Inês Fernandes Gomes

Molecular characterization of new candidate genes for familial thyroid cancer

Dissertação de Mestrado em Biologia Celular e Molecular,
orientada pelo Doutora Branca Cavaco e pela Professora Doutora Maria Carmen Alpoim,
apresentada ao Departamento de Ciências da Vida da Universidade de Coimbra.

Julho 2016

· U C ·

UNIVERSIDADE DE COIMBRA

Inês Fernandes Gomes

Molecular characterization of new candidate genes for familial thyroid cancer

Dissertação apresentada à Universidade de Coimbra para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biologia Celular e Molecular, realizada sob orientação da Doutora Branca Cavaco (Unidade de Investigação em Patobiologia Molecular (UIPM), Instituto Português de Oncologia de Lisboa Francisco Gentil) e da Professora Doutora Maria Carmen Alpoim (Departamento de Ciência da Vida, Universidade de Coimbra).

Setembro de 2016

· U C ·

UNIVERSIDADE DE COIMBRA

O presente trabalho foi realizado no grupo de Endocrinologia Molecular da Unidade de Investigação em Patobiologia Molecular (UIPM) do Instituto Português de Oncologia (IPO) de Lisboa, Francisco Gentil.

AGRADECIMENTOS

Quero começar por agradecer à minha orientadora, Doutora Branca Cavaco pela oportunidade de realizar este trabalho no grupo de Endocrinologia Molecular do IPO-Lisboa, pela sua disponibilidade, simpatia e ensinamentos que contribuíram para a minha evolução tanto a nível académico, como a nível pessoal. Ao Doutor Valeriano por todas as sugestões e entusiasmo demostrado no decorrer deste projeto.

Quero também dirigir um agradecimento muito especial às minhas colegas de grupo, com as quais tive o prazer de trabalhar durante este ano. À Margarida por toda ajuda e essencialmente paciência, sei que posso ter contribuído para alguns cabelos brancos. À Sara, Marta e Inês 1, pela amizade, preocupação, incentivo e ajuda, vocês foram incansáveis comigo e terão sempre um lugar muito especial no meu coração. À Joana, Rita, Ana Luísa, Ana Matias e Ana Pinto, sem vocês esta jornada não teria sido a mesma, obrigada por tudo. Tenho a certeza que não foi fácil para ninguém aturar os meus dramas, mas sei que vão ter tantas saudades minhas, como eu irei ter vossas!

Aos restantes membros da UIPM, que me fizeram sentir em casa neste último ano e que me ajudaram sempre que precisei. A vossa boa disposição e aquelas horas de almoço na vossa companhia tornaram os meus dias mais alegres.

Não posso deixar de dirigir uma palavra de carinho a todos os meus colegas de mestrado e amigos. Um obrigado muito especial às princesas de mestrado, vocês sabem quem são e o quão especial foi ter-vos conhecido, nunca pensei que se pudesse criar uma amizade tão especial em dois anos. E um grande obrigado também à Sara P., Carina e Lena, que já me acompanham a alguns anos e me têm demostrado o verdadeiro significado da amizade.

Gostaria também de agradecer à minha família, especialmente à minha prima/irmã mais velha Susete, por todo apoio e paciência ao longo destes anos. Sei que posso contar com todos vocês, hoje e sempre!

Por último e não menos importante, quero agradecer aos meus pais, pois sem eles nada disto seria possível. Obrigado por me apoiarem em tudo nesta vida, por todos os valores que me transmitiram e por estarem sempre tão presentes para mim, foi graças a vocês que me tornei na mulher que sou hoje.

A todos os que tornaram este trabalho possível, um muito OBRIGADA!

TABLES OF CONTENTS

LIST OF NOMENCLATURES, ABBREVIATIONS AND ACRONYMS

AKT– Protein kinase B

APC – Adenomatous Polyposis coli

ASR- Age-standardized

ATHL1– Acid Trehalase-like 1 (yeast)

AURK- Aurora kinase

B4GALNT4- Galactosaminyl tranferase 4

BLMH– Bleomycin hydrolase

bp– Base pairs

BRAF- B-Raf Proto-Oncogene, Serine/Threonine Kinase

CAMK- Calcium/calmodulin dependent protein kinase

CASP8AP2– Caspase 8 Associated Protein 2

CCDC6- Coiled-Coil Domain Containing 6

CDC42– Cell division control protein 42 homolog

CDKI- Cyclin-dependent kinase inhibitor

CDKN2A– Cyclin-Dependent Kinase Inhibitor 2A

CDKN2AIP– CDKN2A Interacting Protein

CDKN2B– Cyclin-Dependent Kinase Inhibitor 2A

CENPA- centromere protein A

Chr– Chromosome

cM– CentiMorgan

CNVs- Copy Number Variation

CPC- Chromosomal passenger complex

cPTC- classical papillary thyroid carcinoma

CTNNB1– Catenin beta 1

ddNTPs– dideoxyribonucleotide triphosphates

DICER1– Ribonuclease type III

DNA- Deoxyribonucleic acid

DNAH17– Dynein Axonemal Heavy Chain 17

DNAJC28– DnaJ Heat Shock Protein Family (Hsp40) Member C28

dNTPs– deoxyribonucleotide triphosphates

EGFR- Epidermal growth factor receptor

ERICH6B– Glutamate Rich 6B

ExAC- Exome Aggregation Consortium

FFPE – Formalin-fixed paraffin embedded

FNMTC- Familial non medullary thyroid carcinoma

FOXE1– Forkhead box E1

FPKM- Fragments per kilobase gene model and millions reads

fPTC/PRN- Papillary Thyroid Carcinoma and Papillary Renal Neoplasia locus

FTC– Follicular thyroid carcinoma

FTEN- Familial Thyroid Epitelial Neoplasia

fvPTC- follicular-variant of papillary thyroid carcinoma

GDP- Guanosine diphosphate

GTP- Guanosine triphosphate

GXYLT1– Glucoside Xylosyltransferase 1

HABP2– Hyaluronan binding protein 2

HHEX- Hematopoetically expressed homeobox

HIF1α– Hypoxia-inducible factor 1-alpha

HLA-DRB1– Major Histocompatibility Complex, Class II, DR Beta 1

HPA– Human protein atlas

HRAS- Harvey Rat Sarcoma Viral Oncogene Homolog

hTERT- Human Telomerase Reverse Transcriptase

IFITM10- Interferon induced transmembrane protein 10

IGV– Integrative genomics viewer

INCENP- Inner centromere protein antigens

Indels- Insertion or deletion of bases in the DNA

KCTD16– Potassium Channel Tetramerization Domain Containing 16

KRAS- Kirsten Rat Sarcoma Viral Oncogene Homolog

LFNG– Lunatic Fringe Homolog (Drosophila)

LILRB1– Leukocyte Immunoglobulin Like Receptor B1

LOD- Logarithm of odds

LOH– Loss of Heterozygosity

MAPK- Mitogen active protein kinase

$MgCl_2$ – Magnesium chloride

MNG– Multinodular goiter

mRNA– Messenger ribonucleic acid

MTC- Medullary thyroid carcinoma

MTCH2– Mitochondrial Carrier 2

MUC3A– Mucin 3A, Cell Surface Associated

MUC5AC- Mucin 5AC, Oligomeric Mucus/Gel-Forming

NCOA4- Nuclear Receptor Coactivator 4

NFKB- Nuclear factor kappa B

NGS– Next generation sequencing

NKX2.1– NK2 homeobox 1

NMTC- Non medullary thyroid carcinoma

NRAS- Neuroblastoma Rat Sarcome Viral Oncogene Homolog

PAX8- Paired Box 8

PCL- Phospholipase C

PCR– Polymerase chain reaction

PI3K- Phosphoinositide 3-kinase

PIK3CA- Phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha

PPARγ– Peroxisome proliferation-actived receptor type gamma

PRKAR1a– Protein Kinase CAMP-Dependent Type I Regulatory Subunit Alpha

PTC- Papillary thyroid cancer

PTEN– Phosphatase and Tensin Homolog

RAS- Rat Sarcoma Virus Homolog

RET- Rearranged during Transfection

RNA- Ribonucleic acid

RP1L1– Retinitis Pigmentosa 1-Like 1

RTK- Receptor tyrosine kinase

SBS– Sequencing by synthesis

SPRY4– Sprouty RTK Signaling Antagonist 4

SRGAP1– SLIT-ROBO Rho GTPase Activing Protein 1

SRRM2- Serine/Arginine Repetitive Matrix 2

STK33– Serine/Threonine Kinase 33

T_3– Triiodothyronine hormone

T_4– Thyroxine hormone

TARS2- Threonyl-TRNA Synthetase 2, Mitochondrial

TBC1D14– TBC1 Domain Family Member 14

TCO- Thyroid Carcinoma With Cell Oxyphilia

tcPTC- tall cell papillary thyroid carcinoma

TIGD4– Tigger Transposable Element Derived 4

TNF- Tumor necrosis factor

TP53– Tumor protein 53

TRAF2- Tumor necrosis factor receptor-associated factor 2

TRAPP- Transport protein particle

TREH- Trehalase

TRIO– Trio Rho Guanine Nucleotide Exchange Factor

TRK- Tropomyosin-related protein

TSH– Thyroid-Stimulating hormone

TTF1– Thyroid transcriptional factor-1

TTF2- Thyroid transcription factor-2

UV- Ultraviolet

VEGF- Vascular endothelial growth factor

WES- Whole Exome sequencing

WGS– Whole genome sequencing

WNT5A– Wingless-type MMTV integration site family, member 5A

WRN– Werner Syndrome RecQ Like Helicase

ZNF717– Zinc Finger Protein 717

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Thyroid cancer is the most common neoplasia of the endocrine system. The incidence of thyroid cancer is growing, with one of the fastest rates of increase among human cancers. Thyroid carcinomas can occur either as sporadic or familial forms. The familial forms of non medullary thyroid carcinoma (FNMTC) represent about 5% of non medullary thyroid neoplasms, and the molecular mechanisms involved in its aetiology are poorly understood. The few known FNMTC susceptibility genes only account for a very small fraction of familial thyroid cancer genetic susceptibility.

To identify new candidate genes that may confer susceptibility for FNMTC, whole exome sequencing (WES) was performed in leucocyte DNAs from patients with thyroid carcinomas, from a highly informative FNMTC family.

Bioinformatic analyses were undertaken in order to identify, filter and select the genetic variants shared by the different patients, which were subsequently validated by Sanger sequencing.

In order to select the most likely pathogenic variants, that could be associated with FNMTC aetiology in this family, several studies were performed, which included: segregation studies, *in silico* functional characterization and analysis of gene expression (mRNA and protein), and gene mapping. Taking into account all these data, several candidate genes were identified and, among these, the genes *SPRY4*, *STK33*, *TBC1D14* and *TRIO* were considered the most relevant to further proceed to *in vitro* studies for functional characterization.

Hopefully, the results obtained in the *in vitro* studies will lead to the identification of a novel susceptibility gene for FNMTC in this family, which may also be relevant in the aetiology of this disease in other families, as well as to a better understanding of the cellular and molecular mechanisms involved in cancer development.

Keywords: Thyroid, familial non medullary thyroid carcinoma (FNMTC), susceptibility gene, whole exome sequencing (WES).

RESUMO

O cancro da tiróide é a neoplasia mais frequente dos tecidos endócrinos e a sua incidência tem apresentado uma das mais rápidas taxas de aumento entre os tumores mais frequentes em humanos. Os carcinomas da tiróide podem ocorrer tanto esporadicamente como sob formas familiares. As formas familiares de carcinomas não medulares da tiróide (FNMTC) representam cerca de 5% de todas as neoplasias não medulares da tiróide. Os mecanismos moleculares envolvidos na etiologia do FNMTC são ainda pouco conhecidos. Os genes de susceptibilidade já identificados apenas se encontram envolvidos numa pequena fração de famílias com carcinomas não medulares da tiróide.

Com o objetivo de identificar novos genes candidatos que possam conferir susceptibilidade para o FNMTC, foi sequenciado todo o exoma (WES) em DNA de leucócitos de doentes com carcinomas da tiróide, provenientes de uma família muito informativa com FNMTC.

Foram realizadas análises bioinformáticas para identificar, filtrar e selecionar as variantes genéticas partilhadas pelos diferentes doentes, que foram posteriormente validadas por sequenciação de Sanger.

Para selecionar as variantes com maior probabilidade de serem patogénicas e estarem associadas à etiologia do FNMTC nesta família, foram realizados diferentes estudos, que incluíram: estudos de segregação, caracterização funcional e análise da expressão génica (mRNA e proteica) *in silico*, e mapeamento genético. Tendo em consideração todos os dados obtidos nestas análises, vários genes candidatos foram identificados, entre os quais, os genes *SPRY4, STK33, TBC1D14* e *TRIO* foram considerados os mais relevantes para prosseguir para os estudos funcionais *in vitro.*

Espera-se que os resultados dos estudos *in vitro,* conduzam à identificação de um novo gene de susceptibilidade para FNMTC nesta família, e que este gene possa também ser relevante na etiologia desta doença noutras famílias. É também esperado que estes estudos possam levar a uma melhor compreensão dos mecanismos celulares e moleculares envolvidos no desenvolvimento do cancro.

Palavras-Chave: Tiróide, carcinoma não medular da tiróide familiar (FNMTC), gene de susceptibilidade, sequenciação do exoma (WES).

CHAPTER 1

Introduction

1.1 CANCER

Cancer is one of the leading causes of mortality and morbidity worldwide, accounting for millions of deaths every year[1]. It is a complex disease that occurs due to failures in the mechanisms that usually control the growth and proliferation of cells, differentiation, and cellular death[2–4].

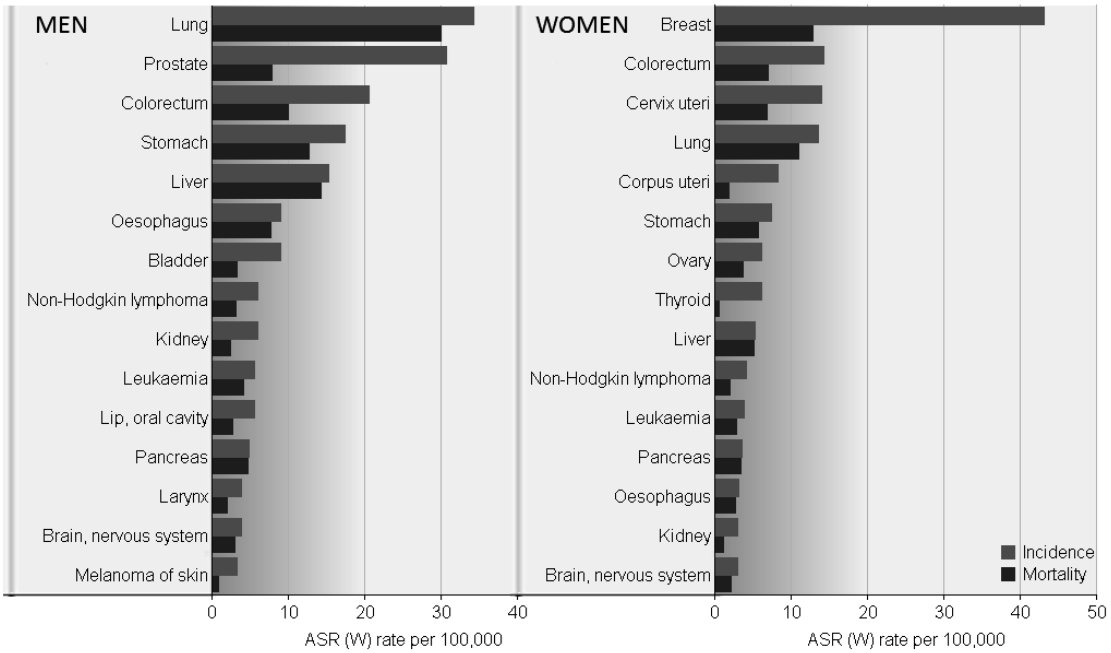


Figure 1- Estimated age-standardized (ASR) cancer incidence and mortality worldwide rates in 2012 for men (left) and women (right). Adapted from[1].

