UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS QUÍMICAS

TESIS DOCTORAL

Searching for new genes involved in familial colorectal cancer type X by whole-exome sequencing

Búsqueda de nuevos genes implicados en el cáncer colorrectal familiar tipo X por secuenciación de exoma completo

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

PRESENTADA POR

Lorena Martín Morales

Directoras

Trinidad Caldés Llopis
Pilar Garre Rubio

Madrid

UNIVERSIDAD COMPLUTENSE DE MADRID

Facultad de Ciencias Químicas

Departamento de Bioquímica y Biología Molecular

TESIS DOCTORAL

Searching for new genes involved in Familial Colorectal Cancer Type X by whole-exome sequencing

Búsqueda de nuevos genes implicados en el Cáncer Colorrectal Familiar Tipo X por secuenciación de exoma completo

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Lorena Martín Morales

DIRECTORAS

Trinidad Caldés Llopis

Pilar Garre Rubio

Madrid, 2019

UNIVERSIDAD COMPLUTENSE MADRID

DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD DE LA TESIS PRESENTADA PARA OBTENER EL TÍTULO DE DOCTOR

D./Dña. Lorena Martín Morales,
estudiante en el Programa de Doctorado de Bioquímica, Biología Molecular y Biomedicina,
de la Facultad de Ciencias Químicas de la Universidad Complutense de Madrid, como autor/a de la tesis presentada para la obtención del título de Doctor y titulada:

Searching for new genes involved in Familial Colorectal Cancer Type X by Whole-Exome Sequencing

Búsqueda de nuevos genes implicados en el Cáncer Colorrectal Familiar Tipo X por secuenciación de exoma completo

y dirigida por: Trinidad Caldés Llopis y Pilar Garre Rubio

DECLARO QUE:

La tesis es una obra original que no infringe los derechos de propiedad intelectual ni los derechos de propiedad industrial u otros, de acuerdo con el ordenamiento jurídico vigente, en particular, la Ley de Propiedad Intelectual (R.D. legislativo 1/1996, de 12 de abril, por el que se aprueba el texto refundido de la Ley de Propiedad Intelectual, modificado por la Ley 2/2019, de 1 de marzo, regularizando, aclarando y armonizando las disposiciones legales vigentes sobre la materia), en particular, las disposiciones referidas al derecho de cita.

Del mismo modo, asumo frente a la Universidad cualquier responsabilidad que pudiera derivarse de la autoría o falta de originalidad del contenido de la tesis presentada de conformidad con el ordenamiento jurídico vigente.

En Madrid, a 3 de septiembre de 2019

Fdo.: Lorena Martín Morales

Firmado digitalmente por Lorena Martín Morales
Nombre de reconocimiento (DN): cn=Lorena Martín Morales, o=Univerrsidad Complutense de Madrid, ou=Universidad, email=lorenamartinmorales@ucm.es, c=ES
Fecha: 2019.09.03 13:15:30 +02'00'

Esta DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD debe ser insertada en la primera página de la tesis presentada para la obtención del título de Doctor.

El trabajo presentado en esta tesis se ha realizado en el Laboratorio de Oncología Molecular del Hospital Clínico San Carlos de Madrid y ha sido posible gracias a dos proyectos del Instituto de Salud Carlos III con la Dra. Trinidad Caldés como investigadora principal (FIS 13/02588 y 16/01292).

La doctoranda ha realizado además dos estancias durante el transcurso del programa de doctorado: la primera en el laboratorio del Dr. Dan Levy en Ben Gurion University (Beer Sheva, Israel), gracias a una beca de la Federación Europea de Sociedades Bioquímicas (FEBS); la segunda en el laboratorio del Dr. Zhenghe Wang en Case Western Reserve University (Cleveland, Estados Unidos) gracias a una beca de la Asociación Europea de Investigación en Cáncer (EACR).

Agradecimientos / Acknowledgements:

No podría empezar de otra manera que dando las gracias a mis directoras de tesis, las doctoras Trinidad Caldés y Pilar Garre. Trini, gracias por haberme acogido en tu laboratorio y por haberme guiado durante todos estos años. Pilar, gracias por tu ayuda y tus valiosos consejos.

Así mismo, me gustaría dar las gracias a mis compañeros del Laboratorio de Oncología Molecular del Hospital Clínico San Carlos, incluyendo a los que ya se han ido. Sois muchos, así que no os voy a mencionar a todos, pero ha sido un placer formar parte de este equipo. Echaré de menos nuestros famosos desayunos, las conversaciones de la comida sobre temas que iban desde series y viajes hasta funiculares, las risas, las cañas de los viernes, y por supuesto las tardes de juegos en Zacatrus y el lado oscuro. Quiero agradecer especialmente a Miguel, Vanesa y Patri por compartir vuestros conocimientos y por asesorarme cuando he acudido a vosotros, así como a mis co-doctorandos Mateo, Victor y Raquel, por hacer más ameno el día a día y por el apoyo mutuo que nos hemos dado. Ahora ya os dejo el relevo… ánimo que ya os queda poco!

Por otro lado agradecer todo lo que han hecho por mí mis padres, sin los que no habría podido llegar hasta aquí. Gracias por vuestro apoyo incondicional. También dar las gracias a mi hermana y el resto de mi familia: primos, sobris, tios, y especialmente a mis abuelos. Soy muy afortunada de teneros a los cuatro, y espero hacer que os sintáis tan orgullosos de mí como yo lo estoy de vosotros.

Y por supuesto, no me puedo olvidar de Marco. Gracias por apoyarme durante las estancias, y especialmente durante estos últimos meses de redacción de la tesis. Probablemente eres el segundo que más ha sufrido esta tesis (después de mí), y sé que tienes tantas ganas como yo de que termine.

I would also like to thank Dr. Dan Levy and his lab for taking me in. Thank you to all the girls for your precious help, especially Michal, Elina, Zlata, Rita and Lital. My stay in Israel has truly been one of the most amazing experiences of my career, and it wouldn't have been the same without all of you. And of course thanks to Pamela Mattar for putting the cherry on the top of a great stay.

In the same way, I would like to acknowledge Dr. Zhenghe Wang and his lab for giving me the opportunity of working with their group, especially Dongxu for making me feel welcome. Thanks also to all my Cleveland housemates from Steiner House, especially Pamela Alva. You were the most amazing roommate I could have ever had, and I feel that this has been the beginning of many great adventures together.

Finally, to my friends, who are spread around the world.

Index

Abbreviations 11

SUMMARY 15

Summary 17

Resumen 21

1. INTRODUCTION 25

1.1 **Colorectal cancer** 27

1.1.1 The colon and rectum 27

1.1.2 Epidemiology of colorectal cancer 28

1.1.3 CRC risk factors 30

1.1.4 Molecular mechanisms of colorectal carcinogenesis 33

1.1.5 CRC classification 37

1.2 **Hereditary CRC syndromes** 41

1.2.1. Polyposis syndromes 41

1.2.2. Hereditary non-polyposis CRC 44

1.3 **Familial Colorectal Cancer Type X** 49

1.3.1. Definition and terminology 49

1.3.2. Clinical, morphological and molecular features 50

1.3.3. Surveillance recommendations and management 54

1.3.4. Approaches to identify CRC risk genes/loci 55

1.3.5. Genetic susceptibility 57

1.4 **Next Generation Sequencing** 59

1.4.1. NGS technologies 59

1.4.2. NGS applications for the search of genetic variants 63

1.5 **Genes of interest: SETD6** 65

1.5.1. Lysine methylation 65

1.5.2. SETD6 65

1.6 **Genes of interest: PTPRT** 69

1.6.1. Tyrosine phosphorylation 69

1.6.2. PTPs in human cancers 70

1.6.3. PTPRT 70

1.7 **Genes of interest: PYGO1** 73

1.7.1. The canonical Wnt signaling pathway 73
1.7.2. Wnt signaling in cancer 75
1.7.3. PYGO1 75

2. HYPOTHESIS AND OBJECTIVES 77
2.1 **Hypothesis** 79
2.2 **Objectives** 79
2.2.1. Main objective 79
2.2.2. Secondary goals 79

3. MATERIALS AND METHODS 81
3.1 **Subjects and sample collection** 83
3.1.1. Study population 83
3.1.2. Control population 83
3.1.3. DNA and RNA extraction 84
3.1.4. Quantification and quality assessment of nucleic acids 84
3.1.6. Reverse transcription PCR 85
3.2 **Whole-exome sequencing strategy for the study of FCCTX** 86
3.2.1. Whole-exome sequencing 86
3.2.2. Variant filtering 86
3.2.3. *In silic*o studies 87
3.2.4. Variant prioritization 88
3.2.5. Validation, segregation and loss of heterozygosity studies 89
3.3 **Other approaches to handle FCCTX families** 91
3.3.1. Tumor whole-exome sequencing 91
3.3.2. Copy number variations 92
3.4 **Characterization studies** 93
3.4.1. Digital PCR 93
3.4.2. Quantitative real-time PCR 93
3.4.3. Promoter methylation assay 94
3.4.4. Pyrosequencing 95
3.4.5. Histone binding affinity assay 95
3.5 **Characterization of a frameshift variant in SETD6** 97
3.5.1. Cloning and plasmids 97
3.5.2. Cell lines and transfection 98
3.5.3. Western blot 99
3.5.4. Biochemical fractionation 100

3.5.5. Recombinant protein expression and purification 100
3.5.6. ELISA 100
3.5.7. Cell-free *in vitro* methylation assay 101
3.5.8. Protein-protein chromatin immunoprecipitation 101
3.5.9. RNA extraction, reverse transcription and qPCR 102
3.5.10. Viability assay 102

4. **RESULTS** 103
4.1 **Whole-exome sequencing strategy for the study of FCCTX** 105
4.1.1. FCCTX study cohort 105
4.1.2. NGS data analysis and filtering 107
4.1.3. Variant prioritization, validation, segregation and loss of heterozygosity 109
4.1.4. Variants detected in CRC susceptibility genes 110
4.2 **Other approaches to handle FCCTX families** 115
4.2.1. Tumor whole-exome sequencing 115
4.2.2. Copy number variations 115
4.3 **Candidate variants** 116
4.3.1. Predicted consequences of the candidate variants 116
4.3.2. Candidate genes 117
4.3.3. Further characterization of candidate variants 119
4.4 **Functional characterization of a frameshift variant in *SETD6*** 120
4.4.1. *SETD6* c.791_792insA (M264Ifs*3) identified in family cc598 120
4.4.2. Other candidate variants 121
4.4.3. *SETD6* M264Ifs*3 results in the loss of the C-terminal half of the protein 122
4.4.4. *SETD6* M264Ifs*3 cosegregates with CRC but shows no LOH in the tumor 122
4.4.5. Mutant SETD6 is stable and can be overexpressed in human cells 124
4.4.6. SETD6-N shows the same localization pattern as wild-type SETD6 125
4.4.7. Recombinant SETD6-N binds its substrates to the same extent as wild-type SETD6 126
4.4.8. Recombinant SETD6-N fails to methylate its substrates 126
4.4.9. SETD6-N binds its substrates but loses its activity in colon cancer cells 128
4.4.10. Both wild-type and mutant *SETD6* are expressed in the tumor and compete for their substrates 129
4.4.11. Downstream effects of SETD6-N 131
4.5 **Characterization of a frameshift variant in *PTPRT*** 134
4.5.1. *PTPRT* c.4099dup (D1367Gfs*24) identified in family cc765 134
4.5.2. Other candidate variants 134
4.5.3. *PTPRT* D1367Gfs*24 affects the second catalytic domain of the protein 136

4.5.4. *PTPRT* D1367Gfs*24 segregation and LOH 136
4.5.5. The expression of the wild-type *PTPRT* allele is decreased in the tumor from a carrier 138
4.5.6. *PTPRT* promoter is hypermethylated in the tumors from two carriers 139
4.5.7. Expression of *PTPRT*'s downstream target genes 140
4.6 **Characterization of a missense variant in *PYGO1*** 144
4.6.1. *PYGO1* c.1084T>C (S362P) identified in family cc28 144
4.6.2. Location of S362P within the Pygo1 protein 144
4.6.3. Other candidate variants 146
4.6.4. *PYGO1* S362P segregation and LOH studies 147
4.6.5. Mutant Pygo1 shows a decreased histone-binding affinity compared to the wild type 148

5. **DISCUSSION** 151
5.1 **Whole-exome sequencing strategy for the study of FCCTX** 153
5.1.1. FCCTX study cohort 153
5.1.2. Suitability of the WES strategy 154
5.1.3. NGS data analysis, filtering and prioritization 156
5.1.4. Validation, segregation and LOH 157
5.1.5. CRC susceptibility genes 157
5.1.6. Other approaches: tumor WES and CNV analysis 159
5.1.7. Candidate variants and genes 159
5.1.8. Variants selected for further characterization 161
5.2 **Characterization of SETD6 M264Ifs*3** 163
5.2.1. *SETD6* frameshift variant in family cc598 163
5.2.2. Selection of *SETD6* M264Ifs*3 as a candidate variant 163
5.2.3. Consequences of *SETD6* M264Ifs*3 164
5.2.4. Downstream effects 164
5.2.5. Involvement of *SETD6* in cancer predisposition 165
5.2.6. Other candidate variants in this family 166
5.2.7. *SETD6* as a candidate CRC predisposing gene 166
5.3 **Characterization of PTPRT D1367Gfs*24** 168
5.3.1. *PTPRT* frameshift variant in family cc765 168
5.3.2. Selection of *PTPRT* D1367Gfs*24 as a candidate variant 169
5.3.3. Consequences of *PTPRT* D1367Gfs*24 169
5.3.4. Downstream effects 171
5.3.5. Involvement of *PTPRT* in cancer predisposition 172
5.3.6. Other candidate variants in the family 173
5.3.7. *PTPRT* as a candidate cancer predisposition gene 173

5.4 **Characterization of PYGO1 S362P** 174
5.4.1. *PYGO1* missense variant in family cc28 174
5.4.2. Selection of *PYGO1* S362P as a candidate variant 174
5.4.3. Consequences of PYGO1 S362P 175
5.4.4. Other candidate variants in the family 175
5.4.5. PYGO1 is not confirmed as a candidate CRC predisposition gene 176
5.5 **Limitations of this study** 177
5.5.1. Cohort and patient selection 177
5.5.2. Family history 177
5.5.3. Methodology 178
5.5.4. Filtering and prioritization strategy 178
5.6 **Future perspectives** 180
5.6.1. Candidate variant screening in a larger cohort 180
5.6.2. Further characterization of SETD6 mutation 180
5.6.3. Further characterization of PTPRT mutation 181
5.6.4. Characterization of other candidate variants 181

6. CONCLUSIONS 183
6.1 Conclusions 185
6.2 Conclusiones 187

7. BIBLIOGRAPHY 189
7.1 List of references 191
7.2 List of websites 207

APPENDIX 209
A1. Pedigrees of the 13 families 211
A2. Information from the families 216
A3. Primers and PCR conditions 220
A4. Plasmids 224
A5. Buffers and antibodies 227
A6. Articles published during this PhD 229

Abbreviations

A	Adenine
ADE	Adenoma
Amp	Ampicillin
Ams I/II	Amsterdam I/II
BC	Breast cancer
BER	Base excision repair
BSA	Bovine serum albumin
BWA	Burrows-Wheeler Aligner
C	Cytosine
cDNA	Complementary DNA
CGH	Comparative Genomic Hybridization
ChIP	Chromatin immunoprecipitation
CNV	Copy number variation
CRC	Colorectal cancer
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide
dPCR	Digital PCR
DTT	Dithiothreitol
Dx	Diagnosis /Age at diagnosis
EC	Endometrial cancer
ECL	Enhanced chemiluminescence
EDTA	Ethylenediaminetetraacetic acid
ELISA	Enzyme-linked immunosorbent assay
EVS	Exome Variant Server
ExAC	Exome Aggregation Consortium
F	Female
FCCTX	Familial colorectal cancer type X
FFPE	Formalin-Fixed Paraffin-Embedded
For	Forward
FS	Frameshift
G	Guanine
GATK	Genome Analysis Toolkit
GC	Gastric cancer
GFP	Green fluorescent protein
H_2SO_4	Sulfuric acid
HA	Human influenza hemagglutinin
HCl	Hydrochloric acid
HCT116	Human colorectal carcinoma cell line 116
HEK293	Human embryonic kidney cell line 293T
HGMD	Human Gene Mutation Database
HGVS	Human Genome Variation Society
His	Polyhistidine tag

HNPCC	Hereditary non-polyposis colorectal cancer
HRM	High resolution melting
HRP	Horseradish peroxidase
HSF	Human Splicing Finder
IF	Inframe
IGV	Integrative Genomics Viewer
IHC	Immunohistochemistry
IP	Immunoprecipitation
IPTG	Isopropyl β-D-1-thiogalactopyranoside
KCl	Potassium chloride
LB	Lysogeny broth
LOH	Loss of heterozygosity
LS	Lynch syndrome
M	Male
MAF	Minor Allele Frequency
MBP	Maltose-binding protein
MEM	Minimal Essential Medium
$MgCl_2$	Magnesium chloride
MMR	Mismatch repair
MS	Missense
MSI	Microsatellite instablility / microsatellite instab
MSS	Microsatellite stable
NaCl	Sodium chloride
NGS	Next generation sequencing
NS	Nonsense
OD	Optical density
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PBST	PBS-Tween
PCR	Polymerase chain reaction
PKMT	Protein lysine methyltransferase
PMSF	Phenylmethylsulfonyl fluoride
PTK	Protein Tyrosine Kinase
PTP	Protein Tyrosine Phosphatase
PVDF	Polyvinylidene fluoride
qPCR	Quantitative PCR (real-time PCR)
RC	Renal cancer
Rev	Reverse
RIPA	Radioimmunoprecipitation assay buffer
RNA	Ribonucleic acid
RT-PCR	Reverse transcription PCR
SAM	S-Adenosyl methionine
SDS	Sodium dodecyl sulfate
SETD6-C	SETD6 mutant with just the C-terminal half of t
SETD6-N	SETD6 M264Ifs*3 mutant (conserves N-termin

SMART	Simple Modular Architecture Research Tool
SNP	Single nucleotide polymorphism
SP	Splicing
STRING	Search Tool for the Retrieval of Interacting Ger
SUMO	Small Ubiquitin-like Modifier
T	Thymine
TAE	Tris Acetate-EDTA buffer
TBS	Tris-buffered saline
TBST	TBS-Tween
TMB	3,3',5,5'-Tetramethylbenzidine
TNF	Tumor necrosis factor
Tris-HCl	Tris hydrochloride
TSG	Tumor suppressor gene
VCF	Variant Call Format
WES	Whole-exome sequencing
WGS	Whole-genome sequencing
Ig	Immunoglobulin
FN	Fibronectin
D1	Phosphatase domain 1
D2	Phosphatase domain 2

Summary

Summary

Introduction:

Hereditary Non-Polyposis Colorectal Cancer (HNPCC) is an autosomal dominant inherited condition characterized by an increased susceptibility to colorectal (CRC) and other associated cancers and defined by the Amsterdam I and II clinical criteria. An important fraction of these families is known as Lynch Syndrome and is caused by germline inactivating mutations in the mismatch repair (MMR) genes. This results in tumors that lack the corresponding proteins and fail to repair DNA through this pathway, causing microsatellite instability (MSI) and leading to an accumulation of somatic mutations. However, it is estimated that almost half of the families that fulfill the Amsterdam criteria do not present any defects in the MMR genes. For this reason, the term Familial Colorectal Cancer Type X (FCCTX) emerged to designate such group of HNPCC families that lack the MMR deficiency and resulting MSI that define Lynch Syndrome. Thus, FCCTX patients have microsatellite stable tumors with normal expression of the MMR genes, and the genetic basis underlying their predisposition to CRC and other related cancers remains to be elucidated.

Thanks to arrival of Next Generation Sequencing (NGS), some new CRC-predisposition genes have been recently identified; however, most FCCTX cases remain unexplained. FCCTX is thought to be a heterogeneous group of families, presumably including different yet-to-be-discovered genetic syndromes involving germline mutations in novel cancer-predisposing genes, but that could also result from a combination of low-penetrance mutations in different genes. For this reason, identifying the genetic cause of the increased cancer susceptibility in FCCTX families is still a challenge.

Objectives:

The ultimate aim of this study is the identification of new genes that could explain the increased cancer susceptibility in FCCTX families. Within this aim are enclosed the following secondary objectives: 1) to obtain a list of potentially pathogenic candidate variants for each

family; 2) to classify and characterize the variants identified; 3) to perform functional and/or expression studies when possible and suitable. All this with the purpose of finding the cause of CRC heritability in the studied families.

Methodology:

The main method on which this work was based is whole-exome sequencing, one of the applications of NGS. The obtained data was strictly analyzed and those prioritized variants were validated by PCR followed by Sanger sequencing. In addition, the segregation and loss of heterozygosity (LOH) were studied. Finally, some of the candidate genes were characterized by different expression and functional essays, involving both recombinant proteins and cell cultures.

Results:

The thorough filtering and *in silico* analysis of the variants identified by whole-exome sequencing in the affected members of 13 FCCTX families allowed the prioritization of a list variants. After their validation, the segregation in other family members and LOH were studied when possible. A total of 44 candidate variants were selected, most of which affected genes involved in important cellular events such as DNA repair, tumor suppression or different cancer-related signaling pathways. On the other hand, based on the relevance of the candidate genes and the nature of the alterations, three of the variants were selected for a more in-depth evaluation.

One of the variants selected for further characterization was *SETD6* c.791_792insA, p.(Met264IlefsTer3), carried by all the CRC-affected members of one of the families and absent in the healthy ones. *SETD6* is a mono-methyltransferase known to regulate the NF-κB and Wnt signaling pathways, among other. The characterization of the truncated version of SETD6 proved that it lacked its enzymatic activity as a methyltransferase, while maintaining other properties such as its expression, localization and substrate-binding ability. In addition, the mutant allele was expressed in the tumor and the resulting mutant protein competes with the wild type for their substrates, pointing to a dominant negative nature. These findings suggest that this mutation impairs the normal function of SETD6, which may result in the deregulation of the

different pathways in which it is involved, contributing to the increased cancer susceptibility in this family.

Another of the studied variants was *PTPRT* c.4099dup, p.(Asp1367GlyfsTer24), which showed a compatible segregation with the different types of cancer of the family (of the colon, breast and endometrium), without being present in two healthy elderly relatives. *PTPRT* is a tumor suppressor gene found to be frequently mutated at a somatic level in CRC and other cancers, having been proven that these mutations contribute to tumor development. The studied mutation resulted in the loss of a significant fraction of the second phosphatase domain of the protein, shown to be essential for its activity. In addition, the tumors of two of the carriers exhibited epigenetic inactivation of the wild-type allele and an altered expression of PTPRT's downstream target genes, consistent with a causal role of this germline mutation in the cancer predisposition of the family.

Last but not least, the third variant was *PYGO1* c.1084T>C, p.(Ser362Pro). The protein encoded by this gene plays an important role in the Wnt pathway, which is essential in CRC. Despite being a missense mutation, this variant was located in one of the zinc fingers from the PHD domain of the protein. Even though the initial segregation study was favorable, a later analysis of the tumor of a third CRC-affected member showed that this patient did not carry the mutation. Interestingly, the study of mutant PHD proved that the mutation significantly decreased this domain's affinity for two of Pygo1's substrates, which would be expected to decrease the Wnt pathway contrary to what is usually seen in a CRC context. Therefore, these results do not support a causal role of this variant in the increased cancer predisposition of this family, although its implication in tumor progression cannot be ruled out.

Conclusions:

Among all the selected candidate variants, this work proves the causality of two germline truncating mutations, in *SETD6* and *PTPRT*, although more studies are necessary to determine the exact role that they play in cancer susceptibility. However, there is not enough evidence supporting a causal role for the variant identified in *PYGO1*. Additional studies will clarify the relevance of these and other candidate genes.

Resumen

Introducción:

El Cáncer Colorrectal Hereditario No Polipósico (HNPCC) es una condición hereditaria autosómica dominante caracterizada por una aumentada susceptibilidad al cáncer colorrectal (CRC) y otros cánceres asociados, y definida por los criterios clínicos de Ámsterdam. Una parte de estas familias, conocida como Síndrome de Lynch, es causada por mutaciones germinales inactivantes en los genes MMR (del inglés "Mismatch Repair"), lo que resulta en tumores que no pueden reparar los errores en el apareamiento de bases del ADN, causando inestabilidad de microsatélites. Sin embargo, se estima que casi la mitad de las familias que cumplen los criterios de Ámsterdam no presenta defectos en los genes MMR. Por esta razón, el término Cáncer Colorrectal Familiar Tipo X (FCCTX) surgió para designar este grupo de familias HNPCC con genes MMR competentes y estabilidad de microsatélites. Desafortunadamente, la base genética de la aumentada susceptibilidad al CRC y otros tumores relacionados en familias FCCTX todavía se desconoce.

Gracias a la llegada de la secuenciación masiva o "Next Generation Sequencing" (NGS), recientemente se han identificado algunos nuevos genes de predisposición al CRC; sin embargo, la mayoría de casos de FCCTX siguen todavía sin explicar. Se piensa que FCCTX es un grupo heterogéneo de familias, que engloba distintos síndromes genéticos causados por mutaciones en diversos genes de predisposición al cáncer aún sin identificar. No obstante, no se descarta que parte de estas familias resulten de una combinación de varias mutaciones de baja penetrancia en distintos genes, lo que hace aún más difícil la búsqueda de los mismos.

Objetivos:

El objetivo final de este estudio es la identificación de nuevos genes que expliquen el incremento en la susceptibilidad al cáncer en familias con FCCTX. Dentro de esta meta se encuentran los siguientes objetivos secundarios: 1) obtener una lista de variantes candidatas potencialmente patogénicas en cada familia; 2) clasificar y caracterizar las variantes encontradas;

3) hacer estudios funcionales y/o de expresión en aquellos casos en que sea recomendable y posible. Todo esto con el fin de encontrar la causa de la heredabilidad del CRC en las familias estudiadas.

Metodología:

El método principal en el que se ha basado este trabajo es la secuenciación de exoma completo, una de las aplicaciones de la NGS. Los datos obtenidos fueron analizados de forma rigurosa, y aquellas variantes priorizadas fueron validadas por PCR seguida de secuenciación Sanger. Además, se realizaron estudios de segregación y pérdida de heterocigosidad (LOH). Por último, para algunos de los genes candidatos se caracterizaron mediante distintos estudios de expresión y ensayos funcionales, tanto con proteínas recombinantes como en cultivos celulares.

Resultados:

El estricto filtrado y análisis *in silico* de las variantes encontradas por secuenciación del exoma en miembros afectos de 13 familias FCCTX permitió la priorización de una lista de variantes. Tras su validación, se estudió la segregación en otros familiares y la LOH en los casos posibles. Un total de 44 variantes candidatas fueron seleccionadas, la mayoría en genes involucrados en mecanismos celulares importantes, como la reparación del ADN, supresión tumoral o distintas vías de señalización esenciales. Por otro lado, en base a la relevancia de los genes candidatos y en la naturaleza de las alteraciones, tres de las variantes fueron seleccionadas para su estudio en mayor profundidad.

Una de las variantes caracterizadas fue *SETD6* c.791_792insA, p.(Met264IlefsTer3), presente en todos los individuos con CRC de una familia y ausente en los sanos. SETD6 es una mono-metiltransferasa conocida por regular, entre otras, las vías de señalización NF-κB y Wnt. La caracterización de la versión truncada de SETD6 demostró que ésta carecía de actividad enzimática como metiltransferasa, mientras que mantenía otras propiedades como la expresión, localización y habilidad de unión a sustratos. Además, mostramos que el alelo mutado está expresado en el tumor y que la proteína mutada compite con la nativa por sus sustratos, lo que apunta a un papel dominante negativo. Estos resultados sugieren que esta mutación altera la

función normal de SETD6, lo que resultaría en la desregulación de las diferentes vías en las que está involucrado, contribuyendo así a la aumentada susceptibilidad al cáncer de esta familia.

Otra variante estudiada fue *PTPRT* c.4099dup, p.(Asp1367GlyfsTer24), que mostró una segregación compatible con los distintos tipos de cáncer de la familia (de colon, mama y endometrio), estando ausente en dos miembros sanos de edad avanzada. *PTPRT* es un gen supresor de tumores frecuentemente mutado a nivel somático en CRC y otros tipos de tumores, habiéndose demostrado que estas mutaciones contribuyen al desarrollo tumoral. La mutación estudiada truncaba gran parte del segundo dominio fosfatasa de la proteína, que es esencial para la actividad de la misma. Además, los tumores de dos portadores mostraron inactivación epigenética de *PTPRT* y alteración de la expresión de los genes diana de las vías en que está involucrado, lo que concuerda con un papel causal de esta mutación en la predisposición al cáncer de esta familia.

Por último, la tercera variante fue *PYGO1* c.1084T>C, p.(Ser362Pro). La proteína codificada por este gen tiene un importante papel en la vía Wnt, esencial en el CRC. A pesar de tratarse de una mutación puntual, esta variante estaba localizada en uno de los dedos de zinc del dominio PHD de la proteína. Aunque el estudio de segregación inicial fue favorable, un análisis posterior del tumor de un tercer familiar afecto reveló que este miembro no era portador. Por otra parte, el estudio del dominio PHD mutado demostró que la mutación disminuía la afinidad de unión a dos de sus sustratos significativamente, lo que conllevaría a una disminución de la vía Wnt que contradice lo esperado en un contexto de CRC. Con estos resultados no podemos concluir que esta variante sea la causa de la predisposición al cáncer en la familia, aunque no se puede descartar que esté jugando un papel en la progresión tumoral.

Conclusiones:

De entre todas las variantes candidatas seleccionadas, este trabajo demuestra la causalidad de dos mutaciones germinales truncantes en *SETD6* y *PTPRT*, aunque más estudios son necesarios para determinar el papel exacto que juegan en la susceptibilidad al cáncer. Sin embargo, no hay evidencias suficientes que apoyen la causalidad de la variante encontrada en *PYGO1*. Estudios adicionales esclarecerán la relevancia de otros de los genes candidatos.

1. Introduction

1.1 Colorectal cancer

Cancer can be defined as a set of diseases in which cells have acquired the ability to divide and grow uncontrollably by disregarding the normal rules of cell division[1,2]. Specifically, colorectal cancer (CRC) refers to the development of an abnormal cell growth in the tissue of the colon or rectum with the potential to invade other parts of the body. Further information regarding its epidemiology, diagnosis, treatment, risk factors and pathways involved is detailed below.

1.1.1 The colon and rectum

The colon and rectum are sections of the large intestine, which is the terminal part of the lower gastrointestinal track and consists of the cecum, appendix, colon, rectum and anal canal[3] (Figure 1.1 A and B). The colon extends upwards from the cecum and is divided into the ascending, transverse, descending and sigmoid colon[3]. The transverse colon is flanked by the right colic flexure, also known as the hepatic flexure for its proximity to the liver, and the left colic flexure or splenic flexure, located just below the spleen[3]. Finally, extending from the sigmoid colon is the rectum, which is in turn followed by the anal canal[3].

Unlike the small intestine, the colon does not play a major role in the absorption of food and nutrients[4]. Nonetheless, being at the end of the digestive system it has a series of essential functions, including the absorption of water and electrolytes from solid wastes, the production and absorption of vitamins[5], the microbiota-aided fermentation of unabsorbed material[4] and the formation of the feces, which are then propelled to the rectum for elimination[5]. The unique architecture of the colon facilitates these functions, its multiple folds resulting in an immense surface area that allows maximal absorption[6]. The intestinal wall is formed by multiple layers joined together by connective tissue and by the neural and vascular elements[6] (Figure 1.1 C): the innermost layer is the mucosa (formed by the epithelium, a layer of connective tissue called lamina propia and a thin muscle layer called muscularis mucosa), underneath it is the submucosa, then the

muscularis propria or thick muscle layers, and finally the serosa, which is covered by the visceral peritoneum and forms a natural barrier from the spread of inflammatory and malignant processes[6].

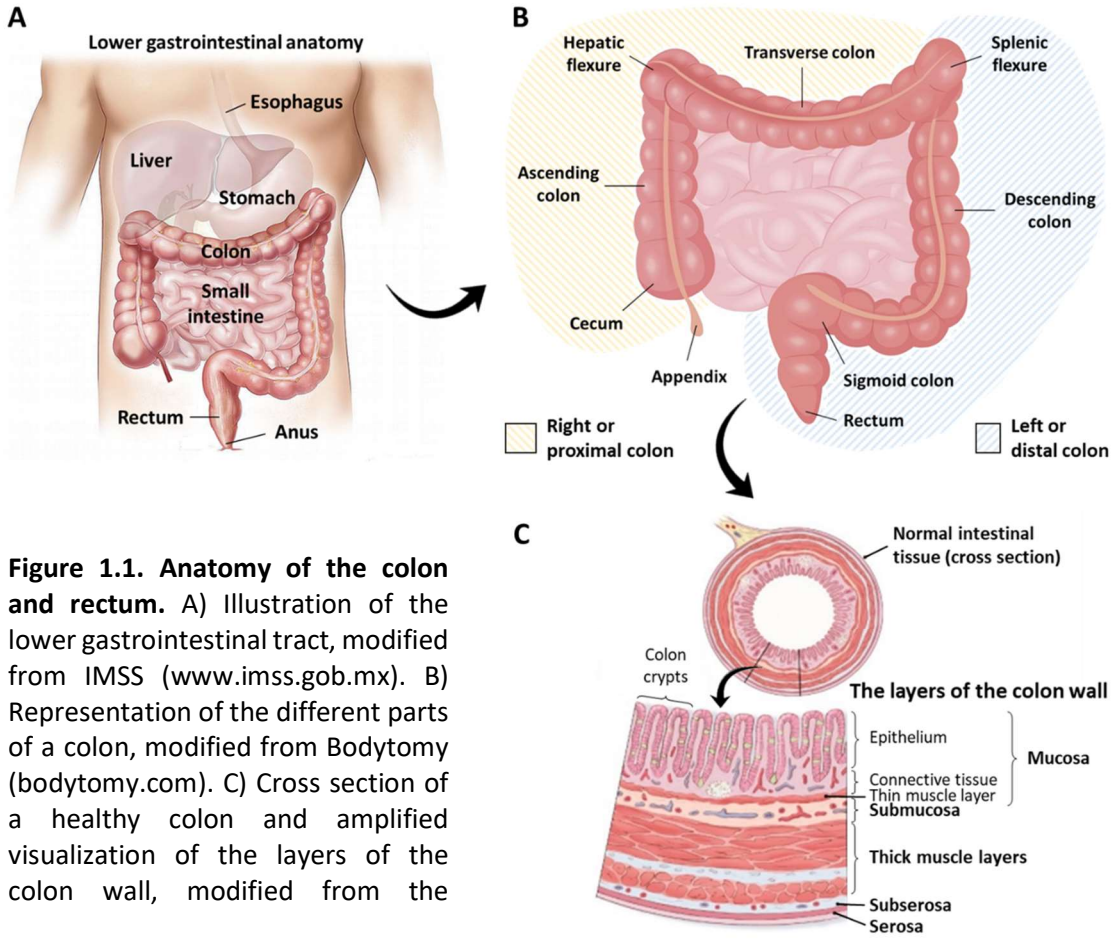


Figure 1.1. Anatomy of the colon and rectum. A) Illustration of the lower gastrointestinal tract, modified from IMSS (www.imss.gob.mx). B) Representation of the different parts of a colon, modified from Bodytomy (bodytomy.com). C) Cross section of a healthy colon and amplified visualization of the layers of the colon wall, modified from the

1.1.2 Epidemiology of colorectal cancer

CRC is the third most common cancer and the fourth leading cancer-related cause of death in the world, with 1,360,602 new cases and 693,933 fatalities registered in 2012[7] (Figure 1.2 A). Regarding the gender, CRC is the third most commonly occurring cancer in men and the second in women, although in most parts of the world the incidence rate of CRC is higher in men than in women[7,8].

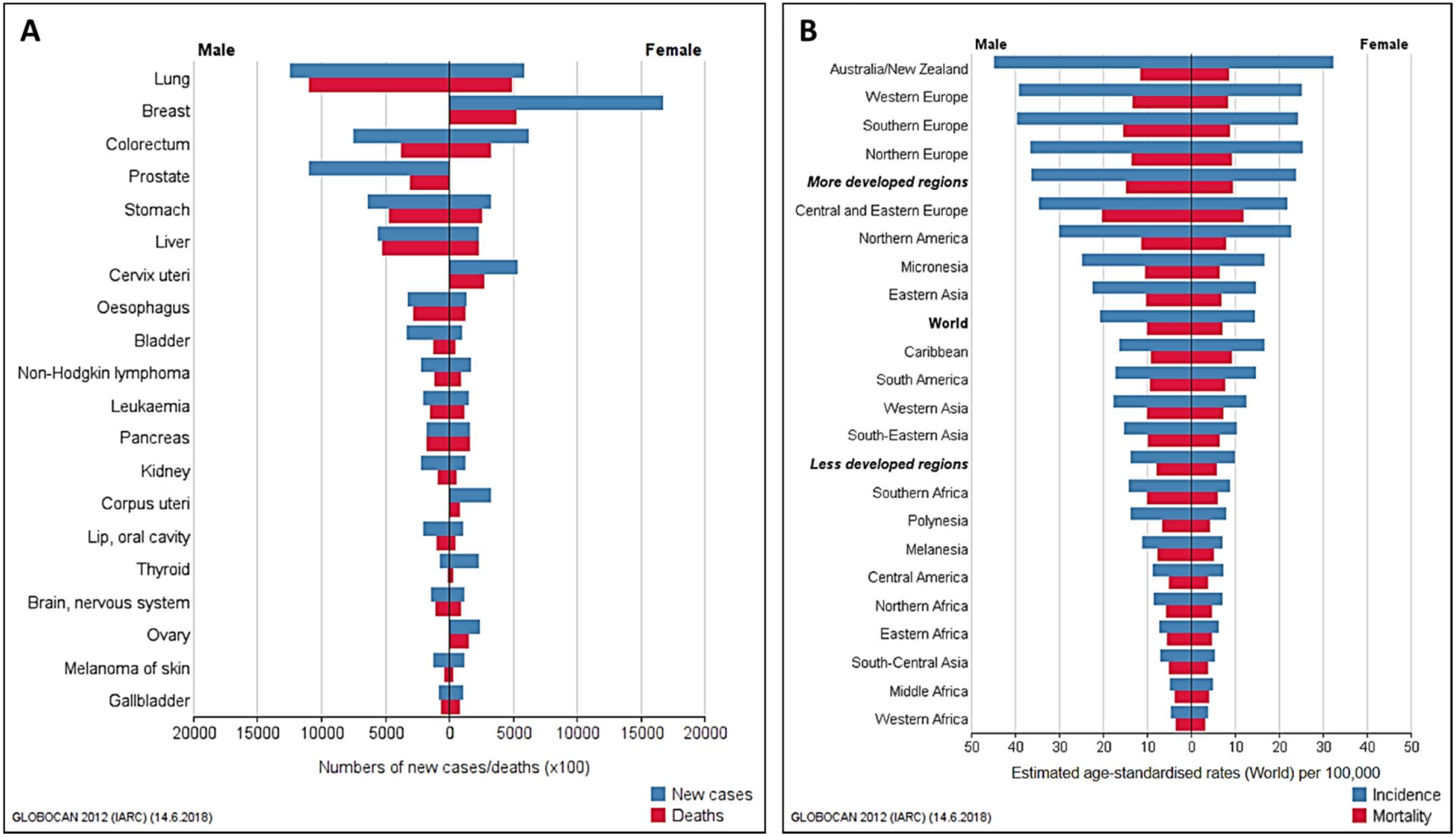


Figure 1.2. World incidence and mortality of CRC. A) Number of new cases (blue) and deaths (red) x100 of the most frequent cancers in the world, in males (left) and females (right). CRC is situated in third position, although it is the fourth when it comes to mortality rates. B) Geographical variation in CRC incidence (blue) and mortality (red) in males (left) and females (right). Estimated age-standardized rates per 100,000. CRC incidence is higher in countries with higher indices of development, but mortality rates are higher in developing countries. Obtained from Globocan 2012 (gco.iarc.fr).

Although the incidence rate of CRC varies greatly worldwide[9], about two thirds of CRC cases occur in countries characterized by high indices of development and/or income[10] (Figure 1.2 B). However, CRC incidence in developing countries has been steadily increasing[11], which indicates a growth in the prevalence of CRC risk factors associated with Western lifestyle, such as unhealthy diet, obesity and smoking[12]. On the other hand, despite being more common in developed areas, mortality rates are higher in developing countries. Mortality rates have shown a descending trend in most economically-developed countries, including Spain and most Western European countries[13]. The decrease in mortality could be due to improved diagnostic procedures and prevention, reduction of risk factors, and access to healthier diets[13]. However, CRC mortality rates are still high in some countries with limited resources, which could be due to the lack of predictive measures and/or the increase in the incidence of CRC[8]. Mortality rates are also 30-40% higher in men than in women[12].

1.1.3 CRC risk factors

No single risk factor accounts for most CRC cases. Instead, many lifestyle-related factors have been linked to CRC[14]. As a matter of fact, the links between diet, weight, and exercise and CRC risk are some of the strongest for any type of cancer[10]. In addition, there are other personal unmodifiable factors that are clearly involved in the lifetime risk of developing CRC, such medical history, age and genetic factors[8,15]. The main risk factors reported to be associated with CRC are summed up below.

Nutritional and lifestyle factors:

Diet plays a key role in the development of CRC[16], and among the most extensively studied is probably the association between a high consumption of red and processed meat[16–20], cholesterol-rich food and saturated fats[8,21] with an increased risk for CRC. On the other hand, several epidemiological studies have shown that increased consumption of vitamin D, folate, fruits, vegetables[16], a high-fiber diet[22], calcium[23] and dairy products reduce the risk for CRC[8,10]. Interestingly, recent studies have reported a dual role of the

gut microbiota in colorectal carcinogenesis[24,25], suggesting it may be an important effector in the relationship between diet and cancer[16,26].

The link between lifestyle and CRC is also well reported. Noteworthy, physical activity has a protector effect[8,27], while obesity and increased body mass index can significantly increase the risk for CRC[8,28]. High alcohol consumption has been also reported to increase CRC incidence in a considerable amount of studies[29–32], as well as tobacco consumption (specially cigarette smoking[8]), which is another common risk factor for CRC[32–34]. Conversely, although a correlation between caffeine consumption and increased risk for CRC has been suggested[35], this relationship has not been definitively confirmed[8]. CRC rates have been also found to be higher in people with lower education or lower socioeconomic status, which could be due to higher incidence of moderating risk factors in those areas (such as immobility, unhealthy diet, smoking and obesity), as well as low screening rates[8]. On the other hand, several studies have shown that the use of aspirin and other nonsteroidal anti-inflammatory drugs (NSAIDs) can play a role in CRC prevention[36–39]. As a matter of fact, aspirin consumption is the only chemoprevention currently approved in patients with a higher genetic predisposition to CRC[37], and can even reduce recurrence, metastasis and mortality[38,39].

Genetic factors:

Inherited susceptibility contributes significantly to CRC predisposition, with an estimated 12-35% of risk attributed to genetic factors[40,41]. Certainly, people with a history of CRC in a first-degree relative (parent, sibling or child) are at a 2.25 increased risk of developing CRC[42–45], and some studies suggest that the extended family history should be also considered[46]. This risk is even higher if the CRC was diagnosed before the age of 45 or if more than one first-degree relative is affected, and can be due to inherited genes, shared environmental factors or the combination of both[8,45]. Having family members who have had adenomatous polyps is also linked to a higher risk for CRC[47], given that 70-90% of CRCs arise from adenomatous polyps and that those larger than 2cm in diameter have a 50% chance of malignancy[8,47].

This inherited susceptibility is caused by germline pathogenic variants in certain cancer risk genes, whose effect depends on the penetrance of the corresponding allele, which is defined as the percentage of individuals who present that variant that will develop the disease. The genetic factors that can predispose to CRC range between rare high-penetrance mutations that confer considerable elevations in the risk for hereditary syndromes and common low-penetrance variants or polymorphisms associated with weak effects on the risk for sporadic CRC[41], including anything in between. In addition, common variants may also modify the risk in individuals with hereditary syndromes[41]. Figure 1.3 shows the main genes that have been associated with CRC predisposition, distributed by their conferred effect and risk allele frequency[41].

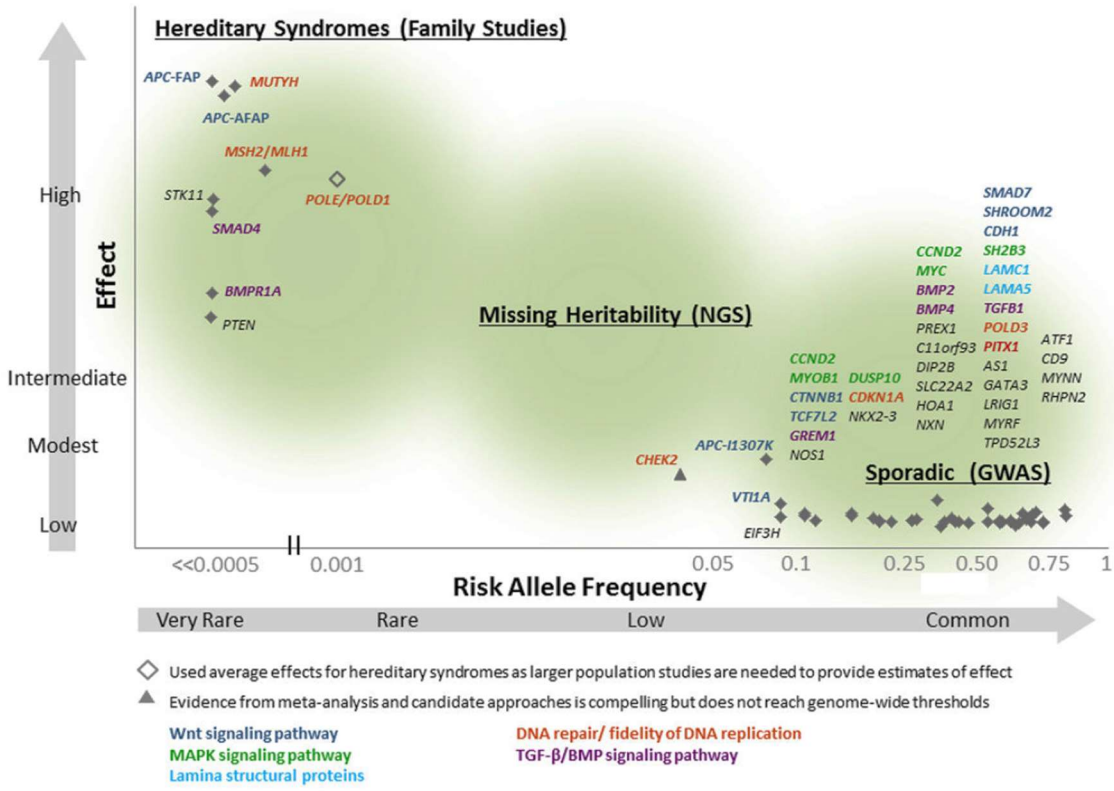


Figure 1.3. Genetic architecture of known CRC genetic susceptibility loci. Risk allele frequency shown for the ethnicity in which the locus was discovered. For variants with a recessive effect (MUTYH), the frequency of the homozygote rare allele is shown. Obtained from Peters et al. 2016.

Personal medical history and other unmodifiable risk factors:

Patients with long-term inflammatory bowel disease (such as ulcerative colitis and Crohn's disease) have an increased risk for CRC[48–50], which depends on the duration and extent of disease, the age of onset and the severity of inflammation[8]. Various studies have also shown that type 2 diabetes mellitus is associated with an increased risk for CRC[51,52]. Other CRC risk factors linked to the personal medical history include having undergone certain medical procedures such as cholecystectomy, ureterocolic anastomosis and pelvic radiotherapy[8]. Age is also a major risk factor for CRC, which is uncommon before the age of 40[9,11]. The incidence begins to increase significantly between the ages of 40 and 50, with rates increasing in each succeeding decade[8]. On the other hand, the role of gender in CRC development has not been definitively confirmed, although in most studies it has been reported that CRC risk is higher in men than in women[7,8]. Interestingly, right-sided CRC has been observed more in women, while left-sided CRC is mostly observed in men[53].

1.1.4 Molecular mechanisms of colorectal carcinogenesis

CRC is a heterogeneous group of diseases with distinctive genetic and epigenetic characteristics[54]. Colorectal carcinogenesis is driven by the sequential activation of proto-oncogenes and inactivation of tumor suppressor genes[55]. Some of the genes that are particularly important in CRC are involved in pathways that are essential for cancer initiation and progression, such as the Wnt, MAPK, PI3K, TGF-β and p53 signaling pathways[56]. Alterations in these pathways can occur due to different molecular mechanisms of tumorigenesis. Most CRCs typically arise from one or a combination of three different mechanisms: chromosomal instability (CIN), microsatellite instability (MSI or MIN) and CpG island methylator phenotype (CIMP)[57]. Moreover, CIN, MSI, and CIMP often overlap, which has important implications for CRC prognosis, diagnosis and treatment[57]. The CIN pathway accounts for the majority of sporadic CRCs, whereas the MSI or mutator pathway accounts for the majority of hereditary cases and about 15% of sporadic CRCs[58]. On the other hand, the CIN pathway is thought to progress through the classic adenoma-carcinoma sequence[55], while the CIMP pathway arises via the serrated

adenoma pathway (derived from sessile serrated adenomas)[56], and MSI tumors have been reported to progress by both of these[55,59].

Chromosomal Instability:

This is a hallmark of classical CRC that is associated with 70-85% of sporadic CRCs[60], and typically begins with the acquirement of inactivating mutations in the *APC* gene, followed by the mutational activation of the *KRAS* oncogene and the inactivation of tumor suppressor gene *TP53*[57]. This pathway is also associated with familial adenomatous polyposis, caused by germline mutations the *APC* gene[61]. APC is a tumor suppressor that negatively regulates the Wnt pathway. Its inactivation triggers the hyperactivation of the Wnt signaling pathway by increasing β-catenin levels and allowing its translocation to the nucleus, where it promotes the transcription of different oncogenes[62]. This inactivation may happen through somatic mutations (which occur in 72% of sporadic CRCs and 30%-70% of sporadic adenomas) or by the hypermethylation of its promoter (which happens in 12% of primary carcinomas and adenomas)[57]. Mutations in other genes of this pathway, especially activating mutations in *CTNNB1* (which encodes β-catenin), are also linked to CIN and are found in 48% of CRCs without *APC* mutations[57].

On the other hand, *KRAS* is the most commonly mutated member of the RAS family in CRC, and is an important component of the MAP kinase pathway, which regulates cell proliferation, differentiation, senescence and apoptosis[62]. Point mutations in exons 2 and 3 of *KRAS* (mainly in codons 12, 13, and 61) activate this kinase, which further activates the Raf-MEK-ERK and PI3K/AKT/PKB pathways, promoting tumorigenesis[57]. *KRAS* mutations occur after *APC* mutations in 30%-60% of CRCs and play an important role in the adenoma-carcinoma sequence[57,63]. Last but not least, *TP53* is a tumor suppressor gene that plays an essential role in the control of the cell cycle and apoptosis. The p53 transcription factor induces G1 cell-cycle arrest and allows DNA repair prior to DNA replication, inducing cell death through apoptosis if DNA repair is unsuccessful[62]. *TP53* mutations are believed to occur at the time of transition from adenoma to cancer[62], and are present in up to 75% of sporadic CRCs and approximately 30% of late adenomas[58]. In

addition, germline *TP53* mutations cause Li-Fraumeni syndrome, which is an inherited condition that predisposes to a wide variety of neoplasms, including CRC[64].

Frequent genetic alterations observed in the CIN pathway include aneuploidy, which is an imbalance in the chromosome number[57], chromosomal genomic amplifications on chromosomes 7, 8q, 13q, 20 and X, and a high frequency of loss of heterozygosity (LOH), which is the loss of one of the copies or alleles of a gene[63]. Broad deletions can occur at chromosomes 1, 4, 5, 8p, 14q, 15q, 17p, 18, 20p and 22q, with the highest LOH frequency affecting the long arm of chromosome 18 (18q)[62], where many tumor suppressor genes are located, including S*MAD4* and *DCC*[58,63]. Frequently affected by partial loss are also chromosome 5q ,where *MCC* and *APC* are located (20-50% of CRCs), 8p (50% of CRCs) and 17p, where *TP53* is located (75% of CRCs but not in adenomas)[63]. In addition, focal gains or losses are found in regions containing important cancer genes[63], while mutations in other genes (such as *TGFBR* and *PIK3CA*) have been also recently described[62].

Microsatellite Instability:

The MSI pathway is caused by the inactivation of the DNA mismatch repair (MMR) system, which is responsible for correcting the errors that occur during DNA replication as an additional system to preserve genomic integrity[62], and whose most important components are ATPases *MLH1*, *MLH3*, *MSH2*, *MSH3*, *MSH6*, *PMS1* and *PMS2*[57]. Both alleles of a MMR gene must be inactivated for a deficient MMR system, which may be caused by somatic mutations in the MMR genes[58] or by MMR promoter hypermethylation, mechanism that is often associated with the CIMP pathway[57]. Germline pathogenic mutations in the MMR genes (mainly *MLH1*, *MSH2*, *MSH6* and *PMS2*) predispose to cancer causing Lynch Syndrome, a hereditary CRC syndrome that will be further explained in section 1.2[57]. MSI is therefore a hallmark of Lynch Syndrome tumors, but is also found in 15-20% of sporadic CRCs[62], usually caused by hypermethylation of the *MLH1* gene promoter[57]. The MSI pathway is often associated with proximal colon tumors, mucinous histology, lymphocytic infiltration and poor differentiation but better prognosis[57,62].

A defective MMR system results in genetic alterations in short tandem repeats (known as microsatellites), whose length variability in the tumor is termed microsatellite instability[62]. MSI can be detected by a five-marker panel which includes mono- or dinucleotide repeats in certain genes (such as *BAT25*, *BAT26*, *D5S346*, *D2S123* and D17S250)[57,63], while the immunohistochemical staining of MMR proteins is useful to identify which MMR protein is deficient[62]. Tumors with instability in ≥30% of these markers are called MSI-high (MSI-H), those with instability in <30% are called MSI-low (MSI-L), while those without MSI are called microsatellite stable (MSS)[57]. It is commonly accepted that the majority of MSS tumors follow the CIN pathway of tumorigenesis[65]. The consequence of MMR deficiency is the accumulation of somatic mutations, reason for which MSI tumors are hypermutated[55] and this pathway is also known as the mutator pathway[58]. *TGFβRII* gene mutations are the most common in this pathway, but other genes targeted for mutations include *BAX*, *CASP5*, *IGF2R*, *E2F4*, *TCF4*, *MSH3*, *MSH6*, *RIZ*, *PTEN*, *RAD50, BRAF* and *CDX2*[58]. LOH and mutations in *APC*, *TP53* and *KRAS* are less frequently observed than in CIN, with *APC* mutations detected in only in 21% of MSI CRC cases[58]. *BRAF* mutations are common in sporadic but not hereditary MSI CRCs[59]. Interestingly, *BRAF* activating mutations activate the same RAS-RAF-MEK-ERK signaling pathway as *KRAS*, while *TGFβRII* mutations inactivate the TGF-β pathway, promoting growth and proliferation in the same way as *SMAD4* loss, and *BAX* mutations can block p53-induced apoptosis as an alternative to *TP53* mutations[59].

CpG Island Methylator Phenotype:

Epigenetic regulation changes gene expression or function without affecting the DNA sequence, usually by DNA methylation or histone modifications[63,65]. DNA methylation commonly occurs at the cytosine within CG dinucleotides (called CpG islands) located around promoter regions near transcription start sites, leading to gene silencing[59,63]. Global DNA hypomethylation and localized promoter CpG island hypermethylation are common epigenetic events in cancer[57], and environmental factors including smoking and advanced age have been shown to correlate with increased methylation[63]. Specifically, in CRC the CIMP pathway is characterized by widespread

promoter hypermethylation resulting in the transcriptional inactivation of multiple genes, including tumor suppressors or genes involved in the cell cycle[57]. This epigenetic gene silencing provides an alternative mechanism for loss of function of various tumor suppressor genes[63], most importantly *MGMT* and *MLH1*[57], but also *APC*, *MCC*, *CDKN2A*, *IGF2*, *RUNX3*, *SOCS1*, *MINTs* and several others[57,59,63]. A classic example is the hypermethylation silencing of *MLH1* in sporadic MSI-H CRC[63].

Although the CIMP phenotype can occur in either MSI-H or MSS tumors[54], promoter hypermethylation usually overlaps with sporadic MSI CRCs[66] and is associated with *BRAF* mutations and MSI[57]. Mutations in the *BRAF* proto-oncogene appear to be an early event in CIMP tumors[57], specially the *BRAF* V600E mutation, which is strongly correlated with *MLH1* hypermethylation and is observed in 18.7% of sporadic cases[57]. *MLH1* hypermethylation has been confirmed in 80% of MSI-H sporadic CRCs, which present loss of *MLH1* expression without germline MMR gene mutations[57]. Different marker gene panels (including *MLH1*, *CDKN2A* and other genes)[57,59,63] have been developed to divide tumors as CIMP-high (CIMP-H), CIMP-low (CIMP-L) and CIMP-negative[65]. Pathologically, CIMP-H tumors are often associated with older age, female gender, proximal tumor location, poor tumor differentiation, wild-type *TP53*, mucinous or signet ring cell morphology, marked peritumoral lymphocytic reaction and high levels of global DNA methylation[56,63], features that are also associated with MSI-H tumors[56]. The precursor lesions are usually sessile serrated adenomas or serrated polyps, which account for 9% of colorectal polyps and have distinct features (flat or minimal elevation, strong predilection for the cecum and ascending colon, *BRAF* mutations and extensive DNA methylation)[63].

1.1.5 CRC classification

Molecular CRC subtypes:

Even though different combinations are possible, the analysis of the molecular mechanisms involved in tumor development has allowed the classification of CRC into 5

main molecular subtypes, each with a different molecular profile and clinico-pathological features[59,63]: **1)** CIMP-H, MSI-H, *MLH1* methylation, *BRAF* mutation, chromosomally stable, often originated in hyperplastic polyps, sessile-serrated lesions, and serrated adenomas, and known as sporadic MSI-H (12%). **2)** CIMP-H, partial *MLH1* methylation, *BRAF* mutation, chromosomally stable, MSS or MSI-L, originated in serrated polyps (8%). **3)** CIMP-L, *KRAS* mutation, *MGMT* methylation, CIN, MSS or MSI-L, with origin in conventional tubular or tubulovillous adenomas or serrated polyps (20%). **4)** CIN, MSS, CIMP-negative, originated in traditional adenomas (57%). **5)** Inherited MMR gene mutation, MSI-H, CIMP-negative, *BRAF* mutation negative, chromosomally stable, with origin in conventional adenomas or occasionally in serrated adenomas (3%)[59,63].

Sporadic, familial and hereditary CRC:

According to the family history, CRC can be classified into 3 types[67], as shown in Figure 1.4: **1) Sporadic CRC** represents the vast majority of CRCs (60-80%) and appears in individuals with no family history of CRC and no cancer-predisposing germline mutations[67]. **2) Familial CRC** refers to a group of cases with some familial aggregation of the disease but without a dominant inheritance pattern and for which no associated genes have been identified. These cases represent 20-40% of CRCs are thought to result from a combination of genetic predisposition due to low-risk alleles and environmental or lifestyle shared factors[67]. **3) Hereditary CRC** comprises those cases that show a Mendelian inheritance of CRC and other associated cancers, and/or present pathogenic mutations in known CRC-predisposition genes. This group is further divided depending on the presence of polyps into polyposis and non-polyposis syndromes, and accounts for 5-10% of all CRCs[67].

CRC staging classification:

CRC stages were created for diagnostic purposes in order to help ascertain the extent to which a certain CRC had developed. Tumor staging is determined by histologic analysis of surgically resected lesions based on a series of parameters, and is important for the estimation of prognosis and the selection of the most suitable treatment for each

patient[54]. The most commonly used staging system for CRC is the American Joint Committee on Cancer (AJCC) TNM system[68], which assigns a number based on 3 categories: the size of the tumor and degree of invasion of the colon wall (T), the degree of spreading to nearby lymph nodes (N) and the degree of spreading to other organs or metastasis (M). Carcinoma in situ (stage 0) is the earliest stage, after which stages range from I to IV, with higher numbers meaning a higher spreading (Figure 1.5). Stages II, III and IV are in addition divided into further subgroups (A, B and C) for a more accurate classification.

Other staging systems that have been used over the years include the Dukes classification[69,70], ranging from dukes A to D, and the Ashton-Coller classification[71], which is an adaptation from the former that further divides stages B and C. However, currently both of these have been largely replaced by the TNM system[68].

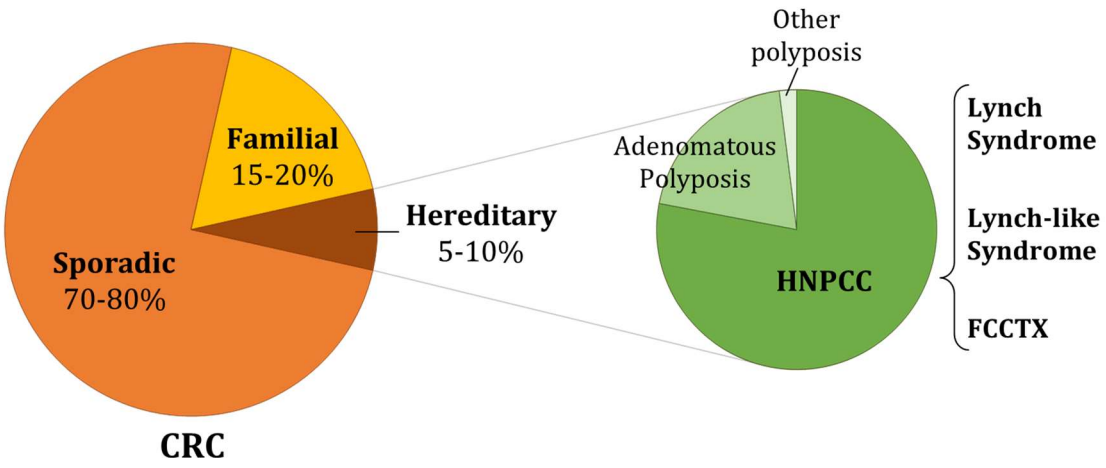


Figure 1.4. CRC classification based on family history. Even though the majority of CRCs are sporadic, it is estimated that about a 20-30% of CRCs have a familial component, with a 5-10% showing a dominant pattern of inheritance or presenting mutations in CRC predisposition genes. That fraction is called hereditary CRC, and is comprised of a few different polyposis syndromes and a majority of Hereditary Non-Polyposis Colorectal Cancer (HNPCC) cases. HNPCC is defined by the fulfillment of the Amsterdam I/II criteria and is further divided into Lynch Syndrome, Lynch-like Syndrome and Familial Colorectal Cancer Type X (FCCTX), which is the group in which this thesis is focused.

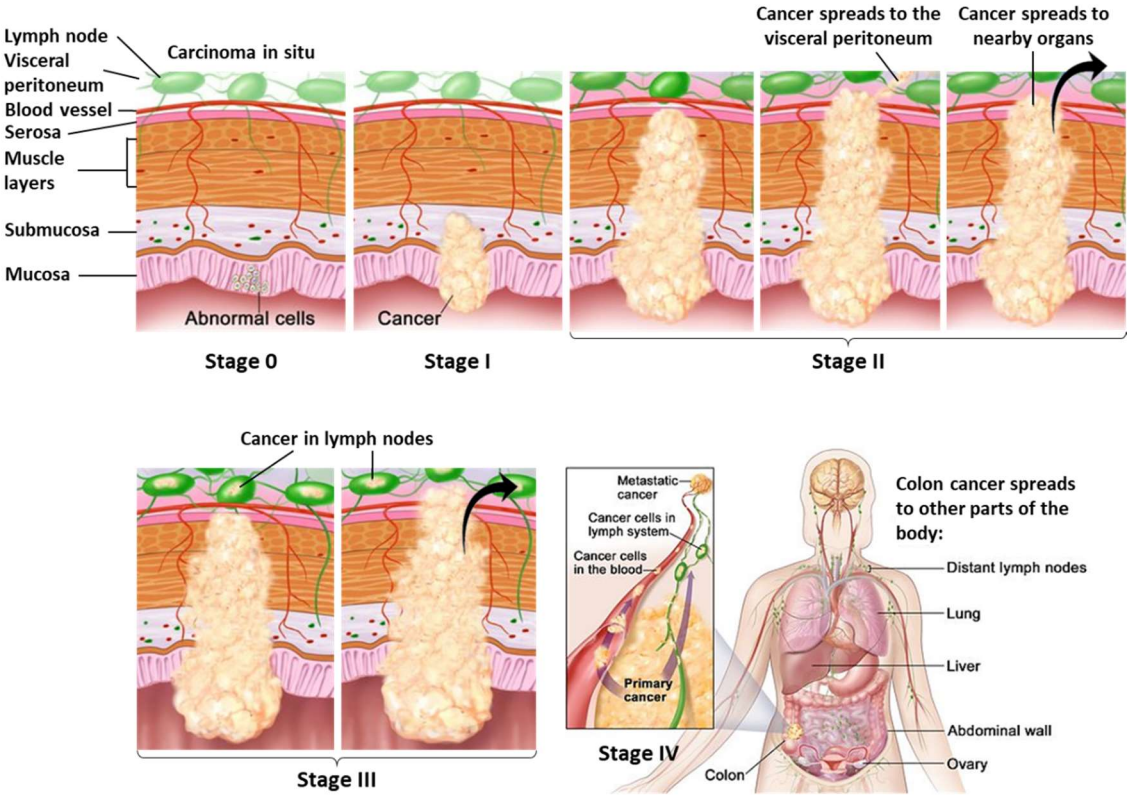


Figure 1.5. CRC staging classification. Simplified representation of the different stages in the development of a colorectal tumor, from a carcinoma in situ (stage 0) until the invasion of lymph nodes (stage III) and distant organs (stage IV). Images have been modified from the National Cancer Institute (www.cancer.gov).

1.2 Hereditary CRC syndromes

Even though the majority of CRCs are sporadic, genetic susceptibility to CRC is more common than previously thought[72]. Several recent studies have identified pathogenic germline variants in a broad spectrum of high and moderate penetrance cancer susceptibility genes in >10% of individuals with CRC[72]. Genetic susceptibility appears to be even more prevalent among young CRC patients, with several studies documenting prevalence of germline mutations of 16-33% among those diagnosed under the age of 50[73,74]. The hereditary CRC syndromes, characterized by dramatic increases in risk for colorectal neoplasia, are phenotypically divided into polyposis and nonpolyposis syndromes, based largely on the number and histology of colorectal polyps[75].

1.2.1. Polyposis syndromes

Polyposis syndromes represent around 1% of all CRC cases and can be divided, based on their histology and genetic basis, into the broad categories detailed below[75,76].

Adenomatous polyposis:

- **Familial adenomatous polyposis (FAP)** is an autosomal dominant genetic disease characterized by the development of multiple colorectal adenomas (usually from dozens to hundreds), and caused by pathogenic germline variants in the *APC* gene[75,77]. Despite its inheritance pattern, approximately 25-30% of the cases have no family history and are explained by *de novo* mutations or somatic mosaicisms[77–80]. The clinical presentation varies significantly, from a classic phenotype (100s-1000s polyps) that usually requires surgical colectomy, to more subtle presentations (10-100 polyps), known as **attenuated polyposis (AFAP)** [77,78]. If untreated, individuals with FAP and AFAP have a lifetime risk of developing CRC of 100% and 69%, respectively[77]. In addition, they can also develop upper gastrointestinal tract lesions (including gastric fundic gland polyps and duodenal adenomas) and extraintestinal manifestations such as increased risk for papillary thyroid cancers and desmoid tumors[75].

- **MUTYH-associated polyposis (MAP)** is an autosomal recessive syndrome associated with biallelic germline variants in the base-excision repair gene *MUTYH*. Individuals with MAP can present a wide range of phenotypes including classic and attenuated polyposis, with an increased risk for CRC[77,81]. Although adenomatous polyps predominate in MAP, hyperplastic polyps are also common[77]. It is still unclear whether monoallelic *MUTYH* variants could be associated with a moderate CRC risk increase[75,82].

- **Polymerase Proofreading-Associated Polyposis (PPAP)** is associated with germline pathogenic variants in the exonuclease proofreading domains of the *POLE* and *POLD1* polymerases[83,84]. PPAP patients may present autosomal dominant classic or attenuated polyposis, and show ultramutated tumors with MMR-proficient phenotypes[75].

- **Other adenomatous polyposis syndromes** include two recently identified rare autosomal recessive forms caused by biallelic mutations in *NTHL1*, a DNA glycosylase involved in base excision repair, and in *MSH3*, a MMR gene not associated with Lynch syndrome[75,85]. The latter has been reported to confer susceptibility to classical or attenuated adenomatous polyposis and possibly extracolonic tumors, including breast cancer[86,87].

<u>Hamartomatous polyposis syndromes</u>:

Although rare, hamartomatous polyposis syndromes are characterized by the presence of gastrointestinal hamartomatous polyps and exhibit autosomal dominant patterns of inheritance[75]. This type of polyposis includes the following syndromes:

- **Peutz-Jeghers syndrome (PJS)** is characterized by multiple hamartomatous polyps throughout the gastrointestinal tract and increased risk for various cancers including gastric, colorectal, pancreatic, breast, lung, and sex cord tumors. Individuals with PJS may have prominent mucocutaneous pigmentation and bowel obstructions due to polyp intussusceptions. Germline pathogenic variants in *STK11* are identified in 50-70% of individuals[75,88].

- **Juvenile polyposis syndrome (JPS)** is characterized by multiple gastric and/or colonic hamartomas and increased risk for gastric cancer and CRC. Germline pathogenic variants in *BMPR1A* and *SMAD4* are identified in 50-70% of affected individuals, and those with *SMAD4* mutations are also at risk for hereditary hemorrhagic telangiectasia[75,89].

- **PTEN-hamartoma tumor syndrome (PHTS)** is associated with increased risk for breast, thyroid, endometrial, renal and gastrointestinal cancers and is caused by germline pathogenic variants in *PTEN*[75]. The gastrointestinal phenotype of PHTS can include gastric and colorectal hamartomas, adenomas, serrated polyps, hyperplastic polyps, lipomas and ganglioneuromas[90]. *PTEN* pathogenic variants confer variable clinical phenotypes, which include different conditions such as Cowden, Bannayan-Riley-Ruvalcaba and Proteus-like syndromes[75,91–93].

Serrated polyposis:

Previously known as hyperplastic polyposis, serrated polyposis is defined by: 1) more than 5 serrated polyps proximal to the sigmoid colon with at least two bigger than 1cm; 2) any number of serrated polyps proximal to the sigmoid colon in an individual with a first-degree relative who has serrated polyposis; or 3) more than 20 serrated polyps of any size throughout the colon[54]. Germline mutations in the tumor suppressor gene *RNF43* have been identified in rare cases of serrated polyposis, but the mutation frequency among affected individuals is low[94,95]. Although germline mutations in *GREM1* and *MUTYH* have been reported, genetic testing is usually uninformative[75,96,97].

Mixed polyposis:

Mixed polyposis is characterized by the presence of multiple colorectal polyps of mixed histological type (including serrated lesions, hamartomas and conventional adenomas), and is associated with an increased risk for CRC. While the genetic cause is

still unknown in most cases, germline variants in and upstream of *GREM1* have been identified in some affected individuals[75,98].

1.2.2. Hereditary non-polyposis CRC

Even though the CRC syndromes associated with polyposis phenotypes are the most easily recognized, the vast majority of individuals affected by genetic predisposition to CRC do not exhibit multiple polyps[75]. Syndromic non-polyposis CRC is called Hereditary Non-Polyposis Colorectal Cancer (HNPCC) and accounts for about 5-10% of all the cases of CRC. HNPCC is defined by the fulfillment of one of two sets of selection criteria, known as the Amsterdam Criteria I and II (AC-I/II), which are shown in Table 1.1[99]. Although HNPCC shows a clear dominant inheritance pattern of CRC and other associated cancers, the genetic basis underlying the increased cancer risk of these families is only known in 50-60% of them. These explained cases are known as Lynch Syndrome, and present tumors with defects in the mismatch repair pathway of DNA repair (known as MMR) and a MSI phenotype. For this reason, HNPCC has been subdivided on the basis of the molecular tumor phenotype as MMR-deficient/MSI (Lynch Syndrome) or MMR-proficient/MSS (Familial Colorectal Cancer Type X)[75]. The pipeline for the diagnosis of the different HNPCC syndromes is shown in Figure 1.6.

Lynch Syndrome:

Lynch Syndrome (LS) is the most common of the hereditary CRC syndromes, accounting for 50-60% of the families that fulfill the AC-I/II clinical criteria and 3-5% of all CRC cases. In these patients, the cancer predisposition is caused by germline pathogenic mutations in one of the MMR genes, mainly *MLH1*, *MSH2*, *MSH6* and *PMS2*[99], although deletions of the 3' end of *EPCAM* (which leads to methylation of the promoter region of *MSH2*) have also been described as a cause behind LS[75]. Germline pathogenic mutations in these genes lead to the lack of expression of the corresponding protein in the tumors, what results in a deficient MMR pathway that fails to repair the errors produced during DNA replication[100]. Stretches of DNA with repetitive sequences of nucleotides known as

microsatellites are particularly susceptible to DNA errors when the MMR genes are damaged[99], leading to microsatellite instability (MSI). For this reason, LS-associated tumors have distinctive molecular phenotypes of MMR-deficiency, high MSI and loss of expression of the corresponding DNA mismatch repair protein as observed by immunohistochemistry (IHC)[75]. Although CRC and endometrial cancer are the most common cancers in families affected by germline mutations in these genes, the International Collaborative Group on HNPCC claims that risk for other cancers is also increased among mutation carriers, including cancers of the stomach, ovary, urinary tract, brain (glioma), hepatobiliary tract and skin (sebaceous tumors)[75]. Lynch-associated colorectal neoplasms tend to develop at younger ages and progress more rapidly compared with sporadic CRCs, requiring specialized surveillance[75]. However, only four HNPCC-associated cancers (cancer of the endometrium, small bowel, ureter or renal pelvis) are included in the AC-II.

Even though the terms HNPCC and LS have been often used as synonyms, it is now suggested that LS is defined as a cancer syndrome caused by a defect in one or more of the MMR genes, while HNPCC is defined based on the fulfillment of the AC[99]. Families who fulfill the AC or the more lenient revised Bethesda Guidelines (which are only used for clinical purposes in order to increase diagnose sensitivity) comprise a risk group and should be tested for MMR deficiency. Sequencing the MMR genes in germline DNA is necessary in order to identify the exact disease-causing germline mutations (point mutations, small insertions, splice site alterations and deletions), but sequencing cannot identify mutations in which the entire gene or whole exons are deleted. These deletions can be identified by multiplex ligation-dependent probe amplification (MLPA). In addition, NGS targeted sequencing platforms have been developed for use in clinical settings. Other strategies applied prior to sequencing include testing for MSI or absence of the MMR proteins by IHC in the tumors. MSI-status is assessed using a panel of microsatellite markers, classifying tumors as MSI-high, MSI-low or MSS. MSI-H or absence of MMR proteins are indicators of LS, but the diagnosis cannot be made just on this basis, because 10-15% of sporadic CRC's also exhibit MSI or absence of *MLH1* due to promoter hypermethylation[99].

Regarding the histopathological features, CRCs linked to LS have a predominant right-sided location and are characterized by poorly differentiated tumors, mucinous differentiation, an expanding growth pattern, abundant peritumoral and tumor-infiltrating lymphocytes, Crohn-like reactions and lack of dirty necrosis[101,102]. The gene expression signature of LS is characterized by upregulation of genes related to antigen processing and presentation, apoptosis, natural killer cell-mediated cytotoxicity and T-cell activation[103].

Selection criteria for HNPCC
Amsterdam I criteria (all must be fulfilled)
• At least 3 relatives with CRC. • 1 should be a first-degree relative of the other two. • At least 2 successive generations affected. • At least 1 of the cancers diagnosed before age 50. • Tumors should be histologically confirmed. • Familial adenomatous polyposis should be excluded.
Amsterdam II criteria (all must be fulfilled)
• At least 3 relatives with CRC or an extracolonic HNPCC-associated cancer (cancer of the endometrium, small bowel, renal pelvis or ureter). • 1 should be a first-degree relative of the other two. • At least 2 successive generations affected. • At least 1 of the cancers diagnosed before age 50. • Tumors should be histologically confirmed. • Familial adenomatous polyposis should be excluded in cases of CRC.
Revised Bethesda guidelines (just one must be fulfilled)
• CRC diagnosed before age 50. • Presence of synchronous or metachronous CRC or other HNPCC-associated cancer regardless of age. • CRC with MSI-H in a patient younger than 60 years old. • CRC diagnosed and one or more first-degree relatives with an HNPCC-related tumor, with one of the cancers diagnose before age 50. • CRC diagnosed and 2 or more first or second-degree relatives with HNPCC-related tumors, regardless of age.

Table 1.1. Selection criteria for Hereditary Non-Polyposis Colorectal Cancer. The Amsterdam I or II criteria (AC-I/II) must be fulfilled by the family for their classification as HNPCC. The only difference between AC-I and II is that AC-II also takes into account extracolonic tumors in certain locations. The more lenient Bethesda criteria is only used to determine those high-risk families that should be tested for Lynch syndrome in the clinical practice in order to increase the sensitivity.

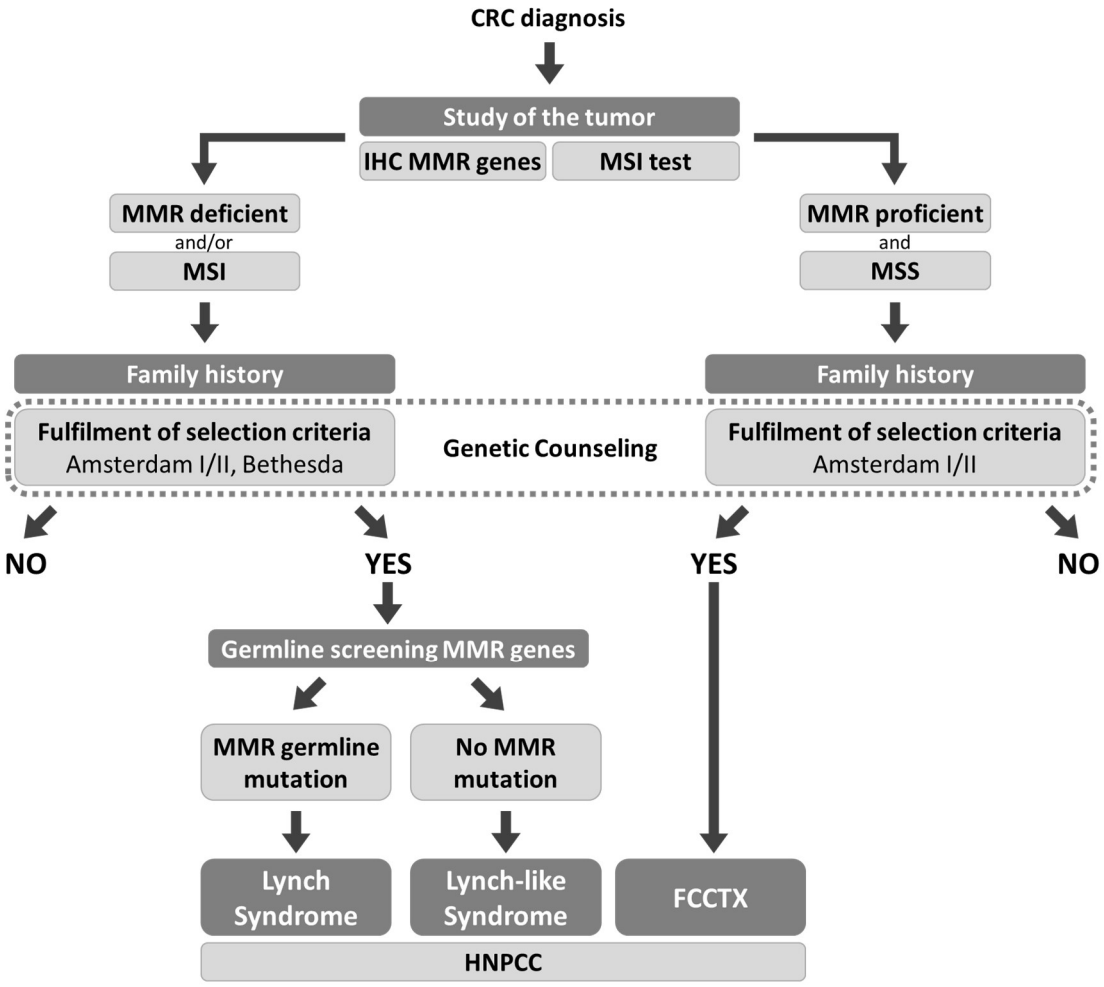


Figure 1.6. Pipeline for the diagnosis of the different HNPCC syndromes. This diagram is a simplification of the steps that follow a CRC diagnosis for its classification as one of the 3 HNPCC entities described to date. Note that other cases such as polyposis syndromes, familial CRC and sporadic CRC have not been included.

Lynch-like Syndrome:

There is a subgroup of HNPCC that presents MMR deficiency and the corresponding MSI in the tumors, but where MMR germline mutations have not been identified. This group of families has been classified as Lynch-like Syndrome, although some of these cases may be explained by false negative results caused by technical errors of the screening method. The arrival of Next Generation Sequencing has reduced the

amount families incorrectly classified, due to a higher efficiency than the previous methods. However, there is still a fraction of HNPCC families that remain under this category. These cases may be caused by promoter hypermethylation of the MMR genes (such as *MLH1*) or by any alteration that is not detectable by the commonly used mutational screening techniques, like variants affecting non-coding regions of the MMR genes. Certainly, many authors believe that patients with Lynch-like Syndrome have actually LS with germline mutations in MMR genes undetectable by current testing, although there are other mechanisms that could inactivate MMR genes resulting in tumors that resemble LS[104].

Familial Colorectal Cancer Type X:

The genetic basis underlying the cancer inheritance is unknown for the remaining 40-50% of HNPCC families, which have MMR-proficient tumors. This group, referred to as Familial Colorectal Cancer Type X (FCCTX), is characterized by being microsatellite stable (MSS) and by the absence of germline mutations in the MMR genes[99]. This group of families is the subject of study of this thesis and is therefore further explained in section 1.3.

1.3 Familial Colorectal Cancer Type X

1.3.1. Definition and terminology

Among patients with clinical characteristics of HNPCC, there is a group known as Familial Colorectal Cancer Type X or FCCTX that is defined based on the fulfilment of the Amsterdam I or II clinical criteria but presenting MMR-proficient, MSS tumors. FCCTX cases lack the MMR germline mutations that define LS, and therefore progress through different mechanisms of carcinogenesis[105]. The name FCCTX is adequate for an entity with an unclear carcinogenesis pathway and unknown genetic basis. Even though the first study that used the term FCCTX only included AC-I families[106], most groups currently include both AC-I and AC-II in their definition of FCCTX. Noteworthy, this group of MMR-proficient, MSS, HNPCC families had already been identified before this name emerged. For that reason, there is a remarkable diversity of abbreviations that have been used to designate these families over the years, including FCCTX, FCC-X, HNPCC-MSS, fCRC-X and MMR-p HNPCC, which hinders the literature search for this condition. However, the term FCCTX is currently the most widespread.

In addition to the MSS phenotype, various clinical, morphological and molecular features differ between FCCTX and LS. Unfortunately, despite the recent progress in clinical detection of FCCTX, its genetic etiology still remains unknown[105]. Current evidence suggests that FCCTX is a highly heterogeneous disease, presumably caused by germline mutations in novel high or moderate-risk cancer predisposition genes. However, it could also result from a combination of low-penetrance alleles in different genes increasing the cancer susceptibility in some families[107], or even from the aggregation of sporadic cases due to shared lifestyle factors[11,106]. This makes the identification of the genetic basis underlying FCCTX tumorigenesis a challenge that many groups are currently facing using different approaches. In addition, there is also heterogeneity regarding the clinical and histopathologic features, epigenetic, genomic profiles and gene expression profiles, and altered pathways[101].

The detection of the genes associated with FCCTX will facilitate the molecular diagnosis of the disease[105], and allow optimal counselling, screening and treatment, as well and prevention programs for other family members. Therefore, investigation of the genetic background of FCCTX is of great importance[99].

1.3.2. Clinical, morphological and molecular features

Clinical and histopathologic features:

FCCTX family members have been reported to show a lower risk for CRC compared to LS[106,108]. In addition, the age of onset in FCCTX patients shows considerable variability between families, and even between individuals within the same family[101]. However, it is well established that CRC is diagnosed at a higher mean age in FCCTX patients than in LS (60.7 years vs 48.7 years, respectively, as described by Lindor et al)[101]. Although this increase in the age at diagnosis is supported by all studies published to date, the specific mean onset age varies according to the study, ranging from 51 to 63 years for FCCTX and from 38 to 54 for LS[108–111,102,112–117]. In the same way, the mean age difference reported between the two syndromes ranges from as low as 3 years to up to 15 years depending on the study[99], but is generally believed to be closer to 10 years. Consistent with this, HNPCC families with higher diagnosis ages show a high proportion of MSS families[118].

On the other hand, many studies claim that FCCTX's tumor spectrum is predominated by CRC[117] and that there is a lower incidence of extracolonic tumors than in LS[108–110], some of them even suggesting that FCCTX families do not show any increased risk for other HNPCC-associated cancers[101,106]. This may be due to the fact that some studies have defined FCCTX as families fulfilling the Amsterdam I criteria with MMR-proficient tumors, while others also included Amsterdam II families[101]. Apart from the lower cumulative incidence of extracolonic cancers, their frequency, number of organs and mean age at diagnosis has also been reported to be lower in FCCTX than in LS[119].

However, unlike other clinical features, the lower risk for extracolonic tumors is not supported by all the groups[120].

Regarding the histopathological features, a number of studies have shown that CRCs from FCCTX and LS show different morphological features[105]. In the first place, CRCs linked to FCCTX are characterized by a predominant distal location[108–111,102,112,113,115,116,121], with 65-75% of the cases developing in the left colon or rectum, which differs from the right-sided preference of LS CRCs[117]. In addition, FCCTX tumors show a higher adenoma/carcinoma ratio[109,121], a lower frequency of multiple CRCs (synchronous or metachronous)[109,121] and more frequent polyp detection (including synchronous or metachronous adenomas)[101,116], which may indicate a slower adenoma-carcinoma progression rate in FCCTX[101,105]. Moreover, FCCTX is associated with a more sporadic-like phenotype, with a higher tumor differentiation[111,102,112,113,121], glandular and infiltrative growth patterns (as opposed to expanding)[101,102], and a higher degree of dirty necrosis (glands filled with necrotic debris) and tumor budding compared to LS[102]. Other features observed in FCCTX tumors include a lower rate of lymphocytic reactions (including peritumorous and tumor-infiltrating lymphocytes[102,111,112,121] and Crohn-like reactions[102,112]), a more heterogeneous and frequently tubular architecture[102] and a lower frequency of mucin production[110,111,102,121]. Unfortunately, despite these generally observed morphological trends, FCCTX tumors lack distinct histopathologic features. This makes the pathological identification of FCCTX-associated CRCs challenging, implying that these cases cannot be recognized based only on histopathological features and that family history needs to be taken into account[102]. Lastly, despite the lower CRC cumulative risk described for FCCTX family members, it has been suggested that this subgroup has a worse prognosis than LS[101], with a lower overall and disease-free survival[122]. Another study has reported that FCCTX's overall survival is similar to LS for stage II tumors, while between LS and sporadic CRC for stage III tumors (LS > FCCTX > sporadic CRC)[116].

Epigenetic profile:

As already described, the CIMP pathway is characterized by CpG island hypermethylation, and is often associated to the MSI phenotype. Conversely, global DNA hypomethylation is

observed in 30-40% of CRCs and has been associated with poor prognosis, shorter survival, younger age of onset and familial CRC risk.[101]. As a matter of fact, a series of studies has demonstrated that FCCTX tumors are highly hypomethylated compared with other CRCs, including LS and sporadic MSS CRC[123,124]. While each of the studied groups showed lower methylation in the tumor than in healthy tissue, the lowest methylation was observed for FCCTX in both tissues[124]. For this reason, Chen et al even suggested a 5-gene marker panel targeted at certain loci as useful biomarkers to distinguish LS from FCCTX[123]. Actually, it has been reported that although most FCCTX tumors present some degree of CpG methylation, none of them showed a high methylation index[125]. Among the hypomethylated loci, *LINE-1* can be pointed out. *LINE-1* hypomethylation has been linked to familial CRC and specifically to FCCTX, with a decreasing trend for *LINE-1* methylation from sporadic CRC to LS to FCCTX[124]. Global hypomethylation can result in *LINE-1* activation, what is thought to interfere with chromosomal segregation leading to CIN[113]. The predisposition to *LINE-1* hypomethylation in FCCTX tumors provides evidence for a link between distinct molecular signatures and phenotypes associated with specific epigenotypes[101].

Genomic profiles:

The genomic profiles of FCCTX tumors also show some similarities to sporadic MSS CRCs, typically presenting from 6 to 8 copy number variations (CNVs)[101]. FCCTX tumors have been reported to show increased and distinct CNVs compared to MSI[126]. Some CNVs associated to FCCTX tumors include recurrent gains of chromosomes 5p, 7p, 7q, 8q, 13q, 17, 19, 20p and 20q, and losses of 8p, 15, 17p, 18p and 18q[101,126–128]. Probably the most extended of all these chromosomal alterations is the gain of the 20q region, which has been specifically linked to FCCTX tumors and is observed in 65-70% of these cases[127,128]. Several candidate target genes are located in this region, including *GNAS*, *AURKA*, *SRC*, *TOP1*, *NELFCD*, *ADRM1*, *ASIP*, *CDH26* and *HNF1A*[101]. Loss of chromosome 18 is another common alteration in FCCTX tumors, which is also observed in sporadic CRCs. Some candidate genes downregulated by this alteration include *SMAD2*, *SMAD4*, *DCC*, *SERPINB5* and *BCL2*, among other[101]. Gain of 20q and loss of chromosome 18 particularly

discriminated between FCCTX and LS tumors, suggesting different preferred tumorigenic pathways[128].

Another frequently observed alteration is high levels of genome-wide copy neutral loss of heterozygosity, detected in 32-40% of FCCTX tumors[101,127]. In addition, low frequency of chromosomal losses (14%) have been demonstrated in FCCTX tumors, which could indicate the involvement of DNA repair mechanisms still unidentified[101]. Mutations in several cancer-related genes such as *TP53*, *KRAS*, *BRAF*, *APC* and *CTNNB1* divide FCCTX tumors into two major groups: one-third of the tumors are characterized by stable genotypes with few genetic changes, retained membranous β-catenin expression and infrequent *TP53* mutations, and two-thirds of the tumors show frequent loss of tumor suppressor gene loci such as *APC*, *TP53*, *SMAD4* and *DCC*, somatic methylation of *APC*, *KRAS*, and *MGMT* and nuclear translocation of β-catenin[101]. These genetic subsets have been suggested to differ in clinical presentation: the first group predominantly develops tumors in the proximal colon at a lower age (mean 54 years), whereas the second group more often develop tumors in the distal colon and are diagnosed at a higher age (mean 59 years)[101].

Gene expression profiles and altered pathways:

Although studies aiming to clarify the molecular characteristics of FCCTX tumors have increased in the past few years, most gene expression studies in CRC have focused on the differences between sporadic – but not hereditary – MSI and MSS tumors. These show distinct profiles with significant deregulation of 65-2070 genes, including those involved in growth factor receptor pathways, transcription, cell cycle, DNA repair, chromatin structure, drug metabolism and chemoresistance[101]. Gene expression data from FCCTX tumors reportedly show some similarities to sporadic MSS CRCs, with upregulation of genes involved in cellular functions such as amino acid modification, enzyme-linked receptor signaling, growth regulation, DNA repair, vascular smooth muscle contraction and G protein-coupled receptor signaling[101]. A total of 2188 genes have been shown to have distinct expression in FCCTX when compared to LS, including *NDUFA9*, *AXIN2*, *MYC*, *DNA2* and *H2AFZ*[113]. This suggests that these two entities have deregulation of different

canonical pathways. For instance, genes regulating the G-protein coupled signaling are reportedly upregulated in FCCTX, while genes relating to the cell cycle, mitosis and oxidative phosphorylation are upregulated in LS[113].

The current understanding of the influence that gene expression profiles have in the development of FCCTX tumors suggests increased proliferation, reduced apoptotic activity, and enhanced migration and invasion, which could contribute to the poor prognosis suggested in this subgroup[101]. Candidate genes involved in proliferation and migration that have been described in FCCTX include in chromosome 20q genes *CDH26*, *SRC* and *ASIP*. On the other hand, upregulation of the SRC-dependent pathway (through *COX-2*/*PGE-2* signaling, *PTGER1* and *ANGPTL4*) may drive FCCTX tumor development during anaerobic conditions, what would explain the frequent dirty necrosis and the downregulation of aerobic oxidative phosphorylation metabolism genes (such as *ATP5L*, *ATP5A1*, *ATP5B* and *ATP5D*)[101]. In addition, copy number gains and EGFR-mediated activation of SRC have been correlated to migration and invasion[101]. A growth advantage may be also achieved in FCCTX tumors by biallelic loss of certain genes, such as *SMAD2* and *SMAD4* (which inhibit TGFβR-mediated growth), *BCL2* (which induces apoptosis) and *SERPINB5* (tumor suppressor activated by *TP53* that induces apoptosis and inhibits migration and invasion)[101].

The majority of FCCTX tumors (61-63%) showed membranous location of β-catenin, indicating normal regulation of the Wnt signaling pathway unike most LS tumors, which showed deregulation of this pathway (81%). In addition, 56-57% of FCCTX tumors with aberrant β-catenin presented APC mutations, while none harbored *CTNNB1* mutations[120,129]. Membranous localization of β-catenin was associated with chromosomal stability (CIN-), younger age of onset, right-sided tumors and infrequent *TP53* mutations, while aberrant β-catenin was associated with older age of onset, left-sided tumors and CIN+ phenotype[129]. Mutations in genes related to the Wnt pathway, such as GNAS, might only explain a fraction of FCCTX tumors[105].

1.3.3. Surveillance recommendations and management

Given the higher age at diagnosis and the lower CRC cumulative risk reported for FCCTX families, most professionals recommend in this group of families a less intense surveillance than for LS[101,130]. This means less frequent and more individualized surveillance[131]. Although the benefits of colonic surveillance are clear, since screening colonoscopies in asymptomatic FCCTX family members significantly decrease CRC risk[132], the interval between colonoscopies is debated for different hereditary CRC subgroups[131]. Surveillance programs in FCCTX are mainly targeted at CRC, with colonoscopies generally recommended with 3-5-year intervals, starting 5-10 years before the earliest age of onset in the family[101]. It has been suggested that FCCTX families should be managed similarly to families with CRC aggregation but diagnosed above 50, since they both have high CRC prevalence and develop high-risk adenomas[133]. Given that the genetic basis responsible for the increased susceptibility is not known, all members from a certain FCCTX family should undergo the same surveillance[101].

On the other hand, therapeutic strategies and surgical management differs in LS, FCCTX and sporadic CRC cases[134]. Patients with a higher risk should ideally be identified in advance in order to perform the best surgical management. However, when the molecular basis of family CRC aggregation are unknown, as is the case of FCCTX, surgical indications and techniques should remain as in standard cases, although an individualization of subjects and their families is mandatory considering age at diagnosis, family cancer history, and characteristic phenotypes[134].

1.3.4. Approaches to identify CRC risk genes/loci

Linkage analysis:

Genome-wide linkage analyses are used for relatively frequent high-risk alleles and have previously helped identify causal genes in monogenic diseases with classical Mendelian inheritance patterns[99]. The method is based on the observation that genes localized close to each other on a chromosome are inherited together and are thus in

linkage disequilibrium[99]. DNA polymorphisms can be used as markers to identify the locus of a disease-causing gene by determining the haplotype and recombination frequency in family members[99]. The logarithm to odds (LOD) score is used to statistically estimate how likely two loci are linked, with LOD>3 meaning a significant link and LOD<-2 an unlikely link[99]. However, this approach has some limitations, such as the loss of power in the presence of heterogeneity given that it is not possible to calculate correct LOD scores in heterogenic diseases, the requirement of multigenerational pedigrees of affected individuals and the potential misleading results misdiagnoses and phenocopies can give[99].

Genome wide association studies:

Genome wide association studies (GWAS) have increased over the last few years and are used for the identification of common low-risk alleles. These are a non-hypothesis driven method with a case-control design based on the 'common-disease, common-variant' model, where the cause of the disease is attributed to various risk-alleles with a relatively high frequency[99]. For this purpose, the prevalence of genetic markers covering both coding and non-coding regions is compared in affected and unaffected (control) individuals[99]. However, a sample size with sufficient statistical power is critical for the success of genetic association studies to detect causal genes in complex diseases, and GWAS require much larger sample sizes to achieve an adequate statistical power[135], which is this method's main limitation. In addition, GWAS is underpowered to detect associations in rare variants, given their small sample size[99].

Next generation sequencing:

Next generation sequencing (NGS) is used for the identification of rare variants in high or moderate-risk alleles, and has replaced previous "classical" approaches, such as Sanger sequencing. In NGS, DNA sequences of interest are captured and sequenced in millions of parallel reactions that are afterwards analyzed, which requires comprehensive bioinformatics work for the identification of causative mutations[99]. However, causal mutations may be discarded in the filtering process, for instance, by focusing on nonsense variations, if the causative mutation is a low/moderate penetrance missense variation,

which might be important in diseases with reduced penetrance such as FCCTX[99]. Since NGS is the method that was used in this thesis for the identification of putative causal mutations in FCCTX, the different NGS technologies and applications are further detailed in section 1.4.

1.3.5. Genetic susceptibility

Despite the high expectations created by NGS and other aforementioned approaches, identifying new CRC genes in FCCTX has remained a challenge[136]. This may be due to the lack of major high-penetrance genes, the heterogeneity of FCCTX families and the difficulty of demonstrating the causality of candidate genes. After high efforts to identify new genes that could potentially explain the dominant inheritance of FCCTX, only a few putative familial CRC genes have been proposed. However, most are extremely uncommon, while others may only moderately increase the risk for CRC, which complicates the assessment of their contribution to CRC predisposition[75,136–139]. So far, most novel candidate genes lack corroborative data[136], and information from additional mutation carriers is required to estimate risks and recommend surveillance measures[140,141].

Some of the genes that have been associated with FCCTX predisposition in recent studies include *RPS20*[140], *BRCA2*[142], *FAN1*[143], *GALNT12*[144], *BMPR1A*[145], *SEMA4*[146], *NTS*[147], *RASSF9*[147], *TGFBR1*[148] and *OGG1*[149]. However, it has been suggested that the only candidate gene that has shown consistent association with HNPCC is *RPS20*, which encodes a ribosomal protein[75,140]. Although scarce, available data suggest high penetrance for *RSP20* mutations and absence of extracolonic manifestations[75,140]. Other studies consider that given that *BRCA2* has the highest mutation rate and is one of the most crucial DNA repair genes, it will be soon considered a big role player in this type of cancer in comparison with other genes[105]. This well-known gene for its essential role in the predisposition to hereditary breast and ovarian cancer has more recently been linked to CRC risk[142], and is also associated with the fanconi anemia pathway. Interestingly, different studies have suggested an association between other genes from this pathway

(such as *FAN1,* BRIP1/FANCJ, FANCC, FANCE, REV3L/POLZ and FANCM) and CRC susceptibility in a considerable number of families[143,150,151].

Some of the other pathways in which reported FCCTX-associated genes are involved include the TGF-β pathway (*TGFBR1* and *BMPR1A*)[145,148] and base excision repair (*NUDT1* and *OGG1*)[149]. In addition, *SEMA4A* mutations were proved to increase MAPK/Erk and PI3K/Akt pathways and cell cycle progression[146]. Nonetheless, certain studies suggest that some of these genes associated with FCCTX are not major contributors of CRC susceptibility in FCCTX families[145,152]. That is the case of *BMPR1A* and *GALNT12*, even though mutations in the latter have been consistently associated with CRC[144]. In fact, that is probably the case of most of these genes and other genes that are still to be found, which implies that candidate susceptibility genes may explain individually a quite reduced number of FCCTX families.

On the other hand, a number of GWAS have addressed low-risk susceptibility loci in hereditary CRC, although not specifically linked to FCCTX. Several loci associated to CRC identified by GWAS include genes implicated in CRC development, such as *BMP2*, *BMP4*, *CDH1* and *RHPN2*[99]. *BMP2* and *BMP4* are part of the TGF-β family that signal intestinal stem cells through suppression of the Wnt/β-catenin signaling pathway, *CDH1* is also involved in the Wnt/β-catenin signaling pathway, while *RHPN2* is involved in the actin cytoskeleton and cell motility, which can promote cancer invasiveness through adherence junction formation[99]. Other candidate genetic variants have been reported in *CENPE*, *CDH18*, *GALNT12*, *ZNF367*, *HABP4* and *GABBR2* on chromosome 9, *GREM1* and *KIF24* on chromosome 15, and *BCR* at 22q11[101]. Current understanding suggests that the risk alleles identified are insufficient to independently account for FCCTX cases, but a combination of moderate and low-risk alleles could contribute to the familial aggregation[101].

1.4 Next Generation Sequencing

The main method on which this study was based is Next-Generation Sequencing – NGS – or massive parallel sequencing, a state-of-the-art sequencing technology that has revolutionized genomic research, increasing our understanding of the genetic basis of Mendelian diseases exponentially[153]. Certainly, the arrival of NGS has been a major breakthrough in the search for new predisposition genes that explain the inheritance of different cancer syndromes, including FCCTX[105,154]. However, NGS does not denote a single technique, but a diverse collection of post-Sanger high-throughput sequencing technologies extended in the last decade. These technologies include sequencing-by-synthesis, sequencing-by-ligation, ion semiconductor sequencing, and others[155]. Regarding its applications, NGS can be used for include *de novo* sequencing, who-genome or targeted resequencing, transcriptome and epigenome analysis, etc[156]. However, the three most common approaches for the search of genetic variants are: 1) whole-genome sequencing (WGS), which sequences the complete genome of an organism; 2) whole-exome sequencing (WES), which is targeted at the coding regions of the genome or exons; and 3) multigene panels, which targeted sequencing of a previously designed set of genes[99]. The benefits and drawbacks of each of these technologies and applications are discussed below

1.4.1. NGS technologies

Pyrosequencing: Roche 454

Roche 454 was he first commercially successful NGS system and is based on pyrosequencing technology, which relies on the detection of the pyrophosphate released during nucleotide incorporation. The DNA library is prepared with specific adaptors and then denatured into a single strand and captured by amplification beads, followed by emulsion PCR[156]. Then on a picotiter plate, one dNTP will complement the bases of the template strand and release pyrophosphate equaling the amount of incorporated nucleotide[156]. The ATP transformed from pyrophosphate drives the luciferin into

oxyluciferin and generates visible light, while the unmatched bases are degraded. Then another dNTP is added into the reaction system and the pyrosequencing reaction is repeated[156]. The biggest advantages of the Roche system are its speed, its high read length and its possible automatization. However, the high cost of reagents remains a challenge, and it has a relatively high error rate with polybases longer than 6 bp[156].

Sequencing-by-ligation: SOLiD

The Applied Biosystems' SOLiD (Sequencing by Oligo Ligation Detection) sequencer adopts a technology based on ligation sequencing. Here, the libraries are sequenced on a flowcell using an 8-base ligation probe, which contains a ligation site (1st base), a cleavage site (5th base) and 4 different fluorescent dyes (linked to the last base)[156]. The fluorescent signal is recorded during the probes' complementary ligation to the template strand and vanished by the cleavage of the probes' last 3 bases. After 5 rounds of sequencing, the sequence of the fragment can be deduced using ladder primer sets[156]. The main advantage of this method is its high accuracy, while its principal drawbacks are the short read length and its incapability of sequencing *de novo*[156]. A complete run can be finished within 7 days, and automation can be used in library preparations. Its applications include WGS, targeted resequencing, transcriptome research and epigenome[156].

Sequencing by synthesis: Illumina

Illumina systems adopt the technology of sequencing by synthesis. The most extended sequencer of this company is called HiSeq, which is comparable to the aforementioned systems. It addition, there are other platforms with different scales, including NextSeq and MiSeq, a compact sequencer that is small in size with fast turnover rates but limited data throughput. All of Illumina's instruments are based on the same principles, and given that the NGS platform used in this thesis was the HiSeq 2000, this technology is shown in Figure 1.7 and will be explained more in depth.

In the first place, the DNA library (which involves generating a collection of DNA fragments for sequencing) is typically prepared by fragmenting a genomic DNA sample

and ligating specialized adapters to both fragment ends. An alternative is called tagmentation, which combines the fragmentation and ligation reactions into a single step. Adapter-ligated fragments are then PCR amplified and gel purified. The library preparation will depend on the NGS application[157]. The DNA library prepared with fixed adaptors is then denatured to single strands and grafted to the flowcell, followed by bridge amplification to form clusters that contain clonal DNA fragments[156]. Before sequencing, the library splices into single strands with the help of a linearization enzyme, and then four kinds of nucleotides containing different cleavable fluorescent dyes and a removable blocking group complement the template one base at a time. Finally, the emitted signals are captured by a charge-coupled device[156]. Compared with Roche 454 and SOLiD, HiSeq 2000 has the lowest reagent cost and features the biggest output, while the SOLiD system has the highest accuracy, and the Roche 454 system has the longest read length[156].

Regarding the data analysis, HiSeq control system and real-time analyzer calculate the number and position of clusters based on their first 20 bases, based on which the output and quality of each sequencing is decided[156]. HiSeq 2000 uses two lasers and four filters to detect the four types of nucleotide, whose emission spectra have cross-talk resulting in images that are not independent and the quality of sequencing affected by the distribution of bases[156]. The standard sequencing output files of the HiSeq 2000 consist of *bcl files, which contain the base calls and quality scores in each cycle. These files are then converted into *_qseq.txt files by the BCL Converter[156].

Ion semiconductor sequencing: Ion Torrent

Ion Torrent has two platforms of different capacity based on semiconductor sequencing technology: the compact Ion Personal Genome Machine (PGM), and the larger Ion Proton. In these sequencers, when a nucleotide is incorporated into the DNA molecules by the polymerase, a proton is released. The instrument recognizes whether the nucleotide is added or not by detecting the change in pH[156]. The chip is flooded with one nucleotide after another, detecting no voltage if it is not the correct nucleotide. On the other hand, if 2 nucleotides are added, the voltage detected would be double[156]. PGM

was the first commercial sequencing machine that did not require fluorescence and camera scanning, resulting in higher speed, lower cost, and smaller instrument size[156]. Ion Torrent has a stable quality along sequencing reads and a good performance on mismatch accuracies, but rather a bias in detection of indels[156].

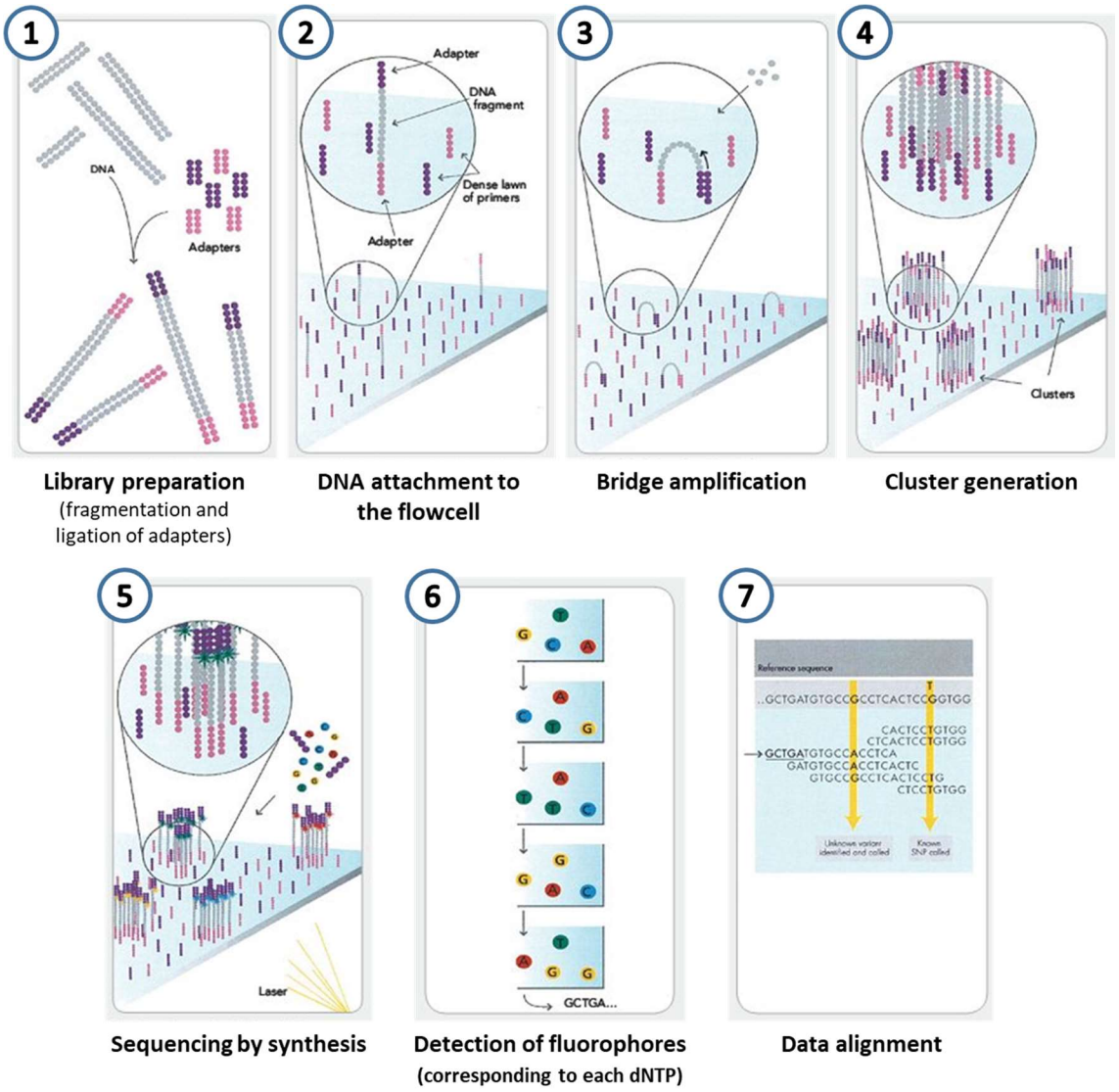


Figure 1.7. Sequencing-by-synthesis NGS technology used by the Illumina platforms. In the first place, the DNA library is prepared by fragmentation and ligation of specific adapters (1). Then, the DNA fragments are attached to the flowcell through those adapters (2) and the clusters are generated after several consecutive bridge amplifications (3 and 4). Finally, single strands are sequenced by synthesis of the complementary strand using fluorescent dNTPs (5). The fluorescence is detected after each round to obtain the sequence of each cluster (6), which is later aligned against a reference genome (7). Adapted from seqanswers.com.

Third generation sequencers:

Third-generation sequencing has two main characteristics. First, PCR is not needed before sequencing, which shortens DNA preparation time[156]. Second, the signal is captured in real time, which means that the signal – no matter whether it is fluorescent (Pacbio) or electric current (Nanopore) – is monitored during the enzymatic reaction of adding nucleotides to the complementary strand[156].

1.4.2. NGS applications for the search of genetic variants

WGS and WES:

WGS and WES are the two main NGS approaches for the search of new genes involved in Mendelian diseases. Although the latter is without doubt more extensively used, they each have their own advantages and disadvantages that should be borne in mind. The main drawback of WES is that it only covers coding regions of the genome, so any defects affecting regulatory regions such as promoters and enhancers would be missed. Conversely, WGS allows the identification of SNVs, insertions, deletions, structural variants and CNVs both in coding and non-coding regions of the genome. In addition, WGS has a more reliable and uniform sequence coverage than WES, given that differences in the hybridization efficiency of WES capture probes can result in regions of the genome with little or no coverage, what means that some information is completely lost. This disparity in coverage among the different areas of the exome obtained by WES also makes WGS a best option for the analysis of CNVs. Finally, WES may enrich preferentially one of the alleles at heterozygous sites, producing false negative SNV calls.

Nonetheless, despite these disadvantages, WES still has one huge advantage that explains its wider use: the fact that WES is targeted to protein coding regions involves that reads represent less than 2% of the genome. This not only reduces the cost of high-depth sequencing, but it also decreases storage and analysis costs. These reduced costs make it feasible to increase the number of samples to be sequenced, enabling large population

based comparisons. Moreover, with our current understanding of the human genome, at the moment the vast majority of data that would be obtained by WGS cannot be interpreted. For these reasons, so far WES has been undoubtedly the most widely used out of the two, although the rapidly decreasing prices suggest that WGS will be the way to go in the future.

Multigene panels:

As mentioned earlier, multigene panels are the third NGS option to search for causal mutations that could potentially explain the inheritance of Mendelian diseases. This targeted sequencing option is best suited for clinical practice, has a lower potential for the identification of new susceptibility genes, limited to those included in the panel. In spite of this, it is still a very effective method to attempt to clarify a part of the large amount of unexplained FCCTX families. And one of its main advantages is that it is even cheaper than WES, which allows the analysis of a higher number of samples.

Another undeniable advantage of multigene panels compared to WES is that you find less uncertain results, given that all the genes that are sequenced have already been linked to cancer predisposition. This eliminates the challenge of dealing with genes with still unknown functions or whose association with cancer is questionable. Therefore, this application extremely simplifies the filtering and prioritization steps, saving a lot of time in data analysis. It also allows a faster translation to the clinic of the results found. For that reason, multigene panels are useful for the screening of candidate genes previously identified by other approaches.

1.5 Genes of interest: SETD6

1.5.1. Lysine methylation

Covalent post-translational modifications (PTMs) of proteins serve as a major mechanism for the regulation of biological processes. Among these modifications, lysine methylation involves the transfer of up to three methyl groups to the ε-amine of a lysine residue, with the potential to modulate the function of the methylated substrate[158]. This modification has been found to regulate protein activity, protein-protein interactions and interplay with other PTMs, playing a vital role in the regulation of many cellular signaling pathways[158]. Interest in this PTM increased by the discovery that histone lysine methylation is involved in heterochromatin formation and that the methylation of histones can either suppress or activate gene transcription depending on which lysine is methylated, which can lead to dramatic effects on gene expression[158]. However, lysine methylation also occurs on various non-histone proteins both nuclear and cytoplasmic[158], but especially transcription- and chromatin-regulating proteins[159]. Some non-histone proteins known to undergo lysine methylation are p53, RB1 and STAT3, which have important roles in human tumorigenesis[160].

Lysine methylation is catalyzed by protein lysine methyltransferases (PKMTs) that transfer a methyl group from a donor S-adenosyl-L-methionine (SAM), which can be in turn removed by the less known demethylases[161]. The SET-domain PKMT superfamily includes all but one of the proteins known to methylate histones on lysine[162]. Even though there are over 70 PKMTs present in the human proteome, little is known about their enzymatic activity beyond histones[159]. The dysregulation of lysine methylation on both histones and non-histone proteins, mediated by dysregulation of methyltransferases and demethylases, is involved in the development and progression of various diseases, including cancer[160].

1.5.2. SETD6

SET domain-containing protein 6 (SETD6) is a member of the PKMT family that is known to monomethylate both histone and non-histone proteins and that has been recently found to regulate many essential biological processes. Figure 1.8 shows some of the main pathways in which this PKMT is involved. SETD6 was initially discovered to monomethylate the NF-κB transcription factor subunit RelA on lysine 310 (RelAK310me1), effectively inactivating it at the chromatin (Figure 1.8 B). Upon recognition of this methyl mark, histone methyltransferase GLP methylates histone H3 on lysine 9, promoting a repressed chromatin state and silencing the transcription of NF-κB target genes[163]. Given that deregulation of NF-κB is linked to pathologic inflammatory processes and cancer[164], SETD6-mediated inhibition of NF-κB signaling may influence tumor suppression and anti-inflammatory responses[165]. In contrast, it was recently reported that SETD6 promotes RelA's transcriptional activity in bladder cancer[166].

Noteworthy, SETD6 has also been reported to participate in the canonical Wnt signaling pathway by forming a complex with PAK4 and β-catenin at the chromatin (Figure 1.8 C)[167]. SETD6 methylates PAK4, which was recently shown to regulate β-catenin signaling[167]. In the presence of SETD6, the physical interaction between PAK4 and β-catenin is dramatically increased, leading to a significant increase in the transcription of Wnt target genes[167]. In this way, SETD6 acts as a key mediator of this pathway by the activation of β-catenin target gene transcription[167]. Interestingly, depletion of SETD6 was reported to significantly hinder the activation of Wnt/β-catenin target genes[167].

Another study demonstrated that SETD6 negatively regulates Nrf2-mediated oxidative stress response through a physical and catalytically-independent interaction with the oxidative stress sensor DJ1 at the chromatin (Figure 1.8 D)[168]. Under basal conditions, DJ1 is directly associated to SETD6 at the chromatin, which inhibits DJ1's activity and leads to the repression of Nrf2-dependent transcription[168]. In response to oxidative stress, Nrf2 expression is raised, SETD6 protein levels are reduced, the interaction between SETD6 and DJ1 at the chromatin weakens, and the transcription of Nrf2 antioxidant genes increases [168]. Depletion of SETD6 from cells results in elevated Nrf2 levels and a significant increase in Nrf2 antioxidant target gene expression[168].

Additionally, it has been suggested that SETD6 may play a role in embryonic stem cell differentiation through monomethylation of histone H2A variant H2AZ at lysine 7 (H2AZK7me1)[169]. H2AZ is an essential chromatin signaling factor significantly increased upon cellular differentiation, and is also a key component of nuclear receptor-dependent transcription. SETD6 was also reported to associate with the estrogen receptor α (ERα), histone deacetylase HDAC1, metastasis protein MTA2, and the transcriptional co-activator TRRAP, acting as a repressor of the estrogen-responsive-dependent transcription. HDAC1 and MTA2 are subunits of the NuRD complex, a group of associated proteins with chromatin remodeling and histone deacetylase activities[170]. However, *SETD6* also acts as a co-activator of several estrogen-responsive genes, such as *PGR* and *TFF1*, in the nuclear hormone receptor signaling pathway[170]. The same study showed that SETD6 is an essential factor for nuclear receptor signaling and cellular proliferation, and that its silencing induces cellular proliferation defects, enhances expression of the cell cycle inhibitor *CDKN1A* and induces apoptosis[170].

Last but not least, SETD6 has been recently identified to methylate PLK1, a key regulator of mitosis that is highly expressed in tumor cells, providing evidence that SETD6 is involved in cell cycle regulation[171]. During mitosis, SETD6 binds and methylates PLK1 on two lysine residues (K209 and K413). Lack of methylation of these two residues results in increased PLK1 kinase activity, leading to accelerated mitosis and faster cellular proliferation[171].

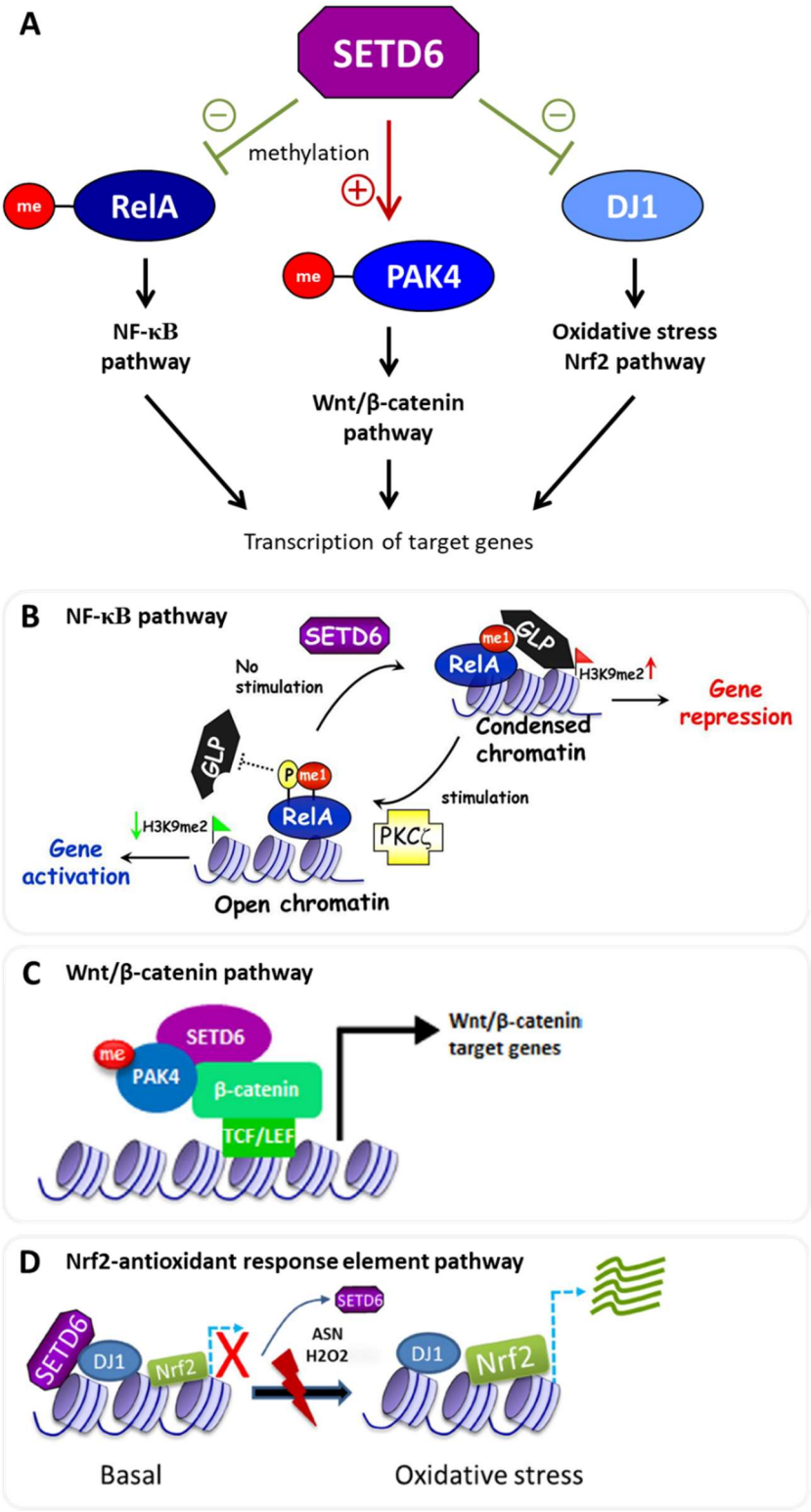


Figure 1.8. Different pathways in which SETD6 is involved. A) Schematic summary of the most relevant pathways regulated by SETD6, together with the effect it has on them: inactivation (-) or activation (+). B) SETD6 in the NF-kB pathway (obtained from lifewp.bgu.ac.il). C) SETD6 in the Wnt pathway (obtained from Vershinin et al. 2016). D) SETD6 in the oxidative stress or Nrf2-antioxidant response element pathway (obtained from Chen et al. 2016).

1.6 Genes of interest: PTPRT

1.6.1. Tyrosine phosphorylation

As previously mentioned, PTMs are involved in a wide variety of cellular activities and phosphorylation is one of the most extensively studied PTM[172]. Since it was discovered, tyrosine phosphorylation has been proven to be an essential mechanism of signal transduction and regulation in all eukaryotic cells, regulating many processes such as cell proliferation, cell cycle progression, metabolic homeostasis, transcriptional activation, neural transmission, differentiation, development and aging[173]. In addition, perturbations in tyrosine phosphorylation are involved in many human diseases, including cancer[173].

Tyrosine phosphorylation is coordinately regulated by protein tyrosine kinases (PTKs), which add a phosphate group to tyrosine residues of their substrates, and protein tyrosine phosphatases (PTPs), which remove such phosphate. The phosphate status of tyrosines is important in many signaling pathways underlying tumorigenesis, such as cell growth, proliferation, differentiation, cell cycle, apoptosis and invasion[174]. Therefore, genetic and epigenetic alterations in genes encoding PTKs and PTPs can result in changes to the equilibrium of kinase-phosphatase activity that might have a deleterious effect, producing abnormal cell proliferation, which could ultimately lead to cancer[174]. The human genome encodes 107 members of the protein tyrosine phosphatase superfamily, which are divided into four classes based on the amino acid sequence of their consensus catalytic domain signatures[174,175]. Class I PTPs comprises by far the biggest group, with 99 members that are divided into two subfamilies, consisting on 38 tyrosine-specific or "classical" PTPs and 61 dual-specificity phosphatases[174,175]. Classical PTPs strictly recognize phospho-tyrosine residues as substrates and can be further divided into two groups: 21 receptor PTPs (PTPRs) and 17 non-receptor PTPs (PTPNs)[174,175].

1.6.2. PTPs in human cancers

Cancer is a genetic disease that is driven by mutations in oncogenes and tumor suppressor genes. Many PTKs have been long known to act as oncogenes in cancer initiation and progression, while most PTPs have been proven to act as tumor suppressors, reversing the negative effects of PTKs[174]. However, some PTPs also have oncogenic properties[174]. Wang et al were the first to discover that PTPs are frequently mutated in CRCs when they identified 6 genes that contained recurrent somatic mutations, including 3 receptor PTPs (*PTPRF*, *PTPRG* and *PTPRT*) and 3 non-receptor PTPs (*PTPN3*, *PTPN13* and *PTPN4*)[176]. Mutations in these 6 genes affected 26% of the CRCs analyzed in the study, none of which were synonymous. It was later discovered that PTPs are somatically mutated in many other cancers, but play particularly important roles in colon and endometrium cancers[175].

Among all PTP genes, *PTPRT* is the most frequently mutated in human cancers, with somatic mutations identified in 11% of colon, 11% of esophagus, 10% of lung, 9% of stomach, 8% of endometrium, 6% of bladder and 6% of head and neck cancers[175]. Moreover, *PTPRT* is also mutated in a smaller fraction of leukemia, breast, ovary, liver, pancreas and prostate cancers[175]. Promoter DNA methylation had been proposed to be another possible mechanism leading to loss of PTPRT function[175], but it was not until recently that *PTPRT*'s promoter was found to be frequently hypermethylated in CRC[177] as well as in head and neck cancer[178].

1.6.3. PTPRT

Protein Tyrosine Phosphatase Receptor Type T or Rho (PTPRT PTPρ), belongs to type IIB subfamily of classical receptor PTPs (class I), which includes PTPRK, PTPRM and PTPRU. This subfamily of PTPs share similar structures that consist of an extracellular region, a transmembrane domain and a specific intracellular structure (Figure 1.9). The extracellular portion of the type IIB PTPRs consists of a MAM domain, an immunoglobulin (Ig) domain and four fibronectin type III (FNIII) repeats. It has been shown that the FNIII

repeats of PTPRK, PTPRM and PTPRT mediate homophilic cell-cell adhesion, while the extracellular region of PTPRU lacks such activity[179]. Consistent with this, Besco et al showed that PTPRT interacts with adherence junction components, such as cadherins and catenins through this extracellular region[180]. Interestingly, most tumor-derived mutations located in the extracellular domain of PTPRT impair cell-cell adhesion[180].

On the other hand, the cytoplasmatic segment of this subfamily consists of a cadherin-like juxtamembrane domain and two phosphatase domains: D1 and D2 (Figure 1.9). It is generally believed that the membrane-proximal PTP domain (D1) is responsible for the tyrosine phosphatase activity per se, whereas the second is a pseudophosphatase domain (named D2) that has no phosphatase activity[181]. However, many tumor-derived mutations are located in the second catalytic domain[176], suggesting that this domain has an important structural function or harbors a still unknown enzymatic activity. As a matter of fact, both catalytic domains have been proven to be essential for the correct function of the protein[182]. While the D1 domain is responsible for the phosphatase activity of receptor PTPs, D2 is important for the regulation of such activity.

Regarding PTPRT's substrates, two main proteins have been reported to be modified by this enzyme: STAT3 and paxillin[182,183]. These two substrates are well-known oncogenes that are inactivated upon dephosphorylation by PTPRT, proving once more its role as a tumor suppressor gene (Figure 1.9). PTPRT dephosphorylates STAT3's well-characterized residue Y705[182], whose phosphorylation is key to STAT3's activation[184]. It has been shown that pY705 STAT3 is up-regulated in a variety of human cancers and that this phosphorylation plays an oncogenic role in tumor development[184]. The relevance of STAT3 Y705 phosphorylation in CRC was demonstrated by Zhang et al, who showed that STAT3 Y705F mutant CRC cells showed reduced tumorigenicity[185]. Furthermore, PTPRT-regulated STAT3 signaling appears to be critical for head and neck tumorigenesis in a recent study that showed that phospho-STAT3 is up-regulated in HNSCC tissues with PTPRT mutations[186].

In contrast to STAT3, the target site of PTPRT on paxillin was the previously uncharacterized residue Y88[183]. Evidence suggesting that the PTPRT-regulated paxillin Y88

phosphorylation may be crucial for colorectal tumorigenesis include the failure of paxillin Y88F knock-in CRC cells to form xenograft tumors in nude mice and the up-regulation of pY88 paxillin in a majority of colorectal tumor tissue[183]. In addition to STAT3 and paxillin, other substrates of PTPRT, including BRC and Syntaxin-binding protein 1[187,188], were identified in brain tissue where PTPRT is abundantly expressed. However, the relevance of these substrates in the context of cancer remains to be determined.

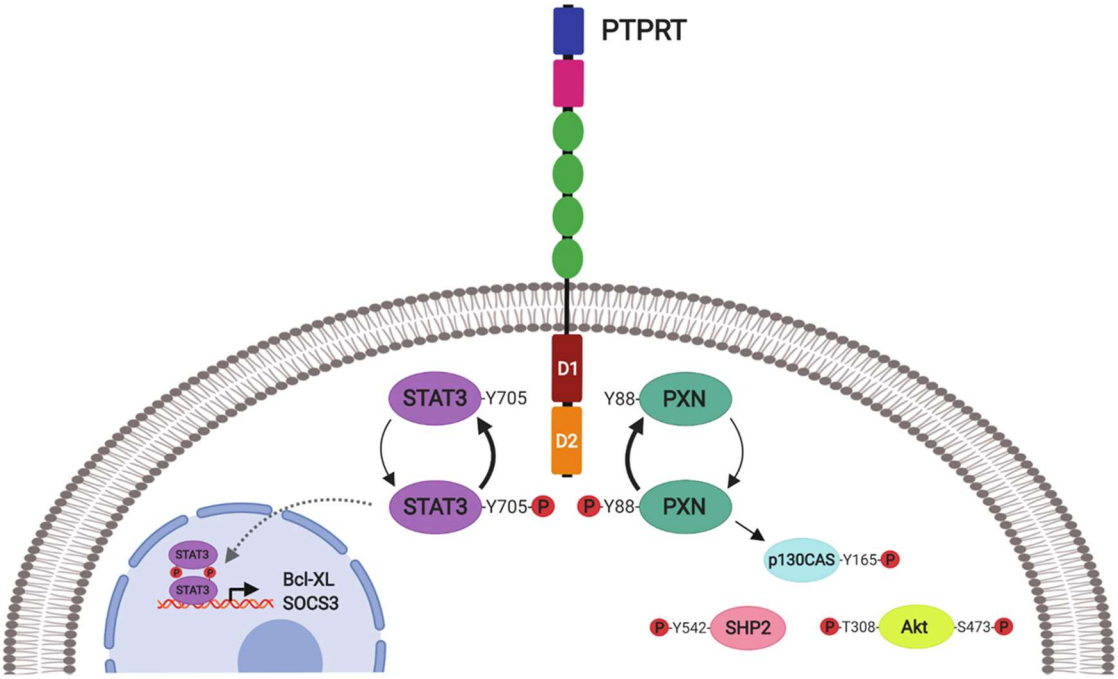


Figure 1.9. PTPRT target genes and pathways. PTPRT dephosphorylates and inactivates two well-known oncogenes, STAT3 and paxillin, resulting in the inhibition of the different downstream pathways in which they are involved. In this way, PTPRT acts as a tumor suppressor by inhibiting the expression of STAT3's target genes (such as Bcl-XL and SOCS3) and inhibiting the phosphorylation of paxillin's target proteins (such as p130CAS, SHP2 and Akt).

1.7 Genes of interest: PYGO1

1.7.1. The canonical Wnt signaling pathway

The Wnt pathway is a highly conserved signaling pathway that regulates many essential cellular processes, including cell fate determination, cell motility and polarity, primary axis formation, organogenesis and stem cell renewal[189]. However, the Wnt family of ligands can stimulate several intracellular cascades[189], based on which this pathway can be classified into a canonical or Wnt/β-catenin dependent pathway and a non-canonical or β-catenin independent pathway, which is in turn divided into a planar cell polarity and a Wnt/Ca^{2+} pathways[189].

When Wnt signals are transduced through the Frizzled (Fzd) family of receptors and LRP5/LRP6 coreceptors, the β-catenin signaling cascade is activated, which regulates cell fate determination[190]. This canonical pathway has a number of positive regulators, such as PAR-1, CKIε, and FRAT, as well as negative regulators, such as APC, AXIN1, AXIN2, CKIα, NKD1, NKD2, βTRCP1, βTRCP2, ANKRD6, NLK, and PPARɣ[190]. The hallmark of canonical Wnt pathway is the accumulation and translocation of β-catenin into the nucleus[189]. Conversely, in the absence of Wnt signals, cytoplasmic β-catenin forms a destruction complex with Axin, APC, PP2A, GSK3 and CK1α[189,190]. This allows its phosphorylation by CK1α and GSK3, which targets it for ubiquitination and subsequent proteolytic degradation by the proteosomal machinery[189] (Figure 1.10 A, left). On the other hand, binding of Wnt ligands triggers series of events that disrupt the APC/Axin/GSK3 complex and lead to the prevention of β-catenin degradation and its consequent stabilization and accumulation in the cytoplasm, followed by its translocation into the nucleus where it acts as a transcriptional co-activator[189] (Figure 1.10 A, right).

Many β-catenin binding partners have been described in the nucleus, including LEF/TCF, Legless/BCL9 and Pygopus (Pygo1 and 2), which influence the nuclear retention and transactivating ability of β-catenin for transcriptional regulation of its target genes. This complex binds to the promoter of the different Wnt target genes and activates their

transcription [189,190]. The canonical Wnt signaling plays a pivotal role in cell fate determination during early embryogenesis[190], and has been shown to play important roles in stem cell renewal and oncogenesis[189].

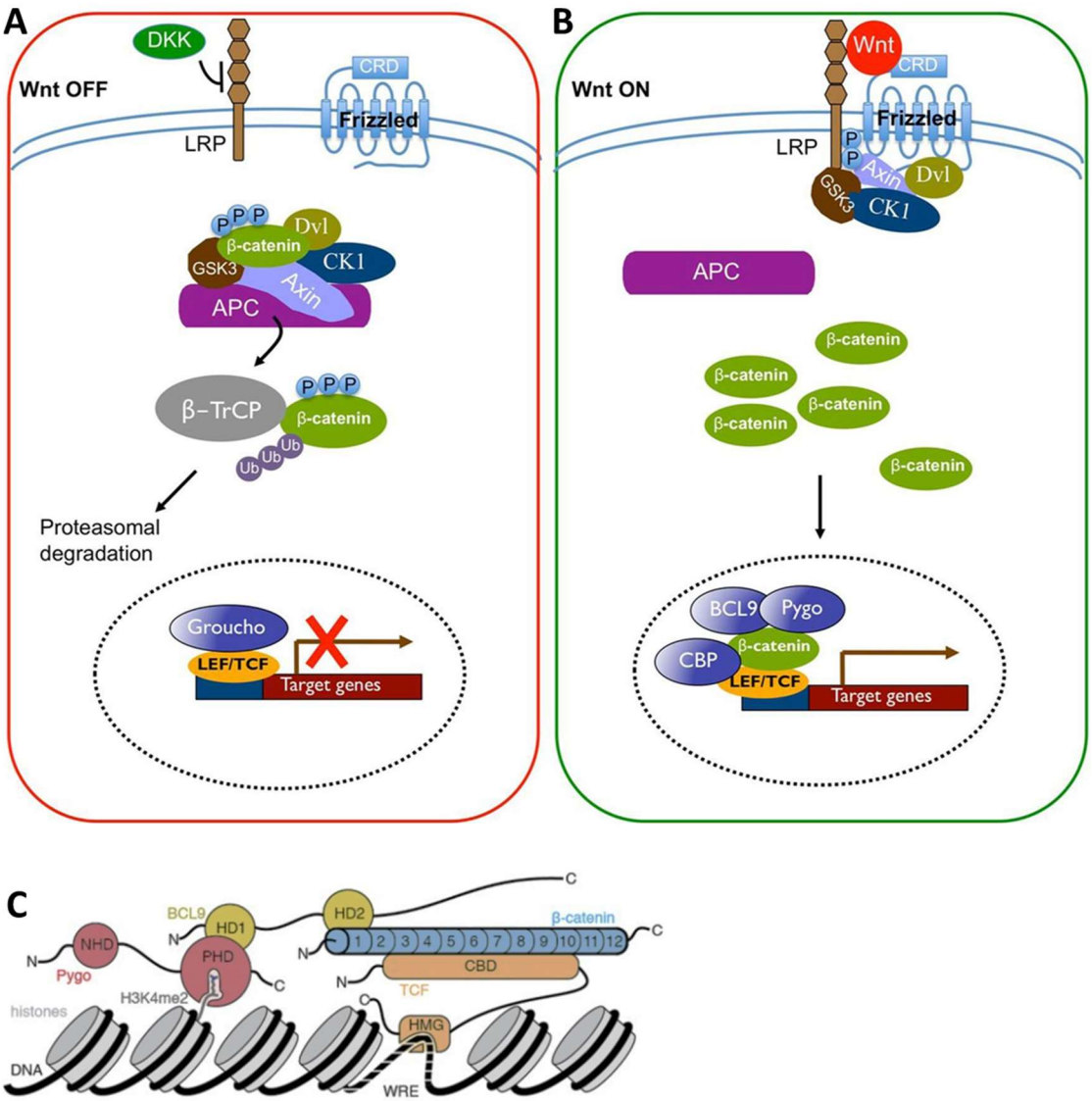


Fig. 1.10. Role of the Pygo proteins in the canonical Wnt pathway. A+B) Canonical Wnt signaling pathway in the presence (A) or absence (B) of Wnt ligands. When the Wnt pathway is inactive, β-catenin forms a complex with APC, Axin and other proteins and is degraded by the proteasomal machinery upon its phosphorylation and ubiquitination. When active, β-catenin expression is stabilized and accumulated in the cytoplasm before translocating to the nucleous, where it promotes the transcription of downstream target genes forming a complex with a number of nuclear proteins, including pygopus, BCL9 and TCF. Image obtained from Yu et al 2014. C) Schematic representation of the Pygo-BCL9 complex and its interaction with β-catenin N-terminus and methylated histone H3 tail (H3K4me2). Recruitment of β-catenin to Wnt target genes requires its binding to TCF factors (bound to specific enhancers through their HMG domain) but also its binding to Pygo-BCL9. Image modified from Miller at all 2014.

1.7.2. Wnt signaling in cancer

Epigenetic silencing and loss-of-function mutations of negative regulators of the canonical Wnt pathway occur in a variety of human cancers[190], and deregulation of Wnt pathway component is associated with a causal or progressive role in cancer[191]. Specifically, the vast majority of CRCs show a hyperactivated Wnt signaling, mainly due to the inactivation of the *APC* gene. Germline variants in this gene also predispose to CRC and are the cause of the most extended polyposis syndrome. In addition, it has been reported that common germline variants in the Wnt/beta-catenin pathway may be involved in CRC development. These variants may be informative in CRC risk assessment to identify individuals at increased risk who would be candidates for screening[192].

The Wnt signaling pathway, together with many other pathways, is implicated in the maintenance of tissue homeostasis by regulating self-renewal of normal stem cells, proliferation and differentiation of progenitor cells. Thus, misregulation of this signaling network leads to carcinogenesis[190]. For that reason, NSAIDs and PPARg agonists with the potential to inhibit the canonical Wnt signaling pathway are candidate agents for chemoprevention, and have been proven to decrease the risk for CRC[8]. In fact, anti-Wnt1 and anti-Wnt2 monoclonal antibodies have shown *in vitro* effects in cancer treatment, suggesting that, after their optimization, small-molecule derivatives and human monoclonal antibodies targeted to the Wnt signaling pathway could be used in cancer medicine[190].

1.7.3. PYGO1

Pygo1 is one of the two pygopus homologs present in mammals[191], and is a nuclear protein involved in signal transduction through the Wnt pathway. Pygo1 forms a complex with β-catenin, LEF/TCF, BCL9 and other components at the chromatin, allowing the transcription of Wnt target genes when the Wnt pathway is activated[193]. In this way, Pygo1 – together with Pygo2, BCL9 and B9L – contributes to efficient β-catenin-mediated transcription in Wnt stimulated mammalian cells and in CRCs with elevated Wnt pathway

activity. Wnt signaling outputs depend on histone decoding by the Pygo-BCL9/Legless complex, suggesting that this complex facilitates an early step in the transition from gene silence to Wnt-induced transcription[194]. However, it has been reported that, in mammals, the Pygo1/Pygo2 genes are not absolutely required for canonical Wnt signaling in most developing systems, but rather function as significant modulators of Wnt signal intensity[195].

Noteworthy, Pygo's interaction with β-catenin is not direct but mediated by BCL9[193]. Pygopus proteins are characterized by a Zn-coordinated PHD finger domain that is critical for the β-catenin-dependent transcriptional switch in normal and malignant tissues[196]. This PHD domain associates with its analogous HD1 domain from BCL9 in order to bind specifically to the histone H3 tail methylated at lysine 4 (H3K4me), preferentially when di-methylated (H3K4me2)[194] (Figure 1.10 B). Although histone-binding residues are identical between the two human Pygo paralogs, Pygo2 complexes exhibit slightly higher binding affinities for methylated histone H3 tail peptides than Pygo1 complexes[196]. BCL9 co-factors binding to Pygo PHD fingers impact indirectly on their histone binding affinity through an allosteric communication, and a highly conserved tryptophan is the signature residue defining the PHD subclass of Zn fingers[197]. Efficient histone binding requires this PHD-HD1 association [194].

As previously mentioned, a vast majority of CRCs bear mutations in APC or β-catenin, resulting in an abnormally high Wnt signaling activity. Knockdown of the pygopus genes in CRC cells containing a mutant APC reduces Wnt reporter gene expression, suggesting that endogenous pygopus proteins modulate signaling output in these cancer cells[191]. Interestingly, experimental evidence has suggested that the pygopus proteins may be involved in neoplastic transformation of multiple cells types[191]. The Pygo-BCL9 complex is a chromatin reader, facilitating β-catenin-mediated oncogenesis, and is thus emerging as a potential therapeutic target for cancer, with small molecules that block protein−protein PHD finger interactions, such as a set of benzothiazoles that bind to the PHD-HD1 interface [197].

2. Hypothesis and Objectives

2.1 Hypothesis

The hypothesis on which the work presented in this thesis is based is that an important number of FCCTX cases comprise yet-to-be-discovered genetic syndromes caused by mutations in high or moderate-penetrance CRC-predisposition genes. Due to the heterogeneity of this group of families, we believe that the best strategy is to address each family individually.

2.2 Objectives

2.2.1. Main objective

The ultimate aim of this study is the identification of new potential CRC-risk genes that can explain the increased cancer susceptibility in FCCTX families. This is meant to be accomplished by whole-exome sequencing of different members from each of the 13 studied families.

2.2.2. Secondary goals

Encompassed within this first objective are the following secondary goals, all this with the purpose of finding the potential cause of CRC heritability in at least a fraction of the studied families:

1) To obtain a list of potentially pathogenic candidate variants that may potentially explain the increased cancer susceptibility of each family subject to study.
2) To classify and characterize the candidate variants identified.
3) To perform functional and/or expression studies for a few selected cases in which it is possible and suitable, in order to further demonstrate the pathogenicity of the corresponding variants.

3. Materials and Methods

3.1 Subjects and sample collection

3.1.1. Study population

The present study included 13 FCCTX families selected from a larger cohort of 37 HNPCC-MSS families that had been recruited over the years at the Genetic Counseling Unit of Hospital Clínico San Carlos (Madrid). The 13 families were selected based on the availability of germline and tumor DNA from more than one patient per family, and are detailed in the Appendix (A1). All of the families fulfilled the Amsterdam I/II clinical criteria for HNPCC (Table 1.1) or a borderline AC-I/II criteria in which we allowed the earliest age at CRC diagnosis to be up to 52 years old. In addition, they all presented MSS tumors with normal expression of the MMR proteins, and did not carry any germline mutation in the MMR genes (Appendix, A2). Between two and three members of each family were selected for whole-exome sequencing, being at least two affected of cancer.

Other family members, whether healthy or affected, were also recruited when possible for segregation studies. In addition, the formalin-fixed paraffin-embedded (FFPE) tumor blocks from the probands and/or their relatives were obtained when available from the Anatomical Pathology Service of the corresponding hospital. Personal and family histories were obtained, and cancer diagnoses were confirmed by medical and pathology reports. The study was approved by the Institutional Review Board of Hospital Clínico San Carlos, and a written informed consent was signed by each participant.

3.1.2. Control population

Healthy individuals with no cancer family history were recruited from the Blood Bank of Hospital Clínico San Carlos (Madrid) and used as healthy controls. The FFPE tumor blocks of sporadic CRC patients were also selected for their use as CRC controls in some experiments. Moreover, FFPE blocks containing non-tumor colon tissue were used as healthy colon controls. A total of 100 germline controls, 10 sporadic CRC controls and 10

healthy colon controls were used. All of these subjects had previously signed a broad informed consent allowing the donation of their biological samples to the hospital's biobank and their use for research.

3.1.3. DNA and RNA extraction

At the time of their initial visit, 10ml of blood were collected from all the participants in BD Vacutainer® K2E (EDTA) Plus Blood Collection Tubes. Germline DNA and RNA were extracted from peripheral blood using the MagNA Pure Compact extractor system (Roche Diagnostics), according to the manufacturer's recommendations. The PAXgene® Blood RNA Kit and Tubes (PreAnalytiX) were used to extract germline RNA when the patient could not come to our hospital.

Tumor DNA and RNA were extracted from 7µm-thick FFPE tissue sections. A hematoxylin and eosin-stained (H&E) section of each block allowed the assessment of tumor cell area and content by two experienced pathologists. Tumor DNA was extracted using the QIAamp® DNA FFPE Tissue Kit from Qiagen, while tumor RNA was isolated employing the RNeasy® FFPE Kit (Qiagen), according to their corresponding protocols.

3.1.4. Quantification and quality assessment of nucleic acids

The NanoDrop® ND-1000 Spectrophotometer was used to assess the DNA and RNA quantity and quality (recommended A260/A280 ratio value >1.8 and A260/A230 ratio >1.9). However, for Next Generation Sequencing purposes, the quality of DNA samples was also tested by agarose gel electrophoresis (0.8% agarose in TAE buffer), and the concentration of the samples was measured in a Qubit® 3.0 Fluorometer (Life Technologies) using the Qubit™ dsDNA HS Asay Kit (Invitrogen).

3.1.6. Reverse transcription PCR

Reverse transcription polymerase chain reaction (RT-PCR) was performed to convert RNA into complementary DNA (cDNA) using the PrimeScript™ RT Reagent Kit (Perfect Real Time, Takara, Clontech), following the kit's instructions. For this purpose, RNA was previously normalized to 100ng/µl, except for the breast tissue samples, for which it could only be normalized to 50ng/µl. The RT-PCR reactions took place in a 2720 Thermal Cycler (Applied Biosystems), with a final volume of 10µl and 650ng of RNA (325ng of RNA for the breast samples).

In order to confirm the absence of genomic DNA (gDNA) in our cDNA samples prior to their use, a PCR targeting exons 2-3 of PALB2 was performed using 1µl of RT-PCR product, and subsequently run on an agarose gel (2% agarose in TAE buffer). This allowed the discrimination between gDNA and cDNA, with a PCR product of 152bp in the absence of intron 2 (cDNA) and a 269bp product when intron 2 was present (gDNA). The PCR was performed with an annealing temperature of 60ºC and 30 cycles of reaction (40 cycles when paraffin-embedded tumor samples were being tested). The primers used for this test and detailed PCR conditions are shown in the Appendix (A3).

3.2 Whole-exome sequencing strategy for the study of FCCTX

3.2.1. Whole-exome sequencing

Whole-exome sequencing (WES) was outsourced to two different companies: NIMGenetics (first 8 families) and Sistemas Genómicos® (last 5 families). The switch of company was made in order to remain with the Illumina sequencing system after the first company replaced their sequencers by another brand. In both places, the exome capture was performed using SureSelectXT Human All Exon V3 (51Mb, Agilent Technologies). The final library size and concentration were determined on an Agilent 2100 Bioanalyzer and a Qubit Fluorometer (Thermo Fisher Scientific), respectively. Finally, the library was sequenced on an Illumina HiSeq 2000 platform with paired-end reads of 101bp and a 50x average coverage depth, following the manufacturer's protocol. Images produced by the HiSeq 2000 were processed using the manufacturer's software to generate FASTQ sequence files. Reads were trimmed and subsequently aligned against the human reference genome version GRCh37/hg19 using the BWA software, creating the BAM files. Low quality reads, PCR duplicates and other sequences that could introduce major biases were removed using Picard-tools and SAMtools[198,199]. Variant calling was performed using a combination of two different algorithms (VarScan[200,201] and GATK[202]) and the identified variants were annotated and described according to the recommendations of the Human Genome Variation Society (HGVD)[203] and Ensembl[204] databases.

3.2.2. Variant filtering

WES data was then thoroughly filtered at our laboratory for each of the different families. The overall filtering strategy used for this study is summarized in Figure 3.1 and comprised the following steps for selection of those variants that were: 1) shared by all the affected members sequenced (allele frequency ≥0.25); 2) carried in heterozygosis

(allele frequency between 0.25 and 0.75); 3) coding (frameshift, inframe, nonsense, splicing or missense) and affecting autosomes; 4) rare (MAF≤0.01 in the general population and not present in 3 or more families); 5) damaging (deleterious effect predicted by at least 4 out of 5 *in silico* tools for missense variants, by 2 out of 2 for inframe variants and predicted to alter splicing for splice region variants by one tool); and 6) absent in healthy elderly relative sequenced (when applicable). On the other hand, an extensive literature search was performed in order to select genes previously associated with CRC predisposition, as well as CRC-susceptibility loci identified by GWAS studies. Coding variants in these genes detected in heterozygosis in all the affected members from each family were then subjected to a more lenient filtering strategy, with a MAF≤0.05, and a damaging effect on the protein predicted by at least 3 out 5 in silico tools for missense variants or 1 out of 2 for inframe variants.

3.2.3. *In silic*o studies

In the first place, the minor allele frequency (MAF) of each variant was checked in three different databases: 1000 Genomes Project[205], Exome Variant Server (EVS)[206] and GnomAD[207]. On the other hand, for every missense variant five online tools were used to predict the impact of each amino acid substitution on the protein: SIFT[208], PolyPhen[209], MutationTaster[210], PROVEAN[211] and finally Condel[212], which is in turn an optimized combination of two other tools, MutationAssessor[213] and FATHMM[214]. Variants were considered damaging when at least four out of these five tools predicted a deleterious effect. All of this was done with the help of the Ensembl Variant Effect Predictor (VEP)[215], and only the canonical transcripts were selected for convenience. When one or more of the predictors failed to give a result, only those variants described as damaging by at least two tools – and neutral by no more than one – were selected. The damage effect of inframe variants could only be predicted by two of the tools mentioned above (PROVEAN[211] and MutationTaster[210]). In this case, any variant with a deleterious prediction by both tools was further considered. Splicing variants were also analyzed *in silico* by the Human Splicing Finder (HSF)[216] in order to predict splicing alterations. These included splice donor/acceptor variants and splice region variants (near the splice site but

without affecting the donor or acceptor sites). Variants with no significant splicing alteration detected by any of the two algorithms used by the HSF tool (HSF and MaxEnt[217,218]) were discarded.

3.2.4. Variant prioritization

The variants that remained after the filtering step were then prioritized based on the gene or locus affected by each variant. For this purpose, all the genes affected by filtered variants were examined in the UniProt[219] and OMIM[220] databases in order to learn about the function of each protein and their relevance in the context of cancer. The Reactome[221] and PathCards[222] tools were then used in order to check them against a list of genes involved in relevant biological functions (e.g. DNA repair, cell cycle, DNA replication, proliferation, apoptosis and important signaling pathways such as the Wnt, MAPK, TGF-β, GPCR, Hedgehog, Notch, mTOR, BMP, PIP3-Akt and NF-κB pathways). This data was used in a first prioritization step to discard those missense, inframe of splice region variants without further evidence pointing to a potential role in cancer development.

For all the remaining genes, references were extensively reviewed using the Pubmed search engine[223]. The STRING database[224] was then consulted so as to check protein-protein interactions, while SMART[225,226] and cBioPortal's MutationMapper[227,228] were used to find out if any described domains were affected by each selected variant. Finally, when possible, the 3D structure of the genes affected by these variants was simulated by SWISS-MODEL[229] or downloaded from the Protein Data Bank[230] and then viewed by the Jmol 3D molecule visualization tool[231]. This was done in order to visualize the effects of a truncating mutation on the structure or to determine the location of a missense variant within the tertiary structure of the protein. After this thorough search, a number of candidate variants were selected based on the relevance of the gene and the location of the variant in the protein structure. This prioritization was especially important for missense variants, and included DNA repair genes, tumor suppressor genes, CRC

susceptibility loci, genes from pathways involved in cancer progression or genes previously associated with cancer.

3.2.5. Validation, segregation and loss of heterozygosity studies

All the candidate variants selected were validated by PCR followed by Sanger sequencing of the corresponding region of each gene using germline DNA. Specific primers for each variant were designed using the online tool Primer3[232] and synthesized by Isogen Life Science (HLPC purification)[233]. The PCRs were carried out in a 2720 Thermal Cycler (Applied Biosystems) using the FastStart™ Taq DNA Polymerase (Roche), with a final volume of 25µl and 30 cycles of reaction. All the primers used, together with the PCR conditions, are summarized in the Appendix (A3). PCR products were analyzed by agarose gel electrophoresis (2% agarose in 0.5x TAE buffer) and subsequently sequenced. Sanger sequencing involved three different steps: thermal cycling, purification and capillary electrophoresis. In the first place, the sequencing reactions were performed in a 2720 Thermal Cycler and a final volume of 10µl, using the fluorescent BigDye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems), together with either a forward or reverse primer (Appendix, A3) and 1-3µl of PCR product depending on the intensity of the agarose gel band. Then, the extension fragments were purified for 30min with the BigDye® Xterminator Purification Kit (Applied Biosystems) and diluted 1:2 in sterile water. Lastly, capillary electrophoresis took place in an ABI 3130 Genetic Analyzer (Applied Biosystems).

