Teresa Patrícia da Silva Gil Duarte

Licenciada em Bioquímica

Candidate Germline Genetic Variants for Familial Colorectal Cancer Type X

Dissertação para obtenção do Grau de Mestre em
Genética Molecular e Biomedicina

Orientador: Dra. Maria Cristina Mantas Albuquerque Valeroso,
Investigadora, Unidade de Investigação em Patobiologia
Molecular (UIPM), Instituto Português de Oncologia de
Lisboa Francisco Gentil, EPE (IPOLFG)

Júri:

Presidente: Prof. Doutora Paula Maria Theriaga Mendes Bernardes Gonçalves
Arguente: Prof. Doutora Maria Alexandra Taquenho Maia e Silva

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro 2017

Teresa Patrícia da Silva Gil Duarte

Licenciada em Bioquímica

Candidate Germline Genetic Variants for Familial Colorectal Cancer Type X

Dissertação para obtenção do Grau de Mestre em
Genética Molecular e Biomedicina

Orientador: Dra. Maria Cristina Mantas Albuquerque Valeroso, Investigadora,
Unidade de Investigação em Patobiologia
Molecular (UIPM), Instituto Português de Oncologia de
Lisboa Francisco Gentil, EPE (IPOLFG)

Júri:

Presidente: Prof. Doutora Paula Maria Theriaga Mendes Bernardes Gonçalves
Arguente: Prof. Doutora Maria Alexandra Taquenho Maia e Silva

Setembro 2017

Candidate Germline Genetic Variants for Familial Colorectal Cancer Type X

Aos meus pais

AGRADECIMENTOS

Em primeiro lugar, não podia deixar de agradecer a todas as pessoas que me ajudaram e contribuíram para a realização deste projeto, esta jornada não seria a mesma sem vocês.

À Dra Cristina Albuquerque por me ter dado a oportunidade de desenvolver este projeto, pela disponibilidade com que esclareceu as minhas dúvidas, por toda a simpatia e por todo o conhecimento que me transmitiu ao longo deste ano.

Ao Dr. Bruno Filipe, um especial agradecimento por toda a incansável ajuda prestada, paciência, disponibilidade para esclarecer todas as minhas dúvidas e por nunca ter perdido a simpatia apesar da constante insistência em lhe "gastar" o nome.

À Mestre Inês Francisco por toda a ajuda prestada na revisão da tese, transmissão de conhecimentos e apoio no trabalho prático realizado, e à Doutoranda Patrícia Silva por toda a ajuda e simpatia com que me acolheu.

À Doutora Branca Cavaco, Coordenadora da unidade de Investigação em Patobiologia Molecular, do Instituto Português de Oncologia, Lisboa, Francisco Gentil, E.P.E, por me ter concedido a oportunidade de participar neste projeto.

Ao Serviço de Gastrenterologia, Clínica de Risco familiar e ao Serviço de Anatomia Patológica do Instituto Português de Oncologia, Lisboa, Francisco Gentil, E.P.E, pela disponibilização do material biológico e informação clínica dos doentes.

À Dra Patrícia Machado, Dra Sidónia Santos e em especial à Dra Sofia Fragoso por toda a boa disposição no laboratório e simpatia.

Aos restantes membros da Unidade de Investigação em Patobiologia Molecular, pela simpatia com que me receberam.

Às minhas colegas de laboratório Marlene Duarte, Ana Magalhães, Ana Hipólito e Carolina Pereira que tornaram todo este percurso mais fácil, seja por um sorriso, uma palavra amiga, uma gargalhada na hora de almoço ou um simples "estou aqui para te ajudar". Sem vocês o laboratório não seria o mesmo.

À Mestre Marlene Duarte por todos os momentos de carinho, amizade e descontração e sobretudo por seres incansável e sempre disposta a ajudar.

À minha Mestre Ana Magalhães por me teres acolhido como tua pupila e sobretudo como tua amiga. Por toda a ajuda, toda a força que me transmitiste e por todos os sorrisos dentro e fora do laboratório. Não podia ter pedido melhor mestre.

À minha Ana Hipólito, por me teres dado uma nova alegria a partir do momento em que entraste no laboratório e que do nada te tornaste o meu apoio. Por todas as noitadas no laboratório, todas as loucuras, sorrisos, tristezas, dias intensivos e acima de tudo, por teres estado sempre aqui para mim.

À minha colega de casa, Marta Fernandes por todo o apoio, paciência para ouvir as minhas lamurias, momentos de descontração, carinho e amizade.

Aos meus “Chucha-rolhas”, “FCTenses” e “Supremas” por toda a amizade e carinho, que mesmo longe estão sempre presentes.

Ao Duarte, as palavras são poucas para tanto que tenho de te agradecer. Obrigada por seres, obrigada por estares e, sobretudo, obrigada por me fazeres sorrir.

Por fim, a vocês, pais, por todo o apoio e amor incondicional, por toda a força e incentivo que deram ao longo da vida e pelos exemplos de coragem que são para mim. Mais uma vez, as palavras não chegam. OBRIGADA.

ABSTRACT

Familial colorectal cancer type X (FCCTX) defines families that fulfill the Amsterdam criteria without evidence of defects in the DNA mismatch repair (MMR) genes and whose tumors do not present microsatellite instability. However, its genetic etiology remains unknown, therefore this study aimed to identify and evaluate novel variants and candidate genes that may play a role in FCCTX susceptibility. Based on a previous whole exome sequencing (WES) study in a FCCTX family, a bioinformatic analysis and a subsequent *in silico* and segregation studies were conducted to identify candidate genes and/or specific variants that may predispose for this syndrome. Since this analysis was already started, variants in 6 genes have already been identified to segregate with the disease.

Therefore, the aim of this project was to continue this work by completing the selection of candidate variants and to characterize and try to clarify the role of these variants for FCCTX susceptibility. In order to elucidate the possible contribution of the corresponding genes for FCCTX, a mutational analysis was performed to search for germline mutations in index patients from FCCTX and FCCTX-like families. In addition, using the WES data, a copy number variation (CNV) analysis was also performed for the family subjected to WES, followed by a bioinformatic and *in silico* analysis to reveal amplicon deletions that may segregate with disease. It was also evaluated the involvement of the *TPP2* gene, previously identified as a possible candidate gene for FCCTX in another family, in healthy and affected FCCTX patients, by mutational/splicing analysis, relative quantification by quantitative PCR and protein truncation test to assess resulting truncating proteins.

The bioinformatic followed by *in silico* and segregation analysis of the variants obtained from WES, revealed 1 variant in the *CACNA1S* gene that segregated with the disease. Adding this variant to the already obtained, a total of 7 variants in different genes were found as possible contributors to FCCTX in this family (*MTMR3*, *NPR2*, *DUSP12*, *LGR6*, *TAS1R1*, *SMG7* and *CACNA1S*). The segregation analysis also revealed the segregation of the previously identified *MTMR3* and *TAS1R1* variants in a patient from an older generation of the family. The CNV analysis revealed, after selective criteria, 22 amplicons of interest with a deletion scenario, for further segregation studies.

The germline mutational analysis in a set of FCCTX and FCCTX-like families uncovered 2 and 3 potentially pathogenic variants for the *MTMR3* and *TAS1R1* genes, respectively. One of the variants found in *MTMR3* was the same found in the WES analysis. Thus far no relevant variants were observed for the remaining candidate genes, however this analysis is not completed.

The *TPP2* study revealed the presence of non-described splicing isoforms. One of these isoforms exhibited a differential expression between healthy and affected individuals and the protein truncation test revealed that this alternative transcript gives rise to a truncated protein.

In conclusion, the identification of more than one genetic variant appears to agree with the suggestion that FCCTX is a heterogeneous entity and the discovery of potentially pathogenic variants in *MTMR3* and *TAS1R1* reinforce their possible involvement in FCCTX. The alternative *TPP2* transcript appears to be involved in the earlier stages of colorectal carcinogenesis.

Keywords: FCCTX, WES, Colorectal Cancer, Germline Genetic Variants, Susceptibility Genes

RESUMO

O cancro do cólon e reto familiar do tipo X (FCCTX) define as famílias que preenchem os critérios de Amesterdão nas quais não é identificada mutação germinal nos genes de reparação de erros de DNA do tipo *mismatch* (MMR) e cujos tumores não apresentam instabilidade de microssatélites. Sendo a sua causa molecular desconhecida. Assim, o presente estudo teve como objetivo identificar e avaliar novas variantes e genes candidatos que possam estar envolvidos na suscetibilidade para o FCCTX. Com base num estudo prévio de *whole exome sequencing* (WES), realizado numa família FCCTX, foi efetuada uma análise bioinformática e *in silico* e um subsequente estudo de segregação, de modo a identificar genes candidatos e/ou variantes específicas que possam predispor para esta condição hereditária. Uma vez que esta análise já tinha sido iniciada, 6 variantes em diferentes genes, que segregaram com a doença, já tinham sido identificadas.

Assim, o objetivo deste trabalho consistiu na continuação deste estudo, completando a seleção das variantes candidatas, e na caracterização e clarificação destas variantes para a suscetibilidade para o FCCTX. De modo a elucidar a possível contribuição destes genes para o FCCTX, foi realizada uma análise mutacional em indivíduos *index* de famílias FCCTX e potenciais famílias FCCTX. Adicionalmente, utilizando os resultados da WES, foi também realizada uma análise de *copy number variantion* (CNV) para a família integrada na análise de WES, seguida de uma análise bioinformática e estudos *in silico* de modo a avaliar a presença de deleções de *amplicons* que pudessem segregar com a doença. O envolvimento de transcritos alternativos do gene *TPP2*, previamente identificado como um possível gene candidato para o FCCTX noutra família, foi também avaliado em indivíduos saudáveis e afetados por análise mutacional/*splicing*, quantificação relativa por PCR quantitativo e teste da proteína truncada, para avaliar a existência de proteínas truncantes.

A análise bioinformática seguida pela análise *in silico* e segregação das variantes obtidas por WES revelou a segregação com a doença de uma variante no gene *CACNA1S*. Tendo em conta as variantes já obtidas, foram identificadas 7 variantes em diferentes genes (*MTMR3*, *NPR2*, *DUSP12*, *LGR6*, *TAS1R1*, *SMG7* e *CACNA1S*) como possíveis intervenientes na suscetibilidade para o FCCTX nesta família. A análise de segregação revelou ainda a segregação das variantes dos genes *MTMR3* e *TAS1R1* num individuo proveniente de uma geração anterior. A análise de CNV revelou, após a introdução de critérios seletivos, 22 *amplicons* de interesse com um cenário de deleção, para estudos de segregação adicionais.

A análise de mutações germinais num conjunto de famílias FCCTX e potenciais famílias FCCTX revelou 2 e 3 variantes potencialmente patogênicas para os genes *MTMR3* e *TAS1R1*, respetivamente. Uma das variantes encontradas no gene *MTMR3* correspondeu à variante encontrada no estudo de WES. Não foram observadas até ao momento variantes relevantes para os restantes genes candidatos, porém esta análise não está completa.

O estudo do gene *TPP2* revelou a presença de isoformas não descritas. Uma destas isoformas apresentou uma expressão diferencial entre o transcrito normal e o alternativo em indivíduos saudáveis e afetados e, o teste da proteína truncada revelou que este transcrito alternativo dá origem a uma proteína truncada.

Em conclusão, a identificação de mais de uma variante genética parece concordar com a sugestão de que o FCCTX é uma entidade heterogénea, e a descoberta de variantes potencialmente patogénicas nos genes *MTMR3* e *TAS1R1* reforçam seu possível envolvimento no FCCTX. O transcrito alternativo do gene *TPP2* parece estar envolvido numa fase inicial da carcinogénese colorretal.

Keywords: FCCTX, *WES*, Cancro Colorectal, Variantes Genéticas Germinais, Genes de Suscetibilidade

TABLE OF CONTENTS

FIGURE INDEX

TABLE INDEX

LIST OF ABBREVIATIONS

AC	Amsterdam criteria
APC	Adenomatous polyposis coli
BER	Base excision repair
BMPR1A	Bone morphogenetic protein receptor type 1A
CACNA1S	Calcium voltage-gated channel subunit alpha1 S
cDNA	Complementary DNA
cGMP	Cyclic guanosine monophosphate
CIMP	CpG island methylator phenotype
CIN	Chromosomal instability
CNV	Copy number variation
CRC	Colorectal cancer
Ct	Threshold cycle
DCC	DCC netrin 1 receptor
ddH_2O	Double-distilled water
ddNTP	Dideoxynucleoside triphosphate
dNTP	Deoxynucleotide triphosphate
DUSP12	Dual specificity phosphatase 12
EDTA	Ethylenediamine tetraacetic acid
EPCAM	Epithelial cell adhesion molecule
FCCTX	Familial colorectal cancer type X
FFPE	Formalin-fixed paraffin-embedded
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GIMAP1	GTPase, IMAP family member 1
GTP	Guanosine triphosphate
HNPCC	Hereditary non-polyposis colorectal cancer
JPS	Juvenile polyposis syndrome
KRAS	KRAS proto-oncogene
LGR6	Leucine rich repeat containing G protein-coupled receptor 6
LOH	Loss of heterozygosity
LS	Lynch syndrome
MAP	*MUTYH*-associated polyposis
***MAPK1* (*ERK2*)**	Mitogen-activated protein kinase 1
***MAPK3* (*ERK1*)**	Mitogen-activated protein kinase 3
MGMT	O-6-methylguanine-DNA methyltransferase
MHC	Major histocompatibility complex
MLH1	MutL Homolog 1
MMR	Mismatch repair
MSH2	MutS Homolog 2
MSH3	MutS Homolog 3
MSH6	MutS Homolog 6
MSI	Microsatellite instability
MSI-H	Microsatellite instability -low
MSI-L	Microsatellite instability - high
MSS	Microsatellite stable
MTMR3	Myotubularin related protein 3
mTOR	Mechanistic target of rapamycin
mTORC1	Mechanistic target of rapamycin complex 1
MUTYH	mutY DNA glycosylase

NGS	Next-generation sequencing
NPR2	Natriuretic peptide receptor 2
PCR	Polymerase chain reaction
PJS	Peutz-Jeghers syndrome
PMS1	PMS1 Homolog 1
PMS2	PMS1 Homolog 2
POLD1	DNA polymerase delta 1
PtdIns(3,5)P2	Phosphatidylinositol 3,5-biphosphate
PtdIns3P	Phosphatidylinositol 3-phosphate
PTT	Protein truncation test
qPCR	Quantitative PCR
RNF213	Ring finger protein 213
RPS20	Ribosomal protein S20
RT	Reverse transcription
RYR1	Ryanodine receptor 1
SD	Standard deviation
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEMA4A	Semaphorin 4A
SMAD2	SMAD family member 2
SMAD4	SMAD family member 4
SMG7	SMG7, nonsense mediated mRNA decay factor
SNP	Single-nucleotide polymorphism
SP	Serrated polyposis
STK11	Serine threonine kinase 11
TAS1R1	Taste 1 receptor member 1
TBE	Tris-Borate-EDTA
TCF	T-cell factor
TGF-β	Transforming growth factor beta
TP53	Tumor protein p53
TPP2	Tripeptidyl peptidase 2
TSG	Tumor suppressor gene
TSG-	Tumor suppressor gene negative
TSG+	Tumor suppressor gene positive
WES	Whole exome sequencing
Wt	Wild type
ZNF717	Zinc finger protein 717

LIST OF UNITS

% (w/v)	Weight per volume percentage
°C	Degrees Celsius
bp	Base pair
Ci	Currie
kDa	Kilodalton
kb	Kilobases
Tm	Melting temperature
V	Volts
mA	Milliamperes
U	Units
L; mL; µL	Liter; milliliter (10^{-3}L); microliter (10^{-6}L)
M; mM; µM; nM	Molar (mol/L); milimolar (10^{-3} M); micromolar (10^{-6} M); nanomolar (10^{-9}M)
mol; pmol	Mole; picomol (10^{-12} mol)

1. INTRODUCTION

1.1 Colorectal cancer

Colorectal cancer (CRC) is the third most common cancer and the fourth most common cause of cancer-related deaths worldwide. (Ferlay *et al.* 2013) In Portugal, CRC has the second highest incidence after breast cancer in female and prostate cancer in men, being the second cause of cancer-related deaths (Miranda *et al.* 2016). CRC has a low incidence at ages younger than 50 years, but strongly increases with age, having a median age at diagnosis of 70 years in developed countries. (Brenner *et al.* 2014)

The pathogenesis of CRC is very complex and diverse and is also influenced by multiple factors, some of which are related to genetic predisposition, while others are attributed to unhealthy lifestyle factors. Several epidemiological studies have confirmed the involvement of numerous environmental and dietary factors, such as cigarette smoking, alcohol abuse, a diet high in fat and low in fiber, a sedentary lifestyle and obesity. (Colussi *et al.* 2013)

About 75% of patients with CRC have sporadic disease with no apparent evidence of having inherited the disorder. The remaining 25% of patients have a family history of CRC that suggests a hereditary contribution, common environmental exposures within the family, or a combination of both. Genetic pathogenic variants have been identified as the cause of inherited cancer risk in some colon cancer–prone families, but these pathogenic variants are estimated to account for only 5% of all CRC. (Daly *et al.* 2017) The remaining 20% of inherited CRC cases do not have a clearly defined mechanism and likely result from a combination of gene polymorphisms and mutations, alterations in multiple susceptibility *loci* and environmental influences. (Popek and Tsikitis 2011)

To develop accurate risk assessment and more precise screening approaches, a clarification of predisposing genes and a better understanding of the pathways and molecular events that drive CRC carcinogenesis is essential. (Daly *et al.* 2017)

1.2 Molecular events associated with colorectal carcinogenesis

CRC results from the accumulation of both acquired genetic and epigenetic changes that transform normal glandular epithelium into invasive adenocarcinoma (Lao and Grady 2011). Presently, CRC can be separated into three categories based on similar molecular genetic features, suggesting divergent pathways of tumorigenesis: chromosomal instability (CIN), microsatellite instability (MSI), and CpG island methylator phenotype (CIMP). (Daly *et al.* 2017)

1.2.1 Chromosomal instability pathway

The majority of CRCs develop through the CIN pathway which is characterized by widespread alterations in chromosome number (aneuploidy) and frequent detectable losses at the molecular level of portions of chromosomes (loss of heterozygosity - LOH), such as 5q, 18q, and 17p. The steps involved

in this process were first described in the classic adenoma-carcinoma progression model proposed by Fearon and Vogelstein in 1990. (Daly *et al.* 2017)

The classic adenoma-carcinoma sequence (figure 1.1), accounts for approximately 80% of sporadic colon tumors and typically includes mutation of the *APC* (adenomatous polyposis coli) tumor suppressor gene early in the neoplastic process. Both copies of the *APC* gene must be functionally inactivated, either by mutation or epigenetic events, for adenomas to develop. *APC* is a component of the WNT signaling pathway (figure 1.2) and a key negative regulator of β-catenin since its normal protein promotes β-catenin degradation. With loss of APC function, β-catenin accumulates and translocates to the nucleus, where it forms a complex with the DNA-binding factor TCF (T-cell factor) and activates the transcription of genes that promote proliferation. Additional mutations accumulate, including activating mutations in *KRAS* (KRAS proto-oncogene), a GTP (guanosine triphosphate)-binding protein that promotes growth and prevents apoptosis through the downstream RAS-RAF-MEK-ERK pathway. Neoplastic progression is also associated with mutations in other tumor suppressor genes such as those encoding *SMAD2* (SMAD family member 2) and *SMAD4* (SMAD family member 4) (18q), which are effectors of *TGF-β* (transforming growth factor beta) signaling. Impairment of *TP53* (tumor protein p53), usually through allelic loss of 17p, often occurs at later stages of tumor progression. (Worthley *et al.* 2010) (Kumar *et al.* 2015)

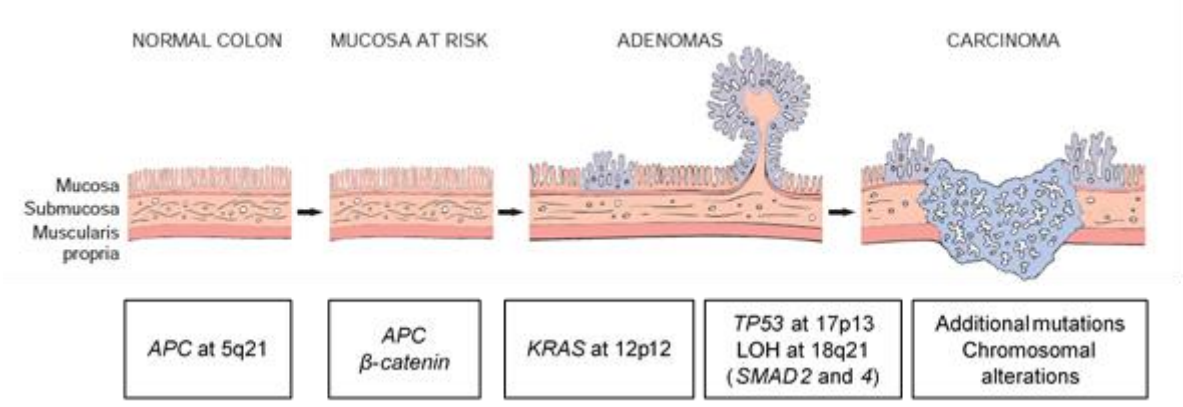


Figure 1.1 - The adenoma-carcinoma sequence. Main histological and molecular changes. (Adapted from: Kumar et al. 2015)

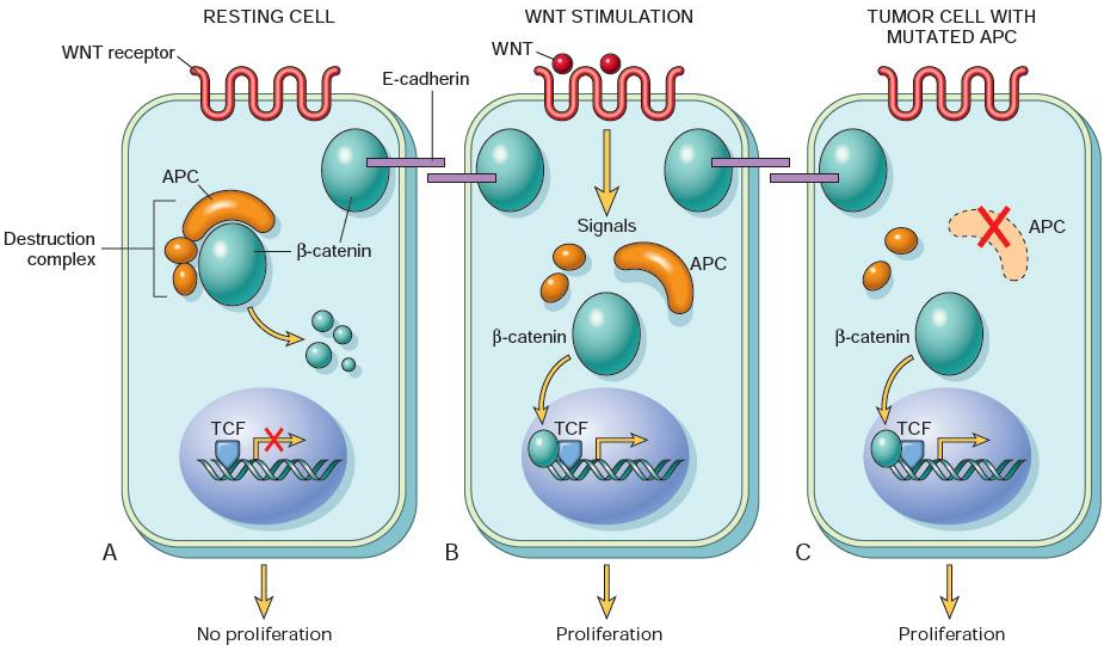


Figure 1.2 - Schematic overview of the WNT signaling pathway and its role in colorectal cancer. (A) In resting colonic epithelial cells (not exposed to WNT), β-catenin forms a complex containing the APC protein, which leads to the destruction of β-catenin, and intracellular levels of β-catenin are low. (B) When normal colonic epithelial cells are stimulated by WNT molecules, the destruction complex is deactivated, β-catenin degradation does not occur, and cytoplasmic levels increase. β-catenin translocates to the nucleus, where it binds to TCF, a transcription factor that activates the transcription of genes that promote proliferation. (C) When APC is mutated or absent, as frequently occurs in colonic polyps and cancer, the destruction of β-catenin cannot occur. β-catenin translocates to the nucleus and coactivates genes that promote entry into the cell cycle and proliferation, and cells behave as if they are under constant stimulation by the WNT pathway. (Adapted from: Kumar et al. 2015)

1.2.2 Microsatellite instability pathway

The microsatellite instability (MSI) pathway represents a form of genomic instability involved in the genesis of approximately 15% of sporadic colorectal cancer and in the majority of hereditary nonpolyposis colorectal cancer (HNPCC) syndrome. (Colussi *et al.* 2013)

Microsatellites are short nucleotide repeat sequences scattered throughout the genome, which are particularly susceptible to slippage by DNA polymerase during replication, resulting in the formation of an insertion–deletion loop that can be identified and corrected by the DNA mismatch repair (MMR) system. If this loop is not repaired it can result in frameshift mutations, which can produce a truncated, nonfunctional protein. MSI is caused by the inactivity of the MMR system, which acts like a proofing machine to increase the fidelity of DNA replications by identifications and direct repair of mismatched nucleotides. The MMR system is composed of multiple interacting proteins including the human MSH2 (MutS Homolog 2), MLH1 (MutL Homolog 1), MSH6 (MutS Homolog 6), MSH3 (MutS Homolog 3), PMS1 (PMS1 Homolog 1) and PMS2 (PMS1 Homolog 2). Inactivation of the MMR system arises from mutations in the *MSH2*, *MLH1*, *MSH6* and *PMS2* genes or by hypermethylation of the *MLH1* gene promoter. (Colussi *et al.* 2013) (Worthley *et al.* 2010)

The rate of adenoma to carcinoma progression appears to be faster in tumors were the MMR system seems to be inactivated. (Daly *et al.* 2017)

1.2.3 CpG island methylator phenotype pathway

A third pathway through which CRC progresses is the CpG island methylator phenotype (CIMP). Regions with a high frequency of cytosine nucleotide followed by a guanine nucleotide (CpG dinucleotide) are termed CpG islands, and are found in the promoter regions of approximately 50% of all genes. The addition of a methyl group to cytosine bases (DNA methylation) in these CpG regions has been associated with silencing of genes involved in cell cycle regulation, apoptosis, angiogenesis, DNA repair, invasion and adhesion. (Roper and Hung 2013) (Colussi *et al.* 2013)

Studies of the epigenomes of CRCs have identified groups of CRCs with distinct epigenotypes. These epigenotypes have unique associations with a variety of mutations and with the genomic instability status of the cancer. The epigenotypes that have been proposed include different classes of CIMP, which is determined by the proportion of methylated *loci* detected. (Lao and Grady 2011) Shen and colleagues have suggested that CIMP CRCs can be divided into two distinct classes, CIMP1 and CIMP2. CIMP1 tumors are often microsatellite unstable (80%) with *BRAF* (B-Raf proto-oncogene) mutations (53%), whereas CIMP2 tumors have *KRAS* mutations (92%), but rarely have MSI or *BRAF* or *TP53* mutations. Tumors that do not have CIMP usually have a high frequency of *TP53* mutations (71%) and an intermediate frequency of *KRAS* mutations (33%) (Shen *et al.* 2007). Another study divided the CIMP phenotype into CIMP-high and CIMP-low. CIMP-high exhibits an exceptionally high frequency of cancer-specific DNA hypermethylation, a strong association with methylated *MLH1* and mutant *BRAF* (p.V600E) and CIMP-low is associated with mutant *KRAS* and lacks in presenting DNA methylation at a unique set of CpG sites. (Hinoue *et al.* 2012)

Although CIMP tumors and the distinct epigenotypes described above, do seem to represent distinct subsets of CRCs, there is still no agreement and validation of a gene methylation panel for designating the epigenotypes. (Lao and Grady 2011)

1.3 Hereditary colorectal cancer syndromes

As previously mentioned, is estimated that about 25% of colorectal cancers are familial with 5% to 6% related to a known genetic syndrome. The hereditary CRCs are broadly divided into polyposis and nonpolyposis syndromes. Individuals with hereditary CRC syndromes are at risk for earlier development of cancer, increased risk of metachronous cancers, and extracolonic manifestations. As such, careful integration of clinical features, family history, endoscopic findings, molecular pathology, and genetic testing are crucial components in establishing the diagnosis of a hereditary cancer syndrome. Recognition and diagnosis of these syndromes is critical for prevention, early detection and treatment of associated malignancies to reduce associated morbidity and mortality. (Gala and Chung 2015) (Wells and Wise 2017)

1.3.1 Polyposis colorectal cancer syndromes

1.3.1.1 Familial adenomatous polyposis

Familial adenomatous polyposis (FAP) is an autosomal-dominant inherited disorder characterized by the early onset of multiple adenomas throughout the colon and rectum. It represents approximately 1% of all CRC cases and is the most common gastrointestinal polyposis syndrome. Two major phenotypes have been described: classic FAP, characterized by hundreds to thousands of adenomas and a risk of CRC of almost 100%; and attenuated FAP (AFAP), characterized by a smaller number of polyps (20–100) and a later age of onset. In many patients, extracolonic manifestations are present, including gastric and duodenal polyps, desmoid tumors, thyroidal and brain tumors, osteomas, congenital hypertrophy of the retinal pigmented epithelium, supernumerary teeth, and epidermoid cysts. (Carballal *et al.* 2014)

FAP and attenuated FAP are caused by germline mutations in the *APC* gene, associated with the CIN pathway. *De novo APC* mutations are responsible for approximately 25% of FAP cases. In addition, approximately 20% of individuals with an apparent *de novo* APC mutation have somatic mosaicism The location of the germline mutation within APC has been associated with the severity of colonic polyposis, the degree of cancer risk, the age at diagnosis, the presence and/or frequency of extra-colonic malignant and non-malignant neoplasms and with the spectra of somatic mutations that lead to tumor formation. (Jasperson *et al.* 2010) (Albuquerque *et al.* 2002)

1.3.1.2 *MUTYH*-associated polyposis

MUTYH-associated polyposis (MAP) is an autosomal recessive disorder characterized by multiple adenomatous polyps, that is caused by mutations in the *MUTYH* (mutY DNA glycosylase) gene. The *MUTYH* gene is a base excision repair (BER) gene and acts in an autosomal recessive manner, with bi-allelic mutations necessary for expression of the phenotype. Although MAP is often associated with a polyposis phenotype, studies have reported bi-allelic mutation carriers displaying young-onset CRC in the absence of polyps, and excess of other extra-intestinal tumors. (Hisamuddin and Yang 2006) (Stoffel and Boland 2015)

1.3.1.3 Hamartomatous polyposis

Intestinal hamartomatous polyposis syndromes are characterized by an overgrowth of cells native to the area in which they normally occur. Peutz-Jeghers syndrome and juvenile polyposis syndrome are hamartomatous polyposis conditions that are both associated with an increased risk for colorectal cancer. (Carballal *et al.* 2014)

Peutz–Jeghers syndrome (PJS) is an autosomal dominant disorder characterized by melanocytic macules of the lips, buccal mucosa, and digits, multiple gastrointestinal hamartomatous polyps, and an increased risk of various neoplasms, including those of the gastrointestinal tract and in extra intestinal sites: breast, ovary, and testicles. Germline mutations of the gene encoding the tumor suppressor gene *STK11* (serine threonine kinase 11), are found in 70% of the patients with PJS. However, not all families

with PJS are linked to this *locus*, suggesting that additional genes are involved in its pathogenesis. (Hisamuddin and Yang 2006)

Juvenile polyposis syndrome (JPS) is a genetically heterogeneous, rare, childhood to early adult onset, autosomal dominant disease that is characterized by the presence of hamartomatous polyps throughout the entire GI tract, mainly in the colon (98%) and stomach (14%). Individuals with JPS have an increased risk for gastric cancer and CRC, with a lifetime risk of 40–50%. A diagnosis of JPS is considered for anyone who fulfills one of the following criteria: five or more colonic juvenile polyps; multiple juvenile polyps throughout the GI tract; or any number of juvenile polyps with a family history of JPS. Germline mutations in the *SMAD4* or *BMPR1A* (bone morphogenetic protein receptor type 1A) genes, both involved in the TGF-β signaling pathway, are identified in approximately 50% of cases. (Carballal *et al.* 2014)

1.3.1.4 Serrated polyposis

Serrated polyposis (SP), initially described as hyperplastic polyposis, is a rare condition characterized by a predisposition to serrated polyps, an increased risk for colorectal and possibly some other extracolonic neoplasms. Serrated polyps include different subtypes of lesions (hyperplastic, sessile serrated polyps/adenomas and traditional serrated adenomas) with a common histological feature: the 'sawtooth' appearance in the crypt epithelium. (Carballal *et al.* 2014)

Diagnosis of SP is based on specific clinical criteria outlined by the World Health Organization, that contemplate the number, size and location of serrated polyps in the colon. SP has largely been considered a genetic disease, but its genetic base and the pattern of inheritance remains unknown and both autosomal recessive and autosomal dominant patterns have been proposed. (Guarinos *et al.* 2012)

1.3.2 Nonpolyposis colorectal cancer syndromes

Hereditary non-polyposis colorectal cancer (HNPCC) is the most prevalent form of hereditary CRC, accounting for at least 2%to 3%of all CRCs, being a dominantly inherited syndrome characterized by significantly increased risks for colon cancer as well as for cancers of the endometrium, stomach, small intestine, hepatobiliary system, kidney, ureter, and ovary. (Wells and Wise 2017) (Lindor 2009a)

Establishing the Amsterdam criteria (AC) was essential to define HNPCC. The AC were introduced for uniform classification based on family history. In 1991 the International Collaborative Group on hereditary nonpolyposis colorectal cancer had established the AC-I (table 1.1). These guidelines turn out to be too strict and could exclude small families or those presenting HNPCC-associated tumors. In response to its stringency, in 1999, the extended AC-II criteria were proposed (table1.1). (Llor *et al.* 2005)

HNPCC families fall into 2 categories: Lynch syndrome, associated with hereditary defects in DNA MMR genes, and familial colorectal cancer type X, with no detectable mutation in DNA MMR genes. (Nieminen *et al.* 2011)

Table 1.1 - Amsterdam criteria I and II (adapted from: Vasen et al. 1999)

Amsterdam criteria I (1991)

There should be at least 3 relatives with CRC, plus all the following:

- One affected patient should be a first-degree relative of the other 2;
- At least 2 successive generations should be affected;
- At least 1 CRC should be diagnosed before the age of 50 years;
- Familial adenomatous polyposis should be excluded.