Nowadays, it is consensual that although cancer is very heterogeneous, the vast majority of tumors share some general features, known as hallmarks of cancer (Figure 2): **(1)** Self-sufficiency in proliferative signals; **(2)** Insensitivity to growth-inhibitory (antigrowth) signals; **(3)** Resistance of programmed cell death (apoptosis); **(4)** Limitless replicative potential; **(5)** Sustained angiogenesis; **(6)** Tissue invasion and metastasis; **(7)** Deregulated metabolism and **(8)** Evasion of the immune system[5,6] (Figure 2). The progressive and successive acquisition of these hallmark capabilities enables the transformation of normal cells into cancer cells, supporting their survival, proliferation and metastatic dissemination[5,6]. These acquisitions are possible due to two enabling characteristics (Figure 2): **(1)** Genomic instability and mutations and **(2)** Tumor-promoting inflammation[6].

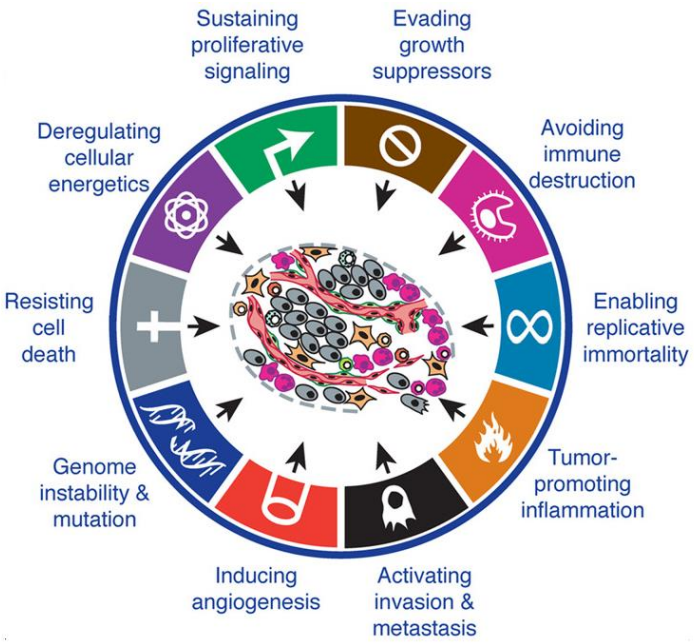


Figure 2- Illustration of the cancer hallmarks and enabling characteristics, proposed by Hanahan and Weinberg in 2011. Adapted from[6].

It is clear that tumorigenesis is a multi-step process resulting from an interplay between environment and genetics. Tumorigenesis can occur due to alterations in three major broad classes of genes[7]:

(1) Oncogenes which result from activation of proto-oncogenes, usually due to a single genetic alteration. Generally involves amplifications, missense mutations, chromosomal rearrangements and epigenetic events that results in a gain-of-function and confer a selective growth advantage on the cell[2,7,8].

(2) Tumor suppressor genes that, when inactivated in both alleles, lose their function of cell growth restrain. The loss-of-function can arise from large chromosomal deletions that can lead to loss of heterozigoty (LOH), point mutations (missense, nonsense, indel), and/or epigenetic events[2,7,9].

(3) Stability genes responsible for DNA repair, control of mitotic recombination and chromosomal segregation. The loss-of-function of these genes increases the rate of mutations in other genes, including proto-oncogenes and tumor suppressor genes, and occurs by the inactivation in both alleles [7,10].

All these alterations can occur in the germline, resulting in hereditary tumor predisposition, and/or in the somatic cells, resulting in sporadic tumors that are not inherited[7].

1.2 THYROID GLAND

The thyroid gland is located in the lower part of the neck, anterior to the trachea and comprises two lateral lobes connected by the isthmus (Figure 3)[11–13].

The thyroid uses iodine to synthesise the hormones triiodothyronine (T3) and thyroxine (T4). These hormones are essential to control some vital functions of the organism, such as the heart rate, blood pressure, body temperature, and basal metabolic rate[12,13].

The thyroid gland is composed by two main types of epithelial cells: **(1)** the C cells (or parafollicular cells), that secrete calcitonin, a hormone responsible for calcium regulation, and **(2)** the follicular cells, that comprises most of the epithelium, responsible for iodine uptake and T3 and T4 secretion[11,12].

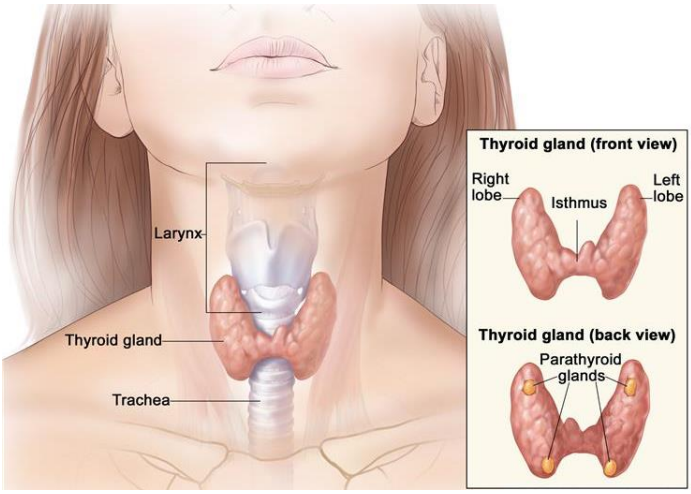


Figure 3- Localization of thyroid gland. Adapted from[14].

1.3 THYROID CANCER

Thyroid cancer accounts for about 1-3% of all malignancies[15], being the most common endocrine neoplasia[16–20]. The incidence of this neoplasia is increasing, being one of the fastest-increasing cancers among common human malignancies (Figure 4), probably due to an improvement in diagnostic techniques and the effect of some environmental factors[16,19,21]. Regardless the rising incidence of this neoplasia, in the great majority of the cases, the prognosis continues to be favourable[16,19,21].

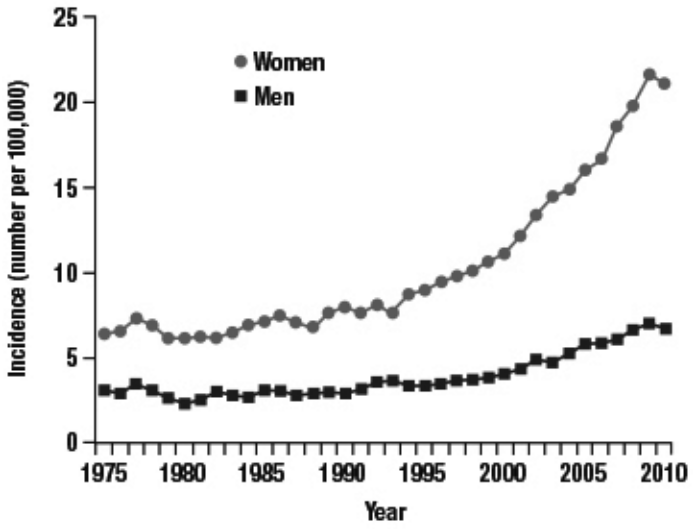


Figure 4- Increasing incidence of thyroid cancer over the years in women and men of United States. Adapted from[22].

The aetiology is not completely known, but some factors have been associated with thyroid cancer development, and these include: environmental factors, such as radiation exposure and iodine deficiency or excess, and genetic alterations.[15]

There is a significant discrepancy in this cancer incidence and prevalence by gender, with a ratio of affected women to men of approximately 3:1 (Figure 4), suggesting also the involvement of hormonal factors[15,23,24].

Thyroid cancer can be divided in two main types of neoplasias, medullary thyroid cancer and non medullary thyroid cancer, depending on the thyroid cells from which it is originated[16,19,21].

Carcinomas originated from parafollicular cells are referred as medullary thyroid carcinomas (MTC) and account for less than 5% of all thyroid cancers. MTC have a familial predisposition in about 25% of cases, which are usually associated whit mutations in the *RET* (*Rearranged during Transfection*) proto-oncogene[19,24–27]. In addition, our group recently showed that in sporadic MTC, besides somatic *RET* mutations that are detected in 61.5% of the cases, *RAS* (*Rat Sarcoma Virus Homolog*) mutations are also involved in the development of these tumors (~ 30%)[28].

Carcinomas originated from follicular cells are designated as non medullary thyroid carcinomas (NMTC), and account for more than 95% of all cases of thyroid cancer. Benign thyroid tumors, such as follicular adenomas, may also

arise from follicular cells [11,20], and may have the potential to serve as precursors for some follicular carcinomas[18]. Thyroid tumors can occur in association with other lesions, such as thyroiditis and nodular hyperplasia[11,20].

In this thesis, the study was focused in neoplasias derived from follicular cells.

1.4 NON MEDULLARY THYROID CARCINOMAS

In the past decade, NMTC has been widely studied and the knowledge of the genetic alterations occurring in these neoplasias has rapidly expanded [18]. It is well established that most genetic alterations in non medullary thyroid cancer involves the effectors of two receptor tyrosine kinase (RTK) signalling pathways: the MAPK (Mitogen Activated Protein Kinase) pathway and the PI3K-AKT (PI3K-Phosphoinositide 3-kinase; AKT- Protein kinase B) pathway[18,29] (Figure 5). Genetic or epigenetic somatic changes in genes encoding the components of the MAPK and PI3K pathways lead to constitutive activation of these classical RTK signalling pathways, playing a fundamental mechanistic role in thyroid tumorigenesis[20,30].

A prominent mutation found in thyroid carcinomas is a transverse point mutation of *BRAF* (*B-Raf Proto-Oncogene, Serine/Threonine Kinase*) (p.V600E), which causes a constitutive activation of the MAPK pathway [18,20,30](Figure 5).

Second in prevalence in thyroid carcinomas are *RAS* mutations. *Ras* genes encode GTPase proteins located at the inner surface of the cell membrane, and transmit signals arising from cell membrane RTKs and G-protein-coupled receptors along the MAPK, PI3K-AKT and other signal pathways[18,29]. There are three human isoforms of *RAS*: *HRAS (Harvey Rat Sarcoma Viral Oncogene Homolog)*, *KRAS (Kirsten Rat Sarcoma Viral Oncogene Homolog)* and *NRAS (Neuroblastoma Rat Sarcoma Viral Oncogene Homolog)*. Activating point mutations usually affect codons 12, 13 and 61 of *RAS* genes[18,29]. Although *RAS* is a classical dual activator of the MAPK and PI3K-AKT pathways, in thyroid tumorigenesis seems to preferentially activate PI3K-AKT pathway[20,30] (Figure 5).

Some chromosomal rearrangements affecting these two signalling pathways are also frequent in thyroid cancer, namely *RET/PTC* rearrangements, which leads to chronic stimulation of MAPK signalling [20,29,30](Figure 5).

The progression and dedifferentiation of thyroid carcinomas involves a number of additional mutations that affect *TP53* (*Tumor Protein 53*), PI3K-ATK pathway and other cell signalling pathways, such as WNT-β-catenin signalling pathway and HIF1α (Hypoxia-inducible factor 1-alpha) pathway[18].

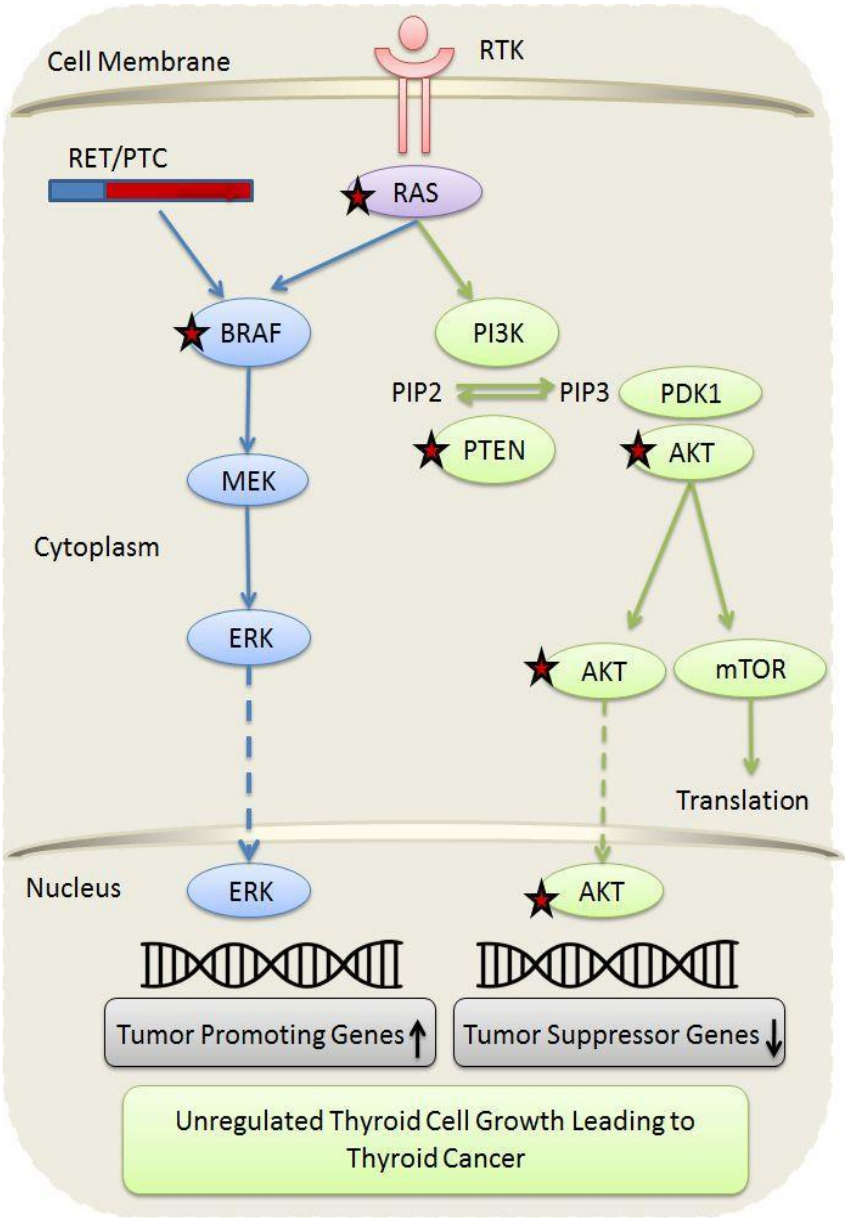


Figure 5- Genetic changes (star) in thyroid cancer affects MAPK and PI3K-AKT pathways. Adapted from[30].