The segregation and LOH studies were also assessed by PCR and subsequent Sanger sequencing. The segregation study was carried out in germline DNA from the available members of each family for all the candidate variants. Tumor DNA was only used in the case of deceased members. For the LOH analyses, tumor DNA was extracted from the FFPE tumor blocks available as described earlier and 40 cycles of reaction were used in the PCR step. Then, the electropherograms of the germline and tumor sequences were compared in order to detect possible variations in the peak height of any of the alleles.

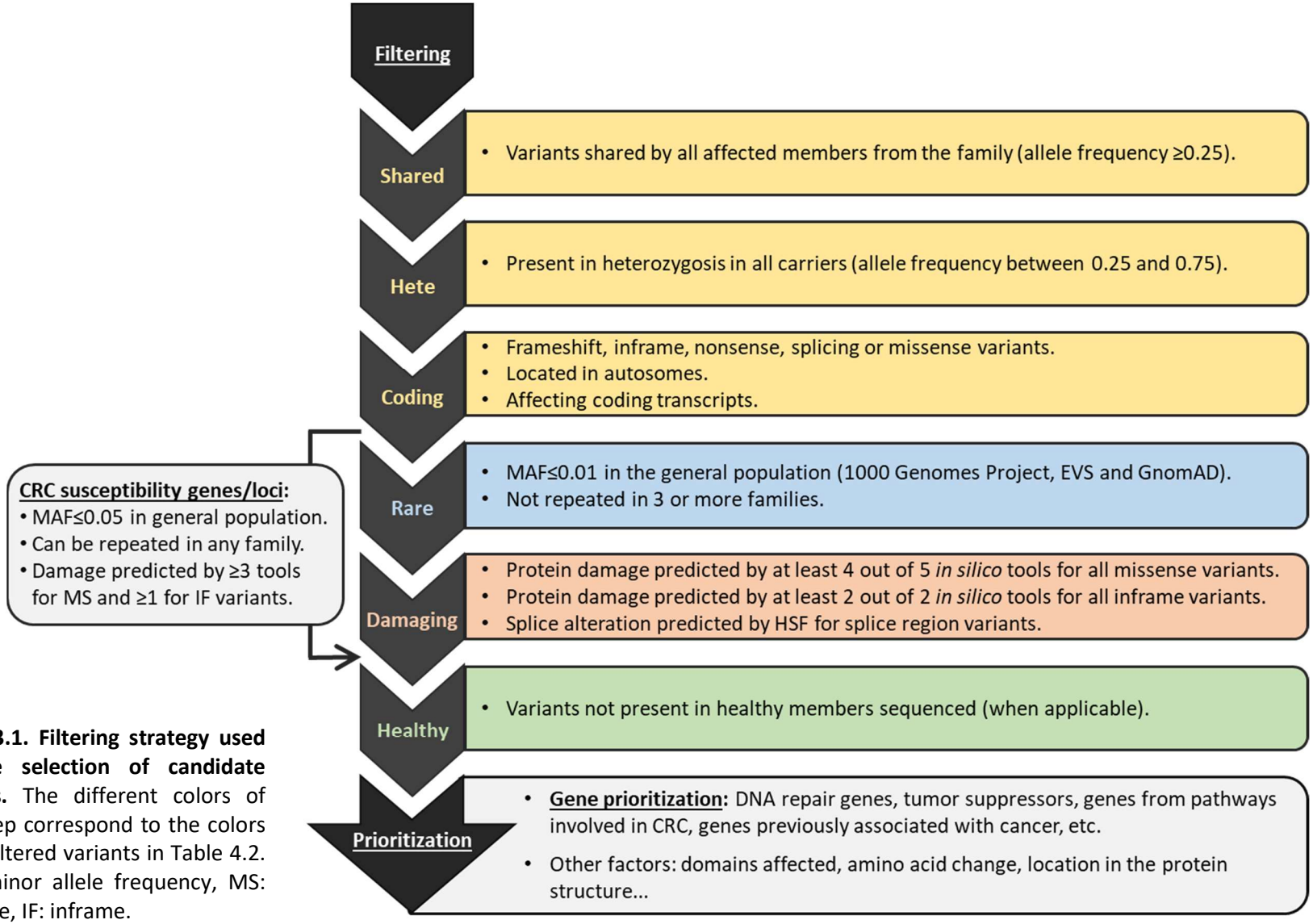


Figure 3.1. Filtering strategy used for the selection of candidate variants. The different colors of each step correspond to the colors of the filtered variants in Table 4.2. MAF: minor allele frequency, MS: missense, IF: inframe.

3.3 Other approaches to handle FCCTX families

3.3.1. Tumor whole-exome sequencing

The whole exome was also sequenced from tumor DNA of two members from two different families. This time, WES was outsourced to the CNAG (Centro Nacional de Análisis Genómico). The quality of the DNA was previously analyzed by agarose gel electrophoresis (1%) and the A260/A280 and A260/A230 ratios were obtained using Nanodrop. Then, the concentration was measured in a Qubit Fluorometer and 6µg of DNA were used for the library preparation. The exome capture was performed using Agilent's SureSelect XT V5, and the DNA library was subsequently sequenced in a HiSeq 2000 (Illumina) with 100bp paired-end reads and a 140x average coverage depth. After the generation of the FASTQ and BAM files, two different somatic variant callers were used, namely Mutect2[234] and Strelka[235].

For the data analysis, the results from the tumor WES were compared to the corresponding germline WES data. This was done with the aim of looking for germline variants that presented either loss of heterozygosity or a second hit in the same gene in the tumor. Therefore, the filtering strategy used for this study sought either the transition of heterozygous to homozygous variants (event 0/1 > 1/1) or the co-occurrence of a heterozygous somatic variant and a heterozygous germline variant in the same gene (0/0 > 0/1 + 0/1 > 0/1), with the allelic frequency of the germline variants in the general population being lower than 0.03 (ExAc[207]). The percentage of altered allele used for the filtering was calculated taking into account the tumor cell content of each tissue sample as determined by an experienced pathologist (70% for patient 275 and 40% for patient 2310).

3.3.2. Copy number variations

The BAM files obtained from germline WES were analyzed in order to look for DNA CNVs. This was achieved with the help of Dr. Juanjo Lozano, head of the Bioinformatics platform of CIBEREHD, who evaluated the number of reads of each area after correcting for biases using the R software[236]. In order to confirm all suspected structural variations, array-based comparative genomic hybridization (CGH) was used in some selected samples, allowing genome-wide CNV profiling. This was done in 7 samples (448, 772, 948, 2695, 6134, 6139 and 6230) in collaboration with IDIBAPS (Institut d'Investigacions Biomèdiques August Pi i Sunyer). The quality of every DNA sample was previously tested by agarose gel electrophoresis and by measuring the A260/A280 and A260/A230 ratios on Nanodrop. The latest version of Agilent Technologies' Human Genome CGH Microarray was then used to analyze 3µg of germline DNA, according to the instructions of the different kits employed. In the first place, the sample preparation was done using the SurePrint G3 Human CGH Microarray Kit. Then, the samples were labeled with different fluorescent dyes for the patient and reference DNAs with the help of the SureTag Complete DNA Labeling Kit. Finally, the Hybridization Gasket Slide Kit was used for the hybridization step, where the two DNAs competed to bind the microarray. The fluorescent signals were then measured in Agilent's SureScan Microarray Scanner System and the data was analyzed by Agilent CytoGenomics Software to generate a plot. For the analysis of the results, copy number ratio cutoffs of 1.25 and 0.75 were used for the identification of regions of gain and loss, respectively, and general population frequencies were also taken into account (MAF≤0.05).

3.4 Characterization studies

3.4.1. Digital PCR

For the allele-specific expression assays, a TaqMan digital PCR (dPCR) was carried out taking advantage of the QuantStudio™ 3D Digital PCR System (Applied Biosystems) and according to the manufacturer's recommendations. To this end, 15µL of a mix containing cDNA, QuantStudio® 3D Digital PCR Master Mix v2 (2x) and specific TaqMan probes (20x) were loaded into a dPCR chip (QuantStudio™ 3D Digital PCR Chip Kit) with the help of the QuantStudio 3D Digital PCR Chip Loader, according to the kit's instructions. The amount of sample used for the reaction was calculated and set up in order to ideally have one molecule of cDNA per well. After the chips had been thoroughly sealed, the dPCRs took place in a GeneAmp® PCR System 9700 (Applied Biosystems), with an annealing temperature of 60ºC and 40 cycles of reaction. Lastly, the fluorescence of each well was measured, allowing the absolute quantification of the samples.

The primers and TaqMan probes used for the dPCR were designed with the Custom TaqMan® Assay Design Tool (Thermo Fisher Scientific) and produced by the same company. The FAM probes specifically identified the mutant transcripts, while the VIC probes only recognized the wild type ones. The dPCR was used to analyze the cDNA from the paraffin-embedded tumors for two variants in SETD6 and PTPRT. Tumor cDNA from sporadic CRC patients served as non-carrier control, and cDNA from healthy colon tissue was also used as a control.

3.4.2. Quantitative real-time PCR

For the quantification of the overall gene expression of two *PTPRT* downstream target genes (BCL-XL and SOCS3), a quantitative real-time PCR (qPCR) was performed in a 7500 Fast Real-Time PCR System (Applied Biosystems). For this purpose, cDNA was analyzed in a 96-well plate with the TaqMan® Gene Expression Master Mix (Applied

Biosystems) and specific TaqMan probes designed by Thermo Fisher Scientific (BCL-XL assay Hs00236329_m1, SOCS3 Hs02330328_s1). PSMB4 was used as a housekeeping gene whose levels served as a reference (assay Hs00160598_m1), and a pool of 5 healthy colon cDNAs was used as a control. All samples were analyzed in triplicates, and the qPCR reaction took place with an annealing temperature of 60ºC and 40 cycles.

The qPCR results were expressed as a Ct value for each sample-probe combination. The Ct (cycle threshold) is defined as the number of cycles required for the fluorescent signal to exceed the background level. Those measurements differing in more than 1 unit with their corresponding replicates were discarded. The quantification of the relative target gene expression was calculated as $2^{-\Delta\Delta Ct}$, where ΔΔCt was the difference between ΔCt of the studied sample and ΔCt of the control pool, and ΔCt was the Ct value of the target gene minus the Ct of the housekeeping gene (PSMB4). The standard deviation was calculated for each sample.

3.4.3. Promoter methylation assay

Tumor DNA from two *PTPRT* mutation carriers was used for the promoter methylation assay. DNA obtained from healthy colon or breast was used as a control, as well as tumor DNA from sporadic CRC patients. The first step of this study was the bisulfite conversion of 1µg of DNA, which was done using the EpiTect® Bisulfite Kit (Qiagen) according to its protocol. A methylation specific PCR (MSP) was then performed with the EpiTect® MSP Kit (Qiagen) and using two different sets of specific primers targeting the promoter region of PTPRT. Each set was composed of 2 pairs of primers, which specifically detected methylated (M and M2) or unmethylated DNA (U and U2). These primers had been previously described by Laczmanska et al.[177] (in CRC, set 1) and by Peyser at al.[178] (in head and neck carcinoma, set 2) (Appendix, A3). The MSP was carried out following the manufacturer's instructions in a final volume of 25µl, with an annealing temperature of 55ºC and 40 cycles of reaction. Finally, the PCR products were analyzed by electrophoresis in a 2.5% agarose gel, and the fractional methylation was quantified by densitometry using the ImageJ software.

3.4.4. Pyrosequencing

Pyrosequencing was used for the measurement of *PTPRT*'s promoter methylation. For this purpose, pyrosequencing primers were designed using the PyroMark Assay Design 2.0 software, consisting in 2 PCR primers targeting a fraction of PTPRT's promoter (with a biotinylated reverse primer) and a biosynthesis primer targeted between them. All the primers were HPLC purified and are shown in the Appendix, A3. DNA from CRC, healthy colon, breast cancer and healthy breast from *PTPRT* mutation carriers was obtained, followed by the bisulfite conversion of 1µg of DNA using the EpiTect® Bisulfite Kit (Qiagen) according to its protocol. A PCR was then performed using 1µL of converted product and a Taq polymerase (Thermo Fisher Scientific), with an annealing temperature of 48ºC and 45 cycles of reaction. Commercial high-methylated and non-methylated DNAs were used as controls (CpGenome Human Methylated/Non-Methylated DNA Standard, Millipore).

The sample preparation was performed using 10µL of each PCR product and PyroMark Annealing Buffer, streptavidin-coated Sepharose beads, Binding Buffer, Wash Buffer and Denaturation Solution (Qiagen), following the manufacturer's instructions and with the help of a vacuum prep work station. Finally, the pyrosequencing took place in a PyroMark Q96 MD sequencer (Biotage, Qiagen) with the help of the Pyro-Q-CpG Software and using PyroMark Gold Q96 reagent Kit (containing the PyroMark Enzyme Mixture, Substrate Mixture and dNTPs for the pyrosequencing reaction) and PyroMark Q96 HS Capillary Tips. The PCRs and subsequent pyrosequencing were performed in triplicates, and the relative methylation (%) and standard deviation were represented.

3.4.5. Histone binding affinity assay

The histone binding affinity of mutant Pygo1 was tested by isothermal titration calorimetry (ITC) at Mariann Bienz's laboratory (Cambridge University, United Kingdom) as described by Miller et al[196]. The ITC was carried out at 25°C with an iTC 200 Microcalorimeter (GE Healthcare) following dialysis of purified wild type and mutant His-tagged PHD_{Pygo1}-$HD1_{BCL9}$ and PHD_{Pygo1}-$HD1_{B9L}$ complexes in 25mM Tris (pH 8.0) and

100mM NaCl. Titrations consisted of 19 consecutive 2-µl injections of peptide solution (following a pre-injection of 0.5µl) into the protein at time intervals of 120s or 150s. The 15-mer H3K4me2 histone tail peptide was used as previously described[194], and its concentration was determined by amino acid analysis.

3.5 Characterization of a frameshift variant in SETD6

The functional characterization of a frameshift variant identified in *SETD6* was performed at Dan Levy's laboratory, in Ben Gurion University and The National Institute for Biotechnology in the Negev (Beer Sheva, Israel). This was done during a research stay that was possible thanks to a fellowship awarded by the Federation of European Biochemical Societies (FEBS).

3.5.1. Cloning and plasmids

Two mutant versions of *SETD6* were cloned into pcDNA3.1 plasmids (Appendix, A4), one mimicking the mutation identified in one of the families (*SETD6-N*) and another one resulting in the C-terminal half of the protein (*SETD6-C*). For this purpose, the two versions were amplified by high-fidelity PCR using the KAPA HiFi HotStart PCR Kit (Kapa Biosystems) according to its instructions, with an annealing temperature of 59ºC and a 1-minute elongation. A wild-type *SETD6* pcDNA3.1 plasmid provided by Dan Levy[163] was used as a template (Appendix, A4). Primers were designed with the help of SnapGene® and synthesized by Isogen Life Science (Appendix, A3). The forward primer included an AscI restriction site tail, while the reverse primer had a PacI restriction site tail attached. PCR products were run on a 1% agarose gel and the bands were extracted from the gel using the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel), according to the manufacturer's recommendations. Then, the extracted products and the vector were simultaneously digested with two restriction enzymes (AscI and PacI, New England BioLabs) for 2h at 37ºC, in a final volume of 20µL with 10µL of sample and 1.5µL from each enzyme. For the digestion of the vector, calf-intestinal alkaline phosphatase (CIP, New England BioLabs) was added after the first hour for an additional hour, so as to prevent self-ligation. The digestion products were run in a 1% agarose gel and the bands were extracted once again.

For the ligation step, 1µL of each product was run on a gel in order to evaluate the amount of vector/insert needed to achieve a 1:1 molecular ratio. The ligation reaction was carried out using the T4 DNA Ligase (Sigma-Aldrich) for 1h at room temperature, in a final volume of 10µL and following the manufacturer's recommendations. A self-ligation control containing no insert was also added. The ligation products were immediately transformed into 100µL of DH5α Competent Bacteria (Thermo Fisher Scientific) by the heat shock method (30min on ice, 45sec at 42ºC, 2min on ice), after which 0.90mL of warm LB media was added and the mix was incubated for 1h at 37ºC. The tubes were then centrifuged at maximum speed and 0.9mL of media were removed before the pellet was resuspended, plated on LB plates containing 100ng/µL of ampicillin and incubated overnight at 37ºC. Finally, plasmid DNA extracted from a few colonies using NucleoSpin Plasmid EasyPure (Macherey-Nagel) was validated by digestion with AscI/PacI followed by visualization on an agarose gel. Positive clones were then sequenced using a primer that targeted the T7 promoter of the plasmids for additional validation.

The pcDNA3.1 plasmids in which the two forms of *SETD6* (N and C) were subcloned for the different experiments included pcDNA3.1-FLAG, pcDNA3.1-HA and pcDNA3.1-GFP. Apart from these pcDNA3.1 plasmids, used for transfection and overexpression assays in human cells, the two versions were also subcloned into a pET-Duet plasmid (Appendix, A4) for their expression in E.coli and protein purification. Wild-type *SETD6* had already been cloned into all these plasmids with the same purposes. In the same way, RelA and PAK4 had been also cloned into pcDNA3.1-FLAG for the overexpression experiments, and pMAL-c2x or pET-Duet plasmids (respectively) for the expression and purification of recombinant proteins (Appendix, A4)[163,167].

3.5.2. Cell lines and transfection

Two different cell lines were used for the characterization of mutant *SETD6*: human embryonic kidney cells (HEK293T) and human colon carcinoma cells (HCT116). Both were maintained in Dulbecco's modified Eagle's medium (Sigma-Aldrich) with 10% fetal bovine serum (Gibco), 2mg/mL L-glutamine (Sigma-Aldrich), 1% penicillin-

streptomycin (Sigma-Aldrich) and 1% MEM non-essential amino acids (Sigma-Aldrich), and they were cultured at 37ºC in a humidified incubator with 5% CO_2. For transient transfection, Mirus transfection reagents (TransIT®-LT1 for HEK293T cells and TransIT®-X2 for HCT116 cells) were used according to the manufacturer's instructions, together with Opti-MEM serum-free medium (Gibco).

3.5.3. Western blot

For western blot analyses, cells were homogenized and lysed in RIPA buffer buffer [50mM Tris-HCl, pH 8, 150mM NaCl, 1% Nonidet P-40, 0.5% sodium deoxycholate, 0.1% SDS, 1mM DTT, and 1:100 protease inhibitor mixture (Sigma-Aldrich)] (Appendix, A5), except for the biochemical fractionation and chromatin immunoprecipitation experiments, in which the cells were lysed as described below. Samples were heated at 95ºC for 5min in Laemmli sample buffer (Sigma-Aldrich), run on an 8-12% SDS-PAGE electrophoresis gel, and then transferred to a polyvinylidene difluoride (PVDF) membrane. Membranes were blocked with either 10% skim milk in PBST or 5% BSA in TBST for 1h on a shaking platform, and subsequently incubated with primary antibody for another hour with agitation. After three washes with the corresponding buffer (PBST or TBST), a 30-minute incubation with HRP-conjugated secondary antibody and three additional washes, a 2-minute reaction with a chemiluminiscent substrate (EZ-ECL, Biological Industries) allowed the visualization of the proteins.

The mouse monoclonal antibodies used were: anti-FLAG M2 (Sigma-Aldrich, F1804), anti-HA (Millipore, 05-904), anti-GAPDH (Abcam, ab8245), anti-histone H3 (Abcam, ab10799) and anti-tubulin (Abcam, ab44928). The rabbit polyclonal antibodies used were: an HRP-conjugated pan methyl lysine antibody (ImmuneChem, ICP0502) and a specific antibody against monomethylated RelA-Lys310 developed by Levy et al.[163]. HRP-conjugated secondary antibodies (goat anti-rabbit and goat anti-mouse) were purchased from Jackson ImmunoResearch (111-035-144 and 115-035-062, respectively). Antibodies were diluted and prepared in PBST with 10% skim milk or in TBST with 5% BSA, according

to the manufacturer's recommendations. All the details regarding the antibodies used are summarized in the Appendix (A5).

3.5.4. Biochemical fractionation

Biochemical fractionation was performed as previously described by Mendez et al.[237], with the addition of a final resuspension of the chromatin pellet for 30min on ice in RIPA buffer with 1mM $MgCl_2$ and benzonase nuclease enzyme (Sigma-Aldrich). The chromatin fraction was obtained by the collection of the supernatant after low-speed centrifugation (5min, 1700g, 4ºC).

3.5.5. Recombinant protein expression and purification

The Escherichia coli BL21-derivative Rosetta host strain was transformed with pET-Duet plasmids containing the gene of interest (*SETD6* wt, *SETD6-N*, *PAK4* or *RelA*) and grown overnight in 3mL LB medium +100µg/mL ampicillin (37ºC, 220rpm). The culture was then expanded to 100mL LB medium and incubated at 37ºC until the absorbance (OD) reached values of 0.6-0.8, when it was induced with 1:10000 IPTG and left overnight at 18ºC and 220rpm. After IPTG induction, the bacteria were harvested by centrifugation (10min, 12000rpm, 4ºC), resuspended in cold lysis buffer (10mM imidazole, 1% PMSF, protease inhibitor cocktail, 0.1% triton and PBS) and then lysed by sonication on ice (25% amplitude, 1min 30s, 10s on/5s off). Finally, the lysate was centrifuged (20min, 4ºC, 18000 rpm) and filtered, and the His-tagged proteins were purified using an ÄKTA™ column.

3.5.6. ELISA

A high-binding 96-well polystyrene microplate (Greiner Bio-One MICROLON®) was coated with 2mg of the recombinant proteins of interest (His-PAK4, MBP-RelA or BSA) diluted in PBS. The plate was blocked with 3% BSA in PBST and subsequently covered with 0.5mg of the recombinant tested proteins (His-SETD6 wt, His-SETD6-N or His-SUMO as a control) diluted in 1% BSA in PBST. A rabbit polyclonal anti-SETD6 primary antibody[163] was

then added, followed by incubation with an HRP-conjugated secondary antibody (goat anti-rabbit, 1:2000; Jackson ImmunoResearch, 111-035-144). All the incubation steps were performed at room temperature with vigorous agitation for 1h, and followed by three washes with PBST. After the final washes, 100µL of TMB reagent were added to each well, succeeded after a few minutes by the same amount of 1N H_2SO_4, in order to stop the reaction. The absorbance at 450nm was then detected using an Infinite® M200 plate reader (Tecan). All samples were analyzed in duplicates.

3.5.7. Cell-free *in vitro* methylation assay

Cell-free *in vitro* methylation reactions with recombinant proteins took place in a final volume of 25µL, containing 4µg of substrate (His-PAK4 or MBP-RelA), 4µg (or increasing amounts for the competition assay) of His-SETD6 (either wt or N), 2mCi of 3H-labeled S-adenosyl-methionine (SAM) (AdoMet, Perkin-Elmer) and PKMT buffer (20mM Tris-HCl, pH 8, 10% glycerol, 20mM KCl, 5mM $MgCl_2$). The reaction tubes were incubated overnight at 30ºC and then resolved by SDS-PAGE electrophoresis and subsequent autoradiogram. For the immunoprecipitation followed by cell-free *in vitro* methylation, human cells were transfected with empty, FLAG-SETD6 wt, FLAG-SETD6-N, FLAG-RelA or FLAG-PAK4 pcDNA3.1 plasmids, and 24h post-transfection they were lysed with RIPA buffer and pulled down overnight with anti-FLAG M2 Affinity gel beads (Sigma-Aldrich, A2220) on a rotor at 4ºC. After two washes in RIPA buffer and another two in PKMT buffer, samples were subjected to an on-beads cell-free *in vitro* methylation assay as described above.

3.5.8. Protein-protein chromatin immunoprecipitation

Protein-protein chromatin immunoprecipitation was modified from a published protocol[238]. After cross-linking, cells were harvested and washed twice with PBS and then lysed in 1mL of lysis buffer (20mM Tris-HCl, pH 8, 85mM KCl, 0.5% Nonidet P-40, and 1:100 protease inhibitor mixture) for 10min on ice. Nuclear pellets were resuspended in 200µl of nuclei lysis buffer (50mM Tris-HCl pH 8, 10mM EDTA, 1% SDS, 1:100 protease inhibitor

mixture) for 10min on ice, and sonicated (Bioruptor, Diagenode) with high power settings for three cycles of 6min each (30s on/off). Samples were then centrifuged (20min, 13000rpm, 4ºC), and the soluble chromatin fraction was collected. The FLAG-labeled substrates present in the soluble chromatin were then immunoprecipitated overnight on a rotor at 4ºC, using 20µl per tube of anti-FLAG M2 Affinity gel beads (Sigma-Aldrich). The beads were then washed according to the published protocol, heated for 30min in Laemmli sample buffer at 95ºC, and resolved on 10-12% SDS-PAGE gels followed by western blot analyses.

3.5.9. RNA extraction, reverse transcription and qPCR

RNA was extracted from cells using the NucleoSpin RNA Kit (Macherey-Nagel), normalized to 100ng/µL and reverse-transcribed to cDNA using the iScript cDNA Synthesis Kit (Bio-Rad), according to the manufacturer's instructions. For the NF-κB target genes, cells were previously induced with 20ng/mL TNF for 1h. Real-time quantitative polymerase chain reaction (qPCR) was carried out in a LightCycler® 480, using the UPL probe library system (Roche). All samples were amplified in triplicates in a 384-well plate, using 0.2µL of cDNA, 0.45µL of each primer, 0.1µL of the corresponding UPL probe and LightCycler480 Probes Master Mix for a final volume of 12µL. The housekeeping gene *GAPDH* was used for the relative analysis of gene expression levels, and qPCR results were analyzed as previously described.

3.5.10. Viability assay

A viability assay of HCT116 cells after 24, 48 and 72h post-transfection with either SETD6 wt or SETD6-N plasmids was performed using the PrestoBlue® Cell Viability Reagent (Thermo Fisher Scientific), with a 30-minute incubation and according to the manufacturer's recommendations. The absorbance at 570nm was detected using an Infinite® M200 plate reader (Tecan), normalized to the 600nm values and plotted in a graph. All samples were analyzed in triplicates.

4. Results

4.1 Whole-exome sequencing strategy for the study of FCCTX

4.1.1. FCCTX study cohort

With the aim of identifying new genes involved in their increased cancer predisposition, a total of 33 members from 13 FCCTX families were selected for their subsequent study (Table 4.1). The pedigrees of the 13 families and additional information about the probands and other family members can be found in the Appendix (A1 and A2). The selected families fulfilled either the Amsterdam I or II criteria for HNPCC[239,240] (46.2% and 53.8%, respectively), except for 4 high-risk families selected as "borderline AC-I/II", in which the lowest age at diagnosis was 50 (cc81 and cc525), 51 (cc440) or 52 years old (cc763). These families were included based on a strong cancer inheritance.

In our cohort of 33 participants, 36.4% were male and 63.6% were female. In addition, 3 were healthy and 30 were affected of a colorectal or HNPCC-associated carcinoma, or a colorectal adenoma. Out of the 30 affected probands, 24 had developed CRC, 1 had an endometrial cancer, 1 had been diagnosed of a renal cancer and 4 had developed colorectal adenomas (without fulfilling the clinical criteria for polyposis). In addition, three of the CRC-affected patients also had other cancers: cc565-III:4 had an endometrial cancer besides the CRC, cc108-III:1 had two CRCs and two breast cancers, and cc7-III:1 had an additional breast cancer. The mean age at diagnosis was 54.3 years old when considering the earliest CRC or HNPCC-associated tumor (without taking into account the adenomas). Among the 4 patients that did not have a diagnosed cancer (cc108-III:2, cc406-III:1, cc406-III:4 and cc525-IV:1), 3 fulfilled the criteria for advanced adenoma (>10mm or villous component or high-grade dysplasia), and the fourth (cc525-IV:1) was selected based on the very early age at diagnosis. The patient with renal cancer (cc7-III:2) had also developed at least 4 colorectal adenomas at ≤50 years old.

			Exome 1				Exome 2				Exome 3				
Fam	**Crit**	**WES**	**ID**	**Tumor**	**Dx**	**Sex**	**ID**	**Tumor**	**Dx**	**Sex**	**ID**	**Tumor**	**Dx**	**Sex**	**Relationship**
cc7	AC-II	2[a]	III:1	CRC	42	F	III:2	ADE/RC	42/47	M					Sibling
cc28	AC-I	2[b]	II:1	CRC	67	F	III:1	CRC	48	M					Child
cc81	AC-I*	3[a,b]	II:5	CRC	66	F	II:6	CRC	69	F	II:7	CRC	50	F	Siblings
cc89	AC-I	2[a]	III:1	CRC	52	F	III:2	CRC	46	M					Sibling
cc108	AC-II	2[a]	III:1	CRC	38/44	F	III:2	ADE	32	M					Sibling
cc122	AC-I	2[a]	III:1	CRC	52	F	IV:1	CRC	28	F					Child
cc406	AC-I	3[a]	II:1	CRC	73	F	III:1	ADE	49	F	III:4	ADE	52	F	Child & niece
cc440	AC-II*	3[a]	III:1	CRC	51	M	III:2	CRC	60	M	III:3	NO		M	Siblings
cc525	AC-II*	3[b]	II:2	CRC	67	M	III:1	CRC	50	F	IV:1	ADE	30	M	Niece & grandnephew
cc565	AC-II	2[a]	III:1	CRC	49	F	III:4	EC/CRC	51/68	F					Cousin
cc598	AC-I	3[b]	II:1	CRC	64	M	II:1	CRC	56	M	II:3	NO		F	Siblings
cc763	AC-I*	3[b]	II:1	CRC	81	F	III:3	CRC	52	F	III:4	CRC	56	M	Daughter & nephew
cc765	AC-II	3[b]	II:6	CRC	70	F	II:7	EC	28	F	II:5	NO		F	Siblings

Table 4.1. Family members selected for exome sequencing. The different fields are specified below. Fam: family identification number; Crit: clinical criteria by which the family was classified (AC-I: Amsterdam I, AC-II: Amsterdam II, *borderline: earliest age of onset allowed up to 52 years old); WES: number of exomes sequenced in that family and company where they were outsourced (a: NINGenetics and b: Sistemas Genómicos); ID: identification number of each member referred to the pedigree (Appendix, A1); Tumor: tumor type (CRC: colorectal cancer, EC: endometrial cancer, RC: renal cancer, ADE: colorectal adenoma); Dx: age at diagnosis; Sex: gender of the patient (F: female, M: male); Relationship: relationship of the second and third members (Exomes 2 and 3) referred to the first one (Exome 1). Only HNPCC-associated carcinomas and colorectal adenomas are displayed in this table.

Regarding the location of the colorectal carcinomas and adenomas, 16.7% were located in the right colon, 16.7% in the left colon, 33.3% in the sigma and 33.3% in the rectum, which means that 83.3% had been developed in the distal colon or rectum. This percentage of distal location goes up to 87% when taking into account CRCs from other family members. On the other hand, there was a wide range of CRC stages among the probands, going from carcinoma in situ to stage IIIC, but none of the patients had metastasis at the time of diagnosis.

4.1.2. NGS data analysis and filtering

The whole exome was sequenced in at least 2 members from each of the 13 FCCTX families, and the data obtained was subsequently filtered. The filtering strategy used is summarized in the diagram of Figure 3.1. The number of variants that remained after each filtering step for the different families is shown in Table 4.2. There were a total of 1646 rare variants (MAF≤0.01), ranging from 34 to 207 per family. After the *in silico* predictions, 532 variants were left, which were reduced to 460 when the healthy relatives sequenced in 3 of the families were taken into account (filtering out those variants that were present in the healthy individuals). This final selection ranged from 11 to 72 per family and comprised a total of 29 frameshift, 24 nonsense, 357 missense, 10 inframe, 13 splice donor/acceptor and 37 splice region variants. Note that 10 missense variants were splice region variants at the same time and are counted in both categories, although just counted once in the total number.

	INITIAL FILTERING				RARE							DAMAGING							ABSENT IN HEALTHY						
	Shared	Hete	Coding	Rare	FS	NS	MS	IF	SP a/d	SP r	Total	FS	NS	MS	IF	SP a/d	SP r	Total	FS	NS	MS	IF	SP a/d	SP r	Total
cc7	61403	24433	3292	148	3	3	133	4	2	4	148	3	3	39	0	2	3	49	3	3	39	0	2	3	49
cc28	51561	19327	3176	178	1[a]	1	169	2	2	8	178	1[a]	1	46	0	2	6	53	1	1	46	0	2	6	53
cc81	52907	13836	2896	122	1	2	95	8[b]	1	17	122	1	2	23	1	1	5	32	1	2	23	1	1	5	32
cc89	65331	26878	3525	207	3	3	192	8	1	4	207	3	3	61	3	1	3	72	3	3	61	3	1	3	72
cc108	66522	26152	3248	159	6[a]	3	142	5[c]	3	4	159	6[a]	3	36	1	3	4	50	6	3	36	1	3	4	50
cc122	46252	16428	2080	145	6	3	126	10[c]	0	2	145	6	3	37	2	0	2	50	6	3	37	2	0	2	50
cc406	36469	7000	1008	34	1	0	30	1	1	3	34	1	0	10	0	1	3	15	1	0	10	0	1	3	15
cc440	65157	27354	3525	196	5	4	177	6	2	5	196	5	4	60	2	2	4	77	4	1	33	2	1	4	45
cc525	40179	8982	1625	53	0	1	49	0	0	3	53	0	1	10	0	0	0	11	0	1	10	0	0	0	11
cc565	55830	17567	2193	39	0	2	32	5	0	0	39	0	2	11	0	0	0	13	0	2	11	0	0	0	13
cc598	50421	20691	3674	182	3	3	154	7	2	15	182	3	3	37	2	2	3	50	2	3	20	1	1	2	29
cc763	39169	8390	1471	37	1	0	35	0	0	2	37	1	0	11	0	0	2	14	1	0	11	0	0	2	14
cc765	50512	20544	3508	146	3	2	128	0	1	14	146	3	2	35	0	1	5	46	1	2	20	0	1	3	27
Total	681713	237582	35221	1646	33	27	1462	56	15	81	1646	33	27	416	11	15	40	532	29	24	357	10	13	37	460

Table 4.2. Number of variants remaining after each filtering step. The colors correspond to the ones used in the diagram of Figure 3.1. The first four columns (yellow) refer to the main initial stages of the variant filtering (Shared: shared variants between affected members of each family; Hete: variants present in heterozygosis in all the carriers; Coding: frameshift, nonsense, inframe or missense or splicing variants, affecting coding transcripts and located in autosomes; Rare: MAF≤0.01 in the general population and not present in more than 3 families). In the next 7 columns (blue), rare variants are separated by their effect (FS: frameshift, NS: nonsense, MS: missense, IF: inframe, SP a/d: splice donor/acceptor, SP r: splice region). The same separation is shown for damaging variants (red columns), which include variants predicted to be deleterious by at least 4 out of 5 (for missense) or 2 out of 2 (for inframe) *in silico* prediction tools and splice region variants predicted to affect the splicing by HSF. Finally, the green columns show the number of rare damaging variants not present in the healthy relative sequenced for the three families in which it was applicable (cc440, cc598 and cc765) a: Including one start-lost variant that produces a shift in the reading frame; b: Including one start-lost variant that does not produce a shift in the reading frame; c: Including one stop-lost variant.

4.1.3. Variant prioritization, validation, segregation and loss of heterozygosity

After this filtering, the variants were further prioritized according to the function and/or previous literature on the genes, the predicted alteration or domains affected by each variant and the frequency in the general population, in order to select the candidate variants that would be further studied. This step was particularly relevant for missense, inframe and splice region variants, while more lenient for frameshift, nonsense and splice donor/acceptor variants. Among the 460 filtered genes, a total of 207 passed a first prioritization step for having any previously reported association with cancer or for being involved in relevant pathways (see Materials and Methods). A more thorough analysis of the variants and the corresponding genes allowed the final prioritization of 44 variants that were successfully validated by Sanger sequencing, after which the segregation and LOH was studied when possible. Table 4.3 shows the results from these studies.

The segregation was used to discard those variants that did not segregate with the disease within the family (not carried by affected members with early onset ages as observed in germline DNA) or to further prioritize variants with a positive or compatible segregation. Only those prioritized variants that showed a positive, compatible or non-informative segregation were further considered as candidate variants, eliminating the ones that clearly did not segregate. Unfortunately, the segregation study was not very informative for most of the families and did not discard many variants. On the other hand, the LOH allowed the detection of significant reductions of either the wild-type or mutant allele in the available tumors of the carriers, but was only used as additional information to further prioritize the variants, not for selection purposes. Most of the tumors did not present a significant loss of any of the alleles, with the exception of those shown in the table as either LOH wt (loss of the wild type allele) or LOH mut (loss of the mutant allele).

4.1.4. Variants detected in CRC susceptibility genes

In order not to miss any potential causal mutation, all the genes that had been previously associated with CRC predisposition, as well as CRC loci identified by GWAS studies, were carefully screened separately from the rest with a more lenient filtering (Figure 3.1). The filtered variants in CRC susceptibility genes are shown in Table 4.4. This analysis allowed the identification of 2 missense variants in *POLE*: c.2083T>A and c.6494G>A. The former was carried by two families (cc28 and cc598) and affected POLE's polymerase domain (POLBc), but it was only predicted to be damaging by 3 out of the 5 *in silico* programs and, in addition, it was classified as benign in ClinVar with 9 submissions (2*). The latter, detected in family cc440, was predicted to be deleterious by all the *in silico* tools used, but did not affect any protein domain and was reported to be either benign or likely benign in ClinVar by 7 submitters (2*). In addition, three missense variants were detected in breast cancer genes previously associated with CRC[75,142]: *BRCA2* c.502C>A, *BRCA2* c.4258G>T and *PALB2* c.2993G>A. However, the two *BRCA2* variants were classified as benign in ClinVar by an expert panel (3*), while the *PALB2* variant was reported to be benign or likely benign by 15 submitters (2*). For these reasons, none of these SNPs were further considered as candidate variants.

Apart from these, only 2 other variants were identified in genes that had been previously reported to be potentially associated with CRC, but with much weaker evidence[136]. *WIF1* c.850A>G involved the substitution of two very similar amino acids[241] (Asn284Asp) not located in any specific secondary structure, while *LAMB4* c.368T>C was predicted to be damaging by just 3 out of the 5 in silico programs. Although both mutations affected protein domains and were rare in the general population, the involvement in CRC predisposition had not been supported for neither of them by their corresponding studies[136]. This, together with the weak predicted effect of these variants, prevented us from including them as candidate variants.

On the other hand, 7 variants affected genes previously associated with CRC through GWAS (Table 4.5). Out of these, 6 were missense and there was only one splice

donor variant. No other truncating mutations were found in CRC predisposition loci. Taking into account the frequency of the variants in the general population, the secondary structure and domains affected, and the relevance of the genes, none of the missense variants were further considered as candidate high-risk variants either. Another gene reported to have its expression altered in FCCTX (*NDUFA9*) also presented a missense variant that was not prioritized based on all of the above. Therefore, among all the variants detected in CRC genes/loci, only *ABCC2* c.1209+2T>G was further considered as a candidate variant.

Fam	Gene	Var Type	HGVSc	Variant	dbSNP	MAF	*In silico*	Segregation			LOH
								HNPCC tumor	ADE	Healthy >60y	
cc7	MUC6	NS	ENST00000421673	c.2109C>A; p.(Cys703Ter)	rs200217410	0,000911	N/A	3/3	N/A	N/A	No LOH
cc7	TXNRD3	FS	ENST00000524230	c.449dup; p.(Glu151GlyfsTer6)	N/A	0	N/A	3/3	N/A	N/A	LOH mut
cc7	GAL3ST2	SP d	ENST00000192314	c.29+1G>A	rs78620448	0,003662	Broken D	3/3	N/A	N/A	No LOH
cc28	CHID1	NS	ENST00000454838	c.3G>A; p.(Met1?)	rs767767312	0,000155	N/A	2/2	N/A	2/2	No LOH
cc28	PYGO1	MS	ENST00000302000	c.1084T>C; p.(Ser362Pro)	N/A	0	5/5	2/3[a]	N/A	1/2	No LOH
cc28	MMP11	MS	ENST00000215743	c.232C>T; p.(Pro78Ser)	rs61752251	0,000884	4/5	3/3	N/A	0/2	No LOH
cc28	AHRR	MS	ENST00000316418	c.680G>C; p.(Cys227Ser)	rs193290532	1,69E-05	4/5	3/3	N/A	2/2	No LOH
cc28	NKD2	MS	ENST00000296849	c.431T>A; p.(Met144Lys)	rs141899564	0,002881	4/5	3/3	N/A	2/2	No LOH
cc81	ECH1	NS	ENST00000221418	c.442C>T; p.(Arg148Ter)	rs201210987	0,000261	N/A	NP	NP	NP	NP
cc81	MAPK15	MS	ENST00000338033	c.106G>A; p.(Gly36Ser)	rs45495391	0,002921	5/5	3/3	NP	NP	NP
cc81	ATR	MS	ENST00000350721	c.2779A>G; p.(Ser927Gly)	rs1159930957	4,06E-06	4/5	2/2[b]	0/1	1/2	LOH mut
cc89	POLQ	FS	ENST00000264233	c.4262_4268del; p.(Ile1421ArgfsTer8)	rs546221341	0,005511	N/A	2/2	N/A	0/1	NP
cc89	C19orf40	MS	ENST00000588258	c.377C>T; p.(Ser126Phe)	rs36017455	0,005403	4/5	2/2	N/A	0/1	NP
cc89	ADIPOQ	MS	ENST00000444204	c.268G>A; p.(Gly90Ser)	rs62625753	0,002988	5/5	2/2	N/A	1/1	NP
cc89	NUP153	MS	ENST00000537253	c.50G>A; p.(Arg17Gln)	rs369696629	5,32E-05	4/5	2/2	N/A	NP	NP
cc108	NME7	FS	ENST00000367811	c.1106_1109dup; p.(Phe370LeufsTer8)	rs755143347	0,00031	N/A	1/1	2/2	N/A	No LOH
cc108	IQGAP2	NS	ENST00000274364	c.1282C>T; p.(Gln428Ter)	rs137988082	0,000218	N/A	1/1	1/2	N/A	NP
cc108	RASSF5	MS	ENST00000579436	c.890C>T; p.(Thr297Ile)	rs782806231	1,62E-05	4/5	1/1	1/2	N/A	NP
cc108	NEK4	MS	ENST00000233027	c.572C>T; p.(Ala191Val)	N/A	0	5/5	1/1	2/2	N/A	No LOH
cc122	MUS81	FS	ENST00000308110	c.1586_1587insGAGGT; p.(Gln531Ter)	N/A	0	N/A	2/2	N/A	N/A	LOH mut
cc122	VEPH1	FS	ENST00000362010	c.1346del; p.(Ser449IlefsTer34)	rs779965301	8,18E-06	N/A	2/2	N/A	N/A	No LOH (III:2) / LOH mut (IV:1)
cc122	RNMT	MS	ENST00000383314	c.938A>G; p.(Tyr313Cys)	rs753507842	1,22E-05	5/5	2/2	N/A	N/A	LOH wt
cc122	ABCB1	MS	ENST00000622132	c.158T>A; p.(Val53Glu)	N/A	0	4/5	2/2	N/A	N/A	No LOH (III:2) / LOH mut (IV:1)
cc122	CTNNAL1	IF	ENST00000325551	c.776_778del; p.(Glu259del)	rs768581936	1,3E-05	2/2	2/2	N/A	N/A	No LOH

Table 4.3. Candidate variants.

Fam	Gene	Var Type	HGVSc	Variant	dbSNP	MAF	*In silico*	Segregation			LOH
								HNPCC tumor	ADE	Healthy >60y	
cc406	SHF	FS	ENST00000290894	c.216_217insA; p.(Asn72LysfsTer42)	N/A	0	N/A	2/2	3/3	N/A	NP
cc406	ABCC2	SP d	ENST00000647814	c.1209+2T>G	N/A	0	Broken D	2/2	3/3	N/A	LOH mut
cc440	GALNT10	FS	ENST00000297107	c.417del; p.(Tyr140ThrfsTer53)	N/A	0	N/A	2/2	N/A	0/1	No LOH
cc440	PHACTR2	FS	ENST00000440869	c.620dup; p.(Asn207LysfsTer7)	rs756307843	2,05E-05	N/A	2/2	N/A	0/1	No LOH
cc440	CASP7	MS	ENST00000369321	c.767G>C; p.(Cys256Ser)	rs201040237	0,000184	5/5	2/2	N/A	0/1	NP
cc440	HELQ	IF	ENST00000295488	c.1764_1766del; p.(Lys588del)	rs772964480	2,05E-05	2/2	2/2	N/A	0/1	No LOH
cc525	ECT2L	NS	ENST00000541398	c.2227C>T; p.(Gln743Ter)	rs199963616	0,000987	N/A	2/2	2/2	0/1	No LOH (III:1) / Partial LOH mut (IV:1)
cc525	EIF4G3	MS	ENST00000602326	c.1026C>G; p.(Cys342Trp)	N/A	0	5/5	2/2	2/2	0/1	No LOH
cc525	C9orf116	MS	ENST00000429260	c.208G>A; p.(Val70Met)	rs533085223	0,000207	5/5	2/2	1/2	0/1	No LOH (III:1) / Partial LOH mut (IV:1)
cc565	ADAMTS 14	NS	ENST00000373208	c.2692A>T; p.(Lys898Ter)	rs761830741	4,07E-06	N/A	4/4	N/A	N/A	NP
cc565	ARMC3	MS	ENST00000298032	c.2333T>C; p.(Phe778Ser)	rs369926323	0,000196	4/5	4/4	N/A	N/A	LOH wt
cc565	RHOBTB1	MS	ENST00000337910	c.1942C>T; p.(Arg648Trp)	rs548990779	0,000106	5/5	4/4	N/A	N/A	LOH wt
cc598	CCDC60	NS	ENST00000327554	c.1558C>T; p.(Arg520Ter)	rs78597191	0,000558	N/A	3/3	N/A	0/1	No LOH
cc598	SETD6	FS	ENST00000219315	c.863_864insA; p.(Met288IlefsTer3)	rs768456822	0,001271	N/A	3/3	N/A	0/1	No LOH
cc598	ST18	MS	ENST00000276480	c.1708A>C; p.(Thr570Pro)	rs757938433	2,44E-05	4/5	2/2[b]	N/A	0/1	No LOH
cc763	WWOX	MS	ENST00000566780	c.358C>T; p.(Arg120Trp)	rs141361080	0,007547	5/5	3/3	N/A	0/1	No LOH
cc765	PTPRT	FS	ENST00000373193	c.4099dup; p.(Asp1367GlyfsTer24)	N/A	0	N/A	2/3[c]	N/A	0/2	No LOH (II:6) / Partial LOH mut (II:4)
cc765	MAP3K6	MS	ENST00000357582	c.2536C>T; p.(Arg846Cys)	rs146598623	3,66E-05	5/5	2/3[c]	N/A	1/2	No LOH (II:6)
cc765	ABTB1	MS	ENST00000232744	c.1213G>C; p.(Ala405Pro)	rs778524281	0	5/5	3/3	N/A	0/2	No LOH (II:6)
cc765	INVS	SP r	ENST00000262457	c.3017-5T>G	rs201018893	0,002786	Broken A	2/3[c]	N/A	0/2	No LOH (II:6)

Table 4.3. Candidate variants (cont.). Final selection of candidate variants per family. The different columns indicate the family, official gene name, type of mutation (NS: nonsense, FS: frameshift, MS: missense, IF: inframe, SP d: splicing donor, SP r: splicing region), transcript used for the nomenclature of the variant, variant name at cDNA and protein level, reference rs number of previously described variants, minor allele frequency (MAF) according to GnomAD, in silico prediction (number of tools with a deleterious prediction, out of 5 for missense and out of 2 for inframe and splicing variants, result from the segregation study (number of members carrying the mutation vs total) and LOH (LOH mut: loss of mutant allele; LOH wt: loss of the wt allele). a: 3rd is tumor DNA, b: 3rd was in tumor DNA and there is no left, c: 3rd was diagnose very late (85y), NP: not possible.

Family	Gene	V. type	Variant (c.)	Variant (p.)	Transcript	dbSNP	MAF	ClinVar	Pred	Domain	Notes
cc7	WIF1	MS	c.850A>G	p.Asn284Asp	ENST00000286574	rs768033163	0,000004	N/A	5/5	EGF	Association not clear
cc28, cc598	POLE	MS	c.2083T>A	p.Phe695Ile	ENST00000320574	rs5744799	0,011230	Benign (2*)	3/5	POLBc	CRC susceptibility gene
cc81	BRCA2	MS	c.502C>A	p.Pro168Thr	ENST00000380152	rs80358726	0,000037	Benign (3*)	3/5	NO	Breast cancer susceptibility gene
cc108	BRCA2	MS	c.4258G>T	p.Asp1420Tyr	ENST00000544455	rs28897727	0,006602	Benign (3*)	3/5	NO	Breast cancer susceptibility gene
cc108	PALB2	MS	c.2993G>A	p.Gly998Glu	ENST00000261584	rs45551636	0,016430	Likely benign (2*)	5/5	WD40	Breast cancer susceptibility gene
cc440	POLE	MS	c.6494G>A	p.Arg2165His	ENST00000320574	rs5745068	0,005857	Likely benign (2*)	5/5	NO	CRC susceptibility gene
cc440	LAMB4	MS	c.368T>C	p.Leu123Ser	ENST00000388781	-	-	N/A	3/5	LamNT	Association not clear

Table 4.4. Variants identified in previously described CRC predisposition genes.

Family	Gene	V. type	Variant (c.)	Variant (p.)	Transcript	dbSNP	MAF	ClinVar	Pred	Domain	Notes
cc7	TPD52L3	MS	c.401G>A	p.Gly134Glu	ENST00000344545	rs117022582	0,004959	N/A	5/5	TDP52	Reported to be testis-specific and involved in spermatogenesis.
cc28	TDGF1	MS	c.307G>A	p.Gly103Arg	ENST00000296145	rs554684465	0,000004	N/A	5/5	EGF	Higher expression associated with CRC and poor prognosis.
cc81	NDUFA9	MS	c.881C>T	p.Pro294Leu	ENST00000266544	rs34076756	0,016300	Likely benign (1*)	3/5	NO	Decreased expression in ecreased expression in FCCTX.
cc89	FSTL5	MS	c.578G>A	p.Gly193Glu	ENST00000306100	rs72689202	0,014950	N/A	4/4	EF hand	Down-regulated promotes cell proliferation through Wnt signaling.
cc108, cc440	ABCC2	MS	c.2546T>G	p.Leu849Arg	ENST00000370449	rs17222617	0,015570	Likely benign (1*)	3/5	NO	Involved in early stages of carcinogenesis.
cc122	CUBN	MS	c.8902G>C	p.Glu2968Gln	ENST00000377833	rs45569534	0,015070	VUS/Benign	3/5	CUB	Involved in vitamin absoption and renal and intestinal biology.
cc406	ABCC2	SP don	c.1209+2T>G	N/A	ENST00000370449	-	-	N/A	2/2	ABC transmb	Involved in early stages of carcinogenesis.
cc440	LAMB1	MS	c.2383C>G	p.Arg795Gly	ENST00000222399	rs80095409	0,011420	Benign (1*)	5/5	Laminin EGF	Higher expression in serum of CRC patients.

Table 4.5. Variants identified in genes associated with CRC in GWAS studies or genes relevant to FCCTX biology.

4.2 Other approaches to handle FCCTX families

4.2.1. Tumor whole-exome sequencing

In addition to the germline WES analysis, the whole exome was also obtained from tumor DNA for two of the families (cc7 and cc440). The corresponding tumor and germline exome pairs were then compared in order to detect germline variants that showed a second hit in the tumor (with either loss of heterozygosity or a second mutation in the same gene). However, this study did not detect any second hits or LOH at the genes affected by rare germline variants in these two families. Unfortunately, no more tumor samples could be studied due to the availability, quality and/or quantity of the paraffin-embedded tumor blocks.

4.2.2. Copy number variations

The raw data obtained by germline NGS also allowed a broad evaluation of the CNVs in the studied individuals. This analysis showed a possible CNV in chromosome 6, where some anomalies could be noted in various samples. However, it was likely that this occurrence was a false positive caused by a systematic error of the methodology, since the area was poorly covered and there were some similarities in samples outsourced to the same sequencing company. For this reason, a few representative DNAs with possible CNVs based on the NGS results (from families cc28, cc89, cc525 and cc763) were selected in order to verify if there was a real structural variation by CGH. Regrettably, no CNVs were detected in the studied samples, so the potential CNVs observed in the WES data could be ruled out by this study. However, we cannot discard the possibility of the existence of germline CNVs in other members of our cohort not analyzed by CGH, since they could have been missed by the WES analysis.

4.3 Candidate variants

As mentioned before, a total of 44 variants were finally selected as candidate variants for the 13 FCCTX families after the prioritization, validation and segregation studies. These variants are shown in Table 4.3 and further described below.

4.3.1. Predicted consequences of the candidate variants

Among the 44 candidate variants, there were a total of 10 frameshift, 7 nonsense, 2 inframe, 22 missense, 2 splice donor and 1 splice region variants. All the frameshift and nonsense variants produced a truncated protein and were therefore predicted to have a relevant effect on the protein structure and/or function. However, the extent of the damage caused by these truncating variants depends on the location of the mutation within the protein. Actually, among the truncating mutations prioritized, 2 were located at the end of the corresponding proteins and did not affect any protein domains. In the same way, although there were 3 splicing variants, only 2 directly affected donor sites. The remaining variant was located in the splice region and predicted by both HSF[216] and MaxEnt[217,218] algorithms to affect the splicing by altering a splicing donor.

Regarding the missense variants selected, although all the filtered mutations were predicted to be deleterious by *in silico* tools, only 14 out of 22 affected known protein domains according to the sources checked, while 1 additional variant affected the proximity of a domain (Table 4.3). However, it is worth noting that in some cases the structure of the corresponding protein was not completely understood yet. In addition, changes in the tertiary structure caused by truncating mutations, as well as the location of missense variants within the protein, were visualized when possible in order to see the overall effect or to check whether the affected amino acids were involved in any secondary structure.

4.3.2. Candidate genes

Regarding the genes affected by the candidate variants, there are some that are worth mentioning. As seen in Table 4.4, among all the candidate variants, a total of 12 were located in genes reportedly involved in DNA damage repair by different sources[219–223]. However, only 4 of them were deleterious: *POLQ* c.4262_4268del in family cc89, *NME7* c.1109_1110insACTT in family cc108, *MUS81* c.1586_1587insGAGGT in family cc122 and *PHACTR2* c.620dup in family cc440. The *NME7* variant only meant the loss of the last 6 amino acids of the protein, though. The vast majority of the remaining variants were either missense (6) or inframe (2), and only 4 of these were situated in previously-described protein domains, although 1 more affected the close vicinity of a domain. In addition, there were 3 other variants reported to be involved in the oxidative stress-dependent Nrf2 pathway: two frameshift (*TXNRD3* c.614dup and *SETD6* c.863_864insA) and one splice donor variant (*ABCC2* c.1209+2T>G).

On the other hand, there were a total of 10 candidate variants in tumor suppressor genes (TSGs) or candidate TSGs. These included not only well established tumor suppressors, but also putative TSGs published in the literature. Out of these, 1 was frameshift (*PTPRT* c.4099dup), 1 nonsense (*IQGAP2* c.1282C>T) and the remaining 8 were missense. Among the missense variants, only 3 were located within a protein domain, although 1 more affected the proximity of a domain.

Other candidate variants selected affected genes that were involved in different relevant signaling pathways. For instance, a total of 9 variants were situated in genes reported to be involved in the Wnt pathway (2 frameshift , 1 nonsense, 5 missense and 1 splice region with broken donor). In addition, 1 missense variant was involved in the BMP pathway, and 5 variants were involved in the TGF-β pathway, including 1 frameshift and 4 missense variants. It is worth noting that in some cases one single gene was involved in more than one of these pathways; that is the case of *SETD6*, which is further addressed in section 4.4.

Fam	Gene	Locus	Var Type	Variant (p.)	Pos/lenght	Domain affected	Pathways/fuction
cc7	MUC6	11p15.5	NS	p.(Cys703Ter)	703 / 2439	TIL, VWD and C8	Cytoprotection
cc7	TXNRD3	3q21.3	FS	p.(Glu151GlyfsTer6)	151 / 643	FAD_binding_2 and Pyr_redox	Nrf2 pathway
cc7	GAL3ST2	2q37.3	SP d	N/A	10 / 398	Gal-3-0_sulfotr	Metabolism
cc28	CHID1	11p15.5	NS	p.(Met1?)	1 / 418	Glyco_18	Metabolism, pathogen sensing
cc28	PYGO1	15q21.3	MS	p.(Ser362Pro)	362 / 419	PHD	Wnt pathway
cc28	MMP11	22q11.23	MS	p.(Pro78Ser)	78 / 488	Cys switch	Metalloprotease
cc28	AHRR	5p15.33	MS	p.(Cys227Ser)	2227 / 719	No	Cell growth, TSG
cc28	NKD2	5p15.3	MS	p.(Met144Lys)	144 / 451	N/A	Wnt path, TGF-β
cc81	ECH1	19q13.2	NS	p.(Arg148Ter)	148 / 328	ECH_2	Metabolism
cc81	MAPK15	8q24.3	MS	p.(Gly36Ser)	36 / 544	S_TKc	VEGF pathway
cc81	ATR	3q23	MS	p.(Ser927Gly)	927 / 2644	No	DNA damage, TSG
cc89	POLQ	3q13.33	FS	p.(Ile1421ArgfsTer8)	1421 / 2590	POLAc	DNA repair
cc89	C19orf40	19q13.11	MS	p.(Ser126Phe)	126 / 215	Rad10	DNA repair
cc89	ADIPOQ	3q27.3	MS	p.(Gly90Ser)	90 / 244	Collagen triple helix repeat	AMPK, NF-κB
cc89	NUP153	6p22.3	MS	p.(Arg17Gln)	17 / 1506	No	DNA repair, Wnt, TSG
cc108	NME7	1q24.2	FS	p.(Phe370LeufsTer8)	370 / 376	NDK	DNA repair
cc108	IQGAP2	5q13.3	NS	p.(Gln428Ter)	428 / 1575	IQ and RasGAP	Wnt path, NF-κB, TSG
cc108	RASSF5	1q32.1	MS	p.(Thr297Ile)	297 / 418	RA	Growth, apoptosis, TSG
cc108	NEK4	3p21.1	MS	p.(Ala191Val)	191 / 841	S_TKc	DNA repair
cc122	MUS81	11q13	FS	p.(Gln531Ter)	531 / 551	No	DNA repair
cc122	VEPH1	3q25.31-25.32	FS	p.(Ser449IlefsTer34)	449 / 833	PH	Wnt path, TGF-β
cc122	RNMT	18p11.21	MS	p.(Tyr313Cys)	313 / 476	Methyltransf_23	Translation
cc122	ABCB1	7q21.12	MS	p.(Val53Glu)	53 / 1280	ABC_mb	Transporter
cc122	CTNNAL1	9q31.2	IF	p.(Glu259del)	259 / 734	Vinculin	DNA repair, apoptosis
cc406	SHF	15q21.1	FS	p.(Asn72LysfsTer42)	72 / 423	SH2	Apoptosis, cell growth
cc406	ABCC2	10q24	SP d	N/A	403 / 1545	ABC_mb	Transporter, Nrf2 pathway
cc440	GALNT10	5q33.2	FS	p.(Tyr140ThrfsTer53)	140 / 603	Pfam Glycos_transf_2 and RICIN	Protein glycosylation
cc440	PHACTR2	6q24.2	FS	p.(Asn207LysfsTer7)	207 / 645	RPEL	DNA repair
cc440	CASP7	10q25	MS	p.(Cys256Ser)	256 / 388	CASc	BMP, apoptosis
cc440	HELQ	4q21.23	IF	p.(Lys588del)	588 / 1101	No	DNA repair
cc525	ECT2L	6q24.1	NS	p.(Gln743Ter)	743 / 904	RhoGEF	Pathways / GEF
cc525	EIF4G3	1p36.12	MS	p.(Cys342Trp)	342 / 1591	No	TGF-β
cc525	C9orf116	9q34.3	MS	p.(Val70Met)	70 / 136	DUF4490	DNA repair
cc565	ADAMTS14	10q22.1	NS	p.(Lys898Ter)	898 / 1226	TSP1	Metallopeptidase
cc565	ARMC3	10p12.23	MS	p.(Phe778Ser)	778 / 872	EDR1	Tumor initiation
cc565	RHOBTB1	10q21.2	MS	p.(Arg648Trp)	648 / 696	No	Rho signaling, TSG

Table 4.6. Candidate genes.

Fam	Gene	Locus	Var Type	Variant (p.)	Pos/lenght	Domain affected	Pathways/fuction
cc598	CCDC60	16q21	NS	p.(Arg520Ter)	520 / 550	N/A	Unknown
cc598	SETD6	12q24.23	FS	p.(Met288IlefsTer3)	288 / 473	Rubis	NF-κB, Wnt and Nrf2
cc598	ST18	8q11.23	MS	p.(Thr570Pro)	570 / 1047	MYT1	Transcription factor, TSG
cc763	WWOX	16q23.1-q23.2	MS	p.(Arg120Trp)	120 / 414	Near KR	DNA repair, Wnt, TSG, TGF-β
cc765	PTPRT	20q12-q13	FS	p.(Asp1367GlyfsTer24)	1345 / 1463	PTPc	STAT3 and PXN path, TSG
cc765	MAP3K6	1p36.11	MS	p.(Arg846Cys)	846 / 1288	S_TKc	MAPK and Wnt path, TSG, TGF-β
cc765	ABTB1	3q21	MS	p.(Ala405Pro)	405 / 478	No	PTEN pathway
cc765	INVS	9q31	SP r	N/A	1006 / 1065	No	Wnt pathway

Table 4.6. Candidate genes (cont.). Final selection of candidate variants per family. The different columns indicate the family, official gene name, locus, type of variant (NS: nonsense, FS: frameshift, MS: missense, IF: inframe, SP d: splicing donor, SP r: splicing region), variant name at protein level, position of the alteration relative to the length of the protein, protein domains affected and finally function or pathways in which the gene is involved (TSG: tumor suppressor gene).

4.3.3. Further characterization of candidate variants

Although a few candidate variants were addressed with some preliminary essays, only three of them were assessed with a more in-depth characterization. These include a frameshift variant in *SETD6*, which is involved in a high number of essential pathways, a frameshift variant in *PTPRT*, which is a tumor suppressor gene, and a missense variant in *PYGO1*, which is involved in the Wnt pathway. The characterization of these variant is further discussed in sections 4.3, 4.4 and 4.5.

4.4 Functional characterization of a frameshift variant in *SETD6*

4.4.1. *SETD6* c.791_792insA (M264Ifs*3) identified in family cc598

The whole-exome sequencing from one healthy and two affected members of FCCTX family cc598 (II:1, II:2 and II:3, Figure 4.1 A) allowed the identification of a frameshift variant in the *SETD6* gene. The corresponding family fulfilled the Amsterdam I criteria, with four CRCs affecting two successive generations and the earliest age at diagnosis being 34 years old. Appendix A2 shows in detail the clinical data available from the different members of this family. The variant detected was *SETD6* c.791_792insA, p.(Met264IlefsTer3) (rs768456822, ENST00000310682), simplified as M264Ifs*3. It is worth noting that this nomenclature was given based on SETD6's main functional transcript, while the list of candidate variants shows variant names based on the longest transcripts, which in this case would be c.863_864insA, p.(Met288IlefsTer3), (ENST00000219315). Despite this discrepancy, this variant will be hereon referred to as M264Ifs*3.

SETD6 M264Ifs*3 is truncating mutation that was therefore predicted to have a damaging effect on the protein. And in spite of not being a novel variant, it is quite rare in the general population, with a minor allele frequency of 0.0012 according to GnomAD. Moreover, it is not found in the 1000 Genomes Project nor observed in homozygous state in neither of these databases. In addition, this frameshift mutation affects the three main protein-coding transcripts of this gene (ENST00000310682, ENST00000219315 and ENST00000394266).

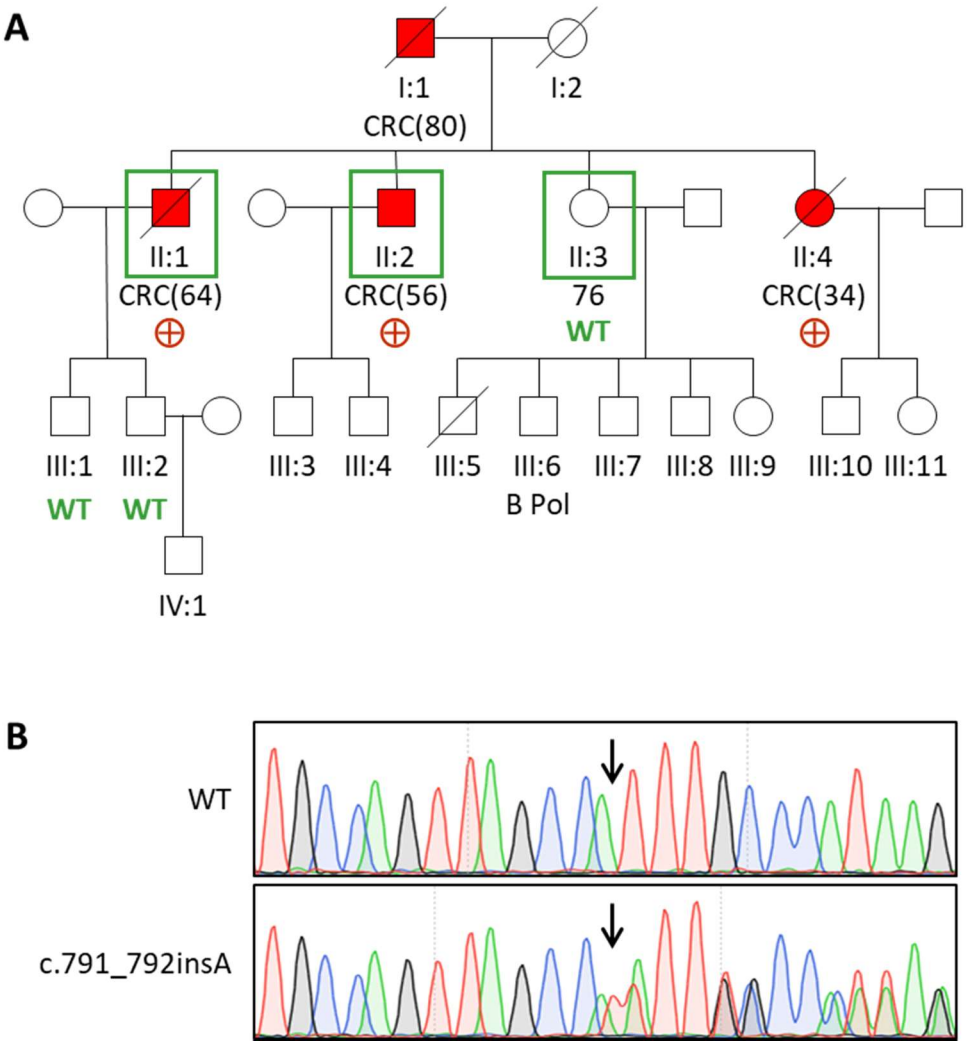


Figure 4.1. *SETD6* M264Ifs*3 cosegregates with colorectal cancer in FCCTX family cc598. A) Pedigree of Amsterdam I family cc598 carrying the *SETD6* deleterious mutation c.791_792insA (p.Met264IlefsTer3). Whole-exome sequencing was performed in family members II:1, II:2 and II:3 (green box), and segregation was studied in members II:4, III:1 and III:2. All affected participants were carriers (+), while the healthy were non-carriers (WT). Colorectal cancer (CRC)-affected members are marked in red. The age at diagnosis or current age of healthy members is included beneath each individual (in years). B) Electropherogram of the reverse wild-type and mutant sequence of the *SETD6* gene. The arrows show the point where the adenine is inserted.

4.4.2. Other candidate variants

Other candidate variants selected for this family included *CCDC60* c.1558C>T (p.Arg20*) and *ST18* c.1708A>C p.(Thr570Pro). Like the selected *SETD6* mutation, both of these variants were present in the two affected members (II:1 and II:2) and absent in the

healthy sibling (II:3). However, among them, the *SETD6* mutation resulted in the most relevant changes for the protein. Given the relevance of the different pathways in which *SETD6* is involved[163,167–169,171] and their implication in cancer development[242,243], *SETD6* c.791_792insA (p.Met264IlefsTer3) was selected among all the candidate variants of this family for additional characterization.

4.4.3. *SETD6* M264Ifs*3 results in the loss of the C-terminal half of the protein

The identified *SETD6* mutation involves the insertion of an adenine between positions 791 and 792 of the cDNA (ENST00000310682), producing a shift in the reading frame of the codons (Figure 4.1 B). This results in the introduction of two new amino acids at positions 264 and 265 — Ile264 and Gly265, instead of the original Met and Ala — followed by a premature stop codon (Figure 4.2 A). The resulting truncated SETD6 has an intact SET domain (responsible for the activity of the protein) but lacks the C-terminal half of the protein, which includes a linker sequence and the whole Rubisco domain, presumably responsible of mediating protein-protein interactions[244] (Figure 4.2 B). Figure 4.2 C shows the broad effect that this mutation has on the tridimensional structure of the protein.

4.4.4. *SETD6* M264Ifs*3 cosegregates with CRC but shows no LOH in the tumor

SETD6 M264Ifs*3 was carried by the two affected members sequenced (II:1 and II:2, with CRC diagnosed at ages 64 and 56, respectively), while not present in the healthy sibling (II:3, 76 years old) (Figure 4.1 A). Although there was only germline DNA of two young healthy relatives to study the segregation, the FFPE tumor block of a deceased CRC-affected member allowed the extension of this analysis. The segregation study confirmed that the mutation was also present in II:4 (CRC diagnosed at 34), while absent in the other

two healthy relatives (III:1 and III:2, 51 and 49 years old, respectively) (Figure 4.1 A and Appendix, Figure A2).

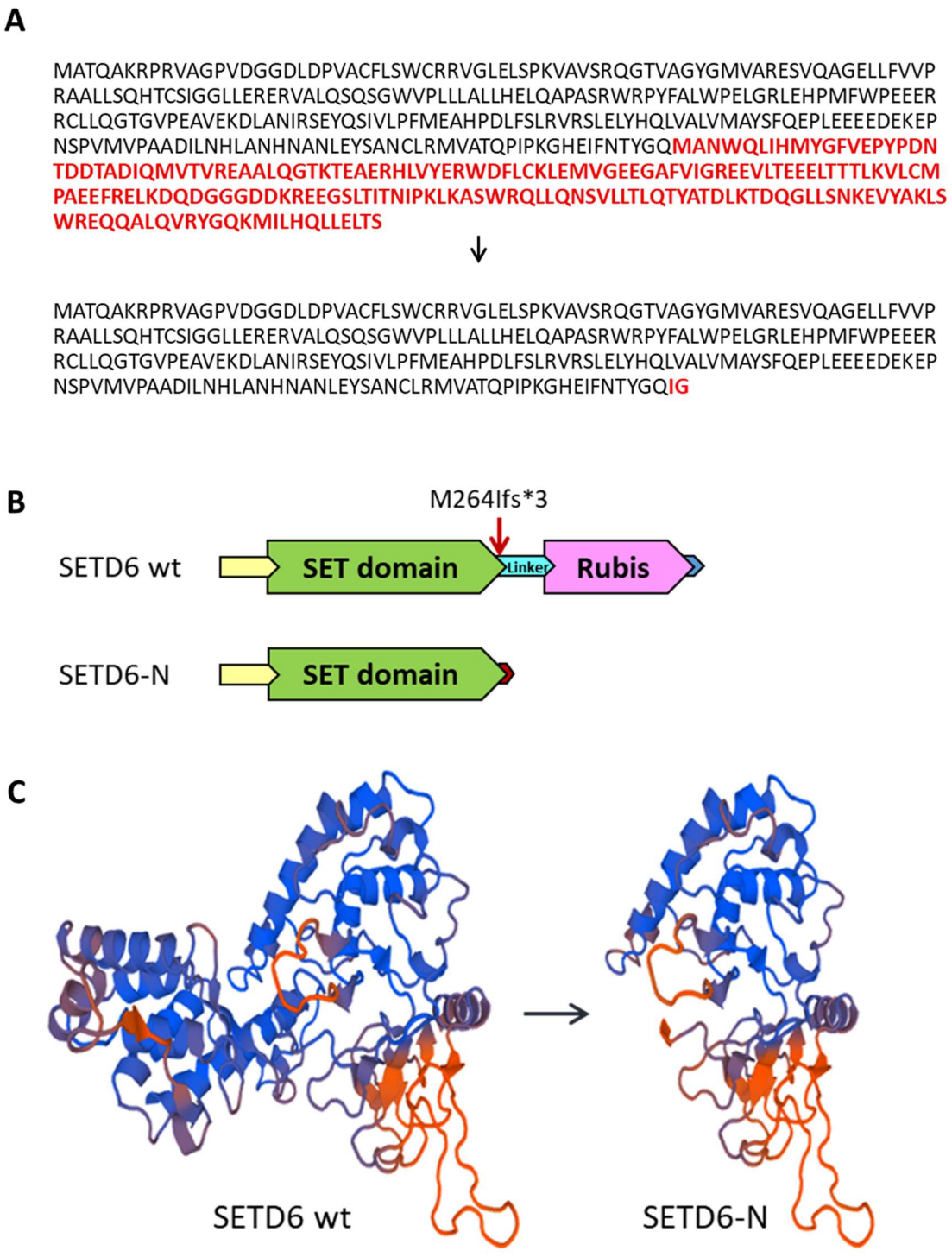


Figure 4.2. *SETD6* M264Ifs*3 causes the loss of the C-terminal half of the protein. A) Amino acid sequence of SETD6 wt and SETD6-N with the changes marked in red. B) Representation of the different domains within the SETD6 protein showing the loss produced by the truncation. C) Effect that the mutation is predicted to have on the 3D structure of SETD6 by SWISS-MODEL.

On the other hand, the LOH analysis revealed that there was no loss of any of the alleles in the tumor of II:2 (Figure 4.3). Although an apparent partial loss of the mutant allele was observed in the tumor of II:4, there was no germline DNA from the same patient to strengthen this observation. Nevertheless, it is noteworthy that member II:4 had a more advanced form of the disease than II:2 (stage IIIB versus I, Appendix, A2).

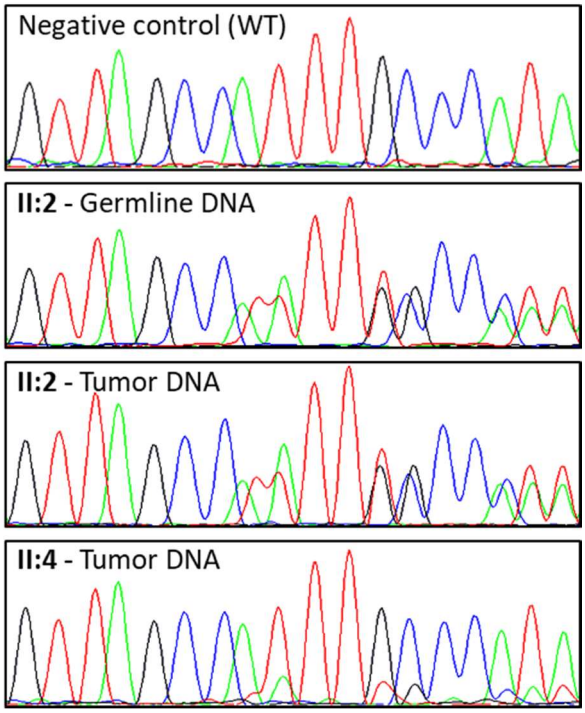


Figure 4.3. LOH study of *SETD6* M264Ifs*3. Electropherograms of the reverse sequences of *SETD6*'s exon 7 (ENST00000310682) from either germline or tumor DNA of two different members of family cc598, obtained by Sanger sequencing. The LOH was analyzed in tumor DNA from members II:2 and II:4, however, only II:2 had germline DNA for the correct comparison of the mutant and wild-type alleles. A negative wild-type control has been added as a reference.

4.4.5. Mutant SETD6 is stable and can be overexpressed in human cells

With the aim of checking whether mutant SETD6 was stable and normally expressed, a truncated version of the protein mimicking the frameshift mutation — hereon referred to as SETD6-N — was cloned and overexpressed in two different human cell lines. SETD6-N was found to be expressed to the same extent as wild-type SETD6 both in HEK293T and HCT116 cells (Figure 4.4 A and B, respectively). It is worth noting that an attempt to overexpress the C-terminal half of SETD6 that contained everything which SETD6-N is missing (known as SETD6-C) was unsuccessful.

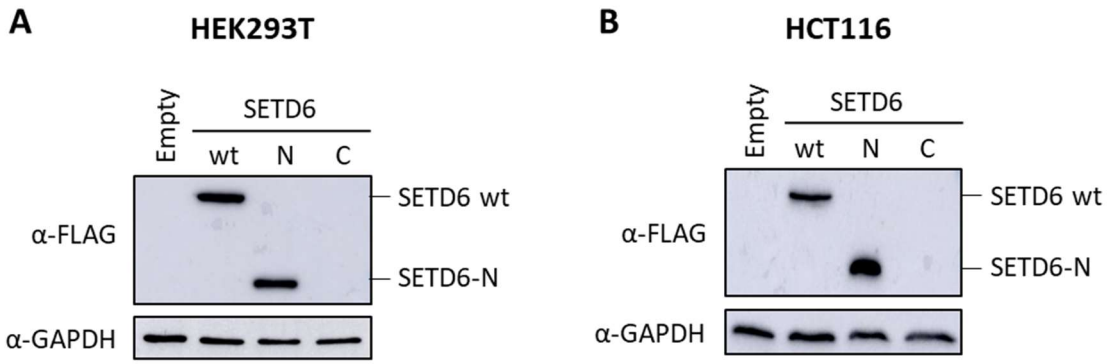


Figure 4.4. Mutant SETD6 is successfully overexpressed. Overexpression of FLAG-SETD6 (wt, N or C) in HEK293T (A) or HCT116 (B) cells. SETD6-N is the truncated version of the protein that mimics the mutation found in our family; SETD6-C is a truncated SETD6 that lacks the N-terminal part of the protein and the SET domain. The different versions of SETD6 were detected by western blot using an anti-FLAG antibody. GAPDH was used as a loading control.

4.4.6. SETD6-N shows the same localization pattern as wild-type SETD6

In order to test if the SETD6 mutation affects its subcellular localization, HEK293T and HCT116 cells were transfected with GFP-tagged SETD6 (either SETD6 wt or SETD6-N). Fluorescence microscopy of the transfected cells revealed that SETD6-N presented the same distribution within the cell as the wild type, both of which were mainly concentrated to the nucleus (Figure 4.5 A and B). We could detect, however, the presence of speckles inside the nucleus of the cells expressing mutant SETD6, which suggested that the association with the chromatin might be altered. To better understand the localization pattern, both proteins were transfected into cells followed by a biochemical fractionation of the cytosolic and chromatin fractions[237]. This experiment confirmed a similar subcellular localization between the mutant and the wild-type, with no significant differences in the cytosolic fraction (Figure 4.5 C and D). Some differences between the wild-type and SETD6-N were only observed at the chromatin level. To further verify these observations, we used an additional method to extract the chromatin (see the Materials and Methods section for more details). This experiment confirmed comparable amounts

of SETD6 wt and SETD6-N at this fraction (Figure 4.5 E and F), suggesting that SETD6-N shows the same chromatin localization pattern as the wild-type enzyme.

4.4.7. Recombinant SETD6-N binds its substrates to the same extent as wild-type SETD6

An ELISA was next performed to test whether the studied mutation would affect the binding ability of SETD6 in a cell-free *in vitro* system. To this end, we compared the direct binding of the wild type and mutant SETD6 to two well-known substrates, RelA and PAK4[163,167]. Recombinant RelA and PAK4 were used to coat an ELISA plate, together with BSA, which was used as a negative control. Recombinant SETD6 wt, SETD6-N and SUMO were used as tested proteins, being the latter another negative control. As shown in Figure 4.6, SETD6-N bound at equal levels to both substrates as did wild-type SETD6, which suggests that the mutation does not impair the interaction between SETD6 and its substrates *in vitro*.

4.4.8. Recombinant SETD6-N fails to methylate its substrates

To compare the enzymatic activity of wild-type SETD6 and SETD6-N, both proteins were subjected to a cell-free *in vitro* methylation assay with RelA and PAK4 as substrates (Figure 4.7 A and B, respectively). As expected, both substrates were specifically methylated by wild-type SETD6. However, no methylation was observed when SETD6-N was present in the reaction instead. It is worth noting that SETD6's automethylation activity was also lost in the truncated mutant. Consistent with these results, we found that while the recombinant wild-type SETD6 methylated immunoprecipitated RelA and PAK4 from HEK293T cells, SETD6-N failed to do so (Figure 4.7 C and D, respectively). In a reciprocal experiment, we observed that recombinant RelA and PAK4 were specifically methylated by wild-type SETD6 isolated from human cells and not by SETD6-N (Figure 4.7 E and F, respectively). These complementary assays further

demonstrate that the *SETD6* truncating mutation identified in this FCCTX family abrogates SETD6's enzymatic activity.

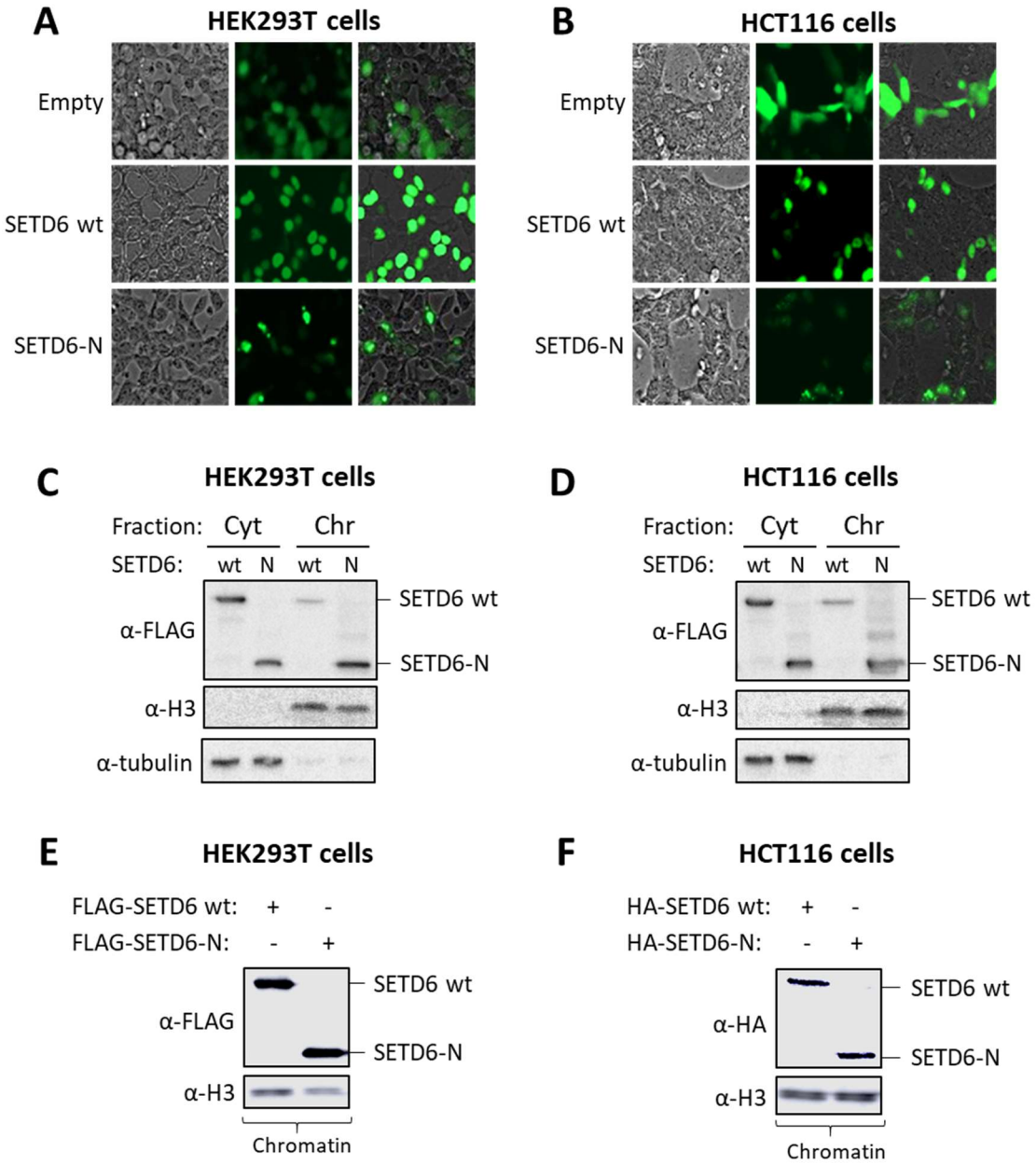


Figure 4.5. SETD6-N shows similar subcellular localization as wild-type SETD6. A+B) HEK293T (A) and HCT116 (B) cells were transfected with GFP-tagged empty, SETD6 wt or SETD6-N plasmids. The distribution of GFP within the cells was observed by fluorescence microscopy. C+D) HEK293T (C) or HCT116 (D) cells were transfected with FLAG-SETD6 (wt or N) and biochemically separated into cytoplasmic (Cyt) and chromatin (Chr) fractions. SETD6 was detected by western blot with an anti-FLAG antibody. Histone H3 and tubulin were used as chromatin and cytoplasmic controls, respectively. E+F) The chromatin fraction of HEK293T (E) or HCT116 (D) cells transfected with FLAG-SETD6 (wt or N) was isolated and SETD6 was detected by western blot using the indicated antibodies. Histone H3 was used as a chromatin control.

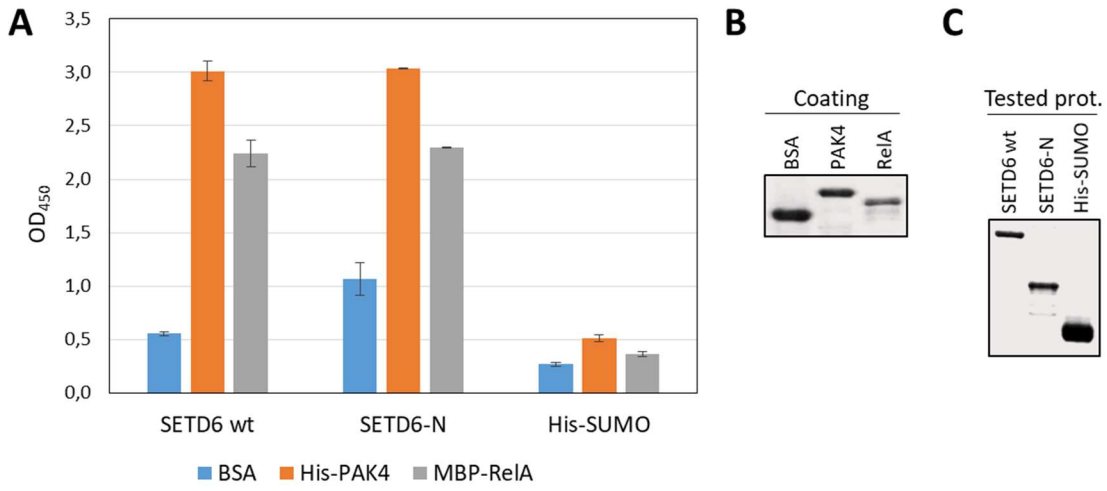


Figure 4.6. Recombinant SETD6-N shows similar binding to its substrates RelA and PAK4 as wild-type SETD6. A) Interaction between recombinant SETD6 (wt or N) and PAK4/RelA determined by an ELISA. The plate was coated with 2µg of His-PAK4, MBP-RelA or BSA (negative control), and then covered with 0.5µg of His-SETD6 wt, His-SETD6-N or His-SUMO (negative control). Bound proteins were detected using a rabbit anti-SETD6 antibody. Data and error bars are from two technical replicates and represent two independent biological experiments. B+D) Coomassie staining for the coating (B) and tested (C) proteins used in the ELISA.

4.4.9. SETD6-N binds its substrates but loses its activity in colon cancer cells

Since the SETD6 mutation was identified in hereditary colon cancer patients, we next aimed to confirm our findings in the colon cancer cell line HCT116. For this purpose, either wild-type SETD6 or SETD6-N were overexpressed in the presence or absence of FLAG-RelA (Figure 4.8 A). As expected, SETD6 wt and SETD6-N physically interacted with RelA at the chromatin to the same extent. Consistent with our cell-free *in vitro* experiments, we also observed that while SETD6 wt methylated RelA, SETD6-N did not. The methylation of RelA at the chromatin was identified using a RelAK310me1 antibody that could specifically recognize monomethylation of RelA at position K310[163]. The same results were obtained when PAK4 was used as the substrate (Figure 4.8 B). These findings support the dominant negative nature of SETD6 c.791_792insA (p.M264Ifs*3) in a colon cancer cellular model.

4.4.10. Both wild-type and mutant *SETD6* are expressed in the tumor and compete for their substrates

We next aimed to check if SETD6's mutant allele was expressed in the tumor, given that the variant was carried in heterozygosis. To this end, a digital PCR was carried out using two TaqMan probes that specifically recognized either the wild-type or the mutant cDNA (with the insertion of the adenine). As observed in Figure 4.9 A and B, the tumor from member II:2 showed a positive expression of both alleles. As expected, all non-carrier controls only expressed the wild-type allele.

The fact that both alleles were expressed raised the question of whether mutant and wild-type SETD6 would compete for their substrates. In order to address this issue, a cell-free *in vitro* methylation competition assay was performed in the presence of recombinant SETD6 wt and different amounts of recombinant SETD6-N, using RelA as the substrate. Figure 4.9 C shows how wild-type SETD6's activity was inhibited in the presence of SETD6-N in a concentration-dependent manner, with a significant reduction in the amount of methylated substrate even when the same amount of each form of the enzyme was present (lanes 2 and 3, respectively). Finally, a competition assay in HCT116 cells confirmed the cell-free *in vitro* results, showing that the methylation of RelA at K310 in cells was reduced upon SETD6-N overexpression in a dose-dependent manner (Figure 4.9 D). Taken together, these results suggest that the expression of the mutant allele in the carriers is expected to compete with the enzymatic activity of SETD6 wt, supporting a dominant negative role of this mutation.

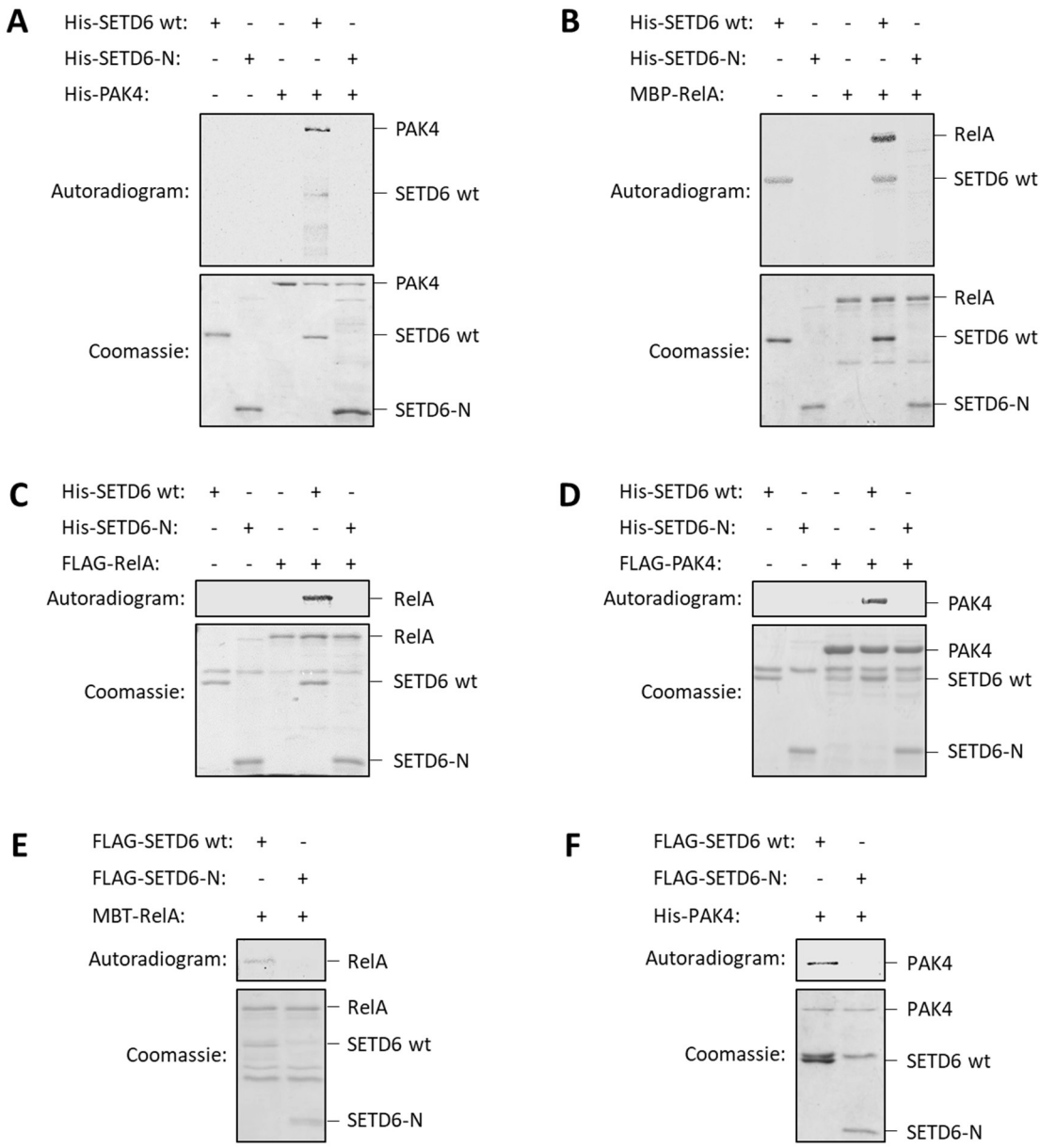


Figure 4.7. SETD6-N fails to methylate its substrates RelA and PAK4. A+B) Cell-free *in vitro* methylation assays in the presence of ^{3}H-SAM, recombinant His-SETD6 (wt or N) and either MBP-RelA (A) or His-PAK4 (B). The methylated proteins were detected by autoradiogram (top images), and the input used in the reactions is shown by Coomassie staining (bottom). C+D) HEK293T cells were transfected with FLAG-RelA (C), FLAG-PAK4 (D) or empty plasmid. Cell extracts were immunoprecipitated with FLAG M2 beads, followed by a radioactive cell-free *in vitro* methylation assay in the presence of ^{3}H-SAM and recombinant His-SETD6 (wt or N). The methylated proteins were detected by autoradiogram (top images), and the input used in the reactions was shown by Coomassie staining (bottom). E+F) HCT116 cells were transfected with either FLAG-SETD6 wt, FLAG-SETD6-N or empty plasmid. Cell extracts were immunoprecipitated with FLAG M2 beads, followed by a radioactive cell-free *in vitro* methylation assay in the presence of ^{3}H-SAM and recombinant MBP-RelA (E) or His-PAK4 (F). The methylated proteins were detected by autoradiogram (top images), and the input used in the reactions was shown by Coomassie staining (bottom).

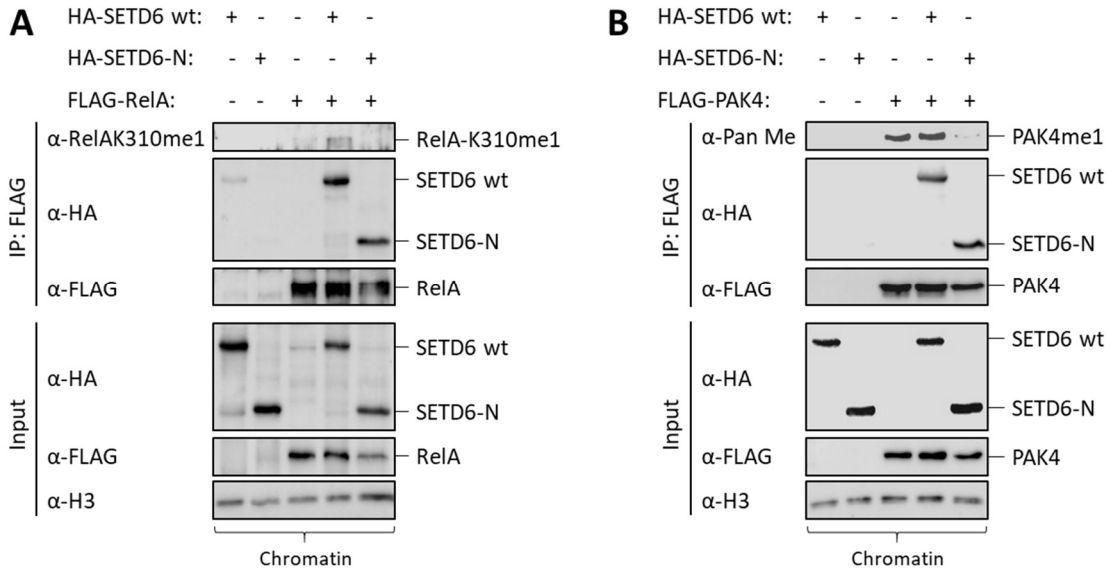


Figure 4.8. SETD6-N binds its substrates RelA and PAK4, while failing to methylate them, in the colon cancer cell line HCT116. A+B) HCT116 cells were transfected with either HA-SETD6 wt or HA-SETD6-N plasmids, in the absence or presence of FLAG-RelA (A) or FLAG-PAK4 (B). The chromatin fraction was then immunoprecipitated with FLAG M2 beads and analyzed by western blot with the indicated antibodies.

4.4.11. Downstream effects of SETD6-N

Last but not least, the downstream effects of the *SETD6* mutation were also addressed, studying the possible impact on the expression of target genes from pathways in which SETD6 is involved (Wnt, NF-κB and Nrf2), or on the viability of the cells. With that goal, a *SETD6* knockout colon cancer cell line was developed using the state-of-the-art CRISPR/Cas9 technology to modify HCT116 cells. However, this process was time-consuming and the knockout cell line was only ready by the very end of the stay in the host laboratory. For that reason, some preliminary assays studying the expression of target genes and the viability were performed meanwhile using normal HCT116 cells.

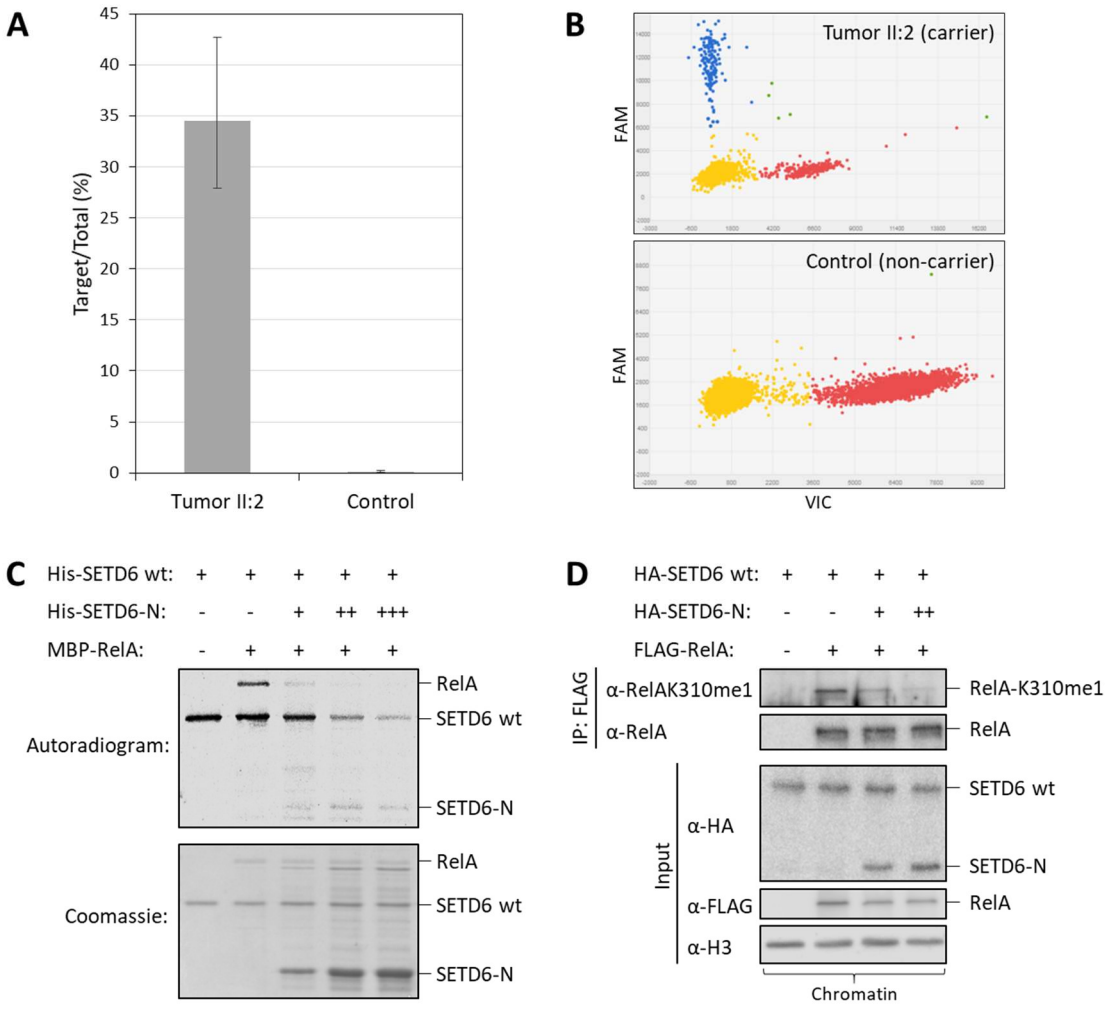


Figure 4.9. Both wild-type and mutant SETD6 are expressed in the tumor and compete for their substrates. (A) Allele-specific expression obtained by digital PCR presented as Target/Total, where "Target" is the mutant *SETD6* allele (detected with the FAM dye). Data from the tumor of II:2 was collected from three independent experiments, and the error bars correspond to the confidence intervals. The Control represents tumor cDNA from four different sporadic CRC patients used as non-carrier controls. (B) Digital PCR visualization of the cDNA from the tumor of member II:2 (upper figure) and a non-carrier control (lower figure). The FAM dye detects the mutant allele (blue), while the VIC dye detects the wild-type allele (red). In green are those wells that showed both dyes (VIC and FAM) and in yellow those that did not amplify. (C) Radioactive cell-free *in vitro* methylation assay in the presence of recombinant wild-type SETD6 and different amounts (+ and ++) of recombinant SETD6-N. The methylated proteins were detected by autoradiogram (top image), and the input used in the reactions is shown by Coomassie staining. (D) HCT116 cells were transfected with HA-SETD6 wt, FLAG-RelA and with or without different concentrations (+ and ++) of HA-SETD6-N. The chromatin fraction was then immunoprecipitated with FLAG M2 beads and analyzed by western blot with the indicated antibodies.

From the target gene expression experiments, some differences were observed between the *SETD6* wt and the *SETD6-N* transfected cells, although not always consistent. Among these, the levels of NF-κB target genes after TNF induction can be pointed out, with an overactivation of this pathway caused by the presence of mutant SETD6, as it had been hypothesized (Figure 4.10). While some differences were also observed for the Wnt and Nrf2 pathway between SETD6-N and wt SETD6, these observations were not consistent, and will have to be confirmed in the future (not shown). The viability of HCT116 cells, on the other hand, did not show any differences when SETD6-N was overexpressed as compared to SETD6 wt (not shown). However, these experiments should be ideally confirmed in the knockout cell line in the future.

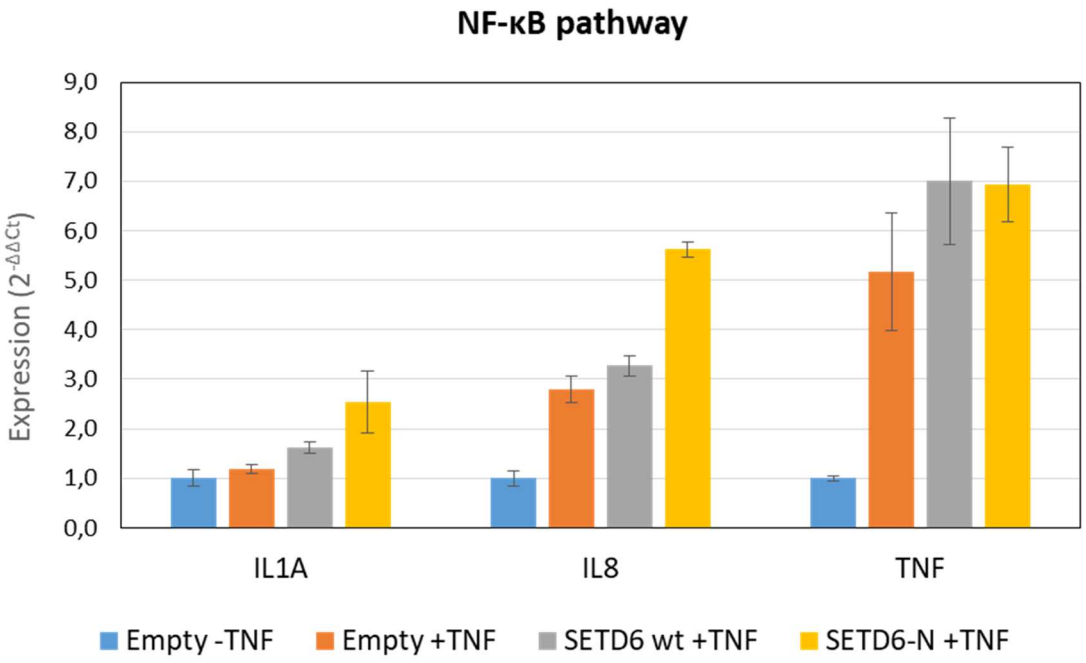


Figure 4.10. The expression of NF-κB downstream target genes is increased with the *SETD6* mutation. HCT116 cells were transfected with either *SETD6* wt, *SETD6-N* or empty plasmids and induced with 20ng/mL of TNF for 1h. RNA was then extracted from the cell lysates and converted to cDNA, and a qPCR was performed to measure the expression of three NF-κB target genes: IL1A, IL8 and TNF. Data was collected from three technical replicates and the expression was calculated as $2^{-\Delta\Delta Ct}$. The error bars represent the standard deviation. An increase in the expression is observed in SETD6-N transfected wells compared to SETD6 wt transfected cells for both IL1A and IL8. However, it is most significant for the latter.

4.5 Characterization of a frameshift variant in *PTPRT*

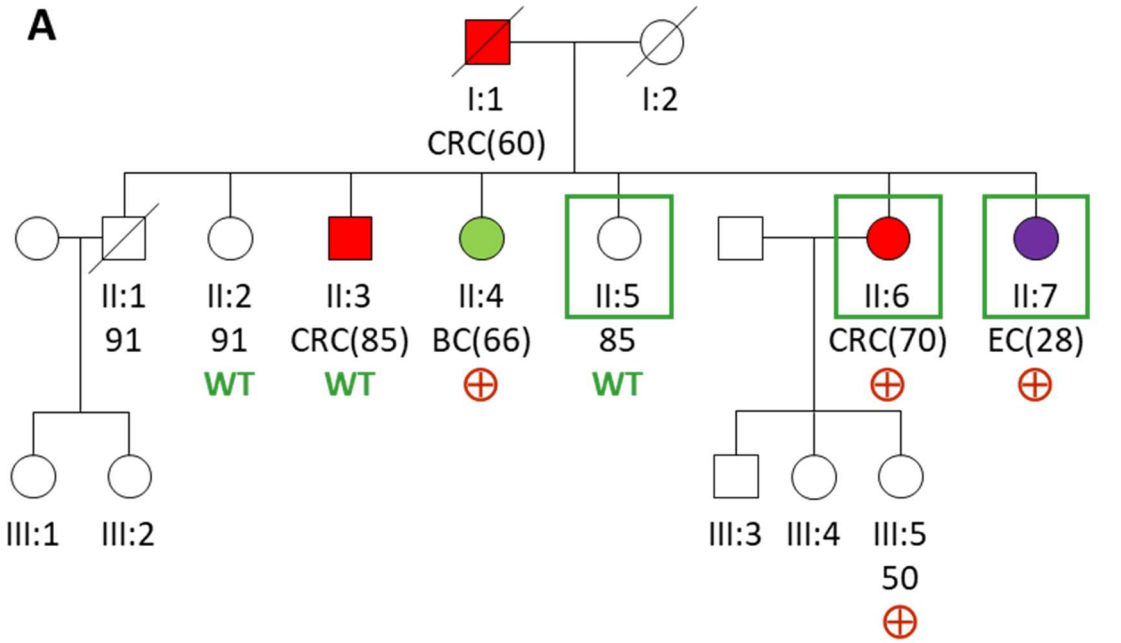
4.5.1. *PTPRT* c.4099dup (D1367Gfs*24) identified in family cc765

The whole-exome sequencing study also revealed a *PTPRT* frameshift mutation carried by family cc765. This family fulfilled the Amsterdam II criteria, with 5 members diagnosed of different cancers in two successive generations (3 CRCs, 1 endometrial and 1 breast cancer), the earliest age of onset being 28 years old (Figure 4.11 A). More information about the different members of this family is detailed in the Appendix (A2). The identified variant was *PTPRT* c.4099dup p.(Asp1367GlyfsTer24) (ENST00000373193), which is abbreviated as D1367Gfs*24. This variant was carried by the two affected members sequenced (II:6 and II:7), while not shared by the elderly healthy relative studied (II:5).

Regarding the frequency in the general population, *PTPRT* D1367G*24 is a novel variant that is not present in any of the public databases. Given the type of mutation, which causes the loss of the N-terminal end of the protein, it was predicted to be deleterious a priori with no need of additional *in silico* analyses. In addition, this variant affects all of *PTPRT*'s long protein-coding transcripts.

4.5.2. Other candidate variants

Other candidate variants in this family included *MAP3K6* c.2536C>T p.(Arg846Cys), *ABTB1* c.1213G>C p.(Ala405Pro) and *INVS* c.3017-5T>G. However, the *PTPRT* variant was undoubtedly the most relevant among all of them, considering both the gene and the effect of the mutation.



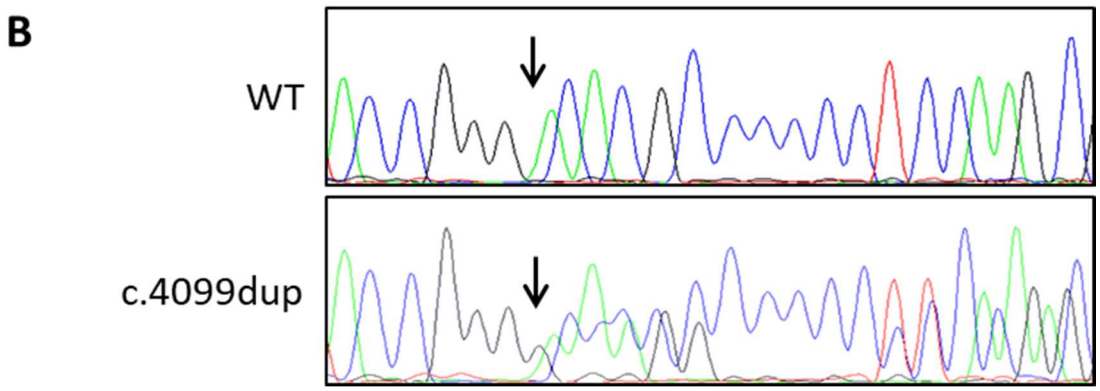


Figure 4.11. *PTPRT* D1367Gfs*24 shows a compatible segregation in family cc765. A) Pedigree of Amsterdam II family cc765, where *PTPRT* c.4099dup, p.(Asp1367GlyfsTer24), was identified. Whole-exome sequencing was done in family members II:5, II:6 and II:7 (green box), and segregation was studied in members II:2, II:3, II:4 and III:5. Three of the affected participants were carriers (+), while the healthy elderly relatives were non-carriers (WT). The only exception was member II:3, who did not carry the mutation and had been diagnosed of colorectal cancer (CRC) at the age of 85. Member III:5 did carry the mutation and was healthy at the moment, but was still young to draw any conclusions. CRC-affected members are marked in red, HNPCC-associated cancers such as endometrial cancer (EC) are marked in purple, and other non-related cancers such as breast cancer (BC) are marked in green. The age at diagnosis or current age of healthy members is included beneath each individual (in years). B) Electropherogram of the wild-type and mutant sequence of the *PTPRT* gene. The arrows show the point where the guanine is inserted.

4.5.3. *PTPRT* D1367Gfs*24 affects the second catalytic domain of the protein

The selected *PTPRT* mutation involves the insertion of a guanine between positions 4091 and 4092 of the cDNA (ENST00000373198) (Figure 4.11 B), causing a shift in the reading frame that is translated as the addition of 24 erroneous amino acids followed by a premature stop codon. Despite the frameshift nature of this mutation, PTPRT D1367Gfs*24 is located at the end of the gene, meaning the loss of just the last 95 amino acids of the protein. However, this mutation affects the second phosphatase domain of PTPRT (also known as D2), which is responsible for the regulation of the enzyme's activity. Figure 4.12 A shows the different protein domains of PTPRT, together with the location of the variant and a schematic visualization of the resulting mutant protein. Figure 4.12 B represents the effect that the truncation is predicted to have on the 3D structure of the cytoplasmatic region of PTPRT (including the D1 and D2 domains), showing a considerable gap caused by the mutation.

4.5.4. *PTPRT* D1367Gfs*24 segregation and LOH

Like all the other candidate genes of this family, PTPRT c.4099dup was carried by the two affected members whose whole-exome was sequenced (II:6 and II:7, with colorectal and endometrial cancer diagnosed at the age of 70 and 28, respectively), while absent in the 85-year-old healthy sibling sequenced. In addition, the segregation of this variant could be studied by Sanger sequencing in 3 other siblings: one with CRC diagnosed at 85, one with breast cancer diagnosed at 66, and one healthy at the age of 91. Although less informative, the 50-year-old healthy daughter of member II:6 was also analyzed (III:5).

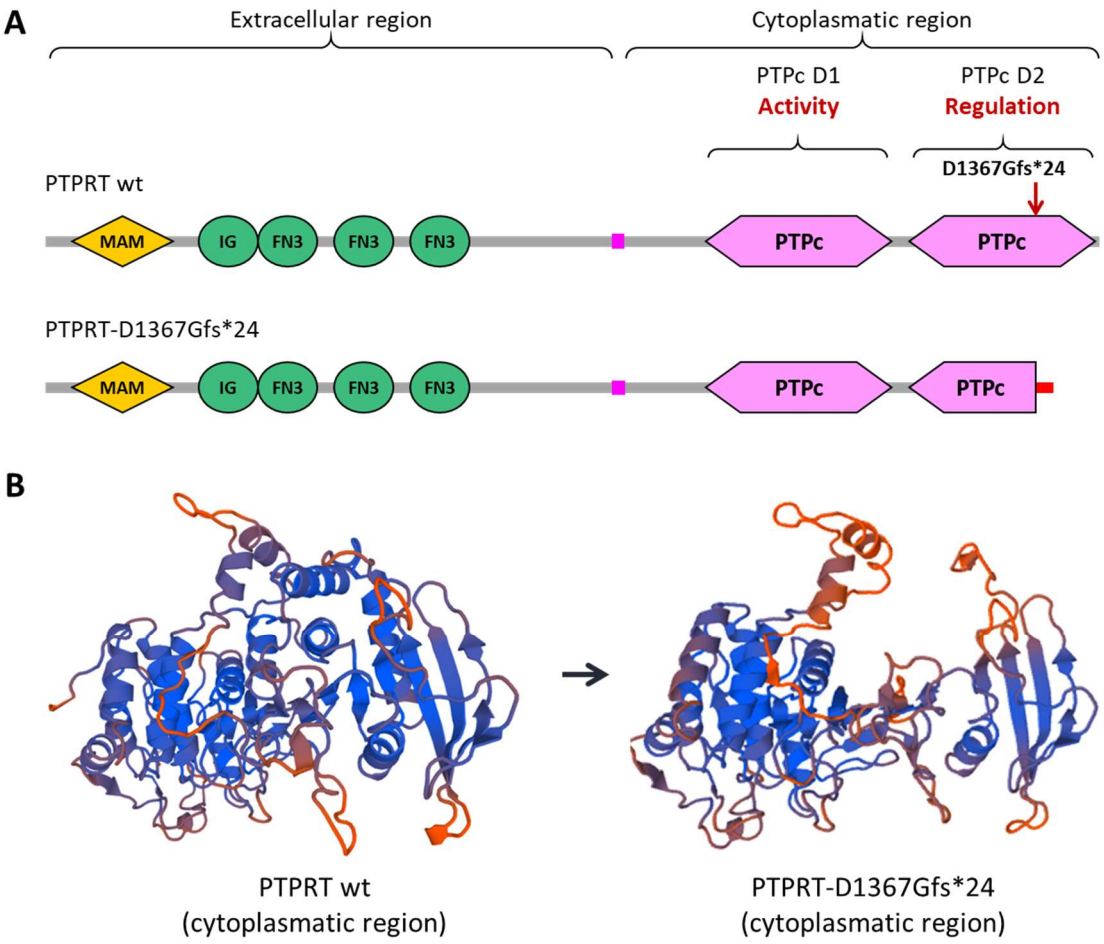


Figure 4.12. Effect of *PTPRT* D1367Gfs*24 on the protein structure. A) Schematic representation of PTPRT protein domains, with the location of the studied mutation (red arrow) and the resulting mutant protein (below). B) 3D representation of the cytoplasmatic region of wild-type (left) and mutant (right) PTPRT, showing the effect of the truncation. 3D models were obtained by SWISS-MODEL.

The result from the segregation study is shown in Figure 4.11 A. Apart from the two affected probands originally studied, only members II:4 and III:5 carried the mutation, while II:2 and II:4 showed a wild-type *PTPRT*. The positive members include the breast cancer patient (II:4) and a healthy but still young relative (III:5), while the negative result was observed in the healthy elderly relative (II:2) and the CRC patient diagnosed at a very late age (II:3), which could be a phenocopy. As far as the tumor LOH is concerned, two FFPE tumor blocks could be studied: the CRC of member II:6 and the breast tumor of member II:4. No significant loss of any of the alleles was observed in the CRC, although a slight reduction of the mutant allele seemed to happen in the breast tumor (Figure 4.13).

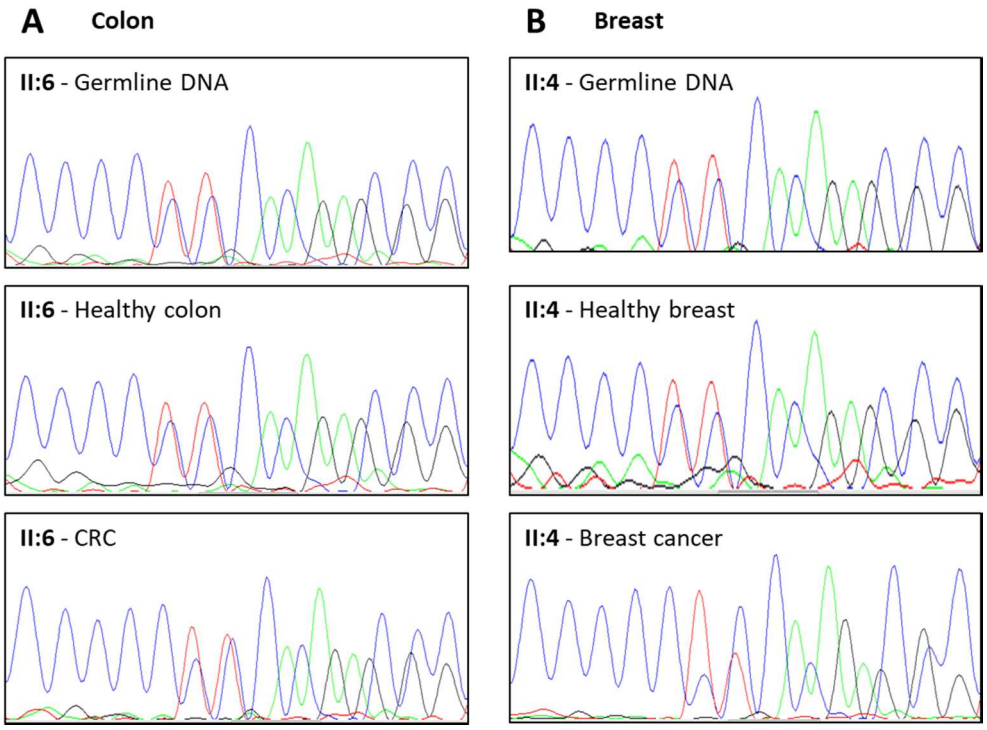


Figure 4.13. LOH study for *PTPRT* D1367Gfs*24. A) Electropherogram obtained by Sanger sequencing of DNA from the CRC-diagnosed member (II:6): germline, healthy colon and CRC**.** B) Electropherogram obtained by Sanger sequencing of DNA from the breast cancer-diagnosed member (II:4): germline, healthy breast and breast tumor**.**

4.5.5. The expression of the wild-type *PTPRT* allele is decreased in the tumor from a carrier

A digital PCR was performed in order to check the allele-specific expression of wild-type and mutant *PTPRT* in the tumor. This assay confirmed that, although rather low, there was expression of both alleles (wild type and mutant) in tumor cDNA from a *PTPRT* mutation carrier (II:6). However, a significant reduction in the expression of the wild-type allele was observed in the CRC compared to healthy colon from the same individual (Figure 4.14). No decrease of the expression of the mutant allele could be appreciated. Unfortunately, no results could be obtained from the breast tumor of member II:4.

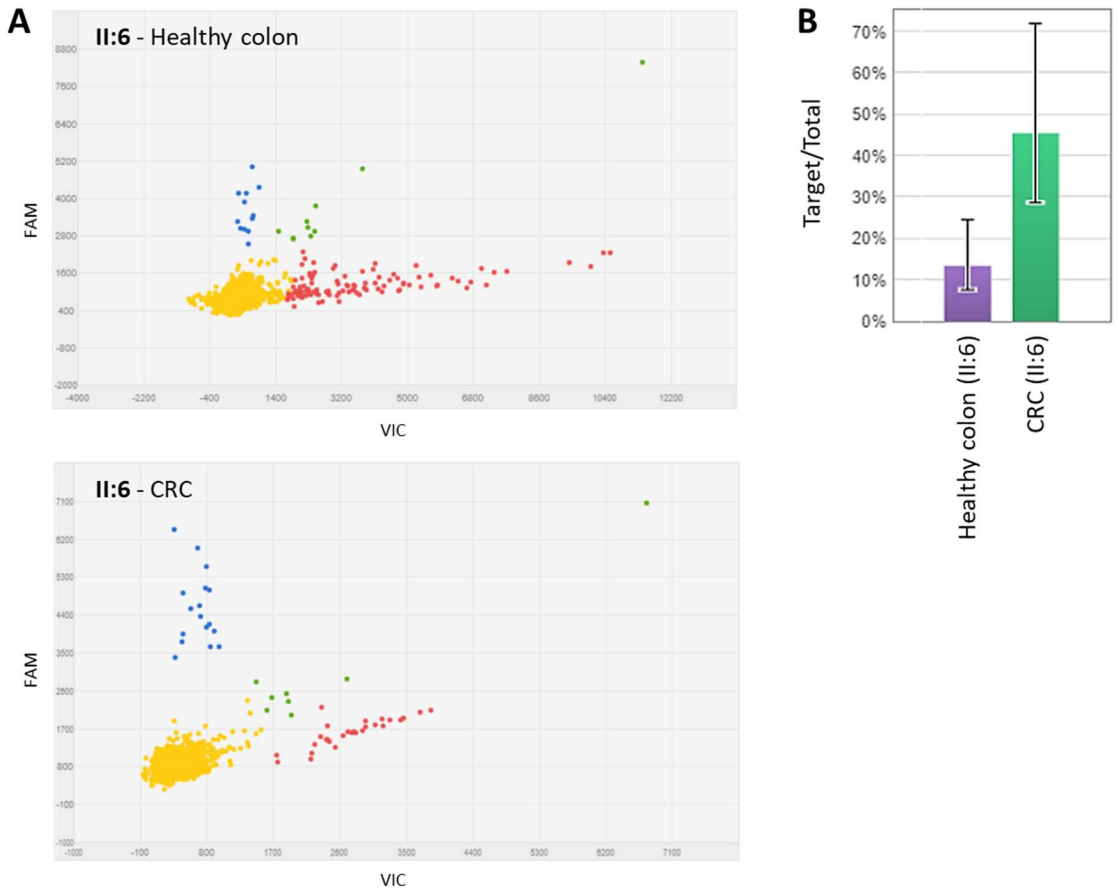


Figure 4.14. The expression of the wild-type *PTPRT* allele is decreased in the tumor from a carrier. A) Digital PCR visualization of the allele-specific expression assay performed in the healthy colon tissue (top) and CRC (bottom) of cc765 family member II:6. The FAM dye detects the mutant allele (blue), while the VIC dye detects the wild-type allele (red). In green are those wells that showed both dyes (VIC and FAM) and in yellow those that did not amplify. B) Quantification of the allele-specific expression obtained by digital PCR presented as Target/Total, where "Target" is the mutant *PTPRT* allele. The relative expression of the mutant allele is increased in the tumor given that the expression of wt PTPRT is decreased. Data was collected from two independent experiments, and the error bars correspond to the confidence intervals.

4.5.6. *PTPRT* promoter is hypermethylated in the tumors from two carriers

The promoter region of *PTPRT* has been reported to be frequently hypermethylated in sporadic CRC[177], and given that LOH of the wild-type allele was not detected, this could be a plausible second hit for the inactivation of the wild-type *PTPRT*

allele in the tumors from the carriers. For this reason, the methylation status of *PTPRT*'s promoter was studied in the tumors available by methylation-specific PCR as previously reported[177,178]. To this aim, two different methylation sites were tested: one published by Laczmanska et al[177] in CRC and another one described by Peyser et al[178] in head and neck cancer.

This assay showed that the CRC of one of the mutation carriers (II:6) was hypermethylated at the two *PTPRT* promoter sites tested, as compared to healthy colon tissue from the same carrier, as well as to a pool of 5 healthy colon samples used as a control (Figure 4.15 A, top). Although to a lesser extent, the same result was observed when the breast tumor and healthy breast from member II:4 were studied at the promoter site described by Peyser et al[178] (Figure 4.15 A, bottom). However, no methylation was observed at all in any of the breast samples at the methylation site reported by Laczmanska et al[177]. These observation are quantified in Figure 4.15 B.

In addition, a pyrosequencing assay targeting a region of *PTPRT*'s promoter confirmed the hypermethylaton observed by MSP in the CRC of member II:6 compared to its healthy colon tissue (Figure 4.15 C). Although once again more subtle, the breast cancer also showed an increased methylation compared to the healthy breast.

4.5.7. Expression of *PTPRT*'s downstream target genes

The lack of a functional PTPRT would be predicted to have consequences on the phosphorylation status of its target proteins STAT3 and paxillin[182,183,245]. This would result, in turn, in an altered expression of STAT3's target genes (such as *BCL2-XL* and *SOCS3*)[182], as well as an abnormal phosphorylation of paxillin's target proteins (such as Akt, p130Cas and Shp2)[183,245]. In order to check if these downstream effects were observed in the tumors from the *PTPRT* mutation carriers, a Taqman real-time qPCR was performed to evaluate the expression of *BCL2-XL* and *SOCS3*.

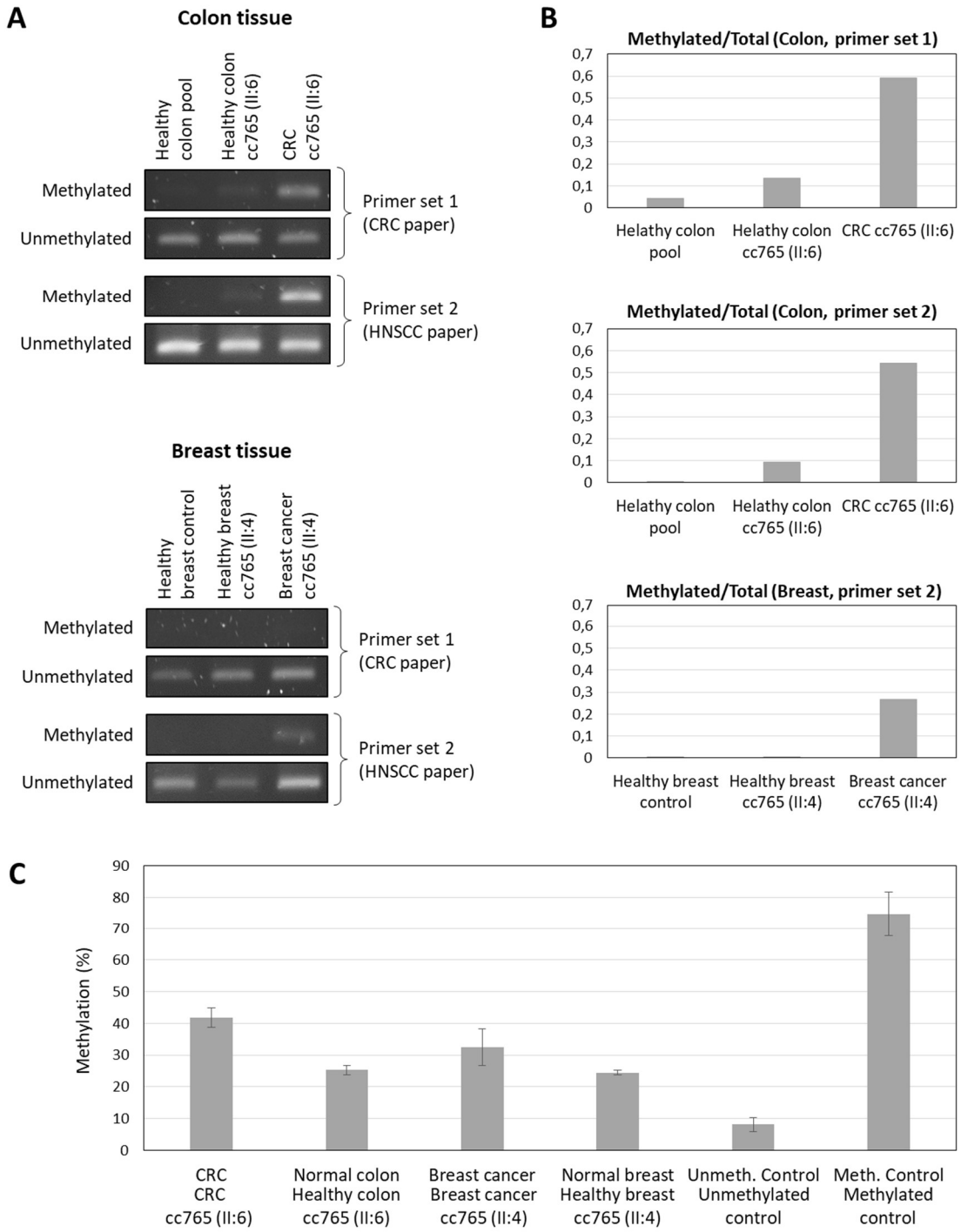


Figure 4.15. *PTPRT*'s promoter is hypermethylated in the colon and breast cancers of two mutation carriers. A) *PTPRT*'s promoter methylation status obtained by MSP of two different promoter sites previously described in CRC (primer set 1) or head and neck carcinoma (primer set 2), in the CRC vs healthy colon of member II:6 (top) or the breast cancer vs healthy breast of member II:4 (bottom). Both tumors show an increase in methylation when compared to their respective healthy tissue and controls. B) Graphical representation of the MSP results presented in A, portrayed as the ratio of methylated product versus total (methylated + unmethylated). Bands were quantified using the ImageJ software. C) *PTPRT*'s promoter methylation status (%) obtained by pyrosequencing for the same samples analyzed in A, together with a methylated and unmethylated commercial controls. The CRC of member II:6 is hypermethylated compared to the corresponding healthy tissue, while a more subtle hypermethylation is observed for the breast cancer of member II:4.

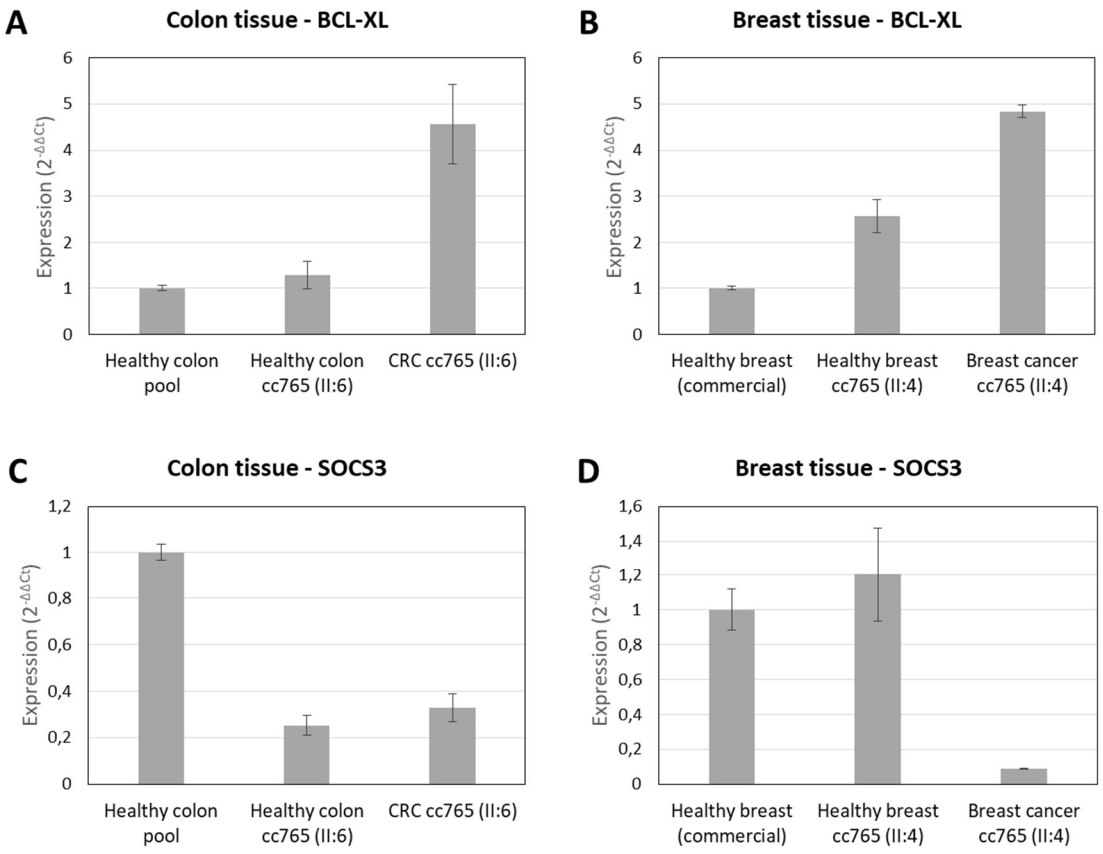


Figure 4.16. Expression of STAT3's target genes *BCL-XL* and *SOCS3* in the tumor or healthy tissue of two *PTPRT* mutation carriers. qPCR data was collected from three technical replicates and represent 2 independent experiments. The expression was calculated as $2^{-\Delta\Delta Ct}$, and the error bars represent the standard deviation. A) The colorectal tumor of a *PTPRT* mutation carrier (II:6) shows a significant increase in the expression of STAT3's target *BCL-XL* when compared to both its corresponding healthy colon and a pool of healthy colon tissue used as a control. B) The same increase in *BCL-XL* levels is shown in the breast tumor of another *PTPRT* mutation carrier (II:4). C) The CRC and healthy colon of member II:6 show a decreased *SOCS3* expression when compared to the healthy colon pool. C) Decreased *SOCS3* is also observed in the breast cancer (but not healthy tissue) of member II:4.

As shown in Figure 4.16 A, *BCL2-XL* was significantly overexpressed in the colon tumor from member II:6 when compared to both the corresponding healthy colon tissue from the same patient and a healthy colon cDNA pool comprised of 4 different normal tissue samples. In the same way, the breast tumor from member II:4 showed a consistently higher *BCL2-X* expression than the commercial healthy breast sample that was used as a control (Figure 4.16 B). The healthy breast tissue from the same patient presented higher expression levels than the healthy colon pool, although lower than the

corresponding tumor. It is worth noting, though, that this patient had an invasive (infiltrating) carcinoma of the breast, reason for which the whole breast had been removed.

On the other hand, *SOCS3* seemed to be considerably underexpressed in both tumors (CRC from II:6 and breast cancer from II:4). Interestingly, this lower expression was also observed in the healthy colon from II:6, while not in the healthy breast tissue from member II:4 (Figure 4.16 C and D).

4.6 Characterization of a missense variant in *PYGO1*

4.6.1. *PYGO1* c.1084T>C (S362P) identified in family cc28

Among all the candidate variants obtained by the whole-exome sequencing study, a missense mutation in *PYGO1* was selected in family cc28. This was an Amsterdam I family with 4 CRCs in two consecutive generations, the earliest diagnosed at the age of 41 (Figure 4.17 A). In addition, there was one additional breast cancer in the family, diagnosed at the age of 49 in the daughter of one of the CRC-affected members. Two affected members of this family were sequenced (II:1 and III:1), both of which carried the mutation.

This prioritized variant was *PYGO1* c.1084T>C, p.(Ser362Pro) (ENST00000302000) – simplified as S362P – which is a novel variant that is therefore not described in any population in the public databases. In addition, this variant is predicted to have a deleterious effect on the protein by the 5 *in silico* tools used and affects all the protein-coding transcripts of *PYGO1* (ENST00000302000, ENST00000563719 and ENST00000645724).

4.6.2. Location of S362P within the Pygo1 protein

As mentioned before, the studied *PYGO1* variant involved the change of a serine for a proline at position 362 of the encoded protein. Although just a missense variant, the five *in silico* tools used predicted a probable damage for the protein (Table 4.3), and it meant the substitution of a polar residue (serine) for a proline, which is a hydrophobic amino acid with very particular properties and structural limitations[241]. In addition, the affected serine was located in one of the Zn fingers of Pygo1's PHD domain (Figure 4.18 A). Pygo1's 3D structure, visualized in Figure 4.18 B, shows how serine 362 (in red) in

situated in close proximity to one of the Zn atoms (in grey). In fact, according to the model published by Fiedler et al in 2008[194], the affected serine would be situated right next to one of the residues (Cys363) that interacts directly with a Zn atom (Figure 4.18 C, red arrow). Although Ser362 is reportedly only semiconserved and is not part of any specific secondary structure (Figure 4.18 D), the different size and nature of the two amino acids (polar/hydrophobic) could be expected to affect the structure of the corresponding Zn finger and, consequently, Pygo's function.

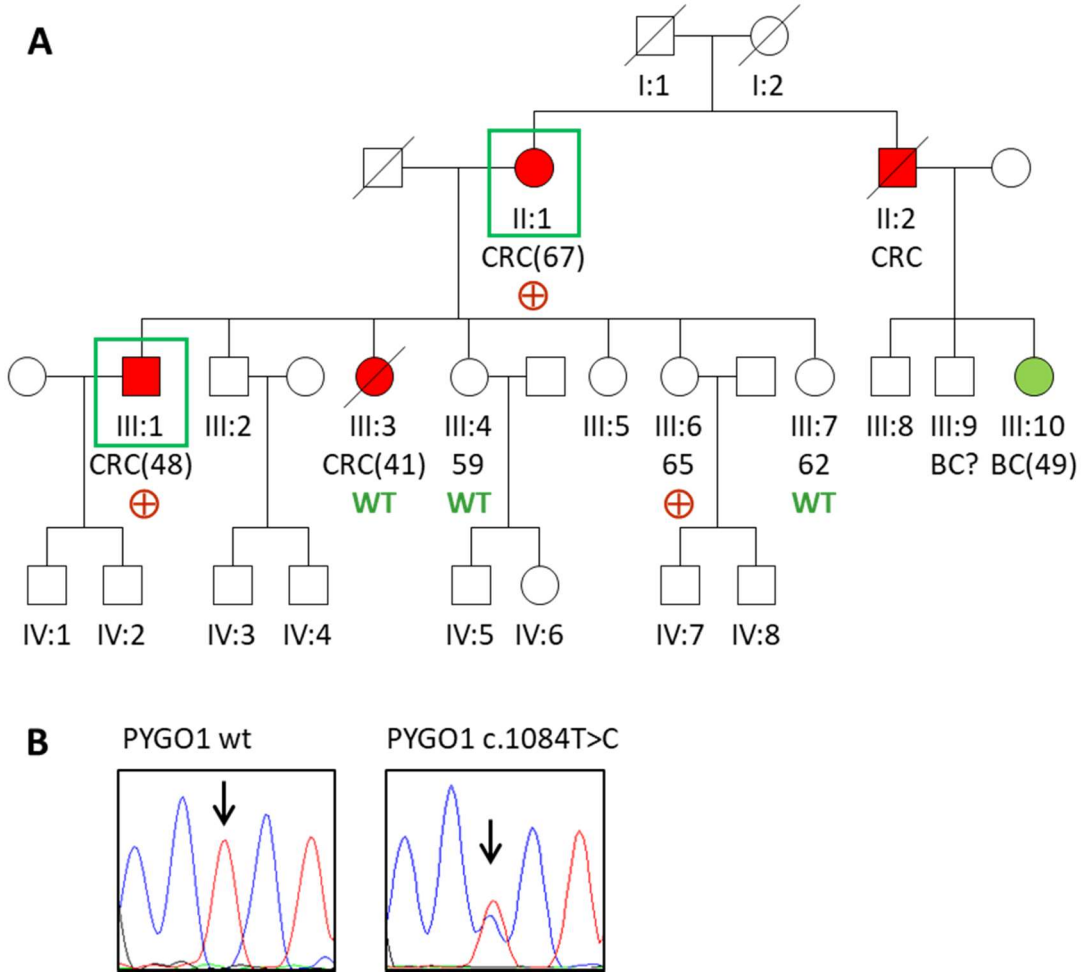


Figure 4.17. *PYGO1* S362P segregation in family cc28. A) Pedigree of Amsterdam I family cc28, where *PYGO1* c.1084T>C, p.(Ser362Pro), was detected. Whole-exome sequencing was done in family members II:1 and III:1 (green box), and segregation was studied in germline DNA from members III:4, III:6, III:7 and in tumor DNA from member III:3. Only two of the affected participants were carriers (+), while the two out of the 3 healthy members studied were non-carriers (WT). Unfortunately, the mutation was not observed in tumor DNA from member III:3, although there was no germline DNA to confirm this observation. CRC-affected members are marked in red, and the age at diagnosis or current age of healthy members is included beneath each individual (in years). B) Electropherogram of the wild-type and mutant sequence of the *PYGO1* gene. The arrows show the point where the missense mutation occurs.

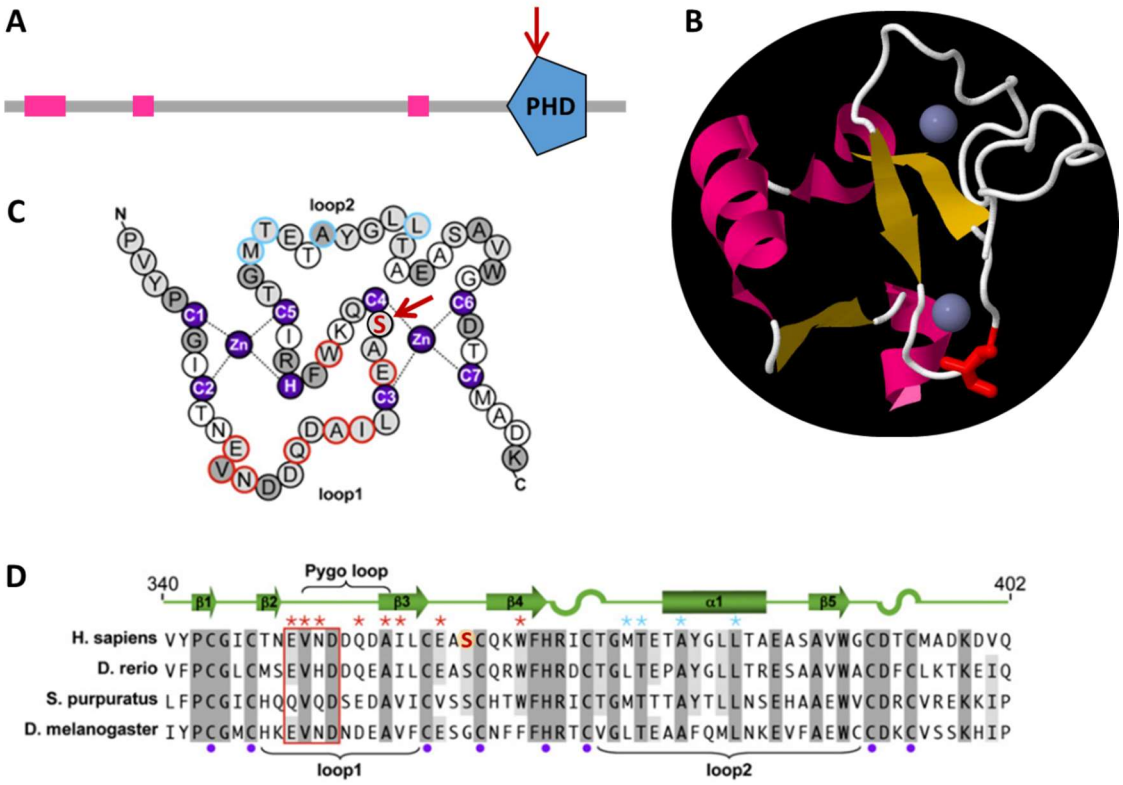


Figure 4.18. Consequences of *PYGO1* S362P on the protein. A) Representation of Pygo1's protein domains, with the location of the affected residue (red arrow). B) 3D visualization of the tertiary and secondary structure of Pygo1 with the substituted serine highlighted in red. The α-helixes are marked in pink, while β-sheets are marked in yellow. The gray spheres represent the zinc atoms. C) Zn finger of Pygo1's PHD domain obtained from Fiedler et al 2008. The red arrow points to the serine residue affected by the studied mutation. In purple are the two Zn atoms together with the 8 amino acids with which they interact directly (7 Cys and 1 His). Serine 362 is situated right next to the 4th Zn-bonded cysteine. D) Conservation of the Zn finger residues (Ser362 is marked in red with a yellow circle).

4.6.3. Other candidate variants

Although there were no truncating variants selected for this family, there were other missense candidate variants, including *CHID1* c.3G>A p.(Met1?), *MMP11* c.232C>T p.(Pro78Ser), *AHRR* c.680G>C p.(Cys227Ser), and *NKD2* c.431T>A p.(Met144Lys). The relevant location of the *PYGO1* variant within the protein, together with Pygo1's role in

the Wnt pathway, was the reason why this variant was selected for further characterization.

4.6.4. *PYGO1* S362P segregation and LOH studies

Apart from the two members who had been studied through NGS (II:1 and III:1, with CRC diagnosed at the age of 67 and 48, respectively), the segregation of this mutation could be initially studied in only 3 additional members of the family: III:4, III:6 and III:7, all of whom where currently healthy at ages of 59, 65 and 62 years old, respectively. The segregation study revealed that out of the 3 healthy relatives, only III:6 carried the *PYGO1* mutation (Figure 4.17 A). However, a later analysis of the tumor of member III:3 (diagnosed of CRC at 41), showed that this member did not carry the mutation (Figure 4.19). It is important to note that this discovery only happened after the variant had already been selected for further characterization, given that this tumor block was originally not available and that obtaining it meant a great effort. Unfortunately, this member had deceased so it was not possible to obtain germline DNA to confirm this result.

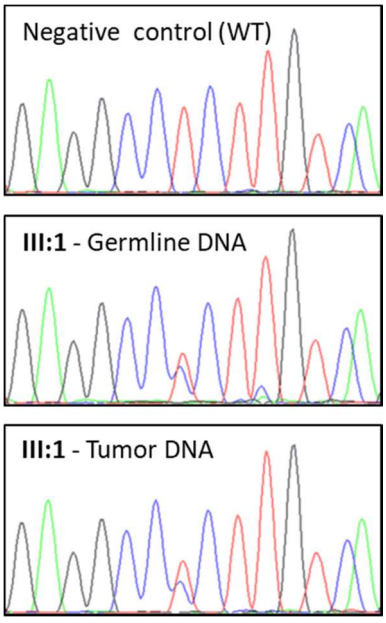


Figure 4.19. LOH analysis of *PYGO1* S362P. Electropherograms of the forward sequences of *PYGO1*'s exon 3 (ENST00000302000) from tumor and germline DNA of member III:1 from family cc28, obtained by Sanger sequencing. There was no significant decrease of any of the alleles. A negative wild-type control has been added as a reference.

On the other hand, the loss of heterozygosity was studied in the tumor of member III:1, which was the only one available from the carriers. As can be seen in Figure 4.19, this analysis showed that there was no apparent loss of any of the alleles in the tumor compared to germline DNA.

4.6.5. Mutant Pygo1 shows a decreased histone-binding affinity compared to the wild type

The relevant location of this missense variant at one of the protein's Zn fingers (PHD domain) prompted us to check if the serine to proline substitution affected Pygo1's histone-binding affinity. In order to do this, the mutant PHD domain was expressed in association with the HD1 domain from either Bcl9 or B9L, and the affinity of the complexes to bind a histone H3K4me2 peptide was tested by Isothermal ITC[196]. The two corresponding wild-type PDH-HD1 complexes were also obtained and tested as a reference.

It is worth noting that although all four protein complexes were expressed simultaneously, those that included mutant PDH gave lower yields of both overall and folded protein (~13.7mg for wt PHD_{Pygo1}-$HD1_{BCL9}$; ~14.7mg for wt PHD_{Pygo1}-$HD1_{B9L}$; ~4.6mg for S362P PHD_{Pygo1}-$HD1_{Bcl9}$ and ~4.7mg for S362P PHD_{Pygo1}-$HD1_{B9L}$). In addition, the proportion of protein that eluted off the gel filtration in the void (which is indicative of misfolded protein) was higher for the S362P mutation, although the overall amount of protein that eluted in the void was about the same for wt and mutant.

The ITC measurements were carried out on freshly purified proteins for optimum results. This experiment showed that the S362P mutation led to a 5.5-fold decrease in affinity for the histone H3K4me2 peptide when Pygo1's PHD formed a complex with Bcl9's HD1 (Figure 4.20 A and B). A 4-fold reduction in affinity was observed when mutant PHD_{Pygo1} was associated with B9L's HD1 (Figure 4.20 D and E).

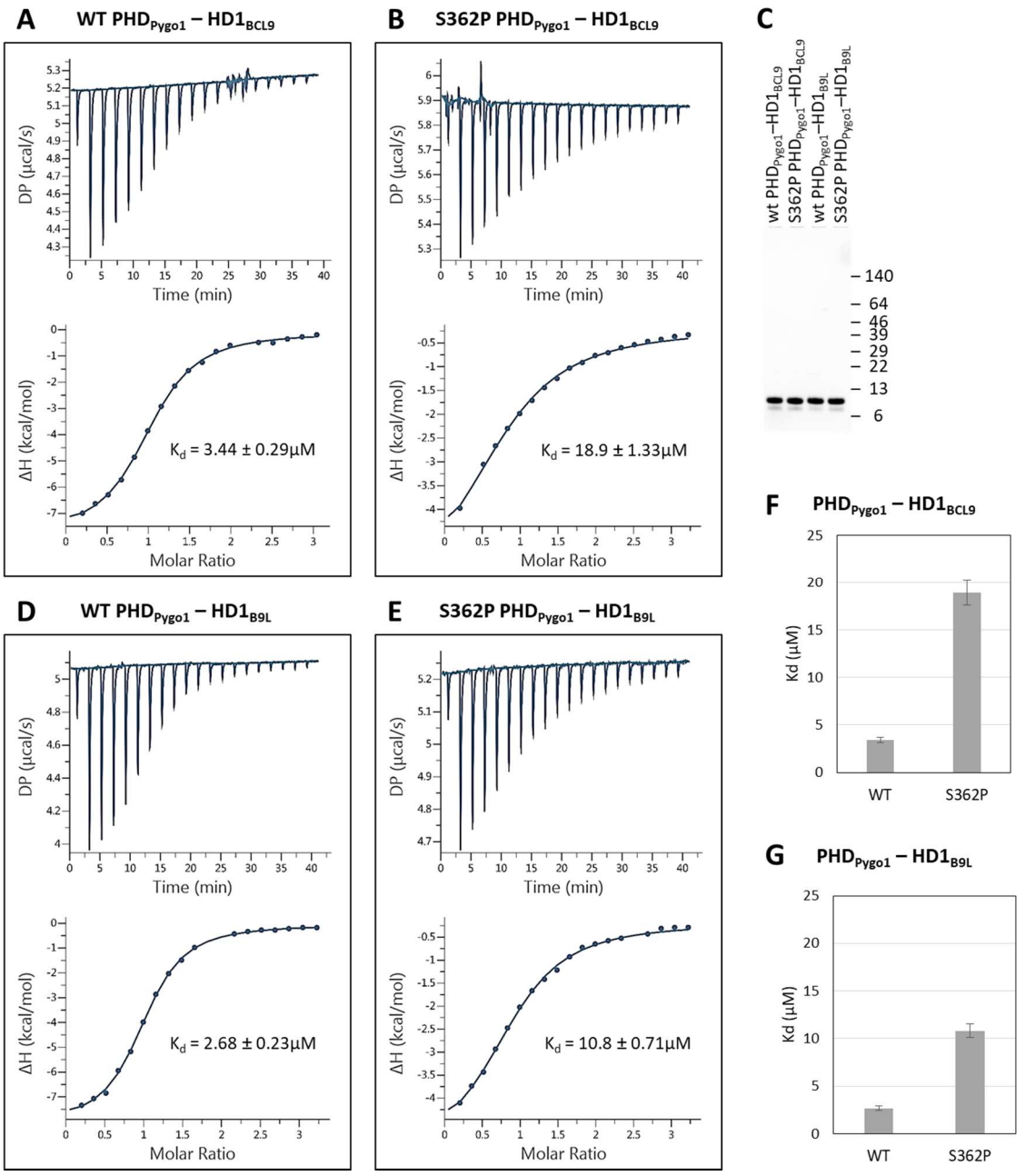


Figure 4.20. Pygo1 S362P shows a decreased histone-binding affinity compared to the wild type. The mutant PHD domain was expressed in association with the HD1 domain from either Bcl9 or B9L, and the affinity of the complexes to bind a histone H3K4me2 peptide was tested by Isothermal Titration Calorimetry (ITC). The two corresponding wild-type PDH-HD1 complexes were also obtained and tested as a reference. The graphs represent the ITC profiles for the histone-binding of PDH-HD1 complexes. Dissociation constant (K_d) values are given in the individual panels (with fitting errors indicated). A+B) The S362P mutation led to a 5.5-fold decrease in affinity for the histone H3K4me2 peptide when Pygo1's PHD formed a complex with Bcl9's HD1. C) Input of each recombinant PDH-HD1 complex used for the ITC. D+E) A 4-fold reduction in affinity was observed when mutant PHD_{Pygo1} was associated with B9L's HD1. F+G) Representation of the histone-binding dissociation constant values for the complex of PHD with HD1 from Bcl9 (F) or B9L (G).

5. Discussion

5.1 Whole-exome sequencing strategy for the study of FCCTX

5.1.1. FCCTX study cohort

The present work focused on the study of FCCTX, which is a term used to describe a group of HNPCC families with a still unknown genetic basis[246]. Although families classified as FCCTX fulfill the same clinical criteria than LS (Amsterdam I/II), they lack germline mutations in the MMR genes and their tumors are MMR proficient with stable microsatellites. In addition, differences in gene expression patterns, epigenetic profiles, genomic alterations and deregulated canonical pathways have been reported between these two entities[101,105,113], further suggesting that different molecular mechanisms mediate the progression of these two pathologies. Other differences between these two groups have been described at the clinical level, including a higher average age of disease onset and lower risk (or even lack of risk) for extracolonic tumors reported in FCCTX patients[101,105,106]. Although the mean age at diagnosis of our cohort was 54.3 years old, it is worth noting that this cohort was selected from a larger group of FCCTX families whose mean age of onset was 55.1 years old, as previously published[116]. Both of these averages are higher than LS's reported average age at diagnosis (38-54 years old)[106,108–111,102,112–117] and agree with the mean age of onset described for FCCTX in some studies (53-55 years)[109,110,112,114], although they are lower than the average reported by some other (58-63 years)[106,108,111,102,113].

On the other hand, the absence of other types of tumors in FCCTX reported by some groups[101,106] is not supported by our cohort, given that 53.8% of the families here studied fulfilled the Amsterdam II criteria and hence presented at least one HNPCC-associated extracolonic cancer. Even though there were just 2 sequenced individuals with only extracolonic tumors and no CRC, over half of the patients had at least one relative with an HNPCC-associated extracolonic tumor, as shown by the fulfillment of the AMS-II criteria. In addition, it had already been reported that the larger cohort of FCCTX families

studied at our laboratory did not show a lower risk for extracolonic tumors than LS[116]. While this contrasts with some studies[101,106], it is in agreement with some others[101]. It is worth noting that the difference in the clinical criteria used to select FCCTX families in the different studies (only Ams-I versus Ams-I/II) entails a bias in the estimation of the extracolonic cancer cumulative risk in these families. Finally, a preferential distal location was observed in 83.3% of the CRC-affected probands, consistent with other studies and even higher than the 65-75% reported in the literature[101,105,117] and the average reported for the larger FCCTX cohort from our hospital (76%)[116].

5.1.2. Suitability of the WES strategy

The different strategies that can be used for the identification of cancer-risk alleles that would explain the cancer susceptibility in FCCTX families include linkage analyses, GWAS and NGS. However, among these, the former presents considerable limitations in the presence of heterogeneity[99]. This questions it's suitability for the study of this condition, given that FCCTX is thought to be a heterogeneous group of families presumably comprising different yet-to-be-discovered genetic syndromes involving high- or moderate-penetrance mutations in novel cancer-predisposing genes[99]. Moreover, FCCTX could also result from a combination of low-penetrance mutations in different genes[107], or even from aggregation of sporadic cases due to shared lifestyle factors[11,106,247,106], which makes even harder the identification of such cancer-predisposing genes. Given the challenge that the study of FCCTX families poses, the two strategies currently used for the study of this group of families are NGS, which focuses on the search for high or medium-penetrance risk genes[248,249], and GWAS, which addresses the low-penetrance multi-gene approach[78]. However, GWAS require much larger sample sizes to achieve an adequate statistical power[135], reason for which NGS was the chosen strategy in our study.

On the other hand, NGS's WES application was selected based on its lower cost-efficiency and more manageable data analysis compared to WGS. Multigene targeted sequencing was not chosen for its reduced number of genes included compared to WES,

which limits the identification of new predisposition genes, and given that most FCCTX families are thought to be inherited by still unknown mechanisms. It is worth noting that, although not included in this thesis, a much larger cohort of 98 patients from high-risk CRC families was assessed by a 94-gene panel at our laboratory. These families fulfilled either the Amsterdam I/II or the Bethesda criteria and presented either MSI or MSS tumors, but without previously detected germline mutations. Among these, 35 were strictly considered FCCTX (AC-I/II and MSS) but could not be included in the present WES study because only one member per family was available, which is the case in most families. Out of these FCCTX cases, pathogenic or likely-pathogenic variants were identified in 4 genes: *CHEK2*, *SMAD4*, *HNF1A* and *TP53*[250] (see publication attached in the Appendix, A6). These explained a small portion of the families, but, as expected, most of them remained unexplained with either no mutation detected or one or more variants of unknown significance[250]. This shows that the multigene panel approach is a good strategy when no more family members are available and for clinical purposes, but a high number of families is necessary in order to ensure the achievements of results.

WES-based studies for the search of germline disease-causing variants require the sequencing of at least 2 (and ideally more) affected members of each family in order to select shared variants. Therefore, the present thesis shows the results obtained by an NGS whole-exome analysis of between 2 and 3 members from 13 FCCTX families. Even though we did not expect to explain all the families, we hoped to find potential causal mutations in at least a fraction of the studied families. The heterogeneity of FCCTX families implied that it was unlikely that the different families here studied would share a common genetic predisposition, reason for which each family was studied individually. Indeed, many research groups with much larger sample sizes have failed to find a common genetic predisposition in this group of families, losing hope of discovering a few shared genes responsible for the increased cancer risk in most FCCTX families[136,146,150,248,249]. Alternatively, claiming that this group of families is excessively diverse, some researchers are now opting for dividing FCCTX into more homogeneous subgroups, hoping to be able to solve the problem this way.

5.1.3. NGS data analysis, filtering and prioritization

One of the most critical points in NGS is the data analysis, reason for which bioinformaticians have become indispensable[99]. Nothing could be done without the correct interpretation of the vast amount of data that NGS generates[249]. However, it is worth noting that there is always a bias attached to the data analysis, the *in silico* studies and the filtering strategy, which highlights the importance of choosing the best filtering approach in order not to miss any potential cancer-risk variant. The strategy used for variant analysis and filtering in this study intended to be as objective as possible, always stablishing cutoffs based on other published studies[249,251].

After the data analysis and thorough filtering of the variants detected, a total of 460 filtered variants were left for the 13 families, ranging from 11 to 73 in each of the families. The discrepancy in the number of filtered variants could be explained based on the family members that were sequenced. Consistent with other studies[248], those families in which 3 members were sequenced had a lower number of filtered variants, and therefore candidate variants, than those in which only 2 members were studied. In the same way, the availability of distant family members resulted in a more stringent filtering and a consequent lower number of shared variants, while the study of first-degree relatives, such as siblings, meant a larger number of shared variants and a more time-consuming prioritization step. Certainly, the only families with less than 20 filtered variants were those in which extended family members had been sequenced (cc406, cc525, cc565, and cc763). On the other hand, sequencing a healthy member from the family was another way of reducing the number of variants left. However, it is worth noting that there is always the risk of missing a variant that does not have a complete penetrance, which could be expected in FCCTX families. For that reason, even when using a healthy relative for the filtering, special attention was paid to all the variants prior to that filtering step, in order not to miss any important mutation.

A further prioritization of these filtered variants, validation and segregation study, led to a final list of candidate genes. The prioritization step based on the search of online databases and literature is unarguably the most biased as well as the most time-consuming step, but was indispensable given the high number of filtered variants. The prioritization step was also in the lines with other similar studies[249,251], and based on the obtained results, we believe that the designed strategy has been successful.

5.1.4. Validation, segregation and LOH

The segregation study was used to discard those variants that clearly did not segregate with the disease when using germline DNA. Unfortunately, it was not very informative in quite some of the cases, mainly for being done in young healthy relatives that may or may not develop CRC and other cancers in the future. Regarding the LOH, it was only used as additional information because it does not predict the pathogenicity of a mutation, since it has been reported to have a dual role in HNPCC families, with no preferential loss of the wild-type or the mutant allele. As a matter of fact, different carriers of well-known pathogenic MMR mutations have been reported to show the three different LOH statuses (no LOH, LOH of the mutant and LOH of the wild-type), even within the same family[252]. In addition, there are other forms of inactivation of the wild-type allele, such as the presence of somatic mutations in the other allele, or hypermethylation of the promoters. It is also worth noting that Sanger sequencing is not a quantitative method for the detection of LOH, so observations of the reduction of one of the alleles would need to be confirmed by other methods, such as the analysis of polymorphic microsatellites close to the region of interest. Interestingly, LOH of the mutant allele was more frequently observed than LOH of the wild-type allele.

5.1.5. CRC susceptibility genes

Even though genes reported to be involved in CRC predisposition were carefully screened with a more lenient filtering strategy, no deleterious mutations in any of the high-risk CRC susceptibility genes reported in previous studies[110] were found in our

cohort. We did detect 5 missense variants in well-characterized cancer genes shared by the affected members sequenced in 6 of the families, including 3 families with *POLE* variants, another 2 with *BRCA2* variants and 1 with a *PALB2* variant. Out of these, *POLE* is the only "typical" CRC gene, and germline mutations in the exonuclease domain of the *POLE* are associated with an increased risk for CRC[253]. However, none of the variants affected the exonuclease domain, which is the one reported to have pathogenic variants. On the other hand, *BRCA2* and *PALB2* are breast cancer genes, but have been previously described in a CRC context[75]. As a matter of fact, a *BRCA2* mutation had been previously identified in our laboratory in another FCCTX family[142]. However, based on the nature of the variants and ClinVar reports, we believe there is enough evidence to claim that these variants are not involved in the increased cancer susceptibility of the families here studied. On the other hand, two missense mutations in two genes that had been considered potentially involved in CRC predisposition (*WIF1* and *LAMB4*) were also found, but based on the effect of the variants and that the causality of neither of them was supported in their corresponding studies, they were not selected as candidate variants either.

In addition, a number of variants were identified in genes previously reported to be associated to CRC risk through GWAS studies. Nonetheless, only one of these variants provided strong enough evidence for its further consideration. This variants was *ABCC2* c.1209+2T>G, a splice donor variant detected in family cc406. ABCC2 is an ABC transporter that has been associated with chemoresistance[254] involved in the ARE–Nrf2 cellular detoxification pathway[255] under oxidative stress[256]. The remaining variants were all missense and did not suggest to play a relevant role as high or moderate-risk alleles, although they may still act as low-risk variants. It is worth noting that this selection only included low frequency variants (MAF≤0.05), given that the purpose of this study was the identification of moderate or high-risk CRC genes. Other common variants in GWAS loci could also be playing a role in the cancer predisposition of these families, most of which are actually located in non-coding regions of the genome not detected by WES.

5.1.6. Other approaches: tumor WES and CNV analysis

Another application of NGS for the search of causal mutations in unexplained families is the study of germline-tumor pairs. To this aim, the tumor of 2 individuals was sequenced and compared to their germline WES data. This was done in order to look for second hits in the tumor that suggest a causality of the corresponding germline variants. However, no second hits were detected in any of the tumors for the filtered variants, which was not a surprise given the low number of samples studied. In order to achieve more results a larger number of samples would need to be analyzed, but that was not possible in our case based on the quality and/or quantity of the FFPE tumors required for optimal results. It is worth noting that this analysis has to take into account the tumor cell content of the FFPE blocks, and can be challenging when a low tumor percentage is observed by the pathologists, which could lead to the loss of some alterations.

NGS data can be also used to look for germline CNVs in the studied samples, and some studies have indeed found alterations this way[138] and reported germline CNVs in FCCTX[257]. For this reason, in addition to the main search for candidate variants in which our study has focused, we also analyzed the WES data searching for CNVs. Unfortunately, our efforts were not fruitful and we did not detect any potential variation. It is worth noting that WES is less reliable than WGS when it comes to this type of study, due to its lower coverage uniformity[153]. Therefore, we cannot rule out the possibility of the presence of CNVs not detected by the exome sequencing in those patients that were not confirmed by CGH.

5.1.7. Candidate variants and genes

A total of 45 candidate variants were selected for the 13 FCCTX families, ranging from 1 to 5 for each of the families. Deleterious mutations were more often prioritized than missense ones based on the relevance of the consequences, which explains the high proportion of this kind of mutations, even though missense changes are more common.

Regarding the pathways in which the genes affected by candidate variants are involved, 12 variants (27.3%) were located in genes involved in DNA repair. Out of these, a truncating variant in *POLQ* was detected, involving the loss of most of the protein. POLQ, also known as polymerase theta, is involved in the response to DNA double-strand breaks[258,259], interstrand-cross-link repair, base excision repair and DNA end-joining[260–262]. And even though *POLQ* has never been linked to CRC before, some other polymerases have (namely *POLE* and *POLD1*)[83]. Interestingly, a case-control study showed an association between a *POLQ* variant (c.-1060A>G) and hereditary breast cancer[263]. On the other hand, higher expression levels of *POLQ* in breast and colorectal cancer are correlated with poor outcomes[264,265]. Another gene that harbored a frameshift variant was *MUS81*, which plays a critical role in the resolution of recombination intermediates during DNA repair after inter-strand cross-links, replication fork collapse and DNA double-strand breaks[219]. In addition, MUS81 is involved in the Fanconi anemia pathway[266], which has been previously reported to be associated to some FCCTX families[150]. These two candidate variants have a high potential of playing a role in the cancer predisposition of the families in which they were found, although the one in *MUS81* is located at the end of the protein and does not affect any protein domains.

HELQ is also involved in the fanconi anemia pathway and interacts directly with the *RAD51* paralogue complex BCDX2. It this way, it functions in parallel to the Fanconi anemia pathway to promote efficient homologous recombination at damaged replication forks, suggesting a critical role for HELQ in replication-coupled DNA repair, germ cell maintenance and tumor suppression in mammals[267]. However, the variant identified in *HELQ* is an inframe variant that only means the elimination of lysine residue. Although not located in any protein domain, this change affected an alpha-helix, but its relevance is still questionable.

On the other hand, other of the reported DNA repair genes affected by deleterious candidate variants do not have such a clear role and may be arguable. That is the case of *NME7*, a gene encoding a nucleoside diphosphate kinase whose 3'-5' exonuclease activity suggests roles in DNA proofreading and repair[268]. Another controversial gene is *PHACTR2*, reported to contribute to the DNA repair capacity in lung cancer[269], but with little

supporting evidence other than that. As for the variants affecting the NRF2 pathway, a frameshift mutation in *TXNRD3* and a splice donor variant in *ABCC2* can be pointed out. *TXNRD3* is a thioredoxin reductase that plays a key role in redox homoeostasis[270], while *ABCC2* is an ATP-binding cassette transporter that acts as a oxidative stress adaptation factor[271]. Apart from these, another frameshift variant was found in *SETD6*, but since this gene was studied more extensively it will be further discussed in section 5.2.

A considerable number of candidate variants (22.7%) also affected putative tumor suppressor genes. Out of these, there were only 2 deleterious variants, in *PTPRT* and in *IQGAP2*. The former is a well reported tumor suppressor that was studied more in depth and will be discussed in section 5.3, while the latter has been more recently identified as a tumor suppressor in hepatocellular carcinoma and prostate cancer[272–274]. In addition, its reduced expression has been related to poor prognosis[275] and it has also been associated to ovarian cancer progression via activating Wnt/β-catenin signaling[276]. It has also been reported to play a role in the NF-κB pathway[277]. There were other relevant pathways affected by candidate genes, including the Wnt pathway, which has very important implications in CRC[243]. Among the variants located in Wnt genes there are two missense variants in *PYGO1* and *NKD2*, which will be further discussed in section 5.4, and a splice region variant in *INVS*, which acts as a molecular switch between the different Wnt signaling pathways, inhibiting the canonical Wnt pathway[278].

5.1.8. Variants selected for further characterization

Based on the nature of the alteration and the relevance of the genes involved, 3 candidate variants were selected for further characterization. These variants were the most promising candidates in their respective families at the time of selection, and were the following affected the genes *SETD6*, *PTPRT* and *PYGO1*. However, this does not mean that other candidate genes from other families have any less causal potential. The results obtained in the more extensive characterization of these 3 variants are further discussed in sections 5.2, 5.3 and 5.4. Although there is obviously not enough evidence to support the investment in time and costs that further studies to characterize all of the candidate

variants would involve, there are some candidates that would undoubtedly deserve a deeper characterization in the future.

5.2 Characterization of SETD6 M264Ifs*3

5.2.1. *SETD6* frameshift variant in family cc598

The first candidate variant selected for further characterization was *SETD6* c.791_792insA, p.(Met264Ilefs*3), identified in FCCTX family cc598. This was a rare frameshift mutation found to cosegregate with the disease within the family, since it was carried by three CRC-affected members (II:1, II:2 and II:4) while absent in three healthy relatives (II:3, III:1 and III:2). Even though two of the healthy individuals were too young to be informative, the segregation of the variant with CRC in this family was as perfect as it could have been.

On the other hand, no LOH was observed in the tumor of member II:2, and while the tumor of deceased member II:4 seemed to show a reduction of the mutant allele, it was not possible to compare the electropherogram with its corresponding germline DNA to confirm this observation. Nevertheless, it is worth noting that member II:4 had a more advanced form of the disease than II:2 and hence such loss may have been a later event during tumorigenesis. Not to mention that in HNPCC tumors, both wild-type and mutant LOH (as wells as no LOH) have been reported in the past, even for well-known pathogenic MMR mutations[252].

5.2.2. Selection of *SETD6* M264Ifs*3 as a candidate variant

SETD6 is a lysine methyltransferase involved in many relevant pathways. In the first place, it has been known for a long time to monomethylate the RelA subunit of transcription factor NF-κB, negatively regulating such pathway[163]. Additionally, SETD6 has been proven to participate in the canonical Wnt signaling cascade by forming a complex with PAK4 and the transcription factor β-catenin, activating the transcription of its target genes[167]. Recent studies have also linked SETD6 to the regulation of the nuclear hormone

receptor signaling pathway[170], embryonic stem cell differentiation[169], oxidative stress response[168,279], mitosis and proliferation[171]. The relevance of many of these pathways in the context of cancer and the predicted impact of the truncation justify the selection of this variant for additional characterization.

5.2.3. Consequences of *SETD6* M264Ifs*3

Regarding the consequences of the studied *SETD6* mutation, this variant results in the introduction of two amino acids at positions 264 and 265 of the protein, followed by a premature stop codon. The truncated protein (SETD6-N) lacks its C-terminal half but retains an intact SET domain, which is responsible for its catalytic activity. Functional assays have shown that mutant SETD6 exhibits dominant negative properties in cell-free *in vitro* systems and in a colon cancer *in vitro* cell line model. It has been proved that, although mutant SETD6 displays similar localization, expression and substrate-binding ability as the wild-type, the mutant protein loses its enzymatic activity. Indeed, unlike wild-type SETD6, SETD6-N lacks both its automethylation activity and the ability to methylate two previously identified substrates, PAK4 and RelA[163,167]. In addition, the two alleles (wild type and mutant) were found to be expressed in the tumor of one of the carriers, and the two forms of SETD6 were shown to compete for their substrates both in cell-free systems and in colon cancer cells, pointing to a dominant negative role.

5.2.4. Downstream effects

As a result of the aforementioned consequences, this mutation may have several downstream effects on the different pathways in which SETD6 is involved[163,167–171]. For example, SETD6-dependent methylation of the NF-κB subunit RelA has been shown to be critical for basal inhibition of NF-κB signaling in the absence of stimulation, so a deficient SETD6 could be predicted to result in the hyperactivation of the NF-κB pathway[163]. As expected, two NF-κB target genes (IL1A and IL8) were shown to be overexpressed in a CRC cell line as a result of the mutation, pointing to the upregulation of the NF-κB signaling pathway. The NF-κB family of transcription factors has an essential role in inflammation

and innate immunity, but it has also been increasingly recognized as a crucial player in many steps of cancer initiation and progression[242]. In fact, activation of the NF-κB pathway has been positively associated with multiple types of cancer, including CRC[280].

In addition, it had been previously demonstrated that methylation of PAK4 by SETD6 promotes the activation of the Wnt/β-catenin pathway[167], whose deregulation has been shown to contribute to CRC development, including in HNPCC[129,281–283]. Although some differences were observed in Wnt target gene expression for the 2 forms of the enzyme, these preliminary results were not consistent and need to be addressed in the future. The loss of SETD6's function and predicted aberrant regulation of NF-κB and/or Wnt signaling could be contributing to the initiation or progression of CRC in the studied FCCTX family. However, SETD6 has several other substrates and is involved in numerous other signaling pathways[166,168,170,279], so future studies are also needed in order to define other downstream consequences of this *SETD6* truncating mutation, ideally using the *SETD6* knock-out HCT116 cell line generated.

5.2.5. Involvement of *SETD6* in cancer predisposition

Given all the results presented earlier, we propose that the presence of *SETD6*'s mutant allele would presumably increase the cumulative risk of developing CRC throughout the life of the carriers, as compared with the general population. More research should be done in order to determine the penetrance of this mutation, whether it is a high or moderate-risk allele, and if there is also an association with other HNPCC-associated cancers. Interestingly, the same *SETD6* frameshift mutation had already been proposed to be associated with ovarian cancer in an extensive study by Kanchi and colleagues[284]. Nonetheless, the search for new genes by exome sequencing in FCCTX families has demonstrated that more than one gene can be involved in the increased cancer, and even when there is a high-risk gene involved, there might be also low-penetrance alleles cooperating in the process as risk modifiers. Moreover, there are other elements that might be modulating the cancer risk, including lifestyle and other environmental factors[8,285]. Although our results are not enough to claim that *SETD6* alone

is responsible for the increased CRC-risk in this FCCTX family, they certainly point to a role of this variant in cancer development.

5.2.6. Other candidate variants in this family

Noteworthy, the WES analysis followed by rigorous filtering identified additional candidate genes beside *SETD6* (Table 4.3). *CCDC60* showed a stop gain variant that results in the loss of 30 amino acids, 20 of which belong to a domain of unknown function, while the remaining 10 belong to a low complexity region. This is a gene of unknown function, but that has been associated with renal cancer[286]. *ST18* showed a missense variant and is tumor suppressor gene that has been associated with breast cancer[287]. Even though just missense, the candidate variant affects a protein domain and involved a change to proline, which is an amino acid with very special conformational properties[241]. It is worth noting that other candidate variants initially considered were later disregarded based on the lack robust evidence supporting their prioritization. Even though *SETD6* M264Ifs*3 was undoubtedly the most promising candidate among them all, we cannot exclude the possibility that the additional candidate variants – and even other filtered variants – may contribute independently or together to the pathology of FCCTX. As previously mentioned, the selection of the *SETD6* variant over the others was based on the effect of the mutation and the function of the protein.

5.2.7. *SETD6* as a candidate CRC predisposing gene

Together, our findings suggest that this truncating dominant negative mutation in *SETD6* could potentially explain the cancer predisposition of this FCCTX family. We therefore propose *SETD6* as a novel gene involved in CRC predisposition. The results here presented certainly point to a pathogenic role of *SETD6* c.791_792insA, p.(M264Ifs*3), though not enough to prove that *SETD6* alone is responsible for their increased cancer risk. Although no other *SETD6* variants were found in the remaining 12 families studied, nor in 109 familial CRC cases provided by the Clinic Hospital (IDIBAPS, Barcelona), the

screening of this gene in a larger group of patients could provide more insights into its role in other FCCTX families.

5.3 Characterization of PTPRT D1367Gfs*24

5.3.1. *PTPRT* frameshift variant in family cc765

The most promising candidate variant for family cc765 was a novel frameshift mutation in the *PTPRT* gene: c.4099dup, p.(Asp1367GlyfsTer24). Although the segregation within the family was not perfect, this variant showed a compatible segregation with the disease. That is because it was not carried by a member who had developed CRC at the age of 85 (II:3). However, if we take into account the elevated age of this patient and the prevalence of the disease in the general population, this could be perfectly explained as a phenocopy. In addition, this variant was carried by 3 other members affected with different cancers at younger ages, two of which belonged to the HNPCC spectrum: II:6 had CRC at 70, II:7 had endometrial cancer at 28, and II:4 had a breast cancer at 66. Moreover, the variant was not present in the two elderly healthy relatives studied (II:2 and II:5, of ages 91 and 85, respectively).

On the other hand, there was no LOH of the wild-type allele in any of the tumors tested (from members II:4 and II:6), although the breast tumor showed a possible partial loss of the mutant allele. This would need to be confirmed, given that Sanger sequencing is not a quantitative method for the detection of LOH. Nonetheless, *PTPRT* has been recently reported to have frequent promoter hypermethylation in CRC among other tumors[177,178], and the promoter region of *PTPRT* was certainly found to be hypermethylated in the two tumors of the carriers that were tested (breast and colon), which could be a different mechanism of inactivation of the wild-type allele in the tumor. Indeed, an allele-specific expression assay showed that the expression of the wild-type *PTPRT* allele was significantly reduced in the colorectal tumor of member II:6 when compared to healthy colon from the same individual, while no decrease was observed in the expression of the mutant allele. This suggests that the aforementioned hypermethylation only affects the wild-type allele and could be considered a second hit

involved in the inactivation of this tumor suppressor, leading to tumor development in the carriers.

5.3.2. Selection of *PTPRT* D1367Gfs*24 as a candidate variant

PTPRT is a tyrosine phosphatase that has been proven to behave as a tumor suppressor through the inactivation of at least two well-known oncogenes, STAT3 and paxillin. Its involvement in relevant pathways (including the paxillin-p130-PI3K-Akt signaling axis and the IL6-JAK-STAT3 pathway) can be pointed out, through which it regulates the expression of genes involved in cell survival, apoptosis, proliferation, growth and cell migration[182,245]. In addition, frequent somatic inactivating alterations in many tumors have been reported to act as driver mutations that promote tumor development and progression[269]. Last but not least, even though it is located quite at the end of the protein, the identified variant is a truncating mutation that affects one of the two catalytic domains of PTPRT. For all of the above, this variant was further prioritized among the other candidate variants for a more in-depth characterization.

5.3.3. Consequences of *PTPRT* D1367Gfs*24

PTPRT D1367Gfs*24 results in the loss of the last 95 amino acids of the protein, including an important portion of the second phosphatase domain, also known as D2. Figure 5.1 A has been adapted from Lui et al[186] to show a representation of the tertiary structure of the D2 domain, showing the region that surrounds the phosphorylated tyrosine. The yellow striped area shows all the residues that are lost with the identified frameshift mutation. The 95 amino acids lost represent 35.7% of the D2 domain, as seen in Figure 5.1 B, which shows the amino acid sequence of such domain with and without the mutation. The different colors mark the proximity to the phosphorylated tyrosine, pointing out that a considerable amount of essential D2 residues are lost with the mutation. Actually, the majority (69.2%) of residues surrounding the phosphotyrosine and in direct contact with the target protein are lost with this mutation (amino acids in red)[186].

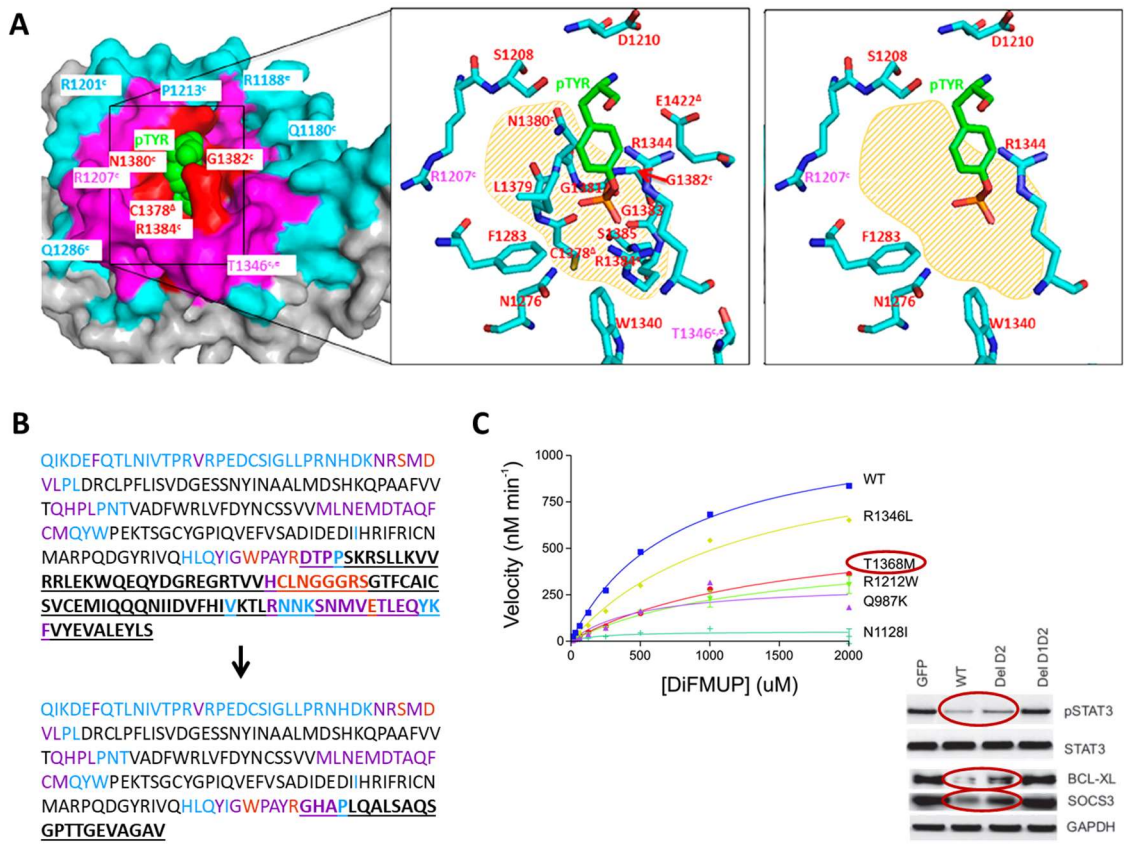


Figure 5.1. Additional consequences of *PTPRT* D1367Gfs*24. A) Visualization of the D2 domain residues surrounding the substrate's phosphotyrosine (pTyr, in green). The colors indicate the proximity to the pTyr, with the closest residues marked in red, followed by the purple and blue ones. The yellow stripped area shows all the residues in contact with the pTyr that are lost with the mutation. The figure has been modified from Lui et al. 2013. B) Amino acid sequence of PTPRT's D2 domain with the effect produced by the mutation. Colors indicate the proximity to the pTyr as in A. Underlined are all the residues that differ between the two forms of PTPRT. C) Previous studies proving that the D2 domain is essential for PTPRT's phosphatase activity, modified from Wang et al. 2004 (top) and Zhang et al. 2007 (bottom). Circled in red are the minimum expected consequences of the studied mutation. The top figure shows how PTPRT's catalytic activity is reduced by half in a missense mutant that affects one on the residues that are lost with our mutation. The bottom figure shows how deletion of the D2 domain has a significant effect on the levels of phosphorylated STAT3 and expression of two of it's target genes (BCL-XL and SOCS3), all of which are increased upon D2 deletion.

Although the first phosphatase domain (D1) is the one known to be responsible for the phosphatase activity of the protein, D2 is responsible for the regulation of this activity and has been proven to be essential for PTPRT's activity[245]. As a matter of fact, Wang et al reported that just a missense mutation affecting D2's residue 1368, which is lost in our mutant, was enough to decrease the enzyme's activity by half (Figure 5.1 C,

top)[176]. The relevance of this second catalytic domain was also pointed out by Zhang et al, who showed how the deletion of just the D2 domain had significant effects on the levels of phosphorylated STAT3 (a well-known PTPRT target), as well as on the expression levels of two of its downstream target genes, *BCL-XL* and *SOCS3* (Figure 5.1 C, bottom)[182]. All things considered, it can be predicted that this mutation will result in significant consequences for the activity of this phosphatase.

5.3.4. Downstream effects

The PTPRT phosphatase is known to dephosphorylate two main target proteins, STAT3 and paxillin[182,183,245], both of which are well-known oncogenes[288]. These two proteins are activated upon phosphorylation by different protein tyrosine kinases, so the PTPRT-mediated removal of the phosphate group results in their inactivation. This results, in turn, in the inhibition of their downstream pathways, through a decreased phosphorylation of paxillin's target proteins[183,245] and a decreased expression of STAT3's target genes[182]. *BCL-XL* and *SOCS3* are two of those STAT3's target genes, which have been proven to show increased expression upon PTPRT depletion[182]. Consistent with our hypothesis that this germline PTPRT variant may be involved in the cancers of this FCCTX family, the tumors from the two carriers tested presented a significantly increased expression of *BCL-XL*, which is an oncogenic driver in CRC[289]. This was observed in both the colon and breast tumors when compared to healthy tissue controls, as well as to the corresponding healthy colon or breast tissue from the same patient.

In contrast, *SOCS3* expression was decreased in both tumors. Nonetheless, SOCS proteins are negative feedback regulators of the JAK-STAT signaling pathway[290,291], and it has been proven that *SOCS3* is usually downregulated in CRC, even when IL-6 and STAT3 are upregulated[292]. This is thought to occur through different mechanisms, such as the hypermethylation of *SOCS3* gene promoters, allowing inflammatory cytokines IL-6 to activate STAT3's signaling pathway while inhibiting the expression of *SOCS3*[292], with the purpose of inactivating its negative feedback. This negative regulation of *SOCS3* expression, mediated by the activation of IL-6/STAT3, leads to imbalance and sustained

activation of STAT3 signaling pathway[292]. The same study by Chu et al also showed that SOCS3 plays an important role inhibiting tumor development and that reduced *SOCS3*'s expression affects tumorigenesis and CRC progression, promoting growth and metastasis[292]. For all the things mentioned, both results are compatible with a pathogenic role of this *PTPRT* variant.

5.3.5. Involvement of *PTPRT* in cancer predisposition

PTPRT has been long known to be frequently mutated at the somatic level in many cancers, including CRC[176]. These mutations have been proposed to be driver mutations involved in the development of the tumors in which they are found[176]. In addition, frequent hypermethylation of PTPRT's promoter has been recently described as a mechanism of genetic inactivation in CRC and head and neck cancer[177,178]. It has been proven that *PTPRT* normally functions as a tumor suppressor gene based on a number of reasons. In the first place, overexpression of wild-type *PTPRT* suppresses colony formation of CRC cells, whereas tumor-derived mutations impair this ability. Furthermore, *PTPRT* knock-out mice are susceptible to carcinogen-induced colon tumor formation[183] and *PTPRT* knock-out increases the size of colon tumors in the $Apc^{min/+}$ genetic background[245]. Such tumor suppressor role may be mediated by the inhibition of two well-known oncogenes (STAT3 and paxillin), both of which have been proven to be targets of PTPRT's phosphatase activity[182,183].

However, this is the first time that *PTPRT* is linked to hereditary cancer, which highlights the importance of the results here presented. Even though the role of this protein in tumor development is undeniable, more studies are needed to confirm its involvement as a cancer susceptibility gene. Studying this gene in a larger FCCTX cohort would help with this task. As a matter of fact, the screening of *PTPRT* in a FCCTX cohort from the Catalan Institute of Oncology (IDIBELL, Barcelona) identified one additional rare missense germline mutation in the D2 domain of PTPRT in one of the CRC-affected patients. However, the effect of this variant cannot be currently established. Even if *PTPRT* is validated as a cancer-risk gene, we cannot discard the possibility that other factors –

either environmental or genetic – may be contributing to the increased cancer susceptibility or modifying its effect.

5.3.6. Other candidate variants in the family

Even though *PTPRT* had the highest potential for the explanation of the increased cancer risk, another 4 candidate variants were prioritized for this family: 2 missense variants in *MAP3K6* and *ABTB1*, and a splice region variant in *INVS*. *MAP3K6* is reported to act as a tumor suppressor, while *ABTB1* is a mediator of PTEN and *INVS* acts as a molecular switch between the different Wnt signaling pathways, inhibiting the canonical Wnt pathway.

5.3.7. *PTPRT* as a candidate cancer predisposition gene

Taken together, all the results here presented point to a probable causal role of *PTPRT* c.4099dup, (p.D1367Gfs*24), in the cancer susceptibility of FCCTX family cc765. For that reason, we propose *PTPRT* as a novel cancer predisposition gene. However, more research is necessary to confirm the causality, penetrance, conferred risk, preferred cancer location and the like.

5.4 Characterization of PYGO1 S362P

5.4.1. *PYGO1* missense variant in family cc28

Even though no truncating candidate variants were prioritized for family cc28, a novel *PYGO1* missense mutation was selected for further characterization. This variant was c.1084T>C, p.(Ser362Pro), and was predicted to be deleterious by all the *in silico* tools used and to affect an important functional domain of the protein. The two CRC-affected studied members of this family (II:1 and III:1) carried this variant, and the initial segregation study done for this mutation showed that it was not present in 2 out of the 3 healthy relatives tested (III:4 and III:7). Unfortunately, a later analysis of the tumor of a third CRC-affected member that was initially not available revealed that this mutation was not shared by this patient. It is worth noting that this result was obtained from tumor DNA, so there is always the possibility that one of the alleles had been previously lost (which is sometimes observed despite the fact that a remainder of the lost allele is expected to be seen). For this reason, a negative segregation result from tumor tissue is always controversial and was not used to discard any variant. As far as the LOH study is concerned, there was no loss of any of the alleles in the tumor of member (III:3), which, as already discussed, is not informative in the context of HNPCC.

5.4.2. Selection of *PYGO1* S362P as a candidate variant

Pygo1 is a nuclear protein that forms a complex with β-catenin, BCL9 and other proteins at the chromatin, promoting the transcription of Wnt target genes[191]. In addition, the canonical Wnt pathway has an important role in cancer, especially in CRC, where it is frequently upregulated[243]. Considering that no truncating mutations were prioritized for this family, the relevance of the Wnt pathway in CRC and that this missense variant affected the most important functional domain of the protein, this candidate variant was selected for further characterization. It is worth noting that at the moment of selection

the segregation study in the tumor of member III:3 had not been performed since the tumor block was still not available.

5.4.3. Consequences of PYGO1 S362P

PYGO1 S362P consists on the substitution of a serine for a proline and affects the PHD domain of the protein, which is a zinc finger domain essential for Pygo1's function. Moreover, the affected residue is situated in a turn between two β-sheets, and is in very close proximity to one of the Zn atoms bound to the protein. As a matter of fact, Ser362 is situated right next to one of the cysteines that directly interacts with such Zn[194], and considering that proline is a very special amino acid it was reasonable to expect an effect on the Zn binding ability and function of the protein[241]. However, based on Pygo1's positive role in the Wnt pathway[191], the original hypothesis was that this change could be increasing the affinity of the protein for its substrates, which, although less likely than the opposite effect, would make more sense in the context of CRC. Nonetheless, mutant Pygo1 showed a significant decrease in the histone-binding affinity when forming a complex with both Bcl9 and B9L, when compared to the wild type protein. This result is contrary to our initial hypothesis and means that mutant Pygo1 could entail a decreased Wnt pathway. The effects that this could have in the tumor development are not known. Interestingly, the proportion of protein that eluted off the gel filtration in the void (which is indicative of misfolded protein) was higher for the S362P mutation, although the overall amount of protein that eluted in the void was about the same for wt and mutant. Although not a strong piece of data, this may indicate a slight reduction in stability of the mutant protein.

5.4.4. Other candidate variants in the family

Other candidate variants in the family included 4 missense variants in *CHID1*, *MMP11*, *AHRR* and *NKD2*. Given that *PYGO1* has not been proven to be responsible for the increased CRC susceptibility of this family, one or more of the other candidate variants selected for family cc28 may be able to explain the increased CRC-susceptibility in this

family. *CHID1* encodes a saccharide- and LPS-binding protein with possible roles in pathogen sensing and endotoxin neutralization that is involved in chronic inflammatory diseases and shows a decreased secretion in CRC[166]. *MMP11* is a matrix metalloprotease or zinc-dependent endopeptidase that is overexpressed in many cancers[293] and plays an important role in the progression of epithelial malignancies, including CRC[294]. Even though it has been shown that *MMP11* knock-down inhibits proliferation and invasion[295], the mutation detected is located in the cysteine switch of the protein, which binds a zinc atom in the inhibited form[296]. Actually, one of the cysteines responsible for the interaction with the Zn atom (Cys80) is only two amino acids away from the affected residue (Pro78), so this substitution may impair the Zn union, which would activate the protein leading to invasion and proliferation. *AHRR* is a tumor suppressor in multiple cancers[168], and has been reported to be involved in CD40/CD40L signaling that inhibits CRC proliferation, induces apoptosis, stalls cells at G0/G1 and influences cell adhesion and metastasis[169]. Last but not least, NKD2 is an antagonist of the Wnt signaling pathway[190], and although no domains have been described so far in this protein, the change affects the only known helix in the area, which makes it a good candidate for future characterization.

5.4.5. PYGO1 is not confirmed as a candidate CRC predisposition gene

All the results presented above do not point to a causal role of *PYGO1* c.1084T>C p.(Ser362Pro) in this family. This is based mainly in the lack of segregation and on the fact that the mutation would be expected to decrease the Wnt pathway, which is the opposite of what is usually seen in CRC. Although we cannot be sure of the consequences that this mutation may have on tumor initiation and progression, the results here presented do not justify additional characterization. Instead, some of the remaining candidate variants may be involved in the increased cancer predisposition and could be further studied in the future.

5.5 Limitations of this study

5.5.1. Cohort and patient selection

One of the limitations of this study is the reduced size of our cohort due to the difficulty of finding families that meet the specific criteria and have enough available members. Although this may not be so relevant given that each family was studied individually, the study of the selected candidate variants in a bigger cohort of FCCTX or even early-onset CRC patients will provide further insights into the potential role of these genes in cancer susceptibility.

The relevance of selecting the best suitable family members for this kind of study should be also pointed out, ideally including three affected relatives as distant as possible. Although some of the families in which the search for candidate genes has been more successful included 2 affected and 1 healthy relative sequenced, the inclusion of a healthy member can be also considered a limitation, given that the penetrance may not be complete and a healthy member could be the carrier of a pathogenic mutation. In addition, in most of the cases extended family members could not be recruited and two or three first-degree relatives were sequenced, which hinders the filtering and prioritization steps given the higher number of shared variants. Lastly, the inclusion of 4 members with adenomas but without a diagnosed cancer is an important limitation. For that reason, apart from the regular filtering strategy, cancer-risk genes were checked in these families without taking into account these patients.

5.5.2. Family history

There are also limitations attached to the family history provided, which may be incomplete and sometimes inaccurate. The achievement of results strongly depends on the veracity of the information supplied. For this reason, it is always essential to confirm all the diagnoses with medical reports. However, there is still some family information

that completely relies on the good communication between the patients and the genetic counselors. The information provided by the family is also used to classify families according to the clinical criteria (Amsterdam, Bethesda and the like), so inaccurate data could result in the erroneous classification of a family. The reduced size of some modern families also hamper the fulfillment of the clinical criteria, and hence could result in the misclassification of high-risk families.

5.5.3. Methodology

Some of the limitations of NGS, and more specifically WES, have already been discussed in the Introduction section. However, it is worth emphasizing that the main limitation of this chosen method is that some areas of the exome are poorly covered or even uncovered, and that WES only sequences the coding regions of the genome, missing any intergenic or intronic regulatory elements[153]. Therefore, any alterations in those regions would not be detected, so we should assume that we are dismissing a fraction of potential causal variants. However, given that the majority of currently known cancer-causing mutations are located within coding regions and that a minimum coverage of x20 was achieved in >95% of the targeted region, we consider that the lost fraction that could be currently interpreted is relatively small. In addition, WES is not the most reliable method for the detection of CNVs, so we cannot rule out the presence of these alterations in the families not studied by CGH.

5.5.4. Filtering and prioritization strategy

Section 5.1 also addressed another of the main limitations of this study: the filtering and prioritization steps. This limitation is mainly due to the bias introduced by *in silico* analyses and the prioritization approach, the cutoffs used in the filtering strategy and the availability of information in public databases. For instance, variants detected with a very low allele frequency due to technical issues (such as the preferential amplification of one of the alleles) would be lost in the first filtering step. In the same way, this allele frequency cutoff can fail to detected mosaicisms, which have been described in

other hereditary syndromes such as polyposis[79,297]. However, lowering this cutoff would significantly increase the amount of false positives in certain regions, complicating the analysis of the vast amount of data obtained.

In addition, low frequency variants (0.01<MAF<0.05) with moderate or modest effects would be also discarded during the filtering stage, as well pathogenic variants with incomplete penetrance carried by healthy members, as previously discussed. For this reason, healthy relatives were only taken into account at the end of the filtering strategy in order to examine the shared variants prior to this step, and genes previously associated with cancer predisposition were allowed to have a MAF≤0.05. Finally, the prioritization of the genes based on previous information can lead to the underestimation of genes that are still not fully understood and have little information in public databases. This is particularly relevant given that there is always a delay in the publication of scientific discoveries. Moreover, the prioritization step based on the search of online databases and literature is unarguably the most biased step.

All this means that although any finding obtained by this type of study is certain, there might always be some information lost. For this reason, the raw data should be always kept available for possible future new data analyses with different approaches.

5.6 Future perspectives

5.6.1. Candidate variant screening in a larger cohort

One of the first things planned for the future is the screening of the most appealing candidate genes in a larger set of FCCTX families, or even in other high-risk CRC cases, such as early onset CRC or families with a weaker but still noticeable CRC family history. This could confirm the implication of these genes in cancer susceptibility and clarify their role. However, none of the genes here presented are thought to be major contributors of CRC susceptibility in FCCTX families, suggesting that a high number of families should be screened in order to increase the odds of finding other carriers. In addition, population-based studies confirming the absence of the identified variants in healthy individuals from a certain population would strengthen any associations observed with CRC.

These screenings should be done specially for *SETD6* and *PTPRT*, but also some of the still uncharacterized candidate genes with the strongest causal potential, such as those involved in DNA repair or those affected by truncating variants. The follow-up of the families will also make the segregation studies more informative, given that many of the available family members were still too young to be taken into account.

5.6.2. Further characterization of SETD6 mutation

Although a quite thorough characterization was performed for the germline mutation identified in *SETD6*, little is still known about its downstream consequences. Future plans include the study of the effects of this mutation on the different pathways in which SETD6 is involved, such as the NF-κB, Wnt and Nrf2 pathways, but also on cell cycle and proliferation. Although some preliminary assays were already performed, the expression of the different target genes should be analyzed by qPCR using the CRIPR-Cas9 SETD6 knock-out CRC cell line that was generated during the stay at Ben Gurion University (Israel). In addition, other approaches could be designed for the study of the cell cycle,

proliferation and response to oxidative stress. These could include wound healing, viability and cellular antioxidant activity assays.

5.6.3. Further characterization of PTPRT mutation

Even though previous literature firmly supports the detrimental effect of the germline mutation identified in *PTPRT* on the function of the protein, future plans for this gene include the functional characterization of this alteration. As a matter of fact, the mutant version of *PTPRT* has already been cloned during a stay at Case Western Reserve University, and assays to determine the effects of this mutation on PTPRT's phosphatase activity, as well as on a newly identified activity of the D2 domain, are currently being addressed. However, this approach has not been included in this thesis given that experiments are still ongoing and it is too early to draw any conclusions. Other future plans could include the study of the expression of a larger number of target genes in the tumor and healthy tissues from the carriers, as well as in cells transfected with the mutant *PTPRT* clone.

5.6.4. Characterization of other candidate variants

Finally, future plans include the characterization of other candidate variants, which will provide insights into their role in cancer predisposition. However, in order to proceed with additional assays, it is crucial to select the best candidates with strong evidence supporting their likely causal role, given the costs and dedication required for these analyses. Some of the candidate genes that are more likely to benefit from these studies include those involved in DNA repair (especially in the fanconi anemia pathway), as well as those affecting well-described tumor suppressors. On the other hand, truncating variants a priori entail better candidates for further characterization.

It is worth noting that some preliminary expression studies have already been started for a few other candidate variants, including two truncating variants in *SHF* and *POLQ*.

6. Conclusions

6.1 Conclusions

1. The strategy used in this study allowed the successful achievement of a list of candidate variants for each of the 13 FCCTX families.

2. A total of 44 candidate variants were selected, out of which:
    - 27.3% affect genes involved in DNA repair, including a 6.8% involved in the fanconi anemia pathway.
    - 22.7% affect reported tumor suppressor genes.
    - The remaining affect, for the most part, genes involved in pathways that modulate proliferation, differentiation and apoptosis.

3. None of the studied families is explained by high or moderate-risk mutations in previously known cancer predisposition genes.

4. The mutation c.791_792insA, p.(Met264IlefsTer3), identified in the *SETD6* gene in FCCTX family cc598:
    - Impairs the protein's methyltransferase activity without affecting its binding ability.
    - Is expressed in the tumor, competing with the wild-type protein in a dominant negative manner.
    - Increases the expression of NF-κB target genes and has the potential of affecting other SETD6-associated cellular events (such as the Wnt pathway, cell cycle and oxidative stress response).
    - Shows a perfect cosegregation with all the cancers in the family.
    - Is consistent with a causal role in the cancer predisposition of this family, reason for which *SETD6* is proposed as a new cancer risk gene.

5. The mutation c.4099dup, p.(Asp1367GlyfsTer24), identified in tumor suppressor gene *PTPRT* in FCCTX family cc765:
    - Affects key residues of the D2 catalytic domain, which is essential for the protein's phosphatase activity.
    - Presents hypermethylation of *PTPRT*'s promoter in the tumor, together with decreased expression of the wild-type allele.
    - Entails the altered expression of *PTPRT*'s downstream target genes in the tumor.
    - Shows a compatible segregation with the different cancers in the family.
    - Is consistent with a causal role in the cancer predisposition of this family, reason for which *PTPRT* is proposed as a new cancer risk gene.

6. The mutation c.1084T>C, p.(Ser362Pro), identified in *PYGO1* in FCCTX family cc28:
    - Involves a drastic amino acid change that affects one of the zinc fingers of the PHD domain.
    - Decreases its histone-binding affinity, which could decrease the expression of Wnt target genes, opposite to what is expected in a colorectal cancer context.
    - Does not segregate with all the cancers in the family.
    - Is not confirmed as the cause underlying the cancer predisposition of this family.

6.2 Conclusiones

1. La estrategia utilizada en este estudio ha permitido la obtención de una lista de genes candidatos para cada una de las 13 familias FCCTX.

2. Un total de 44 variantes candidatas fueron seleccionadas, de las cuales:
   - El 27.3% afecta a genes involucrados en la reparación del ADN, incluyendo un 6.8% implicado en la vía de la anemia de fanconi.
   - El 22.7% afecta a genes descritos como supresores tumorales.
   - El resto afectan, en su mayoría, a genes implicados en vías que regulan los mecanismos de proliferación, diferenciación y apoptosis.

3. Ninguna de las familias estudiadas se explica por mutaciones causales de alto o moderado riesgo en genes conocidos de predisposición al cáncer.

4. La mutación c.791_792insA, p.(Met264IlefsTer3), encontrada en el gen *SETD6* en la familia FCCTX cc598:
   - Conlleva la pérdida de la actividad metiltransferasa de la proteína sin afectar a su habilidad de unión a sustrato.
   - Se expresa en el tumor, compitiendo con la proteína nativa con un efecto dominante negativo.
   - Aumenta la expresión de genes diana de la vía NF-κB, teniendo el potencial de afectar otros mecanismos asociados a SETD6 (como la vía Wnt, el ciclo celular y la respuesta a daño oxidativo).
   - Muestra una perfecta cosegregación con el cáncer en la familia.
   - Concuerda con un papel causal en la predisposición al cáncer en esta familia, por lo que proponemos a *SETD6* como un nuevo gen de riesgo al cáncer.

5. La mutación c.4099dup, p.(Asp1367GlyfsTer24), encontrada en el gen supresor tumoral *PTPRT* en la familia FCCTX cc765:

- Afecta a residuos clave del dominio catalítico D2, que es esencial para su actividad fosfatasa.
- Presenta hipermetilación del promotor de *PTPRT* en el tumor, junto con disminuida expresión del alelo nativo.
- Conlleva la alteración de la expresión de genes diana de vías asociadas a *PTPRT* en el tumor.
- Muestra una segregación compatible con los distintos tipos de cáncer de la familia.
- Concuerda con un papel causal en la predisposición al cáncer de esta familia, por lo que proponemos a *PTPRT* como un nuevo gen de riesgo al cáncer.

6. La mutación c.1084T>C, p.(Ser362Pro), encontrada en el gen *PYGO1* en la familia FCCTX cc28:

- Conlleva un cambio de aminoácido drástico, que afecta a uno de los dedos de zinc del dominio PHD.
- Reduce la afinidad de unión a histonas de la proteína, lo que disminuiría la expresión de genes diana de la vía Wnt, al contrario de lo esperado en un contexto de cáncer de colon.
- No segrega con todos los cánceres de la familia.
- No se puede confirmar su implicación en la susceptibilidad al cáncer de esta familia.

7. Bibliography

7.1 List of references

1. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000 Jan 7;100(1):57–70.
2. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011 Mar 4;144(5):646–74.
3. Richard Drake, A. Wayne Vogl, Adam W. M. Mitchell. Gray's Anatomy for Students. In: Gray's Anatomy for Students. p. 319–24.
4. Krogh D. Biology: a guide to the natural world. 5th ed. Boston: Benjamin Cummings; 2011. 725 p.
5. Azzouz LL, Sharma S. Physiology, Large Intestine. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2018 [cited 2019 Feb 23]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK507857/
6. Rao JN, Wang JY. Intestinal Architecture and Development. In: Regulation of Gastrointestinal Mucosal Growth. San Rafael: Morgan & Claypool Life Sciences; 2010.
7. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer. 2015 Mar 1;136(5):E359-386.
8. Gandomani HS, Yousefi SM, Aghajani M, Mohammadian-Hafshejani A, Tarazoj AA, Pouyesh V, et al. Colorectal cancer in the world: incidence, mortality and risk factors. Biomed Res Ther. 2017 Oct 14;4(10):1656.
9. 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.
10. Haggar F, Boushey R. Colorectal Cancer Epidemiology: Incidence, Mortality, Survival, and Risk Factors. Clin Colon Rectal Surg. 2009 Nov;22(04):191–7.
11. Jemal A, Center MM, DeSantis C, Ward EM. Global Patterns of Cancer Incidence and Mortality Rates and Trends. Cancer Epidemiol Biomarkers Prev. 2010 Aug 1;19(8):1893–907.
12. Favoriti P, Carbone G, Greco M, Pirozzi F, Pirozzi REM, Corcione F. Worldwide burden of colorectal cancer: a review. Updat Surg. 2016 Mar;68(1):7–11.
13. Bosetti C, Levi F, Rosato V, Bertuccio P, Lucchini F, Negri E, et al. Recent trends in colorectal cancer mortality in Europe. Int J Cancer. 2011 Jul 1;129(1):180–91.
14. Brenner H, Kloor M, Pox CP. Colorectal cancer. Lancet Lond Engl. 2014 Apr 26;383(9927):1490–502.
15. Padmanabhan S, Waly MI, Taranikanti V, Guizani N, Rahman MS, Ali A, et al. Modifiable and Non-modifiable Risk Factors for Colon and Rectal Cancer. In: Waly MI, Rahman MS, editors. Bioactive Components, Diet and Medical Treatment in Cancer Prevention [Internet]. Cham: Springer International Publishing; 2018 [cited 2019 Sep 14]. p. 121–30. Available from: http://link.springer.com/10.1007/978-3-319-75693-6_10
16. Song M, Garrett WS, Chan AT. Nutrients, foods, and colorectal cancer prevention. Gastroenterology. 2015 May;148(6):1244-1260.e16.
17. Zhao Z, Feng Q, Yin Z, Shuang J, Bai B, Yu P, et al. Red and processed meat consumption and colorectal cancer risk: a systematic review and meta-analysis. Oncotarget. 2017 Oct 10;8(47):83306–14.
18. Domingo JL, Nadal M. Carcinogenicity of consumption of red meat and processed meat: A review of scientific news since the IARC decision. Food Chem Toxicol Int J Publ Br Ind Biol Res Assoc. 2017 Jul;105:256–61.
19. Demeyer D, Mertens B, De Smet S, Ulens M. Mechanisms Linking Colorectal Cancer to the Consumption of (Processed) Red Meat: A Review. Crit Rev Food Sci Nutr. 2016 Dec 9;56(16):2747–66.
20. Kim E, Coelho D, Blachier F. Review of the association between meat consumption and risk of colorectal cancer. Nutr Res N Y N. 2013 Dec;33(12):983–94.

21. Jorquera Plaza F, Culebras Fernández JM. [Nutrition and colorectal cancer]. Nutr Hosp. 2000 Feb;15(1):3–12.

22. Song Y, Liu M, Yang FG, Cui LH, Lu XY, Chen C. Dietary fibre and the risk of colorectal cancer: a case-control study. Asian Pac J Cancer Prev APJCP. 2015;16(9):3747–52.

23. Yang W, Liu L, Masugi Y, Qian ZR, Nishihara R, Keum N, et al. Calcium intake and risk of colorectal cancer according to expression status of calcium-sensing receptor (CASR). Gut. 2018;67(8):1475–83.

24. Gao R, Gao Z, Huang L, Qin H. Gut microbiota and colorectal cancer. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol. 2017 May;36(5):757–69.

25. Akin H, Tözün N. Diet, microbiota, and colorectal cancer. J Clin Gastroenterol. 2014 Dec;48 Suppl 1:S67-69.

26. Bultman SJ. Interplay between diet, gut microbiota, epigenetic events, and colorectal cancer. Mol Nutr Food Res. 2017;61(1).

27. Brown JC, Winters-Stone K, Lee A, Schmitz KH. Cancer, physical activity, and exercise. Compr Physiol. 2012 Oct;2(4):2775–809.

28. Bardou M, Barkun AN, Martel M. Obesity and colorectal cancer. Gut. 2013 Jun;62(6):933–47.

29. Klarich DS, Brasser SM, Hong MY. Moderate Alcohol Consumption and Colorectal Cancer Risk. Alcohol Clin Exp Res. 2015 Aug;39(8):1280–91.

30. Richardson A, Hayes J, Frampton C, Potter J. Modifiable lifestyle factors that could reduce the incidence of colorectal cancer in New Zealand. N Z Med J. 2016 Dec 16;129(1447):13–20.

31. Dashti SG, Buchanan DD, Jayasekara H, Ait Ouakrim D, Clendenning M, Rosty C, et al. Alcohol Consumption and the Risk of Colorectal Cancer for Mismatch Repair Gene Mutation Carriers. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2017;26(3):366–75.

32. Fagunwa IO, Loughrey MB, Coleman HG. Alcohol, smoking and the risk of premalignant and malignant colorectal neoplasms. Best Pract Res Clin Gastroenterol. 2017 Oct;31(5):561–8.

33. Béjar L, Gili M, López J, Ramírez G, Cabanillas J, Cruz C. [Trends in colorectal cancer in Spain from 1951-2007 and alcohol and cigarette consumption]. Gastroenterol Hepatol. 2010 Feb;33(2):71–9.

34. Botteri E, Iodice S, Bagnardi V, Raimondi S, Lowenfels AB, Maisonneuve P. Smoking and colorectal cancer: a meta-analysis. JAMA. 2008 Dec 17;300(23):2765–78.

35. Schmit SL, Rennert HS, Rennert G, Gruber SB. Coffee Consumption and the Risk of Colorectal Cancer. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2016;25(4):634–9.

36. Singh Ranger G. The role of aspirin in colorectal cancer chemoprevention. Crit Rev Oncol Hematol. 2016 Aug;104:87–90.

37. Drew DA, Cao Y, Chan AT. Aspirin and colorectal cancer: the promise of precision chemoprevention. Nat Rev Cancer. 2016 Mar;16(3):173–86.

38. Tougeron D, Sha D, Manthravadi S, Sinicrope FA. Aspirin and colorectal cancer: back to the future. Clin Cancer Res Off J Am Assoc Cancer Res. 2014 Mar 1;20(5):1087–94.

39. Burn J, Sheth H. The role of aspirin in preventing colorectal cancer. Br Med Bull. 2016;119(1):17–24.

40. Frampton MJE, Law P, Litchfield K, Morris EJ, Kerr D, Turnbull C, et al. Implications of polygenic risk for personalised colorectal cancer screening. Ann Oncol Off J Eur Soc Med Oncol. 2016 Mar;27(3):429–34.

41. Peters U, Bien S, Zubair N. Genetic architecture of colorectal cancer. Gut. 2015 Oct;64(10):1623–36.

42. Lowery JT, Ahnen DJ, Schroy PC, Hampel H, Baxter N, Boland CR, et al. Understanding the contribution of family history to colorectal cancer risk and its clinical implications: A state-of-the-science review. Cancer. 2016 01;122(17):2633–45.

43. Butterworth AS, Higgins JPT, Pharoah P. Relative and absolute risk of colorectal cancer for individuals with a family history: a meta-analysis. Eur J Cancer Oxf Engl 1990. 2006 Jan;42(2):216–27.

44. Henrikson NB, Webber EM, Goddard KA, Scrol A, Piper M, Williams MS, et al. Family history and the natural history of colorectal cancer: systematic review. Genet Med Off J Am Coll Med Genet. 2015 Sep;17(9):702–12.

45. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol. 2001 Oct;96(10):2992–3003.

46. Solomon BL, Whitman T, Wood ME. Contribution of extended family history in assessment of risk for breast and colon cancer. BMC Fam Pract. 2016 01;17(1):126.

47. Colucci PM, Yale SH, Rall CJ. Colorectal polyps. Clin Med Res. 2003 Jul;1(3):261–2.

48. Kim ER, Chang DK. Colorectal cancer in inflammatory bowel disease: the risk, pathogenesis, prevention and diagnosis. World J Gastroenterol. 2014 Aug 7;20(29):9872–81.

49. Stidham RW, Higgins PDR. Colorectal Cancer in Inflammatory Bowel Disease. Clin Colon Rectal Surg. 2018 May;31(3):168–78.

50. Triantafillidis JK, Nasioulas G, Kosmidis PA. Colorectal cancer and inflammatory bowel disease: epidemiology, risk factors, mechanisms of carcinogenesis and prevention strategies. Anticancer Res. 2009 Jul;29(7):2727–37.

51. Guraya SY. Association of type 2 diabetes mellitus and the risk of colorectal cancer: A meta-analysis and systematic review. World J Gastroenterol. 2015 May 21;21(19):6026–31.

52. Piper MS, Saad RJ. Diabetes Mellitus and the Colon. Curr Treat Options Gastroenterol. 2017 Dec;15(4):460–74.

53. Kim S-E, Paik HY, Yoon H, Lee JE, Kim N, Sung M-K. Sex- and gender-specific disparities in colorectal cancer risk. World J Gastroenterol. 2015 May 7;21(17):5167–75.

54. Fleming M, Ravula S, Tatishchev SF, Wang HL. Colorectal carcinoma: Pathologic aspects. J Gastrointest Oncol. 2012 Sep;3(3):153–73.

55. Obuch JC, Ahnen DJ. Colorectal Cancer. Gastroenterol Clin North Am. 2016 Sep;45(3):459–76.

56. Bathe OF, Farshidfar F. From genotype to functional phenotype: unraveling the metabolomic features of colorectal cancer. Genes. 2014 Jul 22;5(3):536–60.

57. Tariq K, Ghias K. Colorectal cancer carcinogenesis: a review of mechanisms. Cancer Biol Med. 2016 Mar;13(1):120–35.

58. Rodriguez-Bigas MA, Lin EH, Crane CH. Genetic Pathways in Colorectal Cancer. In: Holland-Frei Cancer Medicine [Internet]. 6th ed. Hamilton (ON), Canada: BC Decker; 2003. Available from: https://www.ncbi.nlm.nih.gov/books/NBK12839/

59. Arends MJ. Pathways of colorectal carcinogenesis. Appl Immunohistochem Mol Morphol AIMM. 2013 Mar;21(2):97–102.

60. Bosman F, Yan P. Molecular pathology of colorectal cancer. Pol J Pathol Off J Pol Soc Pathol. 2014 Dec;65(4):257–66.

61. Galiatsatos P, Foulkes WD. Familial adenomatous polyposis. Am J Gastroenterol. 2006 Feb;101(2):385–98.

62. Armaghany T, Wilson JD, Chu Q, Mills G. Genetic alterations in colorectal cancer. Gastrointest Cancer Res GCR. 2012 Jan;5(1):19–27.

63. Al-Sohaily S, Biankin A, Leong R, Kohonen-Corish M, Warusavitarne J. Molecular pathways in colorectal cancer: Pathways of colorectal carcinogenesis. J Gastroenterol Hepatol. 2012 Sep;27(9):1423–31.

64. Rengifo-Cam W, Shepherd HM, Jasperson KW, Samadder NJ, Samowitz W, Tripp SR, et al. Colon Pathology Characteristics in Li-Fraumeni Syndrome. Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc. 2018;16(1):140–1.

65. Nazemalhosseini Mojarad E, Kuppen PJ, Aghdaei HA, Zali MR. The CpG island methylator phenotype (CIMP) in colorectal cancer. Gastroenterol Hepatol Bed Bench. 2013;6(3):120–8.

66. Abdel-Rahman WM. Genomic instability and carcinogenesis: an update. Curr Genomics. 2008 Dec;9(8):535–41.

67. Arvelo F, Sojo F, Cotte C. Biology of colorectal cancer. Ecancermedicalscience. 2015;9:520.

68. Edge SB, Compton CC. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. Ann Surg Oncol. 2010 Jun;17(6):1471–4.

69. Dukes CE. The classification of cancer of the rectum. J Pathol Bacteriol. 1932;35(3):323–32.

70. Kyriakos M. The President's cancer, the Dukes classification, and confusion. Arch Pathol Lab Med. 1985 Dec;109(12):1063–6.

71. Astler VB, Coller FA. The prognostic significance of direct extension of carcinoma of the colon and rectum. Ann Surg. 1954 Jun;139(6):846–52.

72. Yurgelun MB, Kulke MH, Fuchs CS, Allen BA, Uno H, Hornick JL, et al. Cancer Susceptibility Gene Mutations in Individuals With Colorectal Cancer. J Clin Oncol. 2017 Apr;35(10):1086–95.

73. Pearlman R, Frankel WL, Swanson B, Zhao W, Yilmaz A, Miller K, et al. Prevalence and Spectrum of Germline Cancer Susceptibility Gene Mutations Among Patients With Early-Onset Colorectal Cancer. JAMA Oncol. 2017 Apr 1;3(4):464.

74. Stoffel EM, Koeppe E, Everett J, Ulintz P, Kiel M, Osborne J, et al. Germline Genetic Features of Young Individuals With Colorectal Cancer. Gastroenterology. 2018 Mar;154(4):897-905.e1.

75. Valle L, Vilar E, Tavtigian SV, Stoffel EM. Genetic predisposition to colorectal cancer: syndromes, genes, classification of genetic variants and implications for precision medicine. J Pathol. 2018 Dec 25;

76. Shussman N, Wexner SD. Colorectal polyps and polyposis syndromes. Gastroenterol Rep. 2014 Feb;2(1):1–15.

77. Jasperson KW, Tuohy TM, Neklason DW, Burt RW. Hereditary and familial colon cancer. Gastroenterology. 2010 Jun;138(6):2044–58.

78. Valle L. Genetic predisposition to colorectal cancer: where we stand and future perspectives. World J Gastroenterol. 2014 Aug 7;20(29):9828–49.

79. Jansen AML, Crobach S, Geurts-Giele WRR, van den Akker BEWM, Garcia MV, Ruano D, et al. Distinct Patterns of Somatic Mosaicism in the APC Gene in Neoplasms From Patients With Unexplained Adenomatous Polyposis. Gastroenterology. 2017;152(3):546-549.e3.

80. Spier I, Drichel D, Kerick M, Kirfel J, Horpaopan S, Laner A, et al. Low-level APC mutational mosaicism is the underlying cause in a substantial fraction of unexplained colorectal adenomatous polyposis cases. J Med Genet. 2016 Mar;53(3):172–9.

81. Balaguer F, Castellví-Bel S, Castells A, Andreu M, Muñoz J, Gisbert JP, et al. Identification of MYH mutation carriers in colorectal cancer: a multicenter, case-control, population-based study. Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc. 2007 Mar;5(3):379–87.

82. Toboeva MK, Shelygin YA, Frolov SA, Kuzminov MA, Tsukanov AS. MutYH-associated polyposis. Ter Arkh. 2019 Mar 18;91(2):97–100.

83. The CORGI Consortium, The WGS500 Consortium, Palles C, Cazier J-B, Howarth KM, Domingo E, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat Genet. 2013 Feb;45(2):136–44.

84. Church JM. Polymerase proofreading-associated polyposis: a new, dominantly inherited syndrome of hereditary colorectal cancer predisposition. Dis Colon Rectum. 2014 Mar;57(3):396–7.

85. Terradas M, Munoz-Torres PM, Belhadj S, Aiza G, Navarro M, Brunet J, et al. Contribution to colonic polyposis of recently proposed predisposing genes and assessment of the prevalence of NTHL1- and MSH3-associated polyposes. Hum Mutat. 2019 Jun 27;

86. Adam R, Spier I, Zhao B, Kloth M, Marquez J, Hinrichsen I, et al. Exome Sequencing Identifies Biallelic MSH3 Germline Mutations as a Recessive Subtype of Colorectal Adenomatous Polyposis. Am J Hum Genet. 2016 Aug 4;99(2):337–51.

87. Olkinuora A, Nieminen TT, Mårtensson E, Rohlin A, Ristimäki A, Koskenvuo L, et al. Biallelic germline nonsense variant of MLH3 underlies polyposis predisposition. Genet Med Off J Am Coll Med Genet. 2019 Aug;21(8):1868–73.

88. Daniell J, Plazzer J-P, Perera A, Macrae F. An exploration of genotype-phenotype link between Peutz-Jeghers syndrome and STK11: a review. Fam Cancer. 2018;17(3):421–7.

89. Cichy W, Klincewicz B, Plawski A. Juvenile polyposis syndrome. Arch Med Sci AMS. 2014 Jun 29;10(3):570–7.

90. Gammon A, Jasperson K, Kohlmann W, Burt RW. Hamartomatous polyposis syndromes. Best Pract Res Clin Gastroenterol. 2009 Apr;23(2):219–31.

91. Pilarski R, Burt R, Kohlman W, Pho L, Shannon KM, Swisher E. Cowden syndrome and the PTEN hamartoma tumor syndrome: systematic review and revised diagnostic criteria. J Natl Cancer Inst. 2013 Nov 6;105(21):1607–16.

92. Pilarski R, Stephens JA, Noss R, Fisher JL, Prior TW. Predicting PTEN mutations: an evaluation of Cowden syndrome and Bannayan-Riley-Ruvalcaba syndrome clinical features. J Med Genet. 2011 Aug;48(8):505–12.

93. Zhou X, Hampel H, Thiele H, Gorlin RJ, Hennekam RC, Parisi M, et al. Association of germline mutation in the PTEN tumour suppressor gene and Proteus and Proteus-like syndromes. Lancet Lond Engl. 2001 Jul 21;358(9277):210–1.

94. Gala MK, Mizukami Y, Le LP, Moriichi K, Austin T, Yamamoto M, et al. Germline mutations in oncogene-induced senescence pathways are associated with multiple sessile serrated adenomas. Gastroenterology. 2014 Feb;146(2):520–9.

95. 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 Sep;66(9):1645–56.

96. Buchanan DD, Clendenning M, Zhuoer L, Stewart JR, Joseland S, Woodall S, et al. Lack of evidence for germline *RNF43* mutations in patients with serrated polyposis syndrome from a large multinational study. Gut. 2017 Jun;66(6):1170–2.

97. Quintana I, Mejías-Luque R, Terradas M, Navarro M, Piñol V, Mur P, et al. Evidence suggests that germline *RNF43* mutations are a rare cause of serrated polyposis. Gut. 2018 Dec;67(12):2230–2.

98. McKenna DB, Van Den Akker J, Zhou AY, Ryan L, Leon A, O'Connor R, et al. Identification of a novel GREM1 duplication in a patient with multiple colon polyps. Fam Cancer. 2019;18(1):63–6.

99. Zetner DB, Bisgaard ML. Familial Colorectal Cancer Type X. Curr Genomics [Internet]. 2017 Jul 26 [cited 2017 Nov 17];18(4). Available from: http://www.eurekaselect.com/150710/article

100. Hsieh P, Yamane K. DNA mismatch repair: molecular mechanism, cancer, and ageing. Mech Ageing Dev. 2008 Aug;129(7–8):391–407.

101. Dominguez-Valentin M, Therkildsen C, Da Silva S, Nilbert M. Familial colorectal cancer type X: genetic profiles and phenotypic features. Mod Pathol Off J U S Can Acad Pathol Inc. 2015 Jan;28(1):30–6.

102. Klarskov L, Holck S, Bernstein I, Nilbert M. Hereditary colorectal cancer diagnostics: morphological features of familial colorectal cancer type X versus Lynch syndrome. J Clin Pathol. 2012 Apr;65(4):352–6.

103. Walkowska J, Kallemose T, Jönsson G, Jönsson M, Andersen O, Andersen MH, et al. Immunoprofiles of colorectal cancer from Lynch syndrome. Oncoimmunology. 2019;8(1):e1515612.

104. DA Silva FC, Wernhoff P, Dominguez-Barrera C, Dominguez-Valentin M. Update on Hereditary Colorectal Cancer. Anticancer Res. 2016;36(9):4399–405.

105. Nejadtaghi M, Jafari H, Farrokhi E, Samani KG. Familial Colorectal Cancer Type X (FCCTX) and the correlation with various genes-A systematic review. Curr Probl Cancer. 2017 Dec;41(6):388–97.

106. 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 Apr 27;293(16):1979–85.

107. Woods MO, Younghusband HB, Parfrey PS, Gallinger S, McLaughlin J, Dicks E, et al. The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. Gut. 2010 Oct 1;59(10):1369–77.

108. Benatti P, Roncucci L, Ganazzi D, Percesepe A, Di Gregorio C, Pedroni M, et al. Clinical and biologic heterogeneity of hereditary nonpolyposis colorectal cancer. Int J Cancer. 2001 Sep 20;95(5):323–8.

109. Mueller-Koch Y, Vogelsang H, Kopp R, Lohse P, Keller G, Aust D, et al. Hereditary non-polyposis colorectal cancer: clinical and molecular evidence for a new entity of hereditary colorectal cancer. Gut. 2005 Dec;54(12):1733–40.

110. Valle L, Perea J, Carbonell P, Fernandez V, Dotor AM, Benitez J, et al. Clinicopathologic and pedigree differences in amsterdam I-positive hereditary nonpolyposis colorectal cancer families according to tumor microsatellite instability status. J Clin Oncol Off J Am Soc Clin Oncol. 2007 Mar 1;25(7):781–6.

111. Llor X, Pons E, Xicola RM, Castells A, Alenda C, Piñol V, et al. Differential features of colorectal cancers fulfilling Amsterdam criteria without involvement of the mutator pathway. Clin Cancer Res Off J Am Assoc Cancer Res. 2005 Oct 15;11(20):7304–10.

112. Shiovitz S, Copeland WK, Passarelli MN, Burnett-Hartman AN, Grady WM, Potter JD, et al. Characterisation of familial colorectal cancer Type X, Lynch syndrome, and non-familial colorectal cancer. Br J Cancer. 2014 Jul 29;111(3):598–602.

113. Dominguez-Valentin M, Therkildsen C, Veerla S, Jönsson M, Bernstein I, Borg A, et al. Distinct gene expression signatures in lynch syndrome and familial colorectal cancer type x. PloS One. 2013;8(8):e71755.

114. Dove-Edwin I, de Jong AE, Adams J, Mesher D, Lipton L, Sasieni P, et al. Prospective results of surveillance colonoscopy in dominant familial colorectal cancer with and without Lynch syndrome. Gastroenterology. 2006 Jun;130(7):1995–2000.

115. Zeinalian M, Hadian M, Hashemzadeh-Chaleshtori M, Salehi R, Emami MH. Familial Colorectal Cancer Type X in Central Iran: A New Clinicopathologic Description. Int J Hematol-Oncol Stem Cell Res. 2017 Jul 1;11(3):240–5.

116. Garre P, Martín L, Bando I, Tosar A, Llovet P, Sanz J, et al. Cancer risk and overall survival in mismatch repair proficient hereditary non-polyposis colorectal cancer, Lynch syndrome and sporadic colorectal cancer. Fam Cancer. 2014 Mar;13(1):109–19.

117. Schiemann U, Müller-Koch Y, Gross M, Daum J, Lohse P, Baretton G, et al. Extended microsatellite analysis in microsatellite stable, MSH2 and MLH1 mutation-negative HNPCC patients: genetic reclassification and correlation with clinical features. Digestion. 2004;69(3):166–76.

118. Palmirotta R, Matera S, Curia MC, Aceto G, el Zhobi B, Verginelli F, et al. Correlations between phenotype and microsatellite instability in HNPCC: implications for genetic testing. Fam Cancer. 2004;3(2):117–21.

119. Yamaguchi T, Furukawa Y, Nakamura Y, Matsubara N, Ishikawa H, Arai M, et al. Comparison of clinical features between suspected familial colorectal cancer type X and Lynch syndrome in Japanese patients with colorectal cancer: a cross-sectional study conducted by the Japanese Society for Cancer of the Colon and Rectum. Jpn J Clin Oncol. 2015 Feb;45(2):153–9.

120. Sánchez-de-Abajo A, de la Hoya M, van Puijenbroek M, Tosar A, López-Asenjo JA, Díaz-Rubio E, et al. Molecular analysis of colorectal cancer tumors from patients with mismatch repair proficient

hereditary nonpolyposis colorectal cancer suggests novel carcinogenic pathways. Clin Cancer Res Off J Am Assoc Cancer Res. 2007 Oct 1;13(19):5729–35.

121. Francisco I, Albuquerque C, Lage P, Belo H, Vitoriano I, Filipe B, et al. Familial colorectal cancer type X syndrome: two distinct molecular entities? Fam Cancer. 2011 Dec;10(4):623–31.

122. Choi Y-H, Lakhal-Chaieb L, Kröl A, Yu B, Buchanan D, Ahnen D, et al. Risks of Colorectal Cancer and Cancer-Related Mortality in Familial Colorectal Cancer Type X and Lynch Syndrome Families. J Natl Cancer Inst. 2018 Oct 30;

123. Chen C-H, Sheng Jiang S, Hsieh L-L, Tang R, Hsiung CA, Tsai H-J, et al. DNA Methylation Identifies Loci Distinguishing Hereditary Nonpolyposis Colorectal Cancer Without Germ-Line MLH1/MSH2 Mutation from Sporadic Colorectal Cancer. Clin Transl Gastroenterol. 2016 Dec 15;7(12):e208.

124. Pavicic W, Joensuu EI, Nieminen T, Peltomäki P. LINE-1 hypomethylation in familial and sporadic cancer. J Mol Med Berl Ger. 2012 Jul;90(7):827–35.

125. Goel A, Xicola RM, Nguyen T-P, Doyle BJ, Sohn VR, Bandipalliam P, et al. Aberrant DNA methylation in hereditary nonpolyposis colorectal cancer without mismatch repair deficiency. Gastroenterology. 2010 May;138(5):1854–62.

126. Chen W, Ding J, Jiang L, Liu Z, Zhou X, Shi D. DNA copy number profiling in microsatellite-stable and microsatellite-unstable hereditary non-polyposis colorectal cancers by targeted CNV array. Funct Integr Genomics. 2017 Jan;17(1):85–96.

127. Chen W, Yuan L, Cai Y, Chen X, Chi Y, Wei P, et al. Identification of chromosomal copy number variations and novel candidate loci in hereditary nonpolyposis colorectal cancer with mismatch repair proficiency. Genomics. 2013 Jul;102(1):27–34.

128. Therkildsen C, Jönsson G, Dominguez-Valentin M, Nissen A, Rambech E, Halvarsson B, et al. Gain of chromosomal region 20q and loss of 18 discriminates between Lynch syndrome and familial colorectal cancer. Eur J Cancer Oxf Engl 1990. 2013 Apr;49(6):1226–35.

129. Abdel-Rahman WM, Ollikainen M, Kariola R, Järvinen HJ, Mecklin J-P, Nyström-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 Feb 24;24(9):1542–51.

130. Engel C, Rahner N, Schulmann K, Holinski-Feder E, Goecke TO, Schackert HK, et al. Efficacy of annual colonoscopic surveillance in individuals with hereditary nonpolyposis colorectal cancer. Clin Gastroenterol Hepatol Off Clin Pract J Am Gastroenterol Assoc. 2010 Feb;8(2):174–82.

131. Lindberg LJ, Ladelund S, Frederiksen BL, Smith-Hansen L, Bernstein I. Outcome of 24 years national surveillance in different hereditary colorectal cancer subgroups leading to more individualised surveillance. J Med Genet. 2017;54(5):297–304.

132. Hatfield E, Green JS, Woods MO, Warden G, Parfrey PS. Impact of colonoscopic screening in Familial Colorectal Cancer Type X. Mol Genet Genomic Med. 2018;6(6):1021–30.

133. Mesher D, Dove-Edwin I, Sasieni P, Vasen H, Bernstein I, Royer-Pokora B, et al. A pooled analysis of the outcome of prospective colonoscopic surveillance for familial colorectal cancer. Int J Cancer. 2014 Feb 15;134(4):939–47.

134. Perea J, Justo I, Alvaro E, Lomas M, Tasende JD, Marín JC, et al. Surgical management of hereditary colorectal cancer: surgery based on molecular analysis and family history. Rev Espanola Enfermedades Dig Organo Of Soc Espanola Patol Dig. 2009 Aug;101(8):536–40.

135. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. Genomics Inform. 2012 Jun;10(2):117–22.

136. Valle L. Recent Discoveries in the Genetics of Familial Colorectal Cancer and Polyposis. Clin Gastroenterol Hepatol. 2017 Jun;15(6):809–19.

137. Bellido F, Sowada N, Mur P, Lázaro C, Pons T, Valdés-Mas R, et al. Association Between Germline Mutations in BRF1 , a Subunit of the RNA Polymerase III Transcription Complex, and Hereditary Colorectal Cancer. Gastroenterology. 2018 Jan;154(1):181-194.e20.

138. Franch-Expósito S, Esteban-Jurado C, Garre P, Quintanilla I, Duran-Sanchon S, Díaz-Gay M, et al. Rare germline copy number variants in colorectal cancer predisposition characterized by exome sequencing analysis. J Genet Genomics. 2018 Jan;45(1):41–5.

139. Martín-Morales L, Feldman M, Vershinin Z, Garre P, Caldés T, Levy D. SETD6 dominant negative mutation in familial colorectal cancer type X. Hum Mol Genet. 2017 Nov 15;26(22):4481–93.

140. Nieminen TT, O'Donohue M-F, Wu Y, Lohi H, Scherer SW, Paterson AD, et al. Germline Mutation of RPS20, Encoding a Ribosomal Protein, Causes Predisposition to Hereditary Nonpolyposis Colorectal Carcinoma Without DNA Mismatch Repair Deficiency. Gastroenterology. 2014 Sep;147(3):595-598.e5.

141. Broderick P, Dobbins SE, Chubb D, Kinnersley B, Dunlop MG, Tomlinson I, et al. Validation of Recently Proposed Colorectal Cancer Susceptibility Gene Variants in an Analysis of Families and Patients—a Systematic Review. Gastroenterology. 2017 Jan;152(1):75-77.e4.

142. Garre P, Martín 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 Jun;87(6):582–7.

143. Seguí N, Mina LB, Lázaro C, Sanz-Pamplona R, Pons T, Navarro M, et al. Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. Gastroenterology. 2015 Sep;149(3):563–6.

144. Evans DR, Venkitachalam S, Revoredo L, Dohey AT, Clarke E, Pennell JJ, et al. Evidence for GALNT12 as a moderate penetrance gene for colorectal cancer. Hum Mutat. 2018 Aug;39(8):1092–101.

145. Evans DR, Green JS, Woods MO. Screening of BMPR1a for pathogenic mutations in familial colorectal cancer type X families from Newfoundland. Fam Cancer. 2018;17(2):205–8.

146. Schulz E, Klampfl P, Holzapfel S, Janecke AR, Ulz P, Renner W, et al. Germline variants in the SEMA4A gene predispose to familial colorectal cancer type X. Nat Commun. 2014 Oct 13;5:5191.

147. Sánchez-Tomé 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 Jun;50(6):657–66.

148. Xicola RM, Bontu S, Doyle BJ, Rawson J, Garre P, Lee E, et al. Association of a let-7 miRNA binding region of *TGFBR1* with hereditary mismatch repair proficient colorectal cancer (MSS HNPCC). Carcinogenesis. 2016 Aug;37(8):751–8.

149. Garre P, Briceno V, Xicola RM, Doyle BJ, de la Hoya M, Sanz J, et al. Analysis of the Oxidative Damage Repair Genes NUDT1, OGG1, and MUTYH in Patients from Mismatch Repair Proficient HNPCC Families (MSS-HNPCC). Clin Cancer Res. 2011 Apr 1;17(7):1701–12.

150. Esteban-Jurado C, Franch-Expósito S, Muñoz J, Ocaña T, Carballal S, López-Cerón M, et al. The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer. Eur J Hum Genet EJHG. 2016;24(10):1501–5.

151. Smith CG, Naven M, Harris R, Colley J, West H, Li N, et al. Exome Resequencing Identifies Potential Tumor-Suppressor Genes that Predispose to Colorectal Cancer. Hum Mutat. 2013 Jul;34(7):1026–34.

152. Seguí N, Pineda M, Navarro M, Lázaro C, Brunet J, Infante M, et al. GALNT12 is not a major contributor of familial colorectal cancer type X. Hum Mutat. 2014 Jan;35(1):50–2.

153. Jamuar SS, Tan E-C. Clinical application of next-generation sequencing for Mendelian diseases. Hum Genomics [Internet]. 2015 Dec [cited 2019 Feb 25];9(1). Available from: https://humgenomics.biomedcentral.com/articles/10.1186/s40246-015-0031-5

154. Ku C-S, Cooper DN, Wu M, Roukos DH, Pawitan Y, Soong R, et al. Gene discovery in familial cancer syndromes by exome sequencing: prospects for the elucidation of familial colorectal cancer type X. Mod Pathol Off J U S Can Acad Pathol Inc. 2012 Aug;25(8):1055–68.

155. Muzzey D, Evans EA, Lieber C. Understanding the Basics of NGS: From Mechanism to Variant Calling. Curr Genet Med Rep. 2015;3(4):158–65.

156. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of Next-Generation Sequencing Systems. J Biomed Biotechnol. 2012;2012:1–11.

157. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: tone down the bias. Exp Cell Res. 2014 Mar 10;322(1):12–20.

158. Lanouette S, Mongeon V, Figeys D, Couture J-F. The functional diversity of protein lysine methylation. Mol Syst Biol. 2014 Apr 8;10:724.

159. Zhang X, Huang Y, Shi X. Emerging roles of lysine methylation on non-histone proteins. Cell Mol Life Sci CMLS. 2015 Nov;72(22):4257–72.

160. Hamamoto R, Saloura V, Nakamura Y. Critical roles of non-histone protein lysine methylation in human tumorigenesis. Nat Rev Cancer. 2015 Feb;15(2):110–24.

161. Shi Y, Whetstine JR. Dynamic regulation of histone lysine methylation by demethylases. Mol Cell. 2007 Jan 12;25(1):1–14.

162. Qian C, Zhou M-M. SET domain protein lysine methyltransferases: Structure, specificity and catalysis. Cell Mol Life Sci CMLS. 2006 Dec;63(23):2755–63.

163. Levy D, Kuo AJ, Chang Y, Schaefer U, Kitson C, Cheung P, et al. Lysine methylation of the NF-κB subunit RelA by SETD6 couples activity of the histone methyltransferase GLP at chromatin to tonic repression of NF-κB signaling. Nat Immunol. 2011 Jan;12(1):29–36.

164. Patel M, Horgan PG, McMillan DC, Edwards J. NF-κB pathways in the development and progression of colorectal cancer. Transl Res J Lab Clin Med. 2018;197:43–56.

165. Gilmore TD. Introduction to NF-kappaB: players, pathways, perspectives. Oncogene. 2006 Oct 30;25(51):6680–4.

166. Mukherjee N, Cardenas E, Bedolla R, Ghosh R. SETD6 regulates NF-κB signaling in urothelial cell survival: Implications for bladder cancer. Oncotarget. 2017 Feb 28;8(9):15114–25.

167. Vershinin Z, Feldman M, Chen A, Levy D. PAK4 Methylation by SETD6 Promotes the Activation of the Wnt/β-Catenin Pathway. J Biol Chem. 2016 Mar 25;291(13):6786–95.

168. Chen A, Feldman M, Vershinin Z, Levy D. SETD6 is a negative regulator of oxidative stress response. Biochim Biophys Acta. 2016 Feb;1859(2):420–7.

169. Binda O, Sevilla A, LeRoy G, Lemischka IR, Garcia BA, Richard S. SETD6 monomethylates H2AZ on lysine 7 and is required for the maintenance of embryonic stem cell self-renewal. Epigenetics. 2013 Feb;8(2):177–83.

170. O'Neill DJ, Williamson SC, Alkharaif D, Monteiro ICM, Goudreault M, Gaughan L, et al. SETD6 controls the expression of estrogen-responsive genes and proliferation of breast carcinoma cells. Epigenetics. 2014 Jul;9(7):942–50.

171. Feldman M, Vershinin Z, Goliand I, Elia N, Levy D. The methyltransferase SETD6 regulates Mitotic progression through PLK1 methylation. Proc Natl Acad Sci U S A. 2019 Jan 22;116(4):1235–40.

172. Singh V, Ram M, Kumar R, Prasad R, Roy BK, Singh KK. Phosphorylation: Implications in Cancer. Protein J. 2017;36(1):1–6.

173. Hunter T. Tyrosine phosphorylation: thirty years and counting. Curr Opin Cell Biol. 2009 Apr;21(2):140–6.

174. Julien SG, Dubé N, Hardy S, Tremblay ML. Inside the human cancer tyrosine phosphatome. Nat Rev Cancer. 2011 Jan;11(1):35–49.

175. Zhao S, Sedwick D, Wang Z. Genetic alterations of protein tyrosine phosphatases in human cancers. Oncogene. 2015 Jul 23;34(30):3885–94.

176. Wang Z, Shen D, Parsons DW, Bardelli A, Sager J, Szabo S, et al. Mutational analysis of the tyrosine phosphatome in colorectal cancers. Science. 2004 May 21;304(5674):1164–6.

177. Laczmanska I, Karpinski P, Bebenek M, Sedziak T, Ramsey D, Szmida E, et al. Protein tyrosine phosphatase receptor-like genes are frequently hypermethylated in sporadic colorectal cancer. J Hum Genet. 2013 Jan;58(1):11–5.

178. Peyser ND, Freilino M, Wang L, Zeng Y, Li H, Johnson DE, et al. Frequent promoter hypermethylation of PTPRT increases STAT3 activation and sensitivity to STAT3 inhibition in head and neck cancer. Oncogene. 2016 Mar 3;35(9):1163–9.

179. Becka S, Zhang P, Craig SEL, Lodowski DT, Wang Z, Brady-Kalnay SM. Characterization of the adhesive properties of the type IIb subfamily receptor protein tyrosine phosphatases. Cell Commun Adhes. 2010 Apr;17(2):34–47.

180. Besco JA, Hooft van Huijsduijnen R, Frostholm A, Rotter A. Intracellular substrates of brain-enriched receptor protein tyrosine phosphatase rho (RPTPrho/PTPRT). Brain Res. 2006 Oct 20;1116(1):50–7.

181. Tonks NK. Protein tyrosine phosphatases: from genes, to function, to disease. Nat Rev Mol Cell Biol. 2006 Nov;7(11):833–46.

182. Zhang X, Guo A, Yu J, Possemato A, Chen Y, Zheng W, et al. Identification of STAT3 as a substrate of receptor protein tyrosine phosphatase T. Proc Natl Acad Sci U S A. 2007 Mar 6;104(10):4060–4.

183. Zhao Y, Zhang X, Guda K, Lawrence E, Sun Q, Watanabe T, et al. Identification and functional characterization of paxillin as a target of protein tyrosine phosphatase receptor T. Proc Natl Acad Sci U S A. 2010 Feb 9;107(6):2592–7.

184. Darnell JE. Validating Stat3 in cancer therapy. Nat Med. 2005 Jun;11(6):595–6.

185. Zhang P, Zhao Y, Zhu X, Sedwick D, Zhang X, Wang Z. Cross-talk between phospho-STAT3 and PLCγ1 plays a critical role in colorectal tumorigenesis. Mol Cancer Res MCR. 2011 Oct;9(10):1418–28.

186. Lui VWY, Peyser ND, Ng PK-S, Hritz J, Zeng Y, Lu Y, et al. Frequent mutation of receptor protein tyrosine phosphatases provides a mechanism for STAT3 hyperactivation in head and neck cancer. Proc Natl Acad Sci U S A. 2014 Jan 21;111(3):1114–9.

187. Park A-R, Oh D, Lim S-H, Choi J, Moon J, Yu D-Y, et al. Regulation of dendritic arborization by BCR Rac1 GTPase-activating protein, a substrate of PTPRT. J Cell Sci. 2012 Oct 1;125(Pt 19):4518–31.

188. Lim S-H, Moon J, Lee M, Lee J-R. PTPRT regulates the interaction of Syntaxin-binding protein 1 with Syntaxin 1 through dephosphorylation of specific tyrosine residue. Biochem Biophys Res Commun. 2013 Sep 13;439(1):40–6.

189. Komiya Y, Habas R. Wnt signal transduction pathways. Organogenesis. 2008 Apr;4(2):68–75.

190. Katoh M, Katoh M. WNT signaling pathway and stem cell signaling network. Clin Cancer Res Off J Am Assoc Cancer Res. 2007 Jul 15;13(14):4042–5.

191. Jessen S, Gu B, Dai X. Pygopus and the Wnt signaling pathway: a diverse set of connections. BioEssays News Rev Mol Cell Dev Biol. 2008 May;30(5):448–56.

192. Hildebrandt MAT, Reyes ME, Lin M, He Y, Nguyen SV, Hawk ET, et al. Germline Genetic Variants in the Wnt/β-Catenin Pathway as Predictors of Colorectal Cancer Risk. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2016 Mar;25(3):540–6.

193. Yu J, Virshup DM. Updating the Wnt pathways. Biosci Rep. 2014 Oct 17;34(5):593–607.

194. Fiedler M, Sánchez-Barrena MJ, Nekrasov M, Mieszczanek J, Rybin V, Müller J, et al. Decoding of methylated histone H3 tail by the Pygo-BCL9 Wnt signaling complex. Mol Cell. 2008 May 23;30(4):507–18.

195. Schwab KR, Patterson LT, Hartman HA, Song N, Lang RA, Lin X, et al. Pygo1 and Pygo2 roles in Wnt signaling in mammalian kidney development. BMC Biol. 2007 Apr 10;5:15.

196. Miller TCR, Rutherford TJ, Johnson CM, Fiedler M, Bienz M. Allosteric remodelling of the histone H3 binding pocket in the Pygo2 PHD finger triggered by its binding to the B9L/BCL9 co-factor. J Mol Biol. 2010 Sep 3;401(5):969–84.

197. Miller TCR, Rutherford TJ, Birchall K, Chugh J, Fiedler M, Bienz M. Competitive Binding of a Benzimidazole to the Histone-Binding Pocket of the Pygo PHD Finger. ACS Chem Biol. 2014;9:2864–74.

198. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma Oxf Engl. 2009 Aug 15;25(16):2078–9.

199. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinforma Oxf Engl. 2011 Nov 1;27(21):2987–93.

200. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinforma Oxf Engl. 2009 Sep 1;25(17):2283–5.

201. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012 Mar 1;22(3):568–76.

202. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep 1;20(9):1297–303.

203. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003 Jun;21(6):577–81.

204. Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, et al. Ensembl 2016. Nucleic Acids Res. 2016 Jan 4;44(D1):D710–6.

205. The 1000 Genomes Project Consortium, Gibbs RA, Boerwinkle E, Doddapaneni H, Han Y, Korchina V, et al. A global reference for human genetic variation. Nature. 2015 Oct;526(7571):68–74.

206. Exome Variant Server [Internet]. NHLBI GO Exome Sequencing Project (ESP), Seattle, WA; Available from: http://evs.gs.washington.edu/EVS/

207. Exome Aggregation Consortium, Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 Aug;536(7616):285–91.

208. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009 Jul;4(7):1073–81.

209. 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 Apr;7(4):248–9.

210. Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. Nat Methods. 2014 Apr;11(4):361–2.

211. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the Functional Effect of Amino Acid Substitutions and Indels. de Brevern AG, editor. PLoS ONE. 2012 Oct 8;7(10):e46688.

212. González-Pérez A, López-Bigas N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. Am J Hum Genet. 2011 Apr;88(4):440–9.

213. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res. 2011 Sep 1;39(17):e118.

214. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013 Jan;34(1):57–65.

215. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol [Internet]. 2016 Dec [cited 2019 Feb 24];17(1). Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4

216. Desmet F-O, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009 May;37(9):e67–e67.

217. Phillips SJ, Dudík M, Schapire RE. A maximum entropy approach to species distribution modeling. In: Twenty-first international conference on Machine learning - ICML '04 [Internet]. Banff, Alberta, Canada: ACM Press; 2004 [cited 2019 May 22]. p. 83. Available from: http://portal.acm.org/citation.cfm?doid=1015330.1015412

218. Phillips SJ, Anderson RP, Schapire RE. Maximum entropy modeling of species geographic distributions. Ecol Model. 2006 Jan;190(3–4):231–59.

219. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–15.

220. Online Mendelian Inheritance in Man, OMIM® [Internet]. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD); Available from: https://omim.org/

221. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018 Jan 4;46(D1):D649–55.

222. Belinky F, Nativ N, Stelzer G, Zimmerman S, Iny Stein T, Safran M, et al. PathCards: multi-source consolidation of human biological pathways. Database [Internet]. 2015 Jan 1 [cited 2019 Jul 9];2015. Available from: https://academic.oup.com/database/article/doi/10.1093/database/bav006/2433138

223. Roberts RJ. PubMed Central: The GenBank of the published literature. Proc Natl Acad Sci U S A. 2001 Jan 16;98(2):381–2.

224. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015 Jan 28;43(D1):D447–52.

225. Letunic I, Doerks T, Bork P. SMART: recent updates, new developments and status in 2015. Nucleic Acids Res. 2015 Jan 28;43(D1):D257–60.

226. Letunic I, Bork P. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. 2018 Jan 4;46(D1):D493–6.

227. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. Cancer Discov. 2012 May;2(5):401–4.

228. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. Sci Signal. 2013 Apr 2;6(269):pl1–pl1.

229. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014 Jul 1;42(W1):W252–8.

230. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.

231. Jmol: an open-source Java viewer for chemical structures in 3D. [Internet]. Available from: http://www.jmol.org/

232. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3--new capabilities and interfaces. Nucleic Acids Res. 2012 Aug;40(15):e115.

233. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics [Internet]. 2012 Dec [cited 2019 Feb 24];13(1). Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-134

234. Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol. 2013 Mar;31(3):213–9.

235. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs. Bioinformatics. 2012 Jul 15;28(14):1811–7.

236. Development Core Team (2008). R: A language and environment for statistical computing. [Internet]. Vienna, Austria.: R Foundation for Statistical Computing; Available from: http://www.R-project.org.

237. Méndez J, Stillman B. Chromatin association of human origin recognition complex, cdc6, and minichromosome maintenance proteins during the cell cycle: assembly of prereplication complexes in late mitosis. Mol Cell Biol. 2000 Nov;20(22):8602–12.

238. Nelson JD, Denisenko O, Bomsztyk K. Protocol for the fast chromatin immunoprecipitation (ChIP) method. Nat Protoc. 2006;1(1):179–85.

239. Vasen HF, Mecklin JP, Khan PM, Lynch HT. The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). Dis Colon Rectum. 1991 May;34(5):424–5.

240. 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 Jun;116(6):1453–6.

241. Matthew J. Betts, Robert B. Russell. Amino Acid Properties and Consequences of Substitutions. In: Bioinformatics for Geneticists. John Wiley & Sons, Ltd.; 2003. p. 289–316. (Chapter 14).

242. Hoesel B, Schmid JA. The complexity of NF-κB signaling in inflammation and cancer. Mol Cancer. 2013;12(1):86.

243. Zhan T, Rindtorff N, Boutros M. Wnt signaling in cancer. Oncogene. 2017;36(11):1461–73.

244. Trievel RC, Flynn EM, Houtz RL, Hurley JH. Mechanism of multiple lysine methylation by the SET domain enzyme Rubisco LSMT. Nat Struct Biol. 2003 Jul;10(7):545–52.

245. Zhao Y, Scott A, Zhang P, Hao Y, Feng X, Somasundaram S, et al. Regulation of paxillin-p130-PI3K-AKT signaling axis by Src and PTPRT impacts colon tumorigenesis. Oncotarget. 2017 Jul 25;8(30):48782–93.

246. Lindor NM. Familial colorectal cancer type X: the other half of hereditary nonpolyposis colon cancer syndrome. Surg Oncol Clin N Am. 2009 Oct;18(4):637–45.

247. Stoffel EM, Kastrinos F. Familial Colorectal Cancer, Beyond Lynch Syndrome. Clin Gastroenterol Hepatol. 2014 Jul;12(7):1059–68.

248. Esteban-Jurado C, Vila-Casadesús M, Garre P, Lozano JJ, Pristoupilova A, Beltran S, et al. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. Genet Med Off J Am Coll Med Genet. 2015 Feb;17(2):131–42.

249. DeRycke MS, Gunawardena SR, Middha S, Asmann YW, Schaid DJ, McDonnell SK, et al. Identification of novel variants in colorectal cancer families by high-throughput exome sequencing. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2013 Jul;22(7):1239–51.

250. Martin-Morales L, Rofes P, Diaz-Rubio E, Llovet P, Lorca V, Bando I, et al. Novel genetic mutations detected by multigene panel are associated with hereditary colorectal cancer predisposition. PloS One. 2018;13(9):e0203885.

251. Esteban-Jurado C, Garre P, Vila M, Lozano JJ, Pristoupilova A, Beltrán S, et al. New genes emerging for colorectal cancer predisposition. World J Gastroenterol. 2014 Feb 28;20(8):1961–71.

252. Sanchez de Abajo A, de la Hoya M, van Puijenbroek M, Godino J, Díaz-Rubio E, Morreau H, et al. Dual role of LOH at MMR loci in hereditary non-polyposis colorectal cancer? Oncogene. 2006 Mar 30;25(14):2124–30.

253. Buchanan DD, Stewart JR, Clendenning M, Rosty C, Mahmood K, Pope BJ, et al. Risk of colorectal cancer for carriers of a germ-line mutation in POLE or POLD1. Genet Med Off J Am Coll Med Genet. 2018;20(8):890–5.

254. To KKW, Poon DC, Wei Y, Wang F, Lin G, Fu L-W. Data showing the circumvention of oxaliplatin resistance by vatalanib in colon cancer. Data Brief. 2016 Jun;7:437–44.

255. Vollrath V, Wielandt AM, Iruretagoyena M, Chianale J. Role of Nrf2 in the regulation of the Mrp2 (ABCC2) gene. Biochem J. 2006 May 1;395(3):599–609.

256. Adachi T, Nakagawa H, Chung I, Hagiya Y, Hoshijima K, Noguchi N, et al. Nrf2-dependent and -independent induction of ABC transporters ABCC1, ABCC2, and ABCG2 in HepG2 cells under oxidative stress. J Exp Ther Oncol. 2007;6(4):335–48.

257. Werdyani S, Yu Y, Skardasi G, Xu J, Shestopaloff K, Xu W, et al. Germline INDELs and CNVs in a cohort of colorectal cancer patients: their characteristics, associations with relapse-free survival time, and potential time-varying effects on the risk of relapse. Cancer Med. 2017 Jun;6(6):1220–32.

258. Ceccaldi R, Liu JC, Amunugama R, Hajdu I, Primack B, Petalcorin MIR, et al. Homologous-recombination-deficient tumours are dependent on Polθ-mediated repair. Nature. 2015 Feb 12;518(7538):258–62.

259. Kent T, Chandramouly G, McDevitt SM, Ozdemir AY, Pomerantz RT. Mechanism of microhomology-mediated end-joining promoted by human DNA polymerase θ. Nat Struct Mol Biol. 2015 Mar;22(3):230–7.

260. Seki M, Marini F, Wood RD. POLQ (Pol theta), a DNA polymerase and DNA-dependent ATPase in human cells. Nucleic Acids Res. 2003 Nov 1;31(21):6117–26.

261. Prasad R, Longley MJ, Sharief FS, Hou EW, Copeland WC, Wilson SH. Human DNA polymerase theta possesses 5'-dRP lyase activity and functions in single-nucleotide base excision repair in vitro. Nucleic Acids Res. 2009 Apr;37(6):1868–77.

262. Arana ME, Seki M, Wood RD, Rogozin IB, Kunkel TA. Low-fidelity DNA synthesis by human DNA polymerase theta. Nucleic Acids Res. 2008 Jun;36(11):3847–56.

263. Brandalize APC, Schüler-Faccini L, Hoffmann J-S, Caleffi M, Cazaux C, Ashton-Prolla P. A DNA repair variant in POLQ (c.-1060A > G) is associated to hereditary breast cancer patients: a case-control study. BMC Cancer. 2014 Nov 19;14:850.

264. Wood RD, Doublié S. DNA polymerase θ (POLQ), double-strand break repair, and cancer. DNA Repair. 2016;44:22–32.

265. Yousefzadeh MJ, Wood RD. DNA polymerase POLQ and cellular defense against DNA damage. DNA Repair. 2013 Jan 1;12(1):1–9.

266. Larin M, Gallo D, Tamblyn L, Yang J, Liao H, Sabat N, et al. Fanconi anemia signaling and Mus81 cooperate to safeguard development and crosslink repair. Nucleic Acids Res. 2014 Sep;42(15):9807–20.

267. Adelman CA, Lolo RL, Birkbak NJ, Murina O, Matsuzaki K, Horejsi Z, et al. HELQ promotes RAD51 paralogue-dependent repair to avert germ cell loss and tumorigenesis. Nature. 2013 Oct 17;502(7471):381–4.

268. Puts GS, Leonard MK, Pamidimukkala NV, Snyder DE, Kaetzel DM. Nuclear functions of NME proteins. Lab Investig J Tech Methods Pathol. 2018;98(2):211–8.

269. Wang L-E, Gorlova OY, Ying J, Qiao Y, Weng S-F, Lee AT, et al. Genome-wide association study reveals novel genetic determinants of DNA repair capacity in lung cancer. Cancer Res. 2013 Jan 1;73(1):256–64.

270. Sun QA, Wu Y, Zappacosta F, Jeang KT, Lee BJ, Hatfield DL, et al. Redox regulation of cell signaling by selenocysteine in mammalian thioredoxin reductases. J Biol Chem. 1999 Aug 27;274(35):24522–30.

271. Tokar EJ, Kojima C, Waalkes MP. Methylarsonous acid causes oxidative DNA damage in cells independent of the ability to biomethylate inorganic arsenic. Arch Toxicol. 2014 Feb;88(2):249–61.

272. Xie Y, Zheng L, Tao L. Downregulation of IQGAP2 Correlates with Prostate Cancer Recurrence and Metastasis. Transl Oncol. 2019 Feb;12(2):236–44.

273. Xie Y, Yan J, Cutz J-C, Rybak AP, He L, Wei F, et al. IQGAP2, A candidate tumour suppressor of prostate tumorigenesis. Biochim Biophys Acta. 2012 Jun;1822(6):875–84.

274. Vaitheesvaran B, Hartil K, Navare A, Zheng null, OBroin P, Golden A, et al. Role of the tumor suppressor IQGAP2 in metabolic homeostasis: Possible link between diabetes and cancer. Metabolomics Off J Metabolomic Soc. 2014 Oct 1;10(5):920–37.

275. Kumar D, Hassan MK, Pattnaik N, Mohapatra N, Dixit M. Reduced expression of IQGAP2 and higher expression of IQGAP3 correlates with poor prognosis in cancers. PloS One. 2017;12(10):e0186977.

276. Deng Z, Wang L, Hou H, Zhou J, Li X. Epigenetic regulation of IQGAP2 promotes ovarian cancer progression via activating Wnt/β-catenin signaling. Int J Oncol. 2016 Jan;48(1):153–60.

277. Brisac C, Salloum S, Yang V, Schaefer EAK, Holmes JA, Chevaliez S, et al. IQGAP2 is a novel interferon-alpha antiviral effector gene acting non-conventionally through the NF-κB pathway. J Hepatol. 2016;65(5):972–9.

278. Bellavia S, Dahan K, Terryn S, Cosyns J-P, Devuyst O, Pirson Y. A homozygous mutation in INVS causing juvenile nephronophthisis with abnormal reactivity of the Wnt/beta-catenin pathway. Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc. 2010 Dec;25(12):4097–102.

279. Cohn O, Chen A, Feldman M, Levy D. Proteomic analysis of SETD6 interacting proteins. Data Brief. 2016 Mar;6:799–802.

280. Moorchung N, Kunwar S, Ahmed KW. An evaluation of nuclear factor kappa B expression in colorectal carcinoma: an analysis of 50 cases. J Cancer Res Ther. 2014 Sep;10(3):631–5.

281. Huang J, Kuismanen SA, Liu T, Chadwick RB, Johnson CK, Stevens MW, et al. MSH6 and MSH3 are rarely involved in genetic predisposition to nonpolypotic colon cancer. Cancer Res. 2001 Feb 15;61(4):1619–23.

282. Morin PJ, Sparks AB, Korinek V, Barker N, Clevers H, Vogelstein B, et al. Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. Science. 1997 Mar 21;275(5307):1787–90.

283. Rowan AJ, Lamlum H, Ilyas M, Wheeler J, Straub J, Papadopoulou A, et al. APC mutations in sporadic colorectal tumors: A mutational “hotspot” and interdependence of the “two hits.” Proc Natl Acad Sci U S A. 2000 Mar 28;97(7):3352–7.

284. Kanchi KL, Johnson KJ, Lu C, McLellan MD, Leiserson MDM, Wendl MC, et al. Integrated analysis of germline and somatic variants in ovarian cancer. Nat Commun [Internet]. 2014 Dec [cited 2019 Feb 25];5(1). Available from: http://www.nature.com/articles/ncomms4156

285. Koh P-K, Kalady M, Skacel M, Fay S, McGannon E, Shenal J, et al. Familial colorectal cancer type X: polyp burden and cancer risk stratification via a family history score. ANZ J Surg. 2011 Aug;81(7–8):537–42.

286. Yusenko MV, Nagy A, Kovacs G. Molecular analysis of germline t(3;6) and t(3;12) associated with conventional renal cell carcinomas indicates their rate-limiting role and supports the three-hit model of carcinogenesis. Cancer Genet Cytogenet. 2010 Aug;201(1):15–23.

287. Jandrig B, Seitz S, Hinzmann B, Arnold W, Micheel B, Koelble K, et al. ST18 is a breast cancer tumor suppressor gene at human chromosome 8q11.2. Oncogene. 2004 Dec 9;23(57):9295–302.

288. Bromberg JF, Wrzeszczynska MH, Devgan G, Zhao Y, Pestell RG, Albanese C, et al. Stat3 as an oncogene. Cell. 1999 Aug 6;98(3):295–303.

289. Scherr A-L, Gdynia G, Salou M, Radhakrishnan P, Duglova K, Heller A, et al. Bcl-xL is an oncogenic driver in colorectal cancer. Cell Death Dis. 2016 Aug;7(8):e2342–e2342.

290. Inagaki-Ohara K, Kondo T, Ito M, Yoshimura A. SOCS, inflammation, and cancer. JAK-STAT. 2013 Jul 15;2(3):e24053.

291. Jiang M, Zhang W, Liu P, Yu W, Liu T, Yu J. Dysregulation of SOCS-Mediated Negative Feedback of Cytokine Signaling in Carcinogenesis and Its Significance in Cancer Treatment. Front Immunol [Internet]. 2017 Feb 8 [cited 2019 May 13];8. Available from: http://journal.frontiersin.org/article/10.3389/fimmu.2017.00070/full

292. Chu Q, Shen D, He L, Wang H, Liu C, Zhang W. Prognostic significance of SOCS3 and its biological function in colorectal cancer. Gene. 2017 Sep;627:114–22.

293. Peruzzi D, Mori F, Conforti A, Lazzaro D, De Rinaldis E, Ciliberto G, et al. MMP11: a novel target antigen for cancer immunotherapy. Clin Cancer Res Off J Am Assoc Cancer Res. 2009 Jun 15;15(12):4104–13.

294. Han H-B, Gu J, Zuo H-J, Chen Z-G, Zhao W, Li M, et al. Let-7c functions as a metastasis suppressor by targeting MMP11 and PBX3 in colorectal cancer. J Pathol. 2012 Feb;226(3):544–55.

295. Kou Y-B, Zhang S-Y, Zhao B-L, Ding R, Liu H, Li S. Knockdown of MMP11 inhibits proliferation and invasion of gastric cancer cells. Int J Immunopathol Pharmacol. 2013 Jun;26(2):361–70.

296. Shetty V, Spellman DS, Neubert TA. Characterization by tandem mass spectrometry of stable cysteine sulfenic acid in a cysteine switch peptide of matrix metalloproteinases. J Am Soc Mass Spectrom. 2007 Aug;18(8):1544–51.

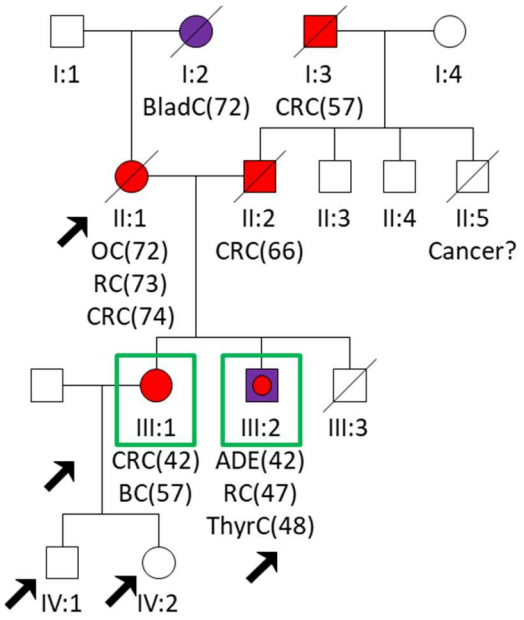
297. Ciavarella M, Miccoli S, Prossomariti A, Pippucci T, Bonora E, Buscherini F, et al. Somatic APC mosaicism and oligogenic inheritance in genetically unsolved colorectal adenomatous polyposis patients. Eur J Hum Genet EJHG. 2018;26(3):387–95.

7.2 List of websites

Ensembl	https://www.ensembl.org/index.html
Condel	https://bbglab.irbbarcelona.org/fannsdb/
Genecards	https://www.genecards.org/
HSF	http://www.umd.be/HSF/
Jmol	http://jmol.sourceforge.net/
Mutation Mapper	https://www.cbioportal.org/mutation_mapper
MutationTaster	http://www.mutationtaster.org/
NCBI	https://www.ncbi.nlm.nih.gov/
OMIM	https://www.omim.org/
Pathcards	https://pathcards.genecards.org/
PolyPhen 2	http://genetics.bwh.harvard.edu/pph2/
Primer 3	http://primer3.ut.ee/
Protein Data Bank	http://www.rcsb.org/
PROVEAN	http://provean.jcvi.org/genome_submit_2.php?species=human
Pubmed	https://www.ncbi.nlm.nih.gov/pubmed/
Reactome	https://reactome.org/
SIFT	https://sift.bii.a-star.edu.sg/
SMART	http://smart.embl-heidelberg.de/
STRING	https://string-db.org/
SWISS-MODEL	https://swissmodel.expasy.org/
UniProt	https://www.uniprot.org/

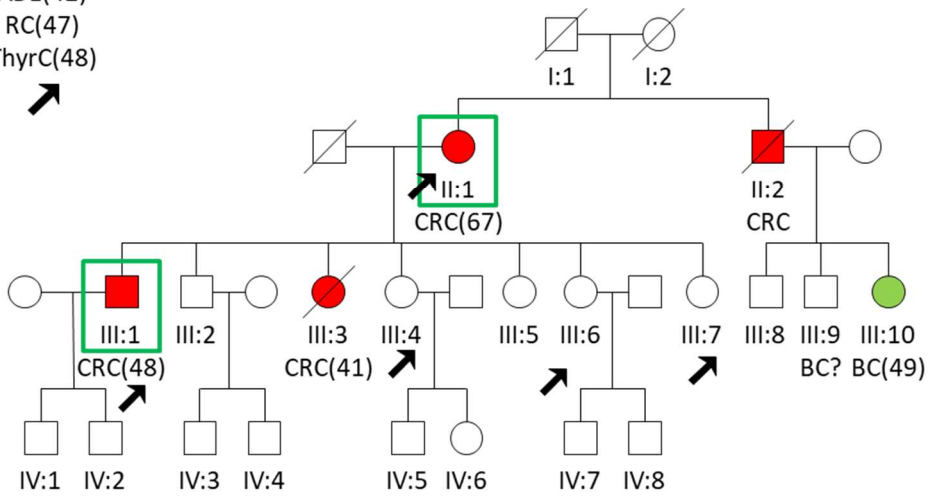
Appendix

A1. Pedigrees of the 13 families

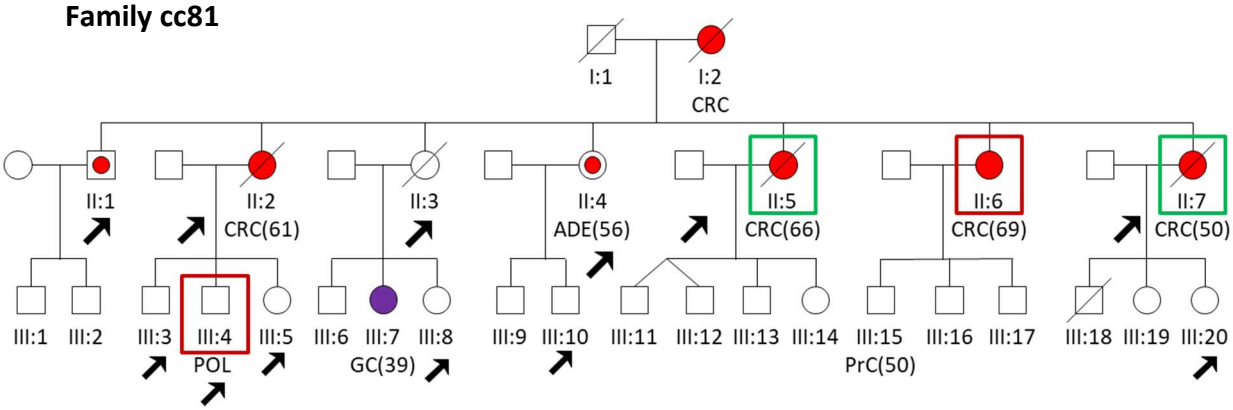


Family cc7

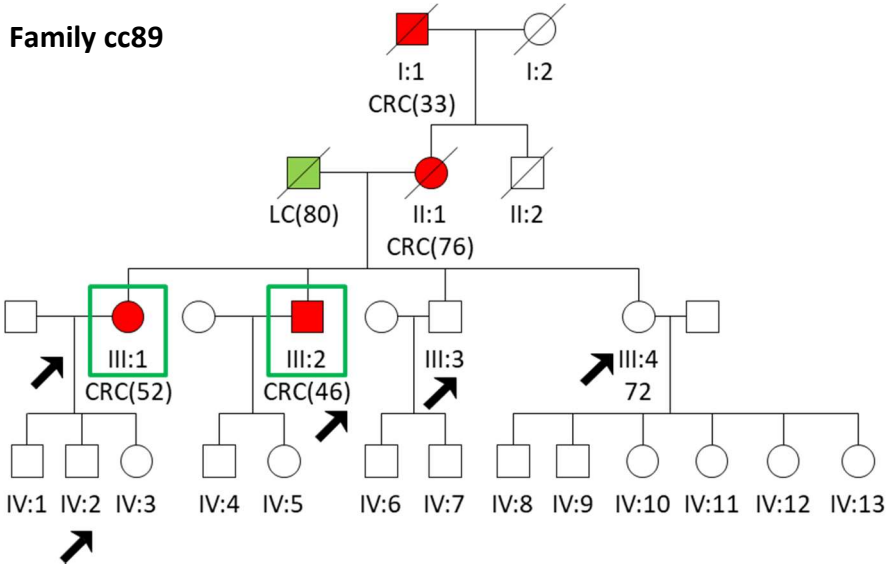
Family cc28



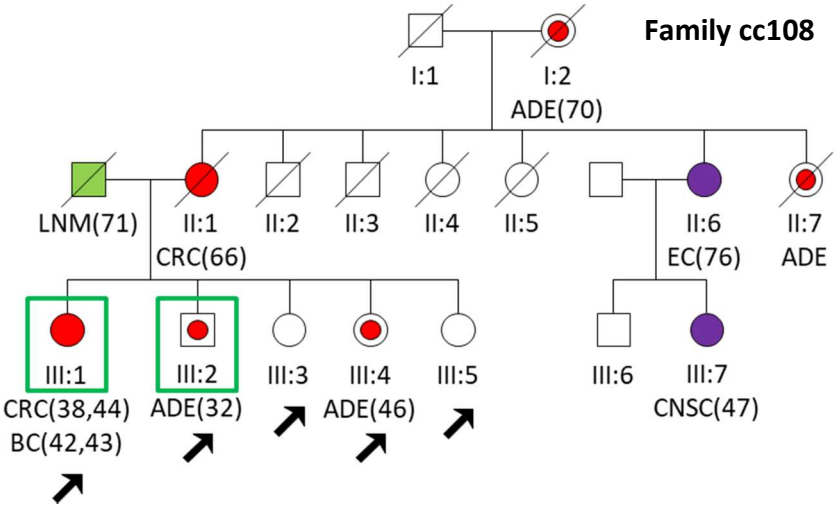
Family cc81



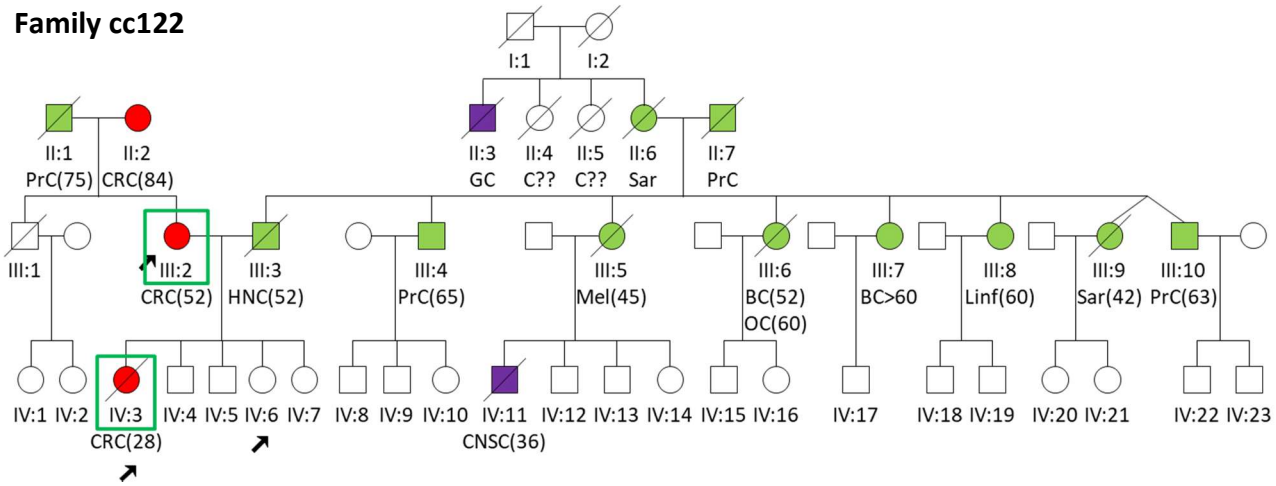
Family cc89



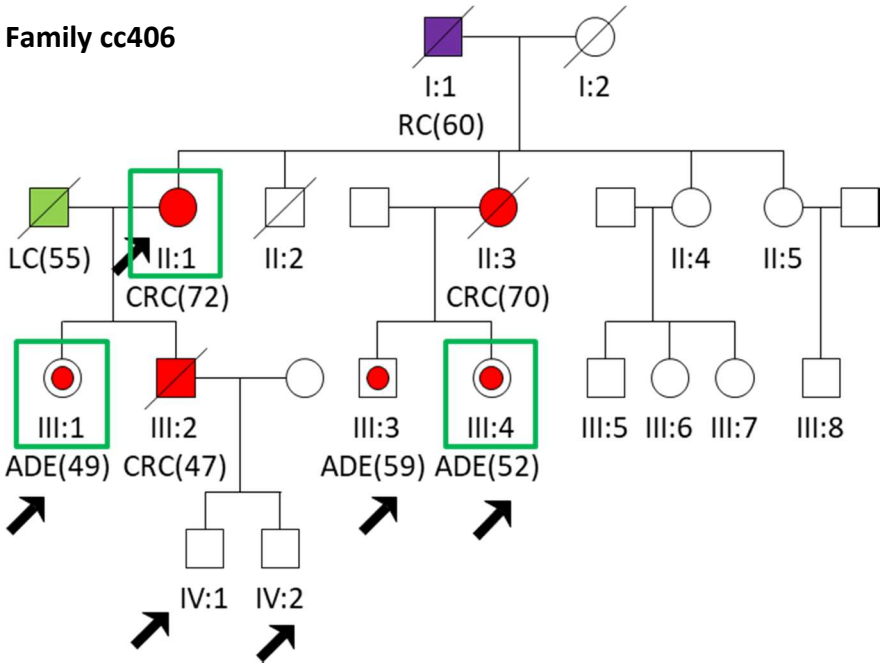
Family cc108



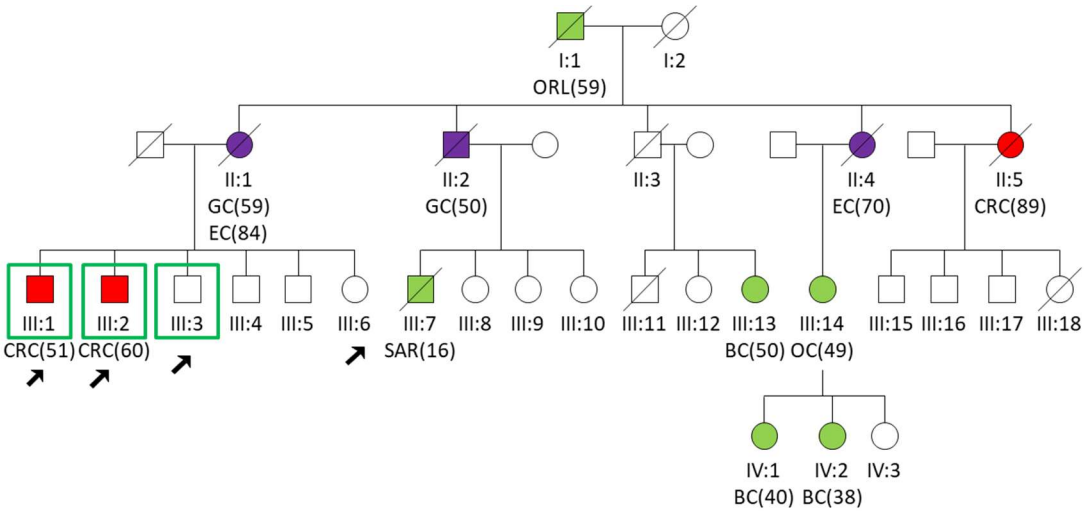
Family cc122



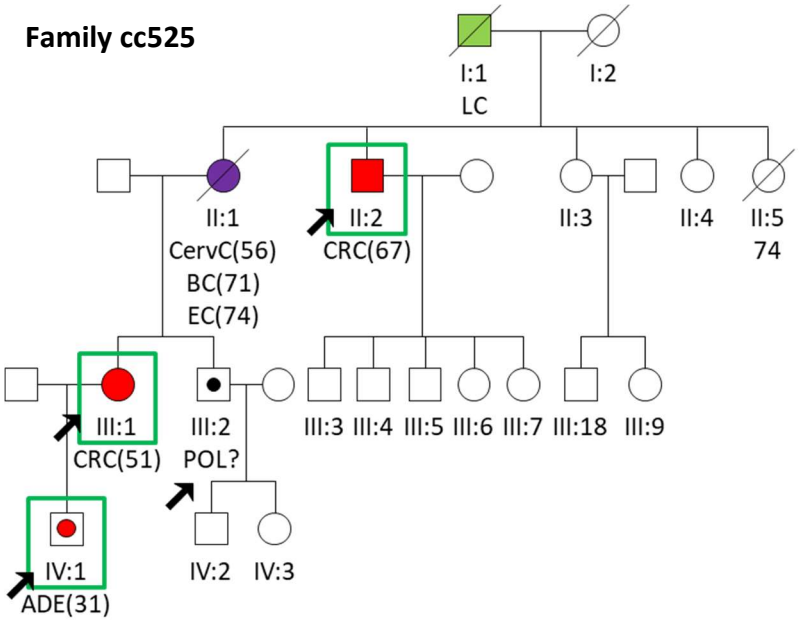
Family cc406



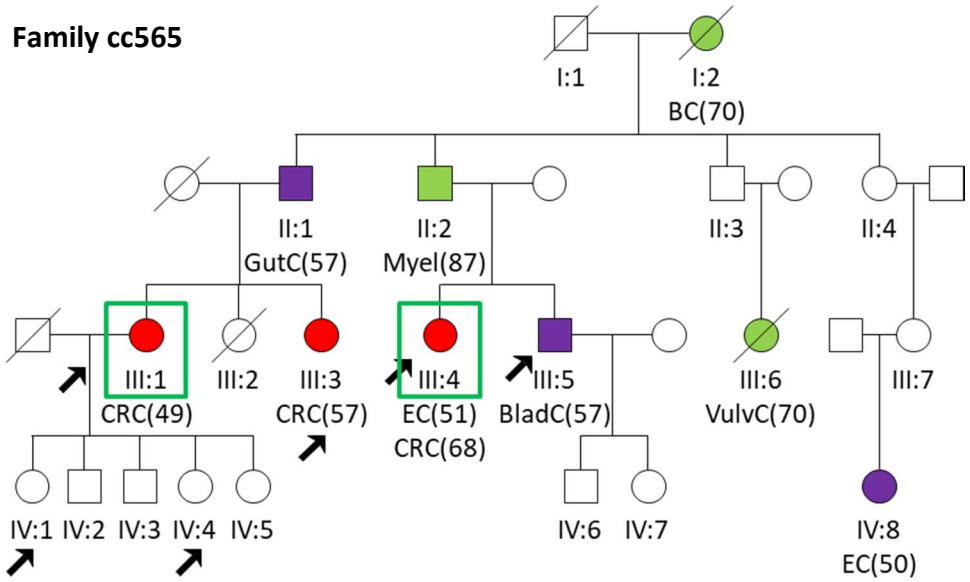
Family cc440



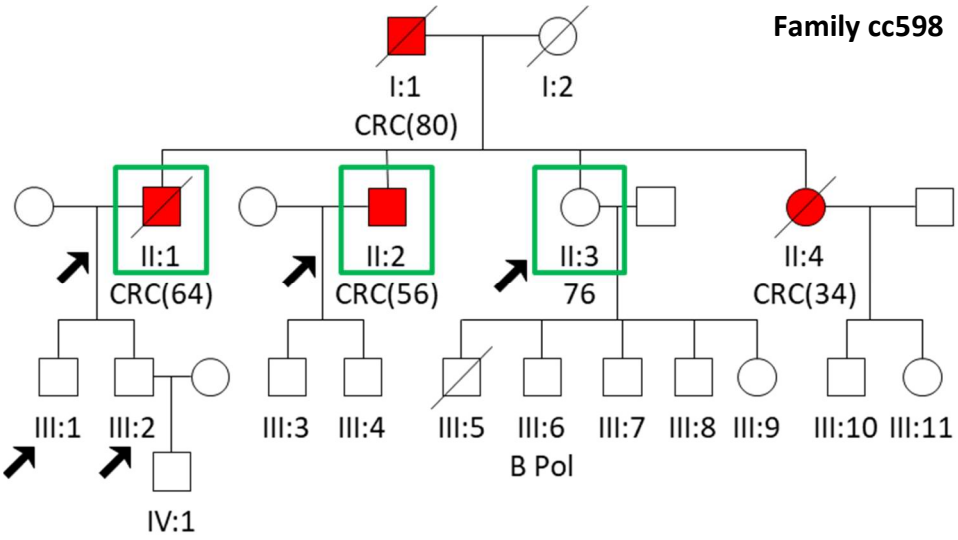
Family cc525



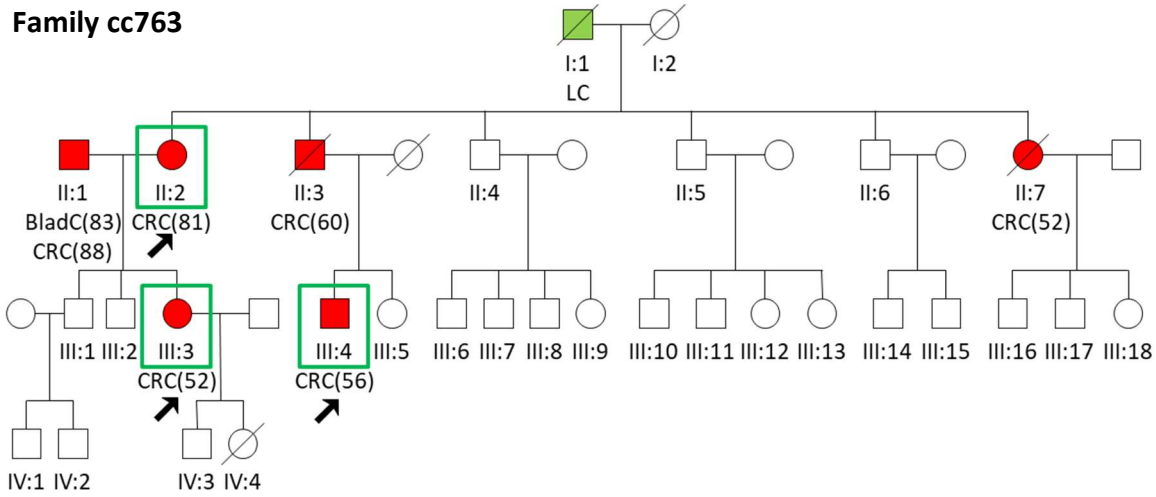
Family cc565

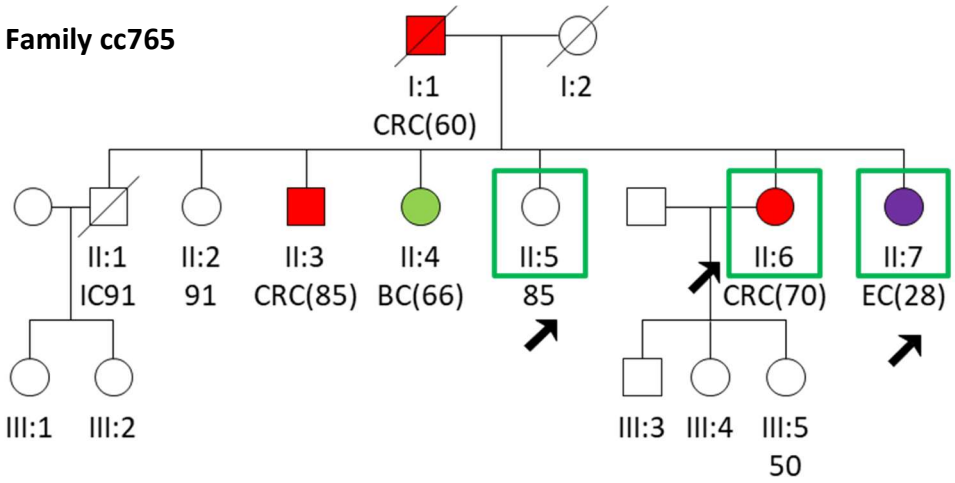


Family cc598



Family cc763





Pedigrees of the 13 families studied in this thesis. Males and females are represented by squares and circles, respectively. Identification numbers are displayed below each family member, with roman numerals indicating the different generations. Reported cancers and ages at diagnosis are detailed underneath and marked in 3 different colors: red for CRC, purple for HNPCC-associated cancers (including all defined by the International Collaborative Group on HNPCC, not just those included in the Amsterdam II criteria), and green for other cancers. Red circles in the center show those patients with adenomas, while black circles designate other polyps. Green boxes indicate those members whose whole-exome was sequenced, and arrows indicate those members whose germline DNA available. CRC: colorectal cancer, ADE: adenoma, EC: endometrial cancer, RC: renal cancer, BladC: bladder cancer, GIC: gastrointestinal cancer, OC: ovary cancer, CNSC: central nervous system cancer, BC: breast cancer, LC: lung cancer, PrC: prostate cancer, HNC: head and neck cancer, ThyrC: thyroid cancer, VulvC: vulvar cancer, Mel: melanoma, Linf: linfoma, Sar: sarcoma, Myel: myeloma.

A2. Information from the families

FAM	CRIT	ID	DNA	WES	FFPE	SEX	BIRTH	AGE	TUMOR	Dx AGE	CRC LOCALIZ	DIFF	CRC STAGE (TNM)	POLYPS	MSI	MMR IHC	MMR MUT
7	AC-II	II.1	2197	NO	NO	F	1931	N/A	OC, RC, CRC	72, 73, 74	Unk	WELL	IIIB-C	NR	MSS	POS	NEG
		III.1	275	YES	YES	F	1956	63	CRC, BC	42, 57	RECTUM	Unk	IIA-B	NR	MSS	POS	NEG
		III.2	284	YES	NO	M	1961	58	ADE, RC, TC	42, 47, 48	N/A	N/A	N/A	YES[d](46)	N/A	N/A	N/A
		IV:1	2588	NO	N/A	M	1987	32	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		IV:2	2587	NO	N/A	F	1985	34	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
28	AC-I	II.1	2226	YES	NO	F	1922	97	CRC	67	RECTUM	Unk	II	NR	MSS	POS	N/A
		III.1	448	YES	YES	M	1943	76	CRC	48	RECTUM	MOD	IIIB-C (T4N1M0)	NR	MSS	POS	NEG
		III:3	N/A	NO	YES	F	1953	N/A	CRC	40	LEFT COLON	MOD	IV	NR	N/A	N/A	N/A
		III.4	2225	NO	N/A	F	1947	72	N/A	N/A	N/A	N/A	N/A	NR	MSS	POS	N/A
		III:6	3185	NO	N/A	F	1948	71	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		III:7	3194	NO	N/A	F	1953	66	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
81	AC-I *	II:1	5955	NO	N/A	M	1945	74	N/A	N/A	N/A	N/A	N/A	YES	N/A	N/A	N/A
		II.2	757	NO	YES	F	1937	N/A	CRC	61	LEFT COLON	Unk	IIIB-C (T3N2M0)	NR	MSS	POS	NEG
		II:4	5957	NO	NO	F	1951	68	ADE	56	N/A	N/A	N/A	YES[m]	N/A	N/A	N/A
		II.5	750	YES	YES	F	1932	N/A	CRC	65	SIGMA	MOD	IIIB (T3N1M0)	YES	MSS	POS	NEG
		II:6	5956	YES	NO	F	1934	85	CRC	69	Unk	Unk	Unk	YES	N/A	N/A	N/A
		II.7	749	YES	YES	F	1941	N/A	CRC	50	LEFT COLON	Unk	IIA (T3N0M0)	YES	MSS	POS	NEG
		III:3	6604	NO	N/A	M	1961	58	N/A	N/A	N/A	N/A	N/A	YES	N/A	N/A	N/A
		III:4	6613	NO	N/A	M	1963	56	N/A	N/A	N/A	N/A	N/A	YES	N/A	N/A	N/A
		III:5	6614	NO	N/A	F	1972	47	N/A	N/A	N/A	N/A	N/A	YES	N/A	N/A	N/A
		III:8	6868	NO	N/A	M	1963	56	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		III:10	6869	NO	N/A	F	1980	39	N/A	N/A	N/A	N/A	N/A	YES	N/A	N/A	N/A
		III:20	6867	NO	N/A	F	1970	49	N/A	N/A	N/A	N/A	N/A	YES	N/A	N/A	N/A

FAM	CRIT	ID	DNA	WES	FFPE	SEX	BIRTH	AGE	TUMOR	Dx AGE	CRC LOCALIZ	DIFF	CRC STAGE (TNM)	POLYPS	MSI	MMR IHC	MMR MUT
89	AC-I	II:1	N/A	NO	YES	F	1917	N/A	CRC	76	SIGMA	WELL	IIB (T4N0M0)	NR	N/A	N/A	N/A
		III.1	772	YES	YES	F	1949	70	CRC	52	SIGMA	WELL	IIA (T3N0M0)	YES	MSS	POS	NEG
		III.2	948	YES	NO	M	1953	66	CRC	46	SIGMA / RECTUM	MOD	IIB-C (T4N0M0)	NR	N/A	N/A	NEG
		III:3	2644	NO	N/A	M	1955	64	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		III:4	5906	NO	N/A	F	1944	75	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		IV:2	2645	NO	N/A	M	1978	41	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
108	AC-II	III.1	972	YES	YES	F	1960	N/A	CRC, BC, BC, CRC	38, 42, 43, 44	RECTUM	MOD	1º: IIIB (T3N1M0), 2º: IIIA (T2N1M0)	YES	MSS	POS	NEG
		III.2	2614	YES	NO	M	1965	N/A	ADE	32	N/A	N/A	N/A	YES$^{t, hd}$(32)	N/A	N/A	N/A
		III.4	2609	NO	N/A	F	1962	57	ADE	46	N/A	N/A	N/A	YES^{tv}(46)	N/A	N/A	N/A
		III:5	2654	NO	N/A	F	1964	55	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		III:6	2656	NO	N/A	F	1961	58	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
122	AC-I	III.2	1042	YES	YES	F	1947	72	CRC	52	SIGMA	MOD	Unk	NR	MSS	POS	NEG
		IV.3	1043	YES	YES	F	1973	N/A	CRC	28	LEFT COLON	Unk	IIIC (T4bNXM0)	NR	MSS	POS	NEG
		IV.6	1044	NO	N/A	F	1971	48	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
406	AC-II	II:1	2439	YES	YES	F	1934	85	CRC	72	RIGHT COLON	Unk	0 (TisN0M0)	YES$^{tv, hd}$	MSS	POS	NEG
		III:1	2767	YES	NO	F	1956	63	ADE	49	N/A	N/A	N/A	YES$^{tv, md}$(49)	N/A	N/A	N/A
		III:2	N/A	NO	YES	M	Unk	N/A	CRC	47	Unk	Unk	Unk	NR	MSS	POS	N/A
		III:3	2765	NO	N/A	M	1948	71	ADE	59	N/A	N/A	N/A	YES$^{tv, d}$(59)	N/A	N/A	N/A
		III:4	2766	YES	N/A	F	1952	67	ADE	52	N/A	N/A	N/A	YES^{tv}(52)	N/A	N/A	N/A
		IV:1	2769	NO	N/A	M	1982	37	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		IV:2	2768	NO	N/A	M	1985	34	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A

FAM	CRIT	ID	DNA	WES	FFPE	SEX	BIRTH	AGE	TUMOR	Dx AGE	CRC LOCALIZ	DIFF	CRC STAGE (TNM)	POLYPS	MSI	MMR IHC	MMR MUT
440	AC-II *	III:1	2310	YES	YES	M	1954	65	CRC	51	Unk	Unk	Unk	NR	MSS	POS	NEG
		III:2	3190	YES	NO	M	1944	75	CRC	60	Unk	Unk	Unk	NR	N/A	N/A	N/A
		III:3	3188	YES	N/A	M	1945	74	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		III:6	3189	NO	N/A	F	1960	59	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
525	AC-II *	II:2	6230	YES	NO	M	1938	81	CRC	67	SIGMA	Unk	Unk	NR	N/A	N/A	N/A
		III:1	2695	YES	YES	F	1956	63	CRC	50	SIGMA	WELL	IIA (T3N0M0)	NR	MSS	N/A	NEG
		III:2	6336	NO	N/A	M	1965	54	N/A	N/A	N/A	N/A	N/A	YES(42)	N/A	N/A	N/A
		IV:1	6139	YES	YES	M	1980	39	ADE	30	N/A	N/A	N/A	YES[t](30)	N/A	N/A	N/A
565	AC-II	III:1	2961	YES	YES	F	1934	85	CRC	49	SIGMA	WELL	II	NR	MSS	POS	N/A
		III:3	5258	NO	NO	F	1939	80	CRC	58	RECTUM	Unk	Unk	YES[t](58)	N/A	N/A	N/A
		III:4	5257	YES	NO	F	1945	74	EC, ADE	51, 66	N/A	N/A	N/A	YES[tv, hd](66)	N/A	N/A	N/A
		III:5	5901	NO	NO	M	1951	68	Blad C	56	N/A	N/A	N/A	NR	N/A	N/A	N/A
		IV:1	5264	NO	N/A	F	1960	59	N/A	N/A	N/A	N/A	N/A	YES[b](39)	N/A	N/A	N/A
		IV:4	5259	NO	N/A	F	1962	57	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
598	AC-I	II:1	3191	YES	YES	M	1938	N/A	CRC	64	LEFT COLON	MOD	IIA (T3N0M0)	YES[tv, hd](64)	MSS	POS	NEG
		II:2	6133	YES	NO	M	1948	71	CRC	56	RIGHT COLON	MOD	I (T2N0M0)	NR	N/A	N/A	N/A
		II:3	6310	YES	N/A	F	1941	78	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		II:4	N/A	NO	YES	F	1951	N/A	CRC	34	RECTUM	WELL	IIIB (T4N1M0)	NR	MSS	POS	N/A
		III:1	5250	NO	N/A	M	1968	51	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		III:2	5251	NO	N/A	M	1966	53	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
763	AC-I *	II:1	6155	YES	NO	F	1929	90	CRC	81	RECTUM	WELL	I (T2N0M0)	NR	N/A	N/A	N/A
		III:3	6134	YES	YES	F	1952	67	CRC	52	Unk	Unk	Unk	NR	MSS	POS	N/A
		III:4	6138	YES	YES	M	1956	63	CRC	56	RECTUM	WELL	IIA (T3N0M0)	YES	MSS	POS	N/A

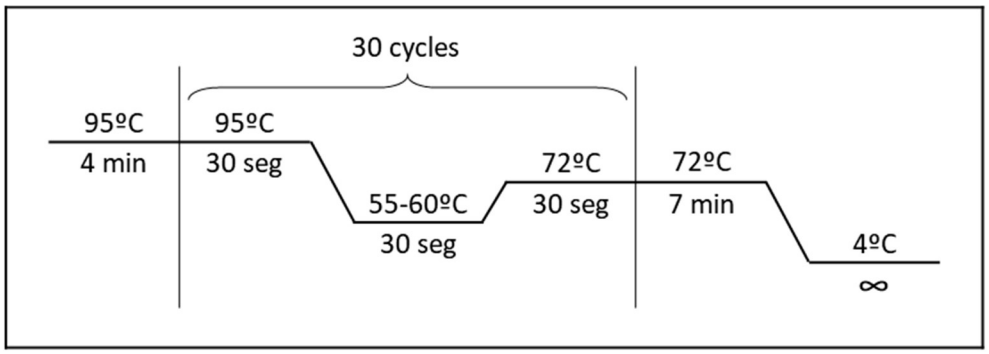
FAM	CRIT	ID	DNA	WES	FFPE	SEX	BIRTH	AGE	TUMOR	Dx AGE	CRC LOCALIZ	DIFF	CRC STAGE (TNM)	POLYPS	MSI	MMR IHC	MMR MUT
765	AC-II	II:2	7707	NO	N/A	F	1924	95	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		II:3	7708	NO	YES	M	1926	93	CRC	85	Unk	Unk	Unk	NR	MSS	POS	N/A
		II:4	7709	NO	YES	F	1928	91	BC	66	N/A	N/A	N/A	NR	N/A	N/A	N/A
		II:5	5322	YES	N/A	F	1930	89	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A
		II:6	5323	YES	YES	F	1932	87	CRC	70	RIGHT COLON	Unk	Unk	NR	MSS	POS	N/A
		II:7	6264	YES	NO	F	1936	83	EC	28	N/A	N/A	N/A	NR	N/A	N/A	N/A
		III:5	7290	NO	N/A	F	1964	55	N/A	N/A	N/A	N/A	N/A	NR	N/A	N/A	N/A

Family members from the 13 families from which either germline or tumor DNA was available at some point of the study. The different columns are described below. FAM: family identification number; CRIT: clinical criteria by which the family was classified (AC-I: Amsterdam I, AC-II: Amsterdam II, *borderline: earliest age of onset allowed up to 52 years old); ID: identification number of each member referred to the pedigree (Appendix Figure A1); DNA: germline DNA assigned number; WES: whether the whole-exome has been sequenced; FFPE: availability of FFPE tumor blocks; SEX: gender of the patient (F: female, M: male); BIRTH: year of birth; AGE: current age in 2019; TUMOR: tumor type (CRC: colorectal cancer, EC: endometrial cancer, RC: renal cancer, ADE: colorectal adenoma, BC: breast cancer, Blad C: bladder cancer, TC: thyroid cancer); Dx AGE: age at diagnosis; CRC LOCALIZ: localization of the CRC; DIFF: differentiation of the CRC; CRC STAGE (TNM): stage of the CRC according to the TNM system; POLYPS: detection of polyps and earliest age (t: tubular, v: villious, tv: tubulovillious, hd: high dysplasia, md: moderate dysplasia, ld: low dysplasia, h: hyperplastic, m: malignant, b: benign); MSI: result from the MSI test; MMR IHC: immunohistochemistry result for MLH1, MSH2 and MSH6; MMR MUT: result from the MMR protein screening for MLH1, MSH2 and MSH6. N/A: not applicable, Unk: unknown, NR: not reported.

A3. Primers and PCR conditions

PCR conditions:

	Stock concentration	Working concentration
Buffer (no $MgCl_2$)	10x	1x
$MgCl_2$	25mM	2mM
dNTPs	2.5mM	0.2mM
Forward primer	10µM	0.2µM
Reverse primer	10µM	0.2µM
Fast-Taq	5U/uL	1U
H_2O	-	-
DNA template	25ng/uL	25ng



Primers for cDNA validation:

Gene	Exon	Primer	Sequence	PCR cond	Product gDNA/cDNA
PALB2	2	FOR	AAATTAGCATTCTTGAAAAGGGAAT	AT=60ºC	269bp / 152bp
PALB2	3	REV	GAGTGTTTTAGCTGCGGTGA		

Primers for validation, segregation and LOH:

Family	Gene	Exon	Primer	Sequence	PCR cond	Product	Sanger
cc7	MUC6	ex17	FOR	TACAACAGCCAAGCCTGTGA	AT=60ºC	132bp	FOR
			REV	ACACTCGCCCTTTTGGTTC			
	TXNRD3	ex4	FOR	TCCATACCAGTTGAACATAGGG	AT=58ºC	165bp	FOR
			REV	CACCACCGATGATGATGAGA			
	GAL3ST2	ex1	FOR	CCAGAGGCCAAGATGATGTC	AT=60ºC	146bp	REV
			REV	AGTCACTTCCTTCGAAAGACAAAC			
cc28	CHID1	ex1	FOR	CCATCAGCAGAGGGGTTG	AT=58ºC	194bp	FOR
			REV	TGGCATCTGACTTTGACAGG			
	PYGO1	ex3	FOR	CTCTTCACCCAAACCGTCAT	AT=58ºC	176bp	FOR
			REV	TGCAGTTAAGAGGCCATAAGC			
	MMP11	ex2	FOR	CCAGACCCTCATGTCATCCT	AT=56ºC, GC rich	194bp	FOR
			REV	GGCACTCAGCCCATCAGA			
	AHRR	ex7	FOR	AGTACTCGGCCTTCCTGACC	AT=58ºC, GC rich	150bp	REV
			REV	CCTCTGTTTCCTTCCACACC			
	NKD2	ex7	FOR	CTGGGGTCTGCTCTGTCAGT	AT=60ºC	177bp	REV
			REV	GAGGACCCTCCTTCCTCTTG			
cc81	ECH1	ex4	FOR	CAAAGGAGATGATGTGGCCC	AT=58ºC	127bp	FOR
			REV	GGCGCGTATGGAGAAAGATC			
	MAPK15	ex2	FOR	TCTGTGGCCTCCTTTCCTC	AT=60ºC	350bp	FOR
			REV	TTCTCCCCACACACTCACCT			
	ATR	ex13	FOR	TAGGGCCGCAAAAGGAGATT	AT=58ºC	206bp	FOR
			REV	GCCTATAGTCCAGACAAACGC			
cc89	POLQ	ex16	FOR	GAGCGACTAAGATAGATCATTTGGA	AT=58ºC	176bp	FOR
			REV	TTGTGAATCAGTAACAGAAACTTCAA			
	C19orf40	ex4	FOR	ACTGTGCTGGACCTTGGAAT	AT=55ºC, 35 cycles	150bp	REV
			REV	GATGGTTGCTTCTGGAGAGC			
	ADIPOQ	ex3	FOR	AGAAGGTTGTGAGTGGGAGC	AT=55ºC	197bp	FOR
			REV	TTCTCCTTTCCTGCCTTGGA			
	NUP153	ex13	FOR	GCTCTCTAAGGCCTCCGC	AT=53ºC, 35 cycles	190bp	FOR
			REV	ATACCTGATGCTGTTGTCGC			
cc108	NME7	ex12	FOR	CTGAAGTCCTGACCAACCTCT	AT=58ºC	157bp	FOR
			REV	TGTCCCAACCTGTGACTTCT			
	IQGAP2	ex12	FOR	TCCACAGTTATGCAAACACACT	AT=55ºC	179bp	REV
			REV	ACCTTCTTCTCCCACTATGCT			
	RASSF5	ex4	FOR	TCCTTCTACCTGCCCCTAGA	AT=55ºC, 35 cycles	159bp	REV
			REV	CCTACCTTGTCCGTCCTTGT			
	NEK4	ex4	FOR	GGCTTACCTTGTCTCTGGGA	AT=58ºC	203bp	FOR
			REV	GGCCATTTCATAGACACAGCA			

Family	Gene	Exon	Primer	Sequence	PCR cond	Product	Sanger
cc122	MUS81	ex15	FOR	GAGACACTGCTGAGCACCAT	AT=60ºC	181bp	REV
			REV	GCTGCAGTAGAGCTGGGATAA			
	VEPH1	ex9	FOR	AGTTTCACACTCCAGAGGAAGT	AT=55ºC	199bp	FOR
			REV	GCTGGTATGTCTCCCCTGTT			
	RNMT	ex7	FOR	GACAAATTTCGTGACCCACA	AT=58ºC	169bp	FOR
			REV	TCAATTCAAAGCTATTGGGAGT			
	ABCB1	ex5	FOR	AAGCAAATCTTCCATGAAACTGT	AT=55ºC	160bp	FOR
			REV	ACACCAGCATCATGAGAGGA			
	CTNNAL1	ex6	FOR	GGACCCAGGAGATTTTGGTA	AT=58ºC	206bp	FOR
			REV	TGACCTTATCCAATGCCACTT			
cc406	SHF	ex2	FOR	CCCCCAAGGGTAGTCTGAGT	AT=60ºC	150bp	FOR
			REV	GCCAGTTGCTTTTCTTCTTCTG			
	ABCC2	ex9	FOR	TCATTCAGTCTTTCTGCCTTCA	AT=58ºC	137bp	FOR
			REV	TGAGGGGATTTTCTTTGGTG			
cc440	GALNT10	ex4	FOR	TGCAGTTAGCCAAACCCTGT	AT=58ºC	150bp	FOR
			REV	GTTGTGGAAGGGGATGATGA			
	PHACTR2	ex5	FOR	AATCACCTGTGCCTCCGAAA	AT=55ºC, 35 cycles	201bp	REV
			REV	AGAGCCAAGCACCCTAGAAG			
	CASP7	ex6	FOR	CATCCTCTTAAGCCATGGAGA	AT=58ºC	183bp	FOR
			REV	TGCAATCTTTCCCATTTTCA			
	HELQ	ex8	FOR	TGGTAGCATTGGTGACAGAAG	AT=58ºC	198bp	REV
			REV	GCATAATTAGTGTTTGCCCAAT			
cc525	ECT2L	ex17	FOR	CCATCCCGAAGATTTGAAGA	AT=58ºC	201bp	FOR
			REV	TGAGTGGGAGTACCACAAAGG			
	EIF4G3	ex12	FOR	CATCCCCAAGAGAAGACACAA	AT=58ºC	198bp	FOR
			REV	CCGTGGCTGATAATTTTTCG			
	C9orf116	ex2	FOR	GCCGTCAACCAAATACAAGG	AT=58ºC	230bp	FOR
			REV	CCCCAGGTTTTTAGCCTCTT			
cc565	ADAMTS14	ex18	FOR	AGTTCACCAAATACGGCTGC	AT=53ºC, 35c, GC rich	167bp	FOR
			REV	CATAGAACCAGACTCCCCGT			
	ARMC3	ex18	FOR	GGTGGTAAGATTCCAAAAGAGAAAC	AT=60ºC	157bp	REV
			REV	ATACCTTACACACCTTGAAAAGCAA			
	RHOBTB1	ex11	FOR	TGCATCTTTCCTTCACCAGA	AT=58ºC	155bp	REV
			REV	CACTTTCGTCTTGAGCGATG			
cc598	SETD6	ex6	FOR	CCACTCAGCCCATTCCTAAA	AT=58ºC	174bp	REV
			REV	TGATACACTCACCCTGTAATGCT			
	CCDC60	ex14	FOR	GCCACCAGTCTCAGTCACAA	AT=60ºC	162bp	FOR
			REV	GCAGCCAGCTGATGTAATCC			
	ST18	ex15	FOR	TCTCTCTTCAACAGCAAAGCA	AT=55ºC, 35 cycles	208bp	FOR
			REV	TGGCTTGTTGGAGAGGATGT			
cc763	WWOX	ex4	FOR	TTCTCTTTTGGGCAGCCATA	AT=58ºC	171bp	FOR
			REV	ACCACAACCACTTTGCCAGT			
cc765	PTPRT	ex29	FOR	CAGCCCTAACAGCCTCTCTG	AT=60ºC	189bp	FOR / REV
			REV	GCAGTGGACCACAGTACGTC			
	MAP3K6	ex20	FOR	TGTCATTGAGATGGCCACAG	AT=56ºC	156bp	REV
			REV	GGGTGAACAAGGGTGACATT			
	ABTB1	ex11	FOR	GCAGCCTATGATGTGCTGAG	AT=60ºC	199bp	FOR
			REV	ACTGGCCCACCTTCTCAAT			
	INVS	ex16	FOR	GGACTTTACCCTGGGCCTAC	AT=60ºC	176bp	FOR
			REV	TTCCCTTCGTGAGAACAACC			

Primers for *SETD6* cloning:

Clone	Primer	Sequence	PCR cond	Product
SETD6-N	SETD6-AscI FOR	GGCGGCGCGCCGCGACCCAGGCGAAGC GTC	AT=59ºC, 1m elongation	817bp
	SETD6-N-PacI REV	GCCTTAATTAATTAGCCTATTTGCCCATA AGTGTTGAAAATCTCATGG		
SETD6-C	SETD6-C-AscI FOR	GCCGGCGCGCCCAAATGGCTAACTGGCA ACTGATTC	AT=59ºC, 1m elongation	586bp
	SETD6-PacI REV	GGCTTAATTAACTAACTTGTCAGTTCCAA CAACTGATG		
Seq	T7 promoter	TAATACGACTCACTATAGGG	N/A	N/A

Primers for *PTPRT* methylation:

MSP primer set 1 (CRC paper)			
Primer	**Sequence**	**PCR cond**	**Product**
PTPRT-FM (methylated 1 FOR)	TGTTCGAGATTTTGGTTAGTAGTC	AT=55ºC, 40 cycles	231bp
PTPRT-RM (methylated 1 REV)	ACATCGTCTACCCTATAAAAACG		
PTPRT-FU (unmethylated 1 FOR)	GTGTTTGAGATTTTGGTTAGTAGTT	AT=55ºC, 40 cycles	233bp
PTPRT-RU (unmethylated 1 REV)	AACATCATCTACCCTATAAAAACAC		

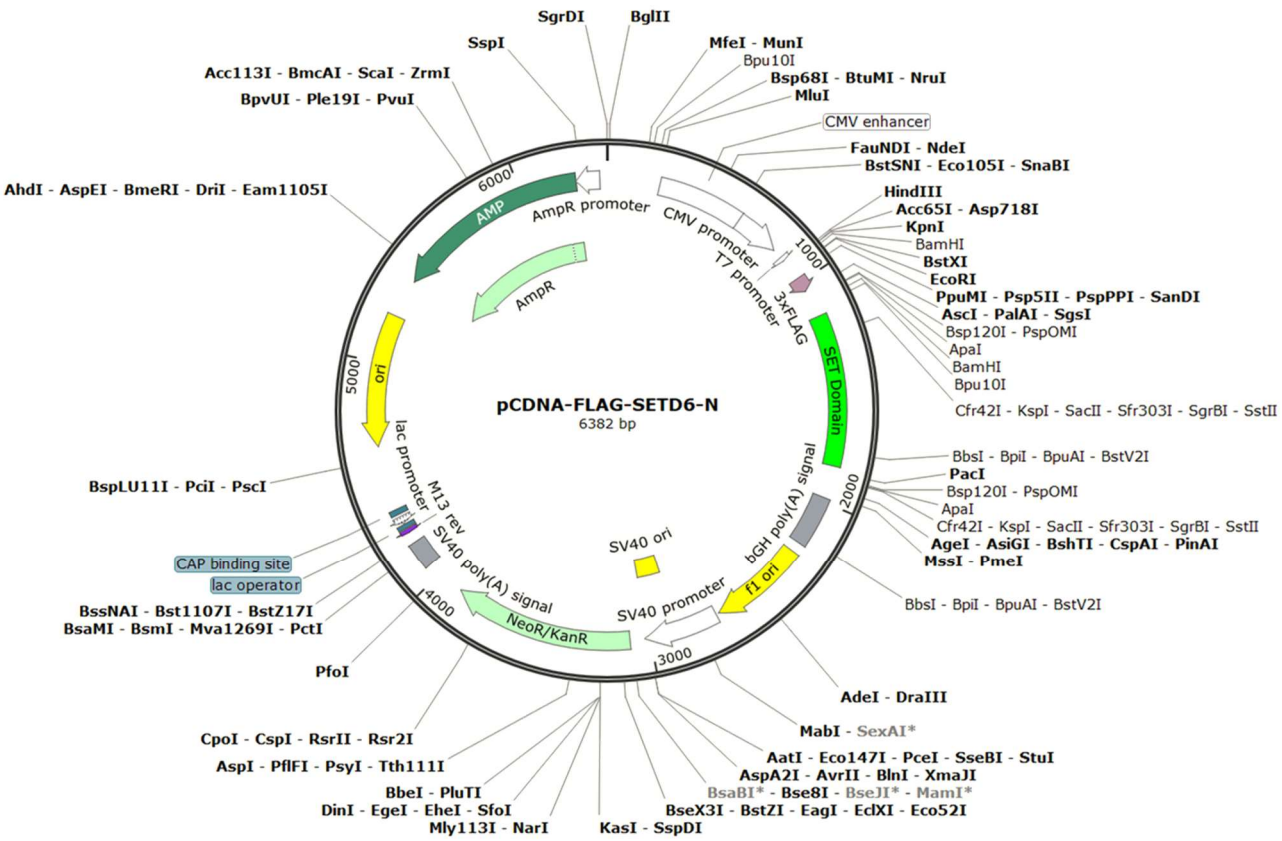
MSP primer set 2 (HNSCC paper)			
Primer	**Sequence**	**PCR cond**	**Product**
PTPRT-FM2 (methylated 2 FOR)	TATTTGCGTTTTTATAGGGTAGACG	AT=55ºC, 40 cycles	120-150bp
PTPRT-RM2 (methylated 2 REV)	GAACCGCTAACACAATCACG		
PTPRT-FU2 (unmethylat. 2 FOR)	ATTTGTGTTTTTATAGGGTAGATGA	AT=55ºC, 40 cycles	120-150bp
PTPRT-RU2 (unmethylat. 2 REV)	ACTCAAACCACTAACACAATCACAC		

Pyrosequencing			
Primer	**Sequence**	**PCR cond**	**Product**
PTPRT-Pyro FOR	TGGGGACAAGAGAGAGATGT	AT=48ºC, 45 cycles	164bp
PTPRT-Pyro REV	GCACACCCTGAGATGAGAATTTCATC		
PTPRT-Pyro SEQ	GGCTAATTGAGGGGGA	N/A	N/A

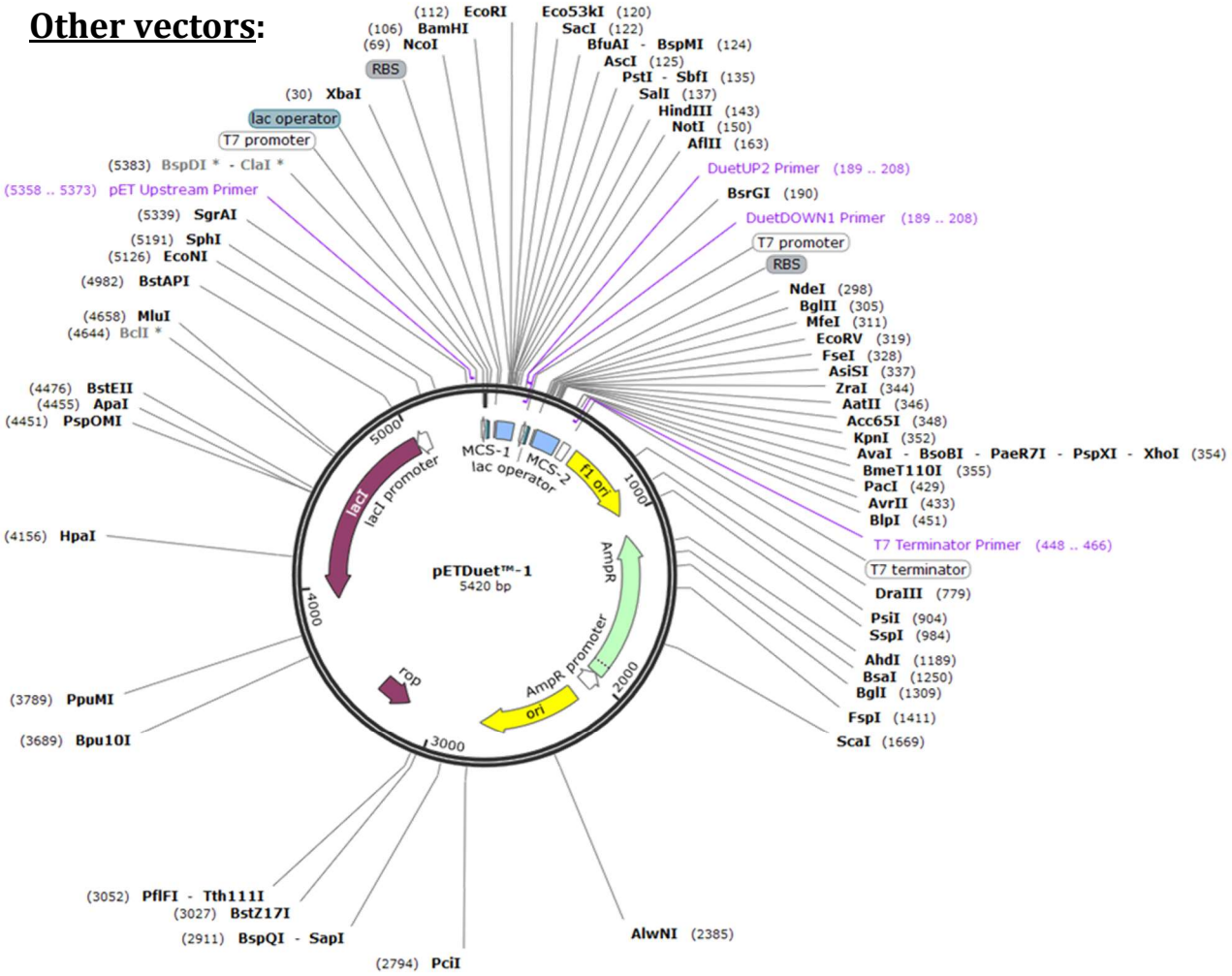
A4. Plasmids

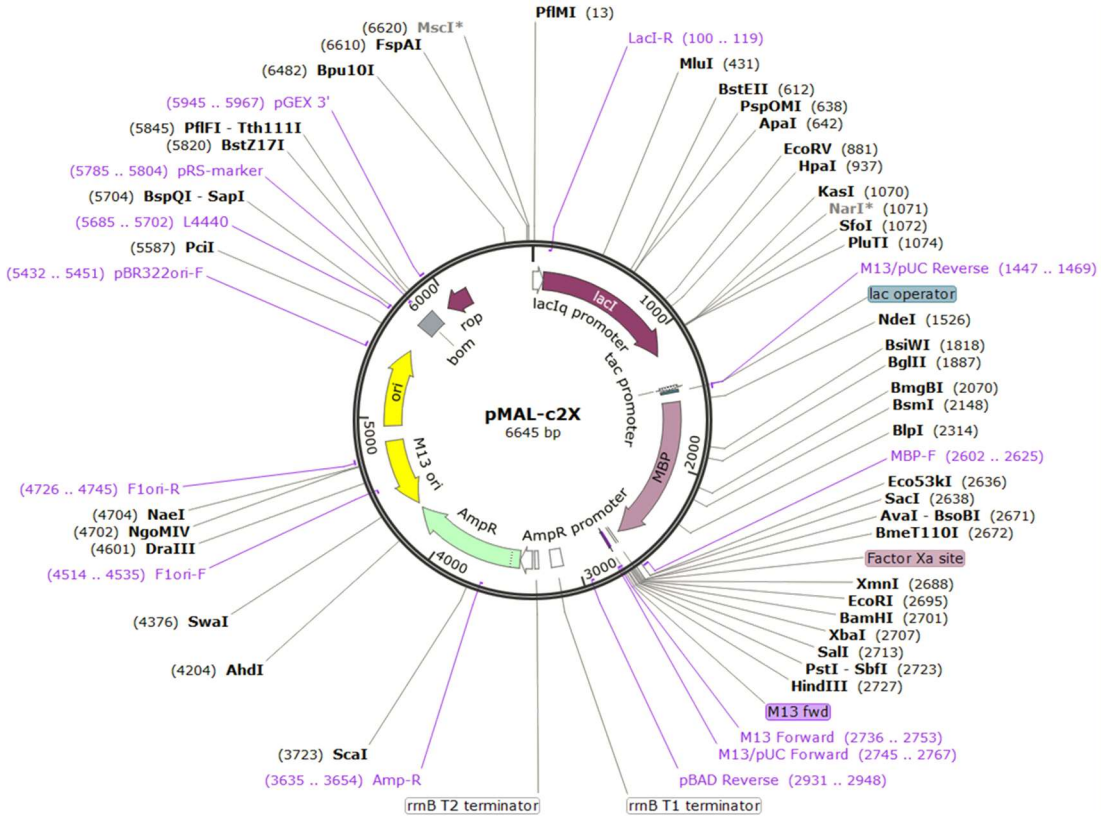
Application	Plasmid	Insert	Tag
Transfection	**pcDNA3.1**	SETD6 wt	FLAG, HA, GFP
		SETD6-N	FLAG, HA, GFP
		SETD6-C	FLAG, GFP
		PAK4	FLAG
		RelA	FLAG
Purification of recombinant proteins	**pET-Duet**	SETD6 wt	His
		SETD6-N	His
		SETD6-C	His
		PAK4	His
	pMAL-c2x	RelA	MBP

pcDNA3.1 plasmid with SETD6-N insert:

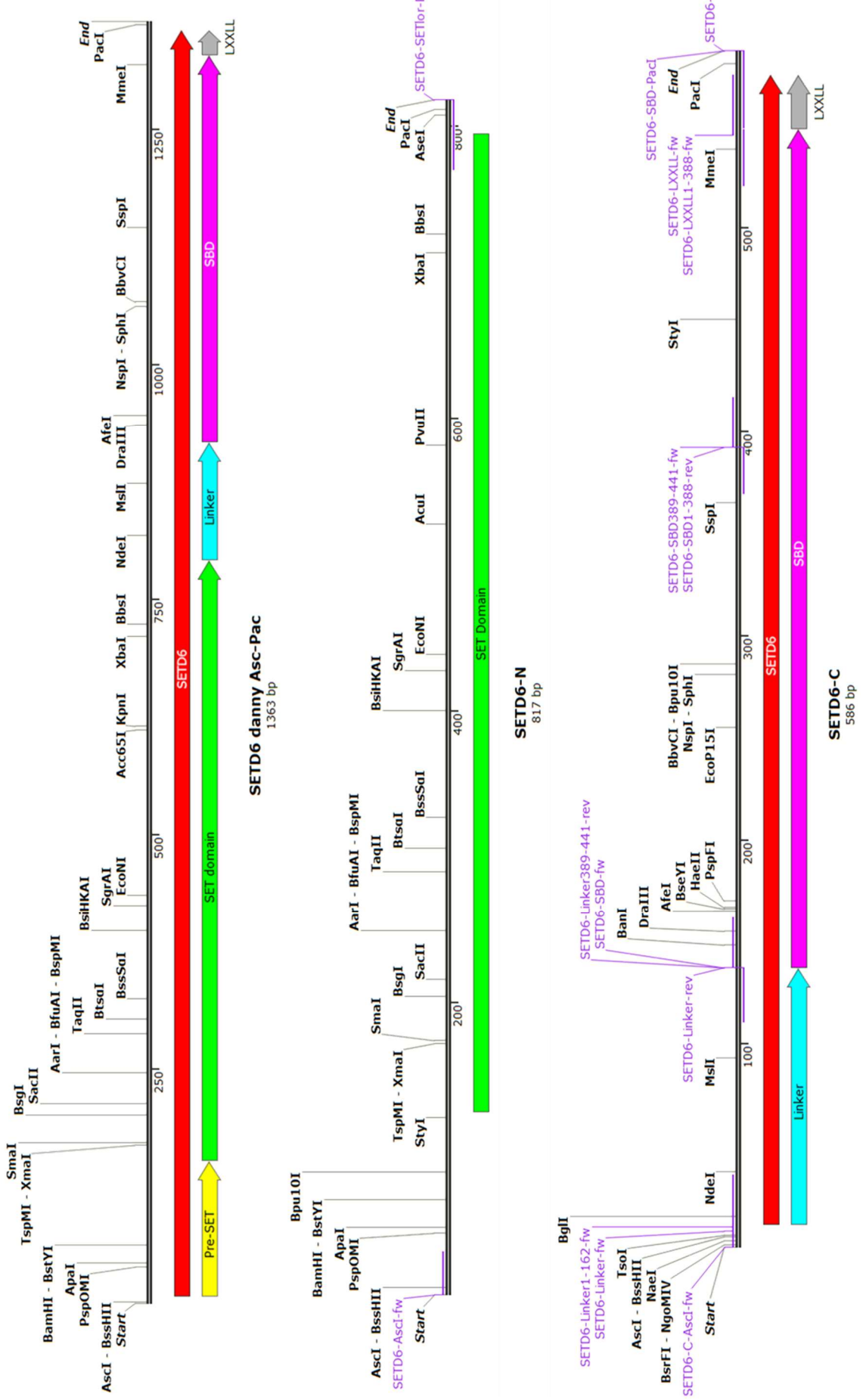


Other vectors:





Inserts:


SETD6 danny Asc-Pac
1363 bp
SETD6
SET domain
Pre-SET
Linker
SBD
LXXLL
AscI - BssHII
Start
BamHI - BstYI
ApaI
PspOMI
TspMI - XmaI
SmaI
BsgI
SacII
AarI - BfuAI - BspMI
TaqII
BtsaI
BssSaI
BsiHKAI
SgrAI
EcoNI
Acc65I
KpnI
XbaI
BbsI
NdeI
MslI
AfeI
DraIII
NspI - SphI
BbvCI
SspI
MmeI
End
PacI
SETD6-N
817 bp
SET Domain
SETD6-SETlor-PacI-rev
AseI
SETD6-Ascl-fw
SETD6-C
586 bp
SETD6-Linker-rev
SETD6-Linker389-441-rev
SETD6-SBD-fw
SETD6-Linker1-162-fw
SETD6-Linker-fw
SETD6-C-Ascl-fw
SETD6-SBD389-441-fw
SETD6-SBD1-388-rev
SETD6-SBD-PacI
SETD6-LXXLL-fw
SETD6-LXXLL1-388-fw
SETD6-PacI-rev

A5. Buffers and antibodies

DNA gels:

TAE buffer x50
2M Tris
50mM EDTA
Acetic acid

Protein gels and westerns:

	Separating gel		
	8%	10%	12%
Acrylamide	2.7mL	3.3mL	4mL
1.5M Tris, pH 8.8	2.5mL	2.5mL	2.5mL
SDS (20%)	50µL	50µL	50µL
APS (10%)	100µL	100µL	100µL
TEMED	10µL	10µL	10µL
H_2O	4.7mL	4.1mL	3.4mL

	Stacking gel
Acrylamide	0.8mL
1M Tris, pH 6.8	0.5mL
SDS (20%)	20µL
APS (10%)	40µL
TEMED	4µL
H_2O	2.7mL

Protein sample buffer (PSB)
250mM Tris-HCl, pH=6.8
10% SDS
30% Glycerol
5% β-mercaptoethanol
Bromophenol blue (very little)

Running buffer x10
250mM Tris
1.92M Glycine
1% SDS
pH 8.1-8.5

Transfer buffer
25mM Tris
192mM Glycine
20% methanol
pH 8.3

PBST 0.1%
3.2mM Na_2HPO_4
0.5mM KH_2PO_4
1.3mM KCl
135mM NaCl
0.1% Tween

TBST 0.1%
50mmM Tris pH=7.5
150mM NaCl
0.1% Tween

Antibody	Dilution	Secondary	Reference	Notes
α-FLAG	1:5000	α-Mouse	Sigma-Aldrich, F1804	10% skim milk in PBST
α-HA	1:3000	α-Mouse	Millipore, 05-904	10% skim milk in PBST
α-H3	1:2000	α-Mouse	Abcam, ab10799	10% skim milk in PBST
α-tubulin	1:3000	α-Mouse	Abcam, ab44928	10% skim milk in PBST
α-GAPDH	1:2000	α-Mouse	Abcam, ab8245	10% skim milk in PBST
α-SETD6	1:2000	α-Mouse	Genetex, GTX629891	10% skim milk in PBST
α-RelAK310me1		α-Rabbit	Levy et al. 2011	
α-Pan Me	1:500	N/A	ImmuneChem, ICP0502	5% BSA in TBST

Protein purification:

Wash buffer
PBS x1
20-25mM Imidazole

Elution buffer
PBS x1
500mM Imidazole

Dialysis buffer
PBS x1
10% Glycerol

Lysis buffer
PBS x1
10mM Imidazole
PMSF (1:100)
Protease inhibitors tablet

Lysis buffer (GST proteins)
50mM Tris pH=7.5
150mM NaCl
0.05% NP-40

PKMT buffer x5
50mM Tris pH 8
10% Glycerol
20mM KCl
5mM $MgCl_2$

Cells:

Ripa buffer
15mM NaCl
50mM Tris-HCl, pH=8
1% NP-40
0.5% DOC
0.1% SDS

A6. Articles published during this PhD

Related publications: (attached)

- **Martín-Morales L**, Feldman M, Vershinin Z, Garre P, Caldés T, Levy D. SETD6 dominant negative mutation in familial colorectal cancer type X. Hum Mol Genet. 2017 Nov 15; 26(22):4481-4493.

- **Martin-Morales L**, Rofes P, Diaz-Rubio E, Llovet P, Lorca V, Bando I, Perez-Segura P, de la Hoya M, Garre P, Garcia-Barberan V, Caldes T. Novel genetic mutations detected by multigene panel are associated with hereditary colorectal cancer predisposition. PLoS One. 2018 Sep 26; 13(9):e0203885.

Other unrelated publications:

- Llovet P, Illana FJ, **Martín-Morales L**, de la Hoya M, Pilar Garre Rubio, Ibáñez MD, Pérez-Segura P, Caldés T, García V. A novel TP53 germline inframe deletion identified in a Spanish series of Li-Fraumeni Syndrome suspected families. Fam Cancer. 2017 Oct; 16(4):567-575.

- Lorca V, Rueda D, **Martín-Morales L**, Poves C, Fernández-Aceñero MJ, Ruiz-Ponte C, Llovet P, Marrupe D, García-Barberán V, García-Paredes B, Pérez-Segura P, de la Hoya M, Díaz-Rubio E, Caldés T, Garre P. Role of GALNT12 in the genetic predisposition to attenuated adenomatous polyposis syndrome. PLoS One. 2017 Nov 2; 12(11):e0187312.

- Lorca V, Rueda D, **Martín-Morales L**, Fernández-Aceñero MJ, Grolleman J, Poves C, Llovet P, Tapial S, García-Barberán V, Sanz J, Pérez-Segura P, de Voer RM, Díaz-Rubio E, de la Hoya M, Caldés T, Garre P. Contribution of New Adenomatous Polyposis Predisposition Genes in an Unexplained Attenuated Spanish Cohort by Multigene Panel Testing. Sci Rep. 2019 Jul 8;9(1):9814.

Human Molecular Genetics, 2017, Vol. 26, No. 22 4481–4493

doi: 10.1093/hmg/ddx336
Advance Access Publication Date: 30 August 2017
Original Article

ORIGINAL ARTICLE

SETD6 dominant negative mutation in familial colorectal cancer type X

Lorena Martín-Morales[1], Michal Feldman[2,3], Zlata Vershinin[2,3], Pilar Garre[1], Trinidad Caldés[1,*] and Dan Levy[2,3,*]

[1]Molecular Oncology Laboratory, Department of Medical Oncology, Hospital Clínico San Carlos, IdISSC, CIBERONC, 28040 Madrid, Spain, [2]The Shraga Segal Department of Microbiology, Immunology and Genetics and [3]National Institute for Biotechnology in the Negev, Ben-Gurion University of the Negev, Be'er-Sheva 84105, Israel

*To whom correspondence should be addressed. Email: trinidad.caldes@salud.madrid.org (T.C.); or Email: ledan@post.bgu.ac.il (D.L.)

Abstract

Familiar colorectal cancer type X (FCCTX) comprises families that fulfill the Amsterdam criteria for hereditary non-polyposis colorectal cancer, but that lack the mismatch repair deficiency that defines the Lynch syndrome. Thus, the genetic cause that increases the predisposition to colorectal and other related cancers in families with FCCTX remains to be elucidated. Using whole-exome sequencing, we have identified a truncating mutation in the *SETD6* gene (c.791_792insA, p.Met264IlefsTer3) in all the affected members of a FCCTX family. SETD6 is a mono-methyltransferase previously shown to modulate the NF-κB and Wnt signaling pathways, among other. In the present study, we characterized the truncated version of SETD6, providing evidence that this SETD6 mutation may play a role in the cancer inheritance in this family. Here we demonstrate that the truncated SETD6 lacks its enzymatic activity as a methyltransferase, while maintaining other properties such as its expression, localization and substrate-binding ability. In addition, we show that the mutant allele is expressed and that the resulting protein competes with the wild type for their substrates, pointing to a dominant negative nature. These findings suggest that the identified mutation impairs the normal function of SETD6, which may result in the deregulation of the different pathways in which it is involved, contributing to the increased susceptibility to cancer in this FCCTX family.

Introduction

Colorectal cancer (CRC) is the third most common cancer and the second leading cancer-related cause of death in the world (1,2). It is estimated that familial risk is involved in up to 30% of all CRC cases (3,4), although not more than 5–6% are caused by known germline mutations in cancer-predisposing genes (5). Hereditary non-polyposis colorectal cancer (HNPCC) is the most common form of inherited CRC. HNPCC is a familial syndrome characterized by an increased incidence of CRC and other extracolonic tumors (6) that has been defined by the Amsterdam I and II clinical criteria (7,8). Approximately half of HNPCC families are referred to as Lynch syndrome, since they are explained by germline inactivating mutations in the mismatch repair (MMR) genes—including MLH1, MSH2, MSH6 and PMS2—or by a large deletion in EPCAM (located upstream of MSH2) (9,10). As a consequence, Lynch syndrome tumors lack the corresponding MMR proteins and fail to repair DNA through the mismatch repair pathway. This, in turn, causes microsatellite instability (MSI) and leads to the accumulation of somatic mutations. It is worth noting that although HNPCC and Lynch Syndrome used to be employed as synonyms, nowadays HNPCC is defined by the clinical criteria (Amsterdam I or II), while Lynch Syndrome refers to the families with MMR defects.

Received: August 6, 2017. **Revised:** August 6, 2017. **Accepted:** August 15, 2017

The other half of HNPCC families are MMR-proficient and present microsatellite-stable (MSS) tumors. These cases have been grouped under the term familial colorectal cancer type X (FCCTX), and the genetic basis underlying their cancer predisposition remains unknown (11). Several studies have reported that the tumors from FCCTX patients have different molecular and clinical features than both Lynch syndrome and sporadic CRC (12,13), suggesting the presence of other forms of genomic instability. FCCTX's tumors also show distinct gene expression profiles and deregulation of different canonical pathways, although some similarities have been described between FCCTX and MSS sporadic tumors (12,14,15). These alterations could result in the deregulation of genes involved in chromosomal segregation, genomic instability, apoptosis, proliferation, growth inhibition and migration (14,16). Nonetheless, FCCTX is a heterogeneous group of families, and we are still far from fully understanding the different events that may be involved in their tumor progression. In the same way, our knowledge regarding the genetic alterations that contribute to FCCTX's heredity is fairly limited. Although previous studies had identified a few genes involved in the cancer susceptibility of these FCCTX families (17–19), it was not until the arrival of Next-Generation Sequencing (NGS) that a larger amount of new cancer-predisposing genes are being discovered (20). In view of the published results, it seems that FCCTX does not form a single entity, since multiple different genes are thought to be involved in their cancer heritability (21). However, together they still explain the inheritance in only a small portion of these FCCTX families. Thus, the identification of new high-risk genes that contribute to the increased cancer susceptibility of FCCTX families remains to be a challenge and a priority.

Covalent post-translational modification (PTM) of proteins is key for the regulation of many biological processes (22). Among these modifications, lysine methylation plays a vital role in the regulation of many cellular signaling pathways (23,24). A lysine residue in a given protein can accept up to three methyl groups, forming mono-, di- and tri-methylated derivatives. The most studied is the methylation of histones, which can either suppress or activate gene transcription depending on which lysine is methylated (25). Similar to histones, non-histone proteins can be also targeted for methylation with diverse biological outcomes, such as changes in gene expression, protein stability and subcellular localization (26–34). Lysine methylation is catalyzed by protein lysine methyltransferases (PKMTs) that transfer a methyl group from a donor S-adenosyl-L-methionine (SAM) (35–39). Even though there are more than 70 PKMTs present in the human proteome, little is known about their enzymatic activity beyond histones.

SETD6 (SET domain-containing protein 6) is a member of the lysine methyltransferase family known to monomethylate RelA—a subunit of the transcription factor NF-κB—on lysine 310 (RelAK310me1). This methyl mark is recognized by GLP (a histone methyltransferase), which in turn methylates histone H3 on lysine 9, promoting a repressed chromatin state and thereby silencing the transcription of NF-κB target genes (40). In contrast, it was recently reported that SETD6 promotes RelA's transcriptional activity in bladder cancer (41). Additionally, SETD6 has been proven to participate in the canonical Wnt signaling cascade by forming a complex with PAK4 (p21-activated kinase 4) and the transcription factor β-catenin at the chromatin, leading to the activation of β-catenin target gene transcription (42). In recent studies, SETD6 has also been linked to the regulation of the nuclear hormone receptor signaling (43), embryonic stem cell differentiation (44) and oxidative stress response (45,46).

In the present study, we sequenced the whole exome from a FCCTX family with the aim of identifying new causative cancer-predisposing genetic variants. Among the different mutations that were detected, a rare frameshift mutation in the *SETD6* gene (resulting in a premature termination of the enzyme) was identified and further characterized, both in cell-free systems and in colon cancer cell line models. This mutation causes the loss of the C-terminal portion of SETD6, leaving an intact catalytic SET domain. Here we demonstrate that, while the truncated protein shows the same subcellular localization and substrate-binding properties as the wild type, it completely loses its catalytic activity. We also show that mutant SETD6 is expressed in the tumor and competes against the wild type in the ability to methylate its substrate, suggesting a dominant negative effect. Taken together, our findings support a model by which a dominant negative mutation in SETD6 may contribute to FCCTX's pathology.

Results

Whole-exome sequencing results from FCCTX family cc598

The whole exome was sequenced from two affected members and one healthy relative of a FCCTX family (II: 1, II: 2 and II: 3, Fig. 1A). The corresponding family—known as family cc598—fulfilled the Amsterdam I criteria (7,8), with four colorectal cancers affecting two successive generations and the earliest age at diagnosis being 34 years old. Supplementary Material, Table S1 shows in detail the clinical data from the different members of this family.

After rigorous filtering of the variants detected aiming to select shared rare heterozygote damaging variants, a number of candidate genes were prioritized. The candidate variants selected were: *SETD6* c.791_792insA (p.Met264IlefsTer3), *CCDC60* c.1558 C > T (p.Arg20*), *L3MBTL1* c.1125 A > G (p.Lys375=), *CCDC62* c.79 C > T (p.Arg27Trp) and *WDR33* c.3266 G > C (p.Arg1089Pro). All the candidate variants (detailed in Supplementary Material, Table S2) were present in the two affected members (II: 1 and II: 2) and absent in the healthy sibling (II: 3). However, among these, the most relevant changes in the protein were observed in *SETD6*.

Rare *SETD6* variant cosegregates with CRC within family cc598

The *SETD6* variant was a frameshift mutation known as c.791_792insA, p.Met264IlefsTer3, or just M264Ifs*3 (rs768456822, ENST00000310682), and it affected the three main protein-coding transcripts of this gene (ENST00000310682.6, ENST00000219315.8 and ENST00000394266.8). This mutation consisted of an insertion of an adenine between positions 791 and 792 of the cDNA and, like the other candidate variants, was carried by the two affected members sequenced (II: 1 and II: 2, with CRC diagnosed at ages 64 and 56, respectively), while not present in the healthy sibling (II: 3, 76 years old) (Fig. 1). In addition, the segregation study confirmed that the mutation was also present in II: 4 (CRC diagnosed at 34), while absent in two other healthy relatives (III: 1 and III: 2, 51 and 49 years old, respectively) (Fig. 1 and Supplementary Material, Fig. S2). The segregation for the other candidate variants is shown in Supplementary Material, Table S3. Loss of heterozygosity (LOH) analysis of *SETD6* c.791_792insA revealed that LOH did not occur in the tumor of II: 2 (Supplementary Material, Fig. S2B).

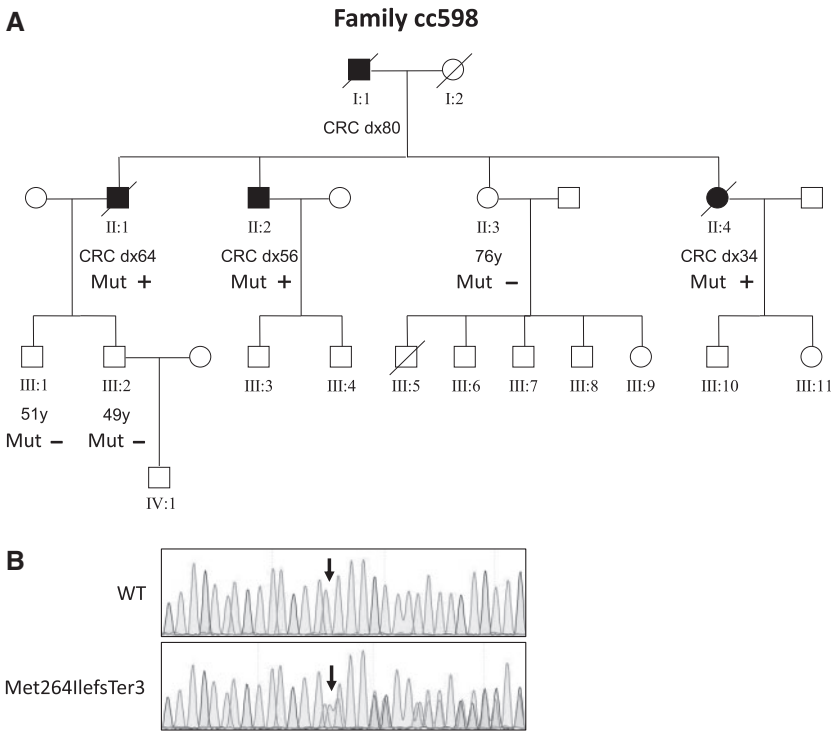


Figure 1. *SETD6* mutation c.791_792insA (p.M264Ifs*3) cosegregates with colorectal cancer in a FCCTX family. (**A**) Pedigree of family cc598 carrying the *SETD6* deleterious mutation c.791_792insA (p.Met264IlefsTer3). Whole-exome sequencing was done in family members II: 1, II: 2 and II: 3, and segregation was studied in members II: 4, III: 1 and III: 2. All affected participants were carriers (Mut+), while the healthy were non-carriers (Mut -). Colorectal cancer (CRC)-affected members are marked in black. The age at diagnosis (dx) or current age of healthy members is included beneath each individual (in years). (**B**) Electropherogram of the reverse wild-type and mutant sequence of the *SETD6* gene. The arrows show the point where the adenine is inserted.

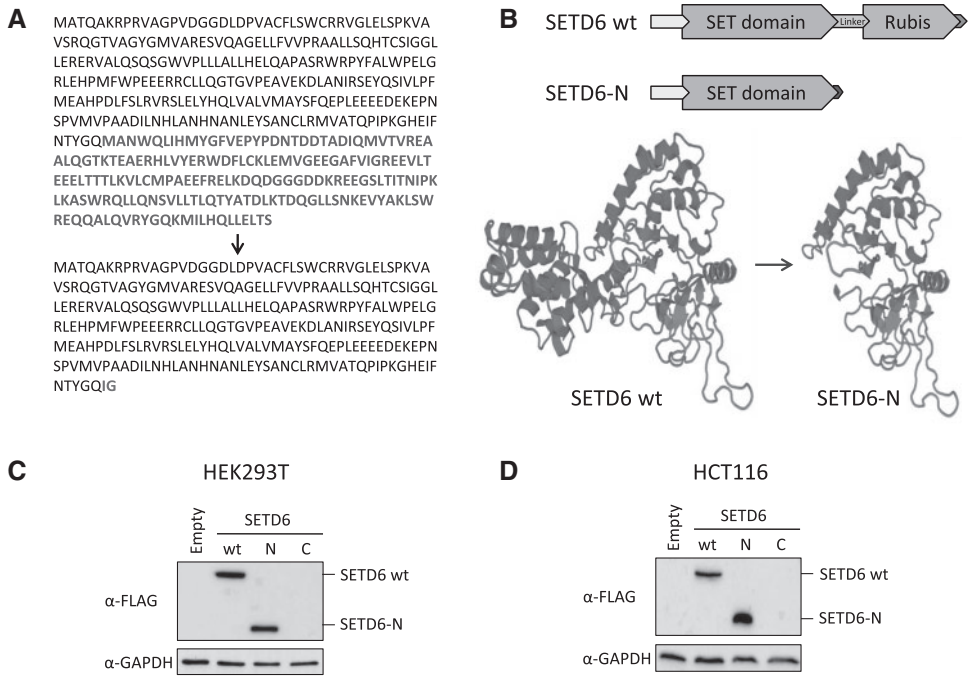


Figure 2. *SETD6* mutation causes the loss of the C-terminal half of the protein, but the mutant protein is still successfully overexpressed. (**A**) Amino acid sequence of SETD6 wt and SETD6-N with the changes marked in bold. (**B**) Representation of the different domains within the SETD6 protein showing the loss produced by the truncation (top), and the effect that the mutation is predicted to have on the 3D structure by SWISS-MODEL (53) (bottom). (**C, D**) Overexpression of FLAG-SETD6 (wt, N or C) in HEK293T (C) or HCT116 (D) cells. SETD6-N is the truncated version of the protein that mimics the mutation found in our family; SETD6-C is a truncated SETD6 that lacks the N-terminal part of the protein and the SET domain. The different versions of SETD6 were detected by western blot using an anti-FLAG antibody. GAPDH was used as a loading control.

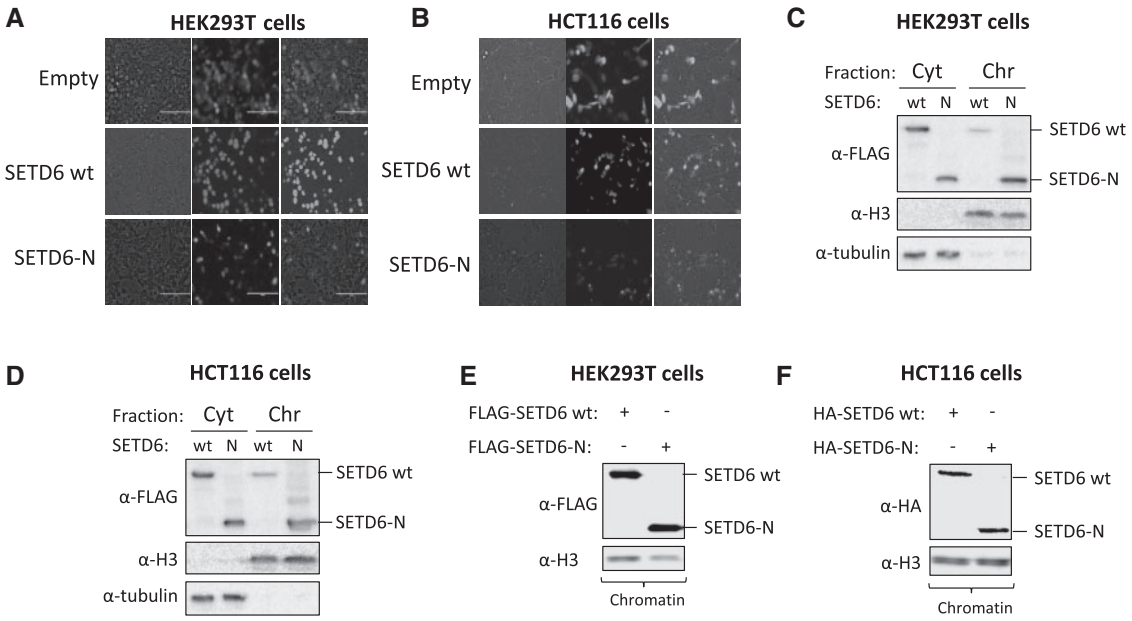


Figure 3. SETD6-N shows similar subcellular localization as wild-type SETD6. (**A, B**) HEK293T (A) and HCT116 (B) cells were transfected with GFP-tagged empty, SETD6 wt or SETD6-N plasmids. The distribution of GFP within the cells was observed by fluorescence microscopy. (**C, D**) HEK293T (C) or HCT116 (D) cells were transfected with FLAG-SETD6 (wt or N) and biochemically separated into cytoplasmic (Cyt) and chromatin (Chr) fractions. SETD6 was detected by western blot with an anti-FLAG antibody. Histone H3 and α-tubulin were used as chromatin and cytoplasmic controls, respectively. (**E, F**) The chromatin fraction of HEK293T (E) or HCT116 (D) cells transfected with FLAG-SETD6 (wt or N) was isolated and SETD6 was detected by western blot using the indicated antibodies. Histone H3 was used as a chromatin control.

On the other hand, although *SETD6* c.791_792insA (p.M264Ifs*3) is not a novel variant, it is quite rare in the general population, with a minor allele frequency (MAF) of 0.001285 according to ExAc and 0.00123 when the TCGA cohort is removed (47). Moreover, it is not found in the 1000 Genomes Project nor observed in homozygous state in neither of these databases. Finally, this variant produces a shift in the reading frame of the codons, and is therefore predicted to be Disease Causing by the MutationTaster software (48) (score = 1), further supporting its pathogenicity.

SETD6 c.791_792insA (p.M264Ifs*3) results in the loss of the C-terminal half of the protein

As mentioned above, the identified *SETD6* mutation (Fig. 1B) results in a frameshift with the consequent introduction of two new amino acids at positions 264 and 265 (ENST00000310682)—Ile264 and Gly265, instead of the original Met and Ala—followed by a premature stop codon (Fig. 2A). The resulting truncated SETD6 has an intact SET domain but lacks the C-terminal half of the protein, which includes a linker sequence and the whole Rubisco domain, presumably responsible of mediating protein-protein interactions (49) (Fig. 2B). With the aim of checking whether mutant SETD6 is stable and normally expressed, a truncated version of the protein mimicking the frameshift mutation—hereon referred to as SETD6-N—was cloned and overexpressed. SETD6-N was found to be expressed to the same extent as wild-type SETD6 both in HEK293T and HCT116 cells (Fig. 2C and D, respectively). It is worth noting that an attempt to overexpress the C-terminal half of SETD6 (containing everything which SETD6-N is missing) was unsuccessful. Our working hypothesis was that the premature termination of the protein would impair SETD6's cellular function.

SETD6-N shows the same localization pattern as wild-type SETD6

In order to test if the SETD6 mutation affects its subcellular localization, HEK293T and HCT116 cells were transfected with GFP-tagged SETD6 (either SETD6 wt or SETD6-N). Fluorescence microscopy of the transfected cells revealed that SETD6-N presented the same distribution within the cell as the wild type, both of which were mainly concentrated to the nucleus (Fig. 3A and B). We could detect, however, the presence of speckles inside the nucleus of the cells expressing mutant SETD6, which suggested that the association with the chromatin might be altered. To better understand the localization pattern, both proteins were transfected into cells followed by a biochemical fractionation of the cytosolic and chromatin fractions (50). This experiment confirmed a similar subcellular localization between the mutant and the wild-type, with no significant differences in the cytosolic fraction (Fig. 3C and D). Some differences between the wild-type and SETD6-N were only observed at the chromatin level. To further establish the presence of mutant SETD6 at chromatin, we used an additional method to extract the chromatin (see the Materials and Methods section for more details). This experiment confirmed comparable amounts of SETD6 wt and SETD6-N at this fraction (Fig. 3E and F). This data suggests that SETD6-N shows the same chromatin localization pattern as the wild-type enzyme.

Recombinant SETD6-N binds its substrates to the same extent as wild-type SETD6

We next performed an ELISA experiment to test whether the studied mutation would affect the binding ability of SETD6 in a cell-free *in vitro* system. To this end, we compared the direct binding of the wild type and SETD6-N to two well-known

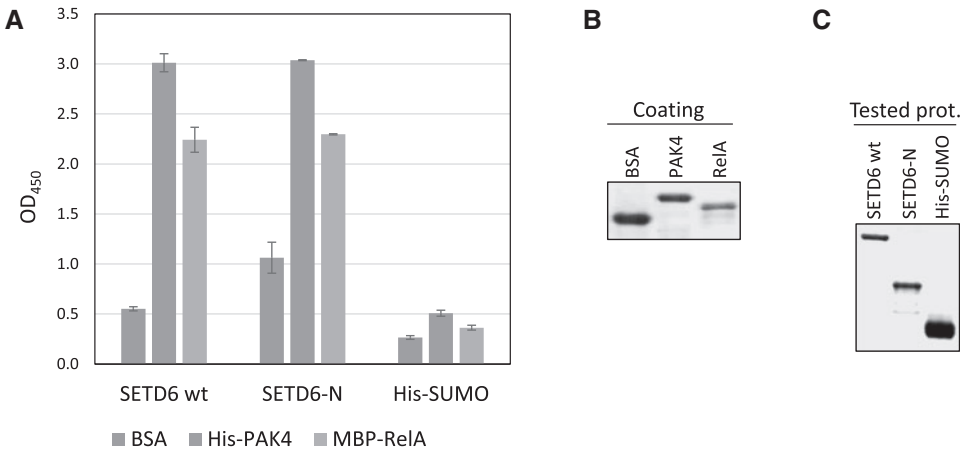


Figure 4. Recombinant SETD6-N shows similar binding to its substrates RelA and PAK4 as wild-type SETD6. (**A**) Interaction between recombinant SETD6 (wt or N) and PAK4/RelA determined by an ELISA. The plate was coated with 2 μg of His-PAK4, MBP-RelA or BSA (negative control), and then covered with 0.5 μg of His-SETD6 wt, His-SETD6-N or His-SUMO (negative control). Bound proteins were detected using a rabbit anti-SETD6 antibody. Data and error bars are from two technical replicates and represent two independent biological experiments. (**B, C**) Coomassie staining for the coating (B) and tested (C) proteins used in the ELISA.

substrates, RelA and PAK4 (40,42). Recombinant RelA and PAK4 were used to coat an ELISA plate, together with BSA, which was used as a negative control. Recombinant SETD6 wt, SETD6-N and SUMO were used as tested proteins, with the latter as another negative control. As shown in Figure 4A–C, SETD6-N bound at equal levels to both substrates as did wild-type SETD6, which suggests that the mutation does not impair the interaction between SETD6 and its substrates *in vitro*.

Recombinant SETD6-N fails to methylate its substrates

To compare the enzymatic activity of wild-type SETD6 and SETD6-N, both proteins were subjected to a cell-free *in vitro* methylation assay with RelA and PAK4 as substrates (Fig. 5A and B, respectively). As expected, both substrates were specifically methylated by wild-type SETD6. However, no methylation was observed when SETD6-N was present in the reaction instead. It is of note that SETD6's automethylation activity was also lost in the truncated mutant. Consistent with these results, we found that while the recombinant wild-type SETD6 methylated immunoprecipitated RelA and PAK4 from HEK293T cells, SETD6-N failed to do so (Fig. 5C and D, respectively). In a reciprocal experiment, we observed that recombinant RelA and PAK4 were specifically methylated by wild-type SETD6 isolated from human cells and not by SETD6-N (Fig. 5E and F, respectively). These complementary assays further demonstrate that the SETD6 truncating mutation identified in the FCCTX family abrogates the enzymatic activity of SETD6.

SETD6-N binds its substrates but loses its activity in colon cancer cells

Since the *SETD6* mutation was identified in hereditary colon cancer patients, we next aimed to confirm our findings in the colon cancer cell line HCT116. For this purpose, either wild-type SETD6 or SETD6-N were overexpressed in the presence or absence of FLAG-RelA (Fig. 6A). As expected, SETD6 wt and SETD6-N physically interacted with RelA at the chromatin to the same extent. Consistent with our cell-free *in vitro* experiments, we also observed that while SETD6 wt methylated RelA, SETD6-N did not. The methylation of RelA at the chromatin was identified using a RelAK310me1 antibody that could specifically recognize monomethylation of RelA at position K310 (40). The same results were obtained when PAK4 was used as the substrate (Fig. 6B). These findings support the dominant negative nature of *SETD6* c.791_792insA (p.M264Ifs*3) in a colon cancer cellular model.

Both wild-type and mutant SETD6 are expressed in the tumor and compete for their substrates

Last but not least, we aimed to check if SETD6's mutant allele was expressed, given that the variant was carried in heterozygosis. To this end, a digital PCR was carried out using two TaqMan probes that specifically recognized either the wild-type or the mutant cDNA (with the insertion of the adenine). As observed in Figure 7A and B, the tumor from member II: 2 showed positive expression of both alleles. As expected, all non-carrier controls only expressed the wild-type allele.

The fact that both alleles were expressed raised the question of whether mutant and wild-type SETD6 would compete for their substrates. In order to address this issue, a cell-free *in vitro* methylation competition assay was performed in the presence of recombinant SETD6 wt and different amounts of recombinant SETD6-N, using RelA as the substrate. Figure 7C shows how wild-type SETD6's activity was inhibited in the presence of SETD6-N in a concentration-dependent manner, with a significant reduction in the amount of methylated substrate even when the same amount of each form of the enzyme was present (lanes 2 and 3, respectively). Finally, a competition assay in HCT116 cells confirmed the cell-free *in vitro* results, showing that the methylation of RelA at K310 in cells is reduced upon SETD6-N overexpression in a dose-dependent manner (Fig. 7D). Taken together, these results suggest that the expression of the mutant allele in the carriers is expected to compete with the enzymatic activity of SETD6 wt, supporting a dominant negative role of this mutation.

Discussion

Familial colorectal cancer type X (FCCTX) is a term used to describe a heterogeneous group of CRC families for whom the genetic basis underlying their cancer predisposition remains unknown. Genome-wide analyses of gene expression patterns

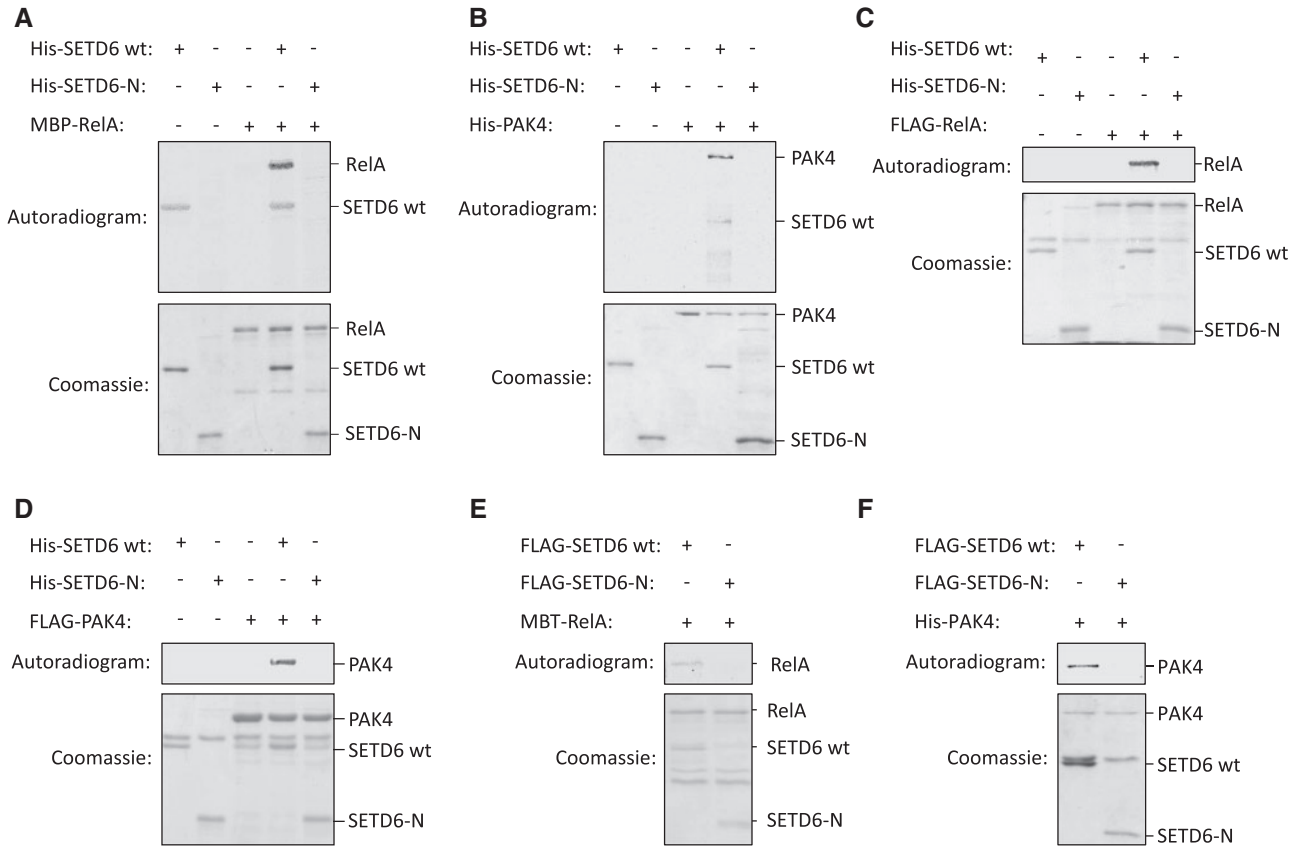


Figure 5. SETD6-N fails to methylate its substrates RelA and PAK4. (**A, B**) Cell-free *in vitro* methylation assays in the presence of ^{3}H-SAM, recombinant His-SETD6 (wt or N) and either MBP-RelA (A) or His-PAK4 (B). The methylated proteins were detected by autoradiogram (top images), and the input used in the reactions is shown by Coomassie staining (bottom). (**C, D**) HEK293T cells were transfected with FLAG-RelA (C), FLAG-PAK4 (D) or empty plasmid. Cell extracts were immunoprecipitated with anti-FLAG M2 beads, followed by a radioactive cell-free *in vitro* methylation assay in the presence of ^{3}H-SAM and recombinant His-SETD6 (wt or N). The methylated proteins were detected by autoradiogram (top images), and the input used in the reactions was shown by Coomassie staining (bottom). (**E, F**) HCT116 cells were transfected with either FLAG-SETD6 wt, FLAG-SETD6-N or empty plasmid. Cell extracts were immunoprecipitated with anti-FLAG M2 beads, followed by a radioactive cell-free *in vitro* methylation assay in the presence of ^{3}H-SAM and recombinant MBP-RelA (E) or His-PAK4 (F). The methylated proteins were detected by autoradiogram (top images), and the input used in the reactions was shown by Coomassie staining (bottom).

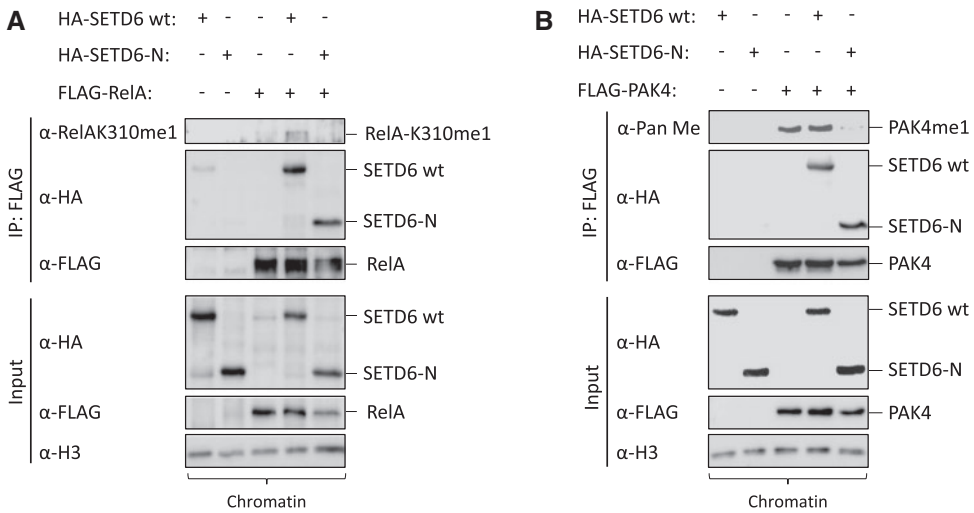


Figure 6. SETD6-N binds its substrates RelA and PAK4, while failing to methylate them, in the colon cancer cell line HCT116. (**A, B**) HCT116 cells were transfected with either HA-SETD6 wt or HA-SETD6-N plasmids, in the absence or presence of FLAG-RelA (A) or FLAG-PAK4 (B). The chromatin fraction was then immunoprecipitated with anti-FLAG M2 beads and analysed by western blot with the indicated antibodies.

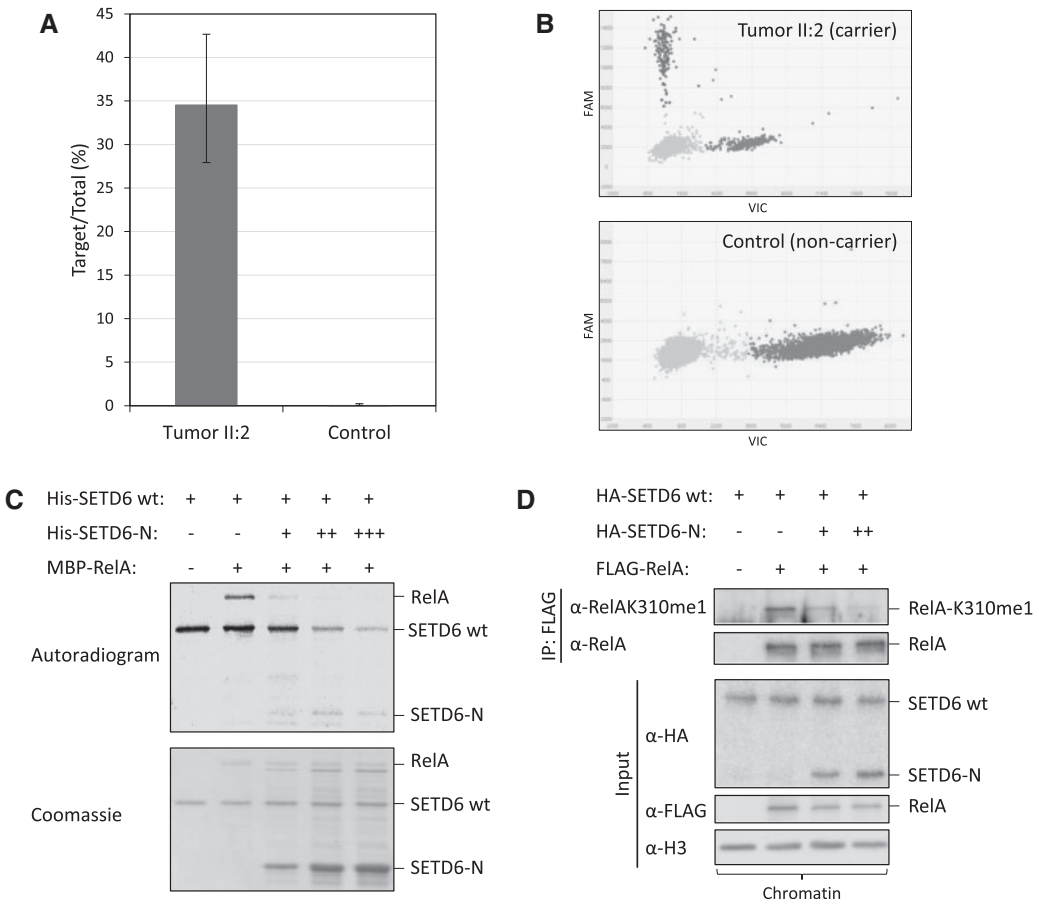


Figure 7. Both wild-type and mutant SETD6 are expressed in the tumor and compete for their substrates. (**A**) Allele-specific expression obtained by digital PCR presented as Target/Total, where "Target" is the mutant SETD6 allele (detected with the FAM dye). Data from the tumor of II:2 was collected from three independent experiments, and the error bars correspond to the confidence intervals. The Control represents tumor cDNA from four different sporadic CRC patients used as non-carrier controls. (**B**) Digital PCR visualization of the cDNA from the tumor of member II:2 (top) and a non-carrier control (bottom). The FAM dye detects the mutant allele, while the VIC dye detects the wild-type allele. (**C**) Radioactive cell-free *in vitro* methylation assay in the presence of recombinant wild-type SETD6 and different amounts (+ and ++) of recombinant SETD6-N. The methylated proteins were detected by autoradiogram (top), and the input used in the reactions is shown by Coomassie staining (bottom). (**D**) HCT116 cells were transfected with HA-SETD6 wt, FLAG-RelA and with or without different concentrations (+ and ++) of HA-SETD6-N. The chromatin fraction was then immunoprecipitated with anti-FLAG M2 beads and analyzed by western blot with the indicated antibodies.

in FCCTX in comparison to Lynch syndrome tumors (15) have shown differences in three major cancer-related pathways: cell cycle progression, oxidative phosphorylation and G protein-coupled receptor signaling, all of which have been previously linked to CRC (51–53). The fact that different genes are expressed in the tumors of FCCTX and Lynch syndrome patients may suggest that different molecular mechanisms mediate the progression of these two pathologies. FCCTX presumably comprises different yet-to-be-discovered genetic syndromes involving high-penetrance mutations in novel cancer-predisposing genes. However, it is thought that some FCCTX cases would result from a combination of low-penetrance mutations in different genes, or even from aggregation of sporadic cases due to shared lifestyle factors (5), what makes their study more difficult. The arrival of Next-Generation Sequencing (NGS) has been an important milestone in the search for new predisposition genes that explain the cancer heritability in these families. Nevertheless, this task is still a challenge and it is believed that the best strategy is to address each family individually, due to their wide diversity.

Using whole-exome sequencing, a rare frameshift mutation in the *SETD6* gene was identified in a FCCTX family. This mutation was found to cosegregate with the disease within the family, since it was carried by three CRC-affected members (II: 1, II: 2 and II: 4) while absent in three healthy relatives (II: 3, III: 1 and III: 2). No LOH was observed in the tumor of member II: 2, suggesting that in this family the *SETD6* mutation does not follow Knudson's two-hit mechanism for tumor suppressor genes, consistent with a dominant negative nature. Nonetheless, the LOH does not predict the pathogenicity of a mutation, since it has been reported to have a dual role in HNPCC families, with no preferential loss of the wild-type or the mutant allele. As a matter of fact, different carriers of well-known pathogenic MMR mutations have been reported to show the three different LOH statuses (no LOH, LOH of the mutant and LOH of the wild-type), even within the same family (54).

Regarding the consequences of the studied *SETD6* mutation, this variant results in the introduction of two amino acids at positions 264 and 265 followed by a premature stop codon. The truncated protein (SETD6-N) lacks its C-terminal half but retains an intact SET domain, which is responsible for its catalytic activity. Here, we show that mutant SETD6 exhibits dominant negative properties in cell-free *in vitro* systems and in a colon cancer *in vitro* cell line model. Although mutant SETD6 displays

similar localization, expression and binding to its known partners as the wild-type, the mutant protein loses its enzymatic activity. Indeed, unlike wild-type SETD6, SETD6-N lacks both its automethylation activity and the ability to methylate two previously identified substrates, PAK4 and RelA (40,42). In addition, the two alleles (wild type and mutant) were found to be expressed in the tumor of one of the carriers, and the two forms of SETD6 were shown to compete for their substrates both in cell-free systems and in a colon cancer cellular model, pointing to a dominant negative role. As a result, this mutation may have several downstream effects on the different pathways in which SETD6 is involved. For example, SETD6-dependent methylation of the NF-κB subunit RelA has been shown to be critical for basal inhibition of NF-κB signaling in the absence of stimulation (40). The NF-κB family of transcription factors has an essential role in inflammation and innate immunity, but it has also been increasingly recognized as a crucial player in many steps of cancer initiation and progression (55). In fact, activation of the NF-κB pathway has been positively associated with multiple types of cancer, including CRC (56). In addition, we had previously demonstrated that methylation of PAK4 by SETD6 promotes the activation of the Wnt/β-catenin pathway (42). On the other hand, deregulation of the Wnt signaling pathway has been shown to contribute to CRC development, including in HNPCC (57–60). Thus, the loss of SETD6's function, together with the aberrant regulation of NF-κB and/or Wnt signaling, could contribute to the initiation or progression of CRC in the studied FCCTX family. However, since SETD6 has several other substrates and is involved in numerous other signaling pathways (41,43,45,46), future studies are needed in order to define the downstream consequences of this SETD6 truncating mutation that we have identified.

Given all the results presented above, we propose that the presence of SETD6's mutant allele would presumably increase the cumulative risk of developing colorectal cancer throughout the life of the carriers, as compared with the general population. More research should be done in order to determine the penetrance of this mutation (that is to say the percentage of individuals who present this variant that will develop the disease, which is not complete even in well-known pathogenic MMR variants), whether it is a high, moderate or low-risk allele, and if there is also an association with other CRC-related cancers. Interestingly, the same SETD6 frameshift mutation had already been proposed to be associated with ovarian cancer in an extensive study by Kanchi and colleagues (61). Nonetheless, the search for new genes by exome sequencing in FCCTX families has demonstrated that more than one gene can be implicated in their increased risk of developing cancer. Actually, even when there is a high-risk gene involved, there might be also low-penetrance alleles that cooperate in the process as risk modifiers. Apart from these polygenic approaches, there are many other factors that might take part in modulating the cancer risk, including lifestyle and other environmental factors. For all these reasons, our results are not enough to claim that SETD6 alone is responsible for the increased CRC-risk in this FCCTX family. Hence, it is still difficult to know the effect that this mutation might have in the other carriers found in the public databases, as much as it is hard to predict if they are going to develop cancer at some point in their life or even if they have already developed it.

Noteworthy, the whole-exome sequencing followed by rigorous filtering identified additional candidate genes beside *SETD6* (Supplementary Material, Table S2). For instance, *CCDC60* showed a stop gain variant that results in the loss of 30 amino acids, 20 of which belong to a domain of unknown function, while the remaining 10 belong to a low complexity region. Another candidate was *L3MBTL1*, located on the long arm of chromosome 20 within a region that has been previously shown to be deleted in several malignancies (62). This gene presented a splicing variant that does not imply the gain or loss of a splicing acceptor/donor, but that is predicted to create a new exonic splicing silencer site instead. Finally, missense mutations were found in *CCDC62*, previously linked to prostate cancer (63), and in the polyadenylation regulating factor *WDR33* (64). Both variants affect low complexity regions of the corresponding proteins, and although they are predicted to be damaging *in silico*, they do not alter any known protein domains. While the current paper focuses on *SETD6*, we cannot exclude the possibility that the additional candidate variants identified may contribute independently or together to the pathology of FCCTX.

Together, our findings suggest that a truncating dominant negative mutation in SETD6 could potentially explain the cancer predisposition of this FCCTX family. These results certainly point to a pathogenic role of *SETD6* c.791_792insA (p.M264Ifs*3), though not enough to prove that *SETD6* alone is responsible for their increased cancer risk. Although no other *SETD6* variants were found in the remaining 12 families that were studied, nor in 109 familial CRC cases provided by Dr. Castellvi-Bel (Hospital Clinic, IDIBAPS, Barcelona, Spain), the screening of this gene in a larger group of patients could provide more insights into its role in other FCCTX families.

Materials and Methods

Study population

The studied family (cc598) was selected among a group of FCCTX families collected in the Genetic Counseling Unit at the Hospital Clínico San Carlos of Madrid. All FCCTX families fulfilled the Amsterdam I or II criteria for HNPCC (7) (at least three relatives with CRC or other related cancers, one being a first degree relative of the other two, with at least two successive generations involved, the earliest age of onset being 50 years old or lower and familial adenomatous polyposis excluded). In addition, all the CRC tumors from these families were MSS and presented normal expression of the MMR proteins. The study was approved by the Institutional Review Boards of the Hospital Clínico San Carlos, and an informed consent was obtained from each participant. Personal and family histories were obtained from the proband and participating relatives, and cancer diagnoses were confirmed by medical and pathology records.

Family cc598

Family cc598 (Fig. 1A), selected for the whole-exome study, is an Amsterdam I family in which the father (I: 1) was diagnosed with CRC at the age of 80. He had two daughters and two sons, three of whom were diagnosed with CRC at ages 64 (II: 1), 56 (II: 2) and 34 (II: 4). Only one daughter was healthy (II: 3). The stage and location of the CRCs developed by II: 1, II: 2 and II: 4 were pT3N0M1 (splenic flexure, left colon), pT2N0M0 (right colon) and pT3N1M0 (rectum), respectively. The tumors from members II: 1 and II: 4 were MSS and showed normal expression of the MMR proteins (MLH1, MSH2, MSH6 and PMS2) (Supplementary Material, Table S1). The whole-exome sequencing was performed in germline DNA from members II: 1, II: 2 and II: 3.

DNA and RNA extraction

Germline DNA and RNA were extracted from peripheral blood using the MagNA Pure Compact extractor system (Roche Diagnostics), according to the manufacturer's protocol. The PAXgene Blood RNA Kit (PreAnalytiX) was used to extract germline RNA conforming to its manual when the patient could not come to our hospital. Tumor DNA and RNA were extracted from 7 μm-thick haematoxylin/eosin-stained sections of the paraffin-embedded tumor tissues with a tumor content of more than 80% as determined by two experienced pathologists. Tumor DNA was extracted using the QIAamp DNA FFPE Tissue Kit from Qiagen, while tumor RNA was isolated employing the RNeasy FFPE Kit (Qiagen), according to their corresponding protocols. A NanoDrop® (ND1000) spectrophotometer was used to assess the DNA quantity and quality.

Whole-exome sequencing

The whole-exome sequencing was outsourced to Sistemas Genómicos®. Library preparation for the capture of selected DNA regions was performed according to Agilent's SureSelect protocol for Illumina paired-end sequencing (SureSelectXT Human All Exon V3, 51 Mb, Agilent Technologies). The final library size and concentration were determined on an Agilent 2100 Bioanalyzer and a Qubit Fluorometer (Thermo Fisher Scientific), respectively. Finally, the library was sequenced on an Illumina HiSeq 2000 platform with paired-end reads of 101 bp, following the manufacturer's protocol. Images produced by the HiSeq 2000 were processed using the manufacturer's software to generate FASTQ sequence files. Reads were aligned against the human reference genome version GRCh37/hg19 using the BWA software, creating the BAM files. Low quality reads, PCR duplicates and other sequences that could introduce major biases were removed using Picard-tools (http://picard.sourceforge.net/) and SAMtools (65). Variant calling was performed using a combination of two different algorithms [VarScan (66) and GATK (67)] and the identified variants were annotated using the HGMD (68) and Ensembl (69) databases.

Filtering and prioritization of the variants

The variants identified by whole-exome sequencing where subsequently filtered as follows: 1. Variants shared by the two affected members but not by the healthy relative were selected. 2. Homozygous variants were discarded, as well as variants present in allosomes. 3. Only coding non-synonymous (missense, stop gain, stop loss, in-frame, frameshift) and splicing variants were selected. 4. Variants with a minor allele frequency (MAF) in the general population greater than 0.01 were eliminated. 5. Missense and in-frame variants not predicted to be damaging by *in silico* programs [SIFT (70), Polyphen (71), MutationTaster (48)] were discarded, as well as splicing variants not directly affecting the donor/acceptor sites nor predicted to alter the splicing by the Human Splicing Finder (HSF) (72). Finally, the filtered variants were prioritized according to the genes and pathways involved.

Variant validation, segregation and loss of heterozygosity (LOH) studies

All the candidate variants were validated by PCR followed by Sanger sequencing of the corresponding region of each gene, using specific primers that were designed with Primer3 (73). The segregation and LOH studies for the *SETD6* variant were also assessed by PCR and subsequent Sanger sequencing of the selected area of the *SETD6* gene (exon 7, ENST00000310682). The segregation study was carried out in germline DNA from the available members of the family (III: 1 and III: 2). However, although no germline DNA was available from the deceased member II: 4, we were able to study the segregation in this member extracting DNA from the paraffin-embedded tumor.

For the LOH analysis, tumor DNA was extracted from the paraffin-embedded tumor available, and the electropherograms of the germline and tumor sequences were compared, allowing the discrimination of the wild-type and the mutant alleles. LOH was considered when the intensity of any allele was reduced by $\geq$50% relative to the other allele. The *SETD6* primers used for the PCRs were CCACTCAGCCCATTCCTAAA (forward) and TGATACACTCACCCTGTAATGCT (reverse).

Plasmids and cloning

The wild-type *SETD6* gene, as well as mutant *SETD6* (with the same variant identified in family cc598), was amplified by high-fidelity PCR and cloned into pcDNA3.1 and pET-Duet plasmids for overexpression and protein purification, respectively. The pcDNA3.1 plasmids in which the two forms of the gene were subcloned include pcDNA3.1-FLAG, pcDNA3.1-HA and pcDNA3.1-GFP, for the different experiments. In the same way, RelA and PAK4 were also cloned into pcDNA3.1-FLAG for the overexpression experiments (42) and pMAL-c2x or pET-Duet plasmids (respectively) for the expression and purification of the recombinant proteins.

Cell lines and transfection

Two different cell lines were used in this study: human embryonic kidney cells (HEK293T) and human colon carcinoma cells (HCT116). Both were maintained in Dulbecco's modified Eagle's medium (Sigma-Aldrich, D5671) with 10% fetal bovine serum (Gibco), 2 mg/ml L-glutamine (Sigma-Aldrich, G7513), 1% penicillin-streptomycin (Sigma, P0781) and non-essential amino acids (Sigma-Aldrich, M7145), and they were cultured at 37 °C in a humidified incubator with 5% CO_2. For transient transfection, Mirus transfection reagents (*Trans*IT®-LT1 for HEK293T cells and *Trans*IT®-X2 for HCT116 cells) were used according to the manufacturer's instructions, together with Opti-MEM serum-free medium (Gibco).

Western blot analysis

For western blot analyses, cells were homogenized and lysed in RIPA buffer [50 mM Tris-HCl, pH 8, 150 mM NaCl, 1% Nonidet P-40, 0.5% sodium deoxycholate, 0.1% SDS, 1 mM DTT, and 1: 100 protease inhibitor mixture (Sigma-Aldrich)], except for the biochemical fractionation and chromatin immunoprecipitation experiments, in which the cells were lysed as described below. Samples were heated at 95 °C for 5 min in Laemmli sample buffer, run on a 8-12% SDS-PAGE electrophoresis gel, and then transferred to a polyvinylidene difluoride (PVDF) membrane. Membranes were blocked with either 10% skim milk in PBST or 5% BSA in TBST for 1 h on a shaking platform, and subsequently incubated with primary antibody for another hour with agitation. After three washes with the corresponding buffer (PBST or TBST), a 30-min incubation with HRP-conjugated secondary antibody and three additional washes, a 2-min reaction with a

chemiluminiscent substrate (EZ-ECL, Biological industries, 20-500-120) allowed the visualization of the proteins.

The mouse monoclonal antibodies used were: anti-FLAG M2 (Sigma-Aldrich, F1804), anti-HA (Millipore, 05–904), anti-GAPDH (Abcam, ab8245), anti-SETD6 (Genetex, GTX629891), and anti-histone H3 (Abcam, ab10799). The rabbit polyclonal antibodies used were: an HRP-conjugated pan methyl lysine antibody (ImmuneChem, ICP0502) and a specific antibody against mono-methylated RelA-Lys310 developed by Levy *et al.* (40). HRP-conjugated secondary antibodies (goat anti-rabbit and goat anti-mouse) were purchased from Jackson ImmunoResearch (111-035-144 and 115-035-062, respectively). Antibodies were diluted and prepared in PBST with 10% skim milk or in TBST with 5% BSA, according to the manufacturer's recommendations.

Biochemical fractionation

Biochemical fractionation was performed as previously described by Mendez *et al.* (50), with the addition of a final resuspension of the chromatin pellet for 30 min on ice in RIPA buffer with 1 mM $MgCl_2$ and benzonase nuclease enzyme (Sigma-Aldrich). The chromatin fraction was obtained by the collection of the supernatant after low-speed centrifugation (5 min, 1700 g, 4 °C).

Recombinant protein expression and purification

The *Escherichia coli* BL21-derivative Rosetta host strain was transformed with pET-Duet plasmids containing the gene of interest (*SETD6* wt, *SETD6-N*, *PAK4* or *RelA*) and grown overnight in 3 ml LB medium +100 μg/ml ampicillin (37 °C, 220 rpm). The culture was then expanded to 100 ml LB medium and incubated at 37 °C until the absorbance (OD) reached values of 0.6–0.8, when it was induced with 1: 10000 IPTG and left overnight at 18 °C and 220 rpm. After IPTG induction, the bacteria were harvested by centrifugation (10 min, 12000 rpm, 4 °C), resuspended in cold lysis buffer (10 mM imidazole, 1% PMSF, protease inhibitor cocktail, 0.1% triton and PBS) and then lysed by sonication on ice (25% amplitude, 1 min 30 s, 10 s on/5 s off). Finally, the lysate was centrifuged (20 min, 4 °C, 18000 rpm) and filtered, and the His-tagged proteins were purified using an ÄKTA™ column.

ELISA

A high-binding 96-well polystyrene microplate (Greiner Bio-One MICROLON®) was coated with 2 μg of the recombinant proteins of interest (His-PAK4, MBP-RelA or BSA) diluted in PBS. The plate was blocked with 3% BSA in PBST and subsequently covered with 0.5 μg of the recombinant tested proteins (His-SETD6 wt, His-SETD6-N or His-SUMO as a control) diluted in 1% BSA in PBST. The rabbit polyclonal anti-SETD6 primary antibody (40) was then added, followed by incubation with an HRP-conjugated secondary antibody (goat anti-rabbit, 1: 2000; Jackson ImmunoResearch, 111-035-144). All the incubation steps were performed at room temperature with vigorous agitation for 1 h, and followed by three washes with PBST. After the final washes, 100 μl of TMB reagent were added to each well, succeeded after a few minutes by the same amount of 1 N H_2SO_4, in order to stop the reaction. The absorbance at 450 nm was then detected using an Infinite® M200 plate reader (Tecan). All samples were analysed in duplicates.

Cell-free *in vitro* methylation assay

Cell-free *in vitro* methylation reactions with recombinant proteins took place in a final volume of 25 μl, containing 4 μg of substrate (His-PAK4 or MBP-RelA), 4 μg (or increasing amounts for the competition assay) of His-SETD6 (either wt or N), 2 mCi of ^{3}H-labeled S-adenosyl-methionine (SAM) (AdoMet, Perkin-Elmer) and PKMT buffer (20 mM Tris-HCl, pH 8, 10% glycerol, 20 mM KCl, 5 mM $MgCl_2$). The reaction tubes were incubated overnight at 30 °C and then resolved by SDS-PAGE electrophoresis and subsequent autoradiogram. For the immunoprecipitation followed by cell-free *in vitro* methylation, human cells were transfected with empty, FLAG-SETD6 wt, FLAG-SETD6-N, FLAG-RelA or FLAG-PAK4 pcDNA3.1 plasmids, and 24 h post-transfection they were lysed with RIPA buffer and pulled down overnight with anti-FLAG M2 Affinity gel beads (Sigma-Aldrich, A2220) on a rotor at 4 °C. After two washes in RIPA buffer and another two in PKMT buffer, samples were subjected to an on-beads cell-free *in vitro* methylation assay as described above.

Protein–protein chromatin immunoprecipitation

Protein-protein chromatin immunoprecipitation was modified from a published protocol (74). After cross-linking, cells were harvested and washed twice with PBS and then lysed in 1 ml of lysis buffer (20 mM Tris-HCl, pH 8, 85 mM KCl, 0.5% Nonidet P-40, and 1: 100 protease inhibitor mixture) for 10 min on ice. Nuclear pellets were resuspended in 200 μl of nuclei lysis buffer (50 mM Tris-HCl pH 8, 10 mM EDTA, 1% SDS, 1: 100 protease inhibitor mixture) for 10 min on ice, and sonicated (Bioruptor, Diagenode) with high power settings for three cycles of 6 min each (30 s on/off). Samples were then centrifuged (20 min, 13000 rpm, 4 °C), and the soluble chromatin fraction was collected. The FLAG-labeled substrates present in the soluble chromatin were then immunoprecipitated overnight on a rotor at 4 °C, using 20 μl per tube of anti-FLAG M2 Affinity gel beads (Sigma-Aldrich, A2220). The beads were then washed according to the published protocol, heated for 30 min in Laemmli sample buffer at 95 °C, and resolved on 10-12% SDS-PAGE gels followed by western blot analyses.

Reverse transcription polymerase chain reaction and digital PCR

Reverse transcription polymerase chain reaction (RT-PCR) was performed to convert RNA into complementary DNA (cDNA) using the PrimeScript RT Reagent Kit (Perfect Real Time) (Takara, Clontech), following the kit's instructions. For the expression assays, a TaqMan digital PCR (dPCR) was carried out taking advantage of the QuantStudio™ 3 D Digital PCR System (Thermo Fisher Scientific), according to the manufacturer's recommendations. The primers and TaqMan probes used in the dPCR were designed with the Custom TaqMan® Assay Design Tool (Thermo Fisher Scientific) and produced by the same company. The FAM probe specifically identified the mutant transcript, while the VIC probe only recognized the wild type. The dPCR was used to analyse the cDNA from the paraffin-embedded tumor of member II: 2. Tumor cDNA from four sporadic CRC patients were used as non-carrier controls.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We thank all the members from Levy's lab for their technical assistance, and Ruth Tennen and Ramon Birnbaum for critical reading of the manuscript. We also thank Sergi Calstellvi-Bel and Sebastia Franch-Exposito (Hospital Clinic, IDIBAPS, Barcelona, Spain) for the screening of *SETD6* in their familial CRC cohort.

Conflict of Interest statement. None declared.

Funding

This work was supported by grants to DL from The Israel Science Foundation (285/14), The Research Career Development Award from the Israel Cancer Research Fund, Marie Curie Career Integration Grant and from the Israel Cancer Association. TC's and LMM's work was supported by grants PI-13/02588, PI-16/01292 and CIBERONC from the Carlos III Health Institute (Spain) and the European Regional Development FEDER funds. LMM's scientific stay in Ben-Gurion University was supported by the Federation of European Biochemical Societies (FEBS).

References

1. Bray, F., Ferlay, J., Laversanne, M., Brewster, D.H., Gombe Mbalawa, C., Kohler, B., Pineros, M., Steliarova-Foucher, E., Swaminathan, R., Antoni, S. *et al*. (2015) Cancer incidence in five continents: inclusion criteria, highlights from Volume X and the global status of cancer registration. *Int. J. Cancer*, **137**, 2060–2071.
2. Steliarova-Foucher, E., O'Callaghan, M., Ferlay, J., Masuyer, E., Rosso, S., Forman, D., Bray, F. and Comber, H. (2015) The European Cancer Observatory: A new data resource. *Eur. J. Cancer*, **51**, 1131–1143.
3. Grady, W.M. (2003) Genetic testing for high-risk colon cancer patients. *Gastroenterology*, **124**, 1574–1594.
4. Lichtenstein, P., Holm, N.V., Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A. and Hemminki, K. (2000) Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl J. Med.*, **343**, 78–85.
5. Stoffel, E.M. and Kastrinos, F. (2014) Familial colorectal cancer, beyond Lynch syndrome. *Clin. Gastroenterol. Hepatol.*, **12**, 1059–1068.
6. Watson, P. and Lynch, H.T. (1993) Extracolonic cancer in hereditary nonpolyposis colorectal cancer. *Cancer*, **71**, 677–685.
7. Vasen, H.F., Mecklin, J.P., Khan, P.M. and Lynch, H.T. (1991) The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). *Dis. Colon Rectum.*, **34**, 424–425.
8. Vasen, H.F., Watson, P., Mecklin, J.P. and 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**, 1453–1456.
9. Lynch, H.T. and de la Chapelle, A. (1999) Genetic susceptibility to non-polyposis colorectal cancer. *J. Med. Genet.*, **36**, 801–818.
10. Lynch, H.T., Riegert-Johnson, D.L., Snyder, C., Lynch, J.F., Hagenkord, J., Boland, C.R., Rhees, J., Thibodeau, S.N., Boardman, L.A., Davies, J. *et al*. (2011) Lynch syndrome-associated extracolonic tumors are rare in two extended families with the same EPCAM deletion. *Am. J. Gastroenterol.*, **106**, 1829–1836.
11. Lindor, N.M., Rabe, K., Petersen, G.M., Haile, R., Casey, G., Baron, J., Gallinger, S., Bapat, B., Aronson, M., Hopper, J. *et al*. (2005) Lower cancer incidence in Amsterdam-I criteria families without mismatch repair deficiency: familial colorectal cancer type X. *JAMA*, **293**, 1979–1985.
12. Sanchez-de-Abajo, A., de la Hoya, M., van Puijenbroek, M., Tosar, A., Lopez-Asenjo, J.A., Diaz-Rubio, E., Morreau, H. and Caldes, T. (2007) Molecular analysis of colorectal cancer tumors from patients with mismatch repair proficient hereditary nonpolyposis colorectal cancer suggests novel carcinogenic pathways. *Clin. Cancer Res.*, **13**, 5729–5735.
13. Garre, P., Martin, L., Bando, I., Tosar, A., Llovet, P., Sanz, J., Romero, A., de la Hoya, M., Diaz-Rubio, E. and Caldes, T. (2014) Cancer risk and overall survival in mismatch repair proficient hereditary non-polyposis colorectal cancer, Lynch syndrome and sporadic colorectal cancer. *Fam. Cancer*, **13**, 109–119.
14. Dominguez-Valentin, M., Therkildsen, C., Da Silva, S. and Nilbert, M. (2015) Familial colorectal cancer type X: genetic profiles and phenotypic features. *Mod. Pathol.*, **28**, 30–36.
15. Dominguez-Valentin, M., Therkildsen, C., Veerla, S., Jonsson, M., Bernstein, I., Borg, A. and Nilbert, M. (2013) Distinct gene expression signatures in lynch syndrome and familial colorectal cancer type x. *PLoS One*, **8**, e71755.
16. Therkildsen, C., Jonsson, G., Dominguez-Valentin, M., Nissen, A., Rambech, E., Halvarsson, B., Bernstein, I., Borg, K. and Nilbert, M. (2013) Gain of chromosomal region 20q and loss of 18 discriminates between Lynch syndrome and familial colorectal cancer. *Eur. J. Cancer*, **49**, 1226–1235.
17. Garre, P., Briceno, V., Xicola, R.M., Doyle, B.J., de la Hoya, M., Sanz, J., Llovet, P., Pescador, P., Puente, J., Diaz-Rubio, E. *et al*. (2011) Analysis of the oxidative damage repair genes NUDT1, OGG1, and MUTYH in patients from mismatch repair proficient HNPCC families (MSS-HNPCC). *Clin. Cancer Res.*, **17**, 1701–1712.
18. Garre, P., Martin, L., Sanz, J., Romero, A., Tosar, A., Bando, I., Llovet, P., Diaque, P., Garcia-Paredes, B., Diaz-Rubio, E. *et al*. (2015) BRCA2 gene: a candidate for clinical testing in familial colorectal cancer type X. *Clin. Genet.*, **87**, 582–587.
19. Palles, C., Cazier, J.B., Howarth, K.M., Domingo, E., Jones, A.M., Broderick, P., Kemp, Z., Spain, S.L., Guarino, E., Salguero, I. *et al*. (2013) Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nat. Genet.*, **45**, 136–144.
20. Segui, N., Mina, L.B., Lazaro, C., Sanz-Pamplona, R., Pons, T., Navarro, M., Bellido, F., Lopez-Doriga, A., Valdes-Mas, R., Pineda, M. *et al*. (2015) Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. *Gastroenterology*, **149**, 563–566.
21. Esteban-Jurado, C., Vila-Casadesus, M., Garre, P., Lozano, J.J., Pristoupilova, A., Beltran, S., Munoz, J., Ocana, T., Balaguer, F., Lopez-Ceron, M. *et al*. (2015) Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. *Genet. Med.*, **17**, 131–142.
22. Pawson, T. and Warner, N. (2007) Oncogenic re-wiring of cellular signaling pathways. *Oncogene*, **26**, 1268–1275.
23. Greer, E.L. and Shi, Y. (2012) Histone methylation: a dynamic mark in health, disease and inheritance. *Nat. Rev. Genet.*, **13**, 343–357.
24. Kouzarides, T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
25. Peterson, C.L. and Laniel, M.A. (2004) Histones and histone modifications. *Curr. Biol.*, **14**, R546–R551.

26. Carr, S.M., Munro, S., Zalmas, L.P., Fedorov, O., Johansson, C., Krojer, T., Sagum, C.A., Bedford, M.T., Oppermann, U. and La Thangue, N.B. (2014) Lysine methylation-dependent binding of 53BP1 to the pRb tumor suppressor. *Proc. Natl Acad. Sci. U S A*, **111**, 11341–11346.
27. Hamamoto, R., Saloura, V. and Nakamura, Y. (2015) Critical roles of non-histone protein lysine methylation in human tumorigenesis. *Nat. Rev. Cancer*, **15**, 110–124.
28. Hamamoto, R., Toyokawa, G., Nakakido, M., Ueda, K. and Nakamura, Y. (2014) SMYD2-dependent HSP90 methylation promotes cancer cell proliferation by regulating the chaperone complex formation. *Cancer Lett.*, **351**, 126–133.
29. Carr, S.M., Munro, S. and La Thangue, N.B. (2012) Lysine methylation and the regulation of p53. *Essays Biochem.*, **52**, 79–92.
30. Chatterjee, S., Senapati, P. and Kundu, T.K. (2012) Post-translational modifications of lysine in DNA-damage repair. *Essays Biochem.*, **52**, 93–111.
31. Huang, J. and Berger, S.L. (2008) The emerging field of dynamic lysine methylation of non-histone proteins. *Curr. Opin. Genet. Dev.*, **18**, 152–158.
32. Shi, X., Kachirskaia, I., Yamaguchi, H., West, L.E., Wen, H., Wang, E.W., Dutta, S., Appella, E. and Gozani, O. (2007) Modulation of p53 function by SET8-mediated methylation at lysine 382. *Mol. Cell*, **27**, 636–646.
33. Zhang, K. and Dent, S.Y. (2005) Histone modifying enzymes and cancer: going beyond histones. *J. Cell Biochem.*, **96**, 1137–1148.
34. Zhang, X., Wen, H. and Shi, X. (2012) Lysine methylation: beyond histones. *Acta Biochim. Biophys. Sin (Shanghai)*, **44**, 14–27.
35. Kuo, A.J., Cheung, P., Chen, K., Zee, B.M., Kioi, M., Lauring, J., Xi, Y., Park, B.H., Shi, X., Garcia, B.A. *et al.* (2011) NSD2 links dimethylation of histone H3 at lysine 36 to oncogenic programming. *Mol. Cell*, **44**, 609–620.
36. Medjkane, S., Cock-Rada, A. and Weitzman, J.B. (2012) Role of the SMYD3 histone methyltransferase in tumorigenesis: local or global effects? *Cell Cycle*, **11**, 1865.
37. Richly, H., Lange, M., Simboeck, E. and Di Croce, L. (2010) Setting and resetting of epigenetic marks in malignant transformation and development. *Bioessays*, **32**, 669–679.
38. Shi, Y. and Whetstine, J.R. (2007) Dynamic regulation of histone lysine methylation by demethylases. *Mol. Cell*, **25**, 1–14.
39. Yeates, T.O. (2002) Structures of SET domain proteins: protein lysine methyltransferases make their mark. *Cell*, **111**, 5–7.
40. Levy, D., Kuo, A.J., Chang, Y., Schaefer, U., Kitson, C., Cheung, P., Espejo, A., Zee, B.M., Liu, C.L., Tangsombatvisit, S. *et al.* (2011) Lysine methylation of the NF-kappaB subunit RelA by SETD6 couples activity of the histone methyltransferase GLP at chromatin to tonic repression of NF-kappaB signaling. *Nat. Immunol.*, **12**, 29–36.
41. Mukherjee, N., Cardenas, E., Bedolla, R. and Ghosh, R. (2017) SETD6 regulates NF-kappaB signaling in urothelial cell survival: Implications for bladder cancer. *Oncotarget*, **8**, 15114–15125.
42. Vershinin, Z., Feldman, M., Chen, A. and Levy, D. (2016) PAK4 methylation by SETD6 promotes the activation of the Wnt/beta-catenin pathway. *J. Biol. Chem.*, **291**, 6786–6795.
43. O'Neill, D.J., Williamson, S.C., Alkharaif, D., Monteiro, I.C., Goudreault, M., Gaughan, L., Robson, C.N., Gingras, A.C. and Binda, O. (2014) SETD6 controls the expression of estrogen-responsive genes and proliferation of breast carcinoma cells. *Epigenetics*, **9**, 942–950.
44. Binda, O., Sevilla, A., LeRoy, G., Lemischka, I.R., Garcia, B.A. and Richard, S. (2013) SETD6 monomethylates H2AZ on lysine 7 and is required for the maintenance of embryonic stem cell self-renewal. *Epigenetics*, **8**, 177–183.
45. Chen, A., Feldman, M., Vershinin, Z. and Levy, D. (2016) SETD6 is a negative regulator of oxidative stress response. *Biochim. Biophys. Acta*, **1859**, 420–427.
46. Cohn, O., Chen, A., Feldman, M. and Levy, D. (2016) Proteomic analysis of SETD6 interacting proteins. *Data in Brief*, **6**, 799–802.
47. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B. *et al.* (2016) Analysis of protein-coding genetic variation in 60, 706 humans. *Nature*, **536**, 285–291.
48. Schwarz, J.M., Cooper, D.N., Schuelke, M. and Seelow, D. (2014) MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods*, **11**, 361–362.
49. Trievel, R.C., Flynn, E.M., Houtz, R.L. and Hurley, J.H. (2003) Mechanism of multiple lysine methylation by the SET domain enzyme Rubisco LSMT. *Nat. Struct. Biol.*, **10**, 545–552.
50. Mendez, J. and Stillman, B. (2000) Chromatin association of human origin recognition complex, cdc6, and minichromosome maintenance proteins during the cell cycle: assembly of prereplication complexes in late mitosis. *Mol. Cell Biol.*, **20**, 8602–8612.
51. Bertucci, F., Salas, S., Eysteries, S., Nasser, V., Finetti, P., Ginestier, C., Charafe-Jauffret, E., Loriod, B., Bachelart, L., Montfort, J. *et al.* (2004) Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. *Oncogene*, **23**, 1377–1391.
52. Wilson, C.H., McIntyre, R.E., Arends, M.J. and Adams, D.J. (2010) The activating mutation R201C in GNAS promotes intestinal tumourigenesis in Apc(Min/+) mice through activation of Wnt and ERK1/2 MAPK pathways. *Oncogene*, **29**, 4567–4575.
53. Dunican, D.S., McWilliam, P., Tighe, O., Parle-McDermott, A. and Croke, D.T. (2002) Gene expression differences between the microsatellite instability (MIN) and chromosomal instability (CIN) phenotypes in colorectal cancer revealed by high-density cDNA array hybridization. *Oncogene*, **21**, 3253–3257.
54. Sanchez de Abajo, A., de la Hoya, M., van Puijenbroek, M., Godino, J., Diaz-Rubio, E., Morreau, H. and Caldes, T. (2006) Dual role of LOH at MMR loci in hereditary non-polyposis colorectal cancer? *Oncogene*, **25**, 2124–2130.
55. Hoesel, B. and Schmid, J.A. (2013) The complexity of NF-kappaB signaling in inflammation and cancer. *Mol. Cancer*, **12**, 86.
56. Moorchung, N., Kunwar, S. and Ahmed, K.W. (2014) An evaluation of nuclear factor kappa B expression in colorectal carcinoma: an analysis of 50 cases. *J. Cancer Res. Ther.*, **10**, 631–635.
57. Huang, J., Kuismanen, S.A., Liu, T., Chadwick, R.B., Johnson, C.K., Stevens, M.W., Richards, S.K., Meek, J.E., Gao, X., Wright, F.A. *et al.* (2001) MSH6 and MSH3 are rarely involved in genetic predisposition to nonpolypotic colon cancer. *Cancer Res.*, **61**, 1619–1623.
58. Morin, P.J., Sparks, A.B., Korinek, V., Barker, N., Clevers, H., Vogelstein, B. and Kinzler, K.W. (1997) Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. *Science*, **275**, 1787–1790.
59. Rowan, A.J., Lamlum, H., Ilyas, M., Wheeler, J., Straub, J., Papadopoulou, A., Bicknell, D., Bodmer, W.F. and Tomlinson, I.P. (2000) APC mutations in sporadic colorectal tumors: A mutational "hotspot" and interdependence of the "two hits". *Proc. Natl Acad. Sci. U S A*, **97**, 3352–3357.

60. Abdel-Rahman, W.M., Ollikainen, M., Kariola, R., Jarvinen, H.J., Mecklin, J.P., Nystrom-Lahti, M., Knuutila, S. and Peltomaki, P. (2005) Comprehensive characterization of HNPCC-related colorectal cancers reveals striking molecular features in families with no germline mismatch repair gene mutations. *Oncogene*, **24**, 1542–1551.
61. Kanchi, K.L., Johnson, K.J., Lu, C., McLellan, M.D., Leiserson, M.D., Wendl, M.C., Zhang, Q., Koboldt, D.C., Xie, M., Kandoth, C. *et al*. (2014) Integrated analysis of germline and somatic variants in ovarian cancer. *Nat. Commun.*, **5**, 3156.
62. Gurvich, N., Perna, F., Farina, A., Voza, F., Menendez, S., Hurwitz, J. and Nimer, S.D. (2010) L3MBTL1 polycomb protein, a candidate tumor suppressor in del(20q12) myeloid disorders, is essential for genome stability. *Proc. Natl Acad. Sci. U S A*, **107**, 22552–22557.
63. Chen, M., Ni, J., Chang, H.C., Lin, C.Y., Muyan, M. and Yeh, S. (2009) CCDC62/ERAP75 functions as a coactivator to enhance estrogen receptor beta-mediated transactivation and target gene expression in prostate cancer cells. *Carcinogenesis*, **30**, 841–850.
64. Schonemann, L., Kuhn, U., Martin, G., Schafer, P., Gruber, A.R., Keller, W., Zavolan, M. and Wahle, E. (2014) Reconstitution of CPSF active in polyadenylation: recognition of the polyadenylation signal by WDR33. *Genes Dev.*, **28**, 2381–2393.
65. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
66. Koboldt, D.C., Chen, K., Wylie, T., Larson, D.E., McLellan, M.D., Mardis, E.R., Weinstock, G.M., Wilson, R.K. and Ding, L. (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, **25**, 2283–2285.
67. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. *et al*. (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
68. Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N.S. and Cooper, D.N. (2009) The Human Gene Mutation Database: 2008 update. *Genome Med.*, **1**, 13.
69. Yates, A., Akanni, W., Amode, M.R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L. *et al*. (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
70. Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
71. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
72. Desmet, F.O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M. and Beroud, C. (2009) Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.*, **37**, e67.
73. Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M. and Rozen, S.G. (2012) Primer3–new capabilities and interfaces. *Nucleic Acids Res.*, **40**, e115.
74. Nelson, J.D., Denisenko, O. and Bomsztyk, K. (2006) Protocol for the fast chromatin immunoprecipitation (ChIP) method. *Nat. Protoc.*, **1**, 179–185.
75. Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Gallo Cassarino, T., Bertoni, M., Bordoli, L. *et al*. (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.*, **42**, W252–W258.
76. den Dunnen, J.T. (2016) Sequence Variant Descriptions: HGVS Nomenclature and Mutalyzer. *Curr Protoc Hum Genet*, **90**, 7 13 11–17 13 19.
77. Smigielski, E.M., Sirotkin, K., Ward, M. and Sherry, S.T. (2000) dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res*, **28**, 352–355.
78. Thorvaldsdottir, H., Robinson, J.T. and Mesirov, J.P. (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*, **14**, 178–192.

PLOS ONE

RESEARCH ARTICLE

Novel genetic mutations detected by multigene panel are associated with hereditary colorectal cancer predisposition

Lorena Martin-Morales[1,2☯], Paula Rofes[1☯], Eduardo Diaz-Rubio[2,3], Patricia Llovet[1,2], Victor Lorca[1], Inmaculada Bando[1,2], Pedro Perez-Segura[3], Miguel de la Hoya[1,2], Pilar Garre[1,2], Vanesa Garcia-Barberan[1,2], Trinidad Caldes[1,2]*

1 Molecular Oncology Laboratory, Hospital Clinico San Carlos, IdISSC, Madrid, Spain, **2** CIBERONC (Centro de Investigacion Biomedica en Red de Cancer), Carlos III Health Institute, Madrid, Spain, **3** Medical Oncology, Hospital Clinico San Carlos, Madrid, Spain

☯ These authors contributed equally to this work.
* trinidad.caldes@salud.madrid.org

OPEN ACCESS

Citation: Martin-Morales L, Rofes P, Diaz-Rubio E, Llovet P, Lorca V, Bando I, et al. (2018) Novel genetic mutations detected by multigene panel are associated with hereditary colorectal cancer predisposition. PLoS ONE 13(9): e0203885. https://doi.org/10.1371/journal.pone.0203885

Editor: Amanda Ewart Toland, Ohio State University Wexner Medical Center, UNITED STATES

Received: April 27, 2018

Accepted: August 29, 2018

Published: September 26, 2018

Data Availability Statement: Sequencing data have been deposited at the NCBI SRA archive with BioProject record PRJNA474807 and SRA accession SUB4117212.

Funding: This work has been supported by grant PI16/01292 from ISCIII, partially supported by the European Regional Development FEDER funds.

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

Abstract

Half of the high-risk colorectal cancer families that fulfill the clinical criteria for Lynch syndrome lack germline mutations in the mismatch repair (MMR) genes and remain unexplained. Genetic testing for hereditary cancers is rapidly evolving due to the introduction of multigene panels, which may identify more mutations than the old screening methods. The aim of this study is the use of a Next Generation Sequencing panel in order to find the genes involved in the cancer predisposition of these families. For this study, 98 patients from these unexplained families were tested with a multigene panel targeting 94 genes involved in cancer predisposition. The mutations found were validated by Sanger sequencing and the segregation was studied when possible. We identified 19 likely pathogenic variants in 18 patients. Out of these, 8 were found in MMR genes (5 in *MLH1*, 1 in *MSH6* and 2 in *PMS2*). In addition, 11 mutations were detected in other genes, including high penetrance genes (*APC*, *SMAD4* and *TP53*) and moderate penetrance genes (*BRIP1*, *CHEK2*, *MUTYH*, *HNF1A* and *XPC*). Mutations c.1194G>A in *SMAD4*, c.714_720dup in *PMS2*, c.2050T>G in *MLH1* and c.1635_1636del in *MSH6* were novel. In conclusion, the detection of new pathogenic mutations in high and moderate penetrance genes could contribute to the explanation of the heritability of colorectal cancer, changing the individual clinical management. Multigene panel testing is a more effective method to identify germline variants in cancer patients compared to single-gene approaches and should be therefore included in clinical laboratories.

Introduction

Hereditary Non-Polyposis Colorectal Cancer (HNPCC) is a familial syndrome with an increased incidence of colorectal (CRC) and other related cancers [1,2], defined by the Amsterdam I or II clinical criteria. It is well established that approximately half of HNPCC

cases are explained by germline mutations in the DNA mismatch repair (MMR) genes, mainly *MLH1*, *MSH2*, *MSH6* and *PMS2*. As a consequence, these cases present MMR pathway defects and microsatellite instability (MSI) in the tumors, and are referred to as Lynch Syndrome (LS) [3,4].

The universal screening for LS currently comprises two different stages. Firstly, the immunohistochemistry (IHC) of the MMR proteins and/or the MSI status is studied in the tumor of every CRC patient, as well as in some endometrial cancers [5,6]. When this result is positive (MSI/absence of MMR) or if the tumor is not available, patients with a family history of CRC are then screened for germline mutations in the MMR genes, which was previously performed by methods such as Denaturation Gradient Gel Electrophoresis (DGGE) or High Resolution Melting (HRM), followed by Sanger sequencing of samples with altered patterns. Initially, candidates for this LS genetic testing were identified based on the Amsterdam criteria [7,8] However, these algorithms may miss some individuals with LS [9], reason for which the more lenient Bethesda guidelines were created to identify high-risk families that should undergo genetic testing. Although all of these are effective screening tools, they may still miss a proportion of patients with LS. It is worth noting that the IHC and MSI tests have lower sensitivity for detecting *MSH6* and *PMS2* mutation carriers in particular [9], and that the screening of *PMS2* is a challenge due to the high number of pseudogenes.

After this screening, only those families in which a germline pathogenic mutation is found in one of the MMR genes are diagnosed with Lynch Syndrome. Those cases that show MSI/MMR defects in the tumor but lack the corresponding germline MMR mutations are classified as unexplained MMR deficiency, whereas the other half of HNPCC families with no evidence of MMR deficiency has been designated Familial Colorectal Cancer Type X (FCCTX) [10–13]. FCCTX patients lack germline MMR mutations and their tumors are microsatellite stable (MSS). The lack of information about the hereditability of cancer risk in all these unexplained families makes it difficult to carry out an individualized genetic counseling.

Next-generation sequencing (NGS) has revolutionized cancer genomics research, and can be used to search for Mendelian disease genes in an unbiased manner by sequencing the entire protein-coding sequence of known predisposition genes [14]. The practice of genetic testing is rapidly evolving owing to the recent introduction of multigene panels for the diagnosis of hereditary cancer [15]. Multigene panels can be a cost and time-effective alternative to sequentially testing multiple genes. Virtually all multigene panels include high-penetrance genes that establish the risk of a particular type of cancer (such as breast or colon), but also many moderate and low-risk genes. This challenges the personalized management of guidelines when a pathogenic mutation is found, since the phenotypic spectrum and penetrance are less defined or unknown for the latter [16]. The TruSight Cancer Sequencing Panel has been developed by Illumina in collaboration with experts in cancer genomics, and targets a set of 94 well-known cancer-predisposing genes.

The purpose of the present study is the use of multigene panel testing for the diagnosis of hereditary cancer in individuals from high-risk colorectal cancer families.

Methods

Patient selection

A total of 1204 high-risk CRC families have been referred for genetic counseling and/or gene testing for Lynch Syndrome at the Cancer Genetic Clinic of Hospital Clinico San Carlos between the years 2000 and 2016. Among them, 393 families fulfilled the Amsterdam I/II or Bethesda clinical criteria and were molecularly characterized by the study of MSI and/or MMR protein expression in the tumor, and the screening of germline MMR gene mutations

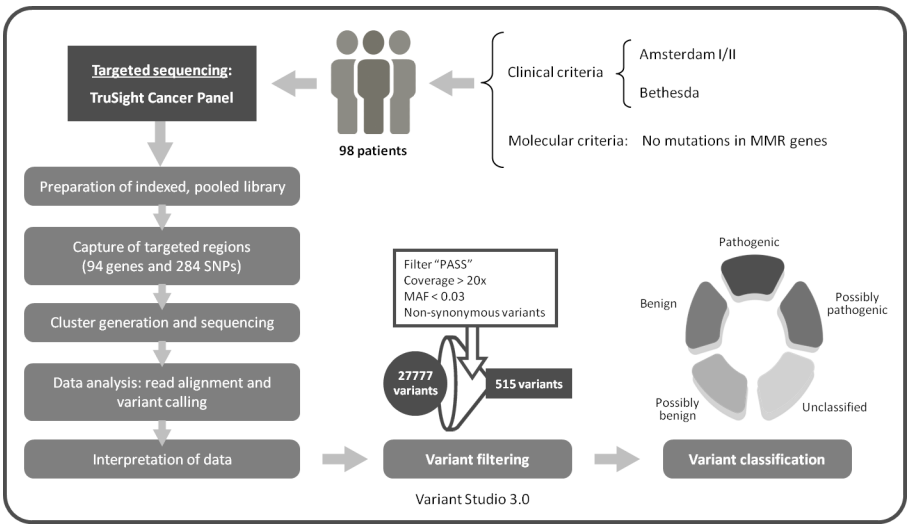


Fig 1. Next generation sequencing workflow using the TruSight cancer sequencing panel.

https://doi.org/10.1371/journal.pone.0203885.g001

(*MLH1*, *MSH2*, *MSH6*, *PMS2* and *EPCAM*, located upstream of *MSH2*) by HRM followed by Sanger sequencing [17,18]. Only 141 families in which a pathogenic germline MMR mutation was found were diagnosed with Lynch Syndrome, while the rest of the families could not be explained by these single-gene analyses. Other genes such as *POLE*, *POLD1* and *NTHL1* were also studied with no positive outcome.

For the present study, we selected 98 patients from those unexplained high-risk CRC families for the test of a multigene cancer panel by NGS (Fig 1). The prioritization of the families was based on the absence of MMR proteins, presence of MSI, lower age at diagnosis or higher number of cancer patients in the pedigree. An *MLH1* mutation carrier was added as a positive control. Participants were asked to donate 10ml of blood at the time of their initial visit. Personal and family histories were obtained from the proband and participating relatives, and cancer diagnoses were confirmed by medical and pathology records. A written informed consent was signed by each participant, and the study was approved wirh an internal code n° 16/204-E_BS by the Clinical Investigation Ethics Committee (CEIC) from the Hospital Clinico San Carlos.

Genomic DNA extraction

Peripheral-blood genomic DNA was extracted with the MagnaPure Compact extractor (Roche, Switzerland) according to the manufacturer's recommended protocol. DNA concentration was measured by Nanodrop (Life Technologies, USA) and Qubit (Life Technologies, USA). The DNA integrity was evaluated by agarose gel electrophoresis.

Next generation sequencing of multiplex PCR amplicons

The patients were tested with the TruSight Cancer Sequencing Panel, targeting 94 genes known to play a role in cancer predisposition (S1 Table). The kit includes more than 5,000 highly-optimized probes (80 mer) that cover the genes and that have been constructed against the human NCBI37/hg19 reference genome. The integrated sample preparation and sequencing was done following the protocol from Illumina, using the Nextera enrichment method and as little as 50ng of DNA for the library. The NGS workflow is summarized in Fig 1.

Data analysis

Sequencing data was analyzed by the MySeq Reporter Software. After the demultiplexing and FASQ file generation, the reads were aligned against the *Homo sapiens* reference genome hg19 to create the BAM files. The genome analysis software Toolkit was used to perform the variant calling and generate the VCF files. After this, Variant Studio 3.0 was used for the variant filtering and annotation. Only variants with 95% of exon covered, labeled as "PASS", with a minimum coverage of 20X and a minor allele frequency (MAF) <0.03 were selected. The variants present in more than 10 different patients were discarded. Regarding the consequence, we only considered missense, nonsense, splicing, in frame and frameshift variants.

Copy number variation analysis

In order to look for DNA copy number variations (CNVs), the BAM files obtained by NGS were analyzed using the Enrichment v3.0.0 app from Illumina's BaseSpace. Those CNVs observed with a quality value greater than 4 and a length greater than 10kb were labeled as "PASS" and selected for further evaluation. Variants present in 4 or more different patients were discarded. The areas affected by potential variations were then examined by the Database of Genomic Variants (DGV) [19] and the Integrative Genomics Viewer (IGV) [20]. CNVs with a frequency above 1% and described by at least two studies were also eliminated. Multiplex ligation-dependent probe amplification (MLPA) was finally carried out in samples 786 and 2954 to confirm *BRCA2*, *MSH2* and *EPCAM* structural variations using SALSA MLPA probe mixes P045, P003 and P072-B2 (MRC-Holland, Netherlands), according to the manufacturer's protocol.

Confirmation of the variants

All clinically actionable variants identified by NGS were validated by Sanger sequencing on a 3130 Genetic Analyzer, with the BigDye Terminator v3.1 Cycle Sequencing Kit (Thermo Fisher Scientific, USA). Sequencing data was aligned against the appropriate reference sequences and analyzed using the Sequencing Analysis Software v5.3.1 (Gene Codes Corp., USA). Unconfirmed variants were eliminated from the results.

Annotation and variant classification

Variants were annotated according to nomenclature recommendations from the Human Genome Variation Society (www.hgvs.org/mutnomen) and further categorized according to the American College of Medical Genetics and Genomics [21] as: benign (class 1), likely benign (class 2), uncertain significance (class 3), likely pathogenic (class 4) and pathogenic (class 5). The following public databases were used for the interpretation of the variants: ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), UMD (http://www.umd.be/), InSight (https://www.insight-group.org/variants/databases/), COSMIC (https://cancer.sanger.ac.uk/cosmic) and Ensembl (http://www.ensembl.org/). Last access: April 2018. Four different *in silico* programs were used for the damage prediction of missense variants (Sift, Provean, PolyPhen-2 and MutationTaster).

Data accesibility

Sequencing data have been deposited at the NCBI SRA archive with BioProject record PRJNA474807 and SRA accession SUB4117212.

Results

Patient samples

Germline DNA was analyzed from 98 patients belonging to high-risk CRC families using the Illumina TruSight Cancer Sequencing Panel, which targets a set of 94 genes known to play a role in cancer predisposition. All patients had been previously tested for Lynch Syndrome by IHC/MSI analyses in the tumor and/or germline MMR single-gene mutation screening, with a negative result for the latter. Characteristics of the studied cohort are detailed in Table 1. The studied families were predominantly affected by CRC; however, other malignancies were observed in some family members, including endometrial, gastric, ovarian, breast, renal and pancreatic cancers. The mean age at diagnosis was 49.1 years old. Most of the carcinomas had an MSS phenotype and presence of the MMR proteins in the tumor. Only 11 patients showed an MSI phenotype, 7 of which had absence of at least one MMR protein (but with no germline MMR mutations detected). The majority of CRCs were Dukes B or C, and nearly all were left-sided.

Results from the NGS targeted sequencing

Among the 98 patients, we found 19 pathogenic or likely pathogenic variants in 18 patients (18.4%), all of which were validated by Sanger sequencing. Table 2 shows the clinical and molecular features of the patients with these mutations. Out of the 18 patients, 8 had *MLH1*/

Table 1. Clinical and molecular characteristics of the 98 probands.

	Amsterdam I (N = 31)	Amsterdam II (N = 22)	Bethesda (N = 45)	Total n° (N = 98)
Gender				
M	14	12	20	46
F	17	10	25	52
Diagnosis age				
<50	19	6	31	56
>50	12	16	14	42
Tumor type				
CRC	31	18	42	91
Breast	0	0	1	1
Gastric	0	1	1	2
Other tumors	0	3	1	4
MSI result				
MSI	6	1	4	11
MSS	19	17	27	63
Unknown	6	4	14	24
IHC result				
MMR-presence	21	16	18	55
MMR-absence	4	1	4	9
Unknown	6	5	23	34
MMR gene test				
MMR wt	28	21	33	82
Unknown	3	1	12	16

N: number of patients; M: male; F: female; CRC: colorectal cancer; MSI: microsatellite instability; MSS: microsatellite stable; IHC: immunohistochemistry; MMR: mismatch repair; wt: wild type.

https://doi.org/10.1371/journal.pone.0203885.t001

Table 2. MMR status in tumors from patients with selected variants identified by the TruSight cancer sequencing panel.

Patient ID	Family Criteria	Cancer Type	Dx Age	MMR status (*MLH1* / *MSH2* / *MSH6* / *PMS2*)			Mutations detected by TruSight Cancer Panel	
				MSI	IHC	HRM	Gene	Variant
499	BETH	CRC	63	MSS	Presence	Wild Type	*BRIP1*	c.903del (p.L301FfsTer2)
555	AMS I	CRC	35	MSI-H	Absence MLH1/PMS2	Wild Type	*MLH1*	c.2050T>G (p.Y684D)
763	AMS I	CRC	47	MSS	Presence	Wild Type	*CHEK2*	c.349A>G (p.R117G)
820	AMS I	CRC	44	MSI-H	Presence	Wild Type	*PMS2*	c.714_720dup (p.F242HfsTer9)
987	AMS I	CRC	62	MSS	Presence	Wild Type	*SMAD4*	c.1194G>A (p.W398Ter)
1041	BETH	CRC	60	MSS	Presence	Wild Type	*MUTYH*	c.1187G>A (p.G396D) c.1227_1228dupGG (p.E410GfsTer43)
1144	AMS I	CRC	45	MSI-H	Absence MLH1/PMS2	Wild Type	*MLH1*	c.2141G>A (p.W714Ter)
1564	AMS II	CRC	58	MSS	Presence	Wild Type	*HNF1A*	c.92G>A (p.G31D)
1652	BETH	CRC	62	MSS	Presence	Wild Type	*XPC*	c.1001C>A (p.P334H)
1756	BETH	CRC	31	MSI-H	Absence MLH1/PMS2	Wild Type	*MLH1*	c.1896+2T>C
1803	AMS II	CRC	79	MSS	Presence	Wild Type	*MUTYH*	c.536A>G (p.Y179C)
1936	AMS I	CRC	47	MSI-H	Absence MSH2/MSH6	Wild Type	*MSH6*	c.1635_1636delAG (p.E546GfsTer16)
2291	AMS I	CRC	51	ND	ND	Wild Type	*PMS2*	c.903G>T (p.K301N)
2456	AMS II	Ovary	35	MSS	Presence	Wild Type	*TP53*	c.783-1G>A
2910	BETH	CRC	39	MSS	Presence	Wild Type	*APC*	c.3199C>T (p.Q1067Ter)
3775	AMS I	CRC	47	MSS	Presence	Wild Type	*MUTYH*	c.1187G>A (p.G396D)
7400	AMS II	CRC	39	MSS	Absence MLH1/PMS2	Wild Type	*MLH1*	c.1731+4A>G
7934	BETH	CRC	35	MSI-H	Presence	ND	*MLH1*	c.677G>T (p.R226L)

BETH: Bethesda; AMS I/II: Amsterdam I and II; CRC: colorectal cancer; Dx Age: age at diagnosis; MMR: mismatch repair; MSS: microsatellite stable; MSI-H: microsatellite instablility-high; IHC: immunohistochemistry; HRM: high resolution melting (for germline screening); ND: not determined.

https://doi.org/10.1371/journal.pone.0203885.t002

MSH6/*PMS2* mutations and 10 carried non-MMR mutations. Four out of the 8 patients with MMR mutations had MSI tumors and absence of the corresponding MMR proteins, while another 3 patients had a discordant tumor screening: one (ID 820) with a frameshift mutation in *PMS2*, c.714_720dup (MSI/presence of PMS2), another (ID 7400) with a splicing mutation in *MLH1*, c.1731+4A>G (MSS/absence of MLH1/PMS2), and the last (ID 7934) with a missense mutation in *MLH1*, c.677G>T (MSI/presence of MLH1/PMS2). Patient ID 555 carried two MMR variants, one in *MLH1* (c.2050T>G, p.Y684D) and another in *PMS2* (c.825A>T, p.Q275H). This patient was diagnosed of CRC at 35 years old, belonged to an Amsterdam I family and had an MSI tumor with absence of MLH1/PMS2. *In silico* studies of both mutations showed that the *PMS2* mutation was neutral while the *MLH1* mutation was predicted to be highly damaging. On the other hand, all of the 10 families with pathogenic variants in non-MMR genes showed MSS tumors, and in one of the families (patient ID 1041) we found two pathogenic variants in *MUTYH* (Table 2). From the remaining patients, 55 (56.1%) were revealed to only carry variants of unknown significance (VUS) in 38 different genes (S2 Table), while 25 (25.5%) just carried polymorphisms.

It is worth noting that, in total, only three splicing variants were not validated by Sanger sequencing (*EZH2* c.1947+1G>T, *MSH2* c.942+2T>G and *MLH1* c.1059-1G>A) and were eliminated from the data. The CNVs were also analyzed as described in Materials and Methods (S3 Table). However, potential structural variations were discarded by the study of SNPs in the corresponding chromosome localization using IGV. It was not possible to analyze some CNVs

due to the low coverage or absence of SNPs in the region. Three CNVs affecting *BRCA2*, *MSH2* and *EPCAM* were not confirmed in 2 of the samples (786 and 2954) by MLPA.

Type, prediction and frequency of likely pathogenic mutations by gene

Out of the 19 germline pathogenic or likely pathogenic mutations detected, 8 (42.1%) were found in MMR genes (5 in *MLH1*, 1 in *MSH6* and 2 in *PMS2*). The remaining 11 mutations were detected in other cancer predisposing genes, including *BRIP1* (n = 1), *CHEK2* (n = 1), *SMAD4* (n = 1), *MUTYH* (n = 4), *HNF1A* (n = 1), *XPC* (n = 1), *TP53* (n = 1), and *APC* (n = 1). The type, prediction and frequency of all the mutations can be observed in Table 3. Among them, there were 9 missense, 4 frameshift, 3 stop-gained and 3 splice site variants. All these variants were rare, and 11 had frequency data not available (NA) in ExAc nor in GnomAD (Table 3). 15 of these mutations were classified as pathogenic or likely pathogenic by the ClinVar and/or InSight databases, following the 5-tier classification system proposed by Plon and colleagues [22]. In addition, 4 of them did not appear in any of the variant databases mentioned earlier, but were considered likely pathogenic due to the type of mutation, *in silico* predictions and/or the molecular features of the tumor. These novel variants were: *SMAD4* c.1194G>A, *PMS2* c.714_720dup, *MSH6* c.1635_1636del and *MLH1* c.2050T>G.

Segregation studies

Unfortunately, segregation studies could not be performed in most of the families, given the unavailability of other family members. Among the few families in which the segregation was

Table 3. Pathogenic and likely pathogenic variants by gene identified by the TruSight cancer sequencing panel.

Patient ID	Family Criteria	Gene	Variant (c.)	Variant (p.)	Type of Mutation	Prediction	Database	Frequency (ExAC)
499	BETH	*BRIP1*	c.903delG	p.Leu301PhefsTer2	Frameshift	Pathogenic	ClinVar	NA
555	AMS I	*MLH1*	c.2050T>G	p.Tyr684Asp	Missense	Likely Pathogenic[#]	Novel	NA
763	AMS I	*CHEK2*	c.349A>G	p.Arg117Gly	Missense	Likely Pathogenic	ClinVar	0.00013
820	AMS I	*PMS2*	c.714_720dup	p.Phe242HisfsTer9	Frameshift	Pathogenic	Novel	NA
987	AMS I	*SMAD4*	c.1194G>A	p.Trp398Ter	Stop gained	Pathogenic	Novel[†]	NA
1041	BETH	*MUTYH*	c.1187G>A	p.Gly396Asp	Miss, SP region	Pathogenic[*]	ClinVar, InSiGHT	0.00280
			c.1227_1228dup	p.Glu410GlyfsTer43	Frameshift	Pathogenic[*]	ClinVar, InSiGHT	0.00015[¶]
1144	AMS I	*MLH1*	c.2141G>A	p.Trp714Ter	Stop gained	Pathogenic	ClinVar, InSiGHT	NA
1564	AMS II	*HNF1A*	c.92G>A	p.Gly31Asp	Missense	Likely pathogenic	ClinVar	0.00071
1652	BETH	*XPC*	c.1001C>A	p.Pro334His	Missense	Likely pathogenic	ClinVar	0.00286
1756	BETH	*MLH1*	c.1896+2T>C	-	Splice donor	Likely pathogenic	ClinVar, InSiGHT	NA
1803	AMS II	*MUTYH*	c.536A>G	p.Tyr179Cys	Missense	Pathogenic[*]	ClinVar, InSiGHT	0.00162
1936	AMS I	*MSH6*	c.1635_1636del	p.Glu546GlyfsTer16	Frameshift	Pathogenic	Novel	NA
2291	AMS I	*PMS2*	c.903G>T	p.Lys301Asn	Miss, SP region	Likely Pathogenic	ClinVar, InSiGHT	0.000008
2456	AMS II	*TP53*	c.783-1G>A	-	Splice acceptor	Pathogenic	ClinVar, IARC *TP53*	NA
2910	BETH	*APC*	c.3199C>T	p.Gln1067Ter	Stop gained	Pathogenic	ClinVar, InSiGHT	NA
3775	AMS I	*MUTYH*	c.1187G>A	p.Gly396Asp	Miss, SP region	Pathogenic[*]	ClinVar, InSiGHT	0.00280
7400	AMS II	*MLH1*	c.1731+4A>G	-	Splice donor	Likely Pathogenic	ClinVar, InSiGHT	NA
7934	BETH	*MLH1*	c.677G>T	p. Arg226Leu	Miss, SP region	Likely Pathogenic	ClinVar, InSiGHT	NA

BETH: Bethesda; AMS I/II: Amsterdam I and II; Miss: missense; SP: splicing; NA: not available

[#]predicted to be probably damaging by *in silico* tools

[*]only causal in homozygosis or in co-occurrence with other mutations

[†]described in COSMIC

[¶]frequency data from gnomAD exomes.

https://doi.org/10.1371/journal.pone.0203885.t003

studied is the one with the *MLH1* c.2050T>G mutation (ID 555). However, this was not very informative, since all the affected members were deceased and only healthy relatives could be tested. Out of the 4 members analyzed, only one carried the variant but was too young to have developed the disease. On the other hand, in the family of ID 1144 (*MLH1* c.2141G>A) 9 relatives were studied, one of whom had developed CRC at the age of 56. As expected, this affected member was shown to carry the mutation, together with another 5 healthy members who were all under 60 (3 of them especially young) and will follow the surveillance recommendations of the Genetic Counseling Unit. For the *CHEK2* variant (c.349A>G), only one distant relative with polyps could be tested and was wild type for the mutation. However, this member also had a CRC history coming from the other side of the family, so no conclusions can be drawn from this result. Finally, the son of participant ID 3775 (monoallelic *MUTYH* c.1187G>A mutation) was also evaluated with a negative result. Although this member had also been reported to have some polyps, not much information was available.

Frequency of VUS in cancer susceptibility genes

All the variants found by NGS were analyzed by ClinVar, and those variants classified as class 3 were selected as VUS. Only those VUS with a very low frequency (<0.005) in ExAc are included in S2 Table. The VUS in cancer genes can be grouped by their functional effect; Fig 2A shows that most of the VUS selected were located in genes involved in DNA repair mechanisms, tumor suppressor genes and proto-oncogenes. Fig 2B shows that the relationship between the number of patients and the number of variants per patient is inversely proportional: 33 patients had 1 VUS, while only one patient (ID 1008) had 5 VUS, and was curiously a patient with a strong family history. Eleven patients (IDs 499, 763, 820, 987, 1041, 1144, 1564, 1652, 1936, 3775 and 7934) with VUS also had a concomitant deleterious mutation (data

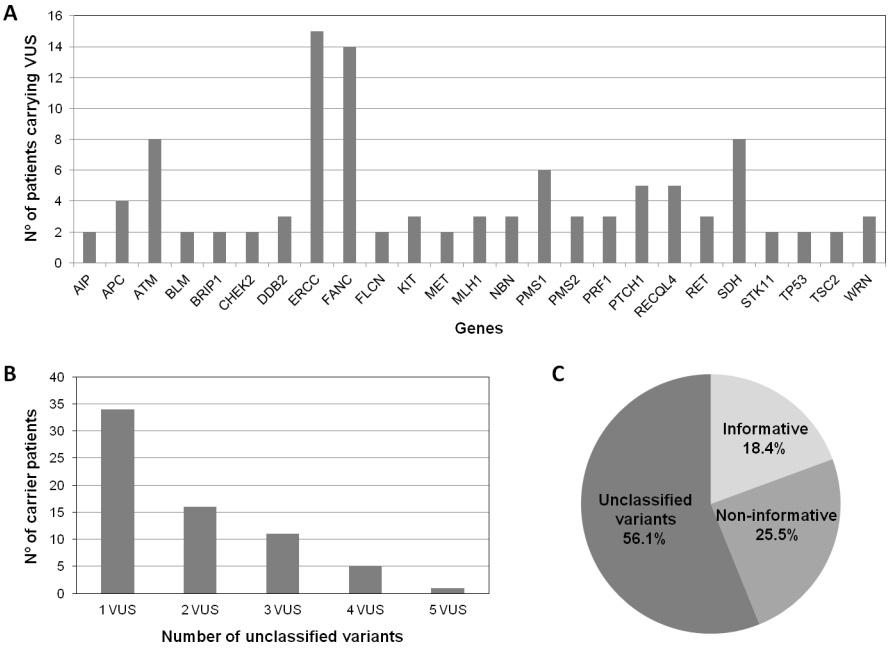


Fig 2. Class 3 variants found in genes implicated in hereditary cancer and clinical experience. A) Number of rare VUS (MAF<0.005) per gene or group of genes, in those cancer susceptibility genes with at least 2 filtered VUS. The different *FANC*, *ERCC* and *SDH* genes were grouped together for simplification. B) Number of unclassified variants per patient of the study cohort. C) Clinical practice experience with multigene panel study. VUS: variants of unknown significance.

https://doi.org/10.1371/journal.pone.0203885.g002

shown in Tables 3 and S2 Table). Among the 55 patients that carried only VUS (56.1%), 31 fulfilled the Amsterdam I/II clinical criteria.

Discussion

Approximately half of HNPCC families carry a germline, pathogenic mutation in one of the MMR genes and are thus considered Lynch Syndrome families. The other half does not present any evidence of MMR deficiency and the genetic basis underlying their cancer predisposition remains unknown, reason for which they are called FCCTX [13,23,24]. However, unexplained families also comprise cases which present MMR defects and/or MSI in the tumors but lack the corresponding germline MMR mutations. All together, these families represent a significant number of cases at the Genetic Counseling Units and they are considered a problem in the clinic. Despite this, the genetic understanding of hereditary CRC syndromes has grown over the years, leading to an increasing request for genetic testing [25]. The limitation of the old screening methods due to their lower sensitivity and the reduced number of genes studied has led to the rise of multigene panel testing in oncology [16,26], and given the advantages of analyzing multiple genes, the benefits of its application in the clinical practice are obvious [15,27].

In this study, 98 unexplained families were reanalyzed using Illumina's NGS TruSight Cancer Sequencing Panel, which targets 94 genes known to play a role in cancer predisposition. The panel testing identified 8 MMR mutations in our cohort (5 in *MLH1*, 1 in *MSH6* and 2 in *PMS2*), 7 of which were found in patients whose tumors showed an altered MMR status (MSI and/or absence of MMR). These cases were missed by our prior screening, and are thus a result from improved testing for these genes. Although these patients represent most of the Lynch-suspected families included in our study, there were another 6 families that were not resolved. The panel also allowed the identification of mutations in other well-known CRC high-penetrance genes, such as *MUTYH* (biallelic), *APC*, *SMAD4* and *TP53*, as well as in moderate-penetrance genes like *MUTYH* (monoallelic), *CHEK2*, *HNF1A*, *BRIP1* and *XPC*. In total, pathogenic or likely pathogenic mutations were found in 18.4% of our cohort, while high-penetrance mutations represented 12.2% of the studied patients.

Four of the identified variants were novel and had not been previously described in any of the variant databases checked (ClinVar, InSiGHT and UMD), although *SMAD4* c.1194G>A had been reported at somatic level in COSMIC. There is enough evidence to claim that the 3 deleterious novel variants (*SMAD4* c.1194G>A, *PMS2* c.714_720dup and *MSH6* c.1635_1636del) are pathogenic, so they will be added to the public databases for future reference. Regarding the missense variant (*MLH1* c.2050T>G), it was present in a patient whose tumor was MSI with absence of MLH1/PMS2. In addition, *in silico* studies showed the change as likely pathogenic. However, the segregation studies performed were not very informative given the lack of affected living relatives, so more studies are needed in order to confirm the pathogenicity of this variant.

Among all the non-MMR genes, *APC*, *MUTYH* and *SMAD4* are well known to be implicated in CRC, specifically in polyposis. It is worth noting that the patient in which the *APC* mutation was found had over 30 polyps. However, *APC* as well as other polyposis-associated genes had already been screened with no positive results (data not shown). For that reason, the family was added to our study on the grounds that it fulfilled the Bethesda criteria. Like the MMR mutations identified, this variant represents a false-negative of previous screenings. Regarding the biallelic *MUTYH* carrier, there was no information of the presence of polyps at the time of patient selection, but a deeper look into the family history revealed that the patient did present multiple polyps. On the contrary, neither the *SMAD4* nor the 2 monoallelic

MUTYH mutation carriers showed any sign of polyposis to our knowledge. Nevertheless, the risk that monoallelic *MUTYH* mutations confer is uncertain, so it is unlikely that they are the only cause of cancer in the corresponding families [27].

The family that carried the *TP53* mutation was not a classic Li-Fraumeni family, but did fulfill the revised Chompret criteria. It had been classified as an Amsterdam II family because there was one member affected with ovary cancer and another 3 with CRC at early ages, but there were also 2 lung cancers within the family and a cousin of the proband had developed a sarcoma very young.

The remaining genes with pathogenic or likely pathogenic mutations had moderate or less defined cancer risks. In the first place, the *BRIP1* gene encodes a member of the RecQ DEAH helicase family that interacts with the BRCT repeats of BRCA1. The bound complex is important in the normal double-strand DNA break repair and appears to be involved in breast and ovarian cancer, where it acts as a tumor suppressor [28]. Rafnar et al. showed that women with *BRIP1* mutations have an increased risk for ovarian cancer that may be as much as 5 times higher than the risk in non-carriers [29]. Noteworthy, one of the *BRIP1* mutations reported by Rafnar et al. (c.1702_1703delAA, p.N568WfsTer9) was also found in a Spanish CRC patient [30], and the tumor showed a loss of the wild type allele in both studies. Here we have identified a different frameshift *BRIP1* mutation (c.903delG, p.L301FfsTer2) in a patient diagnosed with CRC at 63 years old. The family fulfilled the Bethesda criteria but was not very informative due to its reduced size and limited information. Therefore, we still cannot determine the risk for CRC that *BRIP1* mutations confer.

Another gene was *CHEK2*, a tumor suppressor that is activated when DNA is damaged or when DNA strands break. The c.1100delC mutation of *CHEK2* has been confirmed to confer an increased risk of breast cancer in women unselected by family history [31,32,33]. The lifetime risk of developing breast cancer among women with a *CHEK2* mutation has been reported to be approximately 25% [34]. In our cohort, we have found *CHEK2* c.349A>G (p.Arg117Gly), which was considered likely pathogenic by Shoda et al. and proved to produce a nonfunctional protein both by biochemical and bioinformatics analyses [35]. In addition, their results suggest that both of these mutations cannot act in a dominant-negative manner and that tumorigenesis associated with this mutation may be due to haploinsufficiency [35].

We also identified a mutation in *XPC*, a key component of the XPC complex that plays an important role in the early steps of global genome nucleotide excision repair and is involved in damage sensing and DNA binding. The *XPC* mutation found in our study (c.1001C>A, p.Pro334His) had been described as pathogenic in ClinVar by Johns Hopkins University. However, we have identified this mutation for the first time in a CRC patient, who was diagnosed at the age of 50 and did not have any cancer family history. The patient died soon after the diagnosis, not being able to get any additional information. Last but not least, HNF1A is a transcription factor that regulates tissue specific expression of many genes. This gene is implicated in diabetes and had been described in renal cancer, but its role in CRC is unknown. We are also describing the mutation *HNF1A* c.92G>A (p.Gly31Asp) in CRC for the first time.

An increasing number of studies have been published over the past few years addressing the benefits of NGS panel testing for the diagnosis of hereditary cancers as compared with the traditional targeted single-gene screenings. While all of them agree on the huge advantages of panel testing, especially due to its time and cost-efficiency and its higher sensitivity, the only question that remains open to debate is the selection of genes that should be included [36–40]. Here we have used a 94-gene panel, which is to our knowledge the highest number of genes reported in this kind of studies by far. However, this choice was made based on a number of reasons. On the one hand, we have learned from the Genetic Counseling Unit that there is a relatively frequent overlapping of phenotypes among the different hereditary cancer syndromes [36,37],

since the tumor spectrum is sometimes wider than expected [27,41] and the information provided by the family is sometimes incomplete. Therefore, we believe that the best strategy is to group all the syndromes into one single cancer-predisposing multigene panel, as proposed by some other groups [27,36], instead of dividing patients based on their phenotypes.

On the other hand, we strongly support the incorporation of lesser-known genes to NGS panels on a research basis, since the additional cost of adding these genes is minimal [42], and with the intensity of current research the uncertainty of many emerging genes is likely to be resolved soon [38]. This means that some variants that are not informative at the moment may be actionable in the future [38]. In addition, by adding these genes to our panel we are contributing to the accumulation of international research data, which is the only way to continue improving our understanding of CRC genetics [39]. For this reason, we also believe it is highly valuable to include a detailed list of VUS (S2 Table), something that most published studies fail to do [38]. The number of VUS identified cannot be compared with other studies, though, since it is associated with the number of genes on the panel [40]. Despite the high number of genes included in the panel used for the present study, there were some genes that were left out, such as *POLE*, *POLD1*, *NTHL1* or *MSH3*. Although some of them (*POLE*, *POLD1* and *NTHL1*) had already been screened in our cohort, this is a limitation of our study. It goes without saying that we would definitely recommend that future panels used in clinical studies for colorectal cancer families should include those genes as well, for the same reasons that were discussed above.

The clinical practice experience obtained with this multigene panel is shown in Fig 2C. Among all the families that were screened only 18.4% were informative, although this group is underrepresented considering that only unexplained families from previous screenings were included in the study. Out of these, 66.7% carried likely pathogenic mutations in high-penetrance genes and could benefit from a true genetic counseling, taking measures such as reducing the surveillance in non-carriers, who would avoid the stress attached to the lack of awareness. Regarding moderate-penetrance genes, a study with a larger number of patients is needed in order to establish the exact risk they confer. The introduction of NGS panels in the clinical routine of the hospital will help us with this task. Those patients who were just informed of VUS (56.1%) would also take advantage of this measure, since the only thing we can do for now is to keep track of public databases, study the segregation and do functional studies when recommended in order to improve their genetic counseling in the future. The remaining patients (25.5%) were informed that no gene had been found to be involved in their cancer predisposition. Although data regarding lesser-known genes and VUS is highly valuable from a research point of view, participants should be always informed about the limited clinical actionability of testing for genes that are not associated with their phenotype or have moderate penetrance.

In conclusion, the detection of new pathogenic mutations in high-penetrance genes can contribute to the explanation of the cancer heritability in our families, changing the individual clinical management. The NGS panel approach has the advantage of analyzing multiple genes in multiple samples simultaneously, reducing costs and time and increasing the sensitivity in comparison to targeted single-gene screenings. Therefore, multigene panels should be included in clinical laboratories for the screening of all high-risk cancer families regardless of other analyses in the tumor. The number of genes to be included in these panels is debatable, though, and should fit the purposes of each study.

Supporting information

S1 Table. List of genes included in the TruSight cancer sequencing panel.
(DOCX)

S2 Table. Variants of unknown significance identified by the TruSight cancer sequencing panel (Illumina).
(DOCX)

S3 Table. Results from the CNVs analysis.
(XLSX)

Acknowledgments

We are very grateful to the families for their cooperation, and we would also like to acknowledge Alicia Tosar, Isabel Diaz and Paula Diaque for their technical assistance. We acknowledge Jordi Camps for his advice about the CNVs.

Author Contributions

Conceptualization: Trinidad Caldes.

Formal analysis: Lorena Martin-Morales, Paula Rofes, Victor Lorca, Pedro Perez-Segura.

Funding acquisition: Trinidad Caldes.

Investigation: Paula Rofes.

Methodology: Inmaculada Bando.

Supervision: Eduardo Diaz-Rubio, Patricia Llovet, Pilar Garre.

Writing – original draft: Trinidad Caldes.

Writing – review & editing: Lorena Martin-Morales, Miguel de la Hoya, Vanesa Garcia-Barberan, Trinidad Caldes.

References

1. Lynch HT, Smyrk T, Lynch JF. Overview of natural history, pathology, molecular genetics and management of HNPCC (Lynch Syndrome). Int J Cancer, 1996, 69:38–43 https://doi.org/10.1002/(SICI)1097-0215(19960220)69:1<38::AID-IJC9>3.0.CO;2-X PMID: 8600057
2. Watson P, Lynch HT. Extracolonic cancer in hereditary nonpolyposis colorectal cancer. Cancer, 1993, 71:677–85 PMID: 8431847
3. Peltomäki P, Vasen H. Mutations associated with HNPCC predisposition—Update of ICG-HNPCC/INSiGHT mutation database. Dis Markers, 2004, 20:269–76 https://doi.org/10.1155/2004/305058 PMID: 15528792
4. Lynch HT, de la Chapelle A. Genetic susceptibility to non-polyposis colorectal cancer. J Med Genet, 1999, 36:801–18 PMID: 10544223
5. Hampel H, Frankel W, Panescu J, Lockman J, Sotamaa K, Fix D, et al. Screening for Lynch syndrome (hereditary nonpolyposis colorectal cancer) among endometrial cancer patients. Cancer Res, 2006, 66:7810–7 https://doi.org/10.1158/0008-5472.CAN-06-1114 PMID: 16885385
6. Beamer LC, Grant ML, Espenschied CR, Blazer KR, Hampel HL, Weitzel JN, et al. Reflex immunohistochemistry and microsatellite instability testing of colorectal tumors for Lynch syndrome among US cancer programs and follow-up of abnormal results. J Clin Oncol. 2012; 30:1058–63. https://doi.org/10.1200/JCO.2011.38.4719 PMID: 22355048
7. Vasen HF, Mecklin JP, Khan PM, Lynch HT. The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC). Dis Colon Rectum, 1991, 34:424–5 PMID: 2022152
8. 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, 116:1453–6 PMID: 10348829
9. Palomaki GE, McClain MR, Melillo S, Hampel HL, Thibodeau SN. EGAPP supplementary evidence review: DNA testing strategies aimed at reducing morbidity and mortality from Lynch syndrome. Genet Med. 2009; 11:42–65. https://doi.org/10.1097/GIM.0b013e31818fa2db PMID: 19125127

10. Goel A, Xicola RM, Nguyen T-P, Doyle BJ, Sohn VR, Bandipalliam P, et al. Aberrant DNA methylation in hereditary nonpolyposis colorectal cancer without mismatch repair deficiency. Gastroenterology, 2010, 138:1854–62 https://doi.org/10.1053/j.gastro.2010.01.035 PMID: 20102720

11. Lindor NM. Familial colorectal cancer type X: the other half of hereditary nonpolyposis colon cancer syndrome. Surg Oncol Clin N Am, 2009, 18:637–45 https://doi.org/10.1016/j.soc.2009.07.003 PMID: 19793571

12. 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, 293:1979–85 https://doi.org/10.1001/jama.293.16.1979 PMID: 15855431

13. Sánchez-de-Abajo A, de la Hoya M, van Puijenbroek M, Tosar A, López-Asenjo JA, Díaz-Rubio E, et al. Molecular analysis of colorectal cancer tumors from patients with mismatch repair proficient hereditary nonpolyposis colorectal cancer suggests novel carcinogenic pathways. Clin Cancer Res. 2007; 13:5729–35. https://doi.org/10.1158/1078-0432.CCR-06-2996 PMID: 17908962

14. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol, 2008, 26:1135–45 https://doi.org/10.1038/nbt1486 PMID: 18846087

15. Espenschied CR, LaDuca H, Li S, McFarland R, Gau C-L, Hampel H. Multigene Panel Testing Provides a New Perspective on Lynch Syndrome. J Clin Oncol. 2017; 35:2568–75. https://doi.org/10.1200/JCO.2016.71.9260 PMID: 28514183

16. Slavin TP, Niell-Swiller M, Solomon I, Nehoray B, Rybak C, Blazer KR, et al. Clinical Application of Multigene Panels: Challenges of Next-Generation Counseling and Cancer Risk Management. Front Oncol, 2015, 5:208 https://doi.org/10.3389/fonc.2015.00208 PMID: 26484312

17. Godino J, de La Hoya M, Diaz-Rubio E, Benito M, Caldés T. Eight novel germline MLH1 and MSH2 mutations in hereditary non-polyposis colorectal cancer families from Spain. Hum Mutat, 2001, 18:549

18. Caldes T, Godino J, de la Hoya M, Garcia Carbonero I, Perez Segura P, Eng C, et al. Prevalence of germline mutations of MLH1 and MSH2 in hereditary nonpolyposis colorectal cancer families from Spain. Int J Cancer, 2002, 98:774–9 PMID: 11920650

19. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res, 2014, 42:D986–992 https://doi.org/10.1093/nar/gkt958 PMID: 24174537

20. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol, 2011, 29:24–6 https://doi.org/10.1038/nbt.1754 PMID: 21221095

21. 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; 17:405–24. https://doi.org/10.1038/gim.2015.30 PMID: 25741868

22. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, et al. IARC Unclassified Genetic Variants Working Group. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat, 2008, 29:1282–91 https://doi.org/10.1002/humu.20880 PMID: 18951446

23. Garre P, Martín L, Bando I, Tosar A, Llovet P, Sanz J, et al. Cancer risk and overall survival in mismatch repair proficient hereditary non-polyposis colorectal cancer, Lynch syndrome and sporadic colorectal cancer. Fam Cancer, 2014, 13:109–19 https://doi.org/10.1007/s10689-013-9683-2 PMID: 24061861

24. Garre P, Briceño V, Xicola RM, Doyle BJ, de la Hoya M, Sanz J, et al. Analysis of the oxidative damage repair genes NUDT1, OGG1, and MUTYH in patients from mismatch repair proficient HNPCC families (MSS-HNPCC). Clin Cancer Res. 2011; 17:1701–12. https://doi.org/10.1158/1078-0432.CCR-10-2491 PMID: 21355073

25. Lynch HT, Lynch JF, Lynch PM, Attard T. Hereditary colorectal cancer syndromes: molecular genetics, genetic counseling, diagnosis and management. Fam Cancer, 2008, 7:27–39 https://doi.org/10.1007/s10689-007-9165-5 PMID: 17999161

26. Desmond A, Kurian AW, Gabree M, Mills MA, Anderson MJ, Kobayashi Y, et al. Clinical Actionability of Multigene Panel Testing for Hereditary Breast and Ovarian Cancer Risk Assessment. JAMA Oncol, 2015, 1:943–51 https://doi.org/10.1001/jamaoncol.2015.2690 PMID: 26270727

27. Yurgelun MB, Allen B, Kaldate RR, Bowles KR, Judkins T, Kaushik P, et al. Identification of a Variety of Mutations in Cancer Predisposition Genes in Patients With Suspected Lynch Syndrome. Gastroenterology, 2015, 149:604–613.e20 https://doi.org/10.1053/j.gastro.2015.05.006 PMID: 25980754

28. Ramus SJ, Song H, Dicks E, Tyrer JP, Rosenthal AN, Intermaggio MP, et al. Germline Mutations in the BRIP1, BARD1, PALB2, and NBN Genes in Women With Ovarian Cancer. J Natl Cancer Inst, 2015, 107

29. Rafnar T, Gudbjartsson DF, Sulem P, Jonasdottir A, Sigurdsson A, Jonasdottir A, et al. Mutations in BRIP1 confer high risk of ovarian cancer. Nat Genet, 2011, 43:1104–7 https://doi.org/10.1038/ng.955 PMID: 21964575

30. the EPICOLON Consortium, Esteban-Jurado C, Vila-Casadesús M, Garre P, Lozano JJ, Pristoupilova A, et al. Whole-exome sequencing identifies rare pathogenic variants in new predisposition genes for familial colorectal cancer. Genet Med, 2015, 17:131–42 https://doi.org/10.1038/gim.2014.89 PMID: 25058500

31. Ingvarsson S, Sigbjornsdottir BI, Huiping C, Hafsteinsdottir SH, Ragnarsson G, Barkardottir RB, et al. Mutation analysis of the CHK2 gene in breast carcinoma and other cancers. Breast Cancer Res BCR, 2002, 4:R4 PMID: 12052256

32. Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, et al. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat Genet, 2002, 31:55–9 https://doi.org/10.1038/ng879 PMID: 11967536

33. CHEK2 Breast Cancer Case-Control Consortium. CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. Am J Hum Genet, 2004, 74:1175–82 https://doi.org/10.1086/421251 PMID: 15122511

34. Narod SA. Testing for CHEK2 in the cancer genetics clinic: ready for prime time? Clin Genet, 2010, 78:1–7

35. Sodha N, Mantoni TS, Tavtigian SV, Eeles R, Garrett MD. Rare germ line CHEK2 variants identified in breast cancer families encode proteins that show impaired activation. Cancer Res, 2006, 66:8966–70 https://doi.org/10.1158/0008-5472.CAN-06-1990 PMID: 16982735

36. Rosenthal ET, Bernhisel R, Brown K, Kidd J, Manley S. Clinical testing with a panel of 25 genes associated with increased cancer risk results in a significant increase in clinically significant findings across a broad range of cancer histories. Cancer Genet, 2017, 218–219:58–68 https://doi.org/10.1016/j.cancergen.2017.09.003 PMID: 29153097

37. Rohlin A, Rambech E, Kvist A, Törngren T, Eiengård F, Lundstam U, et al. Expanding the genotype-phenotype spectrum in hereditary colorectal cancer by gene panel testing. Fam Cancer, 2017, 16:195–203 https://doi.org/10.1007/s10689-016-9934-0 PMID: 27696107

38. Shirts BH, Casadei S, Jacobson AL, Lee MK, Gulsuner S, Bennett RL, et al. Improving performance of multigene panels for genomic analysis of cancer predisposition. Genet Med Off J Am Coll Med Genet, 2016, 18:974–81

39. Lorans M, Dow E, Macrae FA, Winship IM, Buchanan DD. Update on Hereditary Colorectal Cancer: Improving the Clinical Utility of Multigene Panel Testing. Clin Colorectal Cancer, 2018. https://doi.org/10.1016/j.clcc.2018.01.001

40. Susswein LR, Marshall ML, Nusbaum R, Vogel Postula KJ, Weissman SM, Yackowski L, et al. Pathogenic and likely pathogenic variant prevalence among the first 10,000 patients referred for next-generation cancer panel testing. Genet Med Off J Am Coll Med Genet, 2016, 18:823–32

41. Garre P, Martín 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, 87:582–7 https://doi.org/10.1111/cge.12427 PMID: 24814045

42. Gallego CJ, Shirts BH, Bennette CS, Guzauskas G, Amendola LM, Horike-Pyne M, et al. Next-Generation Sequencing Panels for the Diagnosis of Colorectal Cancer and Polyposis Syndromes: A Cost-Effectiveness Analysis. J Clin Oncol Off J Am Soc Clin Oncol, 2015, 33:2084–91