Amsterdam criteria II (1999)

There should be at least 3 relatives with an HNPCC-associated cancer (CRC, cancer of the endometrium, small bowel, ureter, or renal pelvis), plus all the following:

- One affected patient should be a first-degree relative of the other 2;
- At least 2 successive generations should be affected;
- Cancer in one or more affected relatives should be diagnosed before the age of 50 years;
- Familial adenomatous polyposis should be excluded in any cases of colorectal cancer.

1.3.2.1 Lynch syndrome

Lynch syndrome (LS) is the most common hereditary CRC syndrome, accounting for 1-3% of all CRC cases. LS is an autosomal dominant condition characterized by a germline mutation in one of the MMR genes and a predisposition to develop colorectal and endometrial cancer, among other less frequent tumors, at an early onset (~45 years). (Carballal *et al.* 2014)

LS is the result of a germline mutation in a class of genes involved in the DNA MMR system, previously described in 1.2.2, including *MSH2*, *MLH1*, *MSH6*, and *PMS2*. Approximately 80% of mutations are identified in the *MLH1* and *MSH2* genes, 10–12% in the *MSH6* gene, and *PMS2* may account for 2–3%. (Jasperson *et al.* 2010) Recently, *EPCAM* (Epithelial cell adhesion molecule) germline deletions in the 3′ region, that give rise to EpCAM-MSH2 fusion transcripts, and promoter hypermethylation of *MSH2* in tumors have also been implied in LS. (Kovacs *et al.* 2009)

The AC (I and II) are used in clinical practice to identify individuals at risk for Lynch syndrome who require further evaluation. If either of these criteria is met, genetic testing to detect a mutation in one of the mismatch repair genes can be performed. (Burt 2007) However, AC (I and II) turn out to be too strict and their limited sensitivity hampered decisions of choosing which patients should undergo genetic testing. (Umar et al. 2004) Therefore, the Bethesda guidelines (table 1.2) were developed, in which clinical and pathological features are used to determine which colon cancer specimens should be tested for microsatellite instability, with a goal of identifying patients with germline mutations in DNA MMR genes. (Boland 2005)

To access tumor microsatellite instability (MSI), a standardized panel of microsatellites referred to as the Bethesda panel is used, this panel comprises two mononucleotide (BAT25 and BAT26) and three dinucleotide microsatellites markers (D5S346, D2S123, and D17S250). Considerable MSI or MSI-high (MSI-H) is defined as MSI at ≥2 (40%) of the five specified markers, MSI-low (MSI-L) as MSI only at one marker, and microsatellite stable (MSS) when no instability is demonstrated at these markers. (Worthley *et al.* 2010)

Individuals with LS carry a heterozygous germline MMR mutation and, for malignancy to occur, a second copy of the affected MMR gene must be somatically inactivated. Loss of function of the MMR system may lead to DNA replication errors, especially in microsatellites, which can occur in tumor suppressor genes or proto-oncogenes leading to carcinogenesis. DNA replication errors are propagated through daughter cells, leading to errors in microsatellites, making them unstable (MSI-H). (Carballal *et al.* 2014) (Wells and Wise 2017)

Table 1.2 - The Bethesda guidelines (adapted from: Giardiello et al. 2014)

Bethesda Guidelines
Tumors from individuals should be tested for MSI in the following situations: **1.** CRC diagnosed before the age of 50 years; **2.** Presence of synchronous or metachronous CRC or other HNPCC-associated tumors (colorectal, endometrium, stomach, ovary, pancreas, ureter, renal pelvis, biliary tract, brain, small bowel, sebaceous glands, and kerotoacanthomas), regardless of age. **3.** CRC with MSI-high pathologic-associated features (Crohn-like lymphocytic reaction, mucinous/signet cell differentiation, or medullary growth pattern) diagnosed before the age of 60 years; **4.** CRC diagnosed in one or more first-degree relatives with an HNPCC-related tumor, with one of the cancers being diagnosed before the age of 50 years. **5.** CRC diagnosed in two or more first- or second-degree relatives with HNPCC-related tumors, regardless of age.

1.3.2.2 Familial colorectal cancer type X

In 2005, Lindor and co-workers reported a subset of families who fulfill AC but have no evidence of a DNA MMR defect, tumors did not present microsatellite instability and did not share the same cancer incidence as families with Lynch syndrome. Relatives in such families had a lower incidence of colorectal cancer than those in families with Lynch syndrome, and there was no evident risk for extracolonic tumors. These non-Lynch syndrome clusters of CRC were termed familial colorectal cancer type X (FCCTX). (Lindor *et al.* 2005) Table 1.3 compares and contrasts LS with FCCTX.

Table 1.3 – Contrasts between LS and FCCTX. (adapted from Lindor 2009b)

	LS	FCCTX
CRC		
Cancer risk	Very high	Modestly increased
Age of onset	45 years (average)	50-60 years
Usual location	Proximal colon	Distal colon
Polyps	Few	More
Other cancers		
Endometrial risk	Very high	Low
Other cancer sites	Many	None known
MMR genes		
Germline	Mutations found	No mutations found
Tumor	MSI	MSS
Tumor staining	Loss of MMR expression	Normal expression

As the "type X" label implies, the genetic etiology for FCCTX is largely unknown. Although the identification of the genetic basis of FCCTX has been a topic of intensive research, the cause for the increased risk of cancer in this group remains unknown and evidence suggests that it is a heterogenous grouping. (Carballal *et al.* 2014) (Carethers and Stoffel 2015) It likely includes some families that have a random aggregation of a common tumor; some families may be attributable to shared lifestyle factors, polygenic predisposition and some families likely have a yet to be defined single gene disorders. (Lindor 2009b)

Several linkage studies, next generation sequencing (NGS) and association studies, have been conducted to discover predisposing genes behind FCCTX. A germline mutation in *RPS20* (ribosomal protein S20) (c.147dupA), encoding an ribosomal RNA maturation protein has been identified in a FCCTX family. (Nieminen *et al.* 2014) The product of *RPS20* is required during the late steps of 18S ribosomal RNA formation, and *RPS20* has been associated to TP53 stabilization. Conversely, the constant activation of TP53 consecutive to ribosomal stress induced by *RPS20* mutation could favor, in the long run, the selection of cells that escape regulation by TP53. (Nieminen *et al.* 2014) Three different missense mutations (p.Val78Met, p.Gly484Ala and p.Ser326Phe) in the *SEMA4A* (semaphorin 4A) gene were also identified in three FCCX families. (Schulz *et al.* 2014) The p.Val78Met variant demonstrated significantly increased MAPK/Erk and PI3K/Akt signalling activation as well as cell cycle progression in HCT-116 CRC cells. (Schulz *et al.* 2014) However, the importance of this gene seems to be controversial since another study failed to find evidences to support variations in *SEMA4A* as a determinant of FCCTX risk. (Kinnersley *et al.* 2016) Another gene that has been referred in FCCTX is the *BMPR1A* gene (typically associated with juvenile polyposis), that encodes a protein receptor that

belongs to a family of serine/threonine kinases, where have been reported 2 variants (p.Gludel88 and c.68-10_68+14del). (Nieminen *et al.* 2011) A variant identified in the *POLD1* (DNA polymerase delta 1) gene (p.Pro300Leu) was also found within a FCCTX family. (Duarte 2015)

Despite the efforts made towards a better understanding and characterization of FCCTX, additional studies are still necessary to identify the genetic basis associated with this syndrome. Thus, providing a better screening, surveillance and maintenance of the families.

1.4 Molecular characterization and identification of novel susceptibility genes/variants for FCCTX

1.4.1 Molecular characterization of a group of FCCTX families

In order to achieve a better understanding of the FCCTX syndrome, the Colon Pathology Group from IPOLFG, E.P.E performed a study that aimed to characterize a group of FCCTX families at clinical and molecular level by evaluating the involvement of the CIN pathway in tumors from these families. (Francisco *et al.* 2011) For this purpose, 24 tumors from 15 FCCTX families were analyzed for LOH in known tumor suppressor gene (TSG) *loci* (*APC*, *TP53*, *SMAD4* and *DCC* (DCC netrin 1 receptor), methylation status of MMR genes and *MGMT* (O-6-methylguanine-DNA methyltransferase), and also *APC* and *KRAS* somatic mutations. These analyzes revealed two distinct molecular entities among the tumors of the FCCTX families: one, more prevalent, where tumors presented loss of TSG *loci* (72%) - TSG^+ (tumor suppressor gene positive), suggesting the involvement of the CIN pathway, and a lesser predominant one, with no evidence of TSG loss (28%) - TSG^- (tumor suppressor gene negative). Within the TSG^+ subtype tumors also presented frequent *APC* and *KRAS* somatic mutations, as well as frequent gene promoter methylation, while almost no promoter methylation was found in the TSG^- subtype. (Francisco *et al.* 2011)

1.4.2 Identification of novel FCCTX susceptibility *loci* through linkage analysis

Regarding this FCCTX comprehensive study, a project was developed by the Colon Pathology Group to map new genetic susceptibility *loci* to FCCTX. Thus, a genome-wide linkage study using 50,000 single-nucleotide polymorphisms (SNPs) array and a subsequent LOH analysis using microsatellite markers, was performed in the 2 most informative FCCTX families, one with TGS^+ tumors and one with TGS^- tumors. This analysis allowed the identification of 2 suggestive chromosomal regions (13q and 21q) within the TSG^+ family that may act as potential regions of susceptibility. While for the TSG^- family there was no significant results. (Pereira 2013) (Belo 2010)

Using more microsatellite markers to restrain the identified regions, as well as additional samples of FCCTX tumors, it was possible to determine a minimum region of LOH of 0.84Mb and of 1.3Mb, for regions 13q and 21q respectively (13q32-33 and 21q11). Considering the histological features of the analyzed FCCTX tumors, it was also found that LOH on chromosome 13 were more frequent in adenomas, whereas losses on chromosome 21 were more frequent in carcinomas, suggesting that

events in 13q occur at the level of tumor initiation while events in 21q appear at the level of tumor progression. (Pereira 2013)

Among the various genes contained in these regions of interest, some candidate genes for germline mutation analysis were selected according to the following criteria: genes expressed in the colon; genes encoding proteins involved directly or indirectly in cellular processes associated with tumorigenesis and gene that were not reported as the cause of other pathologies. (Saramago 2014) (Pereira 2014) Using these criteria a set of 13 candidate genes were chosen for germline mutation analysis within the 13q region and 3 for the 21q region. From this large set of genes only few remain as possible contributors for FCCTX, since several studies performed by the Colon Pathology Group have excluded or enhanced their possible involvement. One of the remaining genes that may be involved in FCCTX susceptibility is the Tripeptidyl peptidase 2 (*TPP2*) located in 13q33.1 (figure 1.3), whose study will be continued at the present work.

Figure 1.3 – Schematic view of chromosome 13. Localization of the minimum region of LOH identified through linkage analysis and the genomic location of TPP2. (Adapted from: Genecards (http://www.genecards.org/))

1.4.2.1 Tripeptidyl peptidase 2

TPP2 is an aminopeptidase that removes tripeptides from the free N terminus of longer peptides and is regarded as a housekeeping enzyme of eukaryotic cells. It constitutes the largest cytosolic protease complex in eukaryotes and it plays a role in the ubiquitin-proteasome pathway downstream of the proteasome as well as in the production and destruction of MHC (major histocompatibility complex) class I antigens and degradation of neuropeptides. (Preta *et al.* 2010) TPP2 has been reported to play a role in various biological processes, including antigen processing, cell growth, DNA damage repair and carcinogenesis, fat metabolism, feeding behavior, obesity and most recently in the control of MAPK3 (mitogen-activated protein kinase 3) and MAPK1 (mitogen-activated protein kinase 1) phosphorylation. (Wiemhoefer *et al.* 2015) (Peters *et al.* 2011)

The *TPP2* gene has two protein coding transcripts described in the Ensembl database (Ensembl: http://www.ensembl.org/index.html, accessed in 2017). One wild type (wt) formed by 29 exons and an alternative transcript, that has a different start codon and the presence of an additional 39 bp exon between exons 23 and 24, named 23a.

Molecular analysis of the *TPP2* gene was performed in previous studies of the project described above, 1.4.1.1. A mutational analysis at the mRNA level revealed a new transcript, not described to

date, resulting from an insertion of a 72 bp between exons 13 and 14. (Pereira 2013) This alternative 72 bp exon will be described as 13a throughout this project.

This novel transcript was evaluated in a study conducted by the Colon Pathology Group, where it was observed an increased expression of the 13a transcript in blood samples in 71% of the FCCTX individuals, 69% in LS individuals and 35% in healthy individuals. This increased expression was also followed by a considerable wt transcript reduction in 43%, 38% and 18% of the FCCTX, LS and healthy individuals, respectively. Thus, either 13a overexpression and reduction of wt, or just overexpression of 13a, was more frequent in individuals with FCCTX and LS than in the group of healthy individuals. (Pereira 2014) However, it was unknow if the expression of this transcript would still be observed in the colon. Thus, samples from carcinomas, adenomas and normal mucosa were included in the *TPP2* gene expression analysis. This study showed that the expression of the alternative transcript was significantly reduced when compared to the wt in tissue samples, showing a differential expression between blood and colon samples. (Saramago 2014) However this study was performed in a small and limited number of samples, thus, there is the need to include a larger number of samples in order to confirm these findings.

Given the role of TPP2 within the immune response, these data taken together suggest that this alternative transcript could play a role in FCCTX susceptibility, however further studies are needed to discriminate the role of the alternative transcript when compared to the wt transcript, both in blood and tissue samples.

1.4.3 Identification of novel FCCTX susceptibility variants by whole genome sequencing

Recent advances in the development of sequencing technologies, namely next-generation sequencing (NGS), have provided unprecedented possibilities for genetic analyses. The ability cost-effectively to generate genome-wide sequencing data with deep coverage in a short time frame is replacing approaches that focus on specific regions for gene discovery and clinical testing. While whole genome sequencing remains expensive for most applications, whole exome sequencing (WES), a technique that focuses only on the protein-coding portion of the genome, places many advantages of the emerging technologies into researcher's hands. Recent studies using this technology have uncovered new variants (and genes) for a number of previously unresolved Mendelian disorders. (Petersen *et al.* 2017) (Majewski *et al.* 2011)

NGS platforms share a common technological feature: massively parallel sequencing of clonally amplified or single DNA molecules that are spatially separated in a flow cell. This design is a paradigm shift from that of Sanger sequencing and has allowed scaling-up by orders of magnitude. In NGS, sequencing is performed by repeated cycles of polymerase-mediated nucleotide extensions. As a massively parallel process, NGS generates hundreds of megabases to gigabases of nucleotide sequence from a single instrument run. (Majewski *et al.* 2011)

Following the project that aimed the identification of novel genes for susceptibility to FCCTX, described in 1.4.1 and 1.4.1.1, the TSG^- family (designated L56) was selected for a WES analysis, since

no significant results were obtained through the linkage analysis. Genomic DNA samples from 5 patients of the family were selected for a WES analysis (using Agilent SureSelect Exome version 4 capture kit (*Agilent*) and Illumina TruSeq v3 protocol sample prep in a Hiseq2500 platform (*Illumina*), performed by Center for Biomics at the Erasmus University Medical Center- Rotterdam, Netherlands). This analysis aimed to identify the variants shared by the five individuals and evaluate their pathogenicity and association with the disease. Family pedigree as well as the identification of the patients subjected to WES analysis are represented in figure 1.4.

In order to maximize the identification of potential pathogenic germline mutations and given the high percentage of sporadic CRC in the population, the bioinformatic analysis was built to comprise all possibilities and therefore it was subdivided into five analyzes, each with a unique combination of the patients who shared variants (table 1.4). Thus, analysis 1 contained variants shared by all 5 patients, while the remaining analyzes contained variants shared only by four or three patients: analysis 2 excluded patient 4; analysis 3 excluded patient 2; analysis 4 excluded patients 2 and 4 and analysis 5 excluded patients 3 and 4. This exclusion of shared variants within some family members comes from the hypothesis that patients 2 and 3 may be potential phenocopies and therefore their phenotype may be a sporadic case and not attributed to the share of an inherited variant in the context of an FCCTX family. Patient 4 was also considered as a possible sporadic case because, when comparing the age of adenoma development in other relatives, this patient only started to develop adenomas at age 40.

The results of the five bioinformatic analyzes were provided by *Bioinf2Bio Lda.* using the GRCh38 reference genome and were organized into four outputs (format.*xlsx*), that contain descriptive information about the variants shared for each analysis, each output description can be found in table 1.5. Selection of possibly pathogenic variants that could justify the FCCTX phenotype in the L56 family was done using the variants provided by outputs 3 and 4 (table1.5). Although output 4 describes variants that present high impact according to the software SIFT (Kumar *et al.* 2009) and Polyphen-2 (Ramensky *et al.* 2002) and therefore describing the variants with the greatest pathogenic potential, output 3 analysis was also included in order to avoid the elimination of variants that may be equally important, excluded only by their scores in these software. (Magalhães 2016)

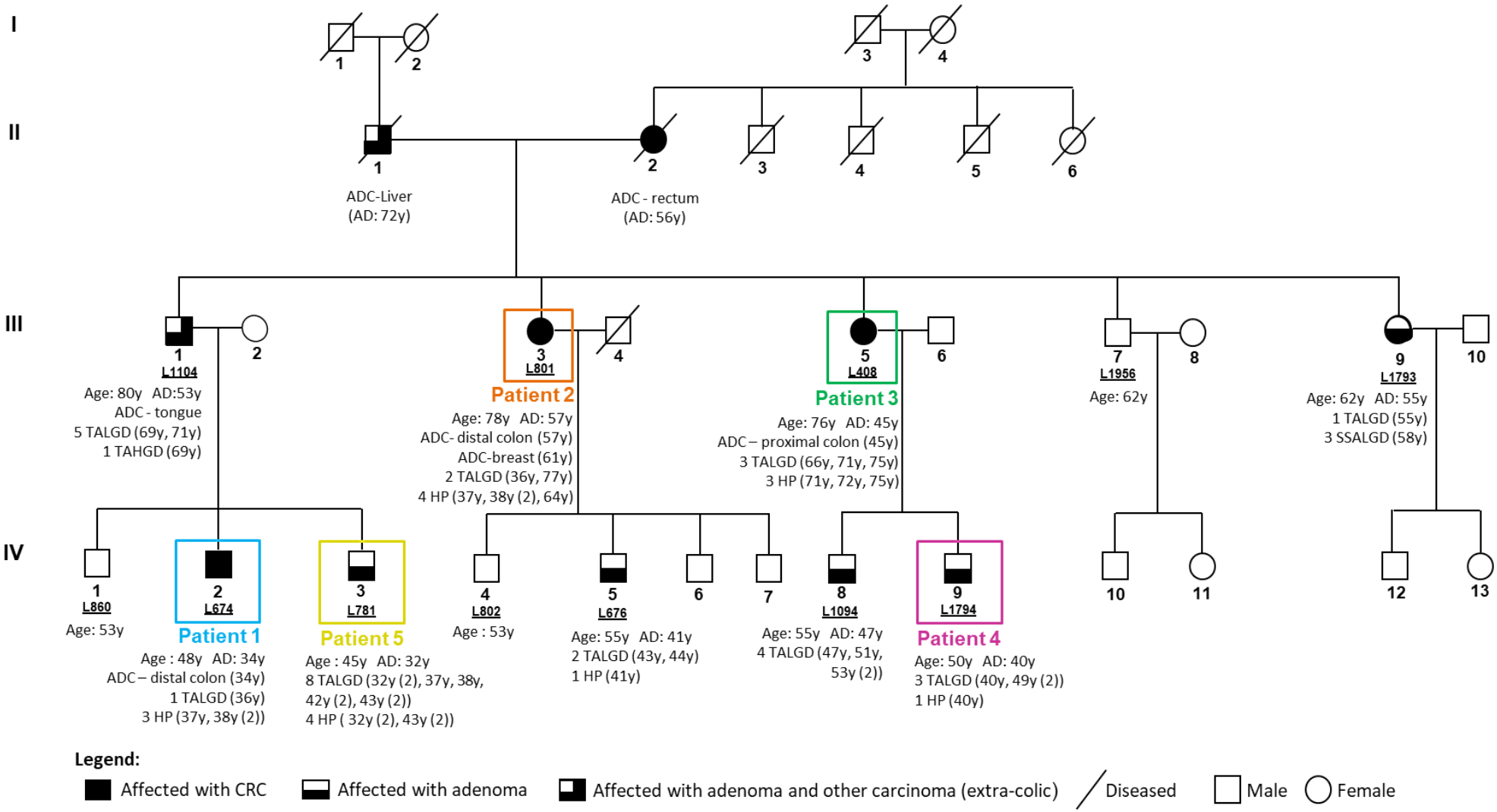


Figure 1.4 – Pedigree of the TSG- family - L56. Patients highlighted in colors were the ones selected for WES analysis. Only individuals with a laboratory identification number (LXXX or LXXXX) have biological material available. The age at diagnosis (AD, in years, y) and respective phenotypic characterization are described for the affected individuals. CA: current age; y: years; ADC: adenocarcinoma; TALGD: tubular adenoma with low-grade dysplasia; TAHGD: tubular adenoma with high-grade dysplasia; HP: hyperplastic polyp; SSALGD: sessile serrated adenoma with low-grade dysplasia.

Table 1.4 - Description of the strategy used in the 5 bioinformatic analyzes using the results obtained by WES for the selected patients from L56, and respective number of variants.

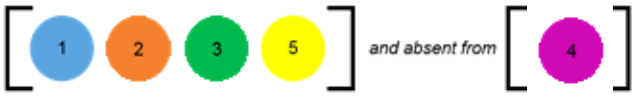
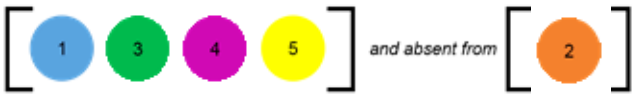
Analysis	Analysis Scheme	Analysis Description	Number of variants
1.		Shared by 1+2+3+4+5 (not phenocopies)	19 104
2.		Shared by 1+2+3+5 (not phenocopies) and absent from 4 (phenocopy)	3 198
3.		Shared by 1+3+4+5 (not phenocopies) and absent from 2(phenocopy)	2 036
4.		Shared by 1+3+5 (not phenocopies) and absent from 2 and 4 (phenocopies)	1 327
5.		Shared by 1+2+5 (not phenocopies) and absent from 3 and 4 (phenocopies)	2 254

Table 1.5 - Description of the 4 outputs obtained after bioinformatic analysis and total number of variants obtained for each output.

Output	Output Description	Total number of variants (Analysis 1+2+3+4+5)
1.	Annotation of all genomic variants matching protein coding transcripts with a: defined genotype; alignment and genotype quality ≥ 20; number of reads ≥ 20.	26 301
2.	Information from COSMIC databases of all genomic variants matching protein coding transcripts with COSMIC ids.	26 301
3.	Annotation of all genomic variants matching protein coding transcripts and with relevant predicted consequences (1 or a combinations of the following): frameshift; splice region; missense; splice acceptor variant; 5' UTR variant; 3' UTR variant; synonymous variant; start lost; stop gained.	18 642
4.	Annotation of all genomic variants derived from output 3 selected for high impact: SIFT: deleterious; PolyPhen: possibly/probably damaging.	604

1.4.3.1 Selection of variants identified through whole exome sequencing analysis

In order to select potential pathogenic variants shared by the affected individuals of the L56 family, for the subsequent segregation study, exclusion criteria were established (figure 1.5), which were applied to the variants described in outputs 3 and 4.

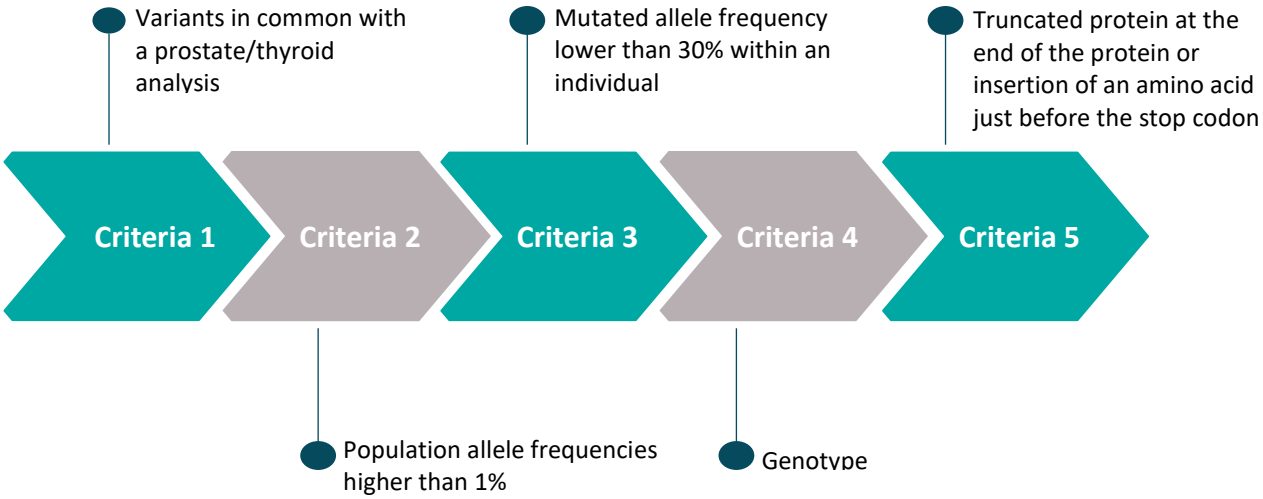


Figure 1.5 - Exclusion criteria used for the selection of variants for segregation studies with the disease within L56 family (Adapted from: Magalhães 2016).

The first criteria consisted in the exclusion of common variants obtained by bioinformatic analysis from three families of familial cancer (colon, thyroid and prostate) analyzed by WES that were included in the IPOFG, EPE project - Exome sequencing project/Familial and Individual cancer risk- identification of novel genes. The second criteria was used to eliminate variants that presented general population frequencies higher than 1%, since these are considered polymorphisms, followed by the exclusion of variants with a the mutated allele frequency, in two or more individuals, with values lower than 30%, which may represent possible artifacts that occurred during the sequencing process. The following criteria took into account the genotypes of the variants identified in the individuals of the family, that is, if they were detected in homozygosity or heterozygosity. Thus, and given the dominant transmission of CRC in the family, variants identified as homozygous for all or almost all individuals were excluded. Finally, the last criteria served to rule out truncating variants whose new stop codon was located at the end of the protein, since they gave rise to truncated proteins with similar size to the wild-type, therefore, with a low pathogenic potential. After applying these criteria to the two outputs (3 and 4), a total of 300 variants were obtained. (Magalhães 2016)

After the use of these exclusion criteria, *in silico* analysis was performed for all of the selected variants to restrict the number of variants and help predict their pathogenicity. General frequencies in the population were confirmed using 1000 Genomes (Auton *et al.* 2015), Exome Aggregation Consortium (ExAC) (ExAC: http://exac.broadinstitute.org/) (Lek *et al.* 2016) and Exome Variant Serve (EVS) (Exome Variant Server, NHLBI Exome Sequencing Project: http://evs.gs.washington.edu/EVS/)

databases. The SIFT/Polyphen-2 values were validated through the VEP software (Jian *et al.* 2014) (VEP: http://www.ensembl.org/Tools/VEP). And an additional analysis was performed using the Mutation Taster software (Schwarz *et al.* 2014) (Mutation Taster: http://www.mutationtaster.org/), in order to predict possible splicing consequences. Finally, a survey on protein function using PubMed database (NCBI Resource Coordinators 2017) (PubMed: https://www.ncbi.nlm.nih.gov/pubmed/) was performed, aiming to access if the proteins, encoded by the genes containing the selected variants, were involved directly or indirectly in cellular processes associated with tumorigenesis. (Magalhães 2016)

Thus, after this analysis, 41 variants, in different genes, were selected for segregation studies of the disease in the family. The segregation analysis highlighted 6 variants in different genes (*MTMR3* (Myotubularin related protein 3), *DUSP12* (Dual specificity phosphatase 12), *LGR6* (Leucine rich repeat containing G protein-coupled receptor 6), *NPR2* (Natriuretic peptide receptor 2), *TAS1R1* (Taste 1 receptor member 1) and *SMG7* (SMG7, nonsense mediated mRNA decay factor) that may affect cellular processes relevant to colorectal tumorigenesis. (Magalhães 2016)

Although a large number of variants were analyzed and refined along this selection sequence, this analysis was not finished, since analysis 1 for output 3 (section 1.4.1.2) was not completed and segregation studies, for two of the selected variants in *CACNA1S* (calcium voltage-gated channel subunit alpha1 S) and *GIMAP1* (GTPase, IMAP family member 1) genes, were not performed. Thus, this study will be carried out at the present work.

1.5 Objectives

The main goal of the present study was to identify and evaluate the potential role of candidate germline genetic variants and genes that may be involved in FCCTX susceptibility.

For this purpose, this project aimed to study specific variants in candidate genes for FCCTX susceptibility, obtained through WES and evaluate their potential pathogenicity, in order to identify mutations that may predispose to this syndrome. In addition, it was also intended to elucidate the possible contribution of the corresponding genes for FCCTX, by searching for germline mutations in index patients from FCCTX and FCCTX-like families, and to evaluate a possible involvement of copy number variations that could eventually segregate with the disease.