NMTC can be subdivided in three subtypes, well differentiated thyroid carcinoma, poorly differentiated thyroid carcinoma and undifferentiated (anaplastic) thyroid carcinoma[16,29,30] (Figure 6).

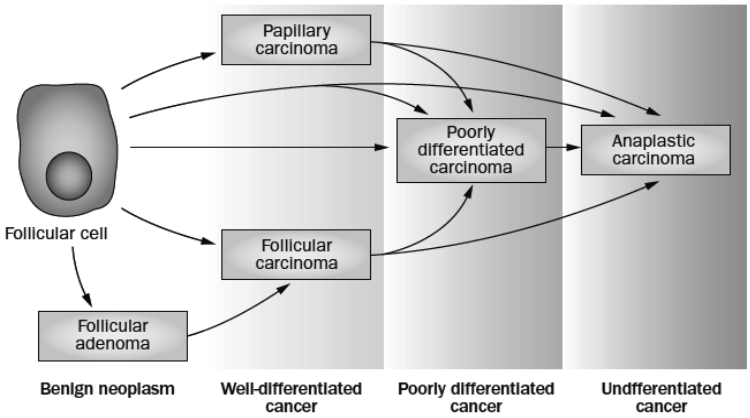


Figure 6- Types of follicular cell-derived thyroid cancer. Adapted from[18].

More than 90% of patients with thyroid cancer present well differentiated thyroid carcinomas, which include papillary thyroid carcinoma (PTC) and follicular thyroid carcinoma (FTC) [16,18,29,30].

PTC is the most common type of thyroid cancer[11]. It is characterized by ill-defined margins and distinctive nuclear features that include enlargement, oval shape, elongation, overlapping and clearing, inclusions and grooves[11,20,29]. PTC can be multifocal and is prone to lymphatic metastasis[11,20]. Subtypes of this carcinoma include classic PTC (cPTC), follicular-variant PTC (fvPTC), tall-cell PTC (tcPTC) and a few rare variants[11,20,29]. The most common genetic alterations in PTCs include mutations in *BRAF, RAS, TRK* (*Tropomyosin-Related Protein*) rearrangements and more than 10 different types of *RET/PTC* translocations with different partner genes, being the most common types *RET/PTC1* and *RET/PTC3,* in which *RET* is fused to either *CCDC6 (Coiled-Coil Domain Containing 6)* or *NCOA4 (Nuclear Receptor Coactivator 4)*, respectively [18,20] (Table I).

FTC usually present vascular or capsular invasion, lacking the nuclear features of PTC[11,20,29]. FTC has propensity to metastasis via the blood stream. Hürthle cell thyroid cancer is a subtype of FTC, characterized by large, mitochondria-rich oncocytic cells with high propensity for metastasis[11,20]. The most common genetic alterations in FTCs include mutations in *RAS* and *PAX8/PPARγ* (*PAX8 – Paired Box 8*; *PPARγ – Peroxisome Proliferation-Activated Receptor type γ*) rearrangements [18] (Table I).

The less common forms of NMTC are the poorly differentiated thyroid carcinomas (PDTC) and the undifferentiated or anaplastic thyroid carcinomas (ATC) [15].

PDTC are neoplasias difficult to categorize, and represent, morphologically and biologically, an intermediate between well differentiated and undifferentiated carcinomas[11,20]. The most common genetic alterations in PDTCs are mutations in *RAS*, *TP53* and CDKIs (Cyclin-dependent kinase inhibitors) [18,31] (Table I).

ATC are the thyroid tumors with the worse prognosis, being extremely metastatic and lethal[11,20]. The cells are completely or almost completely undifferentiated, and the tumor presents as a rapidly enlarging neck mass[11,20]. These neoplasias may occur *de novo* or arise through the process of PTC and FTC dedifferentiation[18,29]. Some of the most common genetic alterations present in these carcinomas are mutations in *RAS*, *TP53*, *BRAF*, CDKIs, *CTNNB1 (Catenin Beta 1)* and *PIK3CA* (*Phosphatidylinositol-4,5-bisphosphate 3-Kinase, Catalytic Subunit Alpha*)[18,31] (Table I).

Table I- Types of non medullary thyroid carcinomas and their mutational profiles. Adapted from[18,31] .

Characteristics	Papillary carcinoma	Follicular carcinoma	Poorly differentiated carcinoma	Anaplastic carcinoma
Prevalence (%)	80-85	10-15	<2	1-2
10-year survival (%)	95-98	90-95	~50	<10
Common mutations and their prevalence (%)	*BRAF* 40-45 *RAS* 10-20 *RET/PTC* 10-20 *TRK* <5	*RAS* 40-50 *PAX/PPAR*γ 30-35 *PIK3CA* <10 *PTEN* <10	*RAS* 20-40 *TP53* 20-30 CDKI 14-20 *BRAF* 10-20 *CTNNB1* 10-20 *PIK3CA* 5-10 AKT1 5-10	*TP53* 50-80 *CTNNB1* 5-60 *RAS* 20-40 *BRAF* 20-40 *PIK3CA* 10-20 CDKI 10-14 *PTEN* 5-15 AKT1 5-10

1.4.1 FAMILIAL NON MEDULLARY THYROID CARCINOMA

The vast majority of thyroid carcinomas occur as sporadic forms, however it can also occur as familial forms[19,24].

NMTC can present together with other pathologies in the context of familial cancer syndromes, or it may be the only/major manifestation in a family [15,17,24].

The first group includes several familial cancer syndromes associated with NMTC, such as Gardner's syndrome (Familial Adenomatous Polyposis), Cowden's disease, Carney complex, and Werner syndrome[15,17,19,32]. However, in the context of these syndromes, NMTC is not the most frequent cancer lesion in the families. The genes involved and the clinical features in these syndromes are well characterized and are summarized in Table II. The second group includes the familial forms of NMTC, where the principal clinical manifestation is thyroid cancer[15,17,19]. In these familial cases, the pathogenesis is not well understood yet.

Cases of familial forms of NMTC represent approximately 5% of all NMTC[15,23,33,34].

Familial non medullary thyroid carcinoma (FNMTC) is currently defined by the diagnosis of two or more first-degree relatives with differentiated thyroid cancer of follicular cell origin, without another familial syndrome, and in the absence of environmental causes[15,17,19,34].

It is thought that the inheritance pattern of FNMTC is autosomal dominant with incomplete penetrance and variable expressivity[17,19,24]. Additionally, it has also been suggested that FNMTC occurs at an early age, and that it is associated to a genetic anticipation in successive generations[35,36].

The PTC subtype is the most common in patients with FNMTC[15,33]. It is also frequent that members of these families present a spectrum of benign thyroid lesions, including follicular adenoma and multinodular goiter[15].

Some authors suggest that familial forms of NMTC are more aggressive and with a less favourable prognosis than sporadic forms[36,37], however, there are studies that reported that the aggressiveness of sporadic and familial forms of NMTC is similar[34,35,38]. In particular, a research developed by our group on this topic supported the latter hypothesis[34].

Table II - Familial syndromes that include NMTC as phenotype. Adapted from[15].

Syndrome	Clinical Feautures	Inheritance pattern	*Locus*	Gene
Gardner's syndrome	Gastrointestinal adenomatous polyps, osteomas, epidermoid cysts, hypertrophy of the retinal epithelium, desmoid tumors, PTC (cribiform morular variant)	Autosomal dominant	5q21	*APC*
Cowden disease	Hamartoma, breast cancer, PTC, FTC	Autosomal dominant	10q22	*PTEN*
Carney complex	Pituitary, gonadal, and adrenal gland cancer, PTC, FTC	Autosomal dominant	17q23-24	*PRKAR1a*
Werner syndrome	Premature aging, soft tissue sarcomas, osteosarcoma, FTC/PTC	Autosomal recessive	8p11-p12	*WRN*

PTC, Papillary thyroid carcinoma; FTC, Follicular thyroid carcinoma.

1.4.1.1 CANDIDATE GENES FOR FNMTC

In order to identify the gene(s) responsible for thyroid cancer predisposition, several studies employing genome wide linkage analysis, using microsatellite markers uniformly distributed across the genome, and large (informative) pedigrees, with multiple affected members, have reported putative susceptibility *loci* for FNMTC[15]. Up to now, with these analyses, several susceptibility chromosomal regions for FNMTC have been mapped and these findings are summarized in Table III.

So far, only two susceptibility genes mapped using *linkage* studies have been identified: *DICER1* (*MNG1 locus*) and *SRGAP1*.

The *MNG1 locus,* where the *DICER1* (*Ribonuclease Type III*) gene was identified, was initially mapped in a Canadian family with 18 cases of MNG (Multinodular Goiter), of which 2 individuals also had PTCs[39]. In posterior studies, germline mutations in *DICER1* were found in this Canadian family, as well, in 4 additional families, 2 of them with familial MNG and another 2 families with MNG cases associated with Sertoli-Leydig cell tumors[40]. Recently, germline mutations in this gene were also associated with the development of NMTC in a family with 6 cases of MNG, 4 of whom also had NTMC[41]. This gene is essential for microRNA (small non-coding RNA molecules) processing, which can negatively regulate gene expression at post-transcriptional level[42].

Regarding *SRGAP1* (*SLIT-ROBO Rho GTPase Activating Protein 1*) gene, two missense variants, p.Q149H and p.R617C, were identified in two different

FNMTC families in a total of 38 FNMTC families. These two genetic alterations were also analysed in sporadic PTC cases and unaffected controls and, the p.R617C alteration was found in 11/3386 sporadic cases and in 7/3395 controls, while p.Q149H was not found in sporadic cases or controls[43]. It was suggested that these variants alter the activity of CDC42 (Cell division control protein 42 homolog), a protein that is involved in intracellular signalling, and is important in cellular processes usually altered in cancer, such as cellular proliferation, migration and survival[43].

Another strategy that has been used to identify susceptibility genes for FNMTC, besides the mapping studies, is the direct study of candidate genes, in which the genes are selected based on their function. Therefore, some genes were chosen to be studied because they are essential for thyroid formation during the embryogenesis, and for its differentiation and maintenance in the adult, as follows:

The thyroid transcription factors *NKX2-1/TTF1* (*NK2 Homeobox 1* or *Thyroid Transcription Factor-1*), *FOXE1/TTF2* (*Forkhead Box E1* or *Thyroid Transcription Factor-2*), *PAX8* and *HHEX (Hematopoietically Expressed Homeobox*) are crucial for the formation and differentiation of thyroid gland, and also for the maintenance of a normal function[44–46]. Mutations in genes encoding those transcription factors could lead to alterations in thyroid morphogenesis and function, contributing to tumorigenesis and therefore, were considered potential candidate genes for FNMTC[44–46]. Accordingly, recent studies reported the involvement of *NKX2.1* and *FOXE1* in FNMTC susceptibility[47,48]. Mutations in *NKX2.1* were associated with increased cell proliferation, loss of expression of differentiated thyroid markers and gain of thyroid-stimulating hormone (TSH) independence, which is required for the growth of normal thyroid cells. Mutations in *NKX2.1* appear to promote the development of benign thyroid nodules and the progression from benign disease to cancer[47]. On the other hand, genetic alterations in *FOXE1* seem to confer proliferative advantage, increase the ability to cellular migration and to potentiate the transcriptional activation of the *WNT5A* (*Wingless-Type MMTV Integration Site Family, Member 5A*) gene, thus contributing to thyroid tumorigenesis[48].

Table III- Susceptibility *loci* to FNMTC. Adapted from[17].

Locus	Location	Studies	Reference
MNG1	14q31	A Canadian family with 18 cases of MNG and 2 cases of NMTC was studied and the *MNG1 locus* was identified. The *DICER1* susceptibility gene was identified in this *locus*. Germline mutations in *DICER1* gene were found in 4 additional families with familial MNG and in 1 family with FNMTC.	Bignell *et al*., 1997[39] Rio Frio *et al.,* 2011[40] Rutter *et al.,* 2016[41]
TCO	19p13.2	A French family with 6 cases of MNG and 3 cases of NMTC was studied. The linkage for this *locus* was confirmed in independent studies, and an interaction between the *TCO* and *NMTC1 loci* was identified.	Canzian *et al*., 1998[49] Bevan *et al.,* 2001[50]
fPTC/PRN	1q21	An American family with 5 members with PTC, 1 member with colon cancer and 2 with papillary renal neoplasms was studied. The involvement of this *locus* with FMNTC has not been confirmed in other independent studies.	Malchoff *et al*., 2000[51]
NMTC1	2q21	A large Tasmanian family with recurrence of PTC was studied. The interaction between *TCO* and *NMTC1 loci* may increase the risk for FNMTC.	Mackay *et al*., 2001[52]
FTEN	8p23.1-p22	A Portuguese family with 11 cases of benign thyroid disease and 5 cases of thyroid cancer was studied by our group. The functional characterization of a candidate gene is ongoing.	Cavaco *et al*., 2008[53]
TERC-hTERT complex	3q26 and 5p15.33	Shorter telomeres, increased hTERT gene copy number, and increased telomerase activity.	Capezzone *et al.,* 2008[54]
FOXE1	9q22.33	This gene encodes a thyroid transcription factor. The involvement of this gene in the aetiology of FNMTC in a Portuguese family was reported by our group.	Pereira *et al.,* 2013[48]
NKX2-1	14q13.3	This gene encodes a thyroid transcription factor. The involvement of this gene in FMNTC was confirmed in 2 different Chinese families.	Ngan *et al.,* 2009[47]
Not named	1q21	American and Italian families with FNMTC were studied and this *locus* was identified.	Suh *et al*., 2009[55]
Not named	6q22	American and Italian families with FNMTC were studied and this *locus* was identified.	Suh *et al*., 2009[55]
Not named	8q24	The involvement of this *locus* was found in 10 of the 26 families that were studied.	He *et al*., 2009[56]
miR-886-3p	5q31.2	Downregulated in FNMTC tumor samples as compared to NMTC. Regulates genes involved in DNA replication and focal adhesion.	Xiong *et al.,* 2011[57]
miR-20a	13q31.3	Downregulated in FNMTC tumor samples as compared to NMTC. In epithelial cells, promotes cellular proliferation and invasion.	Xiong *et al.,* 2011[57]
SRGAP1	12q14	This *locus* was linked to FNMTC in 21 of 38 families with FNMTC analysed. Mutations in the *SRGAP1* gene were found in 2 of the 38 families.	He *et al*., 2013[43]

More recently, a new technology, the next generation sequencing (NGS), that will be further discussed, has been used to discover susceptibility genes in many types of familial cancers. Studies published in 2015, already reported two new susceptibility genes for FNMTC, using NGS.

In one of the studies, a missense variant (p.S346F) in *SRRM2* (*Serine/Arginine Repetitive Matrix 2*) gene was identified in a well-document family with 6 cases of PTC[58]. For the identification of this variant, a combination of different methodologies, such as whole exome sequencing (WES), haplotype analysis and genetic linkage analysis were used[58]. Besides the cosegragation of this variant with the PTC phenotype in the family, the alteration was also found in 7 of 1170 (0.6%) of patients with sporadic PTC, but was not found in 1404 unaffected controls or in additional FNMTC families[58]. The *SRRM2* is involved in pre-mRNA splicing, and the authors of this study suggested that this gene predisposes to PTC by affecting alternative splicing of unidentified target genes[58].

The authors of the other study identified in a family, with 6 cases of PTC and with 1 case of follicular thyroid adenoma, a germline variant (p.G534E) in the *HABP2* (*Hyaluronan Binding Protein 2*) gene[59]. They also identified this variant in 20 of 423 (4.7%) patients with sporadic PTC[59]. Functional studies performed by these authors showed that this variant increased colony formation and cell migration, suggesting that *HABP2* could be a susceptibility gene for FMNTC[59]. However, subsequent studies published by other authors did not support these findings, since the variant was found in several NMTC sporadic cases (0.2-8%) and in unaffected controls (0.5-8.7%), suggesting that it would be unlikely that the variant identified could cause FNMTC[60–62].

As mentioned before, the aetiology of FNMTC, as a distinct syndrome, continues poorly understood and the responsible genes, that may lead to a predisposition to FNMTC have not yet been identified for the great majority of the families[15,17,24]. In the past few years, several studies have been developed in order to clarify the genetic basis of these familial thyroid carcinomas. Establishing a molecular diagnosis for FNMTC may provide the opportunity for its early detection in family members, which may lead to a decrease of the morbidity and mortality associated with these neoplasms, and to the development of new targeted therapies[15].

1.5 TECHNOLOGIES AND APPROACHES USED TO IDENTIFY SUSCEPTIBILITY GENES FOR FAMILIAL CANCER

1.5.1 LINKAGE ANALYSIS

Linkage analysis has been one of the most widely used approach to map susceptibility genes for FNMTC. Linkage studies denotes a group of statistical methods that are used to map the chromosomal region that is most likely to contain the gene responsible for the trait of interest [63,64].

Linkage is defined as the tendency of two or more genetic *loci* to be transmitted together from parents to offspring during meiosis, because they are physically close together on a chromosome[63–65]. The alleles of this two *loci*, that are physically close on the same chromosome and will be co-inherited, are said to be linked [66,67].

The linkage analysis uses polymorphic markers, such as single nucleotide polymorphisms and microsatellites, to track co-segregation of genetic material with a specific phenotype in selected families[64]. It is expected that a gene physically close to a polymorphic marker will be co-segregated, therefore, knowing the chromosomal localization of the marker it is possible to identify a particular chromosomal position that is co-inherited with the trait within the family.

Two types of statistical approaches can be used for these studies: the non-parametric and the parametric linkage analysis[63].

Non-parametric or model-free linkage does not require the specification of a disease model or assumptions about the distribution of the disease in the population[63–65]. While, parametric or model-based linkage analysis requires specification of an inheritance disease model[63–65].

In the FNMTC family studied in the present thesis, we performed a parametric linkage analysis, assuming an autosomal dominant model of inheritance.

Linkage studies allow only the mapping of a region that potentially contains the gene of interest, and these regions can include hundreds of genes. Thus, additional technologies, such Sanger sequencing, have to be performed in order to identify the specific susceptibility gene that is associated with the disease under study[67].

1.5.2 NEXT GENERATION SEQUENCING

For decades, the genome sequencing and the investigation of the genetic background for several diseases was possible using the Sanger method sequencing[68,69]. More recently, a new technology designated Next Generation Sequencing (NGS) has emerged, and it allows the DNA sequencing on a massive scale, producing millions of simultaneous sequences, thereby, generating significantly more sequence reads per instrument run, at a considerably lower cost[68–72]. Thus, NGS is significantly faster, more accurate and requires less amount of DNA to perform sequencing, than the Sanger method.

This new technology includes the whole genome sequencing (WGS), and the sequencing of the coding genome only, whole exome sequencing (WES), that corresponds to approximately 1% of all genome, but is expected to harbour about 85% of pathogenic mutations[68,73–75].

In both WGS and WES is not necessary to define *apriori* a candidate gene, since this technology compares, respectively, the all genome and exome in study, with reference sequences, allowing the identification of genomic alterations[71,76].

One of the biggest challenges associated with WGS and WES is the interpretation of the large amount of genomic data generated[74,75]. Using these technologies, thousands of rare genetic variants are identified, being difficult to determine which variants are indeed associated with the development of the disease in study[74,75].

Therefore, the combination of different approaches, such as *linkage* studies and NGS analysis, can facilitate in the identification of susceptibility genes for familial cancers, namely FNMTC.

CHAPTER 2

Objectives

2.1 OBJECTIVES

The presently known FNMTC susceptibility genes only account for a very small fraction of familial thyroid cancer genetic susceptibility. Therefore, the main objective of this work was the identification of new candidate genes that may confer susceptibility to FNMTC.

Thus, the molecular characterization of a highly informative FNMTC family was performed, using both linkage analysis and sequencing technologies, such as WES and Sanger sequencing, to identify potentially pathogenic genetic variants.

Several bioinformatic analyses were also performed in order to select potentially pathogenic genetic variants that may confer susceptibility for thyroid cancer in the family under study.

Additionally, to investigate the mechanisms that could be involved in the progression of thyroid cancer, an analysis of genes that are frequently mutated in sporadic thyroid carcinomas was performed in tumors samples from patients of the family under study.

This work intended to contribute to a better understanding of the genetic mechanisms involved in the aetiology and development of FNMTC, which, as mentioned already, are at the moment poorly clarified.

CHAPTER 3

Materials and Methods

The experimental approach for this study involved a combination of different strategies and analyses, which are schematized in Figure 7.

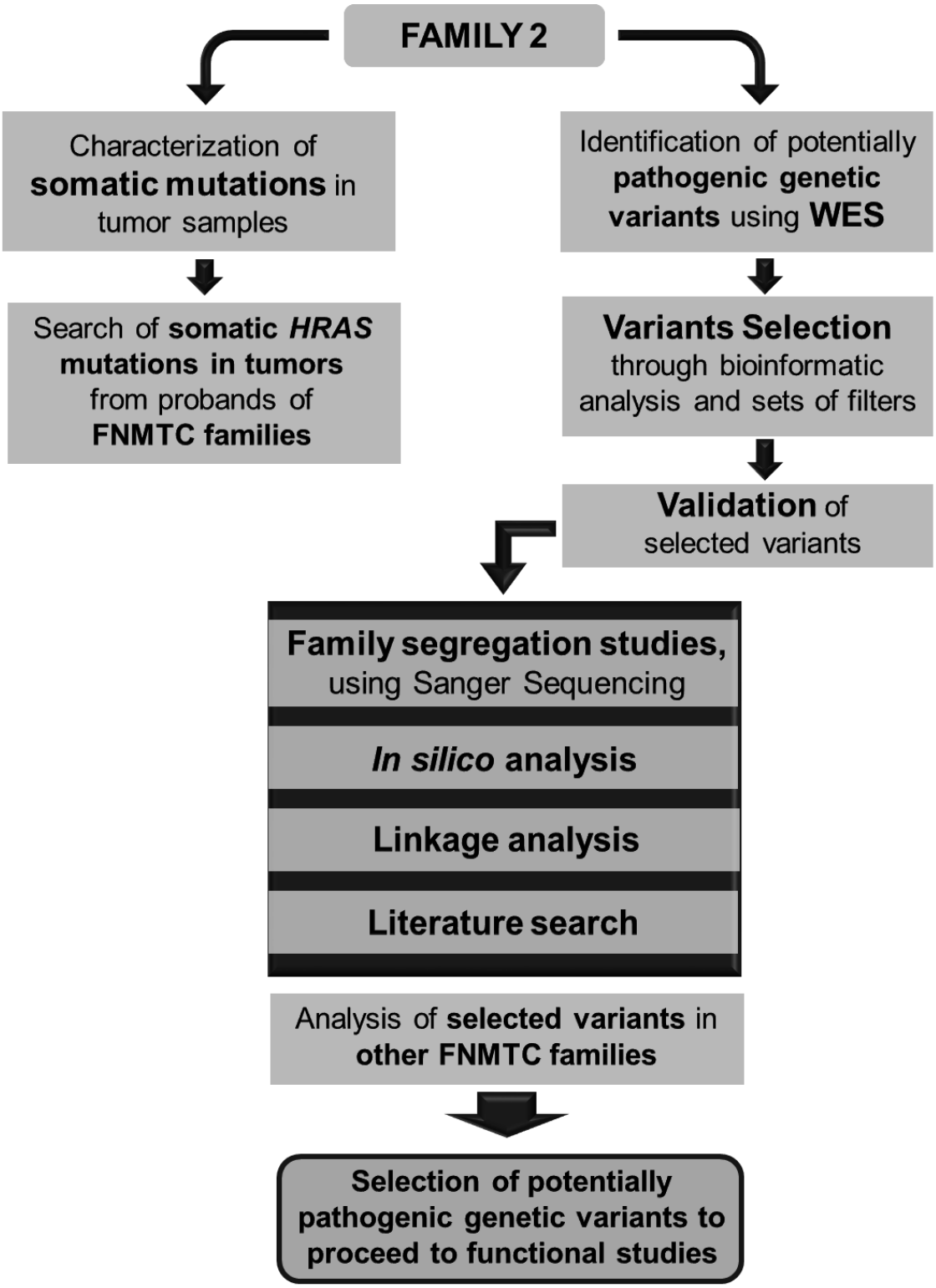


Figure 7- Schematic representation of the steps and strategies used in this work.

3.1 FAMILY 2

The family studied in this work is followed in the Endocrinology Department from I.P.O.L.F.G. This family, represented in Figure 8, has 7 members affected with NMTC (III.10, IV.2, IV.6, IV.8, V.3, V.8 and V.67), 7 patients with nodular goiter only (I.2, III.4, IV.4, IV.25, IV.33, IV.60 and IV.63), and several members affected with other types of cancer, such as bladder cancer, testicular cancer and ovarian cancer. In addition, 5 of the 7 cases affected with NMTC also had colon lesions, which may represent, or not, a syndromic aspect related with the disease under study. Previously, mutations in known susceptibility genes *(FOXE1, NKX2-1)* and positional/functional candidate genes *(HHEX, PAX8, PTEN, CDKN2A, CDKN2B)*, were excluded.

The studies performed in this project were approved by the Ethical Committee of I.P.O.L.F.G. The collection of the samples was undertaken following written informed consent.

3.2 BIOLOGICAL SAMPLES

The biological samples analysed consisted in peripheral blood samples, fresh tissue samples frozen in liquid nitrogen, and formalin-fixed paraffin embedded (FFPE) tumor samples.

A total of 42 peripheral blood samples and 6 FFPE tumor samples from patients and relative members of family 2 were available to study.

From a series of 87 FNMTC families, 56 tumor samples of the probands or relatives (whenever there was no proband sample) were available to study. From this series, a total of 17 FFPE tumor samples and 5 fresh frozen tumor samples of 22 families, were analysed.

A total of 12 peripheral blood samples from healthy controls were also used.

3.3 DNA EXTRACTION

The DNA extraction from peripheral blood, frozen tissue and FFPE tissue samples had been previously performed by researchers of our group.

Extraction and purification of DNA from blood and frozen tissue was performed using the Puregene® Blood Core Kit (Qiagen, Hilden, Germany), according to the manufacture's protocol (Figure 9 a). The extracted DNA was

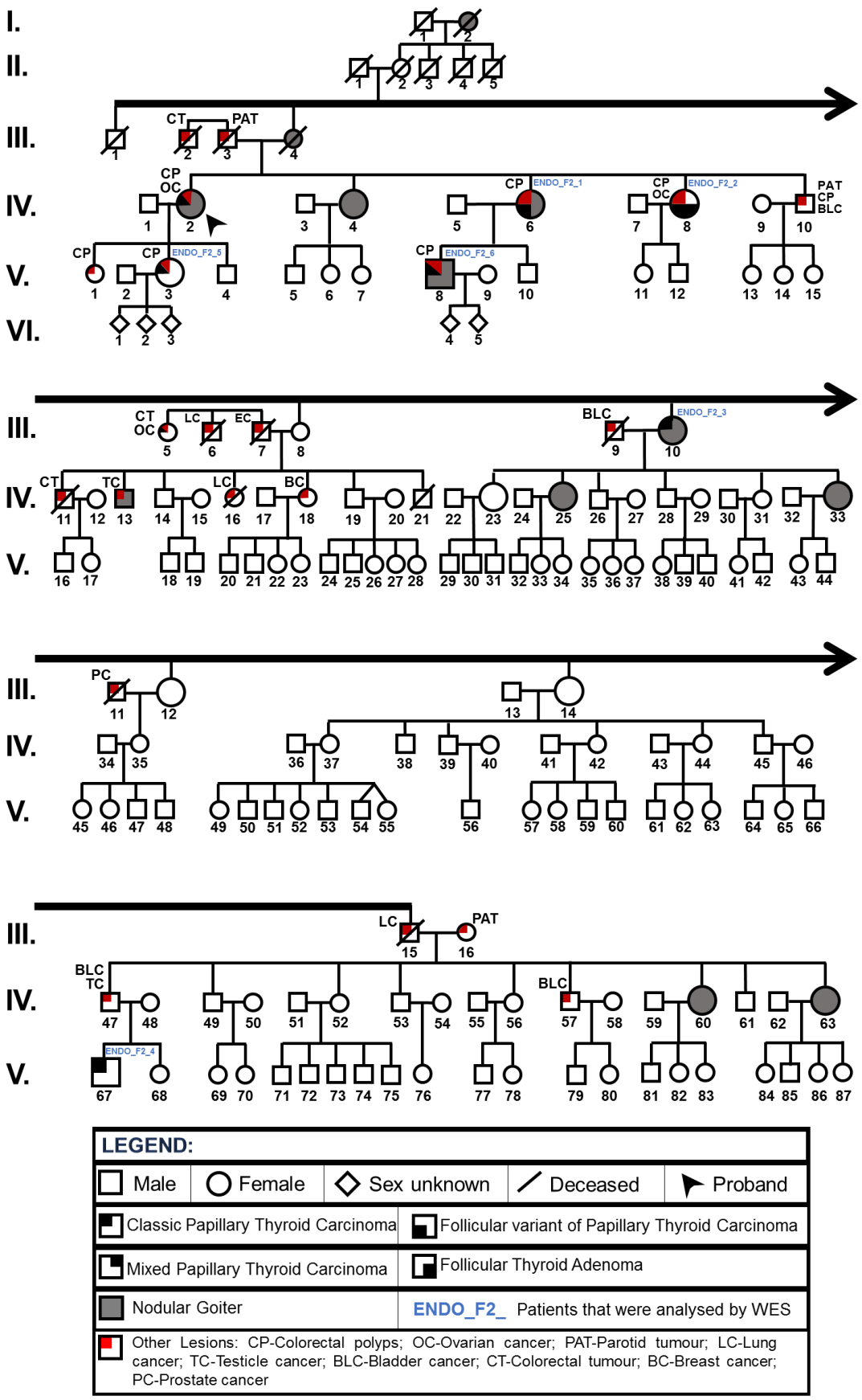


Figure 8- Pedigree of family 2. In this pedigree are represented 6 generations of this family, in which 3 generations are affected with NMTC.

quantified by UV Spectrophotometry (NanoDrop ND-1000, Thermo Fisher Scientific, USA).

Regarding the extraction of DNA from FFPE samples, the GeneRead™ DNA FFPE Kit (Qiagen, Hilden, Germany) was used according to the manufacturer's protocol (Figure 9 b). The DNA integrity and quantification was assessed by agarose gel electrophoresis.

A representative scheme of the experimental procedure of DNA extraction using both kits is shown in Figure 9.

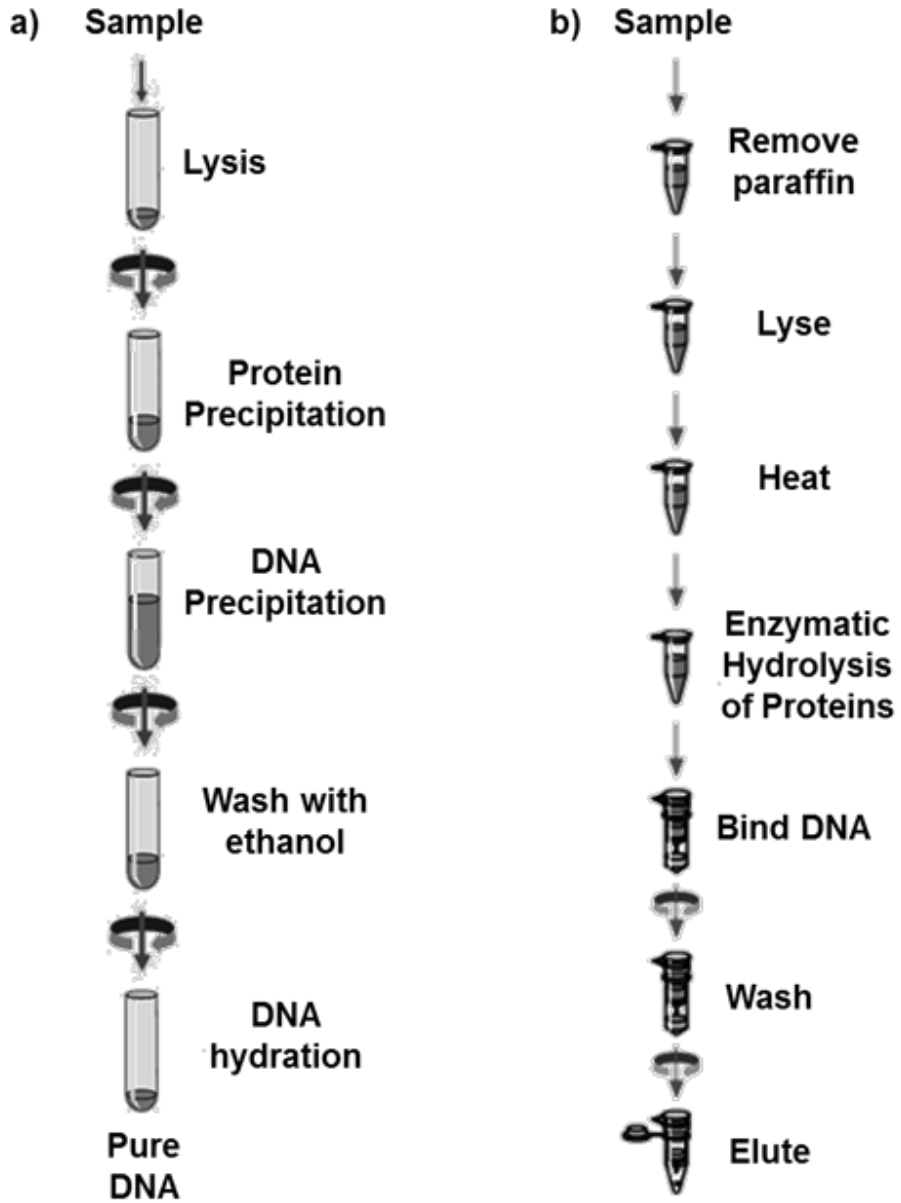


Figure 9- Representative scheme of the experimental procedure of DNA extraction from a) peripheral blood or frozen tissues and b) formalin-fixed paraffin embedded samples. Adapted from[77,78].

3.4 WHOLE EXOME SEQUENCING

Family 2, identified in the Endocrinology Department of I.P.OL.F.G, was chosen for whole exome sequence (WES), as it was a highly informative FNMTC family, presenting a high number of affected members, and because it did not present mutations in known susceptibility genes.

Genomic DNA samples of peripheral blood (3 µg at minimum) from 6 representative members of this family were used for WES. The 6 members sequenced are indicated in Figure 8 as ENDO_F2_1, ENDO_F2_2, ENDO_F2_3, ENDO_F2_4, ENDO_F2_5 and ENDO_F2_6.

The sequencing of whole exome was performed in the Erasmus Center for Biomics, a company placed in the Netherlands. WES was performed using the Agilent version 4 capture kit, followed by the Illumina TruSeq v3 protocol in a HiSeq2000.

The NGS Illumina technology relies in a massively parallel sequencing by synthesis (SBS) approach with fluorescent dyes[79–81]. The workflows of this technology, illustrated in Figure 10, include three main steps: **(1)** library preparation; **(2)** cluster generation and **(3)** sequencing.

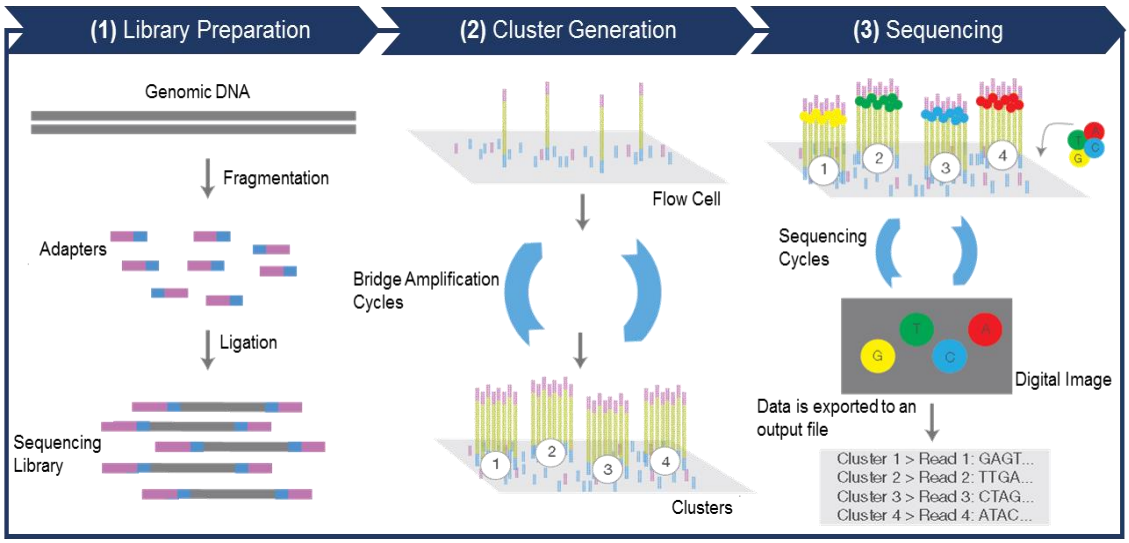


Figure 8- NGS Illumina technology workflows. This technology consists in three main phases: (1) library preparation; (2) cluster generation and (3) sequencing. Adapted from[79].

(1) The sequencing library is prepared by random fragmentation of genomic DNA sample, and attachment of these fragments to adaptors in 3' and 5' ends[79–81]. **(2)** The library is then loaded into a flow cell, where fragments bind to their corresponding adapter sequences and amplify into clonal clusters, trough bridge

amplification[79–81]. **(3)** The templates are sequenced by SBS method. In this method, single bases are detected as they are incorporated into DNA template strands, by using different fluorescently labelled deoxyribonucleotide triphosphates **(**dNTP's). The flow cell is imaged and, the emission wavelength and intensity from each cluster is recorded, allowing the identification of nucleotide bases[79–81].

The raw sequence reads obtained in WES, were then generated in *.fastq files format to allow subsequent data analysis.

3.5 VARIANTS SELECTION

After WES, a large amount of data was obtained and, in order to interpret this sequencing data and select the genetic variants that are not artefacts and that could be associated with familial thyroid cancer on family 2, several analyses were performed.

3.5.1 BIOINFORMATIC ANALYSIS

The bioinformatic analysis of the sequencing data was initially performed by Bioinf2Bio, a bioinformatics data analysis company, placed in Porto, Portugal.

The company converted the *.fastq files in *.SAM and then in *.BAM files to enable easy visualization in the Integrative Genomics Viewer (IGV) from the Broad Institute, a tool for interactive exploration of integrated genomic datasets, and to perform sequencing data analysis.

At our request, four different and independent types of analyses were performed. These different analyses were undertaken because some affected family members could be phenocopies. A phenocopy, in this context, would be a family member with sporadic thyroid cancer, originated from a somatic genetic alteration, and not by the inheritance of the germline mutation causing the disease in this family. Four cases (ENDO_F2_1, ENDO_F2_2, ENDO_F2_5 and ENDO_F2_6) had a strong evidence of hereditary disease transmission, since three siblings (only in two was performed WES) had thyroid carcinomas and two of them had offspring affected with thyroid cancer. Therefore, in this work, the members of this family branch, designated as "nuclear family", were included in all bioinformatic analyses. The remaining two cases (ENDO_F2_3 and ENDO_F2_4) did not have first degree relatives with thyroid cancer, thus, in

some bioinformatics analyses they were assumed as phenocopies and were excluded, as follows:

In analysis 1, all variants shared by all patients (ENDO_F2_1, ENDO_F2_2, ENDO_F2_3, ENDO_F2_4, ENDO_F2_5 and ENDO_F2_6) were considered. In analysis 2 only the variants shared by the four patients of the nuclear family (ENDO_F2_1, ENDO_F2_2, ENDO_F2_5 and ENDO_F2_6) were considered, and the variants of the two possible phenocopies were excluded. Regarding analysis 3, only the variants shared with one of the possible phenocopy (ENDO_F2_4) were excluded. In analysis 4, the variants shared with the other possible phenocopy (ENDO_F2_3) were excluded. These four analyses are schematically represented in Table IV.

Table IV- Schematic representation of the bioinformatic analysis of WES data from family 2

Analysis	Analysis Scheme						Variants shared by:
1	ENDO_F2_1	ENDO_F2_2	ENDO_F2_5	ENDO_F2_6	ENDO_F2_3	ENDO_F2_4	All patients
2	ENDO_F2_1	ENDO_F2_2	ENDO_F2_5	ENDO_F2_6	~~ENDO_F2_3~~	~~ENDO_F2_4~~	Nuclear family patients (1+2+5+6). Exclusion of patients 3 and 4 (phenocopies)
3	ENDO_F2_1	ENDO_F2_2	ENDO_F2_5	ENDO_F2_6	ENDO_F2_3	~~ENDO_F2_4~~	Nuclear family patients and patient 3 (1+2+5+6+3). Exclusion of patient 4 (phenocopy)
4	ENDO_F2_1	ENDO_F2_2	ENDO_F2_5	ENDO_F2_6	~~ENDO_F2_3~~	ENDO_F2_4	Nuclear family patients and patient 4 (1+2+5+6+4). Exclusion of patient 3 (phenocopy)

The nuclear family patients are highlighted in blue, the patients that may be (or not) phenocopies are highlighted in orange, and the patients excluded in each analysis are highlighted in grey with a cross.

All reads were aligned against the reference Human Genome version GRCh38 (hg38), and 94 to 99% of the reads were successfully mapped to the Human Genome. After adequate sorting and indexing, genomic variants were called and defined by its chromosome, coordinate and reference allele. These variants were selected for annotation using Ensemble v80. The first set of criteria applied by the bioinformatics company for variant selection in the four analyses was mainly related with quality parameters:

- **Quality of the alignment of each genomic variant:**

The quality of the alignment of each genomic variant had to be equal or above 20 (threshold), if not, the variants were excluded. This is a phred-scaled quality score for the assertion made for the alternative allele and the values range from 3 to 225. High quality scores indicate high confidence calls, so, for instance, if the quality is equal to 20, the probability of error will be 10^{-2} (1 in 100).

- **Quality of the genotype:**

The quality of the genotype call had to be also equal or above 20. This is a phred-scaled quality call of the probability that the genotype call (homozygous or heterozygous) is wrong and ranges from 3 to 99.

- **Total number of high quality reads:**

The total number of high quality reads had to be equal or above 20. This criterion is mainly related with the subsequent validation methodology. Variants with lower values have more probability of being artefacts, and thus less probability of being confirmed by Sanger sequencing.

- **Undefined genotype:**

The variants detected had to correspond to a defined genotype. When the sample quality is not very good, a given variant may have multiple genotypes, and, in these cases, the variant is excluded.

The next criterion applied by the bioinformatics company was the selection of variants which had biological annotations predicted to match only with protein coding transcripts, and exclusion of variants that, due to its localization in the transcript and predicted consequence, did not have a suggestive impact, *e.g*. downstream/upstream variants and intronic variants not located in splicing regions.

In a subsequent phase, a set of filters were defined and applied in the four analyses above described:

- **Exclusion of frequent variants**

It is expected that the frequency of a genomic variant associated with a disease that is not frequent in the population, will be also rare in the population. For that reason, variants that had a frequency higher than 1% in European population

were excluded. Data on the variant frequency was obtained at 1000 Genomes project[82] and Exome Aggregation Consortium (ExAC)[83], whenever possible.

- **Exclusion of synonymous variants:**

The synonymous variants do not alter the final amino acid, and although might cause changes in protein expression, conformation and function, in the majority of the cases, no consequences are expected[84]. Thus, this type of variants were also excluded, at this stage. Nonetheless, these variants may be reviewed if no reliable pathogenic variants are found in the end of this study.

- **Exclusion of variants present in other cases with distinct forms of familial cancer:**

Two other groups of the UIPM also purchased WES in the same company, but to analyse two unrelated families, one with familial colorectal cancer and another with familial prostate cancer. Bioinf2Bio company performed also the initial part of the bioinformatic analysis in these families.

The results obtained for family 2 were compared and crossed with the results of these two families. The variants that were transversal that is, shared between the distinct families, were excluded, since it is not expected the association of a same variant with different types of familial cancer, being most likely sequencing artefacts.

- **Exclusion of Homozygous variants:**

The homozygous variants were excluded, since is thought that the disease under study has an autosomal dominant pattern of inheritance.

The variants that were not excluded by all the criteria and filters mentioned were posteriorly validated by Sanger sequencing, a methodology that will be further discussed in this thesis (Chapter 3.12).

3.6 *IN SILICO* ANALYSIS

In order to help the selection of the most promising genetic variants, several *in silico* analyses were also performed.

3.6.1 IMPACT PREDICTION

The impact prediction of the genetic alterations detected by WES, and then selected through bioinformatic analysis, was achieved using four different

tools. The bioinformatic company provided the information of two prediction tools, SIFT[85] and PolyPhen[86], and then, the information was complemented using the Mutation Taster[87] and Mutation Assessor[88] prediction tools.

- **SIFT:**

SIFT predicts whether the amino acid alteration affects protein function, based on the degree of conservation of amino acid residues in sequence alignments, derived from closely related sequences. The prediction impact score is displayed from 0 for deleterious up to 1 for tolerated.

- **PolyPhen:**

PolyPhen (Polymorphism Phenotyping) predicts the possible impact of the amino acid alteration in the structure and function of a human protein using physical and comparative considerations. The prediction impact score is displayed from 0 for benign up to 1 for probably damaging**.**

- **Mutation Taster:**

Mutation Taster predicts the possible impact of an amino acid alteration taking into account the physico-chemical characteristics of amino acids and the difference between the original and the new amino acid. The prediction impact is displayed as disease causing for probably deleterious alterations or polymorphism for probably harmless alterations.

- **Mutation Assessor:**

Mutation Assessor predicts the functional impact of an alteration based on the assessment of evolutionary conservation of amino acids residues in a protein family. The functional prediction is displayed as high or medium for predicted functional impact and low or neutral for predicted non-functional impact.

3.6.2 PROTEIN AND mRNA EXPRESSION

The expression profile patterns of the selected genes that harboured the genomic variants was evaluated both at protein and mRNA level in normal tissues, particularly in thyroid gland, using the Human Protein Atlas (HPA) database[89].

- **Protein Expression:**

The protein expression data is derived from antibody-based protein profiling using immunohistochemistry. The protein expression is categorized as not detected, low, medium or high.

- **mRNA Expression:**

The mRNA expression data is derived from deep sequencing of RNA (RNA-seq). The mRNA expression is reported as mean FPKM (Fragments Per Kilobase gene model and Million reads) and the transcript expression levels are categorized into not detected (0-0.5 FPKM), low (0.5-10 FPKM), medium (10-50 FPKM) or high (>50 FPKM).

3.7 LITERATURE INFORMATION

To evaluate if the selected genes harbouring the variants could be of our interest, an examination of the literature information available for these genes was undertaken, performing a search in PubMed and in several databases such as: gene cards[90], OMIM[91], Human Protein Reference Database[92], UNIPROT[93] and STRING Interaction Network[94].

3.8 LINKAGE ANALYSIS

Previously, DNA samples of 15 members of family 2 (III.8, III.10, III.16, IV.2, IV.4, IV.6, IV.8, IV.10, IV.23, IV.25, IV.26, IV.33, IV.56, IV.60, IV.63) were genotyped using the GeneChip® Human Mapping 250K Nsp Array (Affymetrix, USA), which enables the genotyping of more than 250,000 SNPs scattered across the genome.

This data was used to perform linkage analysis with the EasyLINKAGE plus software (version v5.02). For this analysis, the relatives were classified as **affected** (the members with thyroid cancer in nuclear family), **non-affected** (spouses) and **unknown** (members that could be asymptomatic carriers, and the two members with thyroid cancer that could be phenocopies).

A genome-wide two-point parametric linkage analysis, which consists in an analysis that is carried out between each marker and a putative disease *locus,* for the 250,000 markers, was undertaken.

The putative *locus* localization was calculated in centiMorgan (cM), a unit of recombinant frequency, which is used to measure the genetic distance between two *loci*[2]. In average, 1 cM is equivalent to 1 million base pairs (bp).

The likelihood for each localization was calculated as a LOD (logarithm of odds) score. A LOD score ≥ 2 was accepted as suggestive of linkage between

the disease *locus* and the marker involved, whereas LOD scores ≤ - 2.0 excluded linkage.

3.9 POLYMERASE CHAIN REACTION

The polymerase chain reaction (PCR) is a DNA cloning method used to amplify specific DNA regions[2].

The reaction, schematized in Figure 11, consists in cycles of temperature variations, starting with a denaturing for the separation of DNA double strand into two single strands. Subsequently, occurs the hybridization of specific primers to the target sequence that bind to the denatured DNA. The temperature of hybridization depends on the sequence of the primers used. Finally, take place the extension, where the synthesis of new complementary DNA strands occurs. The temperature for extension is set based on optimal operating temperature of the enzyme[95].

These steps constitute a cycle. In a PCR, those steps are repeated several times to increase the number of DNA copies[2,95].

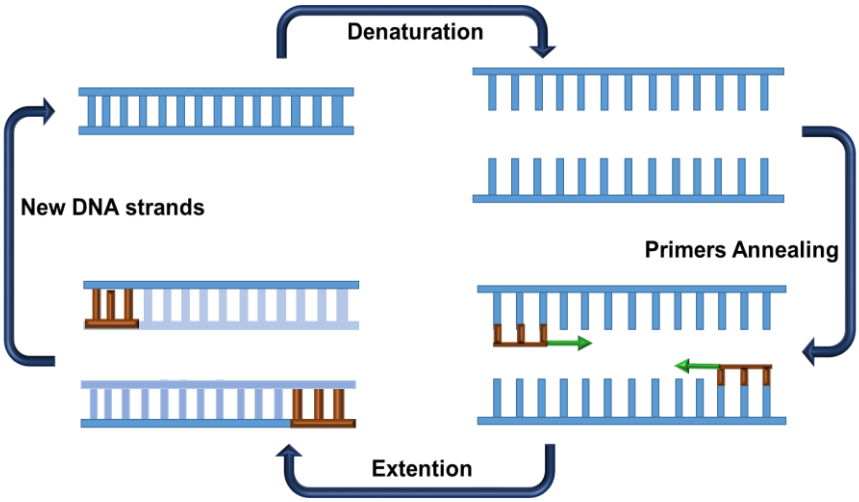


Figure 9- Polymerase Chain Reaction (PCR) technique. The PCR method consists in three main steps: 1) the denaturation of DNA double strand; 2) primers annealing; and 3) extension of new DNA strands. Adapted from[96].

In this work, the PCR reactions were prepared for a final volume of 25 µL and the following compounds were used for the majority of reactions: 2 µL of template DNA (50 ng/µL), 1 µL of forward and reverse primer (10 pmol/µL) (Invitrogen™), 1x buffer solution (Invitrogen™), 2 µL of dNTPs (2.5 mM) (Invitrogen™), 0.2 µL of Taq polymerase (5 U/ µL) (Invitrogen™), 0.5 µL up to 1.075 µL of $MgCl_2$ (50 mM)

(Invitrogen™), depending on the optimal conditions of each reaction, and bidestilled water for add up to the final volume. The primers sequences used are listed in Table XIX in Appendix.

The PCR amplifications of DNA were performed in thermocycler (Biometra, Germany) and comprised the conditions described in Table V:

Table V- Protocol used for PCRs.

Parameter	Stage/step					
	Initial Denaturation	Cycling (35-37 Cycles)				Hold
		Denaturating	Annealing	Extension	Final Extention	
Temperature	95 °C	95 °C	56 °C-60 °C	72 °C	72 °C	4 °C
Time	5 min	1 min	1 min	1min-1min15seg	10 min	until ready to purify

Min, minutes; sec, seconds; ºC, temperature in Celsius.

For all PCR reactions performed, a negative control (blank), with all the compounds mentioned above, except the DNA sample, which was replaced by water, was used to control possible contaminations with DNA.

3.10 CONDITIONS OPTIMIZATION

Optimization of PCR conditions had already been performed for *HRAS*, *NRAS, KRAS* and *BRAF* primers. For the remaining primer pairs used in this work, it was necessary to optimize PCR conditions. For the optimization, leucocyte DNAs from normal controls were used, and different concentrations of $MgCl_2$ and annealing temperatures were tested.

The concentration of $MgCl_2$ is extremely important for the efficiency and specificity of the PCR reaction. The Mg^{2+} ions are a co-factor of the enzyme DNA polymerase[2].

The annealing temperature optimization is important to maximize the primer hybridization with the target DNA region, avoiding unspecific bindings and the synthesis of unwanted products. In general, higher temperatures increase specificity, whereas lower temperatures facilitate hybridization and decrease specificity[2].

For the majority of primers, optimization of PCR conditions was performed testing two different $MgCl_2$ concentrations (1.2 mM and 1.7 mM), along with two different temperatures (57 ºC and 60 ºC).

The sequence of the primers, as well as the PCR conditions chosen for each pair of primers, are listed in Table XIX in Appendix.

3.11 ELECTROPHORESIS OF PCR PRODUCTS IN AGAROSE GEL

In the end of each PCR, the amplified products were analysed by electrophoresis in agarose gel.

The agarose gel electrophoresis enables separation and identification of DNA molecules, by applying an electric field that will separate molecules according to their molecular weight and electric charge[95]. The negatively charged molecules move to the anode, while the positively charged compounds migrate to the cathode[95]. Thus, the use of electrophoresis is effective in determining the success of the PCR reaction, since the DNA is a negatively charged molecule and when applied an electric field, has the ability to migrate towards the positively charged electrode, leading to the separation of the PCR products according to their weight and charge.

The concentration of agarose used was usually 2%. For the gel preparation, the agarose (E agarose, Conda-Pronadisa, Spain) was dissolved in 1x TBE (Tris-Borate-EDTA; National Diagnostics, USA) solution. The preparation included the addition of ethidium bromide (PanReac AppliChem, Germany), in a final concentration of 0.5 µg/ml. The ethidium bromide intercalates in the DNA bases, forming a complex that, when it is irradiated by UV light, emits fluorescence, allowing DNA fragments visualization.

Agarose gels were loaded with PCR products (usually 3 µL) and 0.7 µL of Gene Ruler 50bp DNA ladder (Thermo Fisher Scientific™, EUA), that allowed the molecular weight determination of the amplified products. Prior to gel loading the PCR products and the molecular weight marker were mixed with 1 µL of loading buffer (6x DNA Loading Dye; Thermo Fisher Scientific™, EUA). Depending on the gel size, an electric field of 110 V up to 145 V was applied.

A transilluminator coupled to a video camera and to an informatics system (BioDocAnalyze, Biometra, Germany) was used to expose the agarose gel to UV light radiation, thereby enabling the visualization and image capture of the gel bands.

3.12 AUTOMATED SEQUENCING

The automated DNA sequencing generally known as Sanger sequencing, is a technique that determines the sequence of nucleotides in a DNA fragment, using an automatic sequencer.

This technique, schematized in Figure 12, consists in the synthesis of DNA strands, which possess labelled nucleotides at one end, and differ from each other by a nucleotide[97].

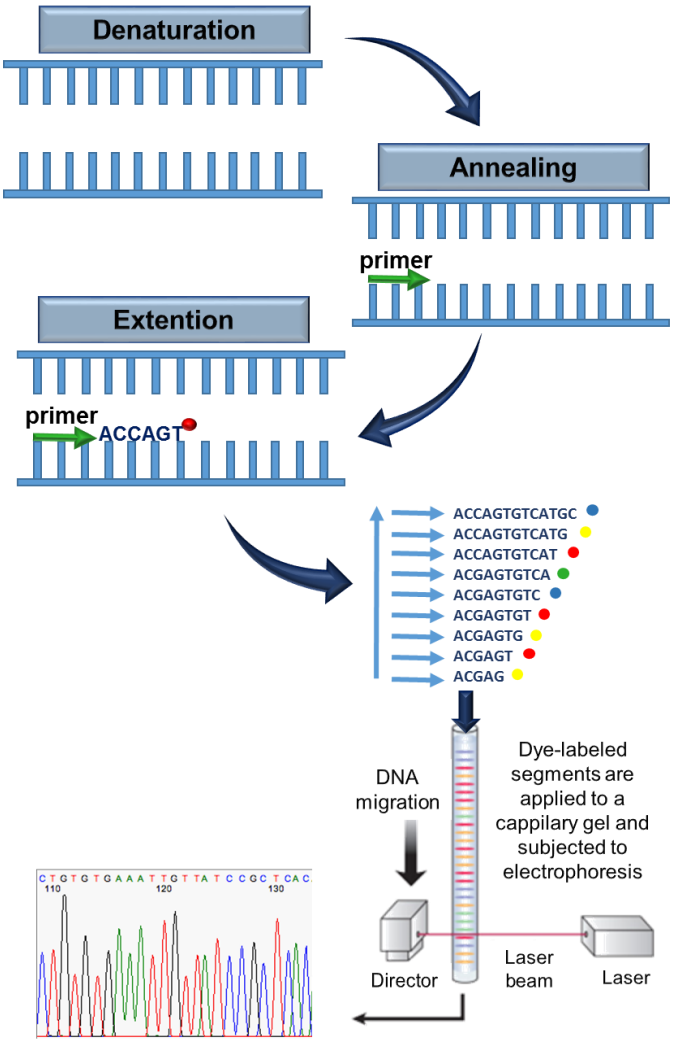


Figure 10- Strategy for automated DNA sequencing reactions. Adapted form[98].

The synthesis of truncated strands is achieved by the use of dideoxynucleotide triphosphates (ddNTPs), labelled with fluorophores[98]. The ddNTPs are identical to dNTP's, with the difference that they lack a 3 ' OH group[97]. When a ddNTP is incorporated into a DNA strand, no additional

nucleotides can be added, since there is no 3' OH group to form a phosphodiester bond with the next nucleotide, stopping the DNA synthesis[99].

Subsequently, the final nucleotide (ddNTP) in the distinct DNA strands will be detected in the reading sequence during capillary electrophoresis, which allows the separation, with high resolution, of fragments that differ in only one nucleotide. The fragments with different sizes are separated in the capillary, that contains a electrophoretic gel, and migrate, passing by a laser and a detector[98,99].

When the fragments move across the path of a laser beam, the fluorophores attached to ddNTPs are excited, emitting a fluorescence of four different wavelengths, and revealing the identity of the terminal nucleotide. So, the order of fluorescent ddNTPs, according to the fragments separation, allows to reveal the DNA sequence, generating a graph of peaks (electropherogram) corresponding to each nucleotide, with four specific colours for each base[99].

After amplification by PCR and confirmation that the DNA fragment of interest was amplified, the preparation of DNA samples for sequencing involves several steps: purification of initial PCR products, PCR sequencing and purification of PCR sequencing products. Then, the sequencing of purified products is executed by capillary electrophoresis, producing electropherograms that allows analysis of sequences. All this steps are schematize in Figure 13.

3.12.1 STEPS INVOLVED IN DNA SEQUENCING

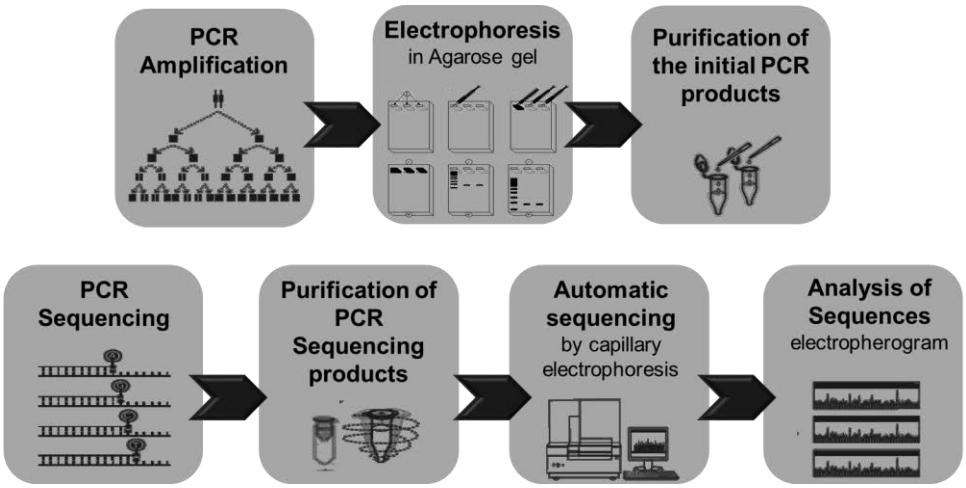


Figure 11- Scheme of the necessary steps for the DNA sequencing. Adapted from [100,101].

3.12.1.1 PURIFICATION OF PCR PRODUCTS

Before performing the PCR sequencing reactions it is necessary purify the PCR products, removing unincorporated primers and unincorporated nucleotides that may interfere with the following reactions.

The purification was carried out using enzymes in the PCR reaction or using GFX™ PCR DNA and Gel Band Purification KIT (Amersham Bioscience, GE Healthcare, EUA).

When no unspecific bands were observed in agarose gel, a combination of two enzymes were used: 0.5 µL (10 U) of Exo I (Exonuclease I; Thermo Scientific, EUA) and 1 µL (1 U) of FastAP (Thermosensitive Alkaline Phosphatase; Thermo Scientific, EUA) by each 10 µL of PCR product. The Exo I degrades single-stranded DNA and FastAP catalyzes the release of 5'- and 3'-phosphate groups from DNA, RNA, and nucleotides. A mix with the PCR reaction and enzymes was prepared and incubated at 37 ºC for 15 minutes, then the reaction was stopped by heat, inactivating the enzymes at 85 ºC for 15 minutes.

When unspecific bands were observed, the DNA band of interest was excised from the agarose gel and purified, using the purification KIT mentioned above.

The DNA band purification protocol is divided in four main steps (Figure 14): **(1)** sample capture; **(2)** sample binding; **(3)** wash and dry and **(4)** elution.

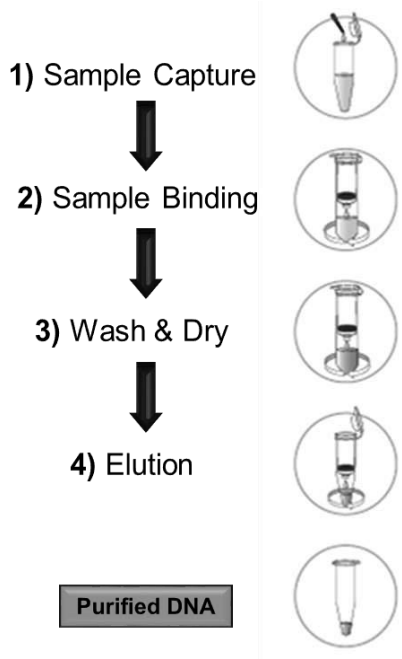


Figure 12- Representative scheme of the experimental procedure of GFX™ PCR DNA and Gel Band Purification protocol. Adapted from [102].

(1) Sample capture:

For the DNA sample capture, 500 µL of capture buffer was added to a tube containing the fragment excised. The mixture was agitated and incubated at 60ºC, until the agarose was completely dissolved.

(2) Sample binding:

Then, the capture buffer-sample mix was transferred to a GFX column (placed previously in a collector tube), incubated for 3 minutes at room temperature and centrifuged at 16,000×g (Eppendorf centrifuge 5415C) for 30 seconds. The liquid at the bottom of the collection tube was discarded and the GFX column was placed back in the collection tube. At this point of protocol, the DNA is bind to the silica membrane of GFX column.

(3) Wash and dry:

A wash buffer (500 µL) was added to GFX column, incubated for 3 min at room temperature and centrifuged at 16,000×g (Eppendorf centrifuge 5415C) for 30 seconds. The liquid at the bottom of the collection tube was discarded and the GFX column was transferred to a new tube.

(4) Elution:

In the center of GFX column, 10-60 µL of elution buffer was added and incubated for 5 minutes at room temperature. To recover the purified DNA, a centrifugation at 16,000 ×g (Eppendorf centrifuge 5415C) for 1 minute was performed.

In the end of protocol, the purified DNA was loaded in agarose gel for visual quantification of DNA intensity.

3.12.1.2 PCR SEQUENCING

The PCR sequencing reactions were performed using ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems, EUA), according to the manufacturer's protocol.

For each sample, the reaction was prepared in ice to a final volume of 20µL. The mixture used comprised the following compounds: 2 µL of 5x BigDye Sequencing Buffer (Applied Biosystems, EUA), 0.4 µL of forward or reverse primer (10 pmol/µL), 1.2 µl of Big Dye (Applied Biosystems, EUA), purified DNA (the quantity was variable, depending on the intensity observed in agarose gel) and sterile distilled water for add up to the final volume**.** The sequencing reactions

were performed in thermocycler (Biometra, Germany) and comprised the conditions described in Table VI:

Table VI- Protocol used for PCR sequencing reactions.

Parameter	Stage/step				
	Incubation	Cycling (25 Cycles)			Hold
		Denaturating	Annealing	Extension	
Temperature	96°C	96°C	50°C	60°C	4°C
Time	5 min	10 sec	5 sec	4 min	until ready to purify

Min, minutes; sec, seconds; ºC, temperature in Celsius.

3.12.1.3 PURIFICATION OF PCR SEQUENCING PRODUCTS

After performing sequencing reactions, it was also necessary purify the PCR sequencing products to remove unincorporated dye terminators (ddNTPs).

The purification was carried out using the Dye Terminator Removal Kit in a 96 well-format (Thermo Fisher Scientific™, EUA) or through precipitation with Ethanol/EDTA/acetate sodium, as recommended by the protocol of BigDye® Terminator v1.1 Cycle Sequencing Kit (Applied Biosystems,EUA).

When no unspecific bands were observed and the amplified DNA bands had a strong intensity, the Dye Terminator Removal Kit was used. The kit includes a 96-well plate, containing pre-hydrated gel-filtration material, that needed to be placed onto a wash plate and centrifuged for 3 minutes at 950xg (Eppendorf centrifuge 5810R) to remove the storage buffer. Then, the separation plate was placed onto a sample plate (the plate to be used in the automated sequencer) and the sequencing reaction products were added to the centre of the well and centrifuged for 3 minutes at 950xg (Eppendorf centrifuge 5810R). During centrifugation, the DNA samples were eluted into the sample plate.

When unspecific bands were observed or the intensity of the samples in agarose gel was weak, a precipitation with Ethanol/EDTA/acetate sodium was undertaken.

For each precipitation reaction a mix with 2 µL of acetate sodium (3 mM, pH 4.6), 2 µL of EDTA (125 mM, pH 8) and 50 µL of absolute ethanol was prepared. In a 1.5 mL tube with this mix, the sequencing reaction product was added and briefly agitated. Next, a centrifugation at 14,000 rpm (eppendorf

Centrifuge 5810R, rotor radius = 180 mm) during 40 minutes, at 4 ºC, was undertaken. After centrifugation, the flow-through was completely removed and 200 µL of 70% ethanol was added to the tube, that was briefly agitated and then centrifuged at 14,000 rpm (eppendorf Centrifuge 5810R, rotor radius = 180 mm) during 15 minutes, at 4 ºC. The supernatant was gently removed and the pellet was dried at 37 ºC during 5 minutes.

Finally the pellet was resuspended in 15 µL of formamide HI-DI (Applied Biosystems,EUA), agitated and transferred to a 96-well plate (Platemax, Axygen, EUA).

After the purification by both methods, the samples in the plate were denaturated at 95 ºC, for 10 minutes.

3.12.1.4 CAPILLARY ELECTROPHORESIS OF SEQUENCING PRODUCTS

After denaturating the samples, the plate was centrifuged at 1,200 rpm (eppendorf Centrifuge 5810R, rotor radius = 180 mm) during 2 minutes. Finally, the plate was set in the ABI Prism 3130 Genetic Analyzer (Applied Biosystems, EUA) automated sequencer and subjected to a capillary electrohoresis during 30 minutes for each sample.

The DNA sequences of each sample were obtained using the Sequencing Analysis 3.4.1 (Applied Biosystems, EUA) program that converts the fluorescent signals into electropherograms. Posteriorly, the analysis of the obtained DNA sequences was performed using the Variant reporter v.1.0 (Applied Biosystems, EUA) software.

CHAPTER 4

Results

4.1 CHARACTERIZATION OF SOMATIC MUTATIONS IN TUMOR SAMPLES FROM FAMILY 2

In family 2, thyroid cancer is likely to be hereditary, and a germline mutation segregating in the family is possibly associated with tumor initiation. In addition, somatic mutations, that are only present in the tumor, and are not inherited, may be involved in progression. To clarify this hypothesis, and the molecular mechanisms that can be involved in the progression of thyroid carcinomas in this family, hotspots of genes that are frequently mutated in sporadic thyroid tumors, were analysed by Sanger sequencing.

The mutational hotspots of *HRAS*, *NRAS*, *KRAS* and *BRAF* were sequenced in the tumor samples of patients: IV.2, IV.6, IV.8, V.3, V.8 and V.67. This analysis was not performed in patient III.10 because the tumor sample was unavailable.

All the familial tumors analysed presented somatic mutations in *HRAS*, except for patient's V.67 tumor that had a *BRAF* mutation (Table VII).

Table VII-Somatic mutations found in the thyroid carcinomas of patients from family 2.

Mutated Gene	Patient					
	IV.2	**IV.6**	**IV.8**	**V.3**	**V.8**	**V.67**
HRAS	c.34 **G**GC>**C**GC p.Gly12Arg transversion	c.37 **G**GT>**C**GT p.Gly13Arg transversion	c.37 **G**GT>**C**GT p.Gly13Arg transversion	c.35 G**G**C>G**A**C p.Gly12Asp transition	c.37 **G**GT>**C**GT p.Gly13Arg transversion	
BRAF						c.1799 G**T**G>G**A**G p.Val600Glu transversion

4.2 SELECTION OF GENOMIC VARIANTS DETECTED BY WES

To identify potentially pathogenic genetic variants that could confer susceptibility for thyroid cancer in family 2, blood samples of 6 patients with thyroid neoplasias (**IV.6:**ENDO_F2_1; **IV.8**:ENDO_F2_2; **V.3**:ENDO_F2_5; **V.8**: ENDO_F2_6; **III.10**: ENDO_F2_3 and **V.67**: ENDO_F2_4) were analysed by WES.

The WES analysis identified thousands of distinct genomic variants. To select the most interesting candidate genes, bioinformatics analyses were performed.

As explained in Chapter 3, four different and independent analyses were undertaken and, depending on the analysis, distinct groups of variants, shared by different patients, were examined.

The number of genomic variants called was above 300,000 for each sample, and the first set of criteria used to reduce this number was: the quality of the alignment (≥20), the quality of the genotype (≥20), the total number of high quality reads (≥20) and defined genotype. These criteria reduced the number of variants to less than 80,000 for each sample.

Subsequently, another set of criteria was also considered, and this consisted in the exclusion of variants that did not affect protein coding transcripts and that, due to its localization and predicted consequence (*e.g.* downstream/upstream variants and intronic variants not located in splicing regions), did not have a suggestive impact. Applying this second set of criteria the number of distinct variants in the four analyses was significantly reduced from thousands to hundreds, however, the number of genomic alterations was still very high.

To refine the bioinformatic analysis, several filters were subsequently applied: exclusion of variants with allele frequency superior to 1%, exclusion of synonymous variants, exclusion of variants that were also detected in WES analysis of unrelated familial colon and prostate cancers cases and, lastly, exclusion of homozygous variants.

The numbers of the distinct variants obtained in the different analyses, using the distinct sets of criteria and filters are summarize in Figure 15.

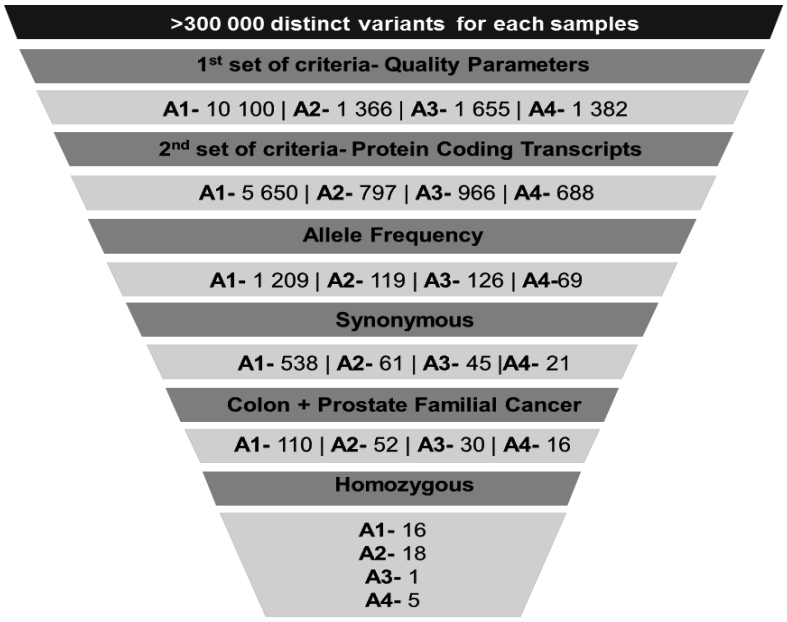


Figure 13- Scheme indicating the number of distinct variants obtained in the different analyses. A1 refers to analysis 1, **A2** to analysis 2, **A3** to analysis 3 and **A4** to analysis 4. Next to A1, A2, A3 and A4 is indicated the number of distinct variants present in the analysis, after applying the criteria or filters mentioned on top.

After bioinformatic analysis, a total of 40 distinct variants in 23 different genes remained for further studies. These candidate genes are listed in Table VIII, according to the analysis in which they were selected.

Table VIII- List of genes selected in analyses 1 to 4, after applying appropriate criteria and filters.

Candidate Genes selected after Bioinformatic Analysis			
Analysis 1	**Analysis 2**	**Analysis 3**	**Analysis 4**
LFNG	*ATHL1*	*DNAH17*	*ERICH6B*
GXYLT1 (2 var)	*MUC5AC*		*ZNF717* (2 var)
MTCH2	*TBC1D14*		*MUC3A* (2 var)
LILRB1	*STK33*		
MUC3A (10 var)	*RP1L1* (2 var)		
CDKN2AIP	*TRIO*		
	BLMH		
	HLA-DRB1 (2 var)		
	DNAJC28		
	ERICH6B		
	CASP8AP2		
	SPRY4		
	KCTD16		
	TARS2		
	TIGD4 (2 var)		

Var, variants.

4.3 STUDY OF THE CANDIDATE GENOMIC VARIANTS SELECTED IN FAMILY 2

The next step in this project was the study of the genomic variants selected by bioinformatic analysis in family 2, validating these variants with Sanger sequencing and then performing segregation studies of these variants in selected members of the family. Some variants could be prioritized for these studies, since there were less relevant variants (*e.g*. variants in *MUC* genes) that appeared hundreds of times in all analyses. The great majority of these latter variants were excluded by the filters, remaining a few that were most likely artefacts or polymorphisms (the information of allele frequency in the population was not available for these variants).

The genes chosen for Sanger sequencing validation are listed in Table IX, as well as the chromosome position of these genes, the corresponding transcripts and the variants under study.

Table IX- List of genes prioritized to perform Sanger sequencing validation in the members of family 2: bioinformatic analysis, chromosome location, transcript and the respective variants.

Gene	Chromosome Location	Transcript	Variant
Analysis 1			
LFNG	7p22.3	ENST00000402506	p.46ins.Glu
GXYLT1	12q12	ENST00000398675	p.Glu249Gly
			p.Glu249Lys
MTCH2	11p11.2	ENST00000302503	p.Lys300Arg
LILRB1	19q13.4	ENST00000396327	p.Arg59His
CDKN2AIP	4q35.1	ENST00000504169	p.241del. Ala
Analysis 2			
ATHL1	11p15.5	ENST00000409548	p.Arg555Trp
TBC1D14	4p16.1	ENST00000409757	p.Glu446Gln
STK33	11p15.4	ENST00000315204	p.Asp442His
TRIO	5p15.2	ENST00000344135	p.Ala19Thr
BLMH	17q11.2	ENST00000261714	p.Glu96Asp
ERICH6B	13q14.13	ENST00000298738	p.Glu138Gly
CASP8AP2	6q15	ENST00000551025	p.Thr211Ile
SPRY4	5q31.3	ENST00000344120	p.Thr234Met
KCTD16	5q31.3	ENST00000507359	p.Ala249Thr
TARS2	1q21.2	ENST00000369064	p.Pro387Leu
TIGD4	4q31.3	ENST00000304337	p.Ala432Thr
			p.Gln489Arg
Analysis 3			
DNAH17	17q25.3	ENST00000389840	p.Val3356Ile
Analysis 4			
ERICH6B	13q14.13	ENST00000298738	p.Tyr139His

4.3.1 VALIDATION OF GENOMIC VARIANTS BY SANGER SEQUENCING

The validation through Sanger sequencing of the variants in the genes mentioned in the last section, was performed in leucocyte DNA of six members from family 2, of whom five had been analysed through WES (**III.10**, **IV.6**, **V.3**, **V.8** and **V.67**) and one member was from a healthy branch (**III.14**).

A total of 20 variants in 17 different genes were sequenced and the results are schematized in Table X.

Table X- Validation by Sanger sequencing of genomic variants selected by bioinformatic analysis.

Genetic Alteration	Family Member III.10	Family Member IV.6	Family Member V.3	Family Member V.8	Family Member V.67	Family Member III.14	Variant Validation
Clinic Information	**TC + MNG**	**TC + MNG + CP**	**TC + CP**	**TC + MNG + CP**	**TC**	**U**	
Analysis 1							
LNFG p.Glu46ins	MUT (HET.)	MUT (HET.)	MUT (HET.)	MUT (HET.)	MUT (HET.)	MUT **(HOMO.)**	**Validated**
GXYLT1 p.Glu249Gly	WT	WT	WT	WT	WT	WT	Not Confirmed
GXYLT1 p.Glu249Lys	WT	WT	WT	WT	WT	WT	Not Confirmed
MTCH2 p.Lys300Arg	WT	WT	WT	WT	WT	WT	Not Confirmed
LILRB1 p.Arg59His	WT	WT	WT	WT	WT	WT	Not Confirmed
CDKN2AIP p.Ala241del	MUT (HET.)	MUT (HET.)	MUT (HET.)	MUT (HET.)	MUT (HET.)	MUT **(HOMO.)**	**Validated**
Analysis 2							
ATHL1 p.Arg555Trp	WT	MUT (HET.)	MUT (HET.)	MUT (HET.)	WT	WT	**Validated**
TBC1D14 p.Glu446Gln	WT	MUT (HET.)	MUT (HET.)	MUT (HET.)	WT	WT	**Validated**
STK33 p.Asp442His	WT	MUT (HET.)	MUT (HET.)	MUT (HET.)	WT	WT	**Validated**
TRIO p.Ala19Thr	WT	MUT (HET.)	MUT (HET.)	MUT (HET.)	WT	WT	**Validated**
BLMH p.Glu96Asp	WT	MUT (HET.)	MUT (HET.)	MUT (HET.)	MUT (HET.)	WT	**Validated**
ERICH6B p.Tyr139His	p.Glu135_ Leu140del (HET.)	p.Glu135_ Leu140del (**HOMO.)**	p.Glu135_ Leu140del (**HOMO.)**	p.Glu135_ Leu140del (**HOMO.)**	p.Glu135_ Leu140del (HET.)	p.Glu135_ Leu140del (HET.)	Not Confirmed
CASP8AP2 p.Thr211Ile	WT	MUT (HET.)	MUT (HET.)	MUT (HET.)	WT	WT	**Validated**
SPRY4 p.Thr234Met	WT	MUT (HET.)	MUT (HET.)	MUT (HET.)	WT	WT	**Validated**
KCTD16 p.Ala249Thr	WT	MUT (HET.)	MUT (HET.)	MUT (HET.)	WT	WT	**Validated**
TARS2 p.Pro387Leu	WT	MUT (HET.)	MUT (HET.)	MUT (HET.)	WT	WT	**Validated**
TIGD4 p.Ala432Thr	WT	MUT (HET.)	MUT (HET.)	MUT (HET.)	WT	WT	**Validated**
TIGD4 p.Gln489Arg	WT	MUT (HET.)	MUT (HET.)	MUT (HET.)	WT	WT	**Validated**
Analysis 3							
DNAH17 p.Val3356Ile	MUT (HET.)	The sample did not amplified in PCR	MUT (HET.)	MUT (HET.)	WT	MUT (HET.)	**Validated**
Analysis 4							
ERICH6B p.Glu138Gly	p.Glu135_ Leu140del (HET.)	p.Glu135_ Leu140del (**HOMO.)**	p.Glu135_ Leu140del (**HOMO.)**	p.Glu135_ Leu140del (**HOMO.)**	p.Glu135_ Leu140del (HET.)	p.Glu135_ Leu140del (HET.)	Not Confirmed

TC, thyroid cancer; **MNG**, multinodular goiter; **CP**, colon polyps; **U**, unaffected; **WT**, Wild Type; MUT (HET.), Mutation in heterozygosity; **MUT (HOMO.)**, Mutation in Homozygosity.

The variants in *GXYLT1, MTCH2* and *LILRB1* genes were not detected by Sanger sequencing, and all family members were wild type, thus, these variants were probably WES artefacts.

Although the variants in *LFNG* and *CDKN2AIP* were confirmed by Sanger sequencing, they were detected in homozygosity in a healthy family member, therefore, since an autosomal dominant pattern of inheritance was assumed in this family, they were excluded in the subsequent studies.

The two missense variants in *ERICH6B* gene were not confirmed, but instead, a deletion of 18 bp, covering the location of the two missense variants, was observed. The presence of this deletion in the samples that had been analysed by WES was subsequently confirmed in the analysis with the IGV (Integrative Genomics Viewer) tool. Although this unexpected deletion in *ERICH6B* was observed in homozygosity, it was also screened in other family members to further confirm that it was not an artefact. The sequencing of this region was undertaken for the remaining members of nuclear family (**IV.4**- heterozygous mutation; **IV.8**- homozygous mutation; **IV.2**- heterozygous mutation) and in two daughters of III.10 patient (**IV.23**- heterozygous mutation; **IV.33**- wild type). This particular alteration seemed to be frequent in this family, but it had not been reported in the Ensemble database. As these results continued to be intriguing, the region was also sequenced in six leucocyte DNAs from healthy controls to clarify if this deletion was common in our population. Among the six samples, two presented the deletion in homozygosity, two in heterozygosity and two were wild type. Thus, this alteration seems to be frequent, at least in our population, both in heterozygosity and homozygosity, and for that reason it was excluded from further studies.

The missense variants in the remaining genes (*ATHL1, BLMH, CASP8AP2, DNAH17, KCTD16, SPRY4, STK33, TARS2, TBC1D14, TIGD4* and *TRIO*) were confirmed by Sanger sequencing, and are listed in Figure 16, with examples of the electropherograms obtained in wild type patients and in those with the alteration. Segregation studies were further performed for these variants.

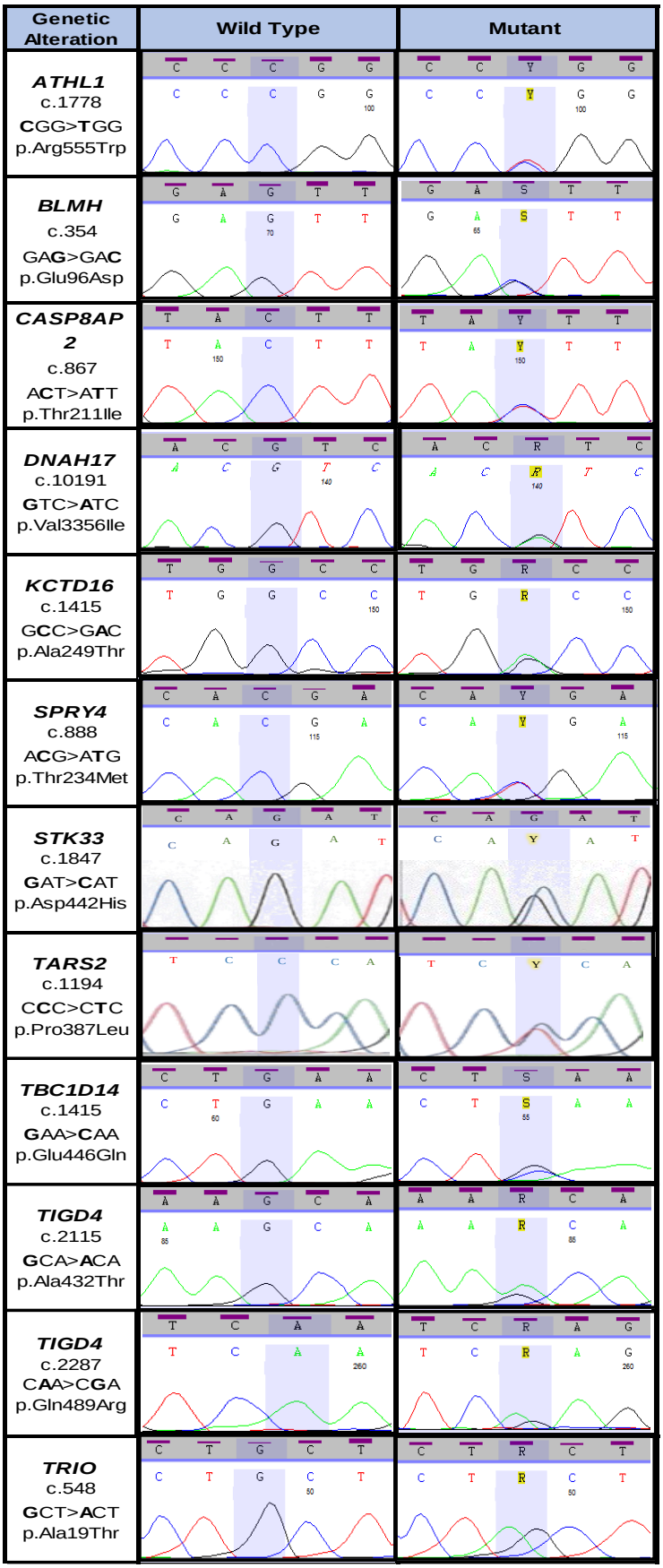


Figure 14- Variants selected for segregation studies in family 2. Examples of the electropherograms obtained in wild type patients and in those with the alteration are shown.

4.3.2 SEGREGATION STUDIES OF GENOMIC VARIANTS VALIDATED BY SANGER SEQUENCING

For the segregation studies, a total of 15 members of family 2 were studied, 5 of them were already analysed in the validation phase.

The sequencing of the genes listed in Figure 16 was undertaken in the leucocyte DNAs from: the nuclear family members with thyroid tumors and/or with MNG (**IV.2**, **IV.4**, **IV.6**, **IV.8**, **V.3** and **V.8**), the female affected with thyroid carcinoma that could be a phenocopy (**III.10**) and her three daughters (**IV.23**, **IV.25**, **IV.33**), in two members from two different family branches with no cases of thyroid cancer (**III.12**, **III.14**), in the male affected with thyroid cancer that could be the other phenocopy (**V.67**) and in two aunts from this patient who had MNG (**IV.60**, **IV.63**).

The results are schematized in Table XI. Four disease segregation models were established, and each variant was assigned to one of these models. This categorization was helpful in the selection of the most relevant variants for future studies. These models are described in Figure 17.

The simplified family pedigrees, showing the segregation of the variants detected in these candidate genes, are shown in the Appendix.

Segregation Models		
Clinic Information	**Thyroid Tumor:** Non medullary thyroid cancer **Colon:** lesions in colon **MNG**: Multinodular Goiter	
MODEL:	**Thyroid Tumor + Colon (Nuclear Family)**	Segregates with Thyroid Tumor and colon lesions in nuclear family only
	Thyroid Tumor + MNG + Colon (Nuclear Family)	Segregates with Thyroid Tumor, MNG and colon lesions in nuclear family only
	Thyroid Tumor + MNG + Colon (Nuclear Family) and MNG (complete family)	Segregates with Thyroid Tumor, MNG and colon lesions in nuclear family and in some MNG in the complete family
	Inconclusive:	Segregates with an affected and unaffected thyroid tumor branches of the family

Figure 15- Disease segregation models established to categorize the most relevant genomic variants identified in family 2.

Table XI- Segregation of variants validated by Sanger sequencing in family 2.

Genetic Alteration	Clinic Information	ATHL1 p.Arg555Trp	BLMH p.Glu96Asp	CASP8AP2 p.Thr211Ile	DNAH17 p.Val3356Ile	KCTD16 p.Ala249Thr	SPRY4 p.Thr234Met	STK33 p.Asp442His	TARS2 p.Pro387Leu	TBC1D14 p.Glu446Gln	TIGD4 p.Ala432Thr	TIGD4 p.Gln489Arg	TRIO p.Ala19Thr
Family Member IV.2	TC + MNG + CP	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT
Family Member IV.4	MNG	WT	WT	MUT	-	WT	WT	WT	WT	MUT	WT	WT	MUT
Family Member IV.6	TC + MNG + CP	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT
Family Member IV.8	TC + CP	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT
Family Member V.3	TC + CP	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT	MUT
Family Member V.8	TC + MNG + CP	WT	MUT	MUT	MUT	WT	WT	WT	WT	WT	WT	WT	WT
Family Member III.10	TC + MNG	WT	WT	WT	MUT	WT	WT	WT	WT	WT	WT	WT	WT
Family Member IV.23	U	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT
Family Member IV.25	MNG	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT
Family Member IV.33	MNG	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT
Family Member III.12	U	WT	MUT	WT	MUT	WT	WT	WT	WT	WT	WT	WT	WT
Family Member III.14	U	WT	MUT	WT	MUT	WT	WT	WT	WT	WT	WT	WