polyp phenotype (33).

Table 4. Cumulative incidence at 75 for various LS-associated cancers by gene

	Cumulative incidence at 75 (% [95% CI]) from Dominguez-Valentin et al. (30)						
	MLH1		***MSH2***		***MSH6***		***PMS2***
Organ	**Females**	**Males**	**Females**	**Males**	**Females**	**Males**	**Both**
Any cancer	**81.0** [74.1–88.4]	**71.4** [62.8–81.3]	**84.3** [77.1–91.0]	**75.2** [65.6–85.7]	**61.8** [47.3–78.3]	**41.7** [25.4–67.1]	**34.1** [19.0–59.6]
Colorectal	**48.3** [40.9–57.4]	**57.1** [48.7–67.9]	**46.6** [39.1–55.4]	**51.4** [41.0–65.0]	**20.3** [11.8–40.5]	**18.2** [7.9–43.2]	**10.4** [2.9–40.8]
Endometrium	**37.0** [30.1–46.5]		**48.9** [40.2–60.7]		**41.1** [28.6–61.5]		**12.8** [5.2–49.5]
Ovaries	**11.0** [7.4–19.7]		**17.4** [11.8–31.2]		**10.8** [3.7–38.6]		**3.0** [0.5–43.3]
Stomach, small bowel, bile duct, gallbladder and pancreas	**11.0** [7.4–16.9]	**21.8** [16.0–29.9]	**12.8** [8.8–19.3]	**19.5** [14.0–27.6]	**4.2** [1.2–26.0]	**7.9** [2.7–30.0]	**3.6** [1.0–33.5]
Ureter and kidney	**3.8** [1.9–8.4]	**4.9** [2.5–10.6]	**18.7** [13.5–26.5]	**17.6** [12.6–25.3]	**5.5** [2.2–26.9]	**1.7** [0.3–24.3]	**3.7** [0.7–33.8]
Prostate		**13.8** [8.8–21.7]		**23.8** [17.2– 33.2]		**8.9** [3.1–31.0]	**4.6** [0.8–67.5]
Breast	**12.3** [8.6–17.9]		**14.6** [10.3– 21.1]		**13.7** [7.4–33.8]		**15.2** [5.9- 51.5]
Brain	**1.6** [0.6–5.3]	**0.7** [0.1–5.2]	**2.9** [1.2–7.9]	**7.7** [4.1–15.2]	**1.2** [0.2–23.4]	**1.8** [0.3–24.4]	**0** [0–30.9]

Clinical Management and diagnosis

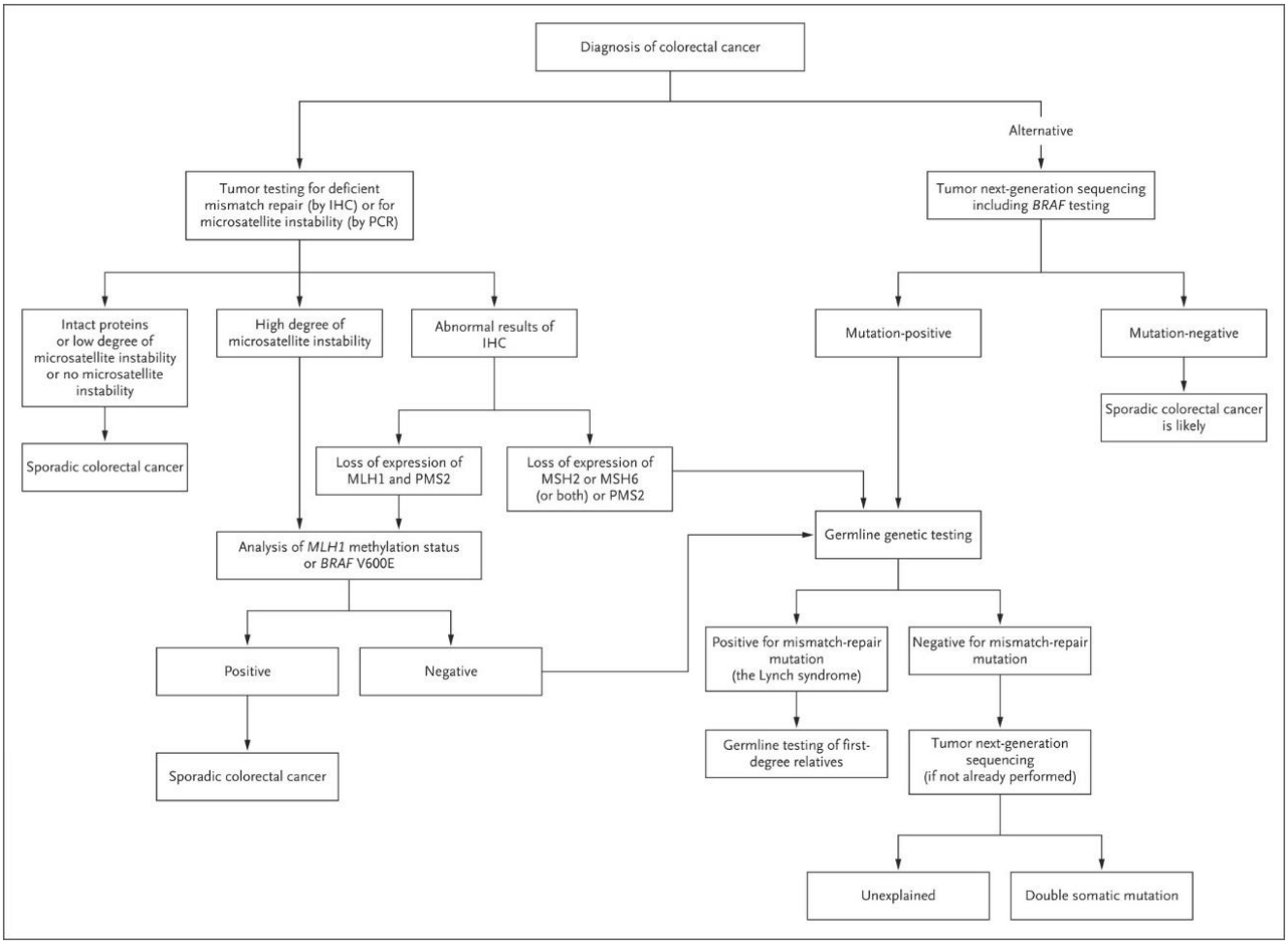

Figure 5. Algorithm for Lynch Syndrome testing in patients with a new diagnosis of CRC. From Sinicrope et al. (34)

The diagnosis of LS is often performed after a new diagnosis of CRC (see Figure 5). The recommended management of LS patients is surveillance through regular colonoscopy. The frequency of the colonoscopies depends on the patient's geographical location, different countries have different practices.

Several different recommendations for risk management exists. In Australia, New South Wales recommends the eviQ guidelines (35), which recommends regular (every 1 to 2 years) colonoscopy, with a starting age depending on the MMR gene affected. There are currently no recommendation for systematic screenings for cancers other than CRC in patients diagnosed with LS (36), mostly due to poor sensitivity, especially for transvaginal ultrasound (endometrial cancer) and CA125 screening (ovarian cancer). Hysterectomy and bilateral salpingo-oophorectomy are recommended for select individuals.

The US preventative taskforce guidelines (37) also recommend regular colonoscopy. But practitioners also offer to the patients the possibility to be screened for endometrial, ovarian, gastric and urinary tracts cancers. The European guidelines, similarly, do not recommend systematic screening for extra-colonic cancers but encourage doctors to act on a case-by-case basis based on patients risks (38), also referencing poor sensitivity of current screening methods.

However, the general recommendation is to test at least every second year after the age of 20 or 10 years younger than the age of first diagnosis in the family (39). There is no known preventative treatment for LS-related CRC although aspirin and other non-steroidal anti-inflammatory drugs (NSAID) have been shown to reduce the risk of CRC in general (9, 40) but in LS in particular (41)

Familial Adenomatous Polyposis and polyposis syndromes

FAP and APC

Familial Adenomatous polyposis (FAP), is the second most commonly diagnosed hereditary CRC syndrome and accounts for around 1% of all diagnosed CRC worldwide.

Individuals are diagnosed with FAP if genetic screening detects pathogenic variants in the gene *APC*. *APC* is a gene involved in the Wnt-β catenin pathway and produces the APC protein. APC and Axin act as scaffolds (42), binding to CK1α and GSKβ, allowing β-catenin to be phosphorylated (and thus creating a β-catenin destruction complex). Phosphorylated β-catenin will then lead to its own degradation through ubiquitination (see Figure 5).

Most of the disease-causing mutations in the *APC* gene are truncating. A truncated APC protein will not allow the CK1α-GSKβ-Axin-APC-β-catenin complex to form, leading to the accumulation of β-catenin in the cell and β-catenin translocation to the nucleus, where it will act as a transcription co-factor to TCF/LEF transcription factors. This leads to the constitutive expression of several oncogenes including cyclin D1 and Axin.

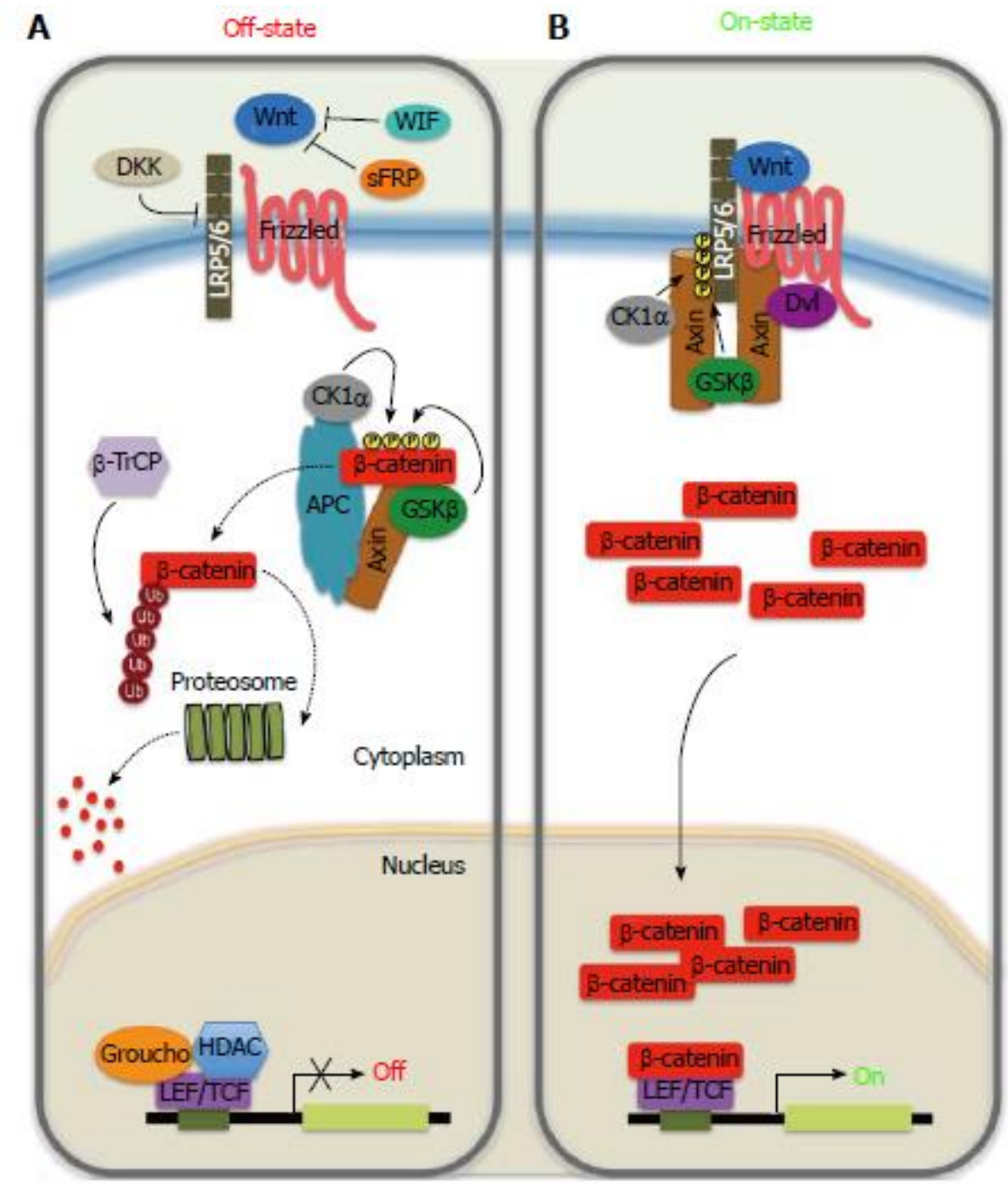

Figure 6. Wnt pathway associated with FAP. Left, Inactive Wnt pathway leading to the ubiquitination of β-Catenin and its processing by the proteasome. Right, active Wnt pathway leading to the accumulation of β-Catenin and the expression of various oncogenes. From Chiurillo et al. (43)

In addition to its role in the β-catenin destruction complex, APC also acts as a scaffold protein for microtubule end-binding proteins (EB) through its c-terminal domain. The APC-EB interactions regulate the dynamics of microtubules during mitosis (spindle formation and chromosome segregation). In the context of CRC, haplo-insufficiency APC results in mis-segregation of chromosome and chromosome instabilities (44).

Pathogenic variants in *APC* will result in a distinct polyp phenotype. Polyps will start as adenomas and then progress along the adenoma-carcinoma sequence, ultimately evolving into a malignant carcinoma (45). Most of the individuals with FAP will accumulate polyps over time, usually presenting with 1000s of polyps at the time of diagnosis. However, an attenuated version of FAP, attenuated FAP (aFAP) exists, where the numbers of polyps remain lower (under 100). Over 60% of cases of aFAP are thought to be caused by *APC* mutations. Patients with a pathogenic variant in the 3' end (46) , 5' end (47) or the alternatively spliced site of exon 9 (48) of *APC* are often associated with an aFAP phenotype. This is often associated with the degradation at either mRNA or APC, leading to haplo-insufficiency of wild-type APC protein.

FAP and aFAP are both caused by pathogenic *APC* variants, which for FAP approaches 100% penetrance and for aFAP less than 100% penetrance (49).

Other Polyposis syndromes

Several other polyposis syndromes have a clearly defined genetic background (50). Two main different families of syndromes can be differentiated, adenomatous and hamartomatous polyposis (see table 5 for classification).

Table 5. Polyp classification and associated causes Adapted from Colucci PM et al. (51)

Histological Classification	Polyp Type	Malignant Potential	Cause
Non-neoplastic	Hyperplastic polyps	No	Sporadic
	Hamartomas		PJS, JPS, PHTS, Sporadic
	Inflammatory polyps		Ulcerative colitis, Crohn's disease
Neoplastic (adenomas)	Tubular adenomas (0–25% villous tissue)	Yes	FAP, NAP, MAP, PPAP, LS, Sporadic
	Tubulovillous adenomas (25–75% villous tissue)		
	Villous adenoma (75–100% villous tissue)		

MUTYH (or *MYH*) Associated Polyposis (MAP) is a well described polyposis syndrome (52). Patients with bi-allelic pathogenic variants in *MUTYH*, which is part of the Base Excision Repair (BER) pathway, have a 28-fold increased risk of CRC and adenomatous polyps (53).

NTHL1-associated polyposis (NAP) is an autosomal recessive syndrome associated with bi-allelic variants in *NTHL1 (54)*. *NTHL1* is, like *MUTYH*, a gene involved in BER. As a result, phenotypes are similar, with a reduced polyp count (less than 100) adenomatous polyps.

Polymerase Proofreading-Associated Polyposis (PPAP), is a polyposis syndrome resulting in adenomatous polyps. It is primarily caused by missense variants in the proofreading domain of the polymerase POLD1 and POLE and not by nonsense or truncating mutations (55, 56). The proofreading domain loses efficiency and promotes the accumulation of genetic aberrations.

LS have also been shown to be associated with a polyp phenotype. However, LS-associated polyps do not constitute a full polyposis phenotype. While most of the individuals with LS will have a relatively low number of polyps (83% under 10 polyps), some individuals can have up to 50 polyps (4%) (57), making the distinction between LS and aFAP difficult. It remains to be determine if these lesions are in fact precursors to disease in LS since the LSDB suggests that this may not be the case (Seppaler et al. 2018)

Peutz–Jeghers syndrome (PJS) is an autosomal dominant syndrome caused by pathogenic variants in *STK11*. It is characterised by pigmented lesions around the mouth, anus, nostrils and fingers arising at an early age. In addition, individuals with PJS develop polyps which are described as hamartomatous. Sporadic hamartomas are generally benign. PJS-related hamartomatous polyps (and all Hamartomatous polyps caused by genetic syndromes) will not evolve into carcinomas but rather are the result of altered stem cell lineage turnover rates that will lead to an acceleration of the progression of cancer (58).

Juvenile Polyposis syndrome (JPS) is an autosomal dominant syndrome caused by pathogenic variants in either *SMAD4* or *BMPR1A* (59). Like PJS, JPS results in a hamartomatous polyp phenotype. JPS is often diagnosed during early life (around 16 to 18) and patients exhibit colorectal polyps 80% of the time. The number of polyps is much lower than FAP, with around 3 to 10 polyps.

PTEN hamartoma tumour syndrome (PHTS) and Cowden syndrome (CS) are autosomal dominant syndromes caused by pathogenic variants in the *PTEN* gene. The PTEN protein is part of the PI3K/AKT pathway, a well described mechanism of tumorigenesis when deregulated. PHTS and CS both result in colorectal hamartomatous polyposis. In addition, individuals affected have an elevated risk of breast, thyroid, endometrial and renal cancers (60) and other non-malignant features.

Serrated polyposis syndromes (SPS), is a syndrome with an unclear genetic origin. However, it is known to have been associated with pathogenic variants in *RNF43* along with pathogenic variants in *BRAF* (61). SPS has a specific phenotype of hyperplastic polyps, sessile or otherwise known as, serrated adenomas).

Clinical management

FAP associated with *APC* pathogenic variants has a nearly 100% penetrance. Recommended guidelines include regular colonoscopy, starting between the age of 15 and 20, every 2 years. If an adenoma is detected, removal should be performed every year until the number of adenomas is too many and then prophylactic colectomy is recommended. For families with a clinical diagnosis of FAP but no *APC* mutation identified, colonoscopy every 2 years is recommended after 20, every 3 to 5 years after 40 and can cease after 50 years of age (62).

Prophylactic colectomy for FAP patients remains the most common intervention and is the most efficient approach to reduce mortality. While there is no recommended age, most patients undergo this procedure between the age of 15 and 25 (63).

The use of aspirin (64) and other NSAID (especially sulindac (65)) have been considered for chemo-prevention of FAP. However, contrary to LS, results were divergent (see review (66)). Aspirin treatment showed a non-significant trend in polyp prevention. Sulindac treatment yielded better results but exhibited side effects such as rectal mucosal erosions. Currently, use of aspirin and other NSAID is not recommended for FAP patients.

1.3) Familial Colorectal Cancer syndromes

Familial Colorectal Cancer syndromes describe cases where there is strong evidence to support a genetic cause, but no clear causative gene or pathogenic variant has been identified.

Risk factor inheritance

While environmental and lifestyle-related risk factors increase the risk for a single individual to develop cancer, it is also important to remember that family members also share (to a certain extent) the same environment and lifestyle and those can also be inherited. Children of smokers will have a higher risk of smoking (67). In a similar manner, children with at least one obese parent will have a higher risk of being obese themselves (68). It is important to keep in mind that all inherited risk-factors are not genetically transmitted. While this thesis will focus only on genomic aberrations (analysis of variants and copy number variations), inherited risk factors for CRC could be epigenetic variations (DNA methylation or histone modification) or inherited mitochondrial diseases (69). Lifestyle or environmental-related factors can also be transmitted from parents to children and could play a role in increased incidence of CRC in particular families.

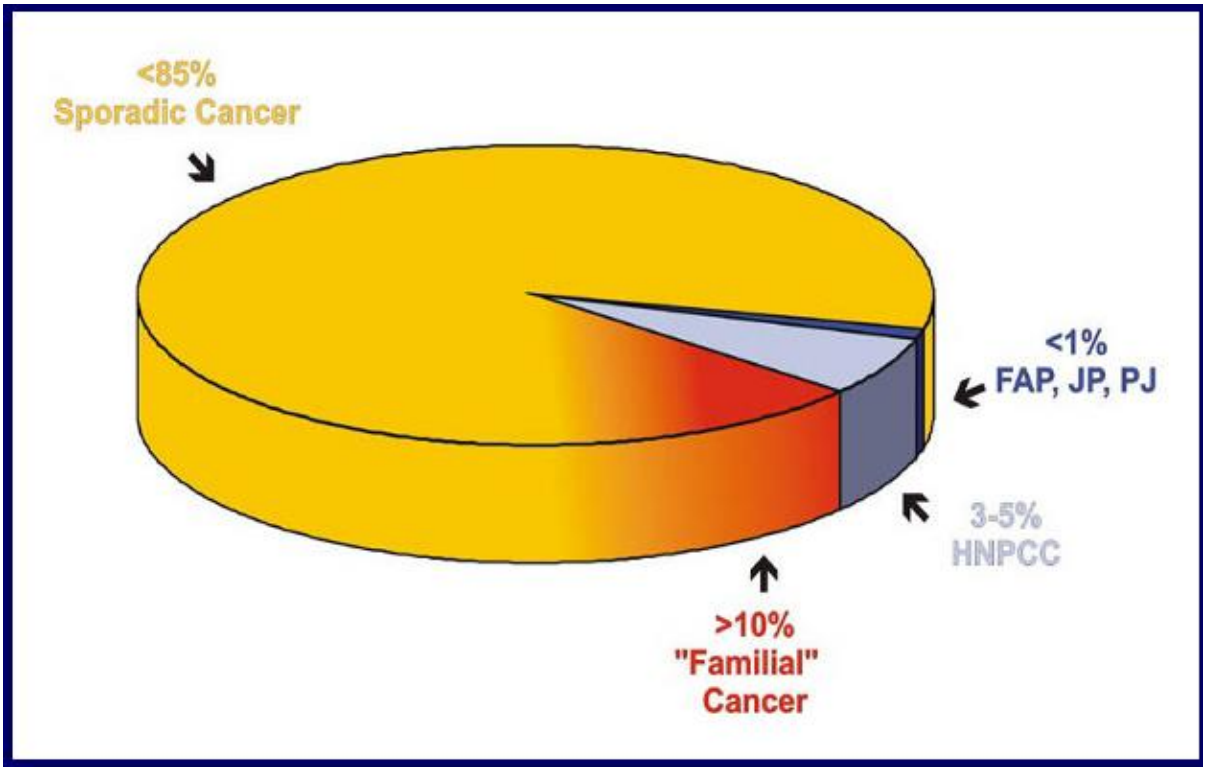

Figure 7. Proportion of the different types of CRC diagnosed each year from Half et al. (70)

Non-Polyposis syndromes

While around 5% of all colorectal cancers diagnosed each year are linked to LS with patients showing impaired MMR, there are still a large proportion (up to 40% (71)) of individuals fulfilling the ACI who come back as MMR mutation-negative after clinical screening. The syndrome affecting these individuals is referred to as Lynch-Like Syndrome (LLS) or Familial Colorectal Cancer Type X (FCCTX).

LLS-derived tumours have a quite heterogeneous molecular profile, but studies suggest that they do share some commonalities.

Patients usually have a later age of cancer onset (> 50 years) and a much lower rate of MSI-high tumours (72) (MSI-high being a marker of MMR deficiency). There has been no identified genetic cause for LLS but there has been research pointing toward other MMR genes (24), BRCA2 (73) and even telomere length (73). However it is most likely that LLS is an heterogeneous disease with numerous low-risk variants associated with disease presentation (74).

Polyposis Syndromes

Most cases of polyposis are caused by inherited pathogenic variants in *APC*. However around 25% of FAP cases are caused by de-novo mutations in *APC*. In additions, several other polyposis syndromes with an obvious genetic

background have been identified lately (MAP, PPAP, NTHL1-associated polyposis, *AXIN1*-associated polyposis (75) etc.), explaining an additional percentage of inherited polyposis.

But there is still, like LLS, a portion of CRCs associated with a polyp phenotype that can be described as "familial", with no known genetic cause. Those syndromes will be referred to as Familial Polyposis Syndrome (FPS) or FAP-like.

Familial CRC

Both FPS and LLS represent around 10% of all diagnosed CRCs each year (including both sporadic and inherited CRCs) (70). Identifying the cause of these familial CRCs is extremely important. It allows individuals with a higher risk of CRC to be better identified and monitored. Although they will be offered colonoscopy due to their family history of cancer, disease management might not be as efficient compared to patients with a genetic predisposition.

Furthermore, a significant proportion of patients with a clinical diagnosis of LS or FAP do not show any pathogenic variant in the usual causative genes. 30-40% (76) of patients diagnosed with LS are mutation-negative in MMR genes and 20-30% (77) of patients clinically diagnosed with FAP lack a pathogenic variant in *APC* (77) respectively. For polyposis syndromes, some of the remaining cases can in part be explained by recently-discovered polyposis syndromes (mentioned above), but their very low incidence only explains a fraction of all cases of FPS.

For these patients, it is important to identify causative genes and pathogenic variants to both explain the cause of CRCs and better treat them. This allows clinicians to distinguish between inherited and sporadic entities, leading to a better management of the disease.

1.4) Next-Generation sequencing and Tools for analysis

Next-Generation Sequencing (NGS) revolutionised how clinicians and researchers approach genomic data. Sanger sequencing allowed to interrogate one locus (up to 600-1000bp) at a time whereas NGS can sequence multiple patients at a time for whole genes, panels of whole genes, full human exomes and even full human genomes with the most recent sequencing platform allowing up to 3000Gbp (around 48 whole genomes) to be sequenced in the same run. This allows researchers to examine the individual genomic status of a patient to better understand the underlying genetic cause of their disease. NGS is qualified as targeted if it only covers a specific panel of genes.

Current advances in next-generation sequencing, both in terms of quality and price, allows researchers to sequence ever more patient samples. The sheer amount of data generated by whole exome (WES) or whole genome sequencing (WGS) requires an extensive downstream analysis. Regardless of the technology used for sequencing (Illumina, IonTorrent, PacBio, Solid, etc.), read sequences will be generated containing hundreds if

not thousands of DNA sequence fragment reads, usually in the form of FASTQ files. In the context of resequencing (sequencing an organism with a known genome, as opposed to de-novo sequencing) the read sequences will be aligned to a reference genome and a call made as to whether a variant has been identified or not.

Standard alignment pipeline

The standard pipeline for resequencing involves:

- Quality control (using FastQC (78))
- Alignment to the reference genome (using BWA (79), bowtie2 (80), etc.)
- Removing/Flagging duplicates reads (especially if PRC is used during library preparation)
- Optional Base Quality Score Recalibration (BQSR using GATK)

The above pipeline generates a Sequence Alignment Map (SAM) file, its binary counterpart BAM file, or its compressed counterpart CRAM file. These represent the standard sequence files used in NGS analysis.

Using those alignment files, researchers can do an array of different analysis. The most obvious and widespread use is to compare the alignment file generated to a reference genome to detect single nucleotide variant/ polymorphism (SNV or SNP) and short insertions and deletions (Indels).

Variant calling

The detection of SNVs and indels is usually performed by variant calling software (HaplotypeCaller (81), Freebayes (82), etc.). These software packages can use a reference genome and compare it to the alignment file to detect SNV insertions and deletions. Most of the popular variant-calling software have measures to minimise false positives and will generate a standardised file called a Variant Calling Format (VCF) file, which contains all the necessary information, allowing the storage of information for multiple samples. VCF files will always contain the following information; chromosome number, the start and end location of the variant, the reference and alternative alleles as well as extra information such as genotype (heterozygous or homozygous for this allele) per sample.

Variant annotation

Variant calling allows researchers to identify variants, but the consequence or pathogenicity of those variants are not always obvious. Variant annotation software such as ANNOVAR (83), VEP (84) or snpEff (85), can annotate VCF files to associate a variant allele and location to several key pieces of information such as: exonic/coding region location, *in-silico* prediction or frequency in the general population.

These annotations are useful so that a prediction of the impact of the identified variant can be made.

Variant pathogenicity prediction and prioritisation

Even with the proper annotation, it is still a delicate exercise to predict the effect of a given variant on an individual. To further help in the classification of genetic variants, the American College of Medical Genetics and the Association for Molecular Pathology (ACMG/AMP) released in 2015 a set of criteria to classify variants into 5 categories (Pathogenic, Probably Pathogenic, Unknown Significance, Probably Benign and Benign, often referred to as "classes" 5 to 1) (86). These criteria are divided in categories, Pathogenic Very Strong, Pathogenic Strong, Pathogenic Moderate and Pathogenic Supporting (PVS, PS, PM and PP) as well as Benign Strong and Benign Supporting (BS, BP) (see Table 6). If a variant fulfils enough pathogenic criteria, it will be classified as pathogenic. In a similar manner, a variant fulfilling enough benign criteria will be classified as benign (see tables 7). The ACMG/AMP criteria gained in popularity as they are relatively simple to compute and are relatively reliable. Software to assign the ACMG/AMP criteria have rapidly been developed (such as CharGer (87) or InterVar (88)).

Table 6. Criteria for pathogenicity prediction developed by the ACMG/AMP

Category	Short Name	Requirement
Pathogenic Very Strong	PVS1	Null variant (nonsense, frameshift, canonical +/−1 or 2 splice sites, initiation codon, single or multi-exon deletion) in a gene where loss of function (LOF) is a known mechanism of disease
Pathogenic Strong	PS1	Same amino acid change as a previously established pathogenic variant regardless of nucleotide change
	PS2	*De novo* (both maternity and paternity confirmed) in a patient with the disease and no family history
	PS3	Well-established *in vitro* or *in vivo* functional studies supportive of a damaging effect on the gene or gene product
	PS4	The prevalence of the variant in affected individuals is significantly increased compared to the prevalence in controls
Pathogenic Moderate	PM1	Located in a mutational hot spot and/or critical and well-established functional domain (*e.g*. active site of an enzyme) without benign variation
	PM2	Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes or ExAC
	PM3	For recessive disorders, detected in *trans* with a pathogenic variant

	PM4	Protein length changes due to in-frame deletions/insertions in a non-repeat region or stop-loss variants
	PM5	Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before
	PM6	Assumed *de novo*, but without confirmation of paternity and maternity
Pathogenic Supporting	PP1	Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease
	PP2	Missense variant in a gene that has a low rate of benign missense variation and where missense variants are a common mechanism of disease
	PP3	Multiple lines of computational evidence support a deleterious effect on the gene or gene product
	PP4	Patient's phenotype or family history is highly specific for a disease with a single genetic aetiology
	PP5	Reputable source recently reports variant as pathogenic but the evidence is not available to the laboratory to perform an independent evaluation

Category	**Short Name**	**Requirement**
Benign Stand-Alone	BA1	Allele frequency is above 5% in Exome Sequencing Project, 1000 Genomes,or ExAC
Benign Strong	BS1	Allele frequency is greater than expected for disorder
	BS2	Observed in a healthy adult individual for a recessive (homozygous), dominant (heterozygous), or X-linked (hemizygous) disorder with full penetrance expected at an early age
	BS3	Well-established *in vitro* or *in vivo* functional studies shows no damaging effect on protein function or splicing
	BS4	Lack of segregation in affected members of a family
Benign Moderate	BP1	Missense variant in a gene for which primarily truncating variants are known to cause disease

	BP2	Observed in *trans* with a pathogenic variant for a fully penetrant dominant gene/disorder; or observed in *cis* with a pathogenic variant in any inheritance pattern
	BP3	In-frame deletions/insertions in a repetitive region without a known function
	BP4	Multiple lines of computational evidence suggest no impact on gene or gene product
	BP5	Variant found in a case with an alternate molecular basis for disease
	BP6	Reputable source recently reports variant as benign but the evidence is not available to the laboratory to perform an independent evaluation
	BP7	A synonymous (silent) variant for which splicing prediction algorithms predict no impact to the splice consensus sequence nor the creation of a new splice site AND the nucleotide is not highly conserved

Table 7. Pathogenicity assignment for the ACMG/AMP criteria

Pathogenic (Class 5)

- 1 Very Strong (PVS1) *AND*
  - ≥1 Strong (PS1–PS4) *OR*
  - ≥2 Moderate (PM1–PM6) *OR*
  - 1 Moderate (PM1–PM6) and 1 Supporting (PP1–PP5) *OR*
  - ≥2 Supporting (PP1–PP5)
- ≥2 Strong (PS1–PS4) *OR*
- 1 Strong (PS1–PS4) *AND*
  - ≥3 Moderate (PM1–PM6) *OR*
  - 2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR*
  - 1 Moderate (PM1–PM6) *AND* ≥4 Supporting (PP1–PP5)

Likely Pathogenic (Class 4)

- 1 Very Strong (PVS1) *AND* 1 Moderate (PM1–PM6) *OR*
- 1 Strong (PS1–PS4) *AND* 1–2 Moderate (PM1–PM6) *OR*
- 1 Strong (PS1–PS4) *AND* ≥2 Supporting (PP1–PP5) *OR*
- ≥3 Moderate (PM1–PM6) *OR*
- 2 Moderate (PM1–PM6) *AND* ≥2 Supporting (PP1–PP5) *OR*
- 1 Moderate (PM1–PM6) *AND* ≥4 Supporting (PP1–PP5)

Likely Benign (Class 2)

- 1 Strong (BS1–BS4) and 1 Supporting (BP1–BP7) *OR*
- ≥2 Supporting (BP1–BP7)

Benign (Class 1)

- 1 Stand-Alone (BA1) *OR*
- ≥2 Strong (BS1–BS4)

It is important to keep in mind that the ACMG/AMP criteria are designed to determine the probability of a variant being pathogenic (or benign). This means that a class 5 variant (Pathogenic) is not more pathogenic than a class 4 variant (Probably Pathogenic) but has a higher probability to be pathogenic, due to the presence of more evidence of pathogenicity.