Also, it was engaged the involvement of the *TPP2* gene, previously identified as a possible candidate gene for FCCTX in another family, in healthy and affected FCCTX patients, by analysis of splicing isoforms and relative quantification by quantitative PCR.

2. MATERIAL AND METHODS

2.1 Biological samples

For segregation analysis of the selected variants, derived from analysis 1 of output 3 (described in 1.4.1.2) and from the previous identified candidate variants in *CACNA1S* and *GIMAP1* genes, in the L56 family, 12 genomic DNA samples obtained from peripheral blood were included, 9 of affected individuals and 3 of unaffected individuals. And for segregation analysis of the *MTMR3* and *TAS1R1* variants in a previously non-analyzed member of the family 1 DNA sample extracted from a formalin-fixed paraffin-embedded normal mucosa tissue was used.

34 index individuals that fulfill the CA or Bethesda guidelines 4 or 5 (criteria that suggests a family history), who fail to present any germline mutation in the MMR genes, were selected for germline mutation analysis in the candidate genes for FCCTX susceptibility. Germline mutation analysis of the *MTMR3* gene was performed in complementary DNA (cDNA) samples of all of the selected patients (analysis of cDNA was motivated by the very large coding region, thus allowing to sequence a small number of fragments), while for *TAS1R1*, *LGR6* and *DUSP12* (for the latter two only 5'UTR and beginning of first exon were analyzed at the DNA level) germline analysis was performed in 26, 12 and 11 genomic DNA samples of the selected patients. For the remaining coding sequence of *LGR6* and *DUSP12* cDNA primers were also designed due to their large coding sequence, as it was performed for *MTMR3*.

For frequency assessment of the c.1933C>T *MTMR3* variant, 50 DNA samples obtained from peripheral blood from Portuguese individuals were used.

In the *TPP2* gene expression analysis 18 cDNA samples obtained from peripheral blood were included: 6 samples from a FCCTX family (L55: that followed the TSG^+ entity), 6 from affected individuals (with carcinoma, adenoma or mutations in the MMR genes) and 6 from healthy individuals (2 of them used as calibrator).

For *TPP2* splicing analysis were included 6 cDNA samples obtained from peripheral blood and for the study of the alternative 13a exon by the protein truncation test technique 16 cDNA samples were used: 6 from patients of the L55 family; 5 with an increased 13a expression; and 5 with a decreased 13a expression.

An informed written consent for the following studies was obtained from all individuals.

2.2 Methods

2.2.1 Nucleic acid extraction/isolation

2.2.1.1 Genomic DNA and RNA isolation from peripheral blood

Genomic DNA was isolated from peripheral blood using the Puregene® Blood Core Kit (*Qiagen*), according to the manufacturer's instructions, and DNA concentration was assessed using Nanodrop2000 (*Thermo Scientific*). After quantification, DNA was diluted in double-distilled water

(ddH_2O) to obtain a working concentration of 80 ng/μL and was stored at -20°C until further use. All DNA samples used in this study had been previously extracted.

RNA extraction from peripheral blood was performed according to the method described for TRI Reagent® Solution (*Ambion, Life Techonologies*). After extraction, RNA concentration was assessed using Nanodrop2000 (*Thermo Scientific*) and afterwards 1μL of the extracted RNA sample together with 9μL of Orange G 1x (loading dye), was subjected to gel electrophoresis in a 0.8% (w/v) agarose gel, stained with ethidium bromide, using 6.5μL of Lambda/HindIII DNA Ladder (*Fermentas*) as a molecular weight marker, in order to assess sample integrity and detect possible genomic DNA contamination. RNA samples were stored at -80°C until further use.

All solutions used in the electrophoretic analysis and agarose gel preparation are described in Appendix D.

2.2.1.2 Genomic DNA isolation from formalin-fixed paraffin-embedded tissue

Isolation of genomic DNA from formalin-fixed paraffin-embedded (FFPE) tissue was performed using the phenol-chloroform method. Some of the reagents used in this isolation method absorb at the same wavelength as the DNA (260 nm), disabling an accurate quantification through spectrophotometric methods. Thus, DNA quantification was performed in a 0.8% (w/v) agarose gel by comparison with DNA standards of known concentrations, also evaluating sample integrity. Gel electrophoresis was performed as described in the previous section using 1μL of DNA instead of 1μL of RNA. Samples were diluted in ddH_2O to a working concentration of 80 ng/μL and stored at -20°C until further use.

The DNA isolation from FFPE samples was performed by a member of the colon pathology group.

2.2.2 Polymerase Chain Reaction

The polymerase chain reaction (PCR) is an enzymatic process that amplifies a DNA template to produce specific DNA fragments *in vitro*. The source of sample DNA can be either genomic or cDNA obtained from RNA samples through reverse transcription (RT).

PCR amplification involves two oligonucleotide primers that flank the DNA segment to be amplified and repeated cycles of 3 consecutive steps. The first PCR step is the denaturation of the double-stranded DNA by heating the sample reaction to approximately 94°C. In the second step, the sample is cooled, which allows the base-specific annealing of primers to complementary strands of DNA. In the third step, temperature is raised to achieve the optimal DNA polymerase temperature, which catalyzes the synthesis of new DNA strands using the primers as a starting point and deoxynucleotide triphosphates (dNTPs) present in the reaction to generate a sequence-specific complementary strand, in a process called elongation. Repetition of these 3 steps results in doubling the amount of targeted sequence (amplicon) in the reaction. The generation of PCR products, therefore, follows an exponential pattern (2^n, where n is the cycle number) and reaches a plateau after approximately 30 to 40 cycles, when most reagents have been consumed and no more PCR product is generated. In addition to two oligonucleotide primers, a DNA polymerase and dNTPs, the PCR reaction requires a buffer solution, providing a suitable chemical environment for optimum activity and stability of the DNA polymerase, and

a bivalent cation, usually magnesium (Mg^{2+}), a cofactor for thermostable DNA polymerases. (Ishmael and Stellato 2008)

2.2.2.1 PCR optimization

In order to obtain an efficient amplification of specific targets an optimization of the amplification conditions was performed. For each amplified fragment, the annealing temperature, Mg^{2+} concentration and the DNA polymerase kit were optimized.

As a starting point, a theoretical annealing temperature for each primer was determined by the equation 1, were A, T, G and C indicate the number of adenines, thymines, guanosines and cytosines, respectively, in each primer sequence.

$$\text{Primer annealing temperature (°C)} = 2 \text{ x } (\text{A} + \text{T}) + 4 \text{ x } (\text{G} + \text{C}) \qquad \text{(Equation 1)}$$

For each pair of primers, the starting annealing temperature was defined by equation 2.

$$\text{Annealing temperature (°C)} = \frac{(\text{Annealing temperature})_{Primer\ 1} + (\text{Annealing temperature})_{Primer\ 2}}{2} - 2 \quad \text{(Equation 2)}$$

This temperature was increased whenever a nonspecific PCR product was observed and decreased when there was no or very few product amplification. Subsequently, a solution containing magnesium ions was titrated to ascertain the optimum concentration to obtain a specific PCR product, since magnesium acts as cofactor of the DNA polymerase. For most of the PCR fragments, conditions were optimized using the enzymatic Biotaq™ kit (*Bioline*), however, in cases where the target region had a high GC content, which hinders the DNA double-strand denaturation process the GC-RICH PCR System kit (*Roche*) was used to improve the amplification of the fragment. For fragments with a size higher or similar to 2000bp, is was necessary to resort to the Expand Long Template PCR System (*Roche*). This PCR system contains thermostable Taq DNA Polymerase with proofreading activity that help to amplify longer templates than the ones produced by the regular DNA polymerase kit (that normally can only amplify products with sizes around 1000bp).

2.2.2.2 Primer design

Since proper primer design is necessary for successful DNA/cDNA amplification, primers were designed considering several guidelines. From the reference gene sequence (Built GRCh38, NCBI: https://www.ncbi.nlm.nih.gov/) of the gene of interest, small sequences between 18 and 25 nucleotides were chosen, in order to flank the desired DNA sequence. *In silico* analysis of the primers was performed using the software NetPrimer (*PREMIER Biosoft*: http://www.premierbiosoft.com/netprimer). This tool facilitates the selection of an optimal primer pair, since all primers are analyzed for secondary structures including hairpins, repeats, self and cross dimers and the stability of this structures. It also calculates primer melting temperature (Tm) and GC content (the number of G's and C's in the primer as a percentage of the total bases). Therefore, primers were designed in order to reduce the formation of

secondary structures and to have a GC content between 40 and 55%, that should be similar between the pair, as well as melting temperatures. After obtaining a primer pair that fulfilled these guidelines, the size of the fragment obtained, as well as the specificity of the primers within the human genome (to ensure that the primers were only annealing with the desired sequence), were accessed using the software Primer-Blast (NCBI: https://www.ncbi.nlm.nih.gov/tools/primer-blast).

All primers used in this study are described in the Appendix A.

2.2.2.3 DNA/cDNA amplification by PCR

All DNA and cDNA samples used in this study were amplified in a reactional volume of 12.5µL.

For samples that were amplified with the Biotaq™ kit (*Bioline*), the reactional mixture consisted of 1µL of template DNA (80 ng/µL) or cDNA, 0.3 µL of each forward and reverse primer (10 pmol/µL), 1µL of dNTPs (200 mM, *Illustra™*, *GE Healthcare*), 1.25µL of the reaction solution buffer, 0.08µL of the DNA polymerase (5 U/µL), the optimized concentration of the $MgCl_2$ solution and ddH_2O to make up the volume.

For the GC-RICH PCR System kit (*Roche*) PCR amplification reaction contained: 1µL of the template; 0.2µL of each forward and reverse primer (10 pmol/µL), 1µL of dNTPs (200 mM, *Illustra™*, *GE Healthcare*); 2.5µL of reaction buffer; 0.25µL of DNA polymerase mix (2 U/µL); the optimized concentration of a $MgCl_2$ solution; and ddH_2O to make up the volume.

And, when using the Expand Long Template PCR System (*Roche*) the reaction was performed using: 1µL of the template; 0.3µL of each forward and reverse primer (10 pmol/µL), 1.25µL of dNTPs (200 mM, *Illustra™*, *GE Healthcare*); 2.5µL of a reaction buffer solution containing Mg^{2+} ions; 0.19µL of DNA polymerase mix (5 U/µL); and ddH_2O to make up the volume.

The reaction buffer solution, DNA polymerase or DNA polymerase mix and $MgCl_2$ solution were provided by the chosen PCR kit.

All PCR reactions were performed in a Veriti® Thermal Cycler (*Applied Biosystems*) or UNO96 thermocycler (*VWR*), according to the conditions described in Appendix B and C.

2.2.2.4 Agarose gel electrophoresis

Agarose gel electrophoresis is a separation method for DNA analysis based on their size and charge. Electrophoresis uses an electrical field to move the negatively charged DNA through an agarose gel matrix towards a positive electrode. Because DNA has a uniform mass/charge ratio, DNA molecules are separated by size within the gel in a pattern such that the distance traveled is inversely proportional to the log of its molecular weight. DNA may be visualized using fluorescent dyes that intercalate into the double-stranded DNA, like ethidium bromide. (Lee *et al.* 2012) Therefore, the PCR efficiency was evaluated by this methodology, in order to observe if there was a specific amplification of the fragment of interest.

When PCR conditions were still in the optimization process described in 2.2.2.1, the efficiency of the PCR reaction was monitored by the addition of 3µL of a loading dye -Orange G 5x- to the total volume of the amplification product and loaded into the gel. When the PCR conditions were optimized,

the efficiency of the PCR reactions of the samples of interest was performed with the addition of 7μL of a loading dye (Orange G 1x) to 3μL of the amplified product, loading the resulting 10μL into the agarose gel. When the fragments had an expected molecular weight below 1000bp electrophoresis was performed in a 2% (w/v) agarose gel, stained with ethidium bromide, followed by separation at 140 volts (V) for 30 min in TBE (Tris-Borate-EDTA) 1x buffer solution, using 6.5μL of the molecular marker GeneRuler 50bp DNA Ladder (*Thermofisher Scientific*). Then the fragments presented a molecular weight around 1500bp electrophoresis was carried out in the same conditions, using an extra molecular weight marker - 1 kb DNA Ladder (*Promega*). For higher molecular weight fragments (≥ 2000bp) electrophoresis was performed in a 1.2% agarose gel (w/v), stained with ethidium bromide, run at 140 V for about 1 hour (this time could vary with the size of the fragments), as a molecular weight marker 6.5μL of both 1 kb DNA Ladder (*Promega*) and Lambda DNA/HindIII Marker (*Thermo Scientific*) were added.

After electrophoresis, the gel was exposed to ultraviolet radiation in a BioDocAnalyze transilluminator (*Biometra*).

All solutions used in the electrophoretic analysis and agarose gel preparation are described in Appendix D.

2.2.3 Sanger sequencing

Sanger sequencing, also known as chain-termination sequencing or dideoxy sequencing, has been the powerhouse of DNA sequencing since its invention in the 1970s. The process is based on the detection of labelled chain-terminating nucleotides that are incorporated by a DNA polymerase during the replication of a template. (GATC Biotech 2017) Dideoxy sequencing is based on synthesis of DNA strands that are complementary to a template DNA strand. The sequencing reaction uses normal dNTPs and modified dideoxynucleoside triphosphates (ddNTPs) for strand elongation. The ddNTPs lack a 3′-OH group that is required for the formation of a phosphodiester bond between two nucleotides, causing DNA polymerase to stop DNA extension whenever a ddNTP is incorporated. The resulting DNA fragments are subjected to capillary electrophoresis, where the fragments flow through a gel-like matrix at different speeds according to their size. Because each ddNTP is labeled with a different fluorescent dye (each of which fluoresces at a different wavelength), the sequencing can be done as a single reaction. As the DNA fragments exit the capillary electrophoresis gel, the dyes are excited by a laser and the emitted light is detected. The result is a electropherogram where bases are represented by a sequence of colored peaks. (Watson *et al.* 2012)

2.2.3.1 Purification of amplified PCR products

After PCR amplification, the PCR product contains leftover dNTPs and primers which were not consumed during amplification. These elements should be removed before sequencing reaction, in order to ensure clean and readable DNA sequences. PCR purification can be performed enzymatically or by gel purification.

Enzymatic purification was performed by adding 1µL of Thermosensitive Alkaline Phosphatase - FastAP (1 U/µL, *Thermo Scientific)* and 0.5µL of Exonuclease I-ExoI (20 U/µL, *Thermo Scientific*) directly to the PCR product. The samples were then subjected to a 15 minute incubation at 37°C (optimal enzyme activity temperature) followed by a 15 minute enzyme inactivation at 85°C in a Veriti® Thermal Cycler (*Applied Biosystems*).

In cases where it was not possible to obtain a specific band, for instance, when the PCR optimization process was inefficient to obtain a specific fragment or, when using cDNA samples, the alternative splicing phenomenon produced more than one fragment (isoform), the band of interest was excised from the agarose gel using a scalpel. The amplified product was then purified using the Cut&Spin Gel Extraction Columns (*GRiSP*) kit. Briefly, the band excised from the agarose gel, was placed in the column provided by the kit and centrifuged at 6000g at room temperature for 10 minutes. To evaluate the purification yield, the eluate was subjected to a gel electrophoresis in a 2% (w/v) agarose gel, as described in section 2.2.2.4, with the addition of 8µL of Orange G 1x to 2µL of the purified product.

2.2.3.2 Sequencing reaction

For each amplified fragment to be sequenced, a reactional mixture was made with the following elements: 2 µL of the forward or reverse primer (1.6 pmol/µL); 2µL of buffer solution - Buffer Sequencing 5x, BigDye® Terminator v1.1 (*Applied Biosystems*) - to maintain a constant pH; a variable amount of mix BigDye™ Terminator v1.1 Cycle (*Applied Biosystems*), depending on the size of the fragment to be sequenced; a variable amount of the purified PCR product, which depends on the yield of the PCR reaction observed on agarose gel and the size of the fragment to be analyzed; and ddH_2O to make up a final 20µl volume. Sequencing reaction was performed on a Veriti® Thermal Cycler (*Applied Biosystems*) using the conditions described in the Appendix C (Table C.5).

The products obtained by the sequencing reaction were stored at 4°C for a maximum of 24 hours until DNA precipitation and purification.

2.2.3.3 DNA precipitation and purification after sequencing reaction

The method used for DNA precipitation and purification, after sequencing reaction, is based on a salting out reaction recommended for the BigDye™ Terminator kit v1.1 Cycle Sequencing Kit (*Applied Biosystems*).

This protocol aims to obtain a purified DNA pellet without contaminants that could interfere with capillary electrophoresis and uses three reagents: absolute ethanol, sodium acetate and EDTA (Ethylenediamine tetraacetic acid). Absolute ethanol and sodium acetate neutralize nucleic acids charges and reduce their solubility, promoting DNA precipitation, while EDTA, a magnesium ions chelator, inhibits the enzyme used in the sequencing reaction. Ethanol 70% (v/v) was used to wash the DNA pellet, which was dried at 37ºC. DNA pellet was stored at 4°C until capillary electrophoresis was performed.

The detailed protocol for this procedure is described in Appendix E.

2.2.3.4 Preparation of sequencing reaction products for capillary electrophoresis

Dry DNA pellets, obtained after precipitation and purification of the sequencing reaction, were resuspended in 17µL of HI-DI formamide (*Applied Biosystems*), used as a capillary injection solvent. Samples were then homogenized in a vortex and the total volume was transferred to a 96-well plate (Platemax, *Axygen*) suitable for the automatic sequencer. The plate was sealed and placed in thermocycler at 95°C for 5 minutes to denature all samples and then incubated on ice for 1-2 minutes. Finally, centrifugation was carried out for 2 minutes at 1200 rpm, for sample deposition and elimination of possible air bubbles. The seal was then removed and the sample plate was placed on a ABI Prism™ 3130 Genetic Analyzer (*Applied Biosystems*) sequencing instrument, where the capillary electrophoresis was run at 50°C and at 15 KV.

2.2.3.5 Analysis of results

Capillary electrophoresis results were obtained in the form of an electropherogram generated by the Sequencing Analysis software 3.4.1 (*Applied Biosystems*). The sequence obtained was manually compared to the reference sequence of the target gene, withdrawal from of the Ensembl database.

2.2.4 cDNA synthesis by reverse transcription reaction

Reverse transcription (RT) is the synthesis of complementary DNA (cDNA) from a single-stranded RNA template, in a process catalyzed by the reverse transcriptase enzyme. First, random primers are used to bind to the mRNA strands so that the reverse transcriptase enzyme is able to convert the RNA into cDNA and then the cDNA is amplificated by PCR.

Each cDNA synthesis comprised two reaction mixtures. The first, for a final volume of 7.75µL, contained: 0.5µL of hexamers (3 µg/µL, *Roche*), that function as random primers; a variable volume of RNA from each sample (equivalent to 1 µg of RNA); and ddH_2O treated with diethylpyrocarbonate (DEPC) (*MERK*), to make up the volume. This first reaction was incubated at 70°C for 10 minutes in a Veriti® Thermal Cycler (*Applied Biosystems*) and then placed on ice in order to add the second reaction mixture, consisting of: 4µL of 5x First Strand Buffer (*Invitrogen*), to maintain a constant pH; 4µL of dNTPs (100mM each, Illustra™ *GE Healthcare*); 2µL of dithiothreitol (DTT) (0.1M, *Invitrogen*), as an enzymatic stabilizer; 1µL of Superscript® II Reverse Transcriptase (200 U/µL, *Invitrogen*); 0.75µL of RnaseOut™ Recombinant Ribonuclease Inhibitor (40 U/µL, *Invitrogen*), for the inactivation of possible RNases present in the reaction; and 0.5µL of ddH_2O treated with DEPC (*MERK*) to make up a final volume of 12.25 µL. Together this 2 reactional mixtures make a final volume of 20µL. The tubes were then placed back into the thermal cycler, containing the second reaction mixture, to terminate the reaction according to the conditions described in the Appendix C (Table C.6). The synthesized cDNA samples were then stored at -20°C.

2.2.5 Quantitative PCR

In conventional PCR, the amplified product is detected by an end-point analysis, by running DNA on an agarose gel after the reaction has finished. In contrast, quantitative PCR (qPCR) allows the accumulation of amplified product to be detected and measured as the reaction progresses, allowing

the quantification of a starting template with accuracy and high sensitivity over a wide dynamic range. This is made possible by including in the reaction a fluorescent molecule that reports an increase in the amount of DNA with a proportional increase in fluorescent signal, like DNA-binding dyes (e.g. SYBR® Green) and fluorescently labeled sequence specific primers or probes (e.g. TaqMan® probe). Specialized thermal cyclers equipped with fluorescence detection modules are used to monitor the fluorescence as amplification occurs. The measured fluorescence reflects the amount of amplified product in each cycle. Because reactions are run and data are evaluated in a closed-tube system, contamination is reduced and the need for post amplification manipulation is practically eliminated. (Bio-Rad Laboratories 2006)

The qPCR amplification plot follows two phases, an exponential phase, were the amount of PCR product approximately doubles in each cycle and a plateau phase, when the reaction components are consumed, and ultimately one or more of the components becomes limiting (figure 2.1).

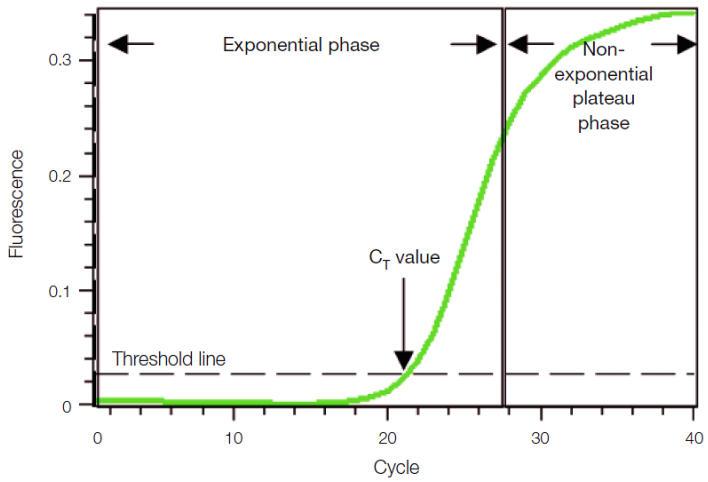


Figure 2.1 – Graphic representation of a qPCR amplification plot (From: Real-Time PCR Applications Guide, Bio-Rad Laboratories).

Even though product accumulates exponentially, in an initial phase, the fluorescence remains at background levels, and cannot be detected. Eventually, enough amplified product accumulates to yield a detectable fluorescent signal. The cycle number at which this occurs is called the threshold cycle, or Ct (figure 2.1). Because Ct values are measured in the exponential phase, when reagents are not limited, real-time qPCR can be used to calculate the initial amount of template present in the reaction. These values depend on the equipment used baseline and the amount of template present at the start of the reaction. With a higher amount of initial template, few amplification cycles will be required to produce enough product to give a fluorescent signal, thus providing a lower Ct value. (Bio-Rad Laboratories 2006) Therefore, the numerical value of the Ct is inversely related to the amount of amplicon in the reaction.

Quantification of gene expression by qPCR can be determined by two methods: absolute and relative quantification. Absolute quantification is used to quantify unknown samples by interpolating their quantity from a standard curve. Relative quantification is used to analyze changes in gene expression

or copy number variation in a given sample relative to a reference sample. In this study, only the relative quantification method was used. (Bio-Rad Laboratories 2006)

A widely-used method for displaying the relative expression of a gene is the comparative method of Ct, also referred to as the $2^{-\Delta\Delta CT}$ method. This method relies on 2 major components, a calibrator sample and a reference gene. One of the samples is usually chosen as the calibrator, and the expression of the target gene in all other samples is expressed as an increase or decrease relative to the calibrator and a reference gene is one whose expression level is constant across all test samples and whose expression is not affected by the experimental treatment under study (e.g. *GAPDH* (glyceraldehyde-3-phosphate dehydrogenase), ***β-actin***, etc.). (Bio-Rad Laboratories 2006) For this method, the Ct of the target gene needs to be normalized to that of the reference gene (ref), for both the test sample(s) and the calibrator sample(s) (more than 1 calibrator sample can be used), according to equation 3:

$$\Delta Ct = (Ct_{\text{target gene}} - Ct_{\text{ref}}) \quad \text{(Equation 3)}$$

Then, the ΔCt of the test sample(s) needs to be normalized to the ΔCt of the calibrator(s), equation 4, and finally calculate the expression ratio, equation 5 (Schmittgen and Livak 2008):

$$\Delta\Delta Ct = \Delta Ct_{(test)} - \Delta Ct_{(calibrator)} \quad \text{(Equation 4)}$$

$$2^{-\Delta\Delta Ct} = \text{Normalized expression ratio} \quad \text{(Equation 5)}$$

The result obtained is the fold increase (or decrease) of the target gene in the test sample relative to the calibrator sample and is normalized to the expression of a reference gene. Normalizing the expression of the target gene to that of the reference gene compensates for any difference in the amount of sample. (Bio-Rad Laboratories 2006)

2.2.5.1 Optimization of qPCR conditions and qPCR reaction

The $2^{-\Delta\Delta Ct}$ method assumes that both target and reference genes are amplified with efficiencies near 100% and within 5% of each other so, before using this method (or for any qPCR assay), it is essential to determine the amplification efficiencies of the target and the reference genes.

To calculate the amplification efficiencies for each target/reference gene, serial dilutions of a reference cDNA sample (cell line HCT 116 (*ATCC® CCL-247™*)) were performed, giving a final concentration of 40, 20, 10, 5 and 2.5 ng/μL. Then, reactions were performed in duplicate on a 96 well plate (*ABI Prism*™ Optical 96 Well Reaction Plate), for a final volume of 15μL, having the following composition: 0.75μL of each primer (foward and reverse) either at 5 or 7.5 pmol/μL, to access optimal primer concentration; 7.5μL of Power SYBR® Green PCR Master Mix (*Applied Biosystems*); 4μL of ddH_2O; and 2μL of cDNA (from the serial dilutions), except in the negative control where the cDNA volume was replaced by ddH_2O.

Afterwards, a standard curve was constructed by plotting the log of the starting quantity of template (40, 20, 10, 5 or 2.5 ng/μL) against the medium Ct value obtained during amplification of each

dilution. From this curve, an equation of the linear regression line was withdrawal and the qPCR efficiency was calculated through equation 6, where m is the slope of the line.

$$Efficiency\ (E) = 10\left(-\frac{1}{m}\right) \quad \text{(Equation 6)}$$

After qPCR optimization, all the reactions were performed to a final volume of 15µL, as described above, using each primer pair and cDNA at the optimized concentration. The reactions were performed in triplicates on the ABI PRISM 7900HT Sequence Detection System (*Applied Biosystems*) and analysis of the results was performed by SDS 2.4 software (*Applied Biosystems*). After the reaction, the Ct values acquired by the software were imported into an Excel spreadsheet and the relative expression analysis was performed according to the $2^{-\Delta\Delta Ct}$ method (described above). Statistical analysis was performed using GraphPad Prism software (version 7.03), using the mean and standard deviation (SD) of the triplicates of each experiment.

2.2.6 Identification of novel FCCTX susceptibility variants by WES

As previously detailed in 1.4.1.2 section, in a previous study, a WES analysis was carried out on 5 affected relatives of the L56 family, to identify candidate genes for FCCTX susceptibility and subsequent bioinformatic analysis for variant identification and stratification according to their possible pathogenicity. Then, this analysis was refined to select and reduce the number of candidate variants for segregation studies with the disease. Although a large number of variants were analyzed and refined, this analysis was not finished, since analysis 1 for output 3 (section 1.4.1.2) was not completed and segregation studies, for two of the selected variants in *CACNA1S* and *GIMAP1* genes, were not performed.

2.2.6.1 Selection of variants identified through WES in the L56 family

The selection of variants, using analysis 1 included in output 3 as a starting point, was made using the guidelines for variant selection previously described in 1.4.1.2.1. After the application of exclusion criteria, the initial 12603 variants were reduced to 116. Which after *in silico* analysis and surveys on protein function, were narrow down to only 2 variants for segregation studies with the disease in the family.

2.2.6.2 Segregation analysis of the selected variants with the disease in the L56 family

After variant selection, segregation studies of the selected variants with the disease in the L56 family were performed, adding the variants already found in the *CACNA1S* and *GIMAP1* genes.

Throughout this study, the colon pathology group was able to gain access to a formalin-fixed paraffin-embedded (FFPE) normal mucosa tissue from the family individual II.2 (now referred as

CAs1555), that was not included in the previous analysis since no genomic DNA was available from this patient. So, a segregation analysis of the already reported variants was performed.

Since the treatment to which the FFPE tissue is subjected until the inclusion in paraffin causes DNA fragmentation, the primers used for the analysis of patient CAs1555 were designed to amplify a fragment with a size inferior to 200bp and, for the segregation of the new variants the primers were designed to amplify a fragment with a size inferior to 600bp, all following the guidelines defined in 2.2.2.2 and are described in Appendix A (Tables A.1 and A.2).

For segregation analysis with the disease of the variants found in the *CACNA1S* and *GIMAP1* genes, plus the 2 found in analysis 1 of output 3, PCR amplification was performed followed by Sanger sequencing of each of the fragments containing the variants, for 12 genomic DNA samples from the L56 family (9 affected individuals and 3 unaffected individuals), as described in sections 2.2.2 and 2.2.3.

Since the FFPE tissue was severely degraded the segregation analysis of patient CAs1555 could only be performed for *MTMR3* and *TAS1R1* genes. This analysis was performed as described above but for 1 DNA sample extracted from a FFPE normal mucosa tissue as described in 2.2.1.2.

PCR conditions used for amplification are described in Appendix B (Tables B.1 and B.2) and C (Tables C.1 and C.2).

2.2.6.3 Identification of novel FCCTX susceptibility genes by copy number variation analysis based on WES data

Copy number variations (CNVs) represent a class of variation in which segments of the genome can be duplicated (gains) or deleted (losses). Research of inherited CNVs have been associated with many disease conditions, including cancer and inherited disorders. Thus, this technique can be a promising approach in the identification of novel FCCTX susceptibility genes.

2.2.6.3.1 Copy number variation analysis in the L56 family

Although a WES analysis had already been performed for the L56 family in the search of candidate genes for FCCTX susceptibility another approach was used – a copy number variation (CNV) analysis in the family. Thus, from the BAM file retrieved from the WES analysis from the L56 family and another family with thyroid cancer (included in the IPOFG, EPE project - Exome sequencing project/Familial and Individual cancer risk- identification of novel genes. (described in 1.4.1.2.1)), a bioinformatic analysis was performed by Bioinf2Bio Lda. Briefly, this CNV bioinformatic analysis was done in order to encompass chromosomal regions of known genes, by comparing the number of reads for a given amplicon in each sample of the L56 family, to the average number of reads observed for the control group (thyroid samples all together). Then, amplicons were classified as having a 'Total Deletion' scenario, when the number of reads for a given amplicon in each L56 family sample were ≤60% of the average number of reads observed for the control group. The resulting number of amplicons with a 'Total Deletion' scenario, comparing with the control group, was 296.

2.2.6.3.1.1 Selection of amplicons derived from CNV analysis

The amplicons identified as a 'Total Deletion' from the bioinformatic analysis, were subjected to an additional analysis for further refinement of amplicons of interest. So, for all the 296 amplicons with a 'Total Deletion' scenario, the number of minimum reads in the control group was analyzed against the number of reads of each of the 5 samples from L56 family, to exclude amplicons were the number of reads of each sample, did not present significant differences when comparing to the minimum number of reads in the control group (i.e. in cases where the number of reads of 1 or more samples were highly similar (although inferior) to the number of minimum reads in the control group). After this selection, 210 amplicons remained. And given the chromosome coordinates (GRCh38) of the amplicons, a survey was conducted using the Ensembl database, where, for each of the 210 amplicons, was reported whether they include: only intronic sections of the gene; parts of exonic and intronic sections of the gene, and how much of the exonic section was included in the amplicon; a full exon; a full exon and parts of the adjacent exon; or more than one full exon.

2.2.7 Germline mutation analysis of candidate genes for FCCTX susceptibility

A previous study, described in 1.4.1.2 and 1.4.1.2.1, identified 6 genetic variants in the *MTMR3*, *DUSP12*, *LGR6*, *TAS1R1*, *NPR2* and *SMG7* genes, as possible FCCTX contributors. To access if these genes, here referred as candidate genes, could play an important role in FCCTX, a germline mutation analysis was conducted in a series of index patients from FCCTX and FCCTX-like families that fulfill the CA or Bethesda guidelines 4 or 5 (criteria that suggests a family history), who fail to present any germline mutation in the MMR genes.

Since all candidate genes present an extensive nucleotide sequence, apart from *TAS1R1* (that presents only 6 exons and mutation analysis was performed using DNA samples), the mutation analysis was performed at a cDNA level instead of DNA, which allowed a reduction of the number of primers used and fragments to be amplified.

2.2.7.1 Primer design for amplification of candidate genes for FCCTX susceptibility

For *TAS1R1* mutation analysis, primers were designed at a DNA level, following the guidelines described in 2.2.2.2, to amplify each one of the 6 exons. Since the last exon (6) as a length of 1105bp, an additional forward primer (with an approximate distance of 500bp of the initial forward primer) was designed to overcome the possibility that the fragment could not be accurately sequenced using only the original forward and reverse primers of this fragment.

For *MTMR3*, *DUSP12*, *LGR6*, *NPR2* and *SMG7*, primers were designed for cDNA amplification in order to reduce the number primer of primers and fragments to be amplified, since the majority of these genes presents more than 10 exons (e.g. the *MTMR3* isoform with the highest exon number as 20 exons). Primers were generated following the guidelines described in 2.2.2.2, to amplify the coding regions of all the known protein coding transcripts (obtained from the Ensembl database) of each candidate gene, in overlapping cDNA fragments, like shown in figure 2.2. Given figure 2.2 as an example, multiple sets of 'external' primers were designed to amplify fragments of approximately 1000bp

to cover all the protein coding region of the gene. The first set of primers of each gene was always designed to incorporate the translation initiation codon: ATG. However, for *DUSP12* (fragment A) and *LGR6* (transcript id ENST00000439764 – fragment E)) the first set of primers had to be designed for DNA amplification, because the sequences (at cDNA) prior to the ATG sequence failed to ensure a good primer design. Each of the fragments, generated with the external primers, were designated at an alphabetic order: first primer pair: A; second primer pair: B, etc., and for each fragment, an additional set of 'internal' primers were designed, a forward and a reverse primer in the middle of the resulting fragment, to prevent cases where the fragment is too long to be accurately sequenced using only the 'external' primers.

All the primers used in the mutational analysis of the candidate genes are described in the Appendix A (Tables A.3, A.4 and A.5).

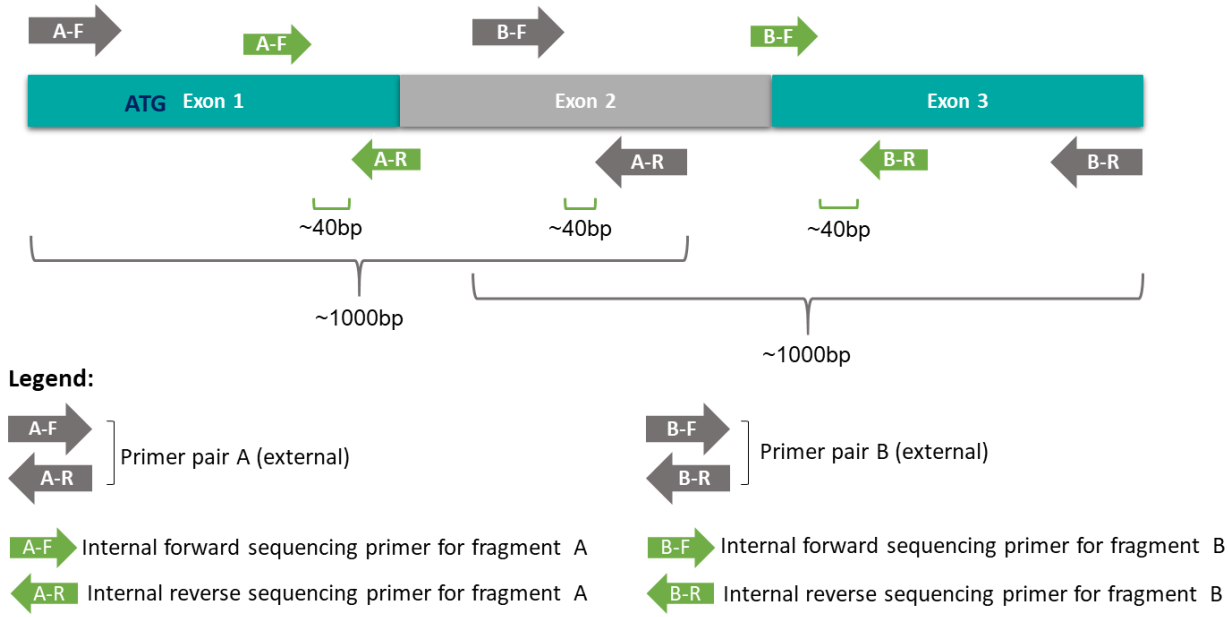


Figure 2.2 – Schematic representation of how the primers for cDNA amplification of the candidate genes were designed. The number and length of the exons are merely representative. F: Forward, R: Reverse, ~: approximately.

2.2.7.2 Mutation analysis of candidate genes for FCCTX susceptibility

Germline mutation analysis of the candidate genes comprised three steps. First cDNA synthesis from patient's RNA samples (for fragments that were designed to be amplified at a cDNA level), was performed as described in 2.2.4, followed by its PCR amplification (of the cDNA or DNA samples) and sequencing analysis according to the methods described in sections 2.2.2 and 2.2.3.

Do to time restrictions, this mutation analysis could not be performed for all the selected genes, thus the analysis was initiated with *MTMR3*, that was one of the most promising genes for FCCTX susceptibility. (Magalhães 2016) Followed by *TAS1R1*, *LGR6* and *DUSP12*. For *MTMR3*, 11 cDNA samples were analyzed for all the fragments, plus an additional 7 samples for fragments B, D and E, and 15 samples for fragment E (that revealed the most promising variants). All 6 exons of *TAS1R1* were

analyzed for 26 genomic DNA samples. For *LGR6* 12 DNA samples were analyzed for fragment E and 11 DNA samples were analyzed for fragment A of *DUSP12*.

All the selected patients for germline mutation analysis are index individuals that fulfill the AC or Bethesda guidelines 4 or 5 (criteria that suggests a family history), that do not present any germline mutation in the MMR genes within the family.

The sequences obtained were manually compared to the reference transcript sequences from the Ensembl database (Built GRCh38) in order to identify alterations.

All reactions were performed in a Veriti® Thermal Cycler (*Applied Biosystems*) or UNO96 thermocycler (*VWR*), according to the conditions described in the Appendix B (Tables B.3 and B.4) and C (Tables C.1 and C.3).

2.2.7.3 Mutational analysis of the *MTMR3* c.1933C>T variant in healthy individuals

The variant c.1933C>T, already identified in the L56 family (Magalhães 2016) was also observed in another sample, thus to rule out the possibility that this variant could have a higher frequency in the population in study than the one reported in the databases, a pool of 50 DNA samples from portuguese individuals 50 portuguese individuals (that had already reproduced allele frequencies described in the databases for other mutations) were analyzed for this variant. PCR amplification and sequencing analysis were performed according to the methods described in sections 2.2.2 and 2.2.3.

All reactions were performed in a Veriti® Thermal Cycler (*Applied Biosystems*), according to the conditions described in the Appendix B (Table B.5) and C (Table C.1), using the primers described in Appendix A (Table A.6).

2.2.8 Potential pathogenic role of *TPP2* gene

Previous studies (reported in 1.4.1.1.1.1) revealed evidence of an alternative non-described transcript (13a) in the *TPP2* gene, that could play a role in FCCTX susceptibility.

To evaluate the potential pathogenic role of this alternative 13a exon, a gene expression analysis by qPCR, a mutation analysis and a protein truncation test were performed.

2.2.8.1 *TPP2* gene expression analysis by qPCR

qPCR was used to quantified *TPP2* gene expression of the alternative 13a transcript as well as the wt transcript (13/14), using *GAPDH* as the reference gene. The primers were designed following the guidelines described in 2.2.2.2 with the addition of 3 rules: for each fragment to be amplified at least one of the primers must be at an exon/exon junction; the fragment must have 60 to 150bp; and each primer's Tm should be as close as possible to 60°C. All primers, except one, were designed prior to this study, and are described in Appendix A (Table A.7).

All RNA samples extracted were subjected to cDNA synthesis by RT-PCR, as described in 1.2.4. In the analysis 18 cDNA samples obtained from peripheral blood were included: 6 from a FCCTX family (L55 – that was characterized with the TSG^{+} subtype), 6 from affected individuals (with carcinoma, adenoma or mutations in the MMR genes) and 6 from healthy individuals (2 of them used as calibrator).

qPCR efficiency was accessed and all the reactions and data analysis were performed as described in 2.2.5.1. Primers were used at 5 pmol/ul and cDNA concentration was 5 ng/µl. qPCR conditions are described in Appendix B (Table B.6) and C (Table C.7).

2.2.8.2 Splicing analysis of *TPP2*

To access any alternative splicing alterations of the *TPP2* transcripts, in the regions that contained the alternative 13a and 23a exons, an analysis of the *TPP2* gene was performed using cDNA, synthetized as described in 2.2.4, from RNA extracted from FCCTX patient's peripheral blood, as described in 2.2.1.1. Subsequently, PCR amplification was done encompassing exon 2 to 29 using the Expand Long Template PCR System (*Roche*), as described in 2.2.2.3. Since this PCR produced 2 distinct bands, both bands were excised from the gel and purified as described in 2.2.3.1. Once these bands were purified they were sequenced using the forward primer in exon 13 and the forward primer in exon 23, as described in 2.2.3.2 to 2.2.3.5. However, this sequencing methodology was not always effective so another approach was used: the elution product obtained from the purified bands was diluted to 1:10 and subjected to a new PCR amplification, a nested PCR, using: 1) a pair of primers (13+14) that amplified the 13a/14 and/or the 13/14 junctions and 2) a pair of primers (23+27) that amplified the 23/24 junction (where the insertion of an exon, 23a, is described), and were subsequently sequenced. PCR amplification and sequencing analysis were performed according to the methods described in sections 2.2.2 and 2.2.3. All the primers used are described in Appendix A (Table A.8) and the PCR conditions in Appendix B (Table B.7) and C (Tables C.1 and C.4).

The sequences obtained were compared to the *TPP2* genomic and transcript sequences from the Ensembl database in order to identify possible splicing alterations.

2.2.8.3 Study of the *TPP2* alternative 13a exon using the protein truncation test technique

The alternative 13a exon found in *TPP2* gene has a stop codon within is sequence. Thus, to see if this stop codon gave rise to a truncated protein, a protein truncation test was performed.

2.2.8.3.1 Protein truncation test

The protein truncation test (PTT) or *in vitro* synthesized protein assay, is a useful approach to detect mutations that generate shortened proteins, mainly premature translation termination, using DNA or RNA as a starting point. PTT is composed of four steps: i) isolation of nucleic acid, either genomic DNA or RNA; ii) amplification of a specific region of the gene of interest; iii) *in vitro* transcription and translation of the product of the amplification reaction; and iv) detection of the translation products. Translation-terminating mutations generate translation products that are shorter and can be easily distinguished from the full-length protein product of the normal allele (figure 2.3). (Beckler 2005)

Amplified sequences for PTT can be generated across the entire protein coding sequence or to amplify specific exons. The key feature of the PTT is a specifically designed PCR primer that allows

coupled *in vitro* transcription/translation of the amplified sequence. The primer (figure 2.3-B) contains a T7 bacteriophage promoter sequence at the 5′ end that directs transcription. A 3-6bp spacer separates the promoter sequence from an optimal eukaryotic translation initiation sequence, referred to as a Kozak consensus sequence, which includes the initiation codon ATG. The bacteriophage promoter (T7), spacer and Kozak sequence are followed by sequences specific to the target. At the 3' end of the target, the primer can include a stop codon if the amplified sequence does not contain the native stop codon. (Beckler 2005)

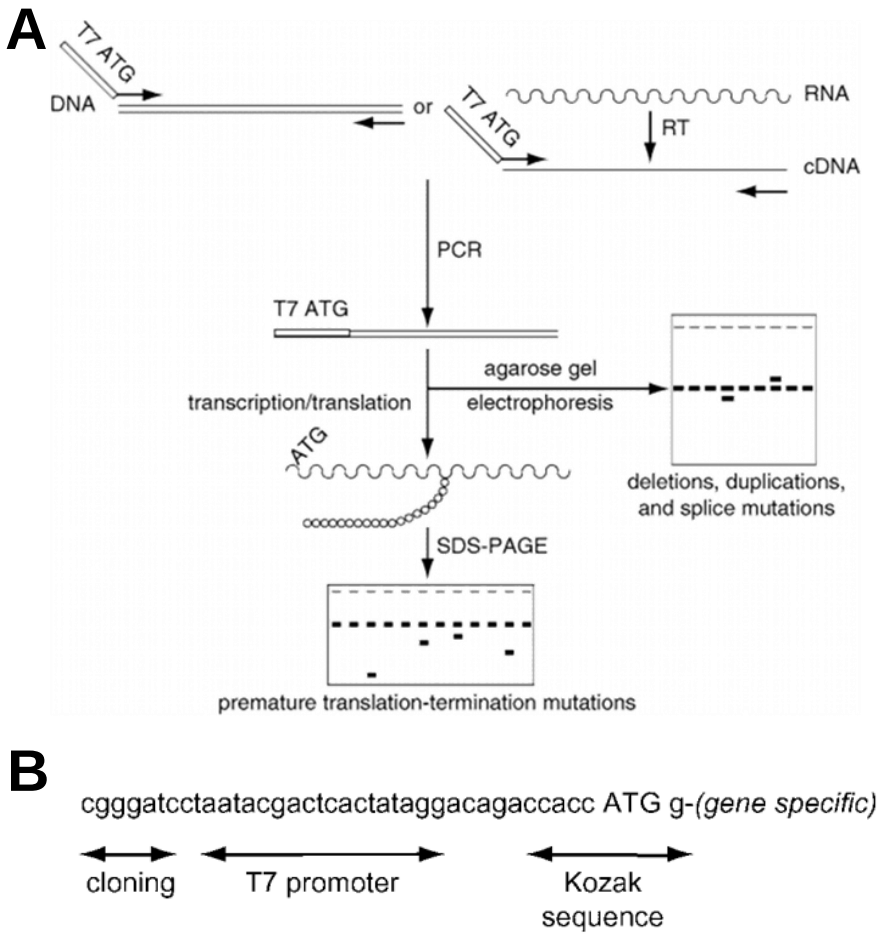


Figure 2.3 - The protein truncation test. (A) Schematic representation of the individual steps of the PTT assay. First, RNA is reverse transcribed (RT) to generate a cDNA copy. Then, the cDNA (or genomic DNA) is amplified with a specifically tailed forward primer facilitating in vitro transcription by T7-RNA polymerase. Products are analyzed on agarose gel to verify amplification, determine yield, and check size. And finally, in vitro transcription/translation is used to generate peptide fragments, analyzed on a SDS-PAGE gel, to detect translation terminating mutations. (B) Sequence of a PTT-specific forward primer facilitating in vitro transcription/translation, containing a T7 RNA polymerase promoter sequence, a spacer, and a translation initiation (Kozak) sequence. (Adapted from: Den Dunnen & Van Ommen 1999)

The detection method for PTT products must be considered when designing primers for amplification. Detection of the translation products is facilitated by incorporation of radioactive labeled amino acids (usually [^{35}S]methionine or [^{3}H]leucine). Thus, the amplified segments should contain one or more of these amino acids. The reactions are resolved on an SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis) gel and either directly dried or fluorographically enhanced and exposed to X-ray film. (Beckler 2005)

2.2.8.3.2 PCR amplification of *TPP2* for PTT

For this study, a PCR amplification containing *TPP2* exons 1 to 19 (which includes the 13a exon) was performed using the previously described T7 forward primer sequence. This primer was designed so that the gene specific sequence would start at the beginning of a normal codon of the open reading frame, to produce the protein's normal reading frame.

The reverse primer was designed as described in 2.2.2.2., while the forward primer was designed given the specific PTT guidelines. PCR amplification was performed as described in 2.2.2.3 using the Expand Long Template PCR System (*Roche*). For this amplification, 16 cDNA samples (extracted and synthesized as described in 2.2.1 and 2.2.4) were used: 6 from patients of the L55 family, 5 with an already reported 13a overexpression and 5 with a decrease in 13a expression.

PCR conditions are described in Appendix B (Table B.8) and C (Table C.4) and primers used in Appendix A (Table A.9).

2.2.8.3.3 Preparation of transcription/translation reaction

After PCR amplification of the desired fragments, transcription and translation was performed using the TNT® Quick Coupled Transcription/Translation Systems (*Promega*) and radioactive methionine L-[^{35}S]-methionine (*Perkin Elmer*) with a specific activity of $\approx 10\mu Ci/\mu l$. Due to the use of radioactive methionine, all manipulation from this step on were performed using a perspex protection, and executed by a member the colon pathology group.

Transcription and translation reactions were prepared in 0.5 mL eppendorf-like tubes consisting of 2-3.5µL of PCR product (depending on band intensity in the agarose gel) and 9.95µL of a reaction mixture containing: 9.3µL of TNT® Quick Master mix (provided by the TNT® Quick Coupled Transcription/Translation Systems (*Promega*)); 0.45µL of radioactive methionine L-[^{35}S]-methionine (*Perkin Elmer*); and 0.25µl of enhancer (also provided by the kit). A positive and a negative control were also prepared: positive control - 0.25µL of luciferase DNA (provided by the TNT® Quick Coupled Transcription/Translation Systems (*Promega*)), 2.5µL of nuclease-free water (also provided by the kit) and 9.95µL of the reactional mixture described above; negative control - 2.7µL of nuclease-free water and 9.95µL of the reactional mixture. This controls allow the evaluation of the L-[^{35}S]-methionine incorporation in proteins synthesized from the lysate. Following the reactional preparation, tubes were incubated at 30°C for 1 hour. After incubation, the reaction was stopped by adding 24µL of a stop solution (Appendix F) and samples were denatured at 100°C for 2 minutes.

2.2.8.3.4 SDS-PAGE gel electrophoresis

Once the transcription/translation reaction was complete the samples were subjected to a SDS-PAGE gel electrophoresis. The SDS-PAGE was composed of a stacking gel (with 4.5% of acrylamide) and a resolving gel (with a percentage of 14% of acrylamide). Total sample volume was applied in each

well, except for the positive control were only 5µl were applied. And 10µl of a molecular weight marker, Protein Molecular Weight Markers [Methyl-^{14}C] methylated (*Perkin Elmer*), was also applied.

The electrophoresis was performed using Tris-glycine/SDS 1x as a buffer, at a current intensity of 30mA during migration of the samples in the stacking gel and, right before the samples start entering the resolving gel, the voltage was changed to 240V. The electrophoresis was proceeded until the bromophenol blue dye (that is present in the stop solution) reached the end of the resolving gel (approximately 2 hours and 30 minutes).

SDS-PAGE gel preparation and all the solutions are described in Appendix F.

2.2.8.3.5 Treatment and SDS-PAGE gel revelation

Once the electrophoresis was finished, the gel was placed in a fixative solution for about 15 minutes. Then the gel was passed through ddH_2O in order to remove solution excesses. The gel was then place in a Amersham Amplify (*GE Healthcare*) solution, to increase the radioactive signal, under agitation for 20 minutes and after removal from the solution, was placed in a Whatmann 3MM sheet covered with a plastic sleeve. The gel was dried under vacuum at 70°C for about 2 hours and 30 minutes. Passed this time, the gel was exposed to an autoradiography film (Biomax MR film, *Kodak*) for 72 hours. The autoradiograph film was then revealed using the X-Ray film processor CURIX 60 (*AGFA healthcare*).

All the solutions used in this process are described in Appendix F.

2.2.8.3.6 Analysis of results

The resulting autoradiography was analyzed based on the PTT presupposition that the translation-terminating mutations generate translation products that are shorter and can be distinguished from the full-length protein product of the normal allele. So, the highest molecular weight band in the gel (also confirmed by the band's molecular weight – since the normal *TPP2* fragment should present a molecular height of approximately 95 KDa) was considered to be the wt protein of the amplified region, while, whenever there was a presence of a lower molecular weight band with the same or higher intensity as the wt, this was considered to be a truncating protein derived from the 13a exon insertion.

3. RESULTS

3.1 Identification of novel FCCTX susceptibility variants by WES

3.1.1 Selection of variants identified through WES in the L56 family

The bioinformatic analysis performed after WES study of the L56 family was performed with the purpose of selecting germline variants that could clarify the susceptibility for FCCTX in the family. As previously reported, this analysis was initiated but was not fully complete for analysis 1 derived from output 3 (described in 1.4.1.2). Thus, after the application of selection criteria (in 1.4.1.2.1) for the 12603 variants comprised in this analysis, 116 remained. A complete characterization, reassessment of the variants frequency and a more complete *in silico* analysis allowed restricting the 116 variants to only 2 - in *ZNF717* (Zinc finger protein 717) and *RNF213* (Ring finger protein 213) genes. Table 3.1 shows the characterization of the 2 variants, plus the addition of 2 other variants derived from the previous bioinformatic analysis but for which segregation analysis has not been performed yet (*CACNA1S* and *GIMAP1*). These 4 variants were all subjected to segregation studies within the family.

Table 3.1 - Characterization of the selected variants after application of selection criteria, review of mutation frequency and in silico analysis.

Output	Analysis	Gene	Identified variant	Variant type	Defined rs	Global allele frequency (%)		European allele frequency (%)		*In silico* prediction	
						Ensembl	ExAC	Ensembl	ExAC	SIFT/PolyPhen	Mutation Taster
3	1	*ZNF717*	c.2381G>A; p.Cys794Tyr	missense	rs139633377	-	-	-	-	Deleterious/ Probably damaging	Polymorphism (AA Score: 194)
		RNF213	c.8084C>T; p.Ala2695Val	missense	rs202096577	0.02	0.06	0.06	0.09	Deleterious/ Benign	Disease causing (AA Score: 64)
	3	*CACNA1S*	c.3607G>A; p.Asp1203Asn	missense	rs367956917	0.001	0.002	0.001	0	Tolerated/Benign	Disease Causing (AA Score: 23)
4	1	*GIMAP1*	c.908T>G; p.Val303Gly	missense	-	-	-	-	-	Deleterious low confidence/Benign	Polymorphism (AA Score: 109)

Database/software used available at: Ensembl database (http://www.ensembl.org) ExAC database (http://exac.broadinstitute.org)
SIFT/PolyPhen software (http://www.ensembl.org/Homo_sapiens/Tools/VEP) Mutation Taster software (http://www.mutationtaster.org/)

3.1.2 Segregation study of the selected variants

The study of segregation of the 4 selected variants with the disease in the family, identified by WES, was performed by assessing the presence or absence of the variant in each sample of the affected and unaffected L56 family members that had biological material available. This information, together with individual's phenotype, allowed to conclude on the segregation pattern of the variants.

The bioinformatic analysis had been divided into 5 analyzes (see 1.4.1.2), excluding shared variants in 1 or 2 affected individuals considering them as possible phenocopies, in the hope of providing a more reliable variant selection in the context of this hereditary syndrome. The 4 selected variants were included in analysis 1 and 3. Analysis 1 comprises variants shared by all affected individuals without the possibility of an existing phenocopy. From the 3 variants derived from analysis 1, none of them segregated with the disease since they were present both at affected and unaffected individuals, and 1

was not found in any of the patients (*GIMAP*), which may be due to a misdetection in the WES. (table 3.2). Thus, these 3 variants do not seem to be responsible for the family's phenotype.

Table 3.2 – Results of the segregation study of the selected variants from analysis 1. The variants are identified as the gene in which they are described. For each individual there is a brief clinical characterization, including the number of carcinomas and/or adenomas and the age at which they were diagnosed.

	Sample	Clinical features	Analysis 1		
			ZNF717	*RNF213*	*GIMAP1*
L56 family	L408 (patient 3)	1 ADC (45y); 3 TALGD (66y, 71y, 75y); 3 HP (71y, 72y, 75y)	Het	Het	Wt
	L674 (patient 1)	1 ADC (34y); 1 TALGD (36y); 3 HP (37y, 38y (2))	Het	Het	Wt
	L676	2 TALGD (43y, 44y) 1 HP (41y)	Het	Het	Wt
	L781 (patient 5)	8 TALGD (32y (2), 37y, 38y, 42y (2), 43y (2)); 4 HP (32y (2), 43y (2))	Het	Het	Wt
	L801 (patient 2)	1 ADC (57y); 2 TALGD (36y, 77y); 4 HP (37y, 38y (2), 64y)	Het	Het	Wt
	L802	Healthy	Het	Het	Wt
	L860	Healthy	Het	Het	Wt
	L1094	4 TALGD (47y, 51y, 53y (2))	Het	Het	Wt
	L1104	5 TALGD (69y, 71y); 1 TAHGD (69y)	Het	Het	Wt
	L1793	1 TALGD (55y); 3 SSALGD (58y)	Het	Het	Wt
	L1794 (patient 4)	3 TALGD (40y, 49y (2)); 1 HP (40y)	Het	Het	Wt
	L1956	Healthy	Het	Wt	Wt

Legend: y-years; ADC-adenocarcinoma; TALGD-tubular adenoma with low-grade dysplasia; TAHGD-tubular adenoma with high-grade dysplasia; HP- hyperplastic polyp; SSALGD-sessile serrated adenoma with low-grade dysplasia; Wt (wild type)- the individual does not present the variant; Het- the individual presents the variant at heterozygosity. Individuals marked in colors were integrated in WES analysis 1. The age at diagnosis (in years, y) and respective phenotypic characterization is described for the affected individuals.

Analysis 3 includes the variants shared by patients 1, 3, 4 and 5, excluding patient 2 from the shared variant analysis, considering it as a possible phenocopy. The segregation study of the *CACNA1S* variant from analysis 3, revealed a segregation of the p.Asp1203Asn variant with the disease, taking the segregation model into account. When considering patient 2 (L801) as a phenocopy, she and her progeny (both of her sons, L676 and L802) should not present the inherited variant, which was observed (table 3.3 and figure 3.1).

Table 3.3 - Results of the segregation study of the selected variants from analysis 3. The variants are identified as the gene in which they are described. For each individual there is a brief clinical characterization, including the number of carcinomas and/or adenomas and the age at which they were diagnosed.

	Sample	Clinical features	Analysis 3 *CACNA1S*
L56 family	L408 (patient 3)	1 ADC (45y); 3 TALGD (66y, 71y, 75y); 3 HP (71y, 72y, 75y)	Het
	L674 (patient 1)	1 ADC (34y); 1 TALGD (36y); 3 HP (37y, 38y (2))	Het
	L676	2 TALGD (43y, 44y) 1 HP (41y)	Wt
	L781 (patient 5)	8 TALGD (32y (2), 37y, 38y, 42y (2), 43y (2)); 4 HP (32y (2), 43y (2))	Het
	L801 (patient 2)	1 ADC (57y); 2 TALGD (36y, 77y); 4 HP (37y, 38y (2), 64y)	Wt
	L802	Healthy	Wt
	L860	Healthy	Wt
	L1094	4 TALGD (47y, 51y, 53y (2))	Het
	L1104	5 TALGD (69y, 71y); 1 TAHGD (69y)	Het
	L1793	1 TALGD (55y); 3 SSALGD (58y)	Het
	L1794 (patient 4)	3 TALGD (40y, 49y (2)); 1 HP (40y)	Het
	L1956	Healthy	Wt

Legend: y-years; ADC-adenocarcinoma; TALGD-tubular adenoma with low-grade dysplasia; TAHGD-tubular adenoma with high-grade dysplasia; HP- hyperplastic polyp; SSALGD-sessile serrated adenoma with low-grade dysplasia; Wt (wild type)- the individual does not present the variant; Het- the individual presents the variant at heterozygosity. Individuals marked in colors were integrated in WES analysis 3. The age at diagnosis (in years, y) and respective phenotypic characterization is described for the affected individuals.

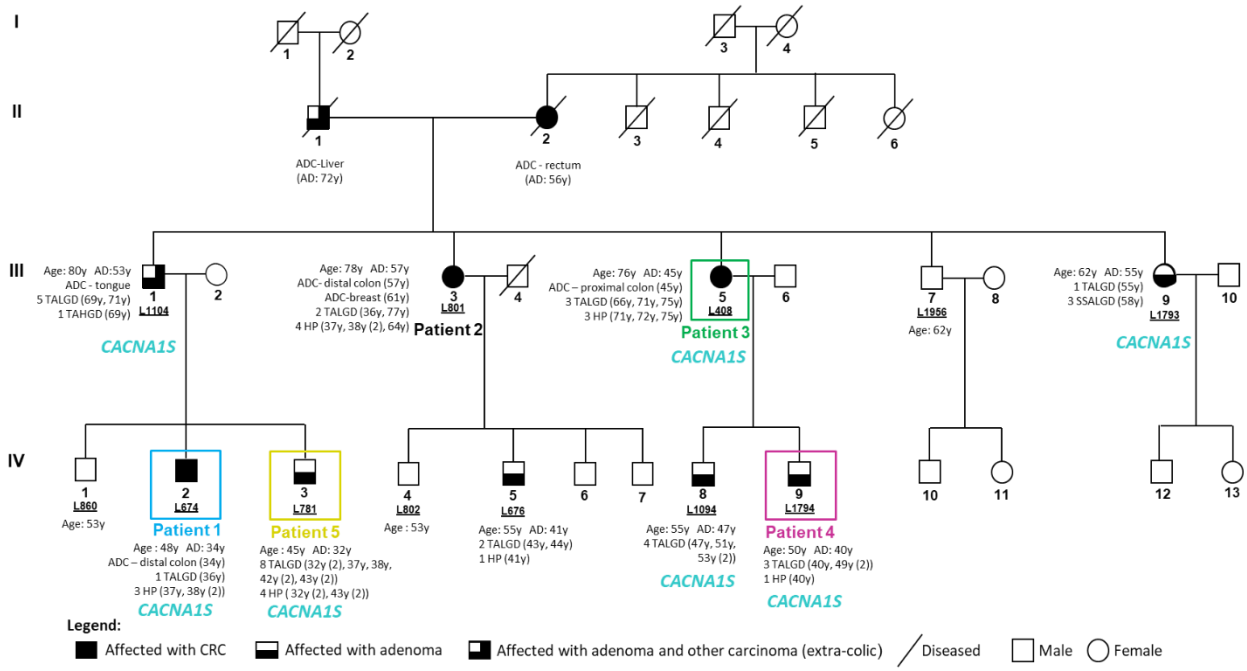


Figure 3.1 – L56 family pedigree with the segregation results for variant the CACNA1S variant - p.Asp1203Asn. Individuals that presented this variant are highlighted with the word CACNA1S in blue. Individuals highlighted in colors were selected for bioinformatic analysis 3. Only individuals with a laboratory identification number (LXXX or LXXXX) had biological material available. The age at diagnosis (AD, in years, y) and respective phenotypic characterization are described for the affected individuals. AD: age at diagnosis (carcinoma or adenoma); y: years; ADC: adenocarcinoma; TALGD: tubular adenoma with low-grade dysplasia; TAHGD: tubular adenoma with high-grade dysplasia; HP: hyperplastic polyp; SSALGD: sessile serrated adenoma with low-grade dysplasia.

Table 3.4 sums up all the selected variants detected by WES that segregated with the disease in L56 family, according to each analysis segregation pattern. The variant found in *CACNA1S* was added

to the previously selected variants in *DUSP12*, *MTMR3*, *NPR2*, *TAS1R1*, *SMG7* and *LGR6* genes. (Magalhães 2016)

Table 3.4 - Characterization of the selected variants that revealed a segregation pattern with the disease according to each analysis.

Output	Analysis	Gene	Variant's chromosomal position (GRCh38)	Identified variant	Variant type	Defined rs (NCBI)	Global allele frequency (%)		European allele frequency (%)		*In silico* prediction		
							Ensembl	ExAC	Ensembl	ExAC	SIFT/ PolyPhen	Mutation taster	Provean
4	1	***MTMR3***	22:30019592	c.1933C>T; p.Arg645Trp	Missense	rs138823197	0.02	0.02	0.03	0.03	Deleterious/ Probably damaging	Disease Causing (AA Score:101)	Deleterious (Score: -4.34)
		TAS1R1	1:6577058	c.1582C>T; p.Leu528Phe	Missense	rs560288110	-	-	-	-	Deleterious/ Probably damaging	Disease Causing (AA Score:22)	Deleterious (Score: -3.13)
3	3	***CACNA1S***	1:201058410	c.3607G>A; p.Asp1203Asn	Missense	rs367956917	0.001	0.002	0.001	0	Tolerated/ Benign	Disease Causing (AA Score: 23)	Neutral (Score: 0.16)
4	4	***NPR2***	9:35807330	c.2644G>A; p.Val882Ile	Missense	rs55700371	0.01	-	0.01	0.01	Deleterious/ Possibly damaging	Disease Causing (AA Score:29)	Neutral (Score: -0.90)
3		***LGR6***	1:202318327	c.2024G>A; p.Arg675Gln	Missense	rs767586087	0.001	-	0.002	-	Tolerated/ Benign	Disease Causing (AA Score:43)	Neutral (Score: -0.86)
4	5	***SMG7***	1:183544963	1970C>T; p.Pro657His	Missense	rs34426362	0.2	0.4	0.7	0.6	Deleterious low confidence/ Probably damaging	Disease Causing (AA Score:77)	Deleterious (Score: -2,66)
		DUSP12	1:161749807	c.6_7insTTG; p.Leu2_Leu3ins	In frame insertion	-	-	-	-	-	-	Regulatory feature	Neutral (Score: -0.47)

Database/software used available at: NCBI: https://www.ncbi.nlm.nih.gov ExAC database: http://exac.broadinstitute.org Ensembl: http://www.ensembl.org
SIFT/PolyPhen: http://www.ensembl.org/Homo_sapiens/Tools/VEP Provean: http://Provean.jcvi.org Mutation Taster: http://www.mutationtaster.org

The segregation analysis of the *MTMR3* and *TAS1R1* variants in the CAs1555 patient, confirmed the presence of these 2 variants. Figure 3.2 exhibits all the selected variants detected by WES found to segregate with the disease in L56 family, according to each analysis segregation pattern, with the addition of the variants found for patient CAs1555 (II.2).

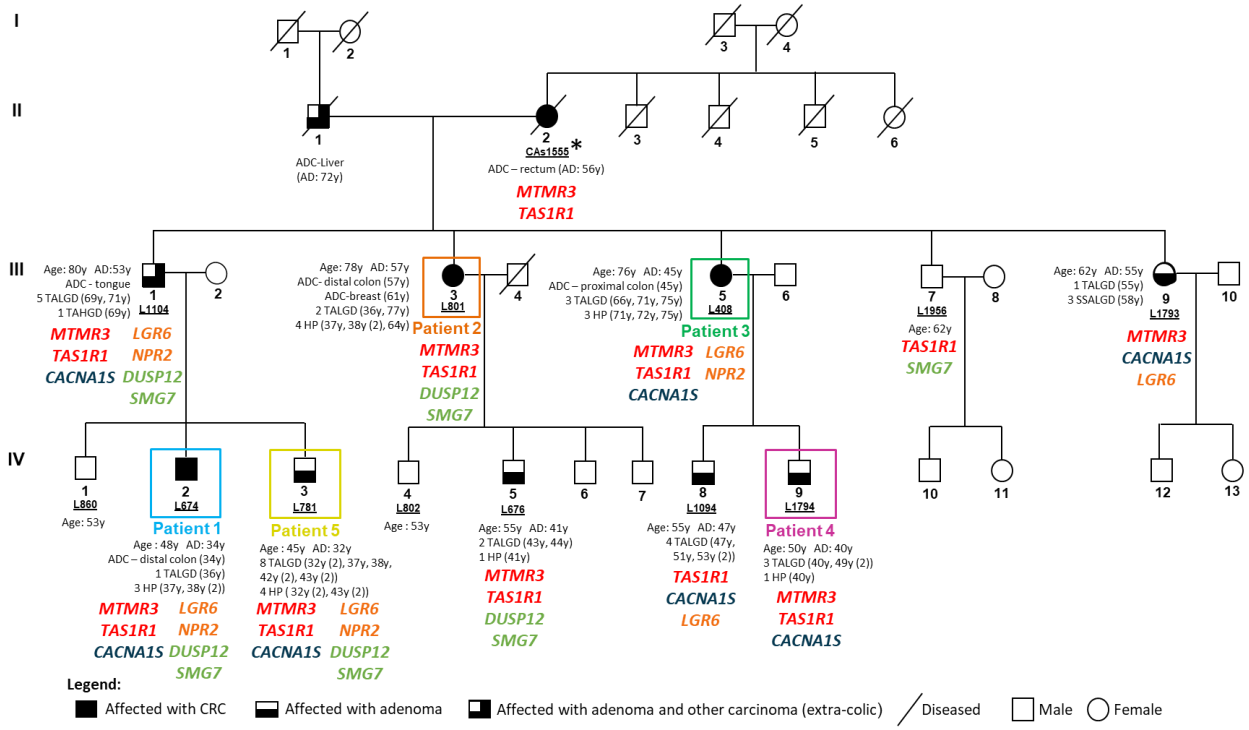


Figure 3.2 - L56 family pedigree with the segregation results for all the selected variants that segregated with the disease. (continues next page)

*(figure 3.2 – continuation) Patients highlighted in colors were selected for WES bioinformatic analysis according to each analysis sub-type. Written in red are the variants derived from analysis 1, in blue from analysis 3, in orange from analysis 4 and in green from analysis 5. Only individuals with a laboratory identification number (LXXX, LXXXX or CAsXXXX) had biological material available. The age at diagnosis (AD, in years, y) and respective phenotypic characterization are described for the affected individuals. * - segregation analysis only performed for MTMR3 and TAS1R1; AD: age at diagnosis (carcinoma or adenoma); y: years; ADC: adenocarcinoma; TALGD: tubular adenoma with low-grade dysplasia; TAHGD: tubular adenoma with high-grade dysplasia; HP: hyperplastic polyp; SSALGD: sessile serrated adenoma with low-grade dysplasia.*

3.1.3 Copy number variation analysis in the L56 family based on WES data

The copy number variation analysis in the L56 family came as an additional approach to exclude the possibility of a given CNV being the cause of the disease in the FCCTX family. The bioinformatic analysis developed for CNV analysis (performed by Bioin2Bio) for the 5 patients that had been previously analyzed by WES, revealed 296 amplicons with possible deletion when compared to a control group (a set of samples from a thyroid cancer family). This number was then reduced to 210 after application of selective criteria that were based on the number of reads for each sample against the minimum number of reads of the control group (section 2.2.6.3). And, after an assessment of the genomic regions included in the 210 selected amplicons, based on a comparison of these two values with the number of reads exhibited by the 5th sample included in the exome sequencing, the final number of amplicons was refined to 22 (table 3.5). This 5th sample (patient 5) was sequenced in a different run from the other 4 samples and, for a given region, despite a lower number of reads for the other 4 samples in comparison with the minimum value observed for thyroid samples, a higher number of reads for patient 5 excluded that region as a possible CNV.

Table 3.5 – Results from the selection of amplicons derived from CNV analysis.

Identified amplicon (GRCh38)		Amplicon length (bp)	Gene
Chromosome	Chromosomal region		
1	10413107 to 10413331	224	*PGD*
1	11677344 to 11677689	345	*MAD2L2*
2	219307734 to 219307949	215	*PTPRN*
2	219333080 to 219333256	176	*RESP18*
2	219491758 to 219491955	197	*SPEG*
2	219611601 to 219612062	461	*STK11IP*
3	52706615 to 52706901	286	*SPCS1*
4	682678 to 682878	200	*MFSD7*
4	1238016 to 1238317	301	*CTBP1*
4	184850967 to 184851170	203	*MIR3945HG*
11	47486700 to 47486906	206	*CELF1*
11	72605126 to 72605296	170	*PDE2A*
12	56021464 to 56021661	197	*IKZF4*
13	32732035 to 32732226	191	*PDS5B*
13	97979806 to 97980021	215	*IPO5*
14	105382795 to 105383001	206	*PACS2*
15	82665416 to 82665536	120	*AP3B2*
16	89187437 to 89187616	179	*CDH15*
16	89510422 to 89510604	182	*SPG7*
17	29618796 to 29619239	443	*CORO6*
19	49624916 to 49625252	336	*PRR12*
20	38641723 to 38641911	188	*ARHGAP40*

3.2 Germline mutation analysis of candidate genes for FCCTX susceptibility

The identification of the variants that segregated with the disease in the L56 family can give new insights on this family susceptibility for FCCTX. In order to clarify the potential contribution of the newly identified candidate genes to this syndrome, a set of 34 FCCTX and FCCTX-like families was selected for a germline mutation analysis.

3.2.1 *MTMR3*

The mutational analysis of the MTMR3 gene in 34 index patients from the selected FCCTX and FCCTX-like families revealed 3 missense variants, in different patients that were not previously reported or displayed a low allele frequency (figure 3.3).

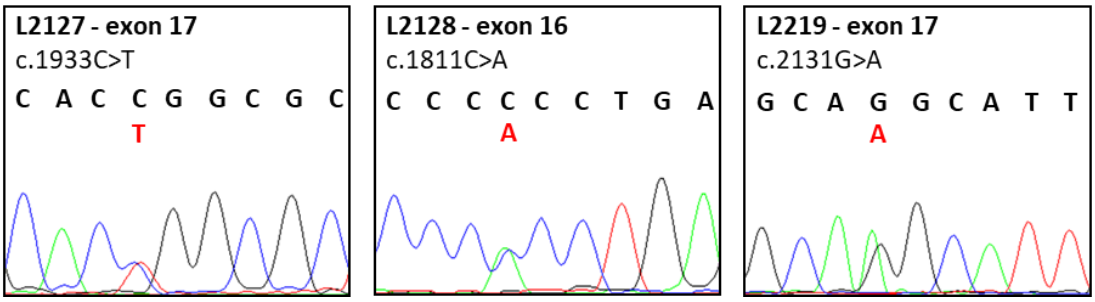


Figure 3.3 - Partial electropherograms representing the 3 variants identified in the MTMR3 gene. In the upper left corner is indicated the sample, the exon where the variant was detected and its characterization. The DNA sequences (forward direction), in black, correspond to the normal gene sequence while the red bases indicate the alterations found. All mutations were detected in heterozygosity.

These 3 variants were all found either at exon 16 or 17, which may indicate a mutation hotspot region in this gene, since for rest of the gene, samples did not show relevant mutations. Table 3.6 resumes the variants found in *MTMR3* and its pathogenic *in silico* prevision.

The variant found in sample L2127, c.1933C>T (p.Arg645Trp), is the same variant found in the L56 family, reinforcing the possibility that this specific variant could be associated with the FCCTX phenotype, since all the predictive software characterizes this variant as pathogenic.

The variants identified in samples L2128 and L2219, have not been described to date in the available mutation databases. The c.1811C>A (Pro604His) variant, found in L2128, is classified as potentially pathogenic in 2/3 of the predictive mutation software (SIFT/PolyPhen and Mutation Taster). In Provean this mutation is referred as neutral, however the score given (-2.43) is really close to the software's default threshold for defining a mutation as deleterious (a mutation is defined as deleterious with a score ≤ -2.50). The c.2131G>A (p.Gly711Ser) variant, found in L2219, is rated as non-pathogenic by all the software used.

Since the variant c.1933C>T (p.Arg645Trp) was found not only in the L56 family but also in sample L2127, although it is described with a low frequency in the European population (19/65366, 0.03%), a control population of 50 Portuguese individuals was analyzed for this variant, to evaluate the possibility that this variant could have a higher frequency in the Portuguese population than the reported in the databases. In this analysis, no individual presented the variant.

Table 3.6 - Characterization of the most relevant variants obtained in the mutational analysis of the MTMR3 gene, for the 34 index patients of FCCTX and FCCTX-like families without an identified mutation in the MMR genes.

Sample	Exon	Identified variant	Variant type	Defined rs (NCBI)	Global allele frequency (%)		European allele frequency (%)		*In silico* prediction		
					Ensembl	ExAC	Ensembl	ExAC	SIFT/ PolyPhen	Mutation taster	Provean
L2127	17	c.1933C>T; p.Arg645Trp	Missense	rs138823197	0.02	0.02	0.03	0.03	Deleterious /Probably damaging	Disease Causing (AA Score:101)	Deleterious (Score: -4.34)
L2128	16	c.1811C>A; Pro604His	Missense	-	-	-	-	-	Deleterious /Possibly damaging	Disease Causing (AA Score:77)	Neutral (Score: -2.43)
L2219	17	c.2131G>A; p.Gly711Ser	Missense	-	-	-	-	-	Tolerated/ Benign	Polymorphism (AA Score:56)	Neutral (Score: -0.13)

Database/software used available at: NCBI: https://www.ncbi.nlm.nih.gov ExAC database: http://exac.broadinstitute.org Ensembl: http://www.ensembl.org
SIFT/PolyPhen: http://www.ensembl.org/Homo_sapiens/Tools/VEP Provean: http://Provean.jcvi.org Mutation Taster: http://www.mutationtaster.org

Besides these 3 variants, almost all the analyzed samples for fragments B and D (see Appendix A: Table A.5) presented alternative splicing, represented in figure 3.4, between exons 9 to 11 and 12 to 15, non-described to date in the available databases. These alterations were always accompanied by the normal transcript. Only samples L1667 and E436, in all the 19 analyzed samples for these fragments, did not present such alterations. Sample L1908 presented the splicing alterations, although with much less peak intensity (in the Sanger sequencing electropherograms), when comparing to the normal transcript, suggesting much lower contribution of the splicing isoform. Samples L1368 and L2208 presented the same proportion between the splicing alterations and the normal transcript and sample L1310 had a superior amount of the spliced isoforms. The remaining samples showed a slight lower contribution of the splicing isoforms in comparison with the normal transcript and all the samples with these alterations presented all the alternative spliced isoforms. No mutations were found that could explain these splicing alterations.

The results of the complete mutational analysis for MTMR3 are summarized in Appendix G.

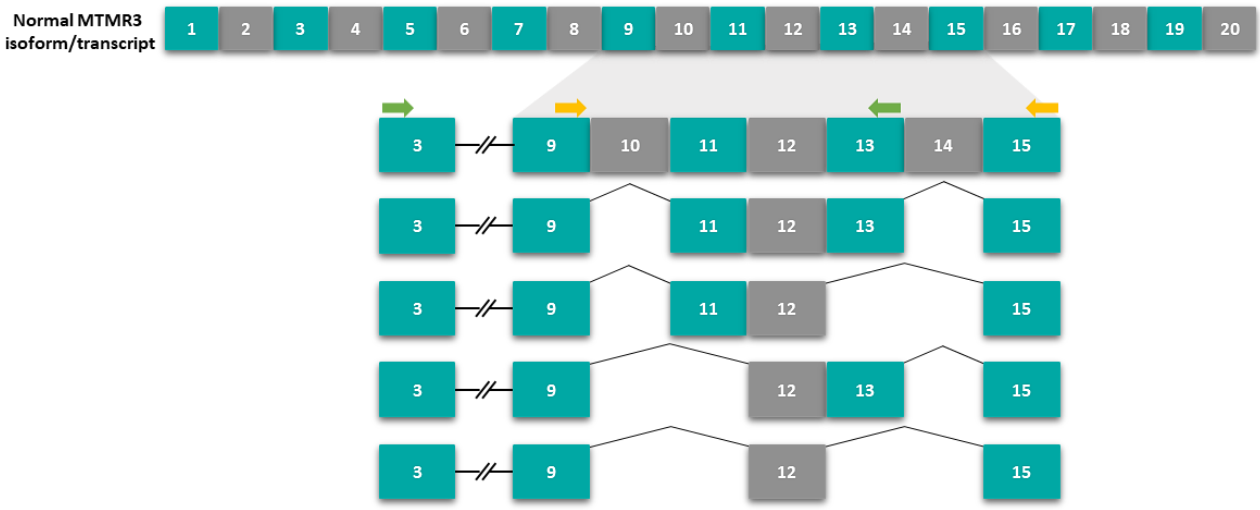


Figure 3.4 – Representative scheme of the alternative spliced isoforms found in fragments B and D of MTMR3. The green arrows represent external primers for fragment B and the yellows for fragment D. The squared boxes represent exons and is respective number. Exons size are not to scale.

3.2.2 *TAS1R1*

For *TAS1R1* gene, the mutational analysis in a set of 26 individuals, revealed 5 relevant variants in different patients, all described in the Ensembl and ExAC databases (table 3.7).

All variants were identified in heterozygosity except for the c.1261-90C>T variant, in intron 3 of sample L1291, that was found in homozygosity. This variant presents an allele frequency of 1.7% which may indicate a polymorphism, however in is homozygotic form, the genotype frequency (T/T) is 0.04%. The Mutation Taster software does not predict a pathogenic consequence and the HSF software (Human Splicing Finder, HSF: http://www.umd.be/HSF3/index.html) does not anticipate an impact on splicing, which makes this variant likely non-pathogenic.

Table 3.7 - Characterization of the most relevant variants obtained in the mutational analysis of the TAS1R1 gene, for the 26 index patients of FCCTX and FCCTX-like families without an identified mutation in the MMR genes.

Sample	Exon	Identified variant	Variant type	Defined rs (NCBI)	Global allele frequency (%)		European allele frequency (%)		*In silico* prediction		
					Ensembl	ExAC	Ensembl	ExAC	SIFT/ PolyPhen	Mutation Taster	Provean
L108	3	c.1067C>G; p.Ser356Cys	missense	rs41307749	0.2	0.2	0.2	0.2	Deleterious/ Probably damaging	Polymorphism (AA Score:112)	Neutral (Score: -1.24)
L701	3	c.1043C>T; p.Ala348Val	missense	rs147969126	0.1	0.07	0	0	Tolerated/ Benign	Polymorphism (AA Score:64)	Neutral (Score: -0.48)
L1121	2	c.329C>T; p.Ala110Val	missense	rs41278020	1.2	1.18	2	1.59	Tolerated/ Probably damaging	Polymorphism (AA Score:64)	Neutral (Score: -2.13)
L1291	Intron 3	c.1261-90C>T *	intronic	rs113004821	1.7	-	1.7	-	-	Polymorphism	-
L1473	3	c.953T>C; p.Ile318Thr	missense	rs754592015	0.001	-	0.001	-	Deleterious/ Probably damaging	Polymorphism (AA Score:89)	Deleterious (Score: -3.05)

* Variant found in homozygosity
Database/software used available at: NCBI: https://www.ncbi.nlm.nih.gov ExAC database: http://exac.broadinstitute.org Ensembl: http://www.ensembl.org
SIFT/PolyPhen: http://www.ensembl.org/Homo_sapiens/Tools/VEP Provean: http://Provean.jcvi.org Mutation Taster software: http://www.mutationtaster.org

The c.1067C>G (p.Ser356Cys) variant, in L108, presents a low frequency and is classified as deleterious/probably damaging by the SIFT/PolyPhen software, although the Mutation Taster and Provean software do not classify this variant as pathogenic, the Mutation Taster AA score is 112. This score evaluates the degree of difference between the original and the new amino acid, by taking into account the physicochemical characteristics of the amino acids, ranging from 0 to 215. So, a higher score like 112, can indicate a significant difference in the physicochemical characteristics of the new amino acid when comparing to the normal one (from a serine to a cysteine). As for potential splicing alterations, the Mutation Taster software does not predict any splicing alterations, however the HSF software predicts the creation of an exonic splicing silencer site, which can lead to splicing alterations.

The variant c.1043C>T (p.Ala348Val), in L701, is not classified as pathogenic by the predictive software, however both the Mutation Taster and HSF software predict possible splicing alterations with a high score variation between wt and mutant (splicing donor increased, wt:0.71/mut:0.91 – Mutation Taster) (activation of an exonic cryptic donor site, wt:46.37/mut:73.2 - HSF). The higher the score variation, more likely it is for the variant to generate a splicing alteration.

Although the c.329C>T (p.Ala110Val) variant, in L1121, is considered as probably damaging by PolyPhen its frequency is higher than 1%, which may indicate a polymorphism.

The c.953T>C (p.Ile318Thr) variant, found in L1473, exhibits the highest pathogenicity predictions, having a deleterious prediction by SIFT and Provean software, and probably damaging by PolyPhen. The Mutation Taster suggests this variant is a polymorphism, but presents an elevated AA score (89) and a possible splicing donor increased (wt:0.59/mut:0.81). The HSF does not predict an impact on splicing for this variant.

The complete mutational analysis for this gene is summarized in Appendix H.

3.2.3 *DUSP12* and *LGR6*

For *DUSP12* and *LGR6*, due to the available time, the mutational analysis was only performed for one of the fragments in 11 index patients of FCCTX and FCCTX-like families. For *DUSP12*, only fragment A (see Appendix A: Table A.5) was analyzed and no alteration was found. For *LGR6*, the only analyzed fragment was E (see Appendix A: Table A.5), which revealed one relevant variant in sample CAs1964: c.213-11224C>T, already reported with the rs: rs79821910 (NCBI). However, this variant displays a global allele frequency of 2% and 3% in the European population, thus being considered a polymorphism. However, these 2 genes should not be discarded as possible FCCTX contributors, since this analysis was only performed in a short portion of the genes and for a small number of samples, thus a more extensive analysis is required.

The complete mutational analysis of both fragments of these 2 genes is summarized in Appendix I.

3.3 *TPP2* role in FCCTX susceptibility

3.3.1 Gene expression analysis

Previous studies have reported differences in gene expression between an alternative *TPP2* transcript (13a) and its wt transcript (here referred as 13/14) in samples derived from patients with different clinical spectra. (Pereira 2014) Thus, in the present study, this expression analysis was performed in 3 distinct groups to observe possible correlations. The first group was composed from 6 affected patients (L950, L905, L926, L447, L1544 and L295) of a FCCTX family, characterized as TSG^{+} (L55); the second group included 6 affected patients (L1591, L1592, L1732, CAs2000, L1642 and L5), with the following clinical and molecular features: L1591 and L1592 are affected with adenomas, but have no evidence of a mutation in the MMR genes (associated with LS); L1732 and L1642 have an identified mutation in the MMR genes, but so far have not presented adenomas; and CAs2000 and L5 both present carcinoma but CAs2000 has a MMR gene mutation. The third group presented 4 healthy individuals (L1568, L1567, L1604 and L1829).

The *TPP2* gene expression analysis (figure 3.5) revealed that for family L55 all individuals present a similar expression between the alternative and the wt transcript except for L447, that presents an overexpression of the wt transcript with almost a total reduction of the alternative, and L1544 that presents an inverse relationship, having the alternative transcript overexpressed over the wt.

In the affected group 4/6 patients presented a 13a overexpression and in 3/4 (L1591, L1592 and CAs2000) there was an almost total reduction of the wt transcript, while for the 2 remaining samples seems to be a slight increase of wt expression. All the patients that developed adenomas, with the exception of L447, seem to have similar expression of the wt and 13a expression or even an increased 13a expression, but never a significative overexpression of the wt transcript.

In all healthy control samples, the wt expression was at least 1-fold higher than the 13a expression, reinforcing the possibility that the *TPP2* gene, more specifically the balance or imbalance between the 13a alternative transcript and the wt could play a role in colorectal tumorigenesis.

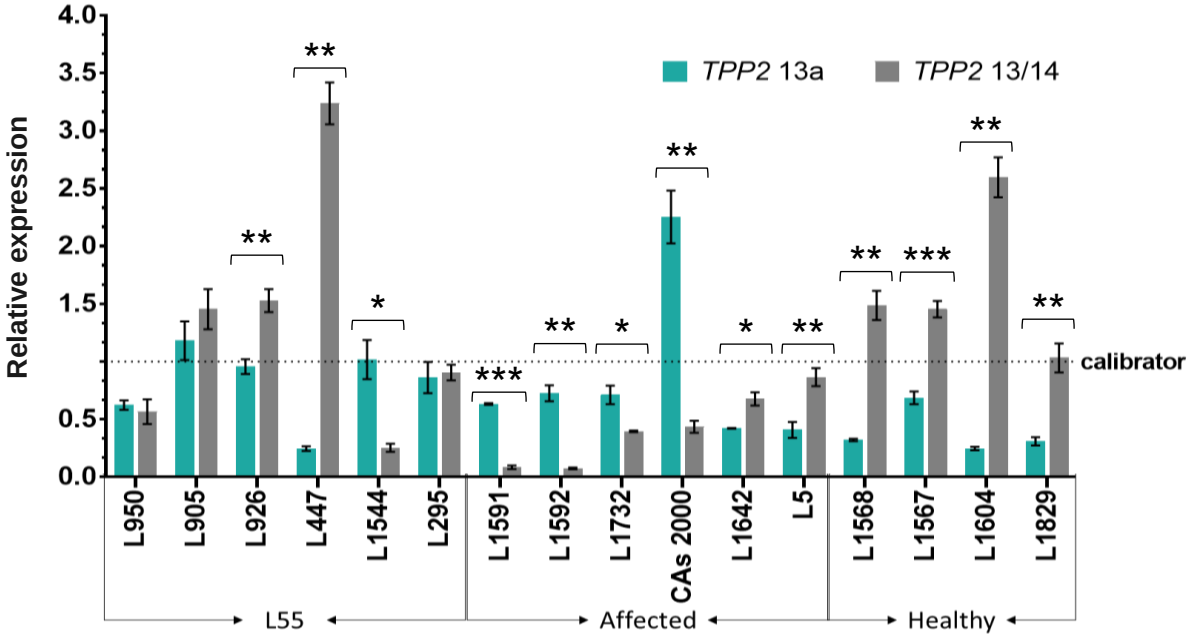


*Figure 3.5 – Relative gene expression analysis by qPCR of the wt (13/14) and an alternative transcript (13a) of the TPP2 gene. Error bars represent SD. Calibrator – mean relative expression of two normal samples used as calibrator. N=2; *p<0,05; **p<0,01; ***p<0,001.*

3.3.2 Splicing analysis

The *TPP2* gene has 2 different transcripts/isoforms reported in the Ensembl database that differ from the insertion of an exon (here referred as 23a) between exons 23 and 24. As already stated, the inclusion of an alternative exon (13a) flanked by the 13 and 14 exons, was reported in previous studies by the colon pathology group. Figure 3.6 schematically represents the reported *TPP2* transcripts, which will be referred as wt, and the novel 13a transcript.

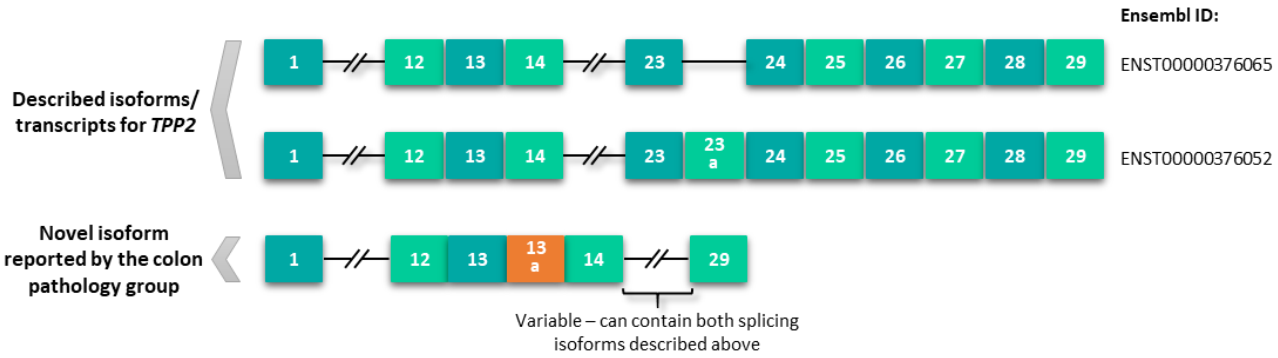


Figure 3.6 – Schematic representation of the described isoforms for TPP2 (ENST00000376065 and ENST00000376052) and a non-described isoform. The squared boxes represent exons and is respective number. Exons size are not to scale.

In order to observe the presence or absence of these isoforms and if there was any exclusivity between the occurrence of the 13a isoform and one of the already reported, blood samples from 6 FCCTX patients were subjected to a splicing examination on these 2 regions (the one that contains the 13a exon and the one that introduces the 23a exon). This analysis started off with an almost full open reading frame amplification (amplification of exons 2 to 29), that for most cases presented the evidence of 2 bands. Then, to perceive the differences between these 2 fragments, nested PCRs were conducted for the 2 regions of interest. Nested PCRs were performed due to insufficient volume of the eluted product after band purification for effective sequencing of the regions of interest.

Sample L950 splicing analysis (figure 3.7) revealed, for the 13 to 14 region, an expression of both the wt and the 13a isoform. For the region covering exons 23 to 27 the highest molecular weight fragment (from PCR 2-29), showed the presence of the isoform that skips the 23a exon, while the lowest molecular weight fragment presented both. Besides these 2 isoforms, another alternative splicing event seems to take form in this sample, which results from the skipping of exon 24 and/or 23a and 24. However, these finding do not explain the presence of two bands with different molecular weights, when almost all the open reading frame of the gene is amplified.

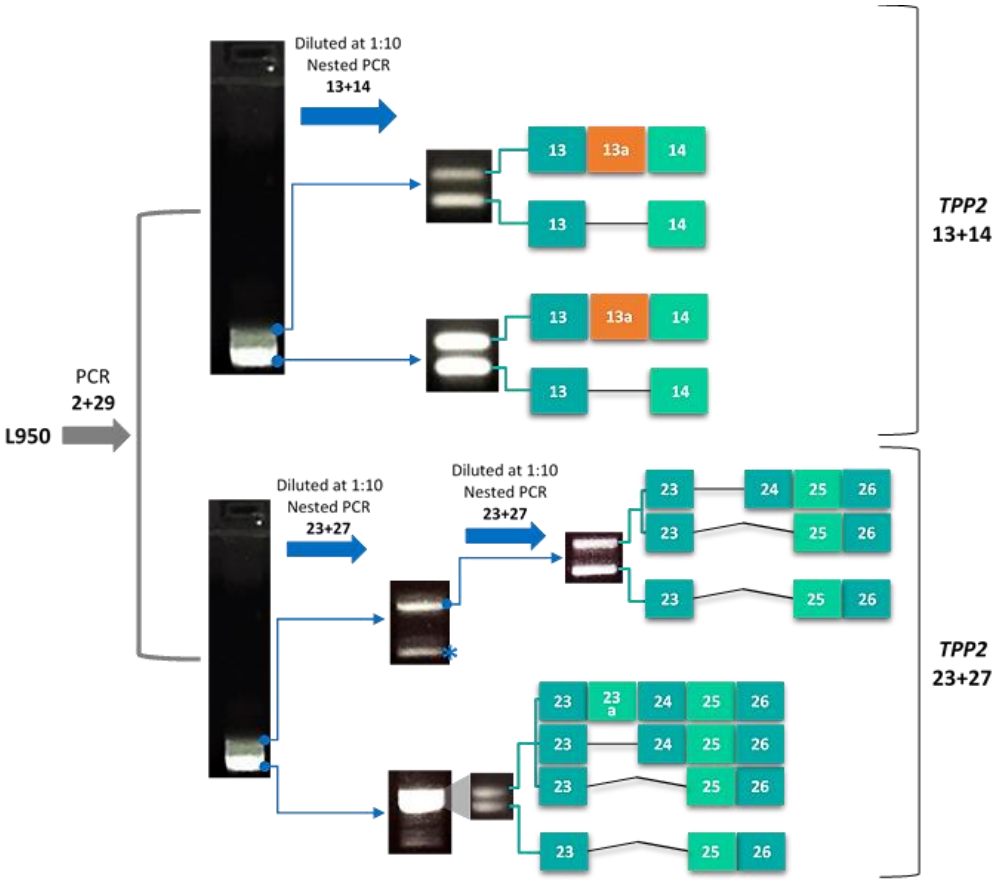


Figure 3.7 – Representative scheme of the steps involved in TPP2 splicing analysis of sample L950 and respective results. The resulting bands (from an 1.2% agarose gel) were excised, purified and subjected to nested PCRs after a dilution of 1:10. Nested PCR 13+14 represents the amplification of exons 13 to 14, and nested PCR 23+27 the exons 23 to 27 (amplification products of the nested PCRs were ran in a 2% agarose gel). Sequencing results are presented as squared boxes that indicate exons and its respective number. Exons size are not to scale. ***** *- sequencing analysis could not be performed due to technical issues.*

The splicing evaluation in sample L1590 (figure 3.8) disclose the presence of only the wt transcript for the 13 to 14 region and the region encompassing exons 23 to 27 showed the wt transcript without the presence of the 23a exon and an alternative splicing derived from the skipping of exon 25 and half of the exon 26. Although the lowest molecular weight fragment, derived from amplification of exons 2 to 29, could not be sequenced for the 23 to 27 region, it appears that the presence of only the wt transcript, deprived of the 23a exon, can be correlated to the absence of the alternative 13a exon, in this sample.

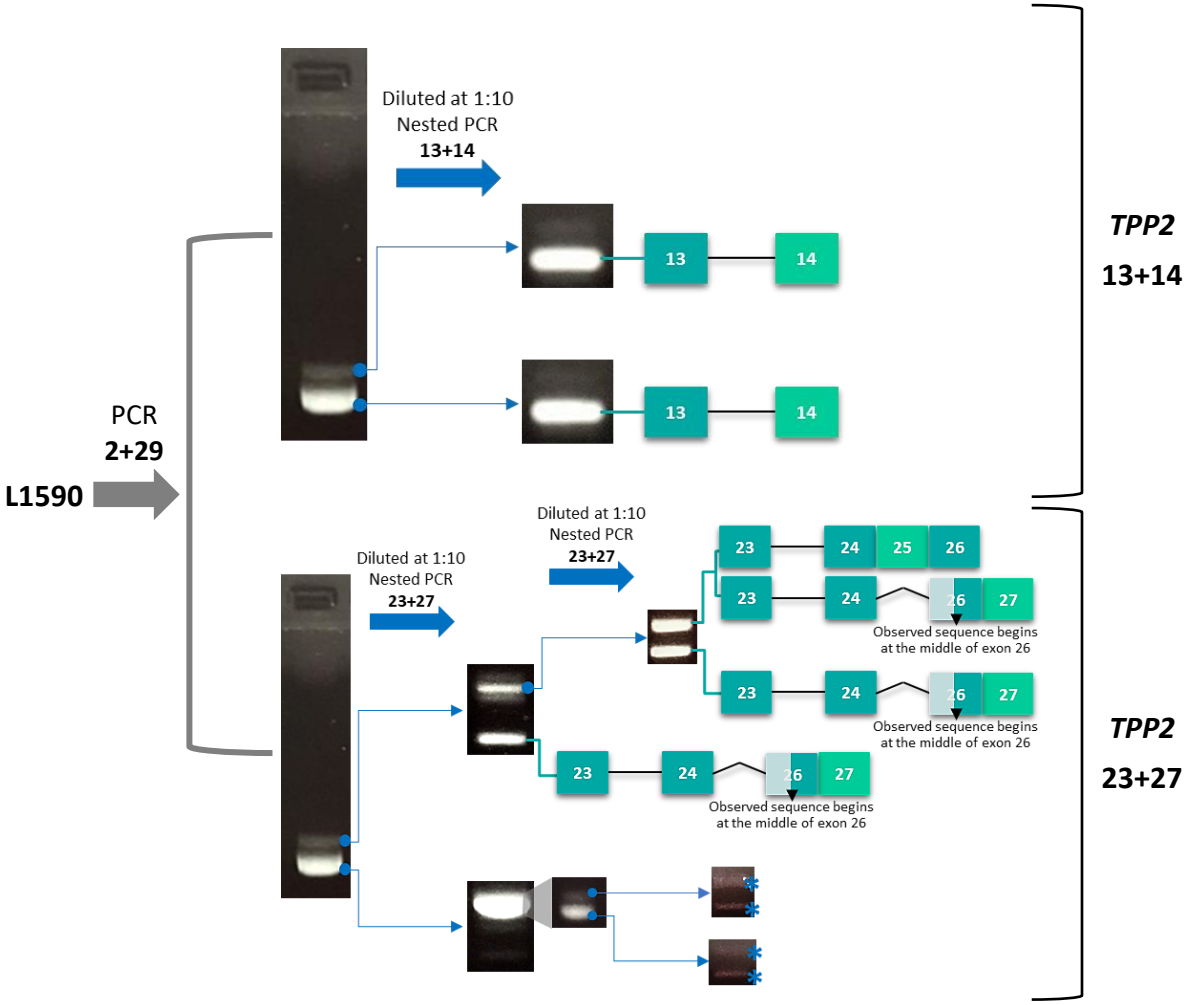


Figure 3.8 - Representative scheme of the steps involved in TPP2 splicing analysis of sample L1590 and respective results. The resulting bands (from an 1.2% agarose gel) were excised, purified and subjected to nested PCRs after a dilution of 1:10. Nested PCR 13+14 represents the amplification of exons 13 to 14, and nested PCR 23+27 the exons 23 to 27 (amplification products of the nested PCRs were ran in a 2% agarose gel). Sequencing results are presented as squared boxes that indicate exons and its respective number. Exons size are not to scale. ***** *- sequencing analysis could not be performed due to technical issues.*

Sample L926 exhibited only one band after amplification of exons 2 to 29 that presents all the wt described transcripts and has no alternative splicing events (figure 3.9).

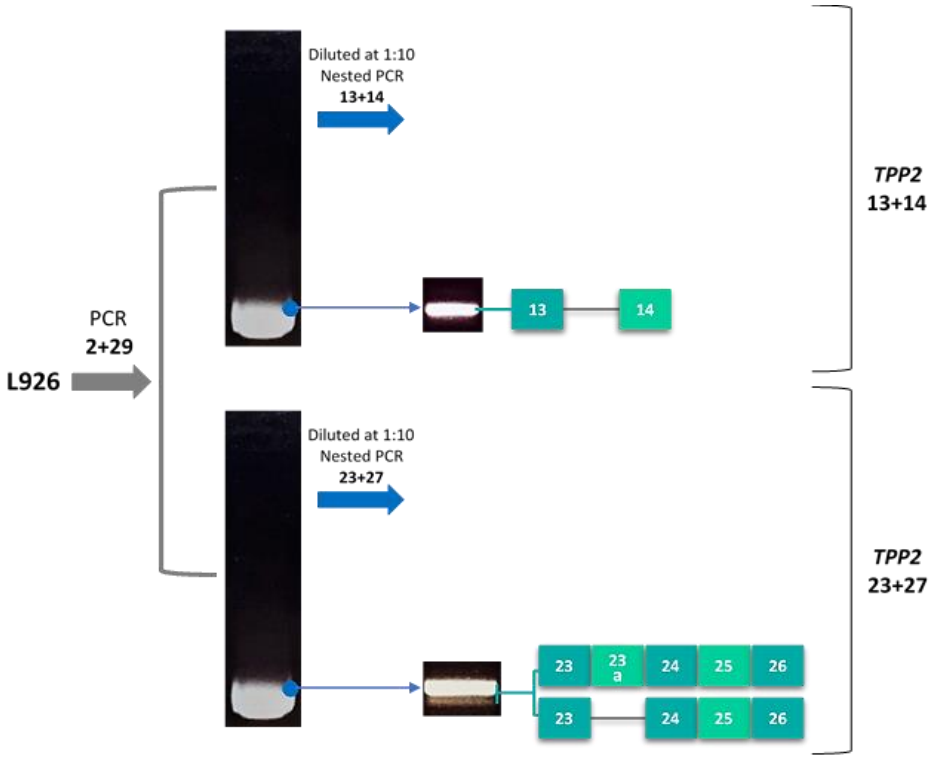


Figure 3.9 - Representative scheme of the steps involved in TPP2 splicing analysis of sample L926 and respective results. The resulting bands (from an 1.2% agarose gel) were excised, purified and subjected to nested PCRs after a dilution of 1:10. Nested PCR 13+14 represents the amplification of exons 13 to 14, and nested PCR 23+27 the exons 23 to 27 (amplification products of the nested PCRs were ran in a 2% agarose gel). Sequencing results are presented as squared boxes that indicate exons and its respective number. Exons size are not to scale.

Sample L447 (figure 3.10) displays all the wt described isoforms, plus the addition of the alternative 13a exon.

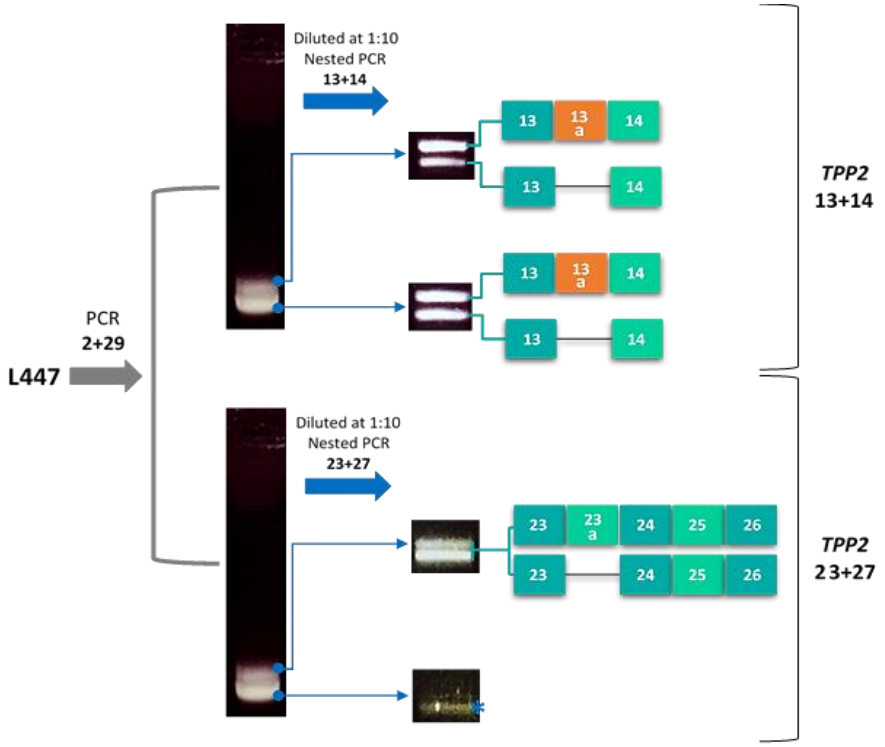


Figure 3.10 - Representative scheme of the steps involved in TPP2 splicing analysis of sample L447 and respective results. (continues next page)

*(figure 3.10 – continuation) PCR 2+29 represents the amplification of exons 2 to 29 of the TPP2 gene. The resulting bands (from an 1.2% agarose gel) were excised, purified and subjected to nested PCRs after a dilution of 1:10. Nested PCR 13+14 represents the amplification of exons 13 to 14, and nested PCR 23+27 the exons 23 to 27 (amplification products of the nested PCRs were ran in a 2% agarose gel). Sequencing results are presented as squared boxes that indicate exons and its respective number. Exons size are not to scale. * - sequencing analysis could not be performed due to technical issues.*

The alternative splicing study was only performed for the region flanked by exons 23 to 27 for samples L905 and L1544 (figure 3.11). Sample 905 only presented the wt transcript deprived of the 23a exon and L1544 exhibited both, with no other alternative splicing events. To notice that sample L905, which only presents one of the transcripts, when amplified from exons 2 to 29 and 23 to 27 only presents one band as well.

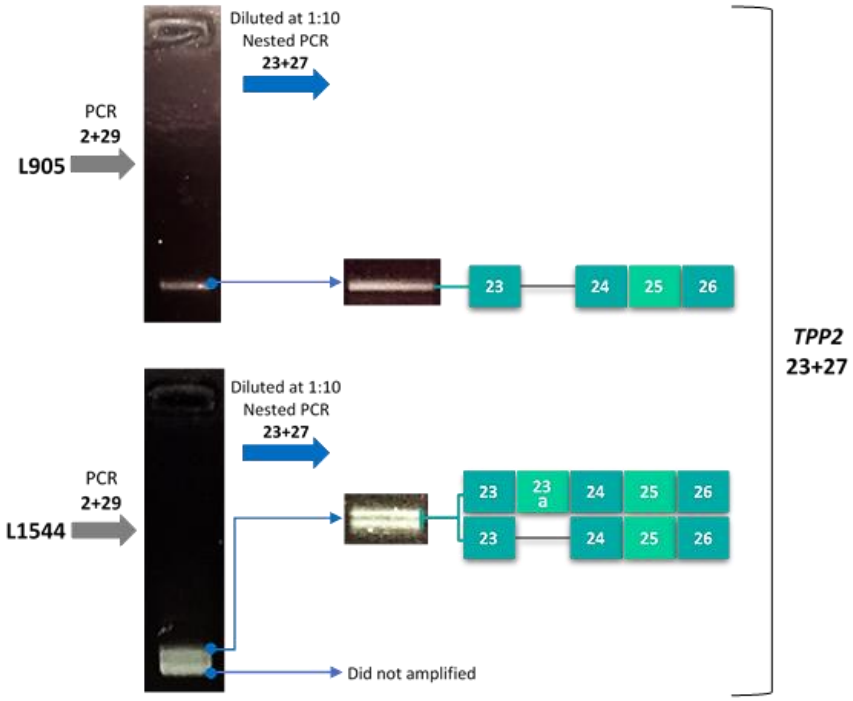


Figure 3.11 - Representative scheme of the steps involved in TPP2 splicing analysis of samples L905 (upper panel) and L1544 (lower panel), and respective results. The resulting bands (from an 1.2% agarose gel) were excised, purified and subjected to a nested PCR after a dilution of 1:10. Nested PCR 23+14 represents the amplification of exons 23 to 27 (amplification products of the nested PCRs were ran in a 2% agarose gel). Sequencing results are presented as squared boxes that indicate exons and its respective number. Exons size are not to scale.

3.3.3 Study of the alternative 13a exon by PTT

The PTT test performed in 16 samples, to access if the insertion of the 13a exon would give rise to a truncated protein, revealed the presence of the truncated protein in the SDS-PAGE gel. The band corresponding to the protein with higher molecular weight corresponded to the wild type protein since it was equivalent to the one expected for the protein obtained from the total amplified fragment, around 95kDa. The second band had a lower molecular weight and appeared to present the same or, in most

cases, a higher intensity than the first band. Given the intensity and the estimated molecular weight (around 61 to 69kDa) of the second band, it is possible to conclude that this band represents a truncated protein due to the 13a exon insertion. Figure 3.12 represents a partial autoradiography where both bands can be observed.

For each of the analyzed samples, the relation between the 13a and the wt transcript had already been observed at the RNA level, thus, a pattern was observed. For samples with an equivalent relationship between the 13a and the wt, or even a higher expression of the wt transcript, the bands presented a similar intensity (as seen in sample L1574 of figure 3.12). However, when samples presented and overexpression of the 13a exon, the band corresponding to the truncated protein presented a much higher intensity, sometimes accompanied by a fading of the wt protein (as seen in samples L1544 and CAs2000 of figure 3.12).

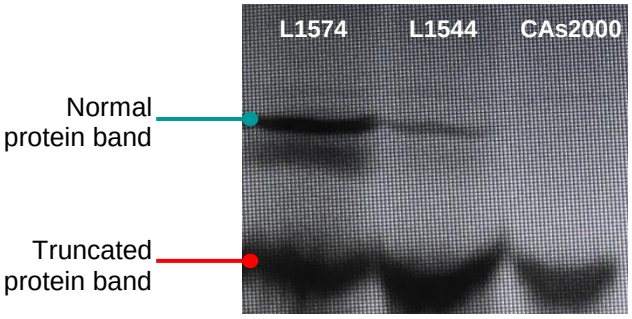


Figure 3.12 – Partial radiography from the PTT, representing the normal and the truncated protein bands for 3 samples: L1574, L1544 and CAs2000.

4. DISCUSSION

4.1 Identification of novel FCCTX susceptibility variants by WES

The WES analysis, performed in the L56 family, had already revealed 6 variants that segregated with the disease in the family, according to each analysis segregation pattern (described in 1.4.1.2). Those variants were found in *MTMR3* (p.Arg645Trp), *TAS1R1* (p.Leu528Phe), *LGR6* (p.Arg675Gln), *DUSP12* (p.Leu2_Leu3ins), *NPR2* (p.Val882Ile) and *SMG7* (p.Pro657His) genes. (Magalhães 2016)

The newly completed examination and segregation studies of the variants obtained from the bioinformatic analysis of WES results in the present study, revealed one more variant that segregated with the disease, arising from analysis 3: *CACNA1S* (c.3607G>A; p.Asp1203Asn). Analysis 3 comprises variants shared by patient 1 (L674), 3 (L408), 4 (L1794) and 5 (L781), excluding patient 2 (L801) as a possible phenocopy. Although the *CACNA1S* variant is described as likely non-pathogenic by Provean and SIFT/Polyphen software, it presents a global and European allele frequency lower than 0.002% and the Mutation Taster software, since this variant is positioned only 3 bases away from the final of exon 28, describes it as disease causing with a possible splicing alteration caused by a loss of the splice-donor.

As for protein function, the *CACNA1S* gene encodes the voltage-gated calcium channel α-subunit Cav1.1, one of the five subunits of the slowly inactivating L-type voltage-dependent calcium channel in skeletal muscle cells, and has an important role in Ca^{2+}-mediated excitation-contraction coupling via interaction with RYR1 (ryanodine receptor 1), which triggers Ca^{2+} release from the sarcoplasmic reticulum and ultimately results in muscle contraction (figure 4.1). (Klingler *et al.* 2014) Mutations in this gene have been associated with hypokalemic periodic paralysis, thyrotoxic periodic paralysis and malignant hyperthermia susceptibility (Orphanet database: http://www.orpha.net), however the variant identified in this study is not linked to any of these disorders.

Even though there is no evident link between CRC and *CACNA1S*, studies have reported that modulation of intracellular calcium levels can affect CRC cell motility (Sundaramoorthy *et al.* 2015). Evidence of a role of Cav1 channels and related regulatory proteins in both CD4 and CD8 T-cell functions such as proliferation, survival, cytokine production, and cytolysis (Badou *et al.* 2013) and, another study (Jacquemet *et al.* 2016), has involved *CACNA1S* directly with filopodia formation and cancer cell invasion in breast cancer cells.

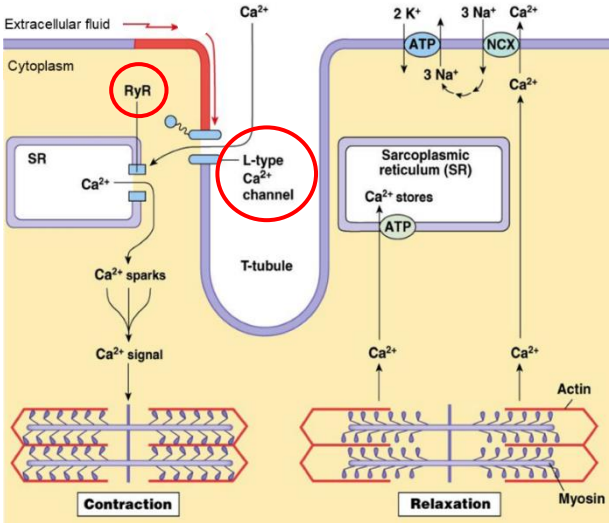


Figure 4.1 – Mechanisms of excitation-contraction coupling and relaxation in skeletal muscle cells. The CACNA1S gene encodes one of the five subunits of the slowly inactivating L-type voltage-dependent calcium channel (highlighted in red), that interacts with ryanodine receptors (RyR – highlighted in red). (Adapted from: http://faculty.pasadena.edu/dkwon/chapter%2014/Chapter%2014_cardiology_files/slide0059_image021.jpg)

According to The Human Protein Atlas database (The Human Protein Atlas: http://www.proteinatlas.org), the protein encoded by the *CACNA1S* gene appears to be rarely or weakly expressed in CRC cells and in normal colon and rectum tissue, which may restrain the connection between alterations in this gene and a CRC syndrome. However, when revising all the variants that segregated with the disease in the family, there is a tendency for mutations to occur in chromosome 1 (from the 7 variants found, 5 were found in chromosome 1: *TAS1R1* - 1:6577058, *SMG7* - 1:183544963, *CACNA1S* - 1:201058410, *LGR6* - 1:202318327 and *DUSP12* - 1:161749807) which may reveal the possibility that this family's phenotype is not associated with only one highly penetrant pathogenic mutation, but the result of an accumulation of mutations in chromosome 1 associated with the *MTMR3* and the *NPR2* variants that revealed an *in silico* pathogenic prevision.

The revelation that the *MTMR3* and the *TAS1R1* variants also segregated in the CAs1555 patient, from an oldest generation, greatly enhances the assumption that these variants may be involved in the FCCTX predisposition in the family.

Both the MTMR3 and the TAS1R1 proteins have been associated with the mTOR (mechanistic target of rapamycin) signaling pathway (Hao *et al.* 2016), one of the pathways linked to CRC (Francipane and Lagasse 2014) and with the autophagy process, a process implied as having a dual role in tumorigenesis, including in CRC carcinogenesis. (Burada *et al.* 2015) The *NPR2* gene encodes natriuretic peptide receptor B that has guanylate cyclase activity which catalyzes the conversion of GTP to cGMP (cyclic guanosine monophosphate) (GeneCards: http://www.genecards.org). cGMP levels help to regulate the turnover of epithelial cells and maintain homeostasis of intestinal mucosa. Abnormalities in this process may lead to the formation of tumors within the intestinal tract. (Camici 2008)

Summing up, the variant found in *MTMR3* seems linked to this family's phenotype and the identification of the *TAS1R1* variant in an affected patient from an older generation and its *in silico* highly pathogenic prevision, reinforces the involvement of this variant in FCCTX. The remaining variants found

in chromosome 1 seem to predispose to the developing of adenomas and the patients that present the most severe phenotypes, with an earlier age of onset, appear to present the *MTMR3* simultaneously with the *NPR2* variant.

4.1.1 Analysis of potential copy number variation (disease associated) from the WES data

The copy number variation analysis for the L56 family after the application of selection criteria was reduced to 22 amplicons with a deletion scenario. Most of these amplicons contain at least one exon or a large portion of an exon of the given gene in the detected region. However, this analysis is not complete. The deletion must be confirmed by another methodology like qPCR. If confirmed, these amplicons need to be subjected to segregation studies in the affected and unaffected individuals of the family, to see if the deletion is consistent with the disease's phenotype.

4.2 Germline mutational analysis of candidate genes for FCCTX susceptibility

4.2.1 *MTMR3*

The *MTMR3* gene encodes a phosphatase protein of the same name that acts on lipids with a phosphoinositol headgroup that can hydrolyze PtdIns3P (phosphatidylinositol 3-phosphate) and PtdIns(3,5)P2 (phosphatidylinositol 3,5-biphosphate) and dephosphorylate proteins phosphorylated on Serine, Threonine and Tyrosine residues. (Zhao *et al.* 2000) The protein contains a PH-GRAM domain at its N-terminus for phosphoinositide-binding, a myotubularin phosphatase domain, for its enzymatic activity, a coiled-coil domain that mediates protein interactions, and a FYVE zinc finger domain at its C-terminus, with the potential to bind phosphoinositide lipids (figure 4.2). (Taguchi-Atarashi *et al.* 2010)

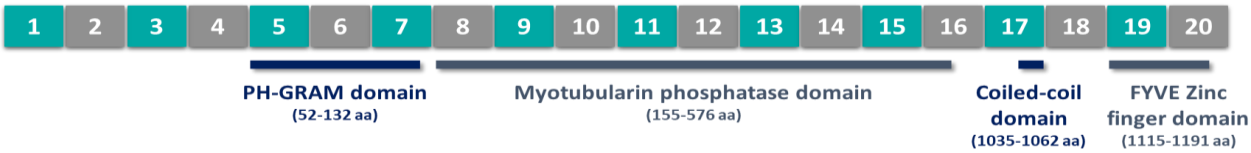


Figure 4.2 - Schematic representation of the MTMR3 transcript (Ensemble ID: ENST00000401950) and localization of the main protein domains. The squared boxes represent exons and is respective number. Exons size are not to scale. aa: amino acid. Protein domains were retrieved from the Ensembl (http://www.ensembl.org) and UniProt (http://www.uniprot.org) database.

The germline mutational analysis for *MTMR3* revealed the presence of 3 missense variants in 3 different samples, all found within exons 16 and 17. 2 of these variants, c.1811C>A (p.Pro604His) and c.1933C>T (p.Arg645Trp) were outlined as possibly pathogenic by the predicted software used.

The c.1811C>A (p.Pro604His) variant is not described in the available databases and was outlined as pathogenic by SIFT and Mutation Taster predictive software and possibly damaging by PolyPhen. Provean defines this mutation as neutral but with a score that is considerably close to the

deleterious threshold. Although this variant is not coexisting with a protein domain, it is located near to a series of phosphorylation residues (PhosphoSitePlus database: https://www.phosphosite.org), in which, the introduction of a different amino acid may induce alterations in protein's arrangement that conceals one or more phosphorylation residues, which may affect is function, due to post-translational changes or protein-protein interactions. Taken these predictions together and since this variant was found in an index patient of a FCCTX-like family, this variant is a suitable candidate for further segregation studies within this patient's family.

The c.1933C>T (p.Arg645Trp) variant, which was also the mutation found to segregate with the disease in the L56 family, has already been extensively characterized in a previous study (Magalhães 2016) and revealed a highly pathogenic prediction, being also located within a series of phosphorylation residues, suggesting to be a strong candidate for the L56 family's FCCTX phenotype. Since this mutation was already ´flagged´ as a strong contender for FCCTX susceptibility an effort was made in trying to access the individual's family pedigree (figure 4.3).

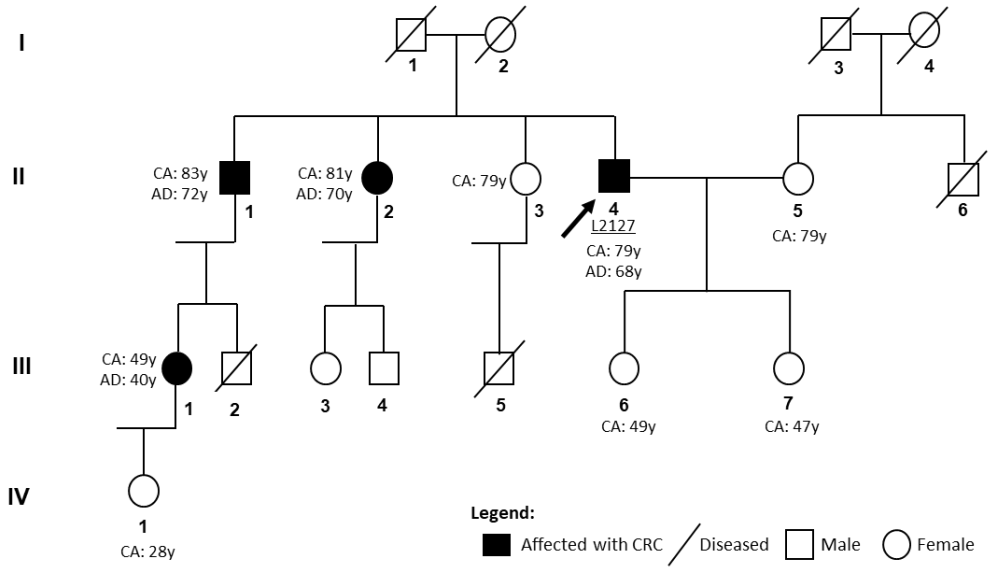


Figure 4.3 – Family's pedigree of the individual were the c.1933C>T (p.Arg645Trp) variant was found (L2127, marked with an arrow). CA: current age; AD: age at diagnosis (carcinoma); y: years.

Looking at this family's pedigree (figure 4.3) there are similar aspects with the L56 family, besides the autosomal dominant inheritance pattern and an apparent phenotypic anticipation, where younger generations appear to develop CRC at an earlier age, which is common to the majority of CRC families, the lack of extra colonic cancers and the good prognosis of the affected members strikingly resemble the L56 family. Accordingly, the affected patients appear to have good prognosis years after their diagnosis, indicating that the mutation and/or combination of mutations in these families may contribute for cancer initiation and progression, but not to its aggressiveness. The only thing that seems to differ between these 2 families is the age at diagnosis, where the L2127's family appears to develop carcinoma at an older age (one of the described characteristics of FCCTX is the development of CRC at an older age when comparing to Lynch syndrome). These findings support that this variant can be in

part responsible for this family's susceptibility for CRC. Thus, a segregation study of this variant with the disease in this family is major concern for defining this variant's possible pathogenicity and involvement in FCCTX susceptibility.

The study of 50 Portuguese controls, for the c.1933C>T (p.Arg645Trp) variant, did not reveal the presence of this mutation in any of the samples, which is against the possibility that this variant was found in individuals from two different families due to a high allele frequency within the Portuguese population, that is not reported in the databases.

The MTMR3 protein plays a role in the autophagy process, a process that has been continuously related to progression and tumor formation in CRC (Zheng *et al.* 2012), and is also believed to be involved in the mTOR signaling pathway, a pathway already associated to CRC carcinogenesis. (Hao *et al.* 2016) In addition, a study has suggested that alterations in the *MTMR3* gene may play a role in the development of gastric and colorectal cancer due to a misregulation of its phosphatase functions (Song *et al.* 2010), and another study demonstrated that *MTMR3* silencing led to decreased cell proliferation, impaired colony formation, arrested cell cycle, and increased apoptosis in CRC cells, thus reinforcing the involvement of *MTMR3* in CRC progression. (Zheng *et al.* 2014) Taken together, these studies and our findings regarding the identified *MTMR3* germline mutations in FCCTX families seem to support the involvement of *MTMR3* as a possible contributor for FCCTX's susceptibility.

Moreover, the mutational analysis of *MTMR3* also revealed the presence of the normal *MTMR3* transcript, co-occurring with alternative transcripts in 17/19 samples. The *MTMR3* gene has 6 protein coding transcripts and another 6 non-coding transcripts described at the Ensembl database, however none of the splicing alterations found within this study have been reported for this gene, and no mutations were found, within this analysis that could explain these findings. The splicing alterations were found between exons 9 and 15, where it was observed the exclusion of exons 10, 11, 13 and 14 without any particular order of exclusion or preferential selection, the only pattern that was observed was the exclusion of exon 14 in all the alternative splicing forms. All the described protein coding transcripts present all these exons, since they are the major components of the myotubularin phosphatase domain, the domain that confers this protein is enzymatic function. (Taguchi-Atarashi *et al.* 2010) So if these alternative splicing isoforms can give rise to a protein coding transcript, the protein will be severely compromised by the impairment of its phosphatase function, and the misregulation of MTMR3 phosphatase function has already been proposed in the development of CRC. (Song *et al.* 2010)

However, the observed splicing alterations were always followed by the presence of the normal transcript, which in the end may be the only one who will give rise to a protein. In addition, these alterations were found in 89% of the analyzed samples, which suggests that these alterations might not be pathogenic. Nevertheless, the observed alterations may have a role not as protein coding, but as a part of a RNA regulatory network has a non-coding transcript, since it is believed that non-coding RNAs can contribute to regulation of mRNA stability, translational efficiency and cellular localization. (Mockenhaupt and Makeyev 2015) (Mattick and Makunin 2006) Thus, additional studies may need to be performed in order to clarify the implication of these splicing alterations.

4.2.2 *TAS1R1*

The TAS1R1 protein is a G protein-coupled receptor that is a component of the heterodimeric amino acid taste receptor TAS1R1/TAS1R3, formed as a dimer of TAS1R1 and TAS1R3 proteins and is responsible for the umami taste perception. Contrary to what was believed, taste receptors, in particular G-protein-coupled, are not restricted to the tongue. They are distributed throughout the stomach, intestine and pancreas, where they aid the digestive process, regulation of energy and glucose homeostasis. (Trivedi 2012) (Depoortere 2014) Like MTMR3, TAS1R1, more specifically the TAS1R1/TAS1R3 G-protein-coupled receptor, has been involved in the autophagy process by interaction with one of the mTOR complexes, mTORC1 (mechanistic target of rapamycin complex 1). (Wauson *et al.* 2012) TAS1R1-TAS1R3 can activate mTORC1 through direct sensing of amino acids, promoting the activation of phospholipase C (PLC), the increase of calcium influx, and the activation of MAPK1-MAPK3, thus stimulating mTORC1 and leading to an inhibition of autophagy (figure 4.4). (Wauson *et al.* 2012) (Wauson *et al.* 2013) TAS1R1-TAS1R3 is also required for the amino acid-induced mTOR localization to the lysosome, a necessary step in mTORC1 activation (figure 4.4). (Wauson *et al.* 2013)

Disruptions in TAS1R1/TAS1R3 have been shown to induce autophagy by inhibiting mTORC1 activity (Wauson *et al.* 2012), and genetic variants in *TAS1R1* were recently associated with an increased risk for gastric cancer. (Choi *et al.* 2016) Given this information and the role of autophagy in CRC, the discovery of variants in *TAS1R1* can give new insights on FCCTX susceptibility.

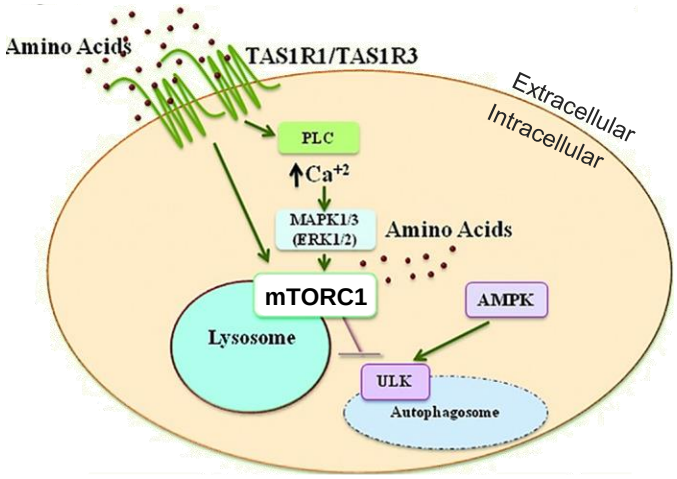


Figure 4.4 –Regulation of autophagy by interaction of G-protein-coupled taste receptor TAS1R1/TAS1R3 with mTORC1. (Adapted from Wauson et al. 2013)

The germline mutational analysis for *TAS1R1* revealed the presence of 3 missense variants, c.1067C>G (p.Ser356Cys), c.1043C>T (p.Ala348Val) and c.953T>C (p.Ile318Thr) with probable pathogenic predictions by the software used. These variants are all present at the ligand binding region domain (figure 4.5) and exhibit a low global and european allele frequency.

Given the localization of these 3 variants in a key domain of the protein, and their *in silico* predictions of possible pathogenicity and splicing alterations, these variants are good candidates for

segregation studies in the respective families of the individuals where the variants were found. And the discovery of these probably pathogenic variants gives new insights as for the probable involvement of the *TAS1R1* gene in FCCTX susceptibility.

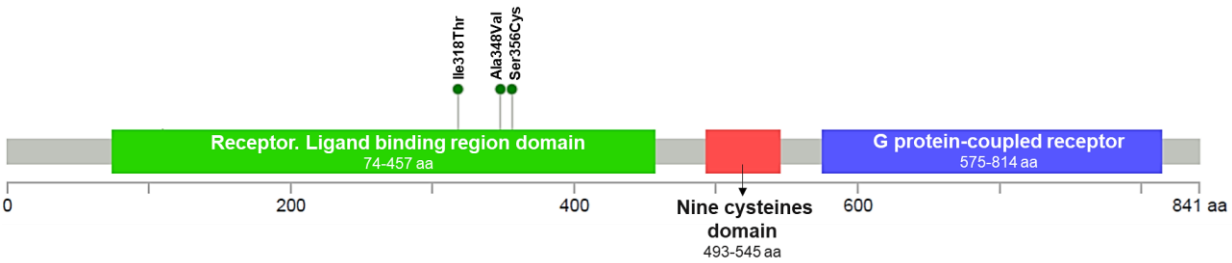


Figure 4.5 – Main domains of the TAS1R1 protein and the localization of 3 of the variants found. aa: amino acid. Adapted from: cBioPortal for Cancer Genomics (http://www.cbioportal.org)

Since both *TAS1R1* and *MTMR3* play a role in autophagy and regulate mTOR, and given the variants found in patients from FCCTX and FCCTX-like families, it would be an interesting approach to correlate the expression of both genes with the expression of mTOR and the participants in the autophagy and mTOR regulation cascade, to search for potential alterations between normal tissue and samples derived from carcinoma and/or adenoma.

4.3 *TPP2* role in FCCTX susceptibility

The gene expression analysis of the undescribed 13a transcript revealed that almost all the analyzed patients of L55 family (all develop adenoma) had a balanced relation between the wt and the 13a transcript, with the exception of L447, that presented an overexpression of the wt transcript, and L1544 that presented an overexpression of the alternative transcript.

In the affected group, 4/6 revealed an overexpression of the 13a transcript where 3/4 had that overexpression followed by an almost total reduction of the wt transcript. The 2/6 remaining samples of this group had a slight increase of the wt expression, one of the individuals develop carcinoma and the other does not present carcinoma or adenoma but has an identified LS mutation in the MMR genes.

As for the samples derived from healthy individuals, all exhibited an overexpression of the wt, that was at least 1-fold higher than the alternative expression. Given these results, it seems like the balance or imbalance between the 13a alternative transcript and the wt, could play a role in adenoma/carcinoma susceptibility, since an overexpression of the alternative transcript was only detected in samples outside the healthy group. As already stated (in section 1.4.1.1), the previous SNP array was suggestive that the chromosomal events in 13q occurred at the level of tumor initiation, which seems consistent with these results considering that all the analyzed patients that developed adenoma but not carcinoma (with the exception of L447) exhibit an equal relationship between the wt and 13a expression or even an increased 13a expression, but never a significant overexpression of the wt transcript, thus implying a possible role of the alternative transcript of the *TPP2* gene in colorectal tumorigenesis.

The presence of the alternative transcript 13a was not only accessed by qPCR, the analysis of splicing isoforms also revealed the presence of this transcript in almost all analyzed samples. In addition to this alternative transcript, another two non-described isoforms were found in two different samples: the alternative splicing with the skipping of exon 24 and/or 23a (sample L950) and the alternative splicing derived from the skipping of exon 25 and the first half of exon 26 (sample L1590). Although these novel isoforms are not present in any of the main domains of the TPP2 protein (figure 4.6), additional studies are needed to perceive if these isoforms can play a role in CRC tumorigenesis, since alternative splicing of cancer-associated genes has been implied in tumor progression and initiation. (Ghigna *et al.* 2008) (Sveen *et al.* 2016)

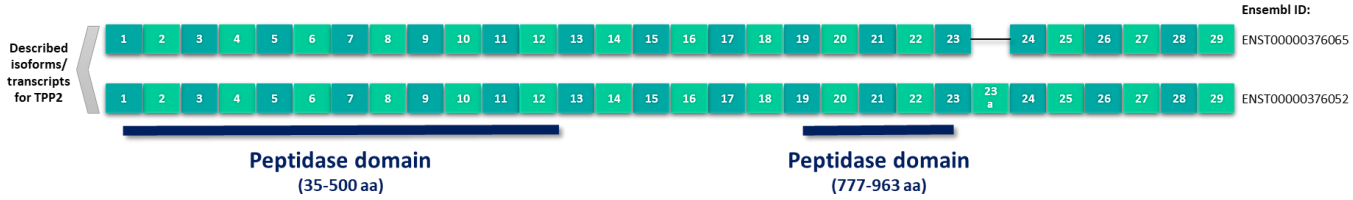


Figure 4.6 - Schematic representation of the described isoforms for TPP2 and localization of the main protein domains. The squared boxes represent exons and is respective number. Exons size are not to scale. aa: amino acid. Protein domains were retrieved from the Ensembl (http://www.ensembl.org) and UniProt (http://www.uniprot.org) database.

In the *TPP2* splicing analysis, most of the samples presented two bands when an almost full open reading frame amplification was performed. The appearance of two distinct molecular weight bands seems to be correlated with the presence of the alternative described and non-described transcripts. One thing that was evident in all the amplified samples was that the detection of the transcript that exhibited the 23a exon was always dependent on the presence of the normal transcript without the 23a exon, considering that the normal transcript was found solely, while the 23a isoform was always found alongside with the normal isoform. And the same was found in relation to the non-described 13a isoform, which was dependent on the normal isoform. Samples L926 and L1590 were the only ones that did not present the 13a isoform (apart from samples L905 and L1544, where the splicing analysis was not performed for region 13 to 14, but have already displayed an expression of the alternative transcript by qPCR), however in the gene expression analysis sample L926 presented only a small overexpression of the wt transcript in comparison to the 13a. Thus, these results do not seem to be in accordance and the splicing analysis for L926 should be repeated, since qPCR analysis is a highly sensitive technique and the results were confirmed by at least two independent experiments.

The PTT confirmed the outlined premise that the insertion of the alternative 13a exon could give rise to a premature stop of the amino acid synthesis, giving rise to a truncated protein, due to an inclusion of a stop codon within its sequence (TAG). Figure 4.7 shows the differences between the amino acid sequences that occurs in a normal translation of the *TPP2* gene and the hypothetic translation that occurs when there is the insertion of the alternative 13a exon.

Figure 4.7 – Amino acid (aa) sequences for the normal TPP2 protein (left panel), with a molecular weight of 138 kDa and the truncated protein generated by the insertion of the 13a exon (right panel), with an approximate (~) molecular weight of 65 kDa. The underlined aa represent the amplified region that was subjected to PTT analysis and red letters represent the additional aa provided by the 13a insertion.

This truncated protein losses the second peptidase domain (figure 4.6), in addition to a major loss by being reduced to half of its original sequence. Being the TPP2 protein the largest known eukaryotic peptidase, its structural assembly is a major concern for its normal function. (Rockel *et al.* 2012) TPP2 assembles into a large homooligomer under physiological conditions and, upon oligomerization, its activity strongly increases, suggesting that activation is assembly induced. Also, TPP2's C-terminal domain contains a structural motif often involved in protein-protein interactions. (Schönegge *et al.* 2012) This truncated protein losses its normal C-terminal domain which may have a critical effect on protein-protein interactions and in the overall assembly and oligomerization of TPP2's homooligomer, which will most likely affect the protein's normal function and assembly.

Although the PTT revealed this truncated protein, it cannot be entirely extrapolated to an *in vivo* context, since there are innumerous regulating factors, that an *in vitro* synthesis cannot predict, but it provides a starting point.

Ultimately, the multiple studies performed on *TPP2* suggest that the alternative expression of non-described isoforms, especially the 13a isoform, could influence the TPP2 protein function and ultimately play a role in CRC tumorigenesis. This assumption is reinforced by the pathways and biological process where TPP2 has been involved and the immune system role in CRC and the interaction of TPP2 with the MHC class I. It is well recognized that cytotoxic lymphocytes (also referred as CD8+ T) constitute one of the most important effector mechanisms of anti-tumor-immunity (Deschoolmeester *et al.* 2011) and MHC class I proteins are responsible of antigen presentation to cytotoxic lymphocytes. (Neefjes *et al.* 2011) However, since the TPP2 protein is one of the major aminopeptidases involved in the presentation of peptides to the MHC class I, a disruption of its functions may contribute to the immune escape of colon-carcinoma cells. (Preta *et al.* 2008)

Still, there is the need to discriminate if these alterations occur in other colorectal cancer entities or if they are strictly correlated to an increased FCCTX susceptibility, since alterations in the expression of the alternative 13a isoform were also found in patients affected with LS.

5. CONCLUSION

A considerable portion of HNPCC families fail to present mutations in the MMR genes and have MSS tumors. These families are classified as FCCTX and the genetic etiology behind this syndrome remains unknown. Therefore, the identification of genetic variants/genes that may contribute to this syndrome predisposition is a major step towards a better screening, surveillance and maintenance of the families.

The WES analysis performed in a FCCTX family, whose tumors presented a molecular TSG^{-} signature, had already revealed the presence of 6 distinct variants that segregated with the disease according to different segregation models, taking into account possible phenocopies in the family. In the present study, another variant (in the *CACNA1S* gene) was found to segregate with the disease according to one of the models. The identification of the *MTMR3* and *TAS1R1* variants in a patient from an older generation favors the involvement of these variants in the family's phenotype.

Moreover, the presence of probably pathogenic mutations in the *MTMR3* and *TAS1R1* genes in index patients from FCCTX and FCCTX-like families and their protein functions that are highly associated with known oncogenic pathways, reinforces the involvement of these genes in FCCTX susceptibility. Also, the finding of the same *MTMR3* mutation observed in the WES analysis, with highly *in silico* pathogenic predictions, in an index patient of an FCCTX-like family, whose families' phenotypic features resemble those of the WES family, emphasizes the involvement of this specific variant in FCCTX susceptibility.

Altogether, these evidences sustain the hypothesis that the FCCTX is likely not linked to a highly penetrant single gene disorder, but is rather a polygenic and heterogeneous condition that is formulated due to the accumulation of mutations in different genes, especially genes involved in metabolic and autophagy processes such as *MTMR3*, *TAS1R1* and *NPR2*, that 'help each other' in the tumorigenesis process.

The finding of alternative splicing isoforms in *MTMR3* does not seem to be associated with a predisposition for FCCTX, since almost all the analyzed patients (89%) presented such alterations.

The analysis of potential CNV, from the exome sequencing data, may reveal variations in amplicons associated with the family's phenotype, however further studies are needed to unveil the influence of the selected amplicons.

The *TPP2* study suggests that the alternative 13a transcript may be involved in the earlier stages of colorectal tumorigenesis, in association with other colorectal cancer entities rather than being exclusively linked with a FCCTX predisposition. Also, *TPP2* may play a role in colorectal cancer due to a disruption in the immune response.

Thinking forward and given the WES analysis results, the colon pathology group has an ongoing project that aims the construction of vectors expressing wild type and mutant candidate genes for *in vitro* expression in normal and colorectal cell models to access their involvement in the colorectal tumorigenesis. Also, an expression analysis is currently ongoing to access the expression of the candidate genes in tumor/matched normal tissues from WES family patients as well as another

expression study aiming to correlate the expression of *MTMR3* and *TAS1R1* with mTOR and the participants in the autophagy regulation cascade in tumor/matched normal tissues. Preliminary results of this analysis have suggested an apparently reduced expression of *MTMR3* in one carcinoma and one adenoma with high-grade dysplasia from the WES family, in comparison with normal mucosa.

6. REFERENCES

Albuquerque, C., Breukel, C., van der Luijt, R., *et al.* 2002. The “just-right” signaling model: APC somatic mutations are selected based on a specific level of activation of the beta-catenin signaling cascade., *Human molecular genetics*, 11(13), pp. 1549–60.

Auton, A., Abecasis, G. R., Altshuler, D. M., *et al.* 2015. A global reference for human genetic variation, *Nature*, 526(7571), pp. 68–74.

Badou, A., Jha, M. K., Matza, D., *et al.* 2013. Emerging Roles of L-Type Voltage-Gated and Other Calcium Channels in T Lymphocytes, *Frontiers in Immunology*. Frontiers, 4, p. 243.

Beckler, G. 2005. The protein truncation test, in *Bridging the Gap Between Genomics and Proteomics: In Vitro Expression Tools for Functional Genomics*. Promega.

Belo, H. 2010. *Papel Do Gene APC e Identificação de Regiões Cromossómicas Envolvidas No Sindroma Familiar de Cancro Do Cólon e Recto Do Tipo X*. Escola Superior de Saúde Egas Moniz.

Bio-Rad Laboratories 2006. Real-Time PCR, in *Applications Guide*.

Boland, C. R. 2005. Evolution of the Nomenclature for the Hereditary Colorectal Cancer Syndromes, *Familial Cancer*. Kluwer Academic Publishers, 4(3), pp. 211–218.

Brenner, H., Kloor, M. and Pox, C. P. 2014. Colorectal cancer, *The Lancet*. Elsevier Ltd, 383(9927), pp. 1490–1502.

Burada, F., Nicoli, E. R., Ciurea, M. E., *et al.* 2015. Autophagy in colorectal cancer: An important switch from physiology to pathology., *World journal of gastrointestinal oncology*. Baishideng Publishing Group Inc, 7(11), pp. 271–84.

Burt, R. 2007. Inheritance of Colorectal Cancer., *Drug discovery today. Disease mechanisms*. NIH Public Access, 4(4), pp. 293–300.

Camici, M. 2008. Guanylin peptides and colorectal cancer (CRC), *Biomedicine & Pharmacotherapy*, 62(2), pp. 70–76.

Carballal, S., Leoz, M. L., Moreira, L., *et al.* 2014. Hereditary colorectal cancer syndromes, *Colorectal Cancer*. Future Medicine Ltd London, UK , 3(1), pp. 57–76.

Carethers, J. M. and Stoffel, E. M. 2015. Lynch syndrome and Lynch syndrome mimics: The growing complex landscape of hereditary colon cancer, *World Journal of Gastroenterology*, 21(31), p. 9253.

Choi, J.-H., Lee, J., Choi, I. J., *et al.* 2016. Variations in TAS1R taste receptor gene family modify food intake and gastric cancer risk in a Korean population, *Molecular Nutrition & Food Research*, 60(11), pp. 2433–2445.

Colussi, D., Brandi, G., Bazzoli, F., *et al.* 2013. Molecular pathways involved in colorectal cancer: implications for disease behavior and prevention., *International journal of molecular sciences*, 14(8), pp. 16365–16385.

Daly, M., Doreen Agnese, Kathleen Calzone, *et al.* 2017. *Genetics of Colorectal Cancer (PDQ®): Health Professional Version*, *PDQ Cancer Information Summaries*. National Cancer Institute (US). Available at: http://www.ncbi.nlm.nih.gov/pubmed/26389505 (Accessed: 13 April 2017).

Depoortere, I. 2014. Taste receptors of the gut: emerging roles in health and disease, *Gut*, 63(1), pp. 179–190.

Deschoolmeester, V., Baay, M., Lardon, F., *et al.* 2011. Immune Cells in Colorectal Cancer: Prognostic Relevance and Role of MSI., *Cancer microenvironment : official journal of the International Cancer Microenvironment Society*. Springer, 4(3), pp. 377–92.

Duarte, M. 2015. *Envolvimento de variantes genéticas específicas na suscetibilidade para o risco aumentado de cancro do cólon e reto e na agressividade tumoral*. Escola Superior de Saúde Egas Moniz.

Den Dunnen, J. T. and Van Ommen, G.-J. B. 1999. The protein truncation test: A review, *Human Mutation*, 14(2), pp. 95–102.

Ferlay, J., Soerjomataram, I., Ervik, M., *et al.* 2013. *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11*, *Lyon, France: International Agency for Research on Cancer*. Available at: http://globocan.iarc.fr/ (Accessed: 13 April 2017).

Francipane, M. G. and Lagasse, E. 2014. mTOR pathway in colorectal cancer: an update, *Oncotarget*, 5(1), pp. 49–66.

Francisco, I., Albuquerque, C., Lage, P., *et al.* 2011. Familial colorectal cancer type X syndrome: two distinct molecular entities?, *Familial Cancer*, 10(4), pp. 623–631.

Gala, M. and Chung, D. C. 2015. Hereditary CRC Syndromes, in *Intestinal Tumorigenesis*. Cham: Springer International Publishing, pp. 1–28.

GATC Biotech 2017. *SANGER SEQUENCING: Overview–What is Sanger sequencing?*, *Expertise*. Available at: https://www.gatc-biotech.com/ (Accessed: 18 April 2017).

Ghigna, C., Valacca, C. and Biamonti, G. 2008. Alternative splicing and tumor progression., *Current genomics*. Bentham Science Publishers, 9(8), pp. 556–70.

Giardiello, F. M., Allen, J. I., Axilbund, J. E., *et al.* 2014. Guidelines on Genetic Evaluation and Management of Lynch Syndrome, *Diseases of the Colon & Rectum*, 57(8), pp. 1025–1048.

Guarinos, C., Sánchez-Fortún, C., Rodríguez-Soler, M., *et al.* 2012. Serrated polyposis syndrome: molecular, pathological and clinical aspects., *World journal of gastroenterology*. Baishideng Publishing Group Inc, 18(20), pp. 2452–61.

Hao, F., Itoh, T., Morita, E., *et al.* 2016. The PtdIns3-phosphatase MTMR3 interacts with mTORC1 and suppresses its activity, *FEBS Letters*. Edited by N. Mizushima, 590(1), pp. 161–173.

Hinoue, T., Weisenberger, D. J., Lange, C. P. E., *et al.* 2012. Genome-scale analysis of aberrant DNA methylation in colorectal cancer, *Genome Research*, 22(2), pp. 271–282.

Hisamuddin, I. M. and Yang, V. W. 2006. Molecular Genetics of Colorectal Cancer: An Overview., *Current colorectal cancer reports*. NIH Public Access, 2(2), pp. 53–59.

Ishmael, F. T. and Stellato, C. 2008. Principles and applications of polymerase chain reaction: basic science for the practicing physician, *Annals of Allergy, Asthma & Immunology*, 101(4), pp. 437–443.

Jacquemet, G., Baghirov, H., Georgiadou, M., *et al.* 2016. L-type calcium channels regulate filopodia stability and cancer cell invasion downstream of integrin signalling, *Nature Communications*, 7, p. 13297.

Jasperson, K. W., Tuohy, T. M., Neklason, D. W., *et al.* 2010. Hereditary and familial colon cancer., *Gastroenterology*. NIH Public Access, 138(6), pp. 2044–58.

Jian, X., Boerwinkle, E., Liu, X., *et al.* 2014. In silico prediction of splice-altering single nucleotide variants in the human genome, *Nucleic Acids Research*. BioMed Central, 42(22), pp. 13534–13544.

Kinnersley, B., Chubb, D., Dobbins, S. E., *et al.* 2016. Correspondence: SEMA4A variation and risk of colorectal cancer, *Nature Communications*, 7, p. 10611.

Klingler, W., Heiderich, S., Girard, T., *et al.* 2014. Functional and genetic characterization of clinical malignant hyperthermia crises: a multi-centre study, *Orphanet Journal of Rare Diseases*, 9(1), p. 8.

Kovacs, M. E., Papp, J., Szentirmay, Z., *et al.* 2009. Deletions removing the last exon of TACSTD1 constitute a distinct class of mutations predisposing to Lynch syndrome, *Human Mutation*, 30(2), pp. 197–203.

Kumar, P., Henikoff, S. and Ng, P. C. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm, *Nature Protocols*, 4(8), pp. 1073–1081.

Kumar, V., Abbas, A. K., Aster, J. C., *et al.* 2015. *Robbins and Cotran Pathologic Basis of Disease*. 9th edn. Philadelphia: Elsevier.

Lao, V. V. and Grady, W. M. 2011. Epigenetics and colorectal cancer., *Nature reviews. Gastroenterology &*

hepatology. Nature Publishing Group, 8(12), pp. 686–700.

Lee, P. Y., Costumbrado, J., Hsu, C.-Y., *et al.* 2012. Agarose Gel Electrophoresis for the Separation of DNA Fragments, *Journal of Visualized Experiments*, (62).

Lek, M., Karczewski, K. J., Minikel, E. V., *et al.* 2016. Analysis of protein-coding genetic variation in 60,706 humans, *Nature*, 536(7616), pp. 285–291.

Lindor, N. M. 2009a. Familial Colorectal Cancer Type X, in *Genetics of Colorectal Cancer*. New York, NY: Springer New York, pp. 183–186.

Lindor, N. M. 2009b. Familial Colorectal Cancer Type X: The Other Half of Hereditary Nonpolyposis Colon Cancer Syndrome, *Surgical Oncology Clinics of North America*, 18(4), pp. 637–645.

Lindor, N. M., Rabe, K., Petersen, G. M., *et al.* 2005. Lower Cancer Incidence in Amsterdam-I Criteria Families Without Mismatch Repair Deficiency, *JAMA*, 293(16), p. 1979.

Llor, X., Pons, E., Xicola, R. M., *et al.* 2005. Differential Features of Colorectal Cancers Fulfilling Amsterdam Criteria without Involvement of the Mutator Pathway, *Clinical Cancer Research*, 11(20).

Magalhães, A. 2016. *Identificação de Novos Genes de Susceptibilidade para o Cancro do Colon e Recto do Tipo X*. Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa.

Majewski, J., Schwartzentruber, J., Lalonde, E., *et al.* 2011. What can exome sequencing do for you?, *Journal of Medical Genetics*, 48(9), pp. 580–589.

Mattick, J. S. and Makunin, I. V. 2006. Non-coding RNA, *Human Molecular Genetics*. Oxford University Press, 15(90001), pp. R17–R29.

Miranda, N., Portugal, C., Nogueira, P. J., *et al.* 2016. *Doenças Oncológicas em Números 2015 - Programa Nacional para as Doenças Oncológicas*. Lisboa.

Mockenhaupt, S. and Makeyev, E. V 2015. Non-coding functions of alternative pre-mRNA splicing in development., *Seminars in cell & developmental biology*. Elsevier, 47–48, pp. 32–9.

NCBI Resource Coordinators 2017. Database Resources of the National Center for Biotechnology Information, *Nucleic Acids Research*, 45(D1), pp. D12–D17.

Neefjes, J., Jongsma, M. L. M., Paul, P., *et al.* 2011. Towards a systems understanding of MHC class I and MHC class II antigen presentation, *Nature Reviews Immunology*, 11(12), pp. 823–36.

Nieminen, T. T., Abdel–Rahman, W. M., Ristimäki, A., *et al.* 2011. BMPR1A Mutations in Hereditary Nonpolyposis Colorectal Cancer Without Mismatch Repair Deficiency, *Gastroenterology*, 141(1), pp. e23–e26.

Nieminen, T. T., O'Donohue, M.-F., Wu, Y., *et al.* 2014. Germline mutation of RPS20, encoding a ribosomal protein, causes predisposition to hereditary nonpolyposis colorectal carcinoma without DNA mismatch repair deficiency., *Gastroenterology*. Elsevier, 147(3), p. 595–598.e5.

Pereira, C. 2013. *Estudo do gene NRIP1 e de novos loci de susceptibilidade para o cancro do cólon e recto familiar do tipo X*. Escola Superior de Saúde Egas Moniz.

Pereira, G. 2014. *Estudo de regiões de susceptibilidade para o cancro do cólon e recto familiar do tipo x: análise de genes candidatos e de ganhos/deleções em tumores*. Escola Superior de Saúde Egas Moniz.

Peters, J., Schönegge, A.-M., Rockel, B., *et al.* 2011. Molecular ruler of tripeptidylpeptidase II: Mechanistic principle of exopeptidase selectivity, *Biochemical and Biophysical Research Communications*, 414(1), pp. 209–214.

Petersen, B.-S., Fredrich, B., Hoeppner, M. P., *et al.* 2017. Opportunities and challenges of whole-genome and -exome sequencing, *BMC Genetics*, 18(1), p. 14.

Popek, S. and Tsikitis, V. L. 2011. Colorectal Cancer : A Review, *Contemporary Oncology*, November.

Preta, G., de Klark, R., Gavioli, R., *et al.* 2010. The Enigma of Tripeptidyl-Peptidase II: Dual Roles in Housekeeping and Stress, *Journal of Oncology*, 2010, pp. 1–10.

Preta, G., Marescotti, D., Fortini, C., *et al.* 2008. Inhibition of Serine-Peptidase Activity Enhances the Generation of a Survivin-Derived HLA-A2-Presented CTL Epitope in Colon-Carcinoma Cells, *Scandinavian Journal*

of Immunology, 68(6), pp. 579–588.

Ramensky, V., Bork, P. and Sunyaev, S. 2002. Human non-synonymous SNPs: server and survey., *Nucleic acids research*, 30(17), pp. 3894–900.

Rockel, B., Kopec, K. O., Lupas, A. N., *et al.* 2012. Structure and function of tripeptidyl peptidase II, a giant cytosolic protease, *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1824(1), pp. 237–245.

Roper, J. and Hung, K. E. 2013. Molecular Mechanisms of Colorectal Carcinogenesis, in *Molecular Pathogenesis of Colorectal Cancer*. New York, NY: Springer New York, pp. 25–65.

Saramago, A. 2014. *Estudo de mutações germinais em genes de suscetibilidade para o Cancro do Colón e Reto Familiar do tipo X*. Faculdade de Ciências e Tecnologia - Universidade Nova de Lisboa.

Schmittgen, T. D. and Livak, K. J. 2008. Analyzing real-time PCR data by the comparative C(T) method., *Nature protocols*, 3(6), pp. 1101–8.

Schönegge, A.-M., Villa, E., Förster, F., *et al.* 2012. The Structure of Human Tripeptidyl Peptidase II as Determined by a Hybrid Approach, *Structure*, 20(4), pp. 593–603.

Schulz, E., Klampfl, P., Holzapfel, S., *et al.* 2014. Germline variants in the SEMA4A gene predispose to familial colorectal cancer type X., *Nature communications*. Nature Publishing Group, 5, p. 5191.

Schwarz, J. M., Cooper, D. N., Schuelke, M., *et al.* 2014. MutationTaster2: mutation prediction for the deep-sequencing age, *Nature Methods*, 11(4), pp. 361–362.

Shen, L., Toyota, M., Kondo, Y., *et al.* 2007. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer, *Proceedings of the National Academy of Sciences*, 104(47), pp. 18654–18659.

Song, S., Kang, M., Yoo, N., *et al.* 2010. Mutational analysis of mononucleotide repeats in dual specificity tyrosine phosphatase genes in gastric and colon carcinomas with microsatellite instability, *APMIS*, 118(5), pp. 389–393.

Stoffel, E. M. and Boland, C. R. 2015. Genetics and Genetic Testing in Hereditary Colorectal Cancer, *Gastroenterology*, 149(5), p. 1191–1203.e2.

Sundaramoorthy, P., Sim, J. J., Jang, Y.-S., *et al.* 2015. Modulation of intracellular calcium levels by calcium lactate affects colon cancer cell motility through calcium-dependent calpain., *PloS one*. Public Library of Science, 10(1), p. e0116984.

Sveen, A., Kilpinen, S., Ruusulehto, A., *et al.* 2016. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes, *Oncogene*. Nature Publishing Group, 35(19), pp. 2413–2427.

Taguchi-Atarashi, N., Hamasaki, M., Matsunaga, K., *et al.* 2010. Modulation of Local PtdIns3P Levels by the PI Phosphatase MTMR3 Regulates Constitutive Autophagy, *Traffic*. Blackwell Publishing Ltd, 11(4), pp. 468–478.

Trivedi, B. P. 2012. Neuroscience: Hardwired for taste, *Nature*, 486(7403), pp. S7–S9.

Umar, A., Risinger, J. I., Hawk, E. T., *et al.* 2004. Testing guidelines for hereditary non-polyposis colorectal cancer, *Nature Reviews Cancer*, 4(2), pp. 153–158.

Vasen, H. F., Watson, P., Mecklin, J. P., *et al.* 1999. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC., *Gastroenterology*, 116(6), pp. 1453–6.

Watson, J., Porter, S. and DeCourcy, K. 2012. *Sequencing and Bioinformatics Module - Instruction Manual*.

Wauson, E. M., Zaganjor, E. and Cobb, M. H. 2013. Amino acid regulation of autophagy through the GPCR TAS1R1-TAS1R3., *Autophagy*. Taylor & Francis, 9(3), pp. 418–9.

Wauson, E. M., Zaganjor, E., Lee, A.-Y., *et al.* 2012. The G Protein-Coupled Taste Receptor T1R1/T1R3 Regulates mTORC1 and Autophagy, *Molecular Cell*, 47(6), pp. 851–862.

Wells, K. and Wise, P. E. 2017. Hereditary Colorectal Cancer Syndromes, *Surgical Clinics of North America*, 97(3), pp. 605–625.

Wiemhoefer, A., Stargardt, A., van der Linden, W. A., *et al.* 2015. Tripeptidyl Peptidase II Mediates Levels

of Nuclear Phosphorylated ERK1 and ERK2., *Molecular & cellular proteomics : MCP*. American Society for Biochemistry and Molecular Biology, 14(8), pp. 2177–93.

Worthley, D.-L., Worthley, D.-L., Leggett, B.-A., *et al.* 2010. Colorectal cancer: molecular features and clinical opportunities., *The Clinical biochemist. Reviews / Australian Association of Clinical Biochemists*, 31(2), pp. 31–38.

Zhao, R., Qi, Y. and Zhao, Z. J. 2000. FYVE-DSP1, a Dual-Specificity Protein Phosphatase Containing an FYVE Domain, *Biochemical and Biophysical Research Communications*, 270(1), pp. 222–229.

Zheng, B., Yu, X. and Chai, R. 2014. Myotubularin-related phosphatase 3 promotes growth of colorectal cancer cells., *TheScientificWorldJournal*. Hindawi, 2014, p. 703804.

Zheng, H.-Y., Zhang, X.-Y., Wang, X.-F., *et al.* 2012. Autophagy enhances the aggressiveness of human colorectal cancer cells and their ability to adapt to apoptotic stimulus., *Cancer biology & medicine*, 9(2), pp. 105–10.

7. APPENDICES

Appendix A – Primers used in this study

Table A.1 - Sequence and expected fragment size of the primers used for segregation analysis of the variants identified by WES in L56 family (expected fragment size was obtained from the primer-BLAST software (primer-BLAST: https://www.ncbi.nlm.nih.gov/tools/primer-blast).

Output	Analysis	Gene	Identified variant	Primer sequence (5' → 3')		Expected size (bp)
3	1	*ZNF717*	c.2381G>A; p.Cys794Tyr	Frw	CACACAGGTGAGAAGCCATTTG	309
				Rev	GTGATTTTGAGGGAACACATAGCC	
		RNF213	c.8084C>T; p.Ala2695Val	Frw	TCTCTCCAAGTCCAGCGTCA	194
				Rev	CCTGTGCCCGTGTTATTTCA	
	3	*CACNA1S*	c.3607G>A; p.Asp1203Asn	Frw	AGCAAAGGGGAGAGGAGTGTT	448
				Rev	TGCCCAGCAGGATGATAGTTC	
4	1	*GIMAP1**	c.908T>G; p.Val303Gly	Frw	TCCAACGAGGTGTATGAGC	559
				Rev	TGCGTTTGTCCCCACTACA	

* Primers were designed prior to this study

Table A.2 - Sequence and expected fragment size of the primers used for segregation analysis of the MTMR3 and TAS1R1 variants identified by WES in the FFPE tissue sample of patient CAs1555 (expected fragment size was obtained from the primer-BLAST software (primer-BLAST: https://www.ncbi.nlm.nih.gov/tools/primer-blast).

Gene	Variant	Primer sequence (5' → 3')		Expected size (bp)
MTMR3	c.1933C>T; p.Arg645Trp	Frw	ACAACACAGTGCCTCTGGCCA	129
		Rev	TGTCGGCAGAAAGGGGATCCT	
TAS1R1	c.1582C>T; p.Leu528Phe	Frw	TTACGGGTTTCCATCACTG	82
		Rev	ATTGCCCACTCACCACTCTT	

Table A.3 – Primer sequences and expected size of the fragments used for amplification of the various exons of the TAS1R1 gene (expected fragment size was obtained from the primer-BLAST software (primer-BLAST: https://www.ncbi.nlm.nih.gov/tools/primer-blast).

Gene	Amplified exon(s)	Primer sequence (5' → 3')		Expected size (bp)
TAS1R1	1	Frw	TTCTGAAACCTCGCAACACC	665
		Rev	GGACAGTTCAAGGGGCAAAG	
	2	Frw	GAGCCCTTGTTGAAGTTGTGG	600
		Rev	GGGACCCCTGACTTGTGACTAA	
	3	Frw	GAAATGGCTGAACGGGACCT	886
		Rev	CCCCCTTGGCATTTAGAGTG	
	4	Frw	CCTAAGAAGGAAGGGACGAGAC	402
		Rev	CTAGGTTGCCCATAAGCCAGT	
	5	Frw	GTAACTGGCTTATGGGCAACC	579
		Rev	ATTCTGTCTGGTAGCAGGTGTCC	
	6	Frw	TTCCTACTCTGCTCATCTGGCT	1198
		Rev	AATCCGTTACACCTCACACTCTC	

Table A.4 – Primer used for TAS1R1 sequencing that differs from the ones described above for exon amplification.

Gene	Primer at exon:	Primer sequence (5' → 3')	
TAS1R1	6	Frw	CTCTTTGCCCTTGGTTTCAC

Table A.5 – Primer sequence and expected fragment size of the external primers used for cDNA and DNA amplification of the MTMR3, DUSP12 and LGR6 candidate genes, and respective internal sequencing primers used (expected fragment size was obtained from the primer-BLAST software (primer-BLAST: https://www.ncbi.nlm.nih.gov/tools/primer-blast).

Gene	Fragment	Amplified exons	Transcript(s) (Ensembl ID)	External Primers		Expected size (bp)	Internal sequencing primers for each fragment	
				Primer sequence (5' → 3')			Primer sequence (5' → 3')	
MTMR3	A	2 to 10	All except: ENST00000406629 and ENST00000445401	Frw	ACTTCCTTGTGAAACCTCCT	908	-	-
				Rev	TACTGACTGTACCAGATGCTCA		-	-
	B	2 to 10	ENST00000406629	Frw	TGGATGAAGAGACTCGGCACA	1072 to 1190	Frw	GAAAGTGTTGAATGCCGAG
				Rev	CCTCCGAATAGAATGAATGTTTG		Rev	TCTTCTATTTTAGCAGGTGGTC
	C	2 to 6	ENST00000445401	Frw	GACTTCACCATAAGGGAAAGAA	462	-	-
				Rev	TCTTCTATTTTAGCAGGTGGTC		-	-
	D	9 to 15	All except: ENST00000406629 and ENST00000445401	Frw	AGTGTATCAAGTTTCAGGTCC	918	Frw	AGAGAGTTTAGCCATCCAA
				Rev	AAACAGGCAGGAATAGGTA		Rev	CCTCCGAATAGAATGAATGTTTG
	E	14 to 17	All except: ENST00000406629 and ENST00000445401	Frw	GTGGAAATGGAGTGGCTGGAT	1011	Frw	ATGTGCGTAACCTGATGCTGTG
				Rev	GACCTGGAGAGAACTGAGGAGAA		Rev	ACAGGCTGTGGTCAGATTGTCG
	F	17 to 20	All except: ENST00000406629 and ENST00000445401	Frw	GAACTGGGTGATGCTGCTCTGA	1463 to 1574	Frw	TGTAAAGAGGGGCTTGTGTGC
				Rev	GAGGAGGATGGAAAAGGTAGGGA		Rev	TCAGTGAGTCCTTGCTCCTACCA
DUSP12	A (for DNA)	1	DNA amplification of exon 1 of all the transcripts	Frw	CCCTCAGCGATAAGTCCAGATTC	339	-	-
				Rev	TCCTCCGAGTCCACTGTTAGCA		-	-
	B *	1 to 6	All	Frw	CAGCCAGGATTGTATTTCGGT	1080 to 1093	Frw	AAGGTTACAGAGAAGTATCCAG
				Rev	GCATAGTTTCCAGGTTACATCAAG		Rev	GAGAACCTCATCTTTCAATCC
LGR6	A *	1 to 15	ENST00000367278	Frw	CATCATGCTGTCTGCCGACT	1204	Frw	TCCATAACAACAACATCAAGGC
				Rev	CATCAAGCCCCCAAGTCCA		Rev	CATAAAAGTGTATCGTCTGTAGC
	B *	1 to 15	ENST00000255432	Frw	GGGAAGACCAAGGTTGACACTT	1180	Frw	TCCATAACAACAACATCAAGGC
				Rev	CATCAAGCCCCCAAGTCCA		Rev	CATAAAAGTGTATCGTCTGTAGC
	C *	2 to 13	ENST00000439764	Frw	GGACCTCAGCATGAACAACCTC	838	Frw	TCCATAACAACAACATCAAGGC
				Rev	CATCAAGCCCCCAAGTCCA		Rev	CATAAAAGTGTATCGTCTGTAGC
	D *	2 to 9	ENST00000423542	Frw	GGACCTCAGCATGAACAACCTC	632	-	-
				Rev	CCTCCAATTTCTGACACCTGTG		-	-
	E (for DNA)	1	DNA amplification of exon 1 of transcript ENST00000439764	Frw	GCATACGAGAGAAGCGGCAGAAC	230	-	-
				Rev	TGCGGAGGACCTGTGGAACTTT		-	-
	F *	14 to 18	ENST00000367278 and ENST00000255432	Frw	CAAGCTGGACCTGACAGACAACCA	1575	Frw	TACTTGGGTCGGAGGCATCGGT
				Rev	AAGCCTCCACCTGCTGGCGTAGAT		Rev	AGGAGTTCATCATCACCAGGGC

* Pcr conditions were not optimized

Table A.6 – Primers used for amplification of the MTMR3 variant p.Arg645Trp.

Gene	Identified variant	Primer sequence (5' → 3')		Expected size (bp)
MTMR3	c.1933C>T; p.Arg645Trp	Frw	CCAAGATTTTCATTTCCCCC	406
		Rev	TTCCTCCTGCTCTCCTTTTC	

All primers were designed prior to this study

Table A.7 - Primers used for TPP2 gene expression of the normal and the alternative 13a transcript.

Gene	Primer at exon:	Primer sequence (5' → 3')		Expected size (bp)
TPP2	13/13a	Frw	CGGAGAACACAGGTATAAGACAG	161
	14 *	Rev	CAAATGGCTGGGACACTGAAC	
	13/14 *	Frw	CCGGAGAACACAGAAAACTCTG	93
	14 *	Rev	CAAATGGCTGGGACACTGAAC	
GAPDH	7 *	Frw	TGCACCACCAACTGCTTAGC	87
	7/8 *	Rev	GGCATGGACTGTGGTCATGAG	

* Primers were designed prior to this study

Table A.8 - Primers used for TPP2 splicing analysis (expected fragment size was obtained from the primer-BLAST software (primer-BLAST: https://www.ncbi.nlm.nih.gov/tools/primer-blast).

Gene	Amplified exons	Primer sequence (5' → 3')		Expected size (bp)
TPP2	2 to 29	Frw	GAAGTGGCGATGTGAATAC	3541
		Rev	GGGAGCCAGTTTTCAGTAA	
	13 to 14	Frw	ATAAAGCCTATGACTACCTCG	245
		Rev	CAAATGGCTGGGACACTGAAC	
	23 to 27	Frw	CCAGCCATTCTTTGTTACT	565
		Rev	TACGAGGGTGGATTTTTGT	

All primers were designed prior to this study

Table A.9 - Primers used in PCR amplification for PTT analysis of TPP2 gene and characteristics of the amplified fragment.

Gene	Amplified exons	Size (kb)	Polypeptide molecular weight (kilodaltons (kDa))	Primer at exon	Primer sequence (5' → 3')
TPP2	1 to 19	2.3	95*	1	[T7]-GCTGCGACTGAGGAGCCCTT
				19	CGCAGTGTTTGGACCCAGTTCTTC

*The polypeptide molecular weight was calculated based on the formula used in Calc Tool database (http://www.calctool.org/CALC/prof/bio/protein_size);
[T7] represents the PTT-specific forward primer sequence, containing a T7 RNA polymerase promoter sequence, a spacer, and the Kozak sequence

Appendix B - PCR and qPCR conditions used in this study

Table B.1 - PCR conditions used for the amplification of the variants selected through bioinformatic analysis, obtained by WES analysis, for the segregation study with the disease in the L56 family.

Output	Analysis	Gene	Identified variant	Temperature (°C)	$MgCl_2$ (µL)
3	1	*ZNF717*	c.2381G>A; p.Cys794Tyr	66	1.5
		RNF213	c.8084C>T; p.Ala2695Val	59	0.5
	3	*CACNA1S*	c.3607G>A; p.Asp1203Asn	63	1
4	1	*GIMAP1*	c.908T>G; p.Val303Gly	60	0.25

All reactions were performed using the Biotaq™ kit (*Bioline*)

Table B.2 - PCR conditions used for amplification of the MTMR3 and TAS1R1 variants identified by WES in the FFPE tissue sample of patient CAs1555.

Gene	Variant	Temperature (°C)	$MgCl_2$ (µL)
MTMR3	c.1933C>T; p.Arg645Trp	68	1
TAS1R1	c.1582C>T; p.Leu528Phe	60	2

All reactions were performed using the AmpliTaq Gold® (*Thermofisher*) kit

Table B.3 – PCR conditions used for amplification of TAS1R1 exons.

Gene	Exon	Temperature (°C)	$MgCl_2$ (µL)
TAS1R1	1	60.5	0.25
	2	65	0.25
	3	61.5	0.25
	4	64.5	0.75
	5	66	1
	6	66.5	0.5

All reactions were performed using the Biotaq™ kit (*Bioline*)

Table B.4 – PCR conditions used for amplification of MTMR3, DUSP12 and LGR6 fragments.

Gene	Fragment	Temperature (°C)	$MgCl_2$ (µL)
MTMR3	A	62	1.5
	B*¥	51.5	0.35
	C	62.5	0.5
	D*	59.5	0.75
	E	64.5	1.25
	F	70	1
DUSP12	A	68	1
LGR6	E	69	1

All reactions were performed using the Biotaq™ kit (*Bioline*) except for fragments marked with *, where it was used the GC-RICH PCR System kit (*Roche*); ¥ - addition of 0.75µL of GC-RICH resolution solution.

Table B.5 – PCR conditions used for amplification of the MTMR3 variant p.Arg645Trp.

Gene	Variant	Temperature (°C)	$MgCl_2$ (µL)
MTMR3	c.1933C>T; p.Arg645Trp	60.5	0.5

The reaction was performed using the Biotaq™ kit (*Bioline*)

Table B.6 – qPCR conditions used for TPP2 gene expression analysis.

Gene	Fragment	Primer concentration	Sample concentration	Efficiency
TPP2	**Wt (13/14)**	5 pmol/µL	5 ng/µL	2.08
	13a			2.02
GAPDH	***GAPDH***			2

Table B.7 – qPCR conditions used for TPP2 splicing analysis.

Gene	Exons	Temperature (°C)	$MgCl_2$ (µL)
TPP2	2 to 29 *	53	1.25**
	13 to 14	60	1
	23 to 27	52	1

All reactions were performed using the Biotaq™ kit (*Bioline*) except for fragments marked with *, where it was used the Expand Long Template PCR System (*Roche*) kit; ** $MgCl_2$ added from buffer 3 of the pcr kit.

Table B.8 – PCR conditions used for TPP2 gene amplification for PTT analysis.

Gene	Exons	Temperature (°C)	$MgCl_2$ (µL)
TPP2	1 to 19	65	1.25*

The reaction was performed using the Expand Long Template PCR System (*Roche*) kit; * $MgCl_2$ added from buffer 2 of the pcr kit.

Appendix C - PCR amplification programs used in this study

Table C.1 - PCR amplification program used with the reagents from the BiotaqTM kit (Bioline). PCR amplification was performed in a UNO96 thermocycler (VWR).

Step	Temperature (°C)	Time	Cycles
Initial denaturation	95	5 min	1
Denaturation	94	50 sec	35
Annealing	Variable	30 sec	
Elongation	72	70 sec	
Final elongation	72	7 min	1
Pause	15	∞	∞

Table C.2 - PCR amplification program used with the reagents from the AmpliTaq Gold® (Thermofisher) kit. PCR amplification was performed in a Veriti® Thermal Cycler (Applied Biosystems).

Step	Temperature (°C)	Time	Cycles
Initial denaturation	95	10 min	1
Denaturation	94	50 sec	40
Annealing	Variable	30 sec	
Elongation	70	1 min	
Final elongation	70	5 min	1
Pause	15	∞	∞

Table C.3 - PCR amplification program used with the reagents from the GC-RICH PCR System (Roche) kit. PCR amplification was performed in a Veriti® Thermal Cycler (Applied Biosystems).

Step	Temperature (°C)	Time	Cycles
Initial denaturation	95	3 min	1
Denaturation	95	30 sec	10
Annealing	Variable	30 sec	
Elongation	72	45 sec	
Denaturation	95	30 sec	25
Annealing	Variable	30 sec	
Elongation	72	45 sec + 5 sec cycle elongation for each successive cycle	
Final elongation	72	7 min	1
Pause	15	∞	∞

Table C.4 - PCR amplification program used with the reagents from the Expand Long Template PCR System (Roche) kit. PCR amplification was performed in a Veriti® Thermal Cycler (Applied Biosystems).

Step	Temperature (°C)	Time	Cycles
Initial denaturation	94	2 min	1
Denaturation	94	10 sec	10
Annealing	Variable	30 sec	
Elongation	68	2 min	
Denaturation	94	30 sec	25
Annealing	Variable	30 sec	
Elongation	68	2 min + 20 sec cycle elongation for each successive cycle	
Final elongation	72	7 min	1
Pause	15	∞	∞

Table C.5 – Program used for sequencing reaction of the studied genes.

Step	Temperature (°C)	Time	Cycles
Initial denaturation	96	5 min	1
Denaturation	95	10 sec	25
Annealing	59	5 sec	
Elongation	60	4 min	
Pause	4	∞	∞

Table C.6 – Reverse transcription program used for cDNA synthesis.

Step	Temperature (°C)	Time	Cycles
Denaturation and binding of hexamers	70	5 min	1
Pause	10	∞	
cDNA synthesis	42	60 min	
Elongation	70	15 min	
Pause	4	∞	∞

Table C.7 – qPCR program used for TPP2 gene expression analysis.

Step	Temperature (°C)	Time	Cycles
Incubation	50	2 min	1
Polymerase activation	95	10 min	
Denaturation	95	15 sec	40
Hybridization and elongation	60	1 min	

Appendix D - Preparation of the reagents used in agarose gel electrophoresis

1. **Preparation of the electrophoresis buffer (TBE 1x):**

   To obtain 1x TBE, a 10x TBE solution (0.89M Tris Borate pH8.3 + 20mM Na2 EDTA, *National Diagnostics*) was diluted to a final volume of 2L in ddH_2O.

2. **Preparation of 2% (w/v) agarose gel:**

   I. Weigh 5g of agarose (Seaken® LE Agarose, *Lonza*) in a 500ml Erlenmeyer;
   II. Add 250ml of TBE 1x buffer;
   III. Dissolve the solution in the microwave;
   IV. Add 12.5µL of ethidium bromide (10mg/ml, *MP biomedicals*) and stir for homogenization;
   V. Place the solution in a polymerization base with 4 well molds and allow to cool until the gel solidifies.

3. **Preparation of 0.8 and 1.2% (w/v) agarose gel:**

   I. Weigh 1.2g for 0.8% or 1.8g for 1.2% of agarose (Seaken® LE Agarose, *Lonza*) in a 500ml Erlenmeyer;
   II. Add 150ml of TBE 1x buffer;
   III. Dissolve the solution in the microwave;
   IV. Add 7.5µL of ethidium bromide (10mg/ml, *MP biomedicals*) and stir for homogenization;
   V. Place the solution in a polymerization base with 1 well mold and allow to cool until the gel solidifies.

4. **Preparation of Orange G 5x and 1x:**

   Orange G 5x: Prepared from 12ml Ficoll (*Sigma*), 125g of Orange G (*Sigma*) and 50ml of ddH_2O, under stirring. Store at -20 °C.

   Orange G 1x: Diluted from Orange G 5x to a final volume of 1ml in ddH_2O. Store at 4ºC.

5. **Preparation of the molecular weight marker GeneRuler 50bp DNA Ladder (*Thermofisher Scientific*):**

   Add 50µL of GeneRuler 50bp DNA Ladder (*Thermofisher Scientific*), 250µL of Orange G 5x and 700µL of ddH_2O.

6. **Preparation of the molecular weight markers Lambda/HindIII DNA Ladder (*Fermentas*):**

   Add 20µL of Lambda/HindIII DNA Ladder (*Fermentas*), 100µL of Orange G 5x and 280µL of ddH_2O.

7. **Preparation of the molecular weight markers Lambda/HindIII DNA Ladder (*Fermentas*):**

   Add 100µL of 1 kb DNA Ladder (*Promega*), 100µL of Orange G 5x and 200µL of ddH_2O.

Appendix E - Precipitation and purification protocol of DNA Ethanol / EDTA / Sodium acetate - BigDye® Terminator v1.1 Cycle Sequencing Kit (*Applied Biosystems*) – after sequencing reaction

1. **Protocol:**

I. In a 2ml eppendorf-like tube add 2µL of EDTA (125mM), 2µL sodium acetate (3M) and 50µL absolute ethanol, for each reaction;

II. Vortex and spin-down;

III. Transfer 54µL of the supernatant to new 1.5ml eppendorf-like tube, add the total volume of the sequencing reaction and homogenize;

IV. Incubate for 15 minutes at room temperature;

V Centrifuge at 14000rpm for 30minutes at 4°C;

VI. Remove the supernatant completely with a micropipette;

VII. Add 100µL of 70% (v/v) ethanol and slightly vortex;

VIII. Centrifuge at 14000rpm for 15minutes at 4°C;

IX. Remove the supernatant completely with a micropipette;

X. Dry the pellet in a dry bath at 37°C for about 10 minutes;

XI. Store the pellet at 4ºC.

2. **Solutions:**

I. EDTA (125 mM, pH 8):

Dissolve 4.65g of EDTA in ddH_2O and make up the volume to 100 mL.

II. Sodium acetate (3M, pH 4.5):

Dissolve 24.8g of sodium acetate in ddH_2O. Adjust the pH with acetic acid and make up the volume to 100 mL.

III. Ethanol 70%:

Add 30mL of ddH_2O to 70mL of absolute ethanol and homogenize.

Appendix F - Preparation of the reagents used in SDS-PAGE

1. **Stop Solution:**

   For 100mL: Add 8µL of glycerol (*SIGMA-ALDRICH®*) to 2.4g of SDS (*SIGMA-ALDRICH®*) (a mask is required during the weighing of this reagent) to 0.004g of bromophenol blue (*SIGMA-ALDRICH®*) and make up the final volume with ddH_2O.

   Just before adding the stop solution to the samples 25µL of β-mercaptoethanol (*SIGMA-ALDRICH®*) are added per 1000µL of the stop solution.

2. **Tris-glycine/SDS buffer:**

   Tris-glycine/SDS buffer 5x: Add 30g of trizma-base (*SIGMA-ALDRICH®*), 144g of glycine (*SIGMA-ALDRICH®*) and 5g of SDS (*SIGMA-ALDRICH®*) (a mask is required during the weighing of this reagent) and make up to a final volume of 1L with ddH_2O (solution prepared under agitation).

   Tris-glycine/SDS buffer 1x: Diluted from Tris-glycine/SDS buffer 5x to a final volume of 1L in ddH_2O.

3. **Resolving buffer (pH 8.9):**

   Add 72.6g of trizma-base (*SIGMA-ALDRICH®*) to 200mL of ddH_2O and adjust pH with a HCl solution.

4. **Stacking buffer (pH 6.7):**

   Add 11.4g of trizma-base (*SIGMA-ALDRICH®*) to 200mL of ddH_2O and adjust pH with a H_3PO_4 solution.

5. **Sealing gel:**

   I. Prepare a solution in the following order, with: 1.5mL of acrylamide-bisacrylamide 30% (*GRISP*), 400µL of the resolving buffer solution, 34.1µL of SDS 10% (*AppliChem*), 1.36mL of ddH_2O, 50µL of an APS 10% solution and 5µL of TEMED (*SIGMA-ALDRICH®*);

   II. Gently mix the solution and add to the polymerization base right away;

   III. Wait at least 30 minutes before adding the resolving gel.

6. **Resolving gel:**

   I. Prepare a solution in the following order, with: 9.1mL of acrylamide-bisacrylamide 30% (*GRISP*), 2.5mL of the resolving buffer solution, 204.6µL of SDS 10% (*AppliChem*), 8.15mL of ddH_2O, 138.6µL of an APS 10% solution and 13.8µL of TEMED (*SIGMA-ALDRICH®*);

   II. Gently mix the solution and add to the polymerization base right away;

   III. Wait at least 30 minutes before adding the stacking gel.

7. Stacking gel:

I. Prepare a solution in the following order, with: 1.12mL of acrylamide-bisacrylamide 30% (*GRISP*), 0.9mL of the stacking buffer solution, 75µL of SDS 10% (*AppliChem*), 5.4mL of ddH_2O, 118µL of an APS 10% solution and 7.9µL of TEMED (*SIGMA-ALDRICH*®);

II. Gently mix the solution, add to the polymerization base and add the mold of the wells right away;

III. Wait at least 1 hour before taking the mold of the wells off.

8. Fixative solution:

Add 30mL of acetic acid (*VWR*), 60mL of methanol (*Merck*) and 210mL of ddH_2O.

Appendix G – Results from the mutation analysis performed on the *MTMR3* gene for 34 index patients from FCCTX and FCCTX-like families

Sample	*MTMR3*					
	Fragment A	**Fragment B**	**Fragment C**	**Fragment D**	**Fragment E**	**Fragment F**
L5	Wt	Splicing	Wt	Splicing	Wt	Wt
L295	Wt	Splicing	Wt	Splicing	Wt	Wt
L1368	Wt	Splicing	Wt	Splicing	Wt	Wt
L1908	Wt	Splicing	Wt	Splicing	Wt	Wt
L2161	Wt	Splicing	Wt	Splicing	Wt	Het (4)
L287	Wt	Splicing	Wt	Splicing	Wt	Wt
CAs1964s	Wt	Splicing	Wt	Splicing	Wt	Wt
L2127	Wt	Splicing	Wt	Splicing	Het (1)	Wt
L2128	Wt	Splicing	Wt	Splicing	Het (2)	Wt
CAs1974s	Wt	Splicing	Wt	Splicing	Wt	-
L1667	Wt	Wt	Wt	Wt	Wt	Wt
2219	-	Splicing	-	Splicing	Het (3)	Wt
L1513	-	Splicing	-	Splicing	Wt	Wt
L1501	-	Splicing	-	Splicing	Wt	Wt
L1490	-	Splicing	-	Splicing	Wt	Wt
L1246	-	Splicing	-	Splicing	Wt	-
L1310	-	Splicing	-	Splicing	Wt	-
L2208	-	Splicing	-	Splicing	Wt	-
E436	-	Wt	-	Wt	-	-
L2121	-	-	-	-	Wt	-
L2136	-	-	-	-	Wt	-
L2141	-	-	-	-	Wt	-
L2153	-	-	-	-	Wt	-
L2180	-	-	-	-	Wt	-
L2185	-	-	-	-	Wt	-
L2195	-	-	-	-	Wt	-
L2156	-	-	-	-	Wt	-
L782	-	-	-	-	Wt	-
L2186	-	-	-	-	Wt	-
L108	-	-	-	-	Wt	-
L2410	-	-	-	-	Wt	-
L1291	-	-	-	-	Wt	-
L1937	-	-	-	-	Wt	-
L1902	-	-	-	-	Wt	-

Wt - Wild type (with no alteration); Het - Heterozygotic; (1) c.1933C>T (global frequency: 0.02%); (2) c.1811C>A (no reported frequency); (3) c.2131G>A (no reported frequency); (4) c.2879A>G (global frequency: 5%); "-" - fragment not amplified.

Appendix H - Results from the mutation analysis performed on the *TAS1R1* gene for 26 index patients from FCCTX and FCCTX-like families

Sample	*TAS1R1*					
	Exon 1	Exon 2	Exon 3	Exon 4	Exon 5	Exon 6
E436	Wt	Wt	Wt	Wt	Wt	Wt
L1038	Wt	Wt	Wt	Wt	Wt	Wt
L1053	Wt	Wt	Wt	Wt	Het (7)	Wt
L108	Wt	Wt	Het (2)	Wt	Wt	Wt
L1121	Wt	Het (1)	Het (3)	Wt	Het (7)	Wt
L1200	Wt	Wt	Wt	Wt	Wt	Wt
L1208	Wt	Wt	Wt	Wt	Het (7)	Wt
L1291	Wt	Wt	Wt	Homo (6)	Wt	Wt
L1310	Wt	Wt	Wt	Wt	Het (7)	Wt
L1424	Wt	Wt	Wt	Wt	Wt	Wt
L1445	Wt	Wt	Wt	Wt	Wt	Wt
L1473	Wt	Wt	Het (4)	Wt	Het (7)	Wt
L1484	Wt	Wt	Het (3)	Wt	Wt	Wt
L2208	Wt	Wt	Wt	Wt	Wt	Wt
L295	Wt	Wt	Wt	Wt	Wt	Wt
L451	Wt	Wt	Wt	Wt	Wt	Wt
L467	Wt	Wt	Wt	Wt	Wt	Wt
L48	Wt	Wt	Wt	Wt	Wt	Wt
L484	Wt	Wt	Wt	Wt	Wt	Wt
L499	Wt	Wt	Wt	Wt	Wt	Wt
L5	Wt	Wt	Wt	Wt	Wt	Wt
L701	Wt	Wt	Het (5)	Wt	Wt	Wt
L927	Wt	Wt	Het (3)	Wt	Wt	Wt
L956	Wt	Wt	Wt	Wt	Het (7)	Wt
L964	Wt	Wt	Wt	Wt	Wt	Wt
L984	Wt	Wt	Het (3)	Wt	Wt	Wt

Wt - Wild type (with no alteration); Het - Heterozygotic; Homo - Homozygotic; (1) c.329C>T (global frequency 1.2%); (2) c.1067 C>G (global frequency: 0.2%); (3) c.1114G>A (global frequency: 10%); (4) c.953T>C (global frequency: 0.001%); (5) c.1043C>T; (6) c.1261-90C>T(global frequency 0.1%); (6) c.1261-90C>T (global frequency: 1.7%); (7) c.1594+41G>A (global frequency: 12%).

Appendix I - Results from the mutation analysis performed on *DUSP12 and LGR6* genes for 11 index patients from FCCTX and FCCTX-like families

Sample	*DUSP12*		*LGR6*					
	Fragment A (DNA)	Fragment B	Fragment A	Fragment B	Fragment C	Fragment D	Fragment E (DNA)	Fragment F
L5	Wt	-	-	-	-	-	Wt	-
L295	Wt	-	-	-	-	-	Wt	-
L1368	Wt	-	-	-	-	-	Wt	-
L1908	Wt	-	-	-	-	-	Wt	-
L2161	Wt	-	-	-	-	-	Het (1)	-
L287	Wt	-	-	-	-	-	Wt	-
CAs1964s	Wt	-	-	-	-	-	Het (2)	-
L2127	Wt	-	-	-	-	-	Wt	-
L2128	Wt	-	-	-	-	-	Het (1)	-
CAs1974s	Wt	-	-	-	-	-	Wt	-
L1667	Wt	-	-	-	-	-	Wt	-

Wt - Wild type (with no alteration); Het - Heterozygotic; (1) c.213-11193del (global frequency: 9%); (2) c.213-11224C>T (global frequency: 2%); "-" - fragment not amplified.