However, when studying the genetic basis of cancer (or other Mendelian disorders) and working with large datasets such as WGS or WES, a large portion of variants remain classified as class 3 variants (i.e. variants of unknown significance (VUS)) even if they nearly fulfil the requirements to be classified as Likely Pathogenic or Pathogenic. This is one of the limitations of current pathogenicity prediction software using the ACMG/AMP criteria: if there is no knowledge about the identified variant (like its frequency in the general population, clinical data, frequency in a diseased cohort), it will most likely be classified as a VUS (the ACMG/AMP criteria PS1, PS3, PM5, PP3 and PP5 require prior knowledge about the variant). Another limitation of this software is that it only analyses variants individually and does not take into account the fact that they might belong to a particular cohort sharing the same haplotype (frequently used to study the underlying genetic cause of Mendelian disease). They will, for example, not notice if a particular gene is frequently mutated or that a variant is greatly enriched in a specific cohort compared to a control population.

1.5) Rationale and hypothesis

As mentioned above, familial CRC syndromes (with no known genetic cause but with strong family history) remain highly prevalent among all CRC diagnoses (around 10-13%). Like most inherited diseases, DNA defects are often the cause of an elevated risk of disease.

I therefore hypothesise that there is a set of yet unidentified genes that increase the risk of inherited CRC.

By identifying a set of genes and variants that could increase the risk of CRC, individuals with a strong family history of CRC could be enrolled into screening programs that have proven benefit to reduce mortality and morbidity of these cancer syndromes (89).

Early detection is important for improved survival (90). In the case of colorectal cancer, individuals have 97.7% one-year survival if detection of disease occurs at stage 1 versus only a 43.9% if detection occurs at stage 4 disease.

In the case of familial CRC, it is important to identify genes and pathogenic variants giving an increased risk of cancer. Individuals with an elevated risk of CRC can then be offered more frequent monitoring, leading to the early detection of potential malignant lesions and an appropriate surgical response, increasing the chances of survival.

1.6) Aims and approach

To identify a set of genes increasing the risk of CRC in familial CRC syndrome patients, we first need to select appropriate individuals. We first distinguished non-polyposis and polyposis familial syndromes as two different entities that need to be studied separately.

In this thesis I will focus both on non-polyposis and polyposis CRC syndromes. For both, we selected a cohort made of individuals that were diagnosed with CRC, but that did not carry a pathogenic variant in a known causative gene for the disease. All individuals also had a strong family history of CRC, suggesting an underlying genetic cause.

Using NGS, we can identify potentially pathogenic variants that were overlooked in these patients. The precise identification of genetic pathogenic variations is key to discover novel genes involved in CRC development. In addition to pathogenic variants and their genes, we can study the copy number variations occurring in these patients, the pathways involved to better understand the underlying mechanisms of the disease and also the genetic predisposition to known CRC risk factors. Furthermore, I aimed to develop new and innovative ways to analyse exome sequencing data in cohorts.

The current research aims to:

I) Investigate the presence of pathogenic variants in all MMR genes associated with Lynch-Like Syndromes using a targeted NGS approach

II) Propose a new and more refined pipeline for pathogenicity prediction in Whole-Exome Sequencing

III) Using the findings from Aim II, evaluate the genetic basis of familial polyposis syndromes

CHAPTER 2:

The MMR pathway in Lynch-Like Syndrome

CHAPTER 2: The MMR pathway in Lynch-Like Syndrome

2.0) Introduction

Lynch-Like Syndromes (LLS) is an umbrella term used to describe all familial non-polyposis CRC emerging from individuals which fulfil the ACII (suggestive of a genetic cause for the disease) but with the confirmed absence of a pathogenic variant in the routinely screened MMR genes.

Aims

The aim of this study was to investigate the presence of potentially pathogenic variants in the currently clinically unscreened MMR genes. LS molecular diagnosis is performed by identifying pathogenic variants in four MMR genes: *MLH1*, *MSH2*, *MSH6* and *PMS2* (with the addition of *EPCAM* deletions that result in MSH2 epigenetic silencing). These genes all express proteins that are involved in post-replication mismatch detection. However, more than 22 genes are involved in the MMR pathway. The presence of pathogenic variants in any of these genes could reduce the efficiency of the MMR pathway. Deficiencies in the MMR pathway, as described in Chapter 1, is observed as MSI and an increased risk of cancer in LS patients.

Identifying pathogenic variants in the remaining key 18 unscreened MMR genes (*MSH3, PMS1, MLH3, EXO1, POLD1, POLD3*, *RFC1, RFC2, RFC3, RFC4, RFC5, PCNA, LIG1, RPA1, RPA2, RPA3, POLD2* and *POLD4)* and re-sequencing the 4 key genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*) using new technology, could reveal novel genes/variants involved in LLS.

There is Increasing evidence to show that the LS definition could be extended to other MMR genes. *MLH3* (91) and *MSH3* (92) are prime examples of new MMR genes linked to CRC.

Approach

The 22 genes involved in the MMR pathway, code for either full proteins or the subunits of larger complexes.

To assess the involvement of the 18 unscreened genes in cancer development and the re-sequencing of the 4 Sanger screened genes with a new technology, we selected a cohort of 274 patients (of Norwegian and Australian origin) diagnosed with CRC or other LS-related cancers, all of whom fulfilled the AC I or II criteria and on screening by Sanger sequencing were deemed to be mutation-negative in at least one of the four known MMR genes as indicate by immunohistochemical assessment.

DNA samples from the 274-constituting cohort were subsequently sequenced using a custom Haloplex design and all variants identified, confirmed using Sanger sequencing. Each variant was annotated to assess its predicted pathogenicity.

This cohort was first analysed for 112 genes known to be involved in CRC (93) (see appendix 1). Even though many cases could be explained by pathogenic variants in key CRC genes, many individuals (75%), did not have any identifiable pathogenic variant that could explain their elevated risk of CRC.

The 22 genes involved in MMR (included in the panel mentioned above but not all analysed as part of that study) were then analysed to identify pathogenic variants. The following study will help better understand the role and the extent of the involvement of the currently unscreened MMR genes in LLS.

2.1)Publication

STATEMENT:

This is a co-author statement attesting to the candidate's contribution to the publication listed below:

I attest that Research Higher Degree candidate Alexandre Xavier contributed to the publication listed below by performing the analysis, the sequencing validation, and writing and managing the manuscript.

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;7(8):e850.

This statement explains the contribution of all authors in the article listed above:

Table illustrating author contribution percentage and description of contribution to the article listed above.

Author	Contribution (%)	Description of contribution to article	Signature	Date
Alexandre Xavier	55%	Performed analysis, sequencing validation and manuscript writing		15/10/2019
Maren Fridtjofsen Olsen	5%	Performed laboratory tests		22/10/2019
Liss A. Lavik, Jostein Johansen, Ashish Kumar Singh	5%	Supplied samples and clinical information. Data analysis interpretation.	Liss Anne Lavik Digitalt signert av Liss Anne Lavik Dato: 2019.10.21 09:47:22 +02'00' Jostein Johansen Digital: signert av Jostein Johansen Dato: 2019.10.21 13:17:34 +02'00' Ashish Kumar Singh Digitally signed by Ashish Kumar Singh Date: 2019.10.21 13:25:53 +02'00'	21/10/2019 21/10/2019 21/10/2019
Wenche Sjursen	10%	Supplied samples and clinical information. Critical revision of the manuscript and study supervision		28.10.2019
Rodney J. Scott	5%	Supplied samples and clinical information. Critical revision of the manuscript and study supervision		15/10/2019
Bente A. Talseth-Palmer	20%	Study design, obtained funding, performed data analysis and interpretation, manuscript writing, critical revision of the manuscript and supervision	Bente Talseth-Palmer Digitally signed by Bente Talseth-Palmer Date: 2019.10.21 22:10:57 +11'00'	21/10/2019

Alexandre Xavier

Date: 15/10/2019

Digitally signed by Dr Lesley MacDonald-Wicks
Date: 2019.11.20 15:40:29 +11'00'

Dr Lesley MacDonald-Wicks

Assistant Dean (Research Training)

Date: 20/11/19

Received: 14 March 2019 | Revised: 18 June 2019 | Accepted: 24 June 2019

DOI: 10.1002/mgg3.850

ORIGINAL ARTICLE

Molecular Genetics & Genomic Medicine Open Access WILEY

Comprehensive mismatch repair gene panel identifies variants in patients with Lynch-like syndrome

Alexandre Xavier[1] | **Maren Fridtjofsen Olsen**[2,3] | **Liss A. Lavik**[3] | **Jostein Johansen**[2] | **Ashish Kumar Singh**[3] | **Wenche Sjursen**[2,3] | **Rodney J. Scott**[1,4] | **Bente A. Talseth-Palmer**[1,5]

[1]University of Newcastle Hunter Medical Research Institute, New Lambton Heights, New South Wales, Australia

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

[3]Department of Medical Genetics, Saint Olavs Hospital University Hospital, Trondheim, Norway

[4]Pathology North, Hunter New England Health, Newcastle, New South Wales, Australia

[5]Møre and Romsdal Hospital Trust, Clinic Research and Development, Molde, Norway

Correspondence
Bente A. Talseth-Palmer, University of Newcastle Hunter Medical Research Institute, Lot 1, Kookaburra Circuit, New Lambton Heights, New South Wales, Australia.
Email: talseth-palmer@hotmail.com

Funding information
Central Norway Regional Health Authority; Norwegian University of Science and Technology; Møre and Romsdal Hospital Trust; Hunter Cancer Research Alliance; Cancer Institute NSW

Abstract

Background: Lynch-like syndrome (LLS) represents around 50% of the patients fulfilling the Amsterdam Criteria II/revised Bethesda Guidelines, characterized by a strong family history of Lynch Syndrome (LS) associated cancer, where a causative variant was not identified during genetic testing for LS.

Methods: Using data extracted from a larger gene panel, we have analyzed next-generation sequencing data from 22 mismatch repair (MMR) genes (*MSH3*, *PMS1*, *MLH3*, *EXO1*, *POLD1*, *POLD3 RFC1*, *RFC2*, *RFC3*, *RFC4*, *RFC5*, *PCNA*, *LIG1*, *RPA1*, *RPA2*, *RPA3*, *POLD2*, *POLD4*, *MLH1*, *MSH2*, *MSH6*, and *PMS2*) in 274 LLS patients. Detected variants were annotated and filtered using ANNOVAR and FILTUS software.

Results: Thirteen variants were revealed in *MLH1*, *MSH2*, and *MSH6*, all genes previously linked to LS. Five additional genes (*EXO1*, *POLD1*, *RFC1*, *RPA1*, *and MLH3*) were found to harbor 11 variants of unknown significance in our sample cohort, two of them being frameshift variants.

Conclusion: We have shown that other genes associated with the process of DNA MMR have a high probability of being associated with LLS families. These findings indicate that the spectrum of genes that should be tested when considering an entity like Lynch-like syndrome should be expanded so that a more inclusive definition of this entity can be developed.

KEYWORDS
Genetics, germline mutation, high-throughput sequencing, Lynch syndrome, MMR gene panel

1 | INTRODUCTION

Lynch Syndrome (LS) is an autosomal dominantly inherited predisposition to colorectal cancer (CRC) and other epithelial malignancies and accounts for approximately 2%–3% of all CRC patients diagnosed annually (de la Chapelle, 2004; Hampel et al., 2005). LS, also known as hereditary nonpolyposis colorectal cancer (HNPCC), is defined by the presence of germline mutations in one of four genes involved in DNA mismatch repair (MMR); *MLH1* (OMIM: 120436), *MSH2* (OMIM: 609309), *MSH6* (OMIM: 600678), and *PMS2* (OMIM: 600259) (Lynch & de la Chapelle, 1999). Deletions in *EPCAM* are also implicated in LS that are associated with epigenetic silencing of *MSH2* (Kuiper et al., 2011). It is

important to identify LS mutation carriers, in order to offer them regular surveillance programs like colonoscopy to improve early detection of cancer.

LS tumor spectrum is quite wide, involving in most cases CRC and, for women, a high risk of endometrial cancer (Lek et al., 2016). In addition, MMR mutations in LS causes an increased risk of ovarian, gastric, urologic tract, kidney, ureter, small bowel, and hepatobiliary tract tumors (Samadder et al., 2017; Watson & Riley, 2005).

LS is an entity that has been diagnosed using the Amsterdam criteria (AC) and Bethesda guidelines (BG), or variants of it (the AC II or revised BG) using the patient's pedigree and family history of cancer (Rodriguez-Bigas et al., 1997; Umar et al., 2004; Vasen, Mecklin, Khan, & Lynch, 1991; Vasen, Watson, Mecklin, & Lynch, 1999). Genetic screening by Sanger sequencing to identify causative variants in MMR genes has been the gold standard to diagnose patients at risk of LS. Individuals fulfilling the AC II or revised BG without a molecular diagnosis are now termed Lynch-like syndrome (LLS) families (Carethers, 2014; Giardiello et al., 2014).

Families that fulfill the AC, where probands have tumors displaying microsatellite instability (MSI) or a loss of MMR genes expression (as judged by immunohistochemistry), are offered screening for pathogenic variants in MMR genes (usually by DNA sequencing) to identify a causative genetic variant. Using this approach, approximately 50% of LS patients remain without a molecular diagnosis after screening the common MMR genes (Bonis et al., 2007; Lindor et al., 2005; Steinke et al., 2014). Early detection and management provide the best likelihood of survival, thus identifying high-risk individuals who could benefit from early detection is a priority. The 50% of patients where pathogenic variants cannot be detected are commonly termed LLS families or familial colorectal cancer type X (FCCTX) as disease segregation is suggestive of an inherited disease but in the absence of any identifiable causative variant. This group appears to have a later age of disease onset compared to LS, suggesting that these families have lower levels of disease penetrance (Lipkin & Afrasiabi, 2007). While the definitions of LLS and FCCTX mostly overlap, LLS is defined by patients with MSI-High tumors but no loss of MMR immunohistochemistry staining (for the four main MMR genes) (Carethers, 2014). On the other hand, FCCTX describes patients fulfilling the AC I but no causative pathogenic variants has been found, and are mostly microsatellite-stable (MSS) (Lipkin & Afrasiabi, 2007).

DNA MMR involves the coordinated response of at least 22 proteins (KEGG pathways (Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2017), Figure 1) that are involved in mismatch recognition, protein recruitment to the lesion, removal of the mismatch and replacement of the incorrect base with the correct one (Fishel, 2015). Thus, the possibility exists that other defects in the DNA MMR pathway may be associated with cancer risk, which manifests as an entity similar to LS. Evidence to support this comes from studies that have examined *MLH3* where it has been proposed to be a candidate gene implicated in LS (Liu et al., 2003). This is supported by evidence that demonstrates *MSH3* variants appear to confer a low risk of disease (associated with various phenotypes (Carethers, Koi, & Tseng-Rogenski, 2015) including a phenotype of polyposis (Adam et al., 2016)) and have a synergistic effect when accompanied by *MSH2* variants (Duraturo et al., 2011). Previously, *POLD1* variants have shown to be associated with an increased risk of CRC that results in a phenotypic disease spectrum, which includes phenotypes, observed in both LS and a polyposis (Buchanan et al., 2017).

To assess the involvement of other MMR genes in LS, we investigated the presence of potentially pathogenic variants in 22 MMR genes (*MSH3*, *PMS1*, *MLH3* (OMIM: 604395), *EXO1* (OMIM: 606063), *POLD1* (OMIM: 174761), *POLD3*, *RFC1* (OMIM: 102579), *RFC2*, *RFC3*, *RFC4*, *RFC5*, *PCNA*, *LIG1*, *RPA1* (OMIM: 179835), *RPA2*, *RPA3*, *POLD2*, *POLD4*, *MLH1*, *MSH2*, *MSH6*, and *PMS2*), using next-generation sequencing (NGS) in patients with a clinical diagnosis of LS. All of whom fulfilled the ACII or the revised BG but lacked a causative variant for the standard MMR gene(s) after genetic testing.

2 | METHODS

2.1 | Ethics approval

Ethics approval was obtained from the Hunter New England Human Research Ethics Committee (04/03/10/3.11) and the University of Newcastle Human Research Ethics Committee (H-2008–0337).

2.2 | Samples

This study used DNA obtained from 82 Norwegian and 192 Australian LLS patients ($n = 274$, see Table 1) previously described (Hansen et al., 2017). In brief, all patients fulfilled the AC II criteria or revised BG and had no pathogenic variant detected during routine genetic screening for the MMR gene(s) tested (*MLH1*, *MSH2*, *MSH6*, and/or *PMS2*). All patients were previously screened for one or more MMR genes as per their practician recommendations.

The sample cohort consisted of unrelated (Australian) and unrelated/related (Norwegian) individuals; eight families with two to three individuals per family were present in the Norwegian cohort (Hansen et al., 2017).

DNA samples from all patients were sequenced as part of the health-care system and all patients have given written informed consent for their samples to be used for research. Ethics approval was obtained from relevant committees.

TABLE 1 Cohort characteristics and screening results for the 274 samples included in the current study

	Total Cohort (*N = 274*)
Nationality	
Norwegian	82
Australian	192
Female	183
Male	91
Median age at first cancer[a]	51.5 [21–86]
Cancer history[b]	
CRC	229
Other cancers	28
Only adenomas	14
Multiple primary cancers	64
Amsterdam Criteria II	
Positive	262
Negative[c]	12
Microsatellite instability status[d]	
MSS	38
MSI-L	6
MSI-H	27
IHC[e]	
Loss of MMR protein staining	83
Normal staining	56

[a]Data missing for six patients.
[b]Data missing for three patients.
[c]Revised Bethesda Guidelines (BG) positive.
[d]Only available for the Norwegian patients. Data missing for 203 patients.
[e]Data available for 68 Norwegian and 71 Australian samples. Data missing for 135 patients.

2.3 | Gene panel sequencing

Sequencing data were generated (See Figure S1) from a 124 multigene panel study described in (Hansen et al., 2017), which contained 22 MMR genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*, *MSH3*, *PMS1*, *MLH3*, *EXO1*, *POLD1*, *POLD3*, *RFC1*, *RFC2*, *RFC3*, *RFC4*, *FRC5*, *PCNA*, *LIG1*, *RPA1*, *RPA2*, *RPA3*, *POLD2*, *and POLD4*).

A custom Haloplex design (Agilent Technologies, Santa Clara, CA) was used for library preparation. Description of both the Haloplex design and the sequencing protocols (HiSeq 2500 and NextSeq, Illumina) have been reported previously (Hansen et al., 2017).

2.4 | Data analysis

The previous (Hansen et al., 2017) study analyzed only 10 MMR genes, (*MLH1*, *MSH2*, *MSH6*, *PMS2*, MSH3, *PMS*1, *MLH3*, *EXO1*, *POLD1*, and *POLD3*). In this current study, we included data from those genes as well as data from the remaining 12 MMR genes (*RFC1*, *RFC2*, *RFC3*, *RFC4*, *FRC5*, *PCNA*, *LIG1*, *RPA1*, *RPA2*, *RPA3*, *POLD2*, and *POLD4*) to create a complete MMR gene panel (See Figure S1 for a full flowchart of the study's design).

Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009) was utilized to align the paired end reads to the human genome (hg19, UCSC assembly, February 2009). BAM files were converted with SAMtools (Li et al., 2009). Variant calling was performed according to GATK Best Practice recommendations using GATK version 3.1 (McKenna et al., 2010) including local realignment around insertion/deletions (indels) and recalibration of quality scores. The variant caller HaplotypeCaller was utilized. Quality control of the called variants was performed using GATK variant filtration with parameter settings according to the recommendations in SEQanswers exome sequencing analysis guide (Van der Auwera et al., 2013). In short, variant quality score recalibration (VQSR) was applied using the recommended set of known variants for both indels and SNP. The tranche threshold of 99.0 was used to select variants. ANNOVAR (Wang, Li, & Hakonarson, 2010) was used to annotate detected variants and filtering of variants was done using the filtering tool FILTUS version 1.0.4 (Vigeland, Gjotterud, & Selmer, 2016).

2.5 | Filtering of variants

FILTUS, a desktop software for fast and efficient detection of disease-causing variants was used on the annotated files (Vigeland et al., 2016). The 10 MMR genes belonging to the gene panel previously analyzed by Hansen et al. (2017 were also included in the current study due to different filtering strategies and for comparative purposes.

Variants were filtered in the 22 MMR genes individually (gene lookup in FILTUS), with function collapse = lists all samples that have same variant together and saved as individual files on gene name, before being combined into one file. Variants with a frequency of more than 0.05 in public databases (ExaC or gNomad (Lek et al., 2016)) were first excluded. Then variants were excluded if detected in more than five unrelated individuals in our cohort (not likely to be pathogenic due to their high frequency) and intronic variants were ignored if they had no variant prediction. Nonsynonymous single-nucleotide variants (SNVs) and indel variants were included in further filtering.

After in silico filtering, using FILTUS, we performed some manual filtering and variant interpretation to remove artifacts and only selecting variants most likely to be causative. We checked detected variants against results reported by Hansen et al. (2017. Further, variant interpretation was performed utilizing Alamut software (Interactive Biosoftware, Rouen, France) and evaluating available literature.

TABLE 1 Cohort characteristics and screening results for the 274 samples included in the current study

	Total Cohort (*N = 274*)
Nationality	
Norwegian	82
Australian	192
Female	183
Male	91
Median age at first cancer[a]	51.5 [21–86]
Cancer history[b]	
CRC	229
Other cancers	28
Only adenomas	14
Multiple primary cancers	64
Amsterdam Criteria II	
Positive	262
Negative[c]	12
Microsatellite instability status[d]	
MSS	38
MSI-L	6
MSI-H	27
IHC[e]	
Loss of MMR protein staining	83
Normal staining	56

[a]Data missing for six patients.
[b]Data missing for three patients.
[c]Revised Bethesda Guidelines (BG) positive.
[d]Only available for the Norwegian patients. Data missing for 203 patients.
[e]Data available for 68 Norwegian and 71 Australian samples. Data missing for 135 patients.

2.3 | Gene panel sequencing

Sequencing data were generated (See Figure S1) from a 124 multigene panel study described in (Hansen et al., 2017), which contained 22 MMR genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*, *MSH3*, *PMS1*, *MLH3*, *EXO1*, *POLD1*, *POLD3*, *RFC1*, *RFC2*, *RFC3*, *RFC4*, *FRC5*, *PCNA*, *LIG1*, *RPA1*, *RPA2*, *RPA3*, *POLD2*, *and POLD4*).

A custom Haloplex design (Agilent Technologies, Santa Clara, CA) was used for library preparation. Description of both the Haloplex design and the sequencing protocols (HiSeq 2500 and NextSeq, Illumina) have been reported previously (Hansen et al., 2017).

2.4 | Data analysis

The previous (Hansen et al., 2017) study analyzed only 10 MMR genes, (*MLH1*, *MSH2*, *MSH6*, *PMS2*, MSH3, *PMS*1, *MLH3*, *EXO1*, *POLD1*, and *POLD3*). In this current study, we included data from those genes as well as data from the remaining 12 MMR genes (*RFC1*, *RFC2*, *RFC3*, *RFC4*, *FRC5*, *PCNA*, *LIG1*, *RPA1*, *RPA2*, *RPA3*, *POLD2*, and *POLD4*) to create a complete MMR gene panel (See Figure S1 for a full flowchart of the study's design).

Burrows-Wheeler Aligner (BWA) (Li & Durbin, 2009) was utilized to align the paired end reads to the human genome (hg19, UCSC assembly, February 2009). BAM files were converted with SAMtools (Li et al., 2009). Variant calling was performed according to GATK Best Practice recommendations using GATK version 3.1 (McKenna et al., 2010) including local realignment around insertion/deletions (indels) and recalibration of quality scores. The variant caller HaplotypeCaller was utilized. Quality control of the called variants was performed using GATK variant filtration with parameter settings according to the recommendations in SEQanswers exome sequencing analysis guide (Van der Auwera et al., 2013). In short, variant quality score recalibration (VQSR) was applied using the recommended set of known variants for both indels and SNP. The tranche threshold of 99.0 was used to select variants. ANNOVAR (Wang, Li, & Hakonarson, 2010) was used to annotate detected variants and filtering of variants was done using the filtering tool FILTUS version 1.0.4 (Vigeland, Gjotterud, & Selmer, 2016).

2.5 | Filtering of variants

FILTUS, a desktop software for fast and efficient detection of disease-causing variants was used on the annotated files (Vigeland et al., 2016). The 10 MMR genes belonging to the gene panel previously analyzed by Hansen et al. (2017 were also included in the current study due to different filtering strategies and for comparative purposes.

Variants were filtered in the 22 MMR genes individually (gene lookup in FILTUS), with function collapse = lists all samples that have same variant together and saved as individual files on gene name, before being combined into one file. Variants with a frequency of more than 0.05 in public databases (ExaC or gNomad (Lek et al., 2016)) were first excluded. Then variants were excluded if detected in more than five unrelated individuals in our cohort (not likely to be pathogenic due to their high frequency) and intronic variants were ignored if they had no variant prediction. Nonsynonymous single-nucleotide variants (SNVs) and indel variants were included in further filtering.

After in silico filtering, using FILTUS, we performed some manual filtering and variant interpretation to remove artifacts and only selecting variants most likely to be causative. We checked detected variants against results reported by Hansen et al. (2017. Further, variant interpretation was performed utilizing Alamut software (Interactive Biosoftware, Rouen, France) and evaluating available literature.

TABLE 2 List of variants identified in LS screened MMR genes (*MLH1*, *MSH2*, and *MSH6*) and other MMR genes (*EXO1*, *POLD1*, *RFC1*, and *RPA1*)

Gene	Hansen et al. (Hansen et al., 2017)	Reference sequence	DNA change	AA change	Domain	Rs ID[a]	gnomad AF[a]	Classification	Type	Sample ID
MLH1	[b]	NM_000249.3	c.514G > A	p.Glu172Lys		NR	NR	VUS	Missense	250
MLH1		NM_001167619.2	c.1130A > C	Lys377Thr		rs63750449	0.004564	VUS	Missense	187 and 214
MLH1	[b]	NM_000249.3	c.2103 + 1G>T	N/A		rs267607888	N/A	Pathogenic	Missense	116
MLH1		NM_000249.3	c.1039–31_1039–29delATA	N/A		rs778381149	0.0003969	VUS	Intronic	295
MSH2	[b]	NM_000251.2	c.138C > G	p.His46Gln		rs33946261	0.0003619	Likely Pathogenic	Missense	9
MSH2	[b]	NM_000251.2	c.1045C > G	p.Pro349Ala		rs267607939	0.00009148	Likely Pathogenic	Missense	281
MSH6		NM_000179.2	c.431G > T	p.Ser144Ile	DNA binding PWWP domain	rs3211299	0.001187	Benign	Missense	201
MSH6		NM_001281492.1	c.892A > G	p.Lys298Glu		rs761822293	0.000003979	VUS	Missense	169
MSH6		NM_001281492.1	c.1054C > T	p.Arg352X		rs63750909	0.00003186	Pathogenic	Nonsense	133
MSH6		NM_001281492.1	c.1118C > G	p.Ser373Cys		rs63750897	0.001165	Likely Benign	Missense	225
MSH6	[b]	NM_000179.2	c.1282A > G	p.Lys428Glu		rs761822293	0.000003979	VUS	Missense	169
MSH6	[b]	NM_000179.2	c.2079dup	p.Cys694Metfs*4		rs267608083	NR	Pathogenic	Frameshift	183
MSH6	[b]	NM_000179.2	c.3261 dup	p.Phe1088Leufs*5		rs748452299	0.0018	Pathogenic	Frameshift	41
EXO1		NM_003686.4	c.1928T > A	p.Leu643X	MLH1 and MSH2 interaction domain removed	NR	N/A	VUS	Nonsense	165
EXO1		NM_003686.4	c.2009A > G	p.Glu670Gly	MSH2 interaction domain	rs1776148	0.78	Benign	Missense	166
EXO1		NM_006027 0.4	c.2485G > T	p.Glu829X	MLH1 interaction domain removed	rs757677420	0.00000292	VUS	Nonsense	154
POLD1		NM_001256849.1	c.1249A > G	p.Thr417Ala	DNA-directed DNA polymerase, family B exonuclease domain	NR	N/A	VUS	Missense	226
POLD1		NM_001256849.1	c.1558insG	p.	DNA-directed DNA polymerase, family B, multifunctional domain	NR	N/A	VUS	Frameshift	109
POLD1		NM_001256849.1	c. 2510G > C	p.Gly811Ala	DNA-directed DNA polymerase, family B multifunctional domain	NR	N/A	VUS	Missense	277

(Continues)

TABLE 2 (Continued)

Gene	Hansen et al. (Hansen et al., 2017)	Reference sequence	DNA change	AA change	Domain	Rs ID[a]	gnomad AF[a]	Classification	Type	Sample ID
RFC1		NM_001204747.1	c.2017G > A	p.Val673Met	ATPase domain	rs28903096	0.0006994	VUS	Missense	174
RFC1		NM_001204747.1	c.2276A > G	p.Lys759Arg	ATPase domain	NR	N/A	VUS	Missense	227
RPA1		NM_002945.4	c.856G > T	p.Val286Phe		rs55800538	0.002942	VUS	Missense	161
RPA1		NM_002945.4	c.1160G > A	p.Gly387Asp	OB domain	NR	N/A	VUS	Missense	239
RPA1		NM_002945.4	c.1165C > T	p.Arg389W	OB domain	rs202068855	0.0005468	VUS	Missense	221
MLH3	[b]	NM_001040108.1	c.885del	p.His296Thrfs*12		NR	N/A	VUS	Frameshift	14

Note: Reported here are the gene symbol, Refseq reference sequence, DNA and amino acid (AA) change, protein domain, the rs ID of the variant if known, then gnomad allele frequency and the classification according to the ACMG 2015 guidelines (Richards et al., 2015).

[a]Not reported (NR) - previously unreported variant.

[b]Variant previously identified in Hansen et al. (2017).

oligosaccharide-binding domain (OB) domain. The OB domain allows RPA1 protein to bind ssDNA in a nonsequence-specific manner and to stabilize the single strand after the damaged strand is excised (Bochkareva, Korolev, Lees-Miller, & Bochkarev, 2002). The remaining variant c.856G > T did not appear to alter the functional domain of the protein.

4 | DISCUSSION

The presence of potentially pathogenic variants in patients diagnosed with LLS shows that there is a clear need to create an exhaustive list of pathogenic or potentially pathogenic genes for inherited CRC in order to identify individuals with a high risk of developing CRC and genes/variants appropriate for functional analysis. MMR genes are good candidates given that they are predicted to be causative in 8% of patients in the current study, a yield comparable to similar studies (Dong et al., 2018; Paulo et al., 2018). In addition, our results suggest that re-screening the four known LS genes in previously variant-negative LS patients with the more sensitive approach of NGS should be undertaken to ensure no pathogenic variants have been missed using less sensitive screening methods.

From the 274 patients enrolled in this study and not previously described we revealed 22 potentially causative variants in nine different MMR genes. Included in the study were the four known LS MMR genes (*MSH2*, *MLH1*, *MSH6*, and *PMS2*) as it is well accepted that older variant detection methods were not as sensitive as approaches that are more contemporary. The number of variants identified in *MLH1* and *MSH6* reflects the sensitivity of older screening methodologies that may not have revealed the presence of these causative variants. We detected additional variants in *MLH1* and *MSH6* compared to Hansen et al. (2017 due to less stringent filtering strategies in FILTUS.

The identification of 11 potentially pathogenic variants through in silico analysis in the extended MMR gene panel does reveal the extent to which the DNA MMR pathway might be associated with the risk of cancer development in families classified as LLS. We show here that families categorized as LLS harbor potentially pathogenic variants in other MMR genes than those already associated with LS. The genes *EXO1*, *POLD1*, *RFC1*, and *RPA1* harbor that variants that were predicted to be pathogenic. None of the variants identified in this study have previously been associated with a cancer phenotype, which is probably due to their extremely low frequency in the general population. However, disrupting the MMR pathway could be a possible cause of cancer development. It is known that *RFC1*, *RPA1*, and *POLD1* are involved in DNA damage repair mechanisms other than MMR and DNA synthesis during replication (see KEGG orthology: K10754, K02999 and K02327). *EXO1* is involved with other

DNA repair and maintenance mechanisms (Keijzers, Liu, & Rasmussen, 2016). Interestingly, the two nonsense *EXO1* variants identified in our study are predicted to affect the binding to *MSH2*/*MLH1*, suggesting that the loss of function would specifically affect *EXO1* MMR-related functions. Moreover, *RPA1* and *POLD1* have been previously described as deleterious when mutated in cancer (Nicolas, Golemis, & Arora, 2016; Wang et al., 2005). Variants in *POLD1* are described in patients presenting a polyposis phenotype termed polymerase-proofreading associated polyposis (PPAP) (Palles et al., 2013). The pathogenic variants identified in *POLD1* in the current study support the notion that the phenotypes of LS and PPAP might overlap, both with a multitumor phenotype.

RPA1 variants have previously been implicated in cancers that are associated with chromosome instability (Hass, Gakhar, & Wold, 2010; Wang et al., 2005). However, a comprehensive genetic study did not show a clear association with CRC (Jokic et al., 2011).

A meta-analysis supports a significant association between RFC1 p.G80A and plasma cell malignancies (Huang et al., 2016). Moreover, the Cancer Genome Atlas (TCGA data) (Cancer Genome Atlas Research et al., 2013) shows that somatic *RFC1* variants occur in 10.2% of uterine cancers and 5.5% of CRCs, which is consistent with tumors identified in LS. RFC1 has been previously described as a member of the BRCA1-associated genome surveillance complex (BASC) (Wang et al., 2000). This complex is involved in DNA damage and abnormal structure detection and more generally in the maintenance of genomic integrity making it a good candidate gene for LLS.

Pathogenic variants in *EXO1* were also found in the current study but its exact role remains to be determined. A previous study suggests that the gene is either associated with low-disease penetrance or influencing a polygenic risk score (Talseth-Palmer et al., 2016). Other reports indicate that even in healthy patients, *EXO1* variants are present, including those that result in a truncated protein (Jagmohan-Changur et al., 2003) and consequently loss of function. Notwithstanding, in our study, both *EXO1* c.1928T > A and c.2485G > T leads to a truncated protein. The variant c.1928T > A has lost the MLH2 and MLH1 interaction domain, whereas the c.2485G > T variant truncates only the MLH1 interaction domain. Lack of either of these domains could affect EXO1 recruitment at the site of the mismatch or DNA damage, impairing the MMR process (Goellner, Putnam, & Kolodner, 2015).

To assess the link between the variants identified in the current study and the development of CRC in LLS families, larger sample cohorts are needed with detailed analysis of the tumor phenotypes to establish if indeed many of the downstream MMR functions are associated with MSI tumors. Furthermore, detailed segregation analysis is required to determine if the variant segregates with disease. Finally, functional analysis would significantly aid in characterizing their respective pathogenic effects.

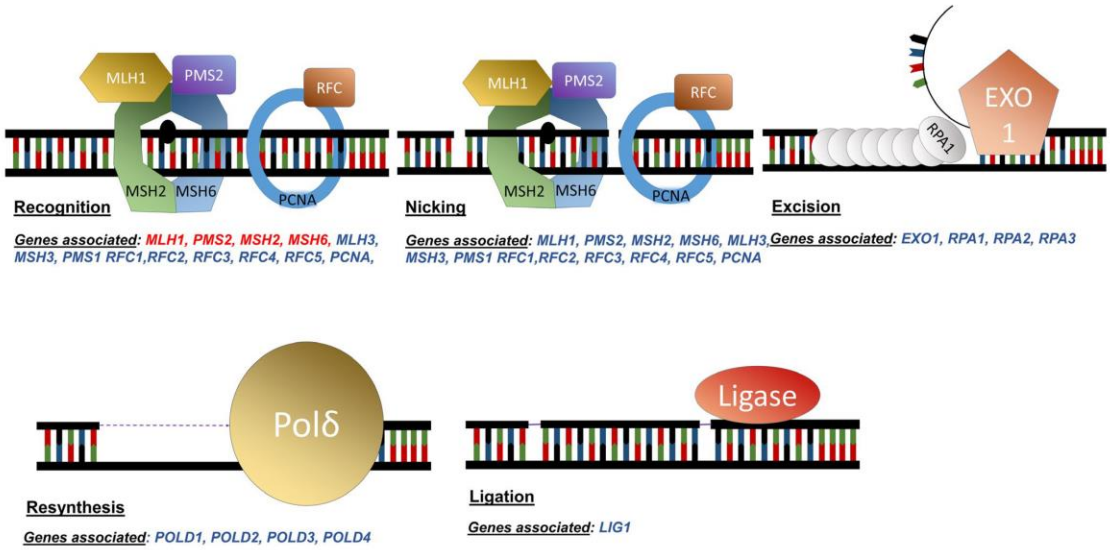

FIGURE 1 Mismatch repair pathway major steps with genes associated. Genes in red are the one usually screened for mutations in a clinical setting. *MLH1*, *PMS2*, *MSH2*, and *MSH6* are all involved in the recognition of DNA damage. PMS2 has an endonuclease function in nicking around the damaged region. EXO1 will then remove the DNA strand containing the error and RPA (Replication Protein A) will protect the remaining single strand of DNA. The DNA polymerase Polδ resynthesises the new DNA strand which is then ligated with a ligase (based on Hsieh & Yamane, 2008)

Limitations of the current study include the relatively small number of patients tested. A larger sample cohort and functional studies of the identified variants are required to confirm the results of this study. Segregation analysis would provide insights into the pathogenicity of these variants but could not be performed as part of this study.

In conclusion, we have shown that other genes associated with the process of DNA MMR have a high probability of being associated with LLS families. In addition, approximately 8% of families that fulfill the ACII or RB criteria in our sample cohort appear to be accounted for by genes involved in the MMR pathway. These findings indicate that these variants are important as they will guide future research focused on the functional impact of newly discovered variants.

ACKNOWLEDGMENTS

This work was supported by the Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU), Norway; Møre and Romsdal Hospital Trust, Norway; the Hunter Cancer Research Alliance (HCRA) and the Cancer Institute NSW, Australia.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Alexandre Xavier https://orcid.org/0000-0002-6397-051X

REFERENCES

Adam, R., Spier, I., Zhao, B., Kloth, M., Marquez, J., Hinrichsen, I., … Aretz, S. (2016). Exome sequencing identifies biallelic MSH3 germline mutations as a recessive subtype of colorectal adenomatous polyposis. *American Journal of Human Genetics*, *99*(2), 337–351. https://doi.org/10.1016/j.ajhg.2016.06.015

Bochkareva, E., Korolev, S., Lees-Miller, S. P., & Bochkarev, A. (2002). Structure of the RPA trimerization core and its role in the multistep DNA-binding mechanism of RPA. *EMBO Journal*, *21*(7), 1855–1863. https://doi.org/10.1093/emboj/21.7.1855

Bonis, P. A., Trikalinos, T. A., Chung, M., Chew, P., Ip, S., DeVine, D. A., & Lau, J. (2007). Hereditary nonpolyposis colorectal cancer: Diagnostic strategies and their implications. *Evid Rep Technol Assess (Full Rep)*, (150), 1–180.

Buchanan, D. D., Stewart, J. R., Clendenning, M., Rosty, C., Mahmood, K., Pope, B. J., … Win, A. K. (2017). Risk of colorectal cancer for carriers of a germ-line mutation in POLE or POLD1. *Genetics in Medicine*, *20*(8), 890–895. https://doi.org/10.1038/gim.2017.185

Carethers, J. M. (2014). Differentiating Lynch-like from Lynch syndrome. *Gastroenterology*, *146*(3), 602–604. https://doi.org/10.1053/j.gastro.2014.01.041

Carethers, J. M., Koi, M., & Tseng-Rogenski, S. S. (2015). EMAST is a form of microsatellite instability that is initiated by inflammation and modulates colorectal cancer progression. *Genes*, *6*(2), 185–205. https://doi.org/10.3390/genes6020185

de la Chapelle, A. (2004). Genetic predisposition to colorectal cancer. *Nature Reviews Cancer*, *4*(10), 769–780. https://doi.org/10.1038/nrc1453

Dong, L. I., Wu, N., Wang, S., Cheng, Y., Han, L., Zhao, J., … Hao, X. (2018). Detection of novel germline mutations in six breast cancer predisposition genes by targeted next-generation sequencing. *Human Mutation*, *39*(10), 1442–1455. https://doi.org/10.1002/humu.23597

Duraturo, F., Liccardo, R., Cavallo, A., De Rosa, M., Grosso, M., & Izzo, P. (2011). Association of low-risk MSH3 and MSH2 variant alleles with Lynch syndrome: Probability of synergistic effects. *International Journal of Cancer*, *129*(7), 1643–1650. https://doi.org/10.1002/ijc.25824

Fishel, R. (2015). Mismatch repair. *Journal of Biological Chemistry*, *290*(44), 26395–26403. https://doi.org/10.1074/jbc.R115.660142

Giardiello, F. M., Allen, J. I., Axilbund, J. E., Boland, R. C., Burke, C. A., Burt, R. W., … Rex, D. K. (2014). Guidelines on genetic evaluation and management of Lynch syndrome: A consensus statement by the US Multi-society Task Force on colorectal cancer. *American Journal of Gastroenterology*, *109*(8), 1159–1179. https://doi.org/10.1038/ajg.2014.186

Goellner, E. M., Putnam, C. D., & Kolodner, R. D. (2015). Exonuclease 1-dependent and independent mismatch repair. *DNA Repair*, *32*, 24–32. https://doi.org/10.1016/j.dnarep.2015.04.010

Hampel, H., Frankel, W. L., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., … de la Chapelle, A. (2005). Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer). *New England Journal of Medicine*, *352*(18), 1851–1860. https://doi.org/10.1056/NEJMoa043146

Hansen, M. F., Johansen, J., Sylvander, A. E., Bjørnevoll, I., Talseth-Palmer, B. A., Lavik, L. A. S., … Sjursen, W. (2017). Use of multigene-panel identifies pathogenic variants in several CRC-predisposing genes in patients previously tested for Lynch Syndrome. *Clinical Genetics*, *92*(4), 405–414. https://doi.org/10.1111/cge.12994

Hass, C. S., Gakhar, L., & Wold, M. S. (2010). Functional characterization of a cancer causing mutation in human replication protein A. *Molecular Cancer Research*, *8*(7), 1017–1026. https://doi.org/10.1158/1541-7786.MCR-10-0161

Hsieh, P., & Yamane, K. (2008). DNA mismatch repair: Molecular mechanism, cancer, and ageing. *Mechanisms of Ageing and Development*, *129*(7–8), 391–407. https://doi.org/10.1016/j.mad.2008.02.012

Huang, X., Gao, Y., He, J., Cai, J., Ta, N., Jiang, H., … Zheng, J. (2016). The association between RFC1 G80A polymorphism and cancer susceptibility: Evidence from 33 studies. *J Cancer*, *7*(2), 144–152. https://doi.org/10.7150/jca.13303

Jagmohan-Changur, S., Poikonen, T., Vilkki, S., Launonen, V., Wikman, F., Orntoft, T. F., … Karhu, A. (2003). EXO1 variants occur commonly in normal population: Evidence against a role in hereditary nonpolyposis colorectal cancer. *Cancer Research*, *63*(1), 154–158.

Jokic, M., Brcic-Kostic, K., Stefulj, J., Catela Ivkovic, T., Bozo, L., Gamulin, M., & Kapitanovic, S. (2011). Association of MTHFR, MTR, MTRR, RFC1, and DHFR gene polymorphisms with susceptibility to sporadic colon cancer. *DNA and Cell Biology*, *30*(10), 771–776. https://doi.org/10.1089/dna.2010.1189

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092

Keijzers, G., Liu, D., & Rasmussen, L. J. (2016). Exonuclease 1 and its versatile roles in DNA repair. *Critical Reviews in Biochemistry and Molecular Biology*, *51*(6), 440–451. https://doi.org/10.1080/10409238.2016.1215407

Kuiper, R. P., Vissers, L. E. L. M., Venkatachalam, R., Bodmer, D., Hoenselaar, E., Goossens, M., … Ligtenberg, M. J. L. (2011). Recurrence and variability of germline EPCAM deletions in Lynch syndrome. *Human Mutation*, *32*(4), 407–414. https://doi.org/10.1002/humu.21446

Lek, M., Karczewski, K. J., Minikel, E. V., Samocha, K. E., Banks, E., Fennell, T., … MacArthur, D. G. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. https://doi.org/10.1038/nature19057

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., & Homer, N., … 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Lindor, N. M., Rabe, K., Petersen, G. M., Haile, R., Casey, G., Baron, J., … Seminara, D. (2005). Lower cancer incidence in Amsterdam-I criteria families without mismatch repair deficiency: Familial colorectal cancer type X. *JAMA*, *293*(16), 1979–1985. https://doi.org/10.1001/jama.293.16.1979

Lipkin, S. M., & Afrasiabi, K. (2007). Familial colorectal cancer syndrome X. *Seminars in Oncology*, *34*(5), 425–427. https://doi.org/10.1053/j.seminoncol.2007.07.008

Liu, H. X., Zhou, X. L., Liu, T., Werelius, B., Lindmark, G., Dahl, N., & Lindblom, A. (2003). The role of hMLH3 in familial colorectal cancer. *Cancer Research*, *63*(8), 1894–1899.

Lynch, H. T., & de la Chapelle, A. (1999). Genetic susceptibility to non-polyposis colorectal cancer. *Journal of Medical Genetics*, *36*(11), 801–818.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Nicolas, E., Golemis, E. A., & Arora, S. (2016). POLD1: Central mediator of DNA replication and repair, and implication in cancer and other pathologies. *Gene*, *590*(1), 128–141. https://doi.org/10.1016/j.gene.2016.06.031

Palles, C., Cazier, J.-B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., … Tomlinson, I. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics*, *45*(2), 136–144. https://doi.org/10.1038/ng.2503

Paulo, P., Maia, S., Pinto, C., Pinto, P., Monteiro, A., Peixoto, A., & Teixeira, M. R. (2018). Targeted next generation sequencing identifies functionally deleterious germline mutations in novel genes in early-onset/familial prostate cancer. *PLoS Genetics*, *14*(4), e1007355. https://doi.org/10.1371/journal.pgen.1007355

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., … Rehm, H. L. (2015). 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. *Genetics in Medicine*, *17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Rodriguez-Bigas, M. A., Boland, C. R., Hamilton, S. R., Henson, D. E., Srivastava, S., Jass, J. R., … Sobin, L. (1997). A National Cancer Institute Workshop on Hereditary Nonpolyposis Colorectal Cancer Syndrome: Meeting highlights and Bethesda guidelines. *Journal of the National Cancer Institute*, *89*(23), 1758–1762. https://doi.org/10.1093/jnci/89.23.1758

Samadder, N. J., Smith, K. R., Wong, J., Thomas, A., Hanson, H., Boucher, K., … Curtin, K. (2017). Cancer Risk in Families Fulfilling the Amsterdam Criteria for Lynch Syndrome. *JAMA Oncol*, *3*(12), 1697–1701. https://doi.org/10.1001/jamaoncol.2017.0769

Steinke, V., Holzapfel, S., Loeffler, M., Holinski-Feder, E., Morak, M., Schackert, H. K., … Engel, C. (2014). Evaluating the performance of clinical criteria for predicting mismatch repair gene mutations in Lynch syndrome: A comprehensive analysis of 3,671 families. *International Journal of Cancer*, *135*(1), 69–77. https://doi.org/10.1002/ijc.28650

Talseth-Palmer, B. A., Bauer, D. C., Sjursen, W., Evans, T. J., McPhillips, M., Proietto, A., … Scott, R. J. (2016). Targeted next-generation sequencing of 22 mismatch repair genes identifies Lynch syndrome families. *Cancer Medicine*, *5*(5), 929–941. https://doi.org/10.1002/cam4.628

Umar, A., Boland, C. R., Terdiman, J. P., Syngal, S., Chapelle, A. D. L., Ruschoff, J., … Srivastava, S. (2004). Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute*, *96*(4), 261–268. https://doi.org/10.1093/jnci/djh034

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., DePristo, M. A.. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*, *43*, 11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Vasen, H. F., Mecklin, J. P., Khan, P. M., & Lynch, H. T. (1991). The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Diseases of the Colon and Rectum*, *34*(5), 424–425. https://doi.org/10.1007/BF02053699

Vasen, H. F., Watson, P., Mecklin, J. P., & Lynch, H. T. (1999). New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. *Gastroenterology*, *116*(6), 1453–1456. https://doi.org/10.1016/S0016-5085(99)70510-X

Vigeland, M. D., Gjotterud, K. S., & Selmer, K. K. (2016). FILTUS: A desktop GUI for fast and efficient detection of disease-causing variants, including a novel autozygosity detector. *Bioinformatics*, *32*(10), 1592–1594. https://doi.org/10.1093/bioinformatics/btw046

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, Y., Cortez, D., Yazdi, P., Neff, N., Elledge, S. J., & Qin, J. (2000). BASC, a super complex of BRCA1-associated proteins involved in the recognition and repair of aberrant DNA structures. *Genes & Development*, *14*(8), 927–939.

Wang, Y., Putnam, C. D., Kane, M. F., Zhang, W., Edelmann, L., Russell, R., … Edelmann, W. (2005). Mutation in Rpa1 results in defective DNA double-strand break repair, chromosomal instability and cancer in mice. *Nature Genetics*, *37*(7), 750–755. https://doi.org/10.1038/ng1587

Watson, P., & Riley, B. (2005). The tumor spectrum in the Lynch syndrome. *Familial Cancer*, *4*(3), 245–248. https://doi.org/10.1007/s10689-004-7994-z

Cancer Genome Atlas Research Network, Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., & Ozenberger, B. A., … Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Xavier A, Olsen MF, Lavik LA, et al. 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

NB. The sentence on the 3rd paragraph of page 6 in this publication should read: "*The genes EXO1, POLD1, RFC1, and RPA1 harbor variants that were predicted to be pathogenic*." (instead of "*The genes EXO1, POLD1, RFC1, and RPA1 harbor **that** variants that were predicted to be pathogenic*."

Additional Discussion

The variants identified in this study were analysed only based on their pathogenicity predicted by in-silico tools (notably SIFT (94), PhyloP (95), Polyphen2 (96), MutationTaster2 (97) and GERP++ (98)). While adding a moderate level of evidence for pathogenicity, in-silico predictions do not guarantee that the variant studied will be pathogenic. This is reflected in the "CLASSIFICATION" column of Table 2 in the publication. Even with the deleterious in-silico predictions, some variants are still classified as variants of unknown significance using the ACMG criteria.

Additionally, the result of tumour immunohistochemistry (IHC) staining for MMR proteins can give us insight about the pathways affected by the variants we are studying. (See Table 8)

For example, the variant in *POLD1,* c.1558insG, was associated with the presence of MLH1 MSH2 and MSH6 staining. In another individual, the variant *POLD1* c. 2510G>C was associated with a loss of MLH1 and PMS2 staining. This might indicate that *POLD1* variants modulate the risk of CRC independently from the MMR pathway (for example a seen in PPAP (55)). Similarly, the two variants identified in *RFC1* were associated with both a loss and a presence of MLH1 and MSH2 staining, indicating an independent role from the MMR pathway.

	Reference sequence	DNA change	AA change	Rs ID	gnomad AF	Classification	Type	LabID	Tumour MMR Immunohistochemistry
MLH1	**NM_001167619.2**	c.1130A>C	Lys377Thr	rs63750449	0.004564	VUS	MISSENSE	02-0836	N/A
	NM_000249.3	c.1039-31_1039-29delATA	N/A	rs778381149	0.0003969	VUS	INTRONIC	10-2139	-ve MSH2, MSH6
MSH6	**NM_000179.2**	c.431G>T	p.Ser144Ile	rs3211299	0.001187	Benign	MISSENSE	03-0491	N/A
	NM_001281492.1	c.892A>G	p.Lys298Glu	rs761822293	3.979E-06	VUS	MISSENSE	01-0597	N/A
	NM_001281492.1	c.1054C>T	p.Arg352X	rs63750909	0.00003186	Pathogenic	NONSENSE	00-0167	N/A
	NM_001281492.1	c.1118C>G	p.Ser373Cys	rs63750897	0.001165	Likely Benign	MISSENSE	04-0768	N/A
EXO1	**NM_003686.4**	c.1928T>A	p.Leu643X	NR	N/A	VUS	NONSENSE	01-0512	N/A
	NM_003686.4	c.2009A>G	p.Glu670Gly	rs1776148	0.78	Benign	MISSENSE	01-0543	-ve MLH1, PMS2, +ve MSH2, MSH6
	NM_006027 .4	c.2485G>T	p.Glu829X	rs757677420	0.00000292	VUS	NONSENSE	01-0079	N/A
POLD1	**NM_001256849.1**	c.1249A>G	p.Thr417Ala	NR	N/A	VUS	MISSENSE	04-0773	N/A
	NM_001256849.1	c.1558insG	p.	NR	N/A	VUS	FRAMESHIFT	98-1929	+ve MLH1, MSH2, MSH6
	NM_001256849.1	c. 2510G>C	p.Gly811Ala	NR	N/A	VUS	MISSENSE	08-1209	-ve MLH1, PMS2
RFC1	**NM_001204747.1**	c.2017G>A	p.Val673Met	rs28903096	0.0006994	VUS	MISSENSE	02-0125	-ve MLH1, MSH2
	NM_001204747.1	c.2276A>G	p.Lys759Arg	NR	N/A	VUS	MISSENSE	04-0822	+ve MLH1, MSH2

RPA1	**NM_002945.4**	c.856G>T	p.Val286Phe	rs55800538	0.002942	VUS	MISSENSE	01-0252	N/A
	NM_002945.4	c.1160G>A	p.Gly387Asp	NR	N/A	VUS	MISSENSE	05-1220	+ve MLH1, MSH2, MSH6, PMS2
	NM_002945.4	c.1165C>T	p.Arg389W	rs202068855	0.0005468	VUS	MISSENSE	04-0670	N/A

Table 8. Variants identified in *Comprehensive mismatch repair gene panel identifies variants in patients with Lynch-like syndrome* published in. Mol Genet Genomic Med Tumours MMR Immunohistochemistry reflect the presence (+ve) or absence (-ve) of staining for a particular MMR protein in the tumour

CHAPTER 3:

TAPES: a Tool for Assessment and Prioritisation in Exome Studies

CHAPTER 3: TAPES: a Tool for Assessment and Prioritisation in Exome Studies

3.0) Introduction

With the evolution of NGS, the amount of data to analyse grew exponentially. Multiple full genomes can be sequenced and analysed in parallel. This highlights the strong need for automated pathogenicity prediction. Researchers need to be able to focus solely on relevant impactful variants and safely discard benign variants.

Automatic annotation of NGS data have been perfected throughout the years with software's like VEP (84), ANNOVAR (83) or SNPeff (85). They allow researcher to associate a variation (characterised by its chromosome location, a reference and an alternative allele, eg. chr5:80873118 G>A) to several characteristics (such as variation type, protein consequence, in-silico predictions, splice variants affected, etc). This helps researchers to make a more informed decision about the variant considered.

In addition to the variant annotation, the ACMG/AMP (American College of Medical Genetics/Association of Molecular Pathology) proposed a set of criteria to predict the overall pathogenicity of variants (see introduction, *1.4-Variant pathogenicity prediction and prioritisation*). These criteria are used to classify genetic variants into 5 classes; Benign, Likely Benign, Unknown Significance, Likely Pathogenic and Pathogenic (or respectively class 1, 2, 3, 4 and 5).

Software were released to automate the assignment of the ACMG/AMP classification such as CharGer (87) or InterVar (88). However, those software's had limitation, such as not being able to properly handle multi-sample variants or not being able to handle trio-data. In addition, the ACMG/AMP prediction is categorical and leaves a lot of variants with the status of Unknown Significance (VUS), even when they are approaching the criteria for inclusion in the "Likely Pathogenic" class.

Aims

Throughout my PhD project, I aimed to refine the pipeline for pathogenicity prediction and WES analysis. The limitations of currently available software were restricting the analysis of our FAP-like cohort WES data. To overcome them, I aimed to develop a system to transform the categorical predictions of the ACMG\AMP into a linear prediction of pathogenicity prediction (see Figure 7). In addition, I wanted to develop a method to assess the enrichment of variants in a cohort without the need of control samples. Finally, the last aim was to create a tool that provides researchers with a powerful reporting and filtering system.

Approach

We first aimed to modify the categorical predictions of the ACMG\AMP. We tested custom scores based on additions (similar to CharGer (87) custom score) or weighted multiplications. They were sub-optimal but were a good approximation that allowed us to rank variants and prioritize the variants researchers have to consider (we later adopted a model developed by Tavtigian et al. 2018 (99)).

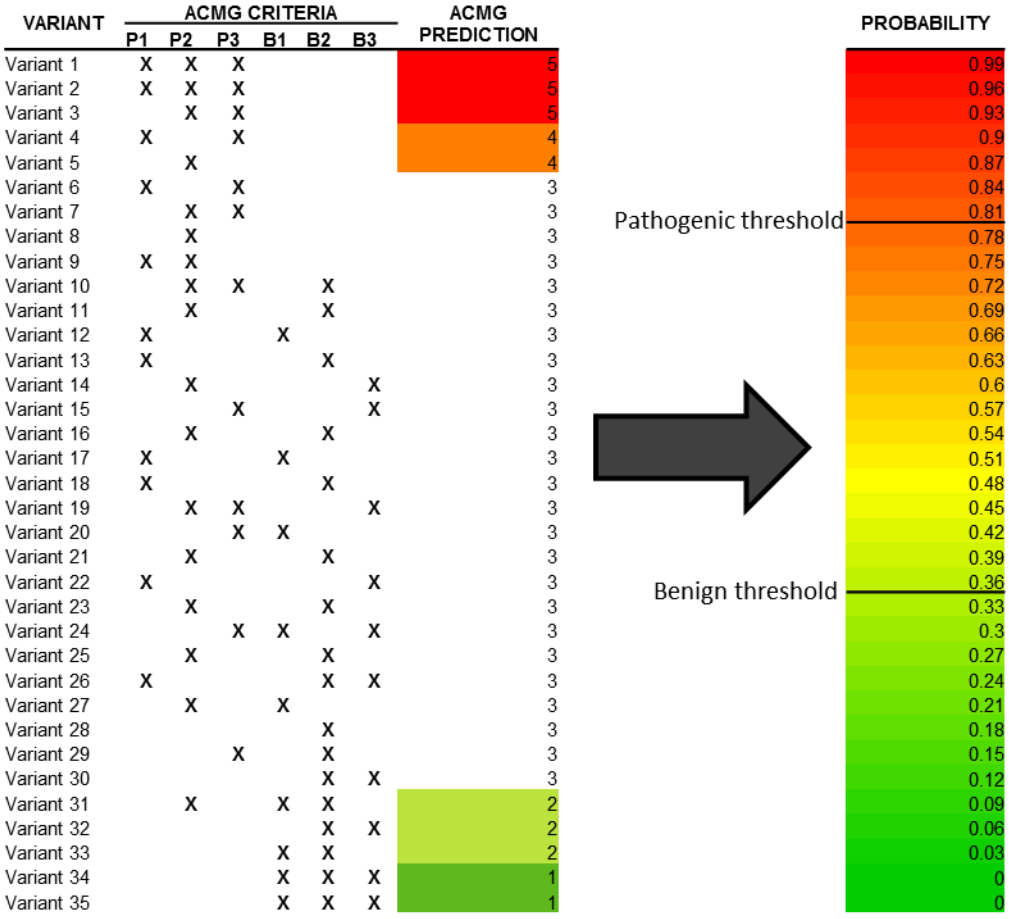

Figure 8. Transformation of the ACMG\AMP categorical prediction into a linear probability of pathogenicity.

We then addressed the issue of variant enrichment in cohorts compared to normal healthy individuals. The idea behind it was that public databases are a very useful resource and have sequenced so many healthy individuals from so many different backgrounds that no other study can compare to them (the gnomad initiative (100) provides 125,748 exomes and 15,708 genome sequenced) and can be utilised by researchers.

In addition to the better pathogenicity prediction, a lacking feature in available software's were the lack of filtering and reporting options. We wanted to provide useful filtering (excluding/including variants based on pathogenicity, disease or gene-lists) and reporting (Polygenic Risk Score for a specific trait, pathway enrichment, or gene burden for the entire cohort). TAPES was the result of a year of work trying to refine the WES analysis pipeline and is the tool I wish had existed when I started my PhD project.

3.1) Publication

STATEMENT I

This is a co-author statement attesting to the candidate's contribution to the publication listed below:

I attest that Research Higher Degree candidate Alexandre Xavier contributed to the publication listed below by performing the design and code, the benchmark and validation as well as writing the manuscript.

Xavier A, Scott RJ, Talseth-Palmer BA. TAPES: A tool for assessment and prioritisation in exome studies. PLOS Computational Biology. 2019;15(10):e1007453.

This statement explains the contribution of all authors in the article listed above:

Table illustrating author contribution percentage and description of contribution to the article listed above.

Author	Contribution (%)	Description of contribution to article	Signature	Date
Alexandre Xavier	85%	Design and code writing, Benchmark and Validation, Manuscript writing		15/10/2019
Rodney J. Scott	5%	Critical revision of the manuscript and study supervision		17/10/2019
Bente A. Talseth-Palmer	10%	Obtained funding, critical revision of the manuscript and study supervision	Bente Talseth -Palmer Digitally signed by Bente Talseth-Palmer Date: 2019.11.18 11:52:11 +11'00'	18/11/2019

Digitally signed by Dr Lesley MacDonald-Wicks
Date: 2019.11.20 15:38:29 +11'00'

Alexandre Xavier

Date: 15/10/2019

Dr Lesley MacDonald-Wicks

Assistant Dean (Research Training)

Date: 20/11/2019

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

TAPES: A tool for assessment and prioritisation in exome studies

Alexandre Xavier[1]*, **Rodney J. Scott**[1,2], **Bente A. Talseth-Palmer**[1,3]

1 School of Biomedical Sciences and Pharmacy, Faculty of Health and Medicine, University of Newcastle and Hunter Medical Research Institute, Newcastle, Australia, **2** NSW Health Pathology North, John Hunter Hospital, Newcastle, Australia, **3** Clinic for Research, Innovation, Education and Development, Møre and Romsdal Hospital Trust, Molde, Norway

* alexandre.xavier@live.fr

OPEN ACCESS

Citation: Xavier A, Scott RJ, Talseth-Palmer BA (2019) TAPES: A tool for assessment and prioritisation in exome studies. PLoS Comput Biol 15(10): e1007453. https://doi.org/10.1371/journal.pcbi.1007453

Editor: Mihaela Pertea, Johns Hopkins University, UNITED STATES

Received: June 30, 2019

Accepted: October 1, 2019

Published: October 15, 2019

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1007453

Data Availability Statement: All source code can be found at: https://github.com/a-xavier/tapes. Documentation is also available through this repository and at: https://github.com/a-xavier/tapes/wiki.

Abstract

Next-generation sequencing continues to grow in importance for researchers. Exome sequencing became a widespread tool to further study the genomic basis of Mendelian diseases. In an effort to identify pathogenic variants, reject benign variants and better predict variant effects in downstream analysis, the American College of Medical Genetics (ACMG) published a set of criteria in 2015. While there are multiple publicly available software's available to assign the ACMG criteria, most of them do not take into account multi-sample variant calling formats. Here we present a tool for assessment and prioritisation in exome studies (TAPES, https://github.com/a-xavier/tapes), an open-source tool designed for small-scale exome studies. TAPES can quickly assign ACMG criteria using ANNOVAR or VEP annotated files and implements a model to transform the categorical ACMG criteria into a continuous probability, allowing for a more accurate classification of pathogenicity or benignity of variants. In addition, TAPES can work with cohorts sharing a common phenotype by utilising a simple enrichment analysis, requiring no controls as an input as well as providing powerful filtering and reporting options. Finally, benchmarks showed that TAPES outperforms available tools to detect both pathogenic and benign variants, while also integrating the identification of enriched variants in study cohorts compared to the general population, making it an ideal tool to evaluate a smaller cohort before using bigger scale studies.

Author summary

New sequencing techniques allow researchers to study the genetic basis of diseases. Predicting the effect of genetic variants is critical to understand the mechanisms underlying disease. Available software can predict how pathogenic a variant is, but do not take into account the abundance of a variants in a cohort. TAPES is a simple open-source tool that can both more accurately predict pathogenicity (using probability over categories) and provide insight on variants enrichment in a cohort sharing the same disease.

Funding: The Hunter Cancer Research Alliance (https://www.hcra.com.au/) funded Bente Talseth-Palmer and Alexandre Xavier. The University of Newcastle (https://www.newcastle.edu.au/) funded Alexandre Xavier. The Cancer Institute NSW (https://www.cancercouncil.com.au/) funded Bente Talseth-Palmer. The 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.

This is a *PLOS Computational Biology* Software paper.

Introduction

With the advances in Next-Generation Sequencing (NGS) technologies and the decline in price over the last few years, exome sequencing has become a standard tool to explore the genetic basis of inherited diseases [1]. It has become easy to annotate the ever-increasing amount of variants identified by such methods, using tools such as VEP [2], snpEff [3] or ANNOVAR [4]. These tools help researchers to better predict the downstream effect of a variant and give insight, for example, on the frequency of the mutation in the general population, the impact on proteins or in-silico predictions of pathogenicity.

In 2015, the American College of Medical Genetics (ACMG) published a set of criteria to assess the probability of a variant pathogenicity, classifying them into five categories [5], from benign to pathogenic, facilitating downstream analysis.

Since then, tools have been developed to assess individual variant pathogenicity using the ACMG criteria (such as CharGer [6] and Intervar [7]) but they do not have the ability to take into account the frequency of variants in a cohort. The categorical nature of the ACMG criteria also leaves a lot of variants classified as "a variant of unknown significance".

Here, we present TAPES, an open-source tool to both assess and prioritise variants by pathogenicity. TAPES can assign the ACMG criteria and by using one of the first implementations of the model described in Tavtigian *et al.* [8], providing a more nuanced and easy to understand estimated probability for a variant to be either pathogenic or benign, thus transforming categorical classification into a more linear prediction. Our goal during development was first to create a simple tool that can better predict pathogenicity and reject benign variants, and then to assess a cohort sharing a phenotype by detecting enriched variants compared to the general population without the need of control samples. In addition, we focused on providing simple yet powerful reporting and filtering systems, while allowing pathway analysis of pathogenic mutations, gene-burden calculations and per-sample reporting.

Design and implementation

ANNOVAR interface and annotated variant file

TAPES sorting option can be used with both ANNOVAR and VEP annotated variant calling files (VCF). However we also provide users with simple wrapping tools for a local installation of ANNOVAR to simplify the workflow (this requires users to download ANNOVAR). Users can annotate VCF, gzipped VCF and binary VCF (BCF) using two simple commands without having to specify the databases and annotations to use.

While there are a set of annotation needed to assign all ACMG criteria (see https://github.com/a-xavier/tapes/wiki/Necessary-Annotations for the full list), TAPES will use as many available annotations as possible to assign the relevant ACMG criteria.

Variant classification

TAPES requires annotated ANNOVAR (VCF or tab/comma-separated values) or VEP (VCF) files to use the sorting module.

Regular ACMG criteria assignment. For most of the ACMG criteria assignment (PVS1, PS1, PS3, PM1, PM2, PM4, PM5, PP2, PP3, PP5, BS1, BS2, BS3, BP1, BP3, BP4, BP6, BP7 and BA1), we tried to stay as true to the original ACMG definition as possible when implementing

the criteria assignment. Please see Richards *et al.* [5] and S1 Table for more information on the ACMG Criteria definition.

Enrichment analysis / PS4 criteria. One of TAPES unique features is the ability to calculate variant enrichment from public frequency data (ExAC or gNomad [9]), without having to sequence control samples. In cohort studies, TAPES require a multi-sample vcf file to extract genotyping data and get frequencies from the cohort studied. It uses a simple one-sided Fisher's exact test to calculate both the Odds Ratio (OR) and the p-value of the enrichment. Only the variant enrichment in the cohort is tested against the general population.

Since OR calculation requires integer numbers and frequency in the general population is given as a 0–1 fraction, TAPES approximates the number of individuals affected using the following formula.

If MAF_c is the Minor Allele Frequency (MAF) in a control population, n_c is the number of individuals affected by the variant in the control population and N_c is the number of individuals without the variant then:

$$MAF_c = y \times 10^{-x}, n_c = \lceil y \rceil \text{ and } N_c = \frac{10^x}{2} - n_c.$$

For example if:

$$MAF_c = 3.23 \times 10^{-5} \text{ then } n_c = 4 \text{ and } N_c = \frac{10^5}{2} - 4.$$

This approximation is only valid if the following assumptions are made; MAF in the control population is under 0.05 and that very rare variants are mostly heterozygous.

The PS4 criteria assignment was designed to be more stringent than a normal study with controls (choosing to overestimate the frequency in the general population) and will only be assigned if OR $\geq$ 20, p-value $\leq$ 0.001 and at least 2 individuals in the cohort share the variant.

Trio analysis / PS2 assignment. TAPES allow researchers to work with trio studies. In trio studies, the user provides information such as sample name, trio ID and pedigree information in a tab-delimited file. Then PS2 will be assigned if a variant is identified as *de-novo* and healthy parents are removed from downstream analysis. PS2 is assigned to a variant if it was found as *de-novo* in any trio but details from each trio will still be provided.

Probability of pathogenicity calculation. TAPES includes the model developed by Tavtigian *et al.* [8] to transform ACMG categorical classification into linear probability of pathogenicity and the method uses the default parameters from (Prior P = 0.10, O_{PVSt} = 350 and X = 2). This allows for a finer pathogenicity prediction and adjustable thresholds to decide variant pathogenicity. It is important to keep in mind that this measure is a probability and not a measure of how pathogenic a variant is.

Cohort reporting

TAPES provides an array of different useful reports.

Filtering. TAPES can easily perform advance filtering. Three different options are available. First, users can provide a custom list of gene symbols (either as a text file or directly on the command line) to only output variants present in those genes. Then users can also do a reverse pathway search by providing the name of a pathway (extracted from KEGG pathways [10]) and output a report with variants in genes involved in that pathway. Finally, users can run searches based on terms contained in the description for each gene, i.e. if the user looks for 'autosomal dominant' genes or 'colorectal cancer' genes. These filtered reports keep the same format as the main report, making it possible to use them with other reporting tools.

By-sample report. For each individual in the cohort, a report containing the variant predicted to be pathogenic with the highest level of confidence will be available. This allows the study of individual samples and their specificity.

By-gene report. TAPES can also calculate, for each gene, a gene burden score. This score helps determining which genes harbour the most potentially pathogenic variants in a cohort. This can be useful when searching for variants in diseases caused by single genes and that cannot be discovered using pathway analysis. The gene burden score is calculated by summing the probability of pathogenicity of a specific variant multiplied by the number of individuals with that genotype in the cohort.

$$Gene\ burden\ score = \sum_{1}^{n} P_i \times N_i$$

Calculated for each gene, where P_i = the probability of pathogenicity of the variant and N_i = Number of samples affected by the variant. If $P_i \leq 0.80$ then the variant is excluded.

This measure is useful to detect which genes in the cohort are particularly enriched in pathogenic and probably pathogenic variants (it is important to remember that this measure is a sum of probabilities). However, there are a few caveats. This measure might be affected by very long genes or genes frequently mutated in exomes (FLAGS [11]). In some cases, poorly mapped reads (for example due to pseudo-autosomal regions in the X or Y chromosome), might impact the result with an excessive number of samples affected by a variant. TAPES provides an appropriate warning for all of those cases.

Pathway analysis. TAPES can also perform a pathway analysis using the EnrichR [12] API. Only genes containing variants that are predicted to be pathogenic are kept as a gene list. The user can then use any library to analyse the gene list but the default is GO_Biological_Process_2018. Pathway analysis is important to understand the possibly disrupted mechanism and the commonalities between variants found in a cohort.

Results

Variant classification

TAPES variant classification was benchmarked against similar tools, CharGer [6] and Intervar [7] using the prediction on the pathogenicity of variants of the expert panel of Zhang et al., 2015 as reference [13] (see S2 Table for the full table). This dataset was also used to benchmark CharGer in their original publication. The 'probably pathogenic' and 'pathogenic' predictions were pooled into one 'pathogenic' group. Similarly the 'probably benign' and 'benign' were pooled into one 'benign' group.

To assess the predictive power of each software, we used Receiver Operating Characteristics (ROC) curves and calculated the area under the curve (AUC) as well as the precision-recall curves and average precision (AP). We compared TAPES ACMG and probability of pathogenicity prediction with CharGer score and InterVar ACMG prediction (see Fig 1).

TAPES probability of pathogenicity, using Tavtigian *et al* [8] modelling, outperformed both software's tested using AUC and AP for prediction of both pathogenic and benign variants.

AUC and AP show that using TAPES ACMG criteria assignment remains less precise than using CharGer custom score (due to the additional information CharGer need to function properly) and closer to InterVar. Using the probability of pathogenicity should be the preferred way to identify pathogenic variants and reject benign variants. Based on ROC curves, a threshold of 0.80–0.85 for probability of pathogenicity seemed to keep high true positive rate (TPR) while low false positive rate (FPR) for predicting pathogenic variants. Similarly, a

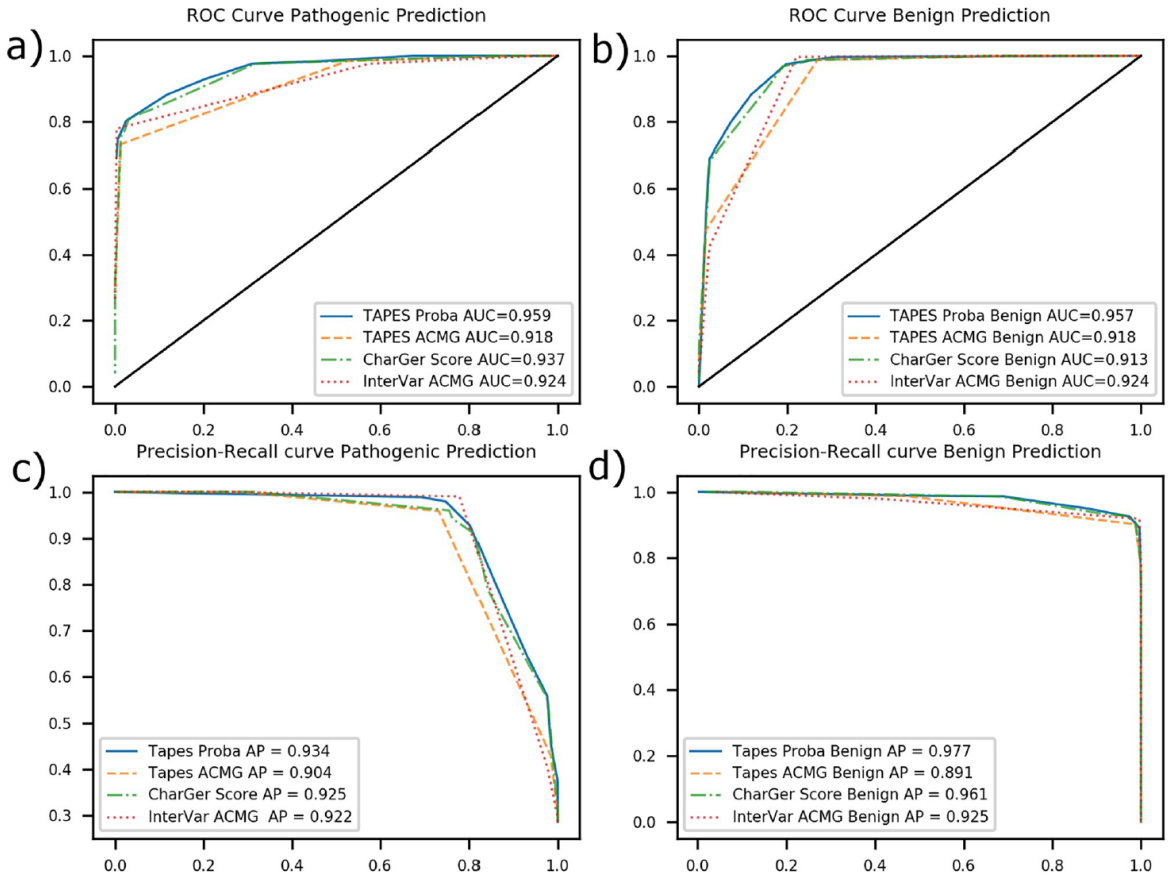

Fig 1. ROC curves and precision recall curves. a) ROC curve of various softwares for pathogenicity prediction AUC b) ROC curve of various softwares for benignity prediction AUC c) Precision-recall curve of various softwares for pathogenicity prediction d) Precision-recall curve of various softwares for benignity prediction (Metrics used; TAPES proba; TAPES probability of pathogenicity prediction, TAPES ACMG: TAPES ACMG prediction, CharGer score: CharGer prediction of pathogenicity based of a custom score, InterVar ACMG: InterVar ACMG prediction).

https://doi.org/10.1371/journal.pcbi.1007453.g001

threshold of 0.20–0.35 for probability of pathogenicity had high TPR and low FPR for predicting benignity.

To validate these findings and choose the best probability thresholds for pathogenicity and benignity, we used TAPES, InterVar and CharGer on a different dataset (**see** S3 Table). Using 530 hand curated variants from ClinGen evidence repository (https://erepo.clinicalgenome.org/evrepo/) as ground truth. TAPES outperformed both InterVar and CharGer (**see** Fig 2). In addition to the precision of the prediction, TAPES also outperformed other software in terms of absolute number of variants correctly identified.

We recommend to use TAPES probability of pathogenicity prediction with either lenient thresholds of 0.8 and 0.35 (respectively for pathogenicity and benignity) or stricter thresholds of 0.85 and 0.20.

Variant enrichment / PS4 benchmark

We compared our method of calculation of ORs compared to the normal method (see Fig 3).

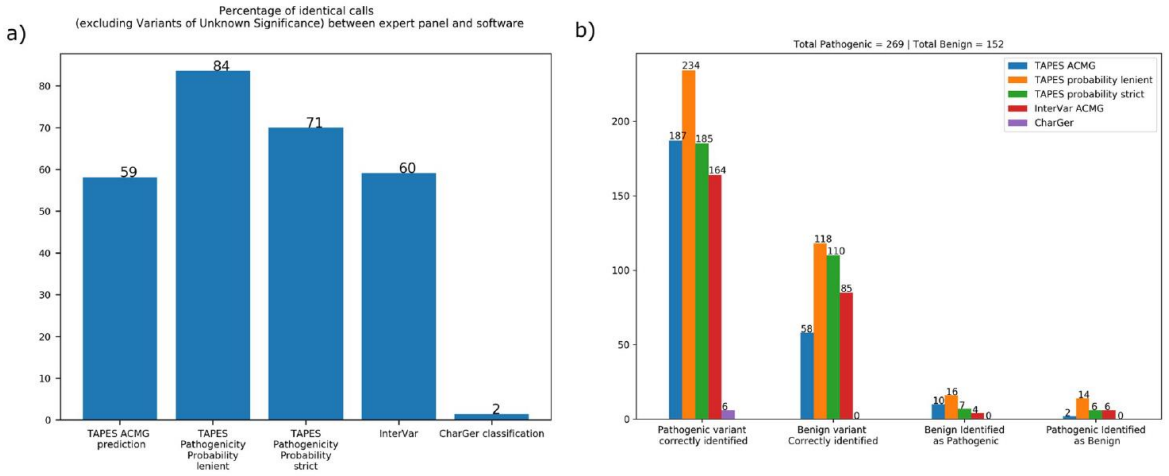

Fig 2. Validation dataset software comparisons. a) Percentage of identical calls between the ClinGen expert panel decisions and software prediction. Lenient thresholds are 0.80 for pathogenicity and 0.35 for benignity. Strict thresholds are 0.85 for pathogenicity and 0.20 for benignity. **b) Absolute number of variants predictions**. Pathogenic and benign variants correctly and incorrectly identified between the panel of expert and various software. (Metrics used; TAPES probability lenient; TAPES probability of pathogenicity prediction 0.35–0.80, TAPES probability strict; TAPES probability of pathogenicity prediction 0.20–0.85 TAPES ACMG: TAPES ACMG prediction, CharGer: CharGer prediction of pathogenicity based of a custom score, InterVar ACMG: InterVar ACMG prediction).

https://doi.org/10.1371/journal.pcbi.1007453.g002

The OR using TAPES extrapolation is always smaller than the normal calculation, making it more stringent. Similarly, the p-value of the Fisher's exact test rises faster with frequency than the normal method. This way, only the most significantly enriched variant are assigned with PS4 to ensure very few false positives.

Reporting options. TAPES reporting options are powerful and easy to use. Using a mock input file with variants from Zhang *et al.* [13] as well as simulated samples to form a cohort, the pathway analysis correctly identified DNA repair as the pathway with the most probable pathogenic variants.

The by-gene report also identified BRCA2 as the gene with the highest gene burden.

See S1 File to see all reports templates.

Availability and future directions

TAPES is available on github at: https://github.com/a-xavier/tapes, under the MIT licence, which allows anyone to both freely download and modify the source code. Help can be found both in the manual (located in the main repository) or on the wiki (https://github.com/a-xavier/tapes/wiki). Examples of inputs can also be found in the main repository. Dependencies can be easily installed using PyPi repositories (pip). All builds are verified through Travis continuous integration on Linux, Windows and macOS. All benchmarks and examples showed in this manuscript were generated using TAPES release 0.1.

All benchmarks and examples were generated using the initial release 0.1 of TAPES (https://github.com/a-xavier/tapes/releases).

TAPES will continue to evolve with the advances in various databases such as ExAC, dnSNP or dbNSFP. As they constantly update their data and the format, TAPES will evolve to be more precise and accurate. In addition, future directions include more statistical measures to detect significant variants in different cohort studies.

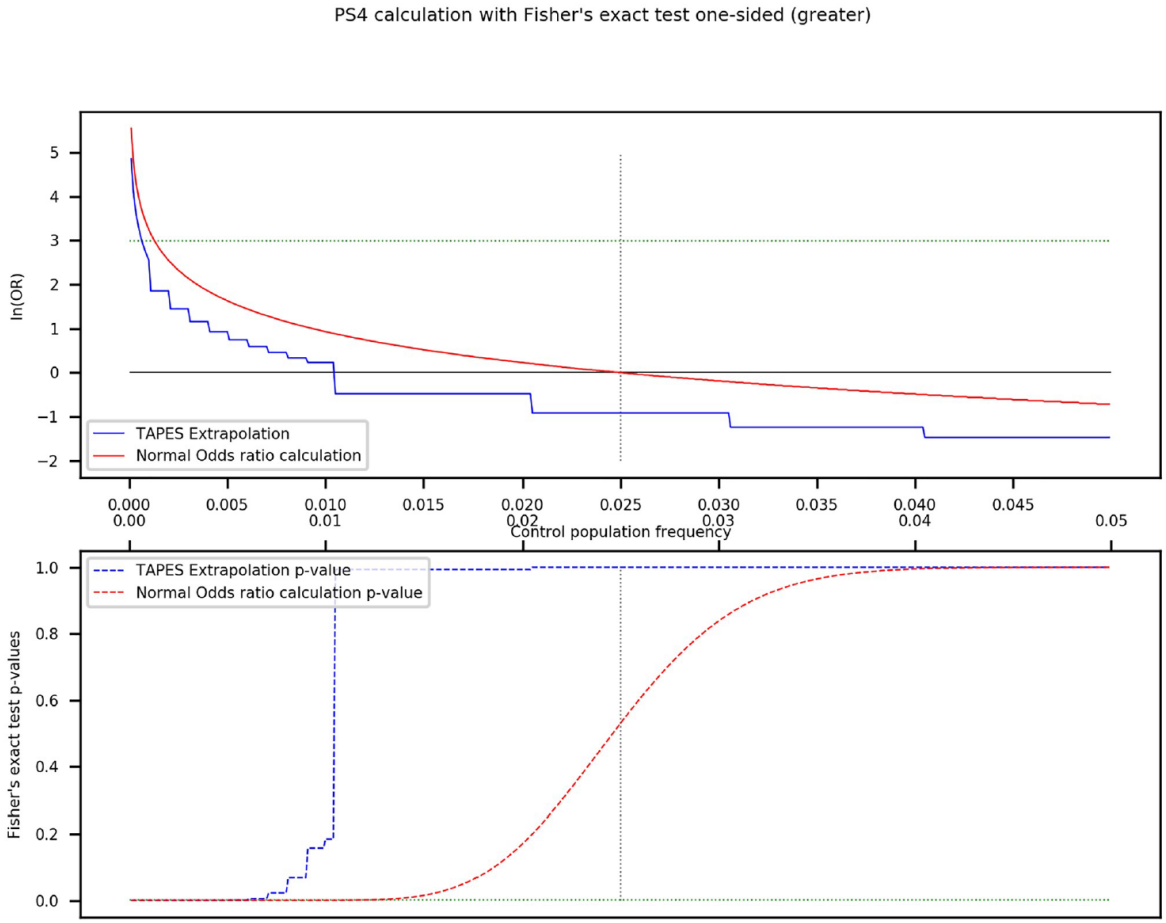

Fig 3. PS4 calculation with Fisher's exact test one sided. Comparison of TAPES extrapolation of odds rations compared to the normal method (top graph). Comparison of the p-value of both methods (bottom graph). The **vertical dotted line** represents the known frequency of the variant in the studied cohort (0.025). The **horizontal green dotted line** represents the thresholds used to assign PS4 (OR = 20 or ln(OR) = 2.9957 (top) and p-value < 0.01(bottom)).

https://doi.org/10.1371/journal.pcbi.1007453.g003

We aim to keep TAPES as simple and useful as possible to make it a perfect endpoint tool to analyse variants from small-scale cohorts.

Supporting information

S1 Table. ACMG criteria assignment in TAPES and definitions from the original Richards et al 2015 article.
(XLSX)

S2 Table. Comparison of Prediction between different pathogenicity assessment software and the expert panel from Zhang J et al. 2015. Comparison between TAPES ACMG and pathogenicity probability prediction, CharGer Prediction Score and InterVar AMCG Prediction.
(XLSX)

S3 Table. Comparison of Prediction between different pathogenicity assessment software and the expert panel from ClinGen evidence repository variants. Comparison between TAPES ACMG and pathogenicity probability prediction, CharGer Prediction Score and InterVar AMCG Prediction.
(TXT)

S1 File. Example reports from TAPES sort option. Generated using the data from: Zhang, J., et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. N Engl J Med 2015;373(24):2336–2346. Using the command: *python tapes.py sort -i ./input.csv -o ./Report/ --tab --by_gene --by_sample --enrichr --disease "autosomal dominant" --kegg "Pathways in cancer"*. This file gives examples for the main report, the by-gene report, the by-sample report, the enrichr report, the disease report and the kegg report.
(XLSX)

S2 File. Files used for TAPES benchmark and validation. The Initial Benchmark folder contains all files used for the original benchmark, CharGer_and_Panel_Benchmark.xlsx: CharGer pathogenicity prediction and expert panel decision from from: Zhang, J., et al. 2015, extracted from CharGer original publication, **Synthetic_VCF_for_Benchmark.vcf.vcf:** Synthetic VCF file created from the **CharGer_and_Panel_Benchmark.xlsx** variants information, **InterVar_Benchmark.txt:** InterVar predictions of pathogenicity after analysis of the synthetic VCF, **TAPES_Benchmark.xlsx**: TAPES prediction of pathogenicity after analysis of the synthetic VCF. The results of all 3 software are compiled in S2 Table. **The Validation folder contains all filed used for the validation of the pathogenicity thresholds and comparison with other software. TAPES_validation_synthetic.vcf**: Synthetic VCF created with data extracted from the ClinGen evidence repository (https://erepo.clinicalgenome.org/evrepo/), **TAPES_validation.charger.txt:** the CharGer predictions of pathogenicity after analysis of the Synthetic VCF, **TAPES_Validation.intervar.txt:** InterVar prediction of pathogenicity after analysis of the synthetic VCF, **TAPES_Validation.tapes.txt:** TAPES prediction of pathogenicity after analysis of the Synthetic VCF. The results of all 3 software are compiled in S3 Table.
(ZIP)

Acknowledgments

The authors would like to thank Mr. Sean Burnard for his helpful advices regarding this manuscript.

Author Contributions

Conceptualization: Alexandre Xavier.

Formal analysis: Alexandre Xavier.

Funding acquisition: Rodney J. Scott, Bente A. Talseth-Palmer.

Methodology: Alexandre Xavier.

Resources: Rodney J. Scott.

Software: Alexandre Xavier.

Supervision: Bente A. Talseth-Palmer.

Writing – original draft: Alexandre Xavier.

Writing – review & editing: Rodney J. Scott, Bente A. Talseth-Palmer.

References

1. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011; 12(11):745–55. Epub 2011/09/29. https://doi.org/10.1038/nrg3031 PMID: 21946919.
2. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17(1):122. Epub 2016/06/09. https://doi.org/10.1186/s13059-016-0974-4 PMID: 27268795
3. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012; 6(2):80–92. Epub 2012/06/26. https://doi.org/10.4161/fly.19695 PMID: 22728672
4. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38(16):e164. Epub 2010/07/06. https://doi.org/10.1093/nar/gkq603 PMID: 20601685
5. 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):405–24. Epub 2015/03/06. https://doi.org/10.1038/gim.2015.30 PMID: 25741868
6. Scott AD, Huang KL, Weerasinghe A, Mashl RJ, Gao Q, Martins Rodrigues F, et al. CharGer: Clinical Characterization of Germline Variants. Bioinformatics. 2018. Epub 2018/08/14. https://doi.org/10.1093/bioinformatics/bty649 PMID: 30102335.
7. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet. 2017; 100(2):267–80. Epub 2017/01/31. https://doi.org/10.1016/j.ajhg.2017.01.004 PMID: 28132688
8. Tavtigian SV, Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med. 2018; 20(9):1054–60. Epub 2018/01/05. https://doi.org/10.1038/gim.2017.210 PMID: 29300386
9. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016; 536(7616):285–91. Epub 2016/08/19. https://doi.org/10.1038/nature19057 PMID: 27535533
10. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017; 45(D1):D353–D61. Epub 2016/12/03. https://doi.org/10.1093/nar/gkw1092 PMID: 27899662
11. Shyr C, Tarailo-Graovac M, Gottlieb M, Lee JJ, van Karnebeek C, Wasserman WW. FLAGS, frequently mutated genes in public exomes. BMC Med Genomics. 2014; 7:64. Epub 2014/12/04. https://doi.org/10.1186/s12920-014-0064-y PMID: 25466818
12. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016; 44(W1):W90–7. Epub 2016/05/05. https://doi.org/10.1093/nar/gkw377 PMID: 27141961
13. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. N Engl J Med. 2015; 373(24):2336–46. Epub 2015/11/19. https://doi.org/10.1056/NEJMoa1508054 PMID: 26580448

CHAPTER 4:

Familial Polyposis Syndromes

CHAPTER 4: Familial Polyposis Syndromes

4.0) Introduction

Familial Polyposis Syndromes (FPS) are inherited CRC syndromes causing a specific phenotype of polyps. The early stage of polyps is called adenomas. These adenomas are benign but, if left untreated, will evolve into malignant carcinomas (45). While polyposis is mostly sporadic, a significant proportion of polyposis is caused by genetic variation. Pathogenic variants in *APC* are the most common cause of inherited polyposis and represent around 1% of all CRCs diagnosed annually. *APC* variants are associated with FAP, aFAP, Gardner syndrome (101) and Turcot syndrome (102, 103).

As described in the general introduction, there are numerous other well-known polyposis syndromes, with a clearly identified genetic background: MAP, NAP, PPAP, PJS, JPS, etc.

However, despite the number of identified syndromes, the majority of familial colorectal polyposis patients still do not have a precise genetic origin associated with their disease. This suggests that there is a large population of individuals with a higher risk of developing polyps and presumably colorectal cancer that remains unidentified.

Aims

The aim of the current study is to identify the genetic aberration present in FAP-like individuals. FAP-like individuals are defined by their strong family history of polyposis or CRC, a clinical diagnosis of FAP but no identified pathogenic variant in *APC or MUTYH*.

The goal is to establish a comprehensive list of genetic factors contributing to the increased risk of polyposis in individuals with such symptoms. With sufficient evidence, genetic screening for polyposis can be extended to include more genes. This would allow more individuals to be identified as at-risk, monitored and offered appropriate disease management to mitigate their risk of presenting with later stages of disease.

Approach

In order to identify the genetic factors contributing to familial polyposis syndromes, whole exome sequencing (WES) was used. WES is a good trade-off between the high throughput of whole-genome sequencing (WGS) and targeted sequencing (using a known gene panel).

WES only focuses on the coding regions of the genome and can include regulatory regions (upstream and downstream of the genes) as well as non-coding RNA sequences, including both miRNA and lncRNA).

Beyond the identification of pathogenic variants, WES also allows the interrogation of the copy number status of genes (using read counts as the main metrics).

A cohort of 48 FAP-like individuals was recruited based on their clinical diagnosis of polyposis, their family history (or CRC) and mutation-negative *APC/MUTYH* screening.

Individuals were selected if they had confirmed polyps and any one of the following:

- At least one first degree relative diagnosed with any type of cancer
- At least two second degree relatives diagnosed with any type of cancer
- At least three third degree relatives diagnosed with any type of cancer

Exceptions were made to these rules. For the samples BS, BP and BT because of their early age of diagnosis (35, 39 and 41). Sample W was selected because of its high number of polyps (more than 40). Sample AQ was selected with 1 confirmed second degree relative and 1 non-confirmed first-degree relative case. Finally, sample BQ was included with only 1 third-degree relative due to confirmed CRC at an early-age (52) followed by a relapse despite being test for both FAP genes and HNPCC genes. The final cohort had an average 3.22 relatives diagnosed with cancer.

Using WES, we identified pathogenic variants, which were analyzed using TAPES and Varsome (104) for pathogenicity prediction, pathway analysis and calculated polygenic risk score for various CRC risk factors. In addition, copy number status was analysed using both XHMM (105) and EXCAVATOR2 (106).

4.1) Publication – Short Report

STATEMENT I

This is a co-author statement attesting to the candidate's contribution to the publication listed below:

I attest that Research Higher Degree candidate Alexandre Xavier contributed to the publication listed below by performing the whole-exome sequencing, the analysis of the data and the manuscript writing.

Xavier A, Scott RJ, Talseth-Palmer BA. Exome sequencing of unexplained familial polyposis identifies both known and novel causative genes To be submitted to Clinical Genetics, August 2020

This statement explains the contribution of all authors in the article listed above:

Table illustrating author contribution percentage and description of contribution to the article listed above.

Author	Contribution (%)	Description of contribution to article	Signature	Date
Alexandre Xavier	65%	Performed whole exome sequencing, Performed analysis and wrote manuscript.		15/20/2019
Rodney J. Scott	10%	Critical revision of the manuscript and study supervision		17/10/2019
Bente A. Talseth-Palmer	25%	Study design, obtained funding, critical revision of the manuscript and study supervision	Bente Talseth-Palmer Digitally signed by Bente Talseth-Palmer Date: 2019.11.18 11:50:54 +11'00'	18/11/19

Alexandre Xavier

Date: 15/10/2019

Dr Lesley MacDonald-Wicks

Assistant Dean (Research Training)

Date: 20/11/19

Short report: Exome sequencing of unexplained familial polyposis identifies both known and novel causative genes

Alexandre Xavier[1], Rodney J. Scott[2] and Bente Talseth-Palmer[1,]

[1] University of Newcastle and Hunter Medical Research Institute, Lot 1, Kookaburra Circuit

New Lambton Heights, NSW, AUS

[2] New South Wales Pathology, Molecular Genetics, John Hunter Hospital, Newcastle, NSW, AUS

Abstract

Inherited polyposis syndromes are predominantly caused by pathogenic variants in *APC* and are linked to Familial Adenomatous Polyposis (FAP). However, after clinical screening, 20% to 30% of individuals diagnosed with FAP do not carry a pathogenic variant in *APC*. Other known inherited polyposis syndromes such as *MUTYH*, *STK11*, *POLD1/E*, or *NTHL1*-associated polyposis only account for a fraction of the remaining cases.

This leaves a large percentage of clinically diagnosed FAP patients without a clear genetic cause. These cases can be categorised as "Familial Polyposis Syndromes (FPS)" who present with colonic polyposis but do not carry any deleterious change in a gene associated with this condition.

A cohort of 48 individuals clinically diagnosed with familial polyposis was selected based on a strong family history of colorectal cancer (CRC) and no pathogenic variant found in *APC* and/or *MUTYH* as a result of genetic screening.

Using whole exome sequencing, FPS patients were found to carry pathogenic variants in MUTYH, APC, RAD50, POLE, NTHL1 and TP53, as well as DNA-repair genes and inflammation related genes. Additionally, a comprehensive assessment of copy number variation (CNV) revealed two loci of interest that were associated with polyposis risk.

Introduction

Familial Adenomatous Polyposis (FAP) is the second most frequently diagnosed inherited colorectal cancer syndrome, representing slightly less than 1% of all colorectal cancers (CRC) diagnosed annually (107). FAP is

primarily caused by the presence of pathogenic variants in *APC* (108), which commonly results in colorectal polyp counts of over a thousand and CRC by forty years of age, if preventative measures are not taken. However, approximately 20% of all clinically diagnosed FAP patients do not exhibit any pathogenic variant in *APC* (109, 110).

Many other inherited polyposis syndromes (both adenomatous and hamartomatous) exist that include *MUTYH* associated polyposis (MAP) (111), Peutz-Jeghers syndrome (PJS) (112), *NTHL1* associated polyposis (54), Polymerase-Proofreading associated polyposis (PPAP) (55) and many others. Collectively they only account for a relatively small proportion of the remaining clinically diagnosed FAP population.

A significant portion of individuals with suspected FAP (with a family history of cancer and confirmed polyposis), do not carry any pathogenic variant in the genes commonly screened for genetic risk. Including non-polyposis CRC, it is estimated that somewhere between 13% and 15 % of CRC are "familial", which suggests there are hereditary components that remains to be identified (107).

After accounting for all the patients that have a genetic diagnosis there remains a large population of individuals with a higher risk of polyposis that have no obvious molecular diagnosis. We will herein refer to these patients with familial polyposis of unknown origin as FAP-like patients and to the polyposis syndromes with no known aetiology as Familial Polyposis Syndrome (FPS).

Identification of the causative genes in FPS is crucial for the accurate diagnosis of FAP-like patients so that their risk of disease is reduced by offering them regular monitoring and/or prophylactic measures to minimise the risk of presenting with late stage (or incurable) disease.

To study the genetic background of FAP-like patients, we recruited a cohort of 48 patients with either a strong history of colorectal cancer or young age of disease onset, a clinical diagnosis of FAP and no pathogenic variants identified in *APC* and/or *MUTYH*. We performed whole exome sequencing on this cohort to identify the presence of pathogenic variants and their copy-number status.

Methods

Cohort selection and inclusion criteria

The 48 samples from the cohort were selected based on their family history of cancer (colorectal and other cancers), their confirmed polyp status, and the absence of a causative variant in APC or both APC and MUTYH, after genetic screening.

Inclusion criteria were as follow: Firstly, confirmed polyposis from histology/colonoscopy reports. The number of polyp's present was not considered; second the absence of a pathogenic variant in *APC*. Additionally, some patients were also screened for pathogenic *MUTYH* variants; third, they included patients with known history of cancer (not restricted to CRC) in their family

Three samples did not have clear pedigree information (BP, BS and BT) but were included in the cohort due to an unusually early age of diagnosis (39, 35, and 41 respectively), suggestive of a genetic basis to their disease. One sample (W) also had no family history available but was kept in the study due to a diagnosis of polyposis (over 40 adenomas).

De-identified DNA was obtained from NSW Health Pathology after genetic screening as part of the standard recommendation for their care. All individuals in this cohort were probands who were not related to one another. See supplementary Table 1 for details on the cohort.

DNA extraction

DNA was extracted from whole blood using the salt-extraction method (113). DNA was quantified using the Qubit Fluorometric Quantification (Invitrogen, USA) with the DNA high-sensitivity kit. DNA quality was assessed using either the 4200 TapeStation System (Agilent, USA) with high-sensitivity D1000 tapes or using the Bioanalyzer System.

Whole exome sequencing protocol

Paired-end library preparation was performed using the Illumina Truseq exome. DNA was sheared to ~150bp using the Bioruptor Pico followed by the recommended protocol using a single index. The final libraries were sequenced using an Illumina Nextseq 500 using 75 bp per read.

Libraries were quantified using Qbit High Sensitivity and were checked using either TapeStation on Bioanalyzer (Agilent, USA) for quality and size.

Whole exome sequencing analysis

FastQ files were generated using the Illumina platform Basespace which also demultiplexed and trimmed the adapters from the reads. Quality control was performed using FastQC. GATK best pratice workflow were followed

for analysis. In short, bam files were generated using BWA-MEM, and then duplicates were marked using Picard. The GATK germline pipeline was followed to generate an analysis-ready vcf file.

Any indels with a mean read-depth of less than 10 and SNP with a mean read depth of less than 20 were filtered out.

Variants were then annotated using both VEP (84) and ANNOVAR (114). TAPES (115) and Varsome (104) were utilised for pathogenicity assessment. Variants were selected and filtered based on their predicted pathogenicity and known biological function.

Variants predicted to be pathogenic or likely pathogenic using the ACMG\AMP criteria are reported in this manuscript. In addition, relevant variants of unknown significance that were detected in at least 2 individuals were also reported. See supplementary Table 2 for a full list of variants considered for analysis.

Copy Number Analysis

Copy number variations were predicted using bam files generated from sequencing. Two different software were used: XHMM (105) and EXCAVATOR2 (106). Regardless of the software used, each individual sample is compared to a normalised panel generated from all 48 samples.

XHMM CNV calls were selected using a phred quality score of at least 30 for the exact CNV, the start point and the end point. For EXCAVATOR2, CNV calls with a probability of more than 0.98 were kept. CNV calls chromosome number, start and end points were extracted to create a bed file for each sample. Intersection between the two bed files for each sample were calculated using Bedtools. Overlaps between the two software calls were considered to be true CNV calls.

Results

Pathogenic variants

Several variants predicted to be pathogenic were identified in genes considered to be associated with polyposis or colorectal cancer: *MUTYH (116)*, *APC (108)*, *POLE (55)*, *TP53 (117)* and *BRCA1 (118),* see Table 1. In addition, numerous pathogenic variants in genes previously linked to cancer or CRC risk-factors were present in the FAP-like patients (*CTSE (119)*, *RAD50 (120)*, *ERCC6 (121)*, *MAP3K9 (122)*, *OGG1* (123), *ERCC2 (124)* and *AXL (125)*).

While CDH23, expressing a cadherin-related protein, is an interesting and recently identified causative gene in cancer (126), it is still classified as a FLAGS genes (127) (frequently mutated in exome studies). Variants identified in it must be treated with caution. However, in our cohort, CDH23 was one of the genes with the highest mutational burden.

Interestingly, most genes harbouring pathogenic variants were related to DNA repair, especially from either Base Excision Repair (BER) or Nucleotide Excision Repair (NER). *MUTYH*, *OGG1* and *NTHL1* belong to BER whereas *ERCC6*, *ERCC2* and *POLE* belong to NER. Both *BRCA1* (with a variant found in two unrelated individuals) and *RAD50* are involved in double strand breaks, through Non-Homologous End Joining.

Additionally, the analysis of reference and alternative allele read depth revealed that most variants had a ratio of ref/alt close to 1:1, suggesting no influence from mosaicism as a mechanism of disease in this cohort.

Table 1. List of variants predicted to be pathogenic identified in FAP-like cohort. Ref=Reference allele, Alt=Alternative allele, * = Homozygous variant for this patient (heterozygous otherwise), Gene oe score = observed/expected loss of function variants. The oe score is a metrics computed by gNomad, the lower the oe score, the more a gene is predicted to be haploinsufficient. Sample: ref/alt indicate the sample affected by the variants with the associated sequencing read depth for the reference and alternative alleles

Variation	Protein	Consequence	Site	Gene Symbol	Protein Domain (InterPro)	Varsome Prediction	Gene oe score	Sample: ref/alt			
NM_001128425.1:c.1477-28G>A	N/A	N/A	intronic	MUTYH	N/A	Uncertain Significance	0.88	BE: 64/19	BF: 59/42		
NM_001128425.1:c.1437_1439delGGA	p.Glu480del	Non-Frameshift	exonic	MUTYH	N/A	Likely Pathogenic	0.88	BF: 78/78			
NM_001128425.1:c.1187G>A	p.Gly396Asp	Missense	exonic	MUTYH	MutY, C-terminal|NUDIX hydrolase domain|NUDIX hydrolase domain-like	Uncertain Significance	0.88	BR: 45/35	T: 24/26		
NM_001128425.1:c.536A>G	p.Tyr179Cys	Missense	exonic	MUTYH	DNA glycosylase|HhH-GPD domain	Uncertain Significance	0.88	BR: 92/109	T: 66/62	BT: 0/242	
NM_001128425.1:c.467G>A	p.Trp156Ter	Stop-gain	exonic	MUTYH	DNA glycosylase|HhH-GPD domain	Pathogenic	0.88	BE: 95/48			
NM_001910.4:c.1033G>A	p.Val345Met	Start-gain	exonic	CTSE	Aspartic peptidase domain|Peptidase family A1 domain	Uncertain Significance	1.12	AM: 7/8	BE: 37/10	BK: 31/27	R: 11/10
NM_002542.5:c.923G>A	p.Gly308Glu	Missense	exonic	OGG1	N/A	Uncertain Significance	0.89	P: 30/27	AM: 19/13		
NM_000038.6:c.637C>T	p.Arg213Ter	Stop-gain	exonic	APC	Adenomatous polyposis coli protein	Pathogenic	0.1	AB: 72/71			
NM_005732.4:c.2789_2792delTCAA	p.Ile930ThrfsTer9	Stop-gain	exonic	RAD50	N/A	Pathogenic	0.7	BT: 75/69			
NM_000124.4:c.422+51G>A	N/A	N/A	intronic	ERCC6	N/A	Uncertain Significance	0.63	BG: 28/36	BP: 26/24		
NM_022124.6:c.3293A>G	p.Asn1098Ser	Missense	exonic	CDH23	Cadherin|Cadherin conserved site|Cadherin-like;Cadherin|Cadherin-like	Uncertain Significance	0.38	AE:18/12	AQ: 44/33		
NM_006231.3:c.1270C>G	p.Leu424Val	Missense	exonic	POLE	DNA-directed DNA polymerase, family B, exonuclease domain|Ribonuclease H-like domain	Uncertain Significance	0.52	BE: 57/15	BK: 40/26		

NM_033141.4:c.1691-1G>A	N/A	Splicing	splicing	MAP3K9	N/A	Pathogenic	0.25	P: 11/9	
NM_000546.5:c.695T>A	p.Ile232Asn	Missense	exonic	TP53	p53, DNA-binding domain|p53-like transcription factor, DNA-binding|p53/RUNT-type transcription factor, DNA-binding domain	Likely Pathogenic	0.2	W: 5/17	
NM_007300.4:c.3979C>T	p.Gln1327Ter	Stop-gain	exonic	BRCA1	N/A	Pathogenic	0.73	BE: 31/10	BK: 33/22
NM_021913.5:c.1711+8A>G	N/A	N/A	intronic	AXL	N/A	Uncertain Significance	0.27	BE: 14/5	BK: 17/9
NM_000400.3:c.1891C>T	p.Arg631Cys	Missense	exonic	ERCC2	ATP-dependent helicase, C-terminal|P-loop containing nucleoside triphosphate hydrolase	Likely Pathogenic	0.59	W: 23/31	
NM_000400.3:c.1847G>C	p.Arg616Pro	Missense	exonic	ERCC2	ATP-dependent helicase, C-terminal|P-loop containing nucleoside triphosphate hydrolase	Likely Pathogenic	0.59	V: 18/8	

Copy-Number analysis

CNV analysis unveiled two loci with aberrant copy numbers, see Table 2. Two of them were pseudogenes and are likely to have little to no effect on phenotype. They were not reported in this manuscript. *CFHR3* harboured a 1.6 kb deletion of exon 4 in two different individuals. Another CNV affected a larger (148kb) section of the HLA locus in two individuals. Both CNVs were predicted to delete only one allele.

CFHR3 harboured CNV deletions in 8% of colonic and rectal cancers in the TCGA cases (https://www.cancer.gov/tcga). Interestingly, in the TCGA colonic and rectal cancers, *HLA-DRB5, HLA-DRB1, HLA-DRB6, HLA-DQA1* and *HLA-DQB1* were all found to be deleted together.

Table 2. Copy Number Variations detected in FAP-like cohort. DEL=deletion, DUP=duplication KB=size of the CNV in kilobase. TCGA Data, COAD = Colon Adenocarcinoma, READ = Rectum Adenocarcinoma

SAMPLE	CNV	INTERVAL hg19	KB	DISEASE/Role	GENE SYMBOL	TCGA COAD/READ CNV %
BJ, AR	DEL	chr1:196757813-196759408	1.6	Atypical Hemolytic Uremic Syndrome	CFHR3	8
BP, BD	DEL	chr6:32485473-32634433	148.96	Histocompatibility	HLA-DRB5, HLA-DRB1, HLA-DRB6,HLA-DQA1, HLA-DQB1	12.5
AV	DEL	chr3:75475601-75478380	2.78	NA / Pseudogene	FAM86DP	0
BJ	DUP	chr8:7153152-7155552	2.4	NA / Pseudogene	FAM90A20P	N/A

Discussion

Pathogenic variants

Several individuals carried pathogenic variants in genes known to cause polyps (*MUTYH*, *APC*, *POLE*, *TP53* and *BRCA1).* The presence of an *APC* variant is most likely due to a lack of sensitivity of older genetic screening methods (performed in 1998 for the sample AB) such as denaturing gradient gel electrophoresis (DGGE) or denaturing high performance liquid chromatography (DHPLC).

The identification of pathogenic variants in *POLE* and *OGG1* may explain the presence of polyps in these samples as these genes have previously been reported to be associated with polyposis. Variants in OGG1 appear to be rare and there are not many reports of variants associated with this gene.

Of interest was the identification of variants in *TP53* and *BRCA1.* In the context of Li-Fraumeni syndrome colorectal cancer with the presence of polyps has rarely been observed and it remains to be determined if indeed TP53 is unequivocally associated with polyposis. Certainly, however, TP53 is linked with the development of colorectal cancer and is considered to be essential for disease progression, suggesting that this germline change may result in CRC if other mutations are acquired that are integral to CRC development. There remains some debate about the role of BRCA1 and colorectal cancer risk. Recent studies confirm the role of *BRCA1* germline variants for heightened CRC risk, however overall risk does not exceed 1% at 50 years of age (versus 0.2% for non-carriers) suggesting it is unlikely to account for many individuals (118) .

ERCC2, *ERCC6*, *RAD50* and *OGG1* are all involved in DNA-repair related pathways. *ERCC6* and *ERCC2* are both involved in Nucleotide Excision Repair (NER). *OGG1* is part of the Base Excision Repair (BER). *RAD50*, as part of the MRN complex, plays a central role in DNA double-strand break repair. Impaired DNA repair is a well-recognised mechanism associated with cancer development and the accumulation of mutations.

Additional pathogenic variants have been identified in *CTSE*, *CDH23*, *MAP3K9* and *AXL*. Interestingly, both *AXL* and *CTSE* have been involved in elevated inflammation. CTSE modulates inflammation and disrupts autophagy and elevates Reactive Oxygen Species (ROS) most likely requiring increased BER. *AXL* inhibition has been shown to suppress the DNA damage response and sensitize cells to PARP inhibition in multiple cancers (128, 129) . AXL activation has been shown to be involved in immune evasion via BCL-1 and Twist (130) whereas loss of AXL function has also been shown to be associated with chronic inflammation and autoimmunity(131) .This outlines the importance of inflammation in cancer development.

MAP3K9 was found to be frequently mutated in melanoma metastasis (132). It was also found that germline SNPs in *MAP3K9* modulate its expression and are important for the development of pancreatic cancer (133). *MAP3K9* regulates the JNK pathway, which is involved in the regulation of inflammation in IBD (134).

Finally, *CDH23,* coding for Cadherin 23, a structural protein, was found to be associated with both familial and sporadic pituitary adenomas (126). The functional description of those variants (using directed mutagenesis in cells/organoids) could further our understanding of polyps' formation.

Copy-Number analysis

The most interesting CNV revealed in this study is the 1.6kb *CFHR3* deletion encompassing exon 4. *CHFR3* encodes complement factor H and has been linked to Atypical Hemolytic Uremic Syndrome (aHUS) (135), which affects platelets and lowers erythrocyte counts. One of the main symptoms of aHUS is enterocolitis, an inflammatory disorder of the intestinal tract, often misdiagnosed as ulcerative colitis, resulting in bloody diarrhoea. This underscores the importance of inflammation in the initiation of malignancy and particularly in the pre-malignant state of polyposis.

The second CNV of interest, a large deletion (148.6kb) affecting mainly HLA class II-related genes (*HLA-DRB5, HLA-DRB1, HLA-DRB6, AK293020, HLA-DQA1* and *HLA-DQB1*) was found in two samples. Previous studies have shown that genetic aberrations (SNPs and CNV) located on the HLA class II locus can be involved in hepatocellular carcinoma development (136). This may be the first example of this type of genomic loss to be associated with CRC.

Conclusion

Our findings show that patients clinically diagnosed with FAP carry pathogenic variants in known CRC-related genes. There is increasing evidence implicating *MUTYH*, *NTHL1*, *POLE*, and *TP53* in the genetic predisposition to CRC such that these genes should be routinely screened for FAP-like patients who do not carry any pathogenic variants in *APC*.

Several individuals remained with no identifiable cause for their polyposis. For these patients, several hypotheses can explain the inherited component of their disease. First, this could be due do methylation profiles associated with CRC risk. Second, it could be due to inherited mitochondrial diseases, leading to modified gastrointestinal manifestations (137). Since only exomes were

interrogated in this study, it cannot be ruled out that other variants residing in the genome could be associated with disease risk. Lastly, it could be an inherited lifestyle and environment that increases the risk of CRC, shared between the parents and their children. Studying these factors or a combination of these factors could explain the remaining unexplained familial polyposis syndromes.

4.2) Publication – Letter to the Editor

STATEMENT I

This is a co-author statement attesting to the candidate's contribution to the publication listed below:

I attest that Research Higher Degree candidate Alexandre Xavier contributed to the publication listed below by performing the whole-exome sequencing, the analysis of the data and the manuscript writing.

Xavier A, Scott RJ, Talseth-Palmer BA. IBD-related markers associate with the age of onset for unexplained familial polyposis patients

This statement explains the contribution of all authors in the article listed above:

Table illustrating author contribution percentage and description of contribution to the article listed above.

Author	Contribution (%)	Description of contribution to article	Signature	Date
Alexandre Xavier	65%	Performed whole exome sequencing, Performed analysis and wrote manuscript.		15/20/2019
Rodney J. Scott	10%	Critical revision of the manuscript and study supervision		17/10/2019
Bente A. Talseth-Palmer	25%	Study design, obtained funding, critical revision of the manuscript and study supervision	Bente Talseth-Palmer Digitally signed by Bente Talseth-Palmer Date: 2019.11.18 11:50:54 +11'00'	18/11/19

Alexandre Xavier

Date: 15/10/2019

Dr Lesley MacDonald-Wicks

Assistant Dean (Research Training)

Date: 20/11/19

IBD-related markers associate with the age of onset for unexplained familial polyposis patients

Alexandre Xavier[1], Rodney J. Scott[2] and Bente Talseth-Palmer[1,]

[1] University of Newcastle Hunter Medical Research Institute, Lot 1, Kookaburra Circuit

New Lambton Heights, NSW, AUS

[2] NSW Health Pathology, Molecular Genetics, John Hunter Hospital, , Newcastle, NSW, AUS

To the Editor:

Knowledge about inherited polyposis syndromes has expanded significantly over the past 20 years. In addition to the most common polyposis syndrome, Familial Adenomatous Polyposis (FAP), several other polyposis syndromes have been thoroughly described with aetiologies linked to *MUTYH*, *NTHL1*, *POLD1/E, PTEN*, *SKT11* or even the mismatch repair (MMR) genes (138). However, a significant proportion of familial polyposis cases remained unexplained after screening for known genes.

Here, we performed whole exome sequencing on 48 unrelated individuals diagnosed with colorectal cancer (CRC) associated with adenomatous polyps and having a family history of cancer. Using the public database GWAScatalog (https://www.ebi.ac.uk/gwas/api/search/downloads/studies_alternative), we explored the possible association between Inflammatory Bowel Disease (IBD) markers and the risk of CRC with associated adenomatous polyps. IBD is a well-known risk factor for CRC and none of the patients enrolled in our study were diagnosed with IBD. Given the significantly increased risk of CRC in patients with IBD a series of IBD related markers were assessed to determine if they were linked to disease.

We extracted the β-values of 31 SNPs known to confer an elevated risk for Inflammatory Bowel Disease (IBD) to establish a Polygenic Risk Score (PRS) (RSID: rs1042058, rs10781499, rs11209026, rs11465804, rs11548656, rs12103, rs1260326, rs12720356, rs1292053, rs2024092, rs2066847, rs2227564, rs2241880, rs2305480, rs2476601, rs2641348, rs3194051, rs34687326, rs34856868, rs3742130, rs3764147, rs3792109, rs3810936, rs4077515, rs4246905, rs501916, rs516246, rs6025, rs6596, rs7076156, rs9868809). We used $PRS = \sum_1^m \beta_i \times SNP_i$ where β represent the beta values for SNP *i* and SNP is the genotype for SNP *i* to construct the IBD-related PRS. PRS can help determine the cumulative effect of several alleles conferring a small risk and has been proved to have a strong predictive power in several diseases (139).

We first compared the IBD-related PRS of our diseased cohort with the PRS of “healthy” publicly genotyped individuals (n= 200) using ENSEMBL REST API. We then examined the relation between PRS and age of diagnosis in our cohort.

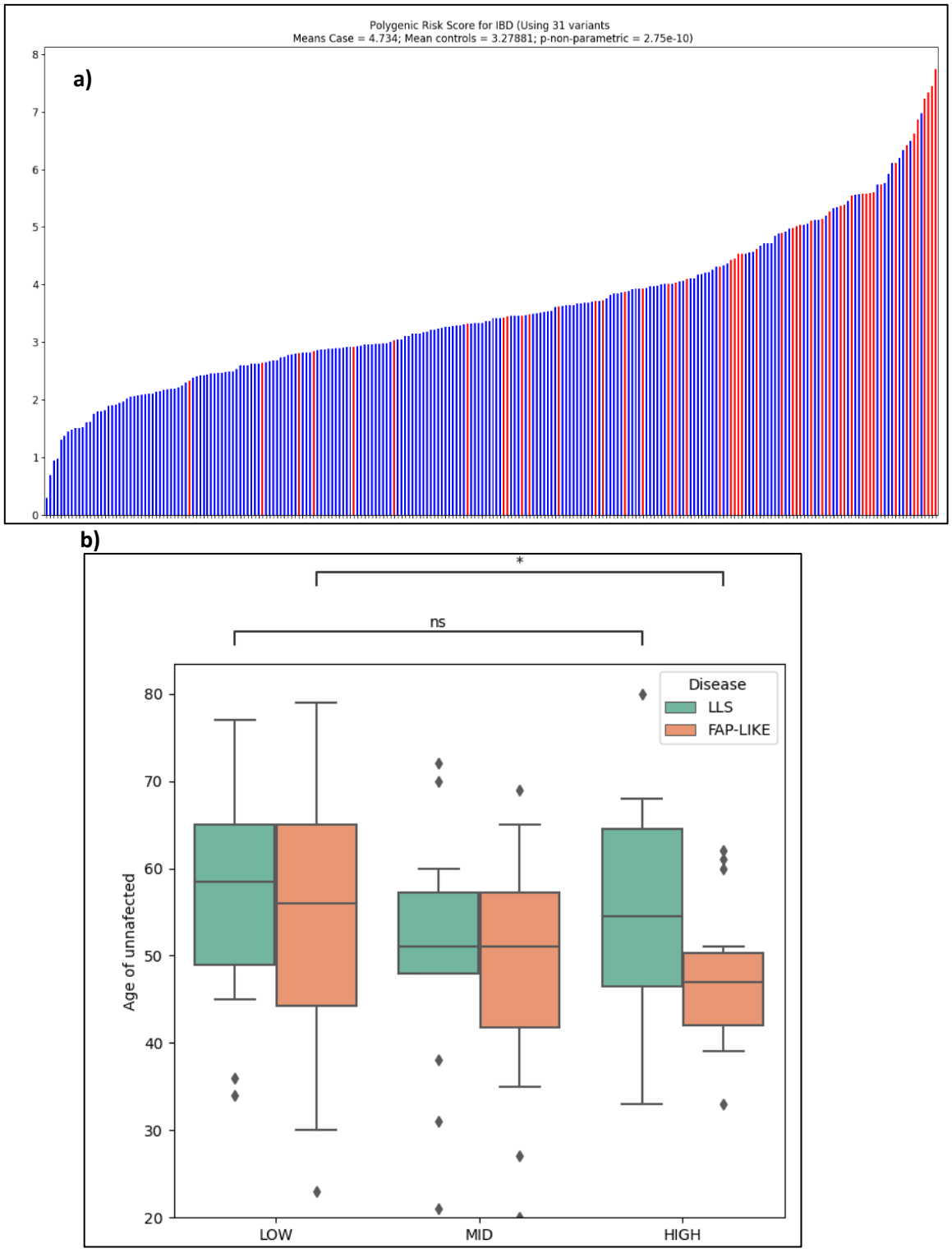

FIGURE 1. a) Ranked IBD-related PRS. Blue: healthy individuals, Red: patient with CRC associated with polyps b) Relationship between IBD-related PRS and age of onset in patient with either polyposis or non-polyposis familial CRC. Low PRS: individuals in the lowest quartile, High PRS: individuals in the highest quartile

First, this study revealed that this polyposis cohort had, on average, a significantly higher PRS compared to healthy population controls ($p = 2.75e^{-10}$). Suggesting an influence of chronic gastrointestinal inflammation in the overall risk of CRC associated with polyps.

We then showed that there is a relation between IBD-related PRS and age of CRC diagnosis. The quartile with the lowest PRS (< 4) had an average age of onset at 56.25 of age. The quartile with the highest PRS (> 5.8) had an average age of onset at 48.08 of age ($p = 0.048$).

These results suggest that the IBD-related PRS could be used, in association with other tools, to predict the trajectory of the disease with at-risk individuals. The association between IBD-related PRS and CRC could also shed a light on the mixed results of Non-steroidal anti-inflammatory drugs (NSAIDs) for the chemo-prevention of polyposis (140), suggesting that some individuals might be more likely to respond to NSAID than others.

Interestingly, performing the same analysis on 48 lynch-like syndrome (LLS) patients (patients diagnosed with hereditary CRC, no polyps and no pathogenic MMR variant) yielded different results. While still showing an elevated IBD-related PRS compared to a healthy population, there was no association between PRS and age of onset (low-PRS age of onset: 55.92, high-PRS age of onset: 52.34, $p = 0.29$). This suggests that LLS has a different aetiology compared to polyposis syndromes.

CHAPTER 5:

General discussion

CHAPTER 5: General discussion

5.1) Overview

Familial CRC remain a significant burden on national health systems since only a fraction of this type of clustering has an unequivocal genetic background. LS/HNPCC is the most frequently diagnosed syndrome accounting for approximately 5% of all CRC. Thereafter it is followed by FAP, representing around 1% of all CRC. Several other genetic syndromes predisposing to CRC have been thoroughly described but they have a very low incidence and do not account for many of the remaining familial CRC syndromes. Currently, "familial CRC" is associated with more than 10% of all CRCs diagnosed annually.

Identifying individuals with an increased risk of CRC due to a genetic predisposition is critical for a number of reasons that include; early detection, monitoring and prophylactic surgery that are necessary to reduce the risk of presenting with incurable disease.

The identification of these at-risk individuals requires knowledge of the genetic and or environmental susceptibility associated with the disease: Environmental factors (i.e. inherited lifestyle choices or environmental exposures) that increase the risk of cancer can include pathogenic changes that disrupt an important pathway; other genomic abnormalities such as an inversion of a large locus; or one or more copy number variants; the accumulation of low risk alleles resulting in an increased risk of CRC; or even an increased risk for a CRC risk-factor (such as IBD, type 1 diabetes, obesity, etc.), or a change in a locus specific methylation pattern that can be inherited from parents.

Throughout this thesis, an examination of only genetic causes such as pathogenic variants (SNPs and indels) and copy number variations in patients with a likely inherited CRC but where no known genetic cause had been identified.

The aims of this thesis were to:

I. Screen for the presence of pathogenic variants in the untested MMR genes in Lynch-like syndrome.
II. Propose a new and more refined pipeline for pathogenicity prediction using whole-exome sequencing.
III. Using the findings from Aim II, evaluate the genetic basis of familial polyposis syndromes

Detection of pathogenic variants in inherited syndromes is one of several key factors to better understand the underlying mechanisms of the disease. A study into the role of the whole MMR pathway in LLS was undertaken to assess whether genes in this pathway outside of the ones well described could be linked to LLS. An attempt was made to refine the prediction, analysis and reporting of pathogenic variants from exome sequencing studies. This allowed for a better understanding of the variants and pathways involved in the development of FAP-like syndromes.

5.2)The untested MMR genes in Lynch-Like Syndromes

This manuscript was the second part of a study examining the possible genetic causes of LLS. The first manuscript (see Appendix 7.1) focused on pathogenic variants in known or possible CRC predisposition genes.

LLS is an umbrella term grouping all patients fulfilling the ACII but where no pathogenic variant has been found in one of the key LS genes after clinical screening: *MLH1*, *MSH2*, *MSH6*, *PMS2* and *EPCAM*.

All of these key genes are related to the MMR pathway and any defect in them results in a high MSI phenotype (typical of a defective MMR), that leads to a higher probability of cancer development.

The MMR pathway is composed of 22 proteins or subunits expressed from 22 different genes; that include MLH1, MSH2, MSH, PMS2 (all associated with LS) and *MSH3*, *PMS1*, *MLH3*, *EXO1*, *POLD1*, *POLD3*, *RFC1*, *RFC2*, *RFC3*, *RFC4*, *RFC5*, *PCNA*, *LIG1*, *RPA1*, *RPA2*, *RPA3*, *POLD2 and POLD4*. Little is known about the role of the other 18 genes and their respective roles in LLS and their contribution to cancer risk.

The first finding of importance was that the older screening methods may not have been as sensitive as NGS with several LLS patients being re-classified as LS patients (24, 93) and having pathogenic variants in *MLH1*, *MSH2* and *MSH6* that were undetected during routine clinical screening. The results of this study suggest that older screening methodologies were not as sensitive as previously thought and that the errors may be more common than previously thought. These findings highlight the need to re-sequence patients with family history of CRC using newer methods. False negative diagnosis put individuals with LS and their descendants at risk, with higher risk of CRC due to the lack of monitoring.

It has also been shown that patients with LLS harboured pathogenic mutation in unscreened MMR genes: *EXO1, POLD1, RFC1, RPA1* and *MLH3*. Any defect in these genes that creates a null allele might lead to MMR malfunction, high-MSI and therefor a higher risk of cancer development. In addition, with the exception of MLH3, all of these genes and associated proteins are also involved in DNA-

related pathways other than MMR, suggesting that the resulting phenotype is likely to be broader than that described for LS.

The involvement of *POLD1 (141)*, *EXO1 (142)* and *MLH3 (143)* in LSS was previously described. Our results adds to the body of knowledge on LLS pathogenic variants and provides more evidence for the role of these genes in cancer predisposition screening. *RFC1* and *RPA1* pathogenic variants are novel findings that needs confirmations before it can be categorically stated that they are involved in LLS.

The role of *EXO1* in CRC predisposition has been largely debated for the past 15 to 20 years. There is molecular evidence that *EXO1* variants occur in healthy individuals without any adverse effects (144). It also has been shown that deletion of *EXO1* in eukaryotes will not stop the MMR pathways, but rather lead to a small increase in mutation rates (from 2 to 4 fold increase) (145, 146). However, many statistical studies show that a few key *EXO1* variants are significantly associated with CRC (147, 148) and even breast cancer (149) . These findings indicate that *EXO1* can act as a modifier of CRC risk rather than a highly penetrant causative gene. Deletion or null variants do not seem to be the mechanisms involved in modulating the risk of CRC.

While all variants identified in these studies are predicted to be pathogenic, functional analysis needs to be undertaken to fully characterise their functional pathogenicity. The outcome of these type of findings (using site directed mutagenesis in cell lines and/or organoids) could provide valuable information on the functional changes conferred by these variants.

5.3)FAP-like cohort analysis

Manuscript findings

After the examination of non-polyposis inherited CRC, we then investigated the genetic background of polyposis entities.

Using TAPES, 48 samples from a FAP-like cohort were analysed. Several pathogenic variants were identified in genes known to be involved in polyposis (*MUTYH (116)*, *APC (108)*, *POLE (55)*, *NTHL1 (54)*, *TP53 (117)* and *BRCA1 (118)*).

All the samples in this study were diagnosed and screened between 1998 and 2009. Between 2009 and 2019, clinical genetic screening has evolved. Sanger sequencing or Multiplex Ligation-dependent Probe Amplification (MLPA, used to detect copy number variation) have been replaced by either NGS or microarrays for a much higher sample throughput. This allows for the rapid screening of more genes for lower costs.

Several new causative genes have been identified in CRC especially in the polyposis syndrome spectrum with the acknowledgement of new polyposis syndromes (PPAP, NAP, JPS, etc.). At present , most commercial CRC gene-panels contain more than 30 genes (typically *APC*, *AXIN2*, *BMPR1A*, *CDH1*, *CHEK2*, *EPCAM*, *GREM1*, *MLH1*, *MSH2*, *MSH3*, *MSH6*, *MUTYH*, *NTHL1*, *PMS2*, *POLD1*, *POLE*, *PTEN*, *SMAD4*, *STK11*, *TP53*, *ATM*, *BLM*, *BUB1B*, *CEP57*, *ENG*, *FLCN*, *GALNT12*, *MLH3*, *RNF43*, *RPS20* https://www.invitae.com, https://www.fulgentgenetics.com), all of which can be screened in a single assay. Moreover, studies have shown that screening can be scaled up for inherited diseases to include all expressed genes in a single assay (i.e. whole exome sequencing), which now can provide more positive results (150).

This means that the samples in the cohort that harbour pathogenic variants in well characterised genes (*MUTYH*, *POLE*, *NTHL1*, *TP53 and BRCA1*) would have been diagnosed correctly in 2019.

However, several pathogenic variants were identified in genes known to be involved in other cancers (*CTSE (119)*, *RAD50 (120)*, *GALNT12 (151)*, *ERCC6 (121)*, *MAP3K9 (122)*, *ERCC2 (124)* and *AXL (125)*). The presence of pathogenic variants in these genes suggests the patients carrying these changes are likely to be susceptible to an increased risk of cancers compared to the general population who do not carry pathogenic variants in these genes and that the actual cancer risk is likely to encompass more than just CRC.

The presence of DNA-repair related pathogenic variants is a well-recognised mechanism of carcinogenesis. DNA replication malfunction will lead to the accumulation of mutations and chromosomal rearrangements, especially in cells with a high turnover rate such as in gastrointestinal tract. In a similar manner, pathogenic variants affecting the BER pathway will result in the accumulation of small indels and SNPs that cannot be appropriately corrected. Over time, pathogenic variants will affect either oncogenes or tumours suppressor genes thereby changing the probability of carcinogenesis.

CNVs were also revealed in FAP-like individuals with interesting results. The most intriguing CNV was the deletion encompassing *CFHR3* that was identified in two individuals. This gene is known to be involved in Atypical Haemolytic Uremic Syndrome, which leads to symptoms similar to ulcerative colitis, a known risk factor for CRC. The second CNV involved the deletion of a large locus in the HLA class II regions, spanning *HLA-DRB5, HLA-DRB1, HLA-DRB6, HLA-DQA1* and *HLA-DQB1.*

Taken together, these results seemed to indicate an important role for both DNA repair and inflammation pathways in the context of familial polyposis. To further investigate the role of

inflammation, we studied the inflammation status of FAP-like individuals using their Inflammatory Bowel Diseases (IBD) PRS as a proxy.

FAP-like individuals had an elevated PRS for IBD compared to the general population. Comparison of IBD-related PRS inside our cohort showed that a high PRS was correlated with an earlier age of cancer onset (regardless of other pathogenic variants). This suggests a strong role of inflammation as a modifier of polyposis. Using PRS on only 31 common SNPs is a powerful tool to predict the trajectory of an individual regarding polyposis.

Chronic inflammation as a risk factor for cancer is not a new concept (152). There have been reports of patients presenting with polyposis mimicking FAP as a result of IBD (153). Our results confirm this finding but also demonstrate that patients with unexplained polyposis do not need to be diagnosed with IBD. These findings are likely to be of great interest to the clinical community if they can be replicated. PRS is a simple test and combined with pathogenic variant assessment can help modulate the diagnosis as well as predict disease trajectory of patients with a strong family of cancer (154).

While interesting, these results would need larger cohorts to determine an IBD-related PRS threshold above which individuals would have a higher risk of CRC or below which they would have a lower risk. A confirmation of this trend would add to the corpus of evidence that anti-inflammatory drug could be effective in the prevention of familial polyposis on select individuals.

Our findings also suggest a possible interaction between inflammation and DNA repair mechanisms as revealed by the presence of genetic variants in *RAD50, ERCC6, ERCC2* and *OGG1* (with the addition of the variants not reported in the manuscript, discussed in the next section). This has been previously highlighted by different studies (for review see (155)) but never in inherited polyposis syndromes. Additionally, given the particularity of the gastrointestinal tract, other interactions should be considered. The most important being the gut microbiota and the immune system. Gut microbiota has been highlighted recently for its role in health and especially in cancer. It has been shown to have a role in: tumour suppression, an inflammation enhancer, an immune system modulator, and many others (for review see (156)). Immune response is also implicated in carcinogenesis. Normal appearing pre-cancerous cells appear to be eliminated by the immune system (157, 158), delaying the appearance of neoplasms, in patients with highly penetrant variants (159). Studying the synergy between chronic inflammation, DNA repair impairment and disturbed gut microbiota together could better define the events that precede and give rise to CRC with respect to the familial polyposis syndromes.

Additional Findings

Variants predicted to be pathogenic with TAPES only (with a probability of pathogenicity over 80% but not confirmed by other Software) were uncovered in additional DNA repair-related genes (see Table 8 in Appendix 7.2). Four genes, *WRN*, *LIG1* (which is also an MMR gene), *POLL* and *LIG3,* involved in the Base Excision Repair pathway (BER) were uncovered. Disruption of BER is a well-known mechanism of polyposis development. *NTHL1* and *MUTYH* both belong to the BER and are responsible for NTHL1-associated polyposis and MUTYH-associated polyposis, making pathogenic variants in these four genes candidates for tumour development.

Additionally, multiple genes involved in DNA-replication were found to harbor potentially pathogenic variants. *GINS1*, *RFC4*, *LIG1*, *PARP2*, *LIG4* and *LIG3* all had variants predicted to be pathogenic. Defects in DNA replications genes will introduce more pathogenic variants, with the risk of not being corrected by DNA-repair pathways. We also showed in Chapter 2 that LLS patients also have *LIG1* germline variants.

It is then not surprising to see that the pathway analysis (see Table 9 in appendix 7.2), showed that both “base-excision repair” (GO:0006284) and “DNA strand elongation involved in DNA replication” (GO:0006271) were among the most disrupted pathways using TAPES pathway analysis.

In addition to DNA-related pathways, pathway analysis showed that two of the most disrupted pathways were cilia-related, namely “intraciliary retrograde transport” (GO:0035721) and “protein localization to cilium” (GO:0061512). This finding is interesting because cilia have been shown to be related to the Wnt-β-catenin-APC pathway (160, 161). Cilia also has prognosis power in CRC, with the loss of primary cilia expression in cancer (161). Recent work uncovered rare disruptive variants in ciliary genes that contributes to testicular cancer susceptibility (162). This makes cilia related pathogenic germline variants a credible marker for polyposis development. However, little is known about the mechanisms underlying polyposis development with a disrupted cilium metabolism.

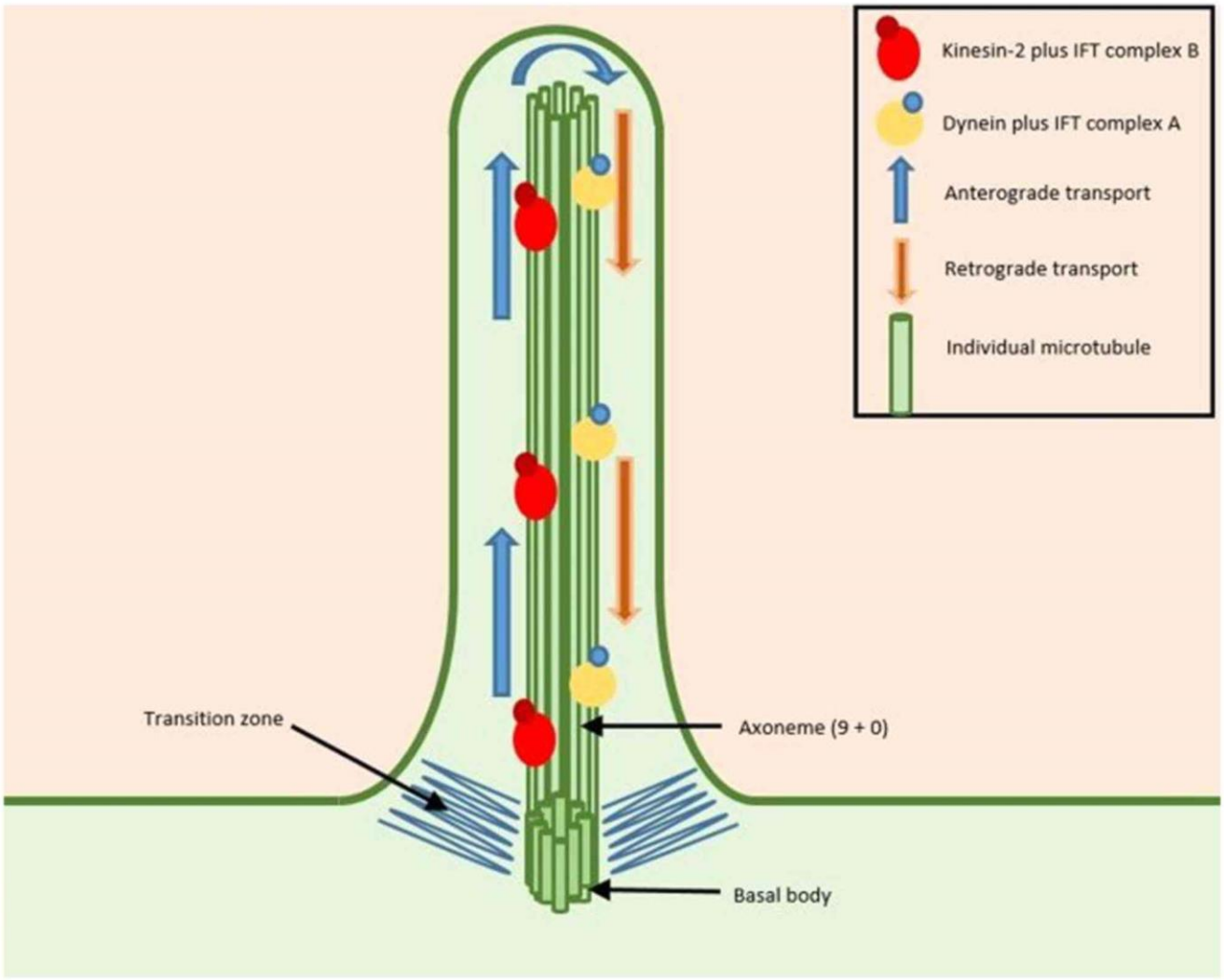

Figure 8. Cilium structure and protein transports from Higgins et al. (161)

Cilia are microtubule-related structures present on the gastrointestinal epithelium. They are known to be involved in the regulation of the Wnt signalling pathway (160). Perturbations of the Wnt pathway is also associated with FAP, where reduced levels of functional APC protein lead to the accumulation of β-catenin and the activation of several oncogenes. Several components of the Wnt signalling pathway have been shown to localise in the primary cilium, such as Frizzled3, Dishevelled2, β-catenin and glycogen synthase kinase-3β. The retrograde cilium transport pathway (see figure 8), if non-functional, will promote the accumulation of vesicles and proteins at the apex of the cilium. The interest in cilia and their involvement in tumorigenesis as well as cancer development has significantly grown lately. These findings could indicate that germline pathogenic variants in cilia-related genes increase the risk of polyposis.

The last pathway that was shown to be enriched in pathogenic variants was the “carbohydrate catabolic process” (GO:0016052). While it is well known that cancerous cells have a modified

metabolism, not much is known about carbohydrate metabolism as a cause of cancer. However, PKLR, PKFM and PGM1 (OMIM: 609712, 610681 and 171900) are known to cause inborn errors of carbohydrate metabolism in an autosomal recessive manner. Heterozygous individuals would need a second hit (somatic variant) to develop a pathology, putting them at a higher risk than homozygous wild-type individuals. Inborn errors of metabolism have also been shown to be determinant in cancer risks, especially breast and liver (163, 164).

Our additional findings suggest that DNA replication impairment and the disruption of cilia could play a role as modifier in CRC risk.

5.4) TAPES: Refining the WES analysis pipeline

During the evaluation of both LLS and FPS cohorts, assessing the pathogenicity of variants remained a major hurdle. The evaluation of variants of unknown significance was one of the main obstacle.

The automated prediction of variants' pathogenicity identified through NGS has always been a challenge. It is reflected by the number of in-silico prediction algorithms each with their own prediction score (the most popular database, dbNSFP, groups 38 different predictors: 29 prediction algorithms and 9 conservation scores).

In 2015, the publication of the ACMG/AMP criteria, proposed a set of criteria to predict the probability of pathogenicity of variants. In 2017, the company Invitae proposed Sherloc (165), a refinement on the ACMG criteria, defining even more rules to predict pathogenicity. However, currently, no free open-source tool has been published using this framework. One of the caveats of the ACMG/AMP criteria is that it is a categorical classification. It will classify variants into categories ranging from benign to pathogenic. This means that a lot of variants have been classified as "Variants of Unknown Significance" (VUS), even if they were only one criteria short of being classified as Likely Pathogenic. A second limitation resides in the tools available for ACMG\AMP criteria assignment. Most of them could not make use of multi-sample variation files (VCF). The primary evidence for a high likelihood of pathogenicity is the fact that a variant is enriched in a diseased population versus the general population (criteria PS4). Very few tools can assign these criteria (some used data from previous GWAS to assign it), implying that the study of 48 individuals with FAP-like syndromes could not be compared to the general population without adding a series of control samples.

When designing TAPES, a conversion from a categorical classification to a more organic prediction was undertaken. Multiple scoring systems were tried with a different weight for each criteria using simple addition and subtraction for pathogenic and benign criteria, respectively. This approach fell short in terms of precision. The model created by Tavtigian et al. (99) was implemented that uses a Bayesian

classification framework and this turned out to be very precise. This allows a researcher to get a very accurate prediction of pathogenicity ranging from 0 to 1 (from 0% to 100% risk to be pathogenic). Using this prediction program researchers can use their own lenient or strict pathogenicity thresholds to determine which variant to further study. This probability model is also very powerful to reject benign variants otherwise classified as VUS, allowing researchers to only focus on more significant variants.

Benchmarks showed that using the probability of pathogenicity outperformed similar tools (CharGer (87) and InterVar (88)) in pathogenicity prediction.

In addition to the scoring system, a simple calculation a simple calculation that can extrapolate the number of individuals affected and unaffected by a variant using only the minor allele frequency from public databases was developed. This allowed the use the PS4 criteria without any need for control samples. Using PS4, it is now possible to detect variants enriched in the FAP-like population studied herein, which was not possible before. This revealed numerous intronic variants, which have been growing in importance for some time (166-168). Most were not included in the final manuscript submitted to the European Journal of Human Genetics because interpreting the consequence of intronic variants is a difficult exercise without further functional studies.

One feature that was added to TAPES after publication is the ability to calculate a Polygenic Risk Score (PRS) for a specific trait or disease using public samples as controls. Using public samples from 1000genome phase 3 (169) as healthy controls, beta values are extracted from GWAS Catalog (170) for each specific trait. Using the cumulative PRS for a trait we can get the average PRS of cases vs controls. This feature was used to estimate the PRS for several CRC risk factors in our FAP-like cohort.

One of the advantages of TAPES is its reporting system, which allows some basic analysis to be done using the predicted pathogenic variants. Using the by-gene report MUTYH was one of the most frequently mutated genes in the cohort of samples studied. This means that MUTYH was mutated in multiple samples initially tested for only for *APC*. This evidence underpins the important of comprehensive gene-panel screening for FAP-like patients.

5.5) Conclusion

Taken all together, the findings from the studies comprised in this thesis indicates that there are still numerous unknown factors contributing to an increased risk of CRC. We have showed that the MMR pathway and the MMR genes not currently clinically screened can have a role in CRC development (especially *EXO1*, *POLD1*, *RFC1*, *RPA1* and *MLH3*).

Furthermore, we added to the knowledge of familial polyposis syndromes by showing both the importance of the DNA-replication and Base excision repair pathways as well as cilium related pathways in tumour development. Numerous pathogenic variants in genes known to be involved in cancer were also identified in FAP-like individuals (*CTSE*, *RAD50*, *GALNT12*, *ERCC6*, *MAP3K9*, *ERCC2* and *AXL*). Copy number analysis and polygenic risk score calculations also appeared to explain the elevated risk of polyposis in individuals with no other known genetic cause.

Throughout my PhD, I refined the analysis NGS bioinformatics pipeline to identify pathogenic variants and get more information out of whole exome sequencing data. TAPES is the result of this process. The program is more precise at detecting pathogenic variants but also at rejecting benign variants, which is important when working with big data. In addition, it can perform analysis without the need for control samples, reducing the cost of sequencing studies. The value of TAPES is in its ability to rapidly and accurately be used for the curation of variants of unknown significance in genetic predispositions to CRC. It can be rapidly applied to other genetic disorders.

REFERENCES

1. Organization WH. Global Health Estimates (GHE) 2016 [Available from: https://www.who.int/healthinfo/global_burden_disease/en/.
2. Dagenais GR, Leong DP, Rangarajan S, Lanas F, Lopez-Jaramillo P, Gupta R, et al. Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study. The Lancet. 2019.
3. Brenner H, Kloor M, Pox CP. Colorectal cancer. Lancet. 2014;383(9927):1490-502.
4. Renehan AG, Tyson M, Egger M, Heller RF, Zwahlen M. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. Lancet. 2008;371(9612):569-78.
5. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. CA Cancer J Clin. 2019;69(1):7-34.
6. Hur SJ, Yoon Y, Jo C, Jeong JY, Lee KT. Effect of Dietary Red Meat on Colorectal Cancer Risk-A Review. Compr Rev Food Sci F. 2019;18(6):1812-24.
7. Fiolet T, Srour B, Sellem L, Kesse-Guyo E, Alles B, Mejean C, et al. Consumption of ultra-processed foods and cancer risk: results from NutriNet-Sante prospective cohort. Bmj-Brit Med J. 2018;360.
8. Wirbel J, Pyl PT, Kartal E, Zych K, Kashani A, Milanese A, et al. Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nature Medicine. 2019;25(4):679-+.
9. Wei EK, Colditz GA, Giovannucci EL, Wu K, Glynn RJ, Fuchs CS, et al. A Comprehensive Model of Colorectal Cancer by Risk Factor Status and Subsite Using Data From the Nurses' Health Study. Am J Epidemiol. 2017;185(3):224-37.
10. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global patterns and trends in colorectal cancer incidence and mortality. Gut. 2017;66(4):683-91.
11. Health AIo, Welfare. Analysis of cancer outcomes and screening behaviour for national cancer screening programs in Australia. Canberra: AIHW; 2018.
12. Minoo P, Zlobec I, Peterson M, Terracciano L, Lugli A. Characterization of rectal, proximal and distal colon cancers based on clinicopathological, molecular and protein profiles. International Journal of Oncology. 2010;37(3):707-18.
13. Gryfe R. Inherited colorectal cancer syndromes. Clin Colon Rectal Surg. 2009;22(4):198-208.
14. Kravochuck SE, Kalady MF, Burke CA, Heald B, Church JM. Defining HNPCC and Lynch syndrome: what's in a name? Gut. 2014;63(9):1525-6.
15. Ligtenberg MJL, Kuiper RP, van Kessel AG, Hoogerbrugge N. EPCAM deletion carriers constitute a unique subgroup of Lynch syndrome patients. Familial Cancer. 2013;12(2):169-74.
16. NSW CI. Lynch syndrome 2019 [Available from: https://www.cancer.nsw.gov.au/learn-about-cancer/cancer-in-nsw/hereditary-cancers/lynch-syndrome.
17. Lynch HT, Shaw MW, Magnuson CW, Larsen AL, Krush AJ. Hereditary factors in cancer. Study of two large midwestern kindreds. Arch Intern Med. 1966;117(2):206-12.
18. Lynch HT, Kimberling W, Albano WA, Lynch JF, Biscone K, Schuelke GS, et al. Hereditary Nonpolyposis Colorectal-Cancer (Lynch Syndrome-I and Syndrome-Ii .1. Clinical Description of Resource. Cancer. 1985;56(4):934-8.
19. Vasen HF, Mecklin JP, Khan PM, Lynch HT. The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). Dis Colon Rectum. 1991;34(5):424-5.
20. Vasen HF, Watson P, Mecklin JP, Lynch HT. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. Gastroenterology. 1999;116(6):1453-6.
21. Rodriguez-Bigas MA, Boland CR, Hamilton SR, Henson DE, Jass JR, Khan PM, et al. A National Cancer Institute Workshop on Hereditary Nonpolyposis Colorectal Cancer Syndrome: meeting highlights and Bethesda guidelines. J Natl Cancer Inst. 1997;89(23):1758-62.

22. Umar A, Boland CR, Terdiman JP, Syngal S, de la Chapelle A, Ruschoff J, et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. J Natl Cancer Inst. 2004;96(4):261-8.
23. Syngal S, Fox EA, Eng C, Kolodner RD, Garber JE. Sensitivity and specificity of clinical criteria for hereditary non-polyposis colorectal cancer associated mutations in MSH2 and MLH1. J Med Genet. 2000;37(9):641-5.
24. 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;7(8):e850.
25. Kloor M, Staffa L, Ahadova A, von Knebel Doeberitz M. Clinical significance of microsatellite instability in colorectal cancer. Langenbecks Arch Surg. 2014;399(1):23-31.
26. Richard GF, Kerrest A, Dujon B. Comparative Genomics and Molecular Dynamics of DNA Repeats in Eukaryotes. Microbiol Mol Biol R. 2008;72(4):686-+.
27. Lander ES, Consortium IHGS, Linton LM, Birren B, Nusbaum C, Zody MC, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.
28. Viguera E, Canceill D, Ehrlich SD. Replication slippage involves DNA polymerase pausing and dissociation. EMBO J. 2001;20(10):2587-95.
29. Sorscher S. The Importance of Distinguishing Sporadic Cancers from Those Related to Cancer Predisposing Germline Mutations. Oncologist. 2018;23(11):1266-8.
30. Dominguez-Valentin M, Sampson JR, Seppala TT, Ten Broeke SW, Plazzer JP, Nakken S, et al. Cancer risks by gene, age, and gender in 6350 carriers of pathogenic mismatch repair variants: findings from the Prospective Lynch Syndrome Database. Genet Med. 2019.
31. Blount J, Prakash A. The changing landscape of Lynch syndrome due to PMS2 mutations. Clinical Genetics. 2018;94(1):61-9.
32. Dowty JG, Win AK, Buchanan DD, Lindor NM, Macrae FA, Clendenning M, et al. Cancer risks for MLH1 and MSH2 mutation carriers. Hum Mutat. 2013;34(3):490-7.
33. Burt RW. Colonic polyps in lynch syndrome. Dis Colon Rectum. 2015;58(4):371-2.
34. Sinicrope FA. Lynch Syndrome-Associated Colorectal Cancer. New Engl J Med. 2018;379(8):764-73.
35. eviQ NG. MMR genes (Lynch syndrome) – risk management 2019 [Available from: https://www.eviq.org.au/cancer-genetics/adult/risk-management/1410-mmr-genes-lynch-syndrome-risk-management.
36. Kanth P, Grimmett J, Champine M, Burt R, Samadder NJ. Hereditary Colorectal Polyposis and Cancer Syndromes: A Primer on Diagnosis and Management. Am J Gastroenterol. 2017;112(10):1509-25.
37. Giardiello FM, Allen JI, Axilbund JE, Boland CR, Burke CA, Burt RW, et al. Guidelines on genetic evaluation and management of Lynch syndrome: a consensus statement by the US Multi-Society Task Force on colorectal cancer. Gastroenterology. 2014;147(2):502-26.
38. Vasen HF, Blanco I, Aktan-Collan K, Gopie JP, Alonso A, Aretz S, et al. Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts. Gut. 2013;62(6):812-23.
39. Vasen HFA, Abdirahman M, Brohet R, Langers AMJ, Kleibeuker JH, van Kouwen M, et al. One to 2-Year Surveillance Intervals Reduce Risk of Colorectal Cancer in Families With Lynch Syndrome. Gastroenterology. 2010;138(7):2300-6.
40. Burn J, Mathers JC, Bishop DT. Chemoprevention in Lynch syndrome. Familial Cancer. 2013;12(4):707-18.
41. Burn J, Gerdes AM, Macrae F, Mecklin JP, Moeslein G, Olschwang S, et al. Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial. Lancet. 2011;378(9809):2081-7.
42. Parker TW, Neufeld KL. APC controls Wnt-induced beta-catenin destruction complex recruitment in human colonocytes. Sci Rep. 2020;10(1):2957.

43. Chiurillo MA. Role of the Wnt/beta-catenin pathway in gastric cancer: An in-depth literature review. World J Exp Med. 2015;5(2):84-102.
44. Serre L, Stoppin-Mellet V, Arnal I. Adenomatous Polyposis Coli as a Scaffold for Microtubule End-Binding Proteins. Journal of Molecular Biology. 2019;431(10):1993-2005.
45. Leslie A, Carey FA, Pratt NR, Steele RJC. The colorectal adenoma-carcinoma sequence. Brit J Surg. 2002;89(7):845-60.
46. van der Luijt RB, Meera Khan P, Vasen HF, Breukel C, Tops CM, Scott RJ, et al. Germline mutations in the 3' part of APC exon 15 do not result in truncated proteins and are associated with attenuated adenomatous polyposis coli. Hum Genet. 1996;98(6):727-34.
47. Spirio L, Olschwang S, Groden J, Robertson M, Samowitz W, Joslyn G, et al. Alleles of the APC gene: an attenuated form of familial polyposis. Cell. 1993;75(5):951-7.
48. van der Luijt RB, Vasen HF, Tops CM, Breukel C, Fodde R, Meera Khan P. APC mutation in the alternatively spliced region of exon 9 associated with late onset familial adenomatous polyposis. Hum Genet. 1995;96(6):705-10.
49. Bisgaard ML, Fenger K, Bulow S, Niebuhr E, Mohr J. Familial adenomatous polyposis (FAP): frequency, penetrance, and mutation rate. Hum Mutat. 1994;3(2):121-5.
50. Byrne RM, Tsikitis VL. Colorectal polyposis and inherited colorectal cancer syndromes. Ann Gastroenterol. 2018;31(1):24-34.
51. Colucci PM, Yale SH, Rall CJ. Colorectal polyps. Clin Med Res. 2003;1(3):261-2.
52. Al-Tassan N, Chmiel NH, Maynard J, Fleming N, Livingston AL, Williams GT, et al. Inherited variants of MYH associated with somatic G : C -> T : A mutations in colorectal tumors. Nature Genetics. 2002;30(2):227-32.
53. Lubbe SJ, Di Bernardo MC, Chandler IP, Houlston RS. Clinical Implications of the Colorectal Cancer Risk Associated With MUTYH Mutation. Journal of Clinical Oncology. 2009;27(24):3975-80.
54. Weren RD, Ligtenberg MJ, Kets CM, de Voer RM, Verwiel ET, Spruijt L, et al. A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. Nat Genet. 2015;47(6):668-71.
55. Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat Genet. 2013;45(2):136-44.
56. Church JM. Polymerase proofreading-associated polyposis: a new, dominantly inherited syndrome of hereditary colorectal cancer predisposition. Dis Colon Rectum. 2014;57(3):396-7.
57. Roberts M, Marshall ML, Webb EM, McGill AK, Susswein LR, Xu ZX, et al. Polyp burden in Lynch syndrome patients ascertained via multigene panel testing. Journal of Clinical Oncology. 2018;36(4).
58. de Leng WWJ, Jansen M, Keller JJ, de Gijsel M, Milne ANA, Morsink FHM, et al. Peutz-Jeghers syndrome polyps are polyclonal with expanded progenitor cell compartment. Gut. 2007;56(10):1475-6.
59. Cichy W, Klincewicz B, Plawski A. Juvenile polyposis syndrome. Archives of Medical Science. 2014;10(3):570-7.
60. Ngeow J, Eng C. PTEN hamartoma tumor syndrome: Clinical risk assessment and management protocol. Methods. 2015;77-78:11-9.
61. Yan HHN, Lai JCW, Ho SL, Leung WK, Law WL, Lee JFY, et al. RNF43 germline and somatic mutation in serrated neoplasia pathway and its association with BRAF mutation. Gut. 2017;66(9):1645-56.
62. Vasen HFA, Moslein G, Alonso A, Aretz S, Bernstein I, Bertario L, et al. Guidelines for the clinical management of familial adenomatous polyposis (FAP). Gut. 2008;57(5).
63. Campos FG. Surgical treatment of familial adenomatous polyposis: Dilemmas and current recommendations. World J Gastroentero. 2014;20(44):16620-9.
64. Burn J, Bishop DT, Chapman PD, Elliott F, Bertario L, Dunlop MG, et al. A randomized placebo-controlled prevention trial of aspirin and/or resistant starch in young people with familial adenomatous polyposis. Cancer Prev Res (Phila). 2011;4(5):655-65.

65. Cruz-Correa M, Hylind LM, Romans KE, Booker SV, Giardiello FM. Long-term treatment with sulindac in familial adenomatous polyposis: a prospective cohort study. Gastroenterology. 2002;122(3):641-5.
66. Lynch PM. Chemoprevention of familial adenomatous polyposis. Fam Cancer. 2016;15(3):467-75.
67. Vuolo M, Staff J. Parent and Child Cigarette Use: A Longitudinal, Multigenerational Study. Pediatrics. 2013;132(3):E568-E77.
68. Bahreynian M, Qorbani M, Khaniabadi BM, Motlagh ME, Safari O, Asayesh H, et al. Association between Obesity and Parental Weight Status in Children and Adolescents. J Clin Res Pediatr E. 2017;9(2):111-7.
69. Afkhami E, Heidari MM, Khatami M, Ghadamyari F, Dianatpour S. Detection of novel mitochondrial mutations in cytochrome C oxidase subunit 1 (COX1) in patients with familial adenomatous polyposis (FAP). Clin Transl Oncol. 2019.
70. Half E, Bercovich D, Rozen P. Familial adenomatous polyposis. Orphanet J Rare Dis. 2009;4.
71. Lindor NM. Familial colorectal cancer type X: the other half of hereditary nonpolyposis colon cancer syndrome. Surg Oncol Clin N Am. 2009;18(4):637-45.
72. Abdel-Rahman WM, Ollikainen M, Kariola R, Jarvinen HJ, Mecklin JP, Nystrom-Lahti M, et al. Comprehensive characterization of HNPCC-related colorectal cancers reveals striking molecular features in families with no germline mismatch repair gene mutations. Oncogene. 2005;24(9):1542-51.
73. Garre P, Martin L, Sanz J, Romero A, Tosar A, Bando I, et al. BRCA2 gene: a candidate for clinical testing in familial colorectal cancer type X. Clin Genet. 2015;87(6):582-7.
74. Sanchez-Tome E, Rivera B, Perea J, Pita G, Rueda D, Mercadillo F, et al. Genome-wide linkage analysis and tumoral characterization reveal heterogeneity in familial colorectal cancer type X. J Gastroenterol. 2015;50(6):657-66.
75. Mazzoni SM, Fearon ER. AXIN1 and AXIN2 variants in gastrointestinal cancers. Cancer Letters. 2014;355(1):1-8.
76. Lindor NM, Rabe K, Petersen GM, Haile R, Casey G, Baron J, et al. Lower cancer incidence in Amsterdam-I criteria families without mismatch repair deficiency: familial colorectal cancer type X. JAMA. 2005;293(16):1979-85.
77. Heinimann K, Mullhaupt B, Weber W, Attenhofer M, Scott RJ, Fried M, et al. Phenotypic differences in familial adenomatous polyposis based on APC gene mutation status. Gut. 1998;43(5):675-9.
78. Brown J, Pirrung M, McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics. 2017.
79. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
80. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-U54.
81. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2017:201178.
82. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv e-prints [Internet]. 2012 July 01, 2012. Available from: https://ui.adsabs.harvard.edu/abs/2012arXiv1207.3907G.
83. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38(16):e164.
84. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17.
85. Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 2012;6(2):80-92.

86. 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. Genetics in Medicine. 2015;17(5):405-24.
87. Scott AD, Huang KL, Weerasinghe A, Mashl RJ, Gao QS, Rodrigues FM, et al. CharGer: clinical Characterization of Germline variants. Bioinformatics. 2019;35(5):865-7.
88. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. American Journal of Human Genetics. 2017;100(2):267-80.
89. Jarvinen HJ, Aarnio M, Mustonen H, Aktan-Collan K, Aaltonen LA, Peltomaki P, et al. Controlled 15-year trial on screening for colorectal cancer in families with hereditary nonpolyposis colorectal cancer. Gastroenterology. 2000;118(5):829-34.
90. Hawkes N. Cancer survival data emphasise importance of early diagnosis. BMJ. 2019;364:l408.
91. Olkinuora A, Nieminen TT, Martensson E, Rohlin A, Ristimaki A, Koskenvuo L, et al. Biallelic germline nonsense variant of MLH3 underlies polyposis predisposition. Genetics in Medicine. 2019;21(8):1868-73.
92. Miao HK, Chen LP, Cai DP, Kong WJ, Xiao L, Lin J. MSH3 rs26279 polymorphism increases cancer risk: a meta-analysis. Int J Clin Exp Patho. 2015;8(9):11060-7.
93. Hansen MF, Johansen J, Sylvander AE, Bjornevoll I, Talseth-Palmer BA, Lavik LA, et al. Use of multigene-panel identifies pathogenic variants in several CRC-predisposing genes in patients previously tested for Lynch Syndrome. Clin Genet. 2017.
94. Vaser R, Adusumalli S, Leng SN, Sikic M, Ng PC. SIFT missense predictions for genomes. Nat Protoc. 2016;11(1):1-9.
95. Hubisz MJ, Pollard KS, Siepel A. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. 2011;12(1):41-51.
96. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248-9.
97. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014;11(4):361-2.
98. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP plus. Plos Computational Biology. 2010;6(12).
99. Tavtigian SV, Greenblatt MS, Harrison SM, Nussbaum RL, Prabhu SA, Boucher KM, et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med. 2018;20(9):1054-60.
100. 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. 2019:531210.
101. Charifa A, Jamil RT, Zhang X. Gardner Syndrome. StatPearls. Treasure Island (FL)2019.
102. Khattab A, Monga DK. Turcot Syndrome. StatPearls. Treasure Island (FL)2019.
103. Hamilton SR, Liu B, Parsons RE, Papadopolous NC, Jen J, Powell SM, et al. The Molecular-Basis of Turcot Syndrome. Gastroenterology. 1995;108(4):A478-A.
104. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Aguilera MA, Meyer R, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35(11):1978-80.
105. Fromer M, Purcell SM. Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. Curr Protoc Hum Genet. 2014;81:7.23.1-1.
106. D'Aurizio R, Pippucci T, Tattini L, Giusti B, Pellegrini M, Magi A. Enhanced copy number variants detection from whole-exome sequencing data using EXCAVATOR2. Nucleic Acids Research. 2016;44(20).
107. Hisamuddin IM, Yang VW. Genetics of colorectal cancer. MedGenMed. 2004;6(3):13-.
108. Groden J, Thliveris A, Samowitz W, Carlson M, Gelbert L, Albertsen H, et al. Identification and Characterization of the Familial Adenomatous Polyposis-Coli Gene. Cell. 1991;66(3):589-600.

109. Aaltonen LA, Peltomaki P, Leach FS, Sistonen P, Pylkkanen L, Mecklin JP, et al. Clues to the Pathogenesis of Familial Colorectal-Cancer. Science. 1993;260(5109):812-6.
110. Galiatsatos P, Foulkes WD. Familial adenomatous polyposis. American Journal of Gastroenterology. 2006;101(2):385-98.
111. Pezzi A, Roncucci L, Benatti P, Sassatelli R, Varesco L, Di Gregorio C, et al. Relative role of APC and MUTYH mutations in the pathogenesis of familial adenomatous polyposis. Scand J Gastroentero. 2009;44(9):1092-100.
112. Boardman LA, Thibodeau SN, Schaid DJ, Lindor NM, McDonnell SK, Burgart LJ, et al. Increased risk for cancer in patients with the Peutz-Jeghers syndrome. Annals of Internal Medicine. 1998;128(11):896-+.
113. Aljanabi SM, Martinez I. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. Nucleic Acids Research. 1997;25(22):4692-3.
114. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc. 2015;10(10):1556-66.
115. Xavier A, Scott RJ, Talseth-Palmer BA. TAPES: A tool for assessment and prioritisation in exome studies. PLOS Computational Biology. 2019;15(10):e1007453.
116. Sampson JR, Dolwani S, Jones S, Eccles D, Ellis A, Evans DG, et al. Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. Lancet. 2003;362(9377):39-41.
117. Rengifo-Cam W, Shepherd HM, Jasperson KW, Samadder NJ, Samowitz W, Tripp SR, et al. Colon Pathology Characteristics in Li-Fraumeni Syndrome. Clin Gastroenterol H. 2018;16(1):140-1.
118. Sopik V, Phelan C, Cybulski C, Narod SA. BRCA1 and BRCA2 mutations and the risk for colorectal cancer. Clinical Genetics. 2015;87(5):411-8.
119. Kawakubo T, Yasukochi A, Toyama T, Takahashi S, Okamoto K, Tsukuba T, et al. Repression of cathepsin E expression increases the risk of mammary carcinogenesis and links to poor prognosis in breast cancer. Carcinogenesis. 2014;35(3):714-26.
120. Fan C, Zhang J, Ouyang T, Li JF, Wang TF, Fan ZQ, et al. RAD50 germline mutations are associated with poor survival in BRCA1/2-negative breast cancer patients. International Journal of Cancer. 2018;143(8):1935-42.
121. Jing JJ, Lu YZ, Sun LP, Liu JW, Gong YH, Xu Q, et al. Epistatic SNP interaction of ERCC6 with ERCC8 and their joint protein expression contribute to gastric cancer/atrophic gastritis risk. Oncotarget. 2017;8(26):43140-52.
122. Chen J, Guo LP, Peiffer DA, Zhou LX, Chan OTM, Bibikova M, et al. Genomic profiling of 766 cancer-related genes in archived esophageal normal and carcinoma tissues. International Journal of Cancer. 2008;122(10):2249-54.
123. Benitez-Buelga C, Vaclova T, Ferreira S, Urioste M, Inglada-Perez L, Soberon N, et al. Molecular insights into the OGG1 gene, a cancer risk modifier in BRCA1 and BRCA2 mutations carriers. Oncotarget. 2016;7(18):25815-25.
124. Diaz-Gay M, Franch-Exposito S, Arnau-Collell C, Park S, Supek F, Munoz J, et al. Integrated Analysis of Germline and Tumor DNA Identifies New Candidate Genes Involved in Familial Colorectal Cancer. Cancers. 2019;11(3).
125. Abdel-Rahman WM, Al-Khayyal NA, Nair VA, Aravind SR, Saber-Ayad M. Role of AXL in invasion and drug resistance of colon and breast cancer cells and its association with p53 alterations. World J Gastroentero. 2017;23(19):3440-8.
126. Zhang QL, Peng C, Song JP, Zhang YC, Chen JH, Song ZJ, et al. Germline Mutations in CDH23, Encoding Cadherin-Related 23, Are Associated with Both Familial and Sporadic Pituitary Adenomas. American Journal of Human Genetics. 2017;100(5):817-23.
127. Shyr C, Tarailo-Graovac M, Gottlieb M, Lee JJY, van Karnebeek C, Wasserman WW. FLAGS, frequently mutated genes in public exomes. Bmc Medical Genomics. 2014;7.
128. Balaji K, Vijayaraghavan S, Diao L, Tong P, Fan Y, Carey JP, et al. AXL Inhibition Suppresses the DNA Damage Response and Sensitizes Cells to PARP Inhibition in Multiple Cancers. Mol Cancer Res. 2017;15(1):45-58.

129. Weinger JG, Brosnan CF, Loudig O, Goldberg MF, Macian F, Arnett HA, et al. Loss of the receptor tyrosine kinase Axl leads to enhanced inflammation in the CNS and delayed removal of myelin debris during Experimental Autoimmune Encephalomyelitis. J Neuroinflamm. 2011;8.
130. Gay CM, Balaji K, Byers LA. Giving AXL the axe: targeting AXL in human malignancy. Br J Cancer. 2017;116(4):415-23.
131. Bosurgi L, Bernink JH, Delgado Cuevas V, Gagliani N, Joannas L, Schmid ET, et al. Paradoxical role of the proto-oncogene Axl and Mer receptor tyrosine kinases in colon cancer. Proc Natl Acad Sci U S A. 2013;110(32):13091-6.
132. Stark MS, Woods SL, Gartside MG, Bonazzi VF, Dutton-Regester K, Aoude LG, et al. Frequent somatic mutations in MAP3K5 and MAP3K9 in metastatic melanoma identified by exome sequencing. Nature Genetics. 2012;44(2):165-9.
133. Tan AC, Fan JB, Karikari C, Bibikova M, Garcia EW, Zhou LX, et al. Allele-specific expression in the germline of patients with familial pancreatic cancer - An unbiased approach to cancer gene discovery. Cancer Biology & Therapy. 2008;7(1):137-46.
134. Roy PK, Rashid F, Bragg J, Ibdah JA. Role of the JNK signal transduction pathway in inflammatory bowel disease. World J Gastroentero. 2008;14(2):200-2.
135. Zipfel PF, Edey M, Heinen S, Jozsi M, Richter H, Misselwitz J, et al. Deletion of complement factor H-related genes CFHR1 and CFHR3 is associated with atypical hemolytic uremic syndrome. Plos Genetics. 2007;3(3):387-92.
136. Clifford RJ, Zhang JH, Meerzaman DM, Lyu MS, Hu Y, Cultraro CM, et al. Genetic Variations at Loci Involved in the Immune Response Are Risk Factors for Hepatocellular Carcinoma. Hepatology. 2010;52(6):2034-43.
137. Finsterer J, Frank M. Gastrointestinal manifestations of mitochondrial disorders: a systematic review. Ther Adv Gastroenter. 2017;10(1):142-54.
138. Kidambi TD, Kohli DR, Samadder NJ, Singh A. Hereditary Polyposis Syndromes. Curr Treat Options Gastroenterol. 2019;17(4):650-65.
139. Khera AV, Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics. 2018;50(9):1219-+.
140. Murff HJ, Shrubsole MJ, Chen Z, Smalley WE, Chen HD, Shyr Y, et al. Nonsteroidal Anti-inflammatory Drug Use and Risk of Adenomatous and Hyperplastic Polyps. Cancer Prev Res. 2011;4(11):1799-807.
141. Jansen AM, van Wezel T, van den Akker BE, Ventayol Garcia M, Ruano D, Tops CM, et al. Combined mismatch repair and POLE/POLD1 defects explain unresolved suspected Lynch syndrome cancers. Eur J Hum Genet. 2016;24(7):1089-92.
142. Wu Y, Berends MJ, Post JG, Mensink RG, Verlind E, Van Der Sluis T, et al. Germline mutations of EXO1 gene in patients with hereditary nonpolyposis colorectal cancer (HNPCC) and atypical HNPCC forms. Gastroenterology. 2001;120(7):1580-7.
143. Silva FC, Valentin MD, Ferreira Fde O, Carraro DM, Rossi BM. Mismatch repair genes in Lynch syndrome: a review. Sao Paulo Med J. 2009;127(1):46-51.
144. 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;63(1):154-8.
145. Goellner EM, Putnam CD, Kolodner RD. Exonuclease 1-dependent and independent mismatch repair. DNA Repair (Amst). 2015;32:24-32.
146. Alam NA, Gorman P, Jaeger EEM, Kelsell D, Leigh IM, Ratnavel R, et al. Germline deletions of EXO1 do not cause colorectal tumors and lesions which are null for EXO1 do not have microsatellite instability. Cancer Genetics and Cytogenetics. 2003;147(2):121-7.
147. Nasserinejad M, Pourhoseingholi MA, Rezasoltani S, Akbari S, Baghestani AR, Shojaee S, et al. Single-nucleotide polymorphism of Exo1 gene is associated with risk of colorectal cancer based on robust Bayesian approach. Gastroenterol Hepatol Bed Bench. 2018;11(Suppl 1):S146-S8.

148. Wu Y, Berends MJW, Post JG, Mensink RGJ, Verlind E, Van Der Sluis T, et al. Germline mutations of EXO1 gene in patients with hereditary nonpolyposis colorectal cancer (HNPCC) and atypical HNPCC forms. Gastroenterology. 2001;120(7):1580-7.
149. Wang HC, Chiu CF, Tsai RY, Kuo YS, Chen HS, Wang RF, et al. Association of Genetic Polymorphisms of EXO1 Gene with Risk of Breast Cancer in Taiwan. Anticancer Research. 2009;29(10):3897-901.
150. Niguidula N, Alamillo C, Shahmirzadi Mowlavi L, Powis Z, Cohen JS, Farwell Hagman KD. Clinical whole-exome sequencing results impact medical management. Mol Genet Genomic Med. 2018;6(6):1068-78.
151. Lorca V, Rueda D, Martin-Morales L, Poves C, Fernandez-Acenero MJ, Ruiz-Ponte C, et al. Role of GALNT12 in the genetic predisposition to attenuated adenomatous polyposis syndrome. Plos One. 2017;12(11).
152. Coussens LM, Werb Z. Inflammation and cancer. Nature. 2002;420(6917):860-7.
153. III GJO, Schraut WH, Peel R, Krasinskas A. Diffuse Filiform Polyposis With Unique Histology Mimicking Familial Adenomatous Polyposis in a Patient Without Inflammatory Bowel Disease. Archives of Pathology & Laboratory Medicine. 2007;131(12):1821-4.
154. Lambert SA, Abraham G, Inouye M. Towards clinical utility of polygenic risk scores. Hum Mol Genet. 2019;28(R2):R133-R42.
155. Kawanishi S, Ohnishi S, Ma N, Hiraku Y, Murata M. Crosstalk between DNA Damage and Inflammation in the Multiple Steps of Carcinogenesis. International journal of molecular sciences. 2017;18(8):1808.
156. Vivarelli S, Salemi R, Candido S, Falzone L, Santagati M, Stefani S, et al. Gut Microbiota and Cancer: From Pathogenesis to Therapy. Cancers (Basel). 2019;11(1).
157. Seppala TT, Ahadova A, Dominguez-Valentin M, Macrae F, Evans DG, Therkildsen C, et al. Lack of association between screening interval and cancer stage in Lynch syndrome may be accounted for by over-diagnosis; a prospective Lynch syndrome database report. Hered Cancer Clin Pr. 2019;17.
158. Dominguez-Valentin M, Seppala TT, Sampson JR, Macrae F, Winship I, Evans DG, et al. Survival by colon cancer stage and screening interval in Lynch syndrome: a prospective Lynch syndrome database report. Hered Cancer Clin Pr. 2019;17(1).
159. Vinay DS, Ryan EP, Pawelec G, Talib WH, Stagg J, Elkord E, et al. Immune evasion in cancer: Mechanistic basis and therapeutic strategies. Semin Cancer Biol. 2015;35 Suppl:S185-S98.
160. Fabbri L, Bost F, Mazure NM. Primary Cilium in Cancer Hallmarks. International Journal of Molecular Sciences. 2019;20(6).
161. Higgins M, Obaidi I, McMorrow T. Primary cilia and their role in cancer. Oncol Lett. 2019;17(3):3041-7.
162. Litchfield K, Levy M, Dudakia D, Proszek P, Shipley C, Basten S, et al. Rare disruptive mutations in ciliary function genes contribute to testicular cancer susceptibility. Nat Commun. 2016;7:13840.
163. da Silva I, da Costa Vieira R, Stella C, Loturco E, Carvalho AL, Veo C, et al. Inborn-like errors of metabolism are determinants of breast cancer risk, clinical response and survival: a study of human biochemical individuality. Oncotarget. 2018;9(60):31664-81.
164. Erez A, Shchelochkov OA, Plon SE, Scaglia F, Lee B. Insights into the Pathogenesis and Treatment of Cancer from Inborn Errors of Metabolism. American Journal of Human Genetics. 2011;88(4):402-21.
165. Nykamp K, Anderson M, Powers M, Garcia J, Herrera B, Ho YY, et al. Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. Genetics in Medicine. 2017;19(10):1105-17.
166. Liu X, Sinn HP, Ulmer HU, Scott RJ, Hamann U. Intronic TP53 Germline Sequence Variants Modify the Risk in German Breast/Ovarian Cancer Families. Hered Cancer Clin Pract. 2004;2(3):139-45.

167. Gatalica Z, Lilleberg SL, Vranic S, Eyzaguirre E, Orihuela E, Velagaleti G. Novel intronic germline FLCN gene mutation in a patient with multiple ipsilateral renal neoplasms. Hum Pathol. 2009;40(12):1813-9.
168. Montalban G, Bonache S, Moles-Fernandez A, Gisbert-Beamud A, Tenes A, Bach V, et al. Screening of BRCA1/2 deep intronic regions by targeted gene sequencing identifies the first germline BRCA1 variant causing pseudoexon activation in a patient with breast/ovarian cancer. J Med Genet. 2019;56(2):63-74.
169. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
170. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47(D1):D1005-D12.

APPENDICES

7.1) Additional Publication

Received: 16 December 2016 | Revised: 6 February 2017 | Accepted: 8 February 2017

DOI: 10.1111/cge.12994

WILEY CLINICAL GENETICS

ORIGINAL ARTICLE

Use of multigene-panel identifies pathogenic variants in several CRC-predisposing genes in patients previously tested for Lynch Syndrome

Maren F. Hansen[1,2] | Jostein Johansen[3] | Anna E. Sylvander[2] | Inga Bjørnevoll[2] | Bente A. Talseth-Palmer[1,4,5] | Liss A. S. Lavik[2] | Alexandre Xavier[4] | Lars F. Engebretsen[2] | Rodney J. Scott[4,6] | Finn Drabløs[3] | Wenche Sjursen[1,2]

[1]Department of Laboratory Medicine, Children's and Women's Health, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[2]Department of Pathology and Medical Genetics, St. Olavs University Hospital, Trondheim, Norway

[3]Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

[4]School of Biomedical Science and Pharmacy, University of Newcastle and Hunter Medical Research Institute, Newcastle, Australia

[5]Clinic for Medicine, Møre and Romsdal Hospital Trust, Molde, Norway

[6]Division of Molecular Medicine Pathology North, NSW Pathology, Newcastle, Australia

Correspondence
Wenche Sjursen, Department of Laboratory Medicine, Children's and Women's Health, Norwegian University of Science and Technology (NTNU), Erling Skjalgssons gt 1, 7006 Trondheim, Norway.
Email: wenche.sjursen@ntnu.no

Funding information
Central Norway Regional Health Authority (RHA); Norwegian University of Science and Technology (NTNU)

Background: Many families with a high burden of colorectal cancer fulfil the clinical criteria for Lynch Syndrome. However, in about half of these families, no germline mutation in the mismatch repair genes known to be associated with this disease can be identified. The aim of this study was to find the genetic cause for the increased colorectal cancer risk in these unsolved cases.

Materials and methods: To reach the aim, we designed a gene panel targeting 112 previously known or candidate colorectal cancer susceptibility genes to screen 274 patient samples for mutations. Mutations were validated by Sanger sequencing and, where possible, segregation analysis was performed.

Results: We identified 73 interesting variants, of whom 17 were pathogenic and 19 were variants of unknown clinical significance in well-established cancer susceptibility genes. In addition, 37 potentially pathogenic variants in candidate colorectal cancer susceptibility genes were detected.

Conclusion: In conclusion, we found a promising DNA variant in more than 25 % of the patients, which shows that gene panel testing is a more effective method to identify germline variants in CRC patients compared to a single gene approach.

KEYWORDS

colorectal cancer, diagnostics, gene panel testing, inherited cancer, Lynch syndrome, next generation sequencing (NGS)

1 | INTRODUCTION

Colorectal cancer (CRC) is one of the most common cancers in the world with approximately 1.3 million new cases diagnosed each year, and is a significant cause of cancer mortality.[1] Inherited factors are estimated to be involved in the development of one third of CRC cases. However, Mendelian CRC syndromes only explain about 5% of these cases.[2] These syndromes are caused by mutations or epimutations in well-known cancer susceptibility genes that include *MLH1*, *PMS2*, *MSH2*, *MSH6*, *EPCAM*, *APC*, *SMAD4*, *BMPR1A*, *STK11*, *MUTYH*, *PTEN*, *KLLN*, *PIK3CA*, *AKT1*, *POLE*, *POLD1*, *AXIN2*, *BUB1* and *BUB3*. Mutations in high penetrance genes such as *TP53* and *CDH1* resulting in other cancer aggregations reveals ambiguous results in terms of their association with colorectal cancer risk.[3,4] Four other genes, *ATM*, *CHEK2*, *MLH3*, and *EXO1* (all associated with some aspect of DNA repair), have been implicated in CRC susceptibility.[5–8]

ATM and *CHEK2* are increasingly being recognised as moderate penetrance genes primarily associated with an increased risk of breast cancer, but they have also been associated with CRC.[5,7] The involvement of *MLH3* and *EXO1* in CRC is still disputed and if any effect at all, they are more likely to modify the risk of other high penetrant genes.[6,8] Previous low-throughput sequencing studies aimed at investigating genes potentially involved in CRC susceptibility have identified candidates like *GALNT12* and *PTPRJ*.[9,10] However, these studies have not been replicated in additional independent cohorts and these genes require further validation before being included in the clinical management of CRC patients.

CRC is also considered as a complex disease, and low penetrant variants together with environmental factors are likely to be associated with the missing heritability apparent for the disease. Genome-wide association studies (GWASs) have identified at least 31 common low-penetrant genetic variants associated with CRC susceptibility (reviewed in[11]). One GWAS has revealed that common variants in *BMP4* influence CRC risk[12] which has been supported by a study that has potentially identified pathogenic germline mutations in *BMP4* in early onset CRC patients with a family history of cancer.[13] It is therefore possible that rare coding variants in genes identified by GWAS can cause hereditary CRC.

Recent advances in sequencing technology have aided a high-throughput approach in the search for new genes involved in hereditary CRC. Four recent exome sequencing studies have identified several potential predisposition alleles.[14–17] However, these studies only implicate potential candidates and require verification before these genes can be considered bone fide hereditary colorectal cancer genes.

In some families there is a clustering of CRC, which is suggestive of a hereditary predisposition. These families typically fulfil the Amsterdam I/II criteria (AM I/II) and/or the revised Bethesda guidelines (RBG), which were devised to help identify patients with Lynch Syndrome (LS) (MIM #609310, #120435, #614350, #614337)[18,19] In this study, we included 274 patients who fulfilled the AM I/II criteria and/or the RBG. The patients had previously been referred for clinical genetic testing of 1 or more of the MMR genes (*MLH1*, *PMS2*, *MSH2*, *MSH6*), but no germline mutations were identified. The aim of this study was to find the genetic cause for the increased CRC risk in these unsolved cases, by using a gene-panel targeting 112 previously known or candidate CRC susceptibility genes.

2 | MATERIALS AND METHODS

2.1 | Samples

This study included DNA samples from 274 (82 Norwegian and 192 - Australian) familial CRC patients. Some of the individuals were related and altogether there were 8 families with 2 to 3 family members each (19 individuals). All patients fulfilled AMI/II and/or RBG and had previously been screened for mutations in 1 or more of the MMR genes (*MLH1*, *PMS2*, *MSH2* and *MSH6*) without any pathogenic findings (80 of the Norwegian samples were also screened by MLPA). Some patients were also tested for other CRC-susceptibility genes, again without any pathogenic germline mutations being identified. Table 1 shows the clinical characteristics of the patients included in the study. The Norwegian samples were screened for mutations as part of their standard patient healthcare, and all genetic testing was performed only after written informed consent from the participants. The Australian patients included in the study had previously given informed consent for their de-identified DNA and clinical records to be used in research related to their condition. Ethics approval was obtained from the Hunter New England Human Research Ethics Committee and the University of Newcastle's Human Research Ethics Committee. DNA was isolated from EDTA-preserved whole blood using iPrep PureLink gDNA Blood kit (Thermo Fisher Scientific, Waltham, Massachusetts) (Norwegian samples) or the salt precipitation method[20] (Australian samples).

2.2 | Gene panel sequencing

We designed a custom HaloPlex (Agilent Technologies, Santa Clara, California) gene panel targeting 112 genes (Table S1, Supporting information) including both well-known CRC genes and candidate CRC susceptibility genes. The design was generated using the webtool SureDesign (Agilent Technologies). Target enrichment was performed according to manufacturer's protocol. Briefly, the samples were quantified on Qubit 2.0 Fluorometer (Life Technologies, Carlsbad, California) using dsDNA BR Assay Kit (Life Technologies). DNA was fragmented by restriction digestion, hybridized to HaloPlex probes containing indexes and purified using magnetic beads. Fragments were then ligated and amplified through 18 PCR cycles. Each library was quantified on Agilent 2100 Bioanalyzer (Agilent Technologies) using the High Sensitivity DNA kit (Agilent Technologies) and finally equimolarly pooled into sequencing ready libraries. The Norwegian samples were sequenced using an Illumina HiSeq 2500 using HiSeq Rapid SBS kit v2 (200 cycles) (Illumina, San Diego, CA). The Australian samples were sequenced on a NextSeq (Illumina) using NextSeq 500 High Output Kit (300 cycles).

2.3 | Data analysis

Analysis of sequencing data was performed as previously described [21], with only minor variation. PCR duplicates were not removed from these datasets due to the use of restriction enzymes in the HaloPlex library preparation, leading to non-random fragmentation. Removing PCR duplicates at this step can lead to removal of ~90% of reads.[22] The variant caller used was HaplotypeCaller. For filtering variants we used the filtering tool FILTUS version 0.99-91.[23]

2.4 | Filtering of variants

Our aim was to detect potentially pathogenic variants and therefore our filtering strategy aimed at removing neutral variants and sequencing errors. First, we selected variants tagged as 'PASS' after quality control, present in 1000 Genomes Project with MAF <0.01 and with a sequencing depth >10. To remove systematic sequencing errors and variants common in the patients included in this study, we excluded all variants detected in ≥10 individuals in these datasets (if over 10 individuals carry a specific variant it can be regarded as

TABLE 1 Clinical characteristics of the patients included in this study

Nationality		Total cohort (N = 274)
	Norwegian	82
	Australian	192
Female		183
Male		91
Median age at first cancer[a]		51.5 (21-86)
Cancer history[b]		
	CRC	229
	Other cancers[c]	28
	Only adenomas	14
	Multiple primary cancers[d]	64
Amsterdam criteria		
	Positive	262
	Negative[e]	12
Microsatellite instability status[f]		
	MSS	38
	MSI-L	6
	MSI-H	27
IHC[g]		
	Loss of MMR protein staining	83
	Normal staining	56

Abbreviations: CRC, colorectal cancer; RBG, revised Bethesda guidelines; MSS, Microsatellite stable; MSI-L, Microsatellite instability low; MSI-H, Microsatellite instability high; MMR, mismatch repair.

[a] Data missing for 6 patients.

[b] Data missing for 3 patients.

[c] Cancer in locations other than colon and rectum.

[d] Patients with more than 1 case of cancer, regardless of location.

[e] AM negative patients were RBG positive.

[f] Only available for the Norwegian patients. Data missing for 203 patients.

[g] Data available for 68 Norwegian and 71 Australian samples. Data missing for 135 patients

common and therefore not likely to be pathogenic). Further, we included non-synonymous, splice-site and frameshift variants. The selected non-synonymous variants were located in conserved regions based on phastCons score, predicted to be at conserved sites by PhyloP and to be deleterious by SIFT, Polyphen2, LRT and MutationTaster. We also included all frameshift and splice-site variants. Following is a brief explanation of the thresholds used to define what is conserved: Annovar uses UCSC phastCons 46 species alignment to annotate variants that fall within conserved regions. It assigns a score ranging from 0 to 1000. The higher score, the more conserved. We selected all variants with any score. In addition, we used PhyloP for base level conservation scores where a score >0.95 is conserved.

The next steps in the filtering process was to review bam files to discover and remove artifacts and variant interpretation to only select variants most likely to be pathogenic. Variant interpretation was performed utilizing Alamut software (Interactive Biosoft-ware, Rouen, France) and evaluating the available literature. Detected variants were classified into 5 classes according to the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology (AMP) guidelines.[24]

2.5 | Validation and segregation analysis by Sanger sequencing

Sanger sequencing was used to confirm detected variants remaining after applying filtering steps described above and to test for detected variants in additional family members. Sanger sequencing was done as previously described.[21] The variants confirmed were submitted to Leiden Open Variation Database 3.0 (http://databases.lovd.nl/shared/genes).

3 | RESULTS

3.1 | Filtering results

The 95 Norwegian samples had a mean coverage of: 256.03. The 192 Australian samples had a mean coverage of: 320.26. This is per base coverage in the targeted sequenced regions. Prior to filtering we identified 13 783 unique variants in the 274 samples, and after in silico filtering 148 unique variants remained. Manual filtering and interpretation to remove artefacts and to select variants most likely to be causal left 92 unique variants. Validation by Sanger sequencing confirmed 73 variants. Of these, 37 were found in known CRC susceptibility genes (Tables 2 and 3). The other 36 variants were found in candidate genes, where the association to CRC is yet to be clarified (Table 4). The 19 variants not confirmed by Sanger sequencing were mostly false positive frameshift variants, due to the remaining adapter sequences. All but 1 of the patients with Sanger validated variants fulfilled the Amsterdam criteria.

3.2 | Pathogenic variants in known CRC susceptibility genes

We found 17 pathogenic variants in 21 samples (Table 2). Of these, there were 4 mono-allelic MUTYH mutation carriers and 1 mono-allelic *BLM* mutation carrier. The mono-allelic BLM mutation carrier did not fulfil the Amsterdam criteria. One patient (no. 203) was bi-allelic for MUTYH mutation (NM_001128425; c.1187G>A and c.1227_1228dup). When excluding the mono-allelic *MUTYH* and BLM mutation carriers, we found a most probable genetic explanation for the increased cancer risk in 16 (6%) of the patients' families using this multigene panel.

We identified 3 pathogenic (class 5) variants in the MMR genes *MLH1* and *MSH6* in 3 patient samples. The *MSH6* (NM_000179.2) variant, c.3261dup (p.Phe1088Leufs*5) had previously been identified in a diagnostic setting and was included as a positive control. The 2 other samples were originally classed as mutation negative for the MMR genes.

Two patients had pathogenic mutations in *ATM*, which is known to be a moderate penetrance gene that confers an increased risk of breast cancer. Both patients had a personal and family history of CRC, and 1 of the patients (no. 154) had breast cancer in the family. The *ATM* variant c.8584+2T>C (NM_000051.3) was also tested, but

TABLE 2 Pathogenic variants in well-known cancer susceptibility genes

Sample ID	Gene	Ref. seq.	DNA	Protein	dbSNP138	ExAC (ALL)	1000 genomes	Class	Affected[a]	Unaffected[a]	ClinVar ID and interpretation
154	ATM	NM_000051.3	c.8494C>T	p.Arg2832Cys	rs587779872	8.24e-06		5			127459: P/LP
34	ATM	NM_000051.3	c.8584+2T>C	p.?	rs730881326	NR		4			181899: P/LP
112	AXIN2	NM_004655.3	c.1987dup	p.Trp663Leufs*44	NR	NR		5			NR
82	BLM	NM_000057.3	c.2824-2A>T	p.?	rs745538883	1.65e-05		4			371621: LP
7	BRCA1	NM_007294.3	c.4096+3A>G	p.?	rs80358015	NR		4			RCV000048442: LP
157	BRCA2	NM_000059.3	c.4415_4418del	p.Lys1472Thrfs*6	rs748716604	NR		5			37902: P
164	BRCA2	NM_000059.3	c.2808_2811del	p.Ala938Profs*21	rs80359351	1.65e-05		5			9322: P
291	CHEK2	NM_007194.3	c.1100del	p.Thr367Metfs*15		0.0018	0.0009984	5			RCV000123265: P
116	MLH1	NM_000249.3	c.2103+1G>T	p.?	rs267607888	NR		5			RCV000075531: LP
183	MSH6	NM_000179.2	c.2079dup	p.Cys694Metfs*4	rs267608083	NR		5			RCV000210176: P
41	MSH6	NM_000179.2	c.3261 dup	p.Phe1088Leufs*5	rs748452299	0.0018		5	2 of 3	0 of 6	89364: P
135, 203[b], 230, 245	MUTYH	NM_001128425.1	c.1187G>A	p.Gly396Asp	rs36053993	0.0028	0.00239617	5			5294: P
203[b]	MUTYH	NM_001128425.1	c.1227_1228dup	p.Glu410Glyfs*43	rs587780078	0.0001		5			127831: P
186	MUTYH	NM_001128425.1	c.934-2A>G	p.?	rs77542170	0.0010	0.0029951	5			41766: LP
4, 27, 28[c]	POLE	NM_006231.3	c.1373A>T	p.Tyr458Phe	NR	NR		5			ref. 21
42	POLE	NM_006231.3	c.824A>T	p.Asp275Val	NR	NR		5	1 of 1	0 of 1	NR
33	PTEN	NM_000314.4	c.377C>T	p.Ala126Val	NR	NR		4			ref 26, 27

Abbreviations: P, pathogenic, LP, likely pathogenic; NR, not reported.

[a] Variant found in additional affected and unaffected individuals from the same family.

[b] Patient 203 has 2 pathogenic mutations in the MUTYH gene.

[c] Patient 4, 27 and 28 belong to the same family.

TABLE 3 VUS in well-known cancer susceptibility genes

Sample ID	Gene	Ref.seq.	DNA	Protein	dbSNP138	ExAC	ESP	1000 genomes	Affected[a]	Unaffected[a]	ClinVar ID and interpretation
213	APC	NM_000038.5	c.6136G>A	p.Ala2046Thr	rs770406711	1.65e-05					185089: US
256	APC	NM_000038.5	c.1139G>A	p.Arg380Gln	rs587782886	5.79e-05					143004: LB/US
45	BLM	NM_000057.3	c.2983T>C	p.Tyr995His	rs142723411	NR	0.000077				NR
175	BRCA2	NM_000059.3	c.714_716dup	p.Glu238_Ser239insArg	rs80359640	NR					126202: US
249	BUB1	NM_001278616.1	c.2458A>G	p.Arg820Gly	NR	NR					NR
127	FANCD2	NM_001018115.1	c.3269C>T	p.Ala1090Val	NR	NR					NR
73	FLCN	NM_144997.5	c.1508G>C	p.Cys503Ser	rs778904029	1.65e-05					NR
83	FLCN	NM_144997.5	c.1523A>G	p.Lys508Arg	rs199643834	0.0002	0.000308				41856: LB/US
250	MLH1	NM_000249.3	c.514G>A	p.Glu172Lys	NR	NR					RCV000075700: US
9	MSH2	NM_000251.2	c.138C>G	p.His46Gln	rs33946261	0.0003			0 out of 3	3 out of 9	90654: US
281	MSH2	NM_000251.2	c.1045C>G	p.Pro349Ala	rs267607939	9.06e-05	0.000077				90512: US
169	MSH6	NM_000179.2	c.1282A>G	p.Lys428Glu	rs761822293	8.24e-06					NR
242	PIK3CA	NM_006218.2	c.1729A>G	p.Arg577Gly	NR	NR					NR
24	PMS2	NM_000535.5	c.1004A>G	p.Asn335Ser	rs200513014	0.0003			0 out of 1		127751: US
3, 21, 37[b]	POLE	NM_006231.3	c.229C>T	p.Arg77Cys	NR	NR			1 out of 1		NR
147	POLE	NM_006231.3	c.844C>T	p.Pro282Ser	rs138207610	0.0001	0.000231	0.000399361			RCV000229770: US
172	POLE	NM_006231.3	c.4168C>T	p.Arg1390Cys	rs768504121	1.65e-05					246319: US
29	PTEN	NM_000314.4	c.-491_-486del	p.?	NR	NR					NR
45 , 74	PTEN	NM_000314.4	c.-488_-486del	p.?	NR	NR					NR

Abbreviations: US, uncertain significance; NR, not reported; LB, likely benign.

[a] Variant found in additional affected and unaffected individuals from the same family.

[b] Patient 4, 27 and 28 belong to the same family.

TABLE 4 Potential pathogenic variants in candidate CRC susceptibility genes

Sample ID	Gene	Ref. seq.	DNA	Protein	ExAC	dbSNP138	ESP	1000 genomes	Affected[a]	Unaffected[a]	ClinVar ID and interpretation
204	**AXIN1**	**NM_003502.3**	**c.497G>T**	**p.Ser166Ile**	**NR**	NR					NR
190	**BMP4**	**NM_001202.3**	**c.250C>T**	**p.Arg84Trp**	**NR**	NR					NR
174	**CCDC18**	**NM_206886.4**	**c.3662_3663del**	**p.Leu1221Glnfs*23**	**NR**	rs761268563					NR
21	DCC	NM_005215.3	c.1817C>G	p.Pro606Arg	1.647e-05	rs773588703					NR
164	DCC	NM_005215.3	c.3370C>T	p.Arg1124Cys	0.00016	rs547920182		0.00019968			NR
194	DCC	NM_005215.3	c.4028G>A	p.Arg1343His	0.00012	rs149118168	0.000308				NR
60, 131	DCLRE1A	NM_014881.3	c.412C>T	p.Arg138*	0.0028	rs41292634	0.002384	0.00199681	0 out of 1	1 out of 2	NR
113	DUSP4	NM_001394.6	c.824G>A	p.Arg275His	6.88e-05	rs372203752	0.000077				NR
66	FAM166A	NM_001001710.1	c.41C>T	p.Pro14Leu	5.06e-05	rs140737708	0.000077				NR
146	HELQ	NM_133636.2	c.2225G>T	p.Cys742Phe	8.29e-06	rs374570294	0.000077				NR
79	LAMA3	NM_198129.2	c.8693A>G	p.Asn2898Ser	7.413e-05	rs779888893					NR
213	LAMA3	NM_198129.2	c.3712dup	p.Tyr1238Leufs*3	0.0001	rs758832093					NR
276	LAMA3	NM_198129.2	c.1273+26_1273+41del		0.0003	rs751342972	0.0008				NR
223	LAMA5	NM_005560.3	c.3964G>A	p.Gly1322Ser	0.00035	rs150741810	0.000389				NR
136	LAMB4	NM_007356.2	c.2468G>A	p.Gly823Glu	NR	NR					NR
249	LAMB4	NM_007356.2	c.1525G>C	p.Asp509His	NR	NR					NR
76	LAMC1	NM_002293.3	c.2426A>G	p.Asp809Gly	NR	NR					NR
259	LAMC1	NM_002293.3	c.1088A>G	p.His363Arg	NR	NR					NR
9	MAML3	ENST00000509479.3	c.1139C>T	p.Ser380Phe	0.0003	rs200202141	0.000724	0.00019968	0 out of 1	0 out of 1	NR
14	MLH3	NM_001040108.1	c.885del	p.His296Thrfs*12	NR	NR			1 out of 2	1 out of 2	5563: P
195	MRPL3	NM_007208.3	c.506G>T	p.Gly169Val	0.00024	rs369657581	0.000384	0.00019968			NR
97	MYH11	NM_002474.2	c.4603C>T	p.Arg1535Trp	0.00012	rs143402648	0.000077	0.000199681			372423: US
149, 262	NUDT7	NM_001105663.1	c.178C>T	p.Arg60Trp	0.00021	rs199760367	0.000336	0.00019968			NR
185	**NUDT7**	**NM_001105663.1**	**c.272G>A**	**p.Arg91Gln**	**0.00012**	rs768311455					NR
276	**PICALM**	**NM_001008660.2**	**c.130T>A**	**p.Tyr44Asn**	**NR**	NR					NR
97, 167	PSPH	NM_004577.3	c.115G>A	p.Gly39Ser	0.00089	rs147077540	0.000769				NR
123	**PTPRJ**	**NM_002843.3**	**c.3878_3879del**	**p.Gln1293Leufs*28**	**NR**	NR					NR
141	PTPRJ	NM_002843.3	c.3793G>A	p.Val1265Met	2.47e-05	rs550632588		0.00019968			NR
295	**PTPRJ**	**NM_002843.3**	**c.3208C>A**	**p.Arg1070Ser**	**NR**	NR					NR
275	**PTPRJ**	**NM_002843.3**	**c.1085del**	**p.Phe362Serfs*14**	**NR**	NR					NR
110	**SLC5A9**	**NM_001011547.2**	**c.1475del**	**p.Gly492Alafs*13**	**0.00037**	rs777247762					NR

(*Continues*)

TABLE 4 (Continued)

Sample ID	Gene	Ref. seq.	DNA	Protein	ExAC	dbSNP138	ESP	1000 genomes	Affected[a]	Unaffected[a]	ClinVar ID and interpretation
189	**TLR2**	**NM_003264.3**	**c.2029C>T**	**p.Arg677Trp**	7.44e-05	**rs121917864**					6663: RF
213	TLR4	NM_138557.2	c.1543G>A	p.Gly515Ser	8.43e-05	rs199930089		0.00019968			NR
189	**TWSG1**	**NM_020648.5**	**c.583T>C**	**p.Trp195Arg**	**NR**	NR					NR
53	**UBAP2**	**NM_018449.2**	**c.2501G>A**	**p.Arg834Gln**	8.25e-06	rs777110723					NR
198	**USP6NL**	**NM_001080491.2**	**c.874C>T**	**p.Arg292Cys**	**4.34e-05**	rs749286362					NR
99	**ZFP14**	**NM_020917.2**	**c.1006G>T**	**p.Gly336Cys**	**0.00016**	rs749848475					NR

Abbreviations: CRC, colorectal cancer; NR, not reported; RF, risk factor.
Variants marked in bold are interesting candidates to be looked further into for their potential role in CRC susceptibility.
[a] Variant found in additional affected and unaffected individuals from the same family.

not found, in a maternal cousin with 3 synchronous cancers and multiple polyps. The unaffected mother of the index patient has now been tested, and did not harbour the ATM variant. Therefore, the cousin might have another predisposing genetic variant leading to his high cancer burden.

One patient diagnosed with CRC at age 65 had a frameshift mutation in *AXIN2*. This patient is deceased, but abnormal dentition was reported, consistent with Oligodontia-colorectal cancer syndrome (MIM #608615).

One patient had a mutation in *BRCA1* (no. 7) and 2 individuals in *BRCA2* (no. 157 and 164). These 3 female patients were affected with early onset CRC. Two of them (nos 7 and 164) had a family history of CRC, breast and ovarian cancer, whereas the third (no. 157) had no family history of breast or ovarian cancer.

Two unique pathogenic variants were detected in 4 patients in *POLE* (NM_006231.3). In 3 of these patients a pathogenic *POLE* mutation c.1373A>T (p.Tyr458Phe) previously reported by Hansen et al[21] was observed. These individuals are all related and belong to the previously reported family.[21] Variant c.824A>T (p.Asp275Val) was identified in individual no. 42 affected with bilateral ovarian cancer at age 37. She was included in this study because of lack of blood sample from her deceased mother. The mother was affected with endometrial cancer at age 49 and CRC at age 88, and the *POLE* variant (c.824A>T) was detected in paraffin-embedded tissue sample from her surgery. This variant is previously found as a somatic change in endometrial cancer [25], but not as a germline variant. Asp275 forms the exonuclease catalytic site of *POLE* and is involved in binding of metal ions important for exonuclease activity.

We found 1 *PTEN* (NM_000314.4) variant c.377C>T (p.Ala126Val) in a patient diagnosed with 4 metachronous tumours (CRC, clear cell renal carcinoma, thymoma and parathyroid adenoma), some of which overlap with the tumour spectrum of Cowden Syndrome (MIM #158350). CRC was the first cancer, diagnosed at 46 years of age. The *PTEN* missense variant is within a highly conserved catalytic domain, and it is reported to give rise to completely inactive protein.[26,27]

The *CHEK2* (NM_007194.3) variant (c.1100del, p.Thr367Metfs*15) was found in a patient who was diagnosed with CRC at age 37. This *CHEK2* variant is a well described, lower penetrant mutation, mainly associated with breast cancer, but also CRC and prostate cancer.[28,29]

3.3 | Variants of unknown significance (VUS) in known CRC susceptibility genes

A total of 19 variants of unknown clinical significance were detected in 21 samples in known cancer susceptibility genes, and some of these may also prove to be pathogenic (Table 3).

MLH1 variant c.514G>A (p.Glu172Lys) was found in a patient diagnosed with CRC at age 51 who has several family members affected with CRC. Residue Glu172 is highly conserved and located in the ATPase domain of *MLH1*, although not at the ATP binding site. This variant has previously been observed 3 times in the COSMIC database. Two times as a somatic change in breast and endometrial cancer and once in a cell culture from the large intestine. A *MSH6*

variant c.1282A>G (p.Lys428Glu) was found in a patient diagnosed with cancer at age 41 with a family history of CRC and uterine cancer. Lys428 is highly conserved and located in the MutS I domain. The variant has not been previously reported.

The *POLE* variant, c.229C>T (p.Arg77Cys), was identified in 3 affected individuals from the same family and in 1 obligate carrier. All 4 family members had early onset CRC and 1 had polyposis. Most of the previously identified pathogenic mutations in *POLE* are found in the DNA binding sites within the exonuclease domain.[21,30,31] *POLE* p.Arg77is conserved (up to *S. cerevisiae*), and there is a large physicochemical difference between Arg and Cys (Grantham distance 180). However, it is not located in any exonuclease domain or at an active site, thus further investigation is needed in order to decide whether it is a causative variant.

A *BUB1* (NM_001278616.1) variant c.2458A>G (p.Arg820Gly) was found in a patient affected with CRC at age 42. Residue Arg820 is highly conserved and located in the protein kinase catalytic domain of BUB1. The mutant residue potentially disturbs the domain and is predicted to abolish its function. Although, the physiochemical difference between Arg (positively charged) and Gly (no charge) is moderate (Grantham distance 125), the difference in size, hydrophobicity and charge between the wild-type and mutant residue is predicted to disturb hydrogen bonds (Cys891 and Asp932) and ionic interactions (salt bridges) (Glu819, Glu892 and Asp932) between residue 820 and these other internal residues. The loss of charge can also cause loss of interaction with other molecules.[32] The mutation is therefore likely to affect the function of the protein.

PIK3CA (NM_006218.2) VUS c.1729A>G (p.Arg577Gly) was found in a patient diagnosed with CRC at age 58 and 3 metachronous melanomas. Arg577 is highly conserved, it is predicted to be pathogenic by 6 prediction programs (PolyPhen, SIFT, MutationTaster, Align GVD, SNPs3D and UMD Predictor), and it located in the PIK domain which has been suggested to be involved in substrate presentation. As described above for the BUB1 mutation, the physicochemical difference between Arg and Gly is moderate (Grantham distance 125). However, this change is predicted to disturb ionic interactions (salt bridges) between PIK3CA residue 577 and Aspartic acid at position 395 and 578, indicating an effect on the protein's function.[32]

Two *PTEN* variants c.-491_-486del and c.-488_-486del are located in 5′ UTR (or exon 1 in transcript NM_001304717) at a binding site for RNA Polymerase II. Detecting mutations in this region in 3 unrelated Norwegian individuals suggests that these variants are common in the Norwegian population. However, because these patients are highly selected the 2 *PTEN* variants may be pathogenic if they disrupt RNA Polymerase II binding, but this needs further investigation.

The variants in Table 3 with reported minor allele frequencies are less likely to be pathogenic, except for that identified in *BLM*, which is associated with recessive disease. In addition, segregation analysis of the *MSH2* variant c.138C>G (p.His46Gln) and *PMS2* c.1004A>G (p.Asn335Ser) does not support pathogenicity. However, *PMS2* is found to have much lower penetrance for CRC than the other MMR genes, and therefore mutations may not always be associated with disease.[33] For the remaining variants listed in Table 3, there is no further information indicating whether they are pathogenic or benign.

3.4 | Variants in candidate CRC genes

We identified 37 unique variants in 36 different patients in candidate genes that have a potential role in CRC susceptibility (Table 4). There was no evidence of autosomal recessive disease identified in this dataset. Variants with a reported allele frequency are less likely to cause a highly penetrant disorder, although moderately penetrant disorders are possible but more difficult to identify. Laminins are essential components of connective tissue basement membranes and influence cell differentiation, migration, and adhesion. Laminin is vital for the maintenance and survival of tissues and defective laminins can lead to the autosomal recessive disorders such as congenital muscular dystrophy (MIM #607855), junctional epidermolysis bullosa (MIM #226700 and #226650) and Pierson Syndrome (MIM #609049).[34] We identified 8 variants in laminin genes; *LAMA3*, *LAMA5*, *LAMB4* and *LAMC1*. Based on Laminins function, these variants are not the most probable candidates to play a role in CRC susceptibility.

Segregation analysis was only possible for the variants *DCLRE1A* (NM_014881.3) c.412C>T (p.Arg138*), *MAML3* (ENST00000509479.3) c.1139C>T (p.Ser380Phe) and *MLH3* (NM_001040108.1) c.885del (p.His296Thrfs*12) due to the availability of samples from additional family members. However, none of these variants seemed to segregate with disease. The *MLH3* variant has previously been found in 2 CRC patients, 1 endometrial cancer patient and 1 unaffected below the age of 75 in a family [35], suggesting the variant to have reduced penetrance. They further suggested *MLH3* to be a low risk gene for CRC. *DCC* variant c.1817C>G (p.Pro606Arg) identified in patient no. 21 was not found in 2 affected family members (nos 3 and 37) who also were included in this study. Instead, these 3 family members all had the *POLE* VUS c.229C>T described above. Another *DCC* variant, c.3370C>T (p.Arg1124Cys), was identified in patient no. 164 who also has a pathogenic *BRCA1* mutation. Consequently, these 2 *DCC* variants are not likely to be associated with a predisposition to CRC.

The remaining 14 variants in the genes *AXIN1*, *BMP4*, *CCDC18*, *NUDT7*, *PICALM*, *PTPRJ*, *SLC5A9*, *TLR2*, *TWSG1*, *UBAP2*, *USP6NL* and *ZFP14* have a potential role in CRC susceptibility (marked bold in the table). Of these, the missense variants in *AXIN1*, *BMP4*, *NUDT7*, *PICALM*, *PTPRJ*, *TLR2*, *TWSG1*, *USP6NL and ZFP14* are located in protein functional domains and the residue (Arg91) affected in *NUDT7* is a putative active site. Four variants in *CCDC18*, *PTPRJ* and *SLC5A9* are frameshift variants. The most interesting candidates are the 2 frameshift and the missense variant (marked bold) in the *PTPRJ* gene. Epigenetic silencing of this gene due to an inherited duplication in a CRC family has previously been reported[10] suggesting that this may be a new CRC susceptibility gene. The 2 frameshift mutations are predicted to disrupt the function of this gene and the missense variant alters a highly conserved amino acid involved in 2 functional domains (PTP type protein phosphatase and protein-tyrosine phosphatase-like). All the patients with *PTPRJ* alterations in this study were diagnosed with CRC above the age of 50 years and have several family members affected with CRC. Unfortunately no samples from additional family members were available at this stage.

4 | DISCUSSION

In this study, we found several pathogenic or likely pathogenic (class 4-5) variants in known cancer susceptibility genes, which validates our approach for identifying disease causing variants. Some of the VUS's revealed in this study may also prove to be pathogenic, as more becomes known about the functional impact of these variants.

Three variants in *MLH1* and *MSH6* as well as a number of variants of unknown significance (VUS) were identified in our sample set. The most likely explanation for this finding is the accuracy of some of the screening protocols that were used to identify variants in known MMR genes. Using high-throughput screening approaches that are significantly more accurate than previous methodologies it is to be expected some additional mutations in these genes will come to light. We recommend that samples screened by methodologies that do not employ direct DNA sequencing be re-evaluated by better more cost-effective and accurate assays.

The phenotype of hereditary cancer syndromes often overlap, because of the pleiotropy of cancer genes. For example in LS a wide spectrum of cancer types are associated with mutations in MMR genes, like ovary cancer. Increased risk of ovary cancer is also associated with mutations in *BRCA1* and *BRCA2*. The spectra of cancer types associated with each cancer syndrome are not always totally determined either. Whether breast cancer is a part of the LS spectrum have been widely debated. There has also been discussed whether there is an increased risk for CRC in *BRCA* mutation carriers, and recent studies have shown that there is an increased for CRC in female *BRCA1* mutation carriers below the age of 50 years (reviewed in[36]). This makes it more difficult to choose the appropriate gene(s) to test. By using multigene panels, all relevant genes can be tested simultaneously, increasing the probability of finding a causal variant. An example in this study is patient no. 7 in which we discovered the pathogenic *BRCA1* variant c.4096+3A>G. This patient and a first degree relative were both affected with CRC and consequently this patient was, at that time, only tested for MMR genes. There was also a case of bilateral breast cancer and 2 cases of ovarian cancer in this family, but the 2 CRC cases in the index patient and her parent suggested a CRC predisposition rather than a breast ovarian cancer family.

Another advantage by using a broader gene panel testing approach is that it may reveal whether there is more than 1 pathogenic variant in a high-risk family. Mutations in different genes in 1 family may explain an untypical spectrum spectre of cancer types in a family.

For LS there are several aspects that can lead to misguided genetic testing of MMR genes. Loss of MMR gene expression may be a result of somatic inactivation mimicking that observed in LS tumours.[37] These patients do not have LS, but a mutation in another CRC-predisposing gene may be associated with their increased cancer risk. This may well be the case for many of the patients included in this study because 83 showed a lack of MMR protein staining in their tumours, 27 were MSI-High and 6 were MSI-Low. The tumours from 4 of the patients with pathogenic mutations identified in *POLE* (nos 4 and 28), *BRCA1* (no. 7) and *ATM* (no. 34), were MSI-High (nos 28 and 34) or MSI-Low (nos 4 and 7), and some had aberrant MMR expression. Nos 28 and 34 did not express MLH1 and PMS2 (no promoter methylation), no.7 did not express MSH6, while no. 4 had normal MMR staining. Tumour immunohistochemical analyses can fail to indicate LS. In previous studies we have shown that some pathogenic MMR variants do not affect protein staining or MSI.[38,39] These patients are at risk for not being tested for LS.

We identified several potentially pathogenic variants in previously proposed candidate CRC susceptibility genes thereby increasing the evidence that they are associated with disease risk. Notwithstanding, additional studies on these genes are required to unequivocally define them as CRC susceptibility genes. Although we have narrowed the list down to some interesting candidates (indicated in Table 4), we could not confirm any of the proposed candidate CRC susceptibility genes due to the absence of additional family members participating in this study. The *POLE* variant c.229C>T (p.Arg77Cys) exemplifies this point, where additional family members appeared to confirm the association. Owing to the paucity of data on what it actually means to harbour a potential causative variant in any of the genes we have identified, we do not recommend the inclusion of candidate genes in a diagnostic setting, as they would only confuse an already complex situation.

For many of the patients we did not find any genetic explanation for their increased CRC risk. The cause for CRC susceptibility in these patients may be found in non-coding regions of the genes of interest or could be explained by copy number variations, which were not addressed in this study. Alternatively, the mutational yield was not particularly high in this study suggesting that other variants are located in genes not targeted by our panel design. These unexplained cases are candidates for exome and whole-genome sequencing.

In conclusion, we have identified a most probably genetic cause for the increased risk of CRC for 17 (6%) of the patients included in this study. We have also identified some variants both in known- and candidate CRC susceptibility genes which should be the subject of further research to determine their involvement in CRC risk. Overall, the results show that gene panel sequencing is a more effective method by which to identify pathogenic germline variants in CRC patients compared with a single gene approaches.

ACKNOWLEDGMENTS

The sequencing service was provided by the Genomics Core Facility, Norwegian University of Science and Technology (NTNU). The bioinformatics analyses were performed at the Bioinformatics core facility, Norwegian University of Science and Technology (NTNU). This work was supported by grants from the Liaison Committee between the Central Norway Regional Health Authority (RHA) and the Norwegian University of Science and Technology (NTNU); and travel support for scientific meetings and courses from COST Action BM1206.

Conflict of interest

The authors declare no conflict of interest.

ORCID

Maren F. Hansen http://orcid.org/0000-0001-5231-4034
Wenche Sjursen http://orcid.org/0000-0003-3880-6440

REFERENCES

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136:E359-E386.
2. Jasperson KW, Tuohy TM, Neklason DW, Burt RW. Hereditary and familial colon cancer. *Gastroenterology*. 2010;138:2044-2058.
3. Wong P, Verselis SJ, Garber JE, et al. Prevalence of early onset colorectal cancer in 397 patients with classic Li-Fraumeni syndrome. *Gastroenterology*. 2006;130:73-79.
4. Richards FM, McKee SA, Rajpar MH, et al. Germline E-cadherin gene (CDH1) mutations predispose to familial gastric cancer and colorectal cancer. *Hum Mol Genet*. 1999;8:607-610.
5. Thompson D, Duedal S, Kirner J, et al. Cancer risks and mortality in heterozygous ATM mutation carriers. *J Natl Cancer Inst*. 2005;97:813-822.
6. Liberti SE, Rasmussen LJ. Is hEXO1 a cancer predisposing gene? *Mol Cancer Res*. 2004;2:427-432.
7. Xiang HP, Geng XP, Ge WW, Li H. Meta-analysis of CHEK2 1100delC variant and colorectal cancer susceptibility. *Eur J Cancer*. 2011;47:2546-2551.
8. Kim JC, Roh SA, Yoon YS, Kim HC, Park IJ. MLH3 and EXO1 alterations in familial colorectal cancer patients not fulfilling Amsterdam criteria. *Cancer Genet Cytogenet*. 2007;176:172-174.
9. Guda K, Moinova H, He J, et al. Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers. *Proc Natl Acad Sci U S A*. 2009;106:12921-12925.
10. Venkatachalam R, Ligtenberg MJ, Hoogerbrugge N, et al. Germline epigenetic silencing of the tumor suppressor gene PTPRJ in early-onset familial colorectal cancer. *Gastroenterology*. 2010;139:2221-2224.
11. Esteban-Jurado C, Garre P, Vila M, et al. New genes emerging for colorectal cancer predisposition. *World J Gastroenterol*. 2014;20:1961-1971.
12. Tomlinson IP, Carvajal-Carmona LG, Dobbins SE, et al. Multiple common susceptibility variants near BMP pathway loci GREM1, BMP4, and BMP2 explain part of the missing heritability of colorectal cancer. *PLoS Genet*. 2011;7:e1002105.
13. Lubbe SJ, Pittman AM, Matijssen C, et al. Evaluation of germline BMP4 mutation as a cause of colorectal cancer. *Hum Mutat*. 2011;32:E1928-E1938.
14. Gylfe AE, Katainen R, Kondelin J, et al. Eleven candidate susceptibility genes for common familial colorectal cancer. *PLoS Genet*. 2013;9:e1003876.
15. Smith CG, Naven M, Harris R, et al. Exome resequencing identifies potential tumor-suppressor genes that predispose to colorectal cancer. *Hum Mutat*. 2013;34:1026-1034.
16. DeRycke MS, Gunawardena SR, Middha S, et al. Identification of novel variants in colorectal cancer families by high-throughput exome sequencing. *Cancer Epidemiol Biomarkers Prev*. 2013;22:1239-1251.
17. Esteban-Jurado C, Vila-Casadesus M, Garre P, et al. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. *Genet Med*. 2015;17:131-142.
18. Umar A, Boland CR, Terdiman JP, et al. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J Natl Cancer Inst*. 2004;96:261-268.
19. Vasen HF, Watson P, Mecklin JP, Lynch HT. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. *Gastroenterology*. 1999;116:1453-1456.
20. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*. 1988;16:1215.
21. Hansen MF, Johansen J, Bjornevoll I, et al. A novel POLE mutation associated with cancers of colon, pancreas, ovaries and small intestine. *Fam Cancer*. 2015;14:437–448.
22. Coonrod EM, Durtschi JD, VanSant Webb C, Voelkerding KV, Kumanovics A. Next-generation sequencing of custom amplicons to improve coverage of HaloPlex multigene panels. *Biotechniques*. 2014;57:204-207.
23. Vigeland MD, Gjotterud KS, Selmer KK. FILTUS: a desktop GUI for fast and efficient detection of disease-causing variants, including a novel autozygosity detector. *Bioinformatics (Oxford, England)*. 2016;32:1592-1594.
24. Richards S, Aziz N, Bale S, 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;17:405-424.
25. Briggs S, Tomlinson I. Germline and somatic polymerase epsilon and delta mutations define a new class of hypermutated colorectal and endometrial cancers. *J Pathol*. 2013;230:148-153.
26. Rodriguez-Escudero I, Oliver MD, Andres-Pons A, Molina M, Cid VJ, Pulido R. A comprehensive functional analysis of PTEN mutations: implications in tumor- and autism-related syndromes. *Hum Mol Genet*. 2011;20:4132-4142.
27. Costa HA, Leitner MG, Sos ML, et al. Discovery and functional characterization of a neomorphic PTEN mutation. *Proc Natl Acad Sci USA*. 2015;112:13976-13981.
28. Gronwald J, Cybulski C, Piesiak W, et al. Cancer risks in first-degree relatives of CHEK2 mutation carriers: effects of mutation type and cancer site in proband. *Br J Cancer*. 2009;100:1508-1512.
29. Meijers-Heijboer H, Wijnen J, Vasen H, et al. The CHEK2 1100delC mutation identifies families with a hereditary breast and colorectal cancer phenotype. *Am J Hum Genet*. 2003;72:1308-1314.
30. Palles C, Cazier JB, Howarth KM, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat Genet*. 2013;45:136-144.
31. Rohlin A, Zagoras T, Nilsson S, et al. A mutation in POLE predisposing to a multi-tumour phenotype. *Int J Oncol*. 2014;45:77-81.
32. Venselaar H, Te Beek TA, Kuipers RK, Hekkelman ML, Vriend G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinf*. 2010;11:548.
33. Moller P, Seppala T, Bernstein I, et al. Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database. *Gut*. 2015;66:446–472.
34. Yurchenco PD, Patton BL. Developmental and pathogenic mechanisms of basement membrane assembly. *Curr Pharm Des*. 2009;15:1277-1294.
35. Liu HX, Zhou XL, Liu T, et al. The role of hMLH3 in familial colorectal cancer. *Cancer Res*. 2003;63:1894-1899.
36. Sopik V, Phelan C, Cybulski C, Narod SA. BRCA1 and BRCA2 mutations and the risk for colorectal cancer. *Clin Genet*. 2015;87:411-418.
37. Mensenkamp AR, Vogelaar IP, van Zelst-Stams WA, et al. Somatic mutations in MLH1 and MSH2 are a frequent cause of mismatch-repair deficiency in Lynch syndrome-like tumors. *Gastroenterology*. 2014;146:643-646.e648.
38. Grindedal EM, Aarset H, Bjornevoll I, et al. The Norwegian PMS2 founder mutation c.989-1G > T shows high penetrance of microsatellite instable cancers with normal immunohistochemistry. *Hered Cancer Clin Pract*. 2014;12:12.
39. Sjursen W, Haukanes BI, Grindedal EM, et al. Current clinical criteria for Lynch syndrome are not sensitive enough to identify MSH6 mutation carriers. *J Med Genet*. 2010;47:579-585.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Hansen MF, Johansen J, Sylvander AE, et al. Use of multigene-panel identifies pathogenic variants in several CRC-predisposing genes in patients previously tested for Lynch Syndrome. *Clin Genet*. 2017;92:405–414. https://doi.org/10.1111/cge.12994

7.2) Additional Tables from the familial polyposis syndrome study

Table 9. List of additional variants predicted to be pathogenic identified in FAP-like cohort (not included in the above manuscript). Ref=Reference allele, Alt=Alternative allele, VUS=Variant of Unknown Significance

Variation	Protein	Consequence	Site	Gene Symbol	Protein Domain	TAPES Prediction	Sample
NM_015967.7:c.727A>G	p.Thr243Ala	Missense	exonic	PTPN22	Protein-tyrosine phosphatase-like;PTP type protein phosphatase	VUS	AV
NM_021133.4:c.471_474delAAAG	p.Lys158Argfs Ter6	Stop-gain	exonic	RNASEL	N/A	Likely Pathogenic	AR
NM_000179.2:c.3260C>A	p.Pro1087His	Missense	exonic	MSH6	DNA mismatch repair protein MutS, clamp|DNA mismatch repair protein MutS, core;DNA mismatch repair protein MutS, core;DNA mismatch repair protein MutS, core|P-loop containing nucleoside triphosphate hydrolase	Likely Pathogenic	BV
ENST00000302036.7:c.208G>A	p.Glu70Lys	Missense	exonic	OGG1	8-oxoguanine DNA glycosylase, N-terminal	VUS	BN
NM_002916.4:c.866T>C	p.Leu289Pro	Missense	exonic	RFC4	DNA polymerase III, clamp loader complex, gamma/delta/delta subunit, C-terminal|Replication factor C, C-terminal	VUS	R
NM_001163213.1:c.1321C>T	p.Arg441Cys	Missense	exonic	FGFR3	Immunoglobulin subtype|Immunoglobulin subtype 2|Immunoglobulin-like domain|Immunoglobulin-like fold	VUS	BF
NM_001166108.2:c.1394G>A	p.Arg465His	Missense	exonic	PALLD	Immunoglobulin I-set|Immunoglobulin subtype|Immunoglobulin subtype 2|Immunoglobulin-like domain|Immunoglobulin-like fold;Immunoglobulin I-set|Immunoglobulin-like domain|Immunoglobulin-like fold;Immunoglobulin-like fold	Likely Pathogenic	BV
NM_182925.5:c.3214G>A	p.Gly1072Ser	Missense	exonic	FLT4	Protein kinase domain|Protein kinase-like domain|Serine-threonine/tyrosine-protein kinase, catalytic domain|Tyrosine-protein kinase, catalytic domain	Likely Pathogenic	M

NM_002944.2:c.4892A>G	p.Tyr1631Cys	Missense	exonic	ROS1	Fibronectin type III|Immunoglobulin-like fold	Likely Pathogenic	R
NM_000553.5:c.1717A>G	p.Thr573Ala	Missense	exonic	WRN	DEAD/DEAH box helicase domain|Helicase superfamily 1/2, ATP-binding domain	Likely Pathogenic	AS
NM_024642.5:c.796G>A	p.Glu266Lys	Missense	exonic	GALNT12	Glycosyltransferase 2-like|Nucleotide-diphospho-sugar transferases	VUS	BV
NM_022124.6:c.5647A>C	p.Asn1883His	Missense	exonic	CDH23	Cadherin|Cadherin-like	VUS	BD
NM_022124.6:c.9932C>T	p.Ser3311Leu	Missense	exonic	CDH23	Cadherin|Cadherin-like	VUS	AC
NM_001174084.2:c.1090C>T	p.Arg364Cys	Missense	exonic	POLL	DNA polymerase lambda, fingers domain|DNA-directed DNA polymerase X	VUS	BN
NM_206937.2:c.560T>C	p.Ile187Thr	Missense	exonic	LIG4	DNA ligase, ATP-dependent, N-terminal	VUS	BG
NM_005484.3:c.709C>T	p.Arg237Trp	Missense	exonic	PARP2	Poly(ADP-ribose) polymerase, regulatory domain	VUS	BE
NM_013975.4:c.2078T>A	p.Val693Glu	Missense	exonic	LIG3	DNA ligase, ATP-dependent, central|Nucleic acid-binding, OB-fold	VUS	AX
NM_000234.3:c.1003C>T	p.Leu335Phe	Missense	exonic	LIG1	DNA ligase, ATP-dependent, N-terminal	VUS	BX
NM_021067.5:c.247C>T	p.Arg83Cys	Missense	exonic	GINS1	GINS subunit, domain A	Pathogenic	BQ

Table 10. List of Pathways significantly enriched in pathogenic variants. Using the GO Biological Process library. Ranked by adjusted p-value.

Name	*P-value*	*Z-score*	*Combined score*	*Genes*	*Adjusted p-value*
intraciliary retrograde transport (GO:0035721)	*2.26E-07*	*-2.68077347*	*41.01726813*	*['ICK', 'DYNC2LI1', 'IFT43', 'TTC21B', 'IFT122', 'TTC21A', 'WDR35']*	*0.00074313*
DNA strand elongation involved in DNA replication (GO:0006271)	*1.02E-05*	*-2.485462*	*28.56344354*	*['GINS1', 'RFC4', 'LIG1', 'PARP2', 'LIG4', 'LIG3', 'POLE']*	*0.01096476*
base-excision repair (GO:0006284)	*1.34E-05*	*-1.892174*	*21.23526286*	*['WRN', 'LIG1', 'NTHL1', 'OGG1', 'POLL', 'LIG3', 'ERCC6', 'POLE', 'TP53', 'MUTYH']*	*0.01096476*
protein localization to cilium (GO:0061512)	*1.02E-05*	*-1.623930*	*18.66014461*	*['TUB', 'ARL6', 'TTC21B', 'IFT122', 'TULP3', 'TTC21A', 'TULP1', 'WDR35']*	*0.01096476*
carbohydrate catabolic process (GO:0016052)	*5.64E-05*	*-2.142826*	*20.96336459*	*['HK3', 'PKLR', 'MAN2B2', 'NAGA', 'MAN2C1', 'PGK2', 'ENO2', 'PFKM', 'PGM1']*	*0.0370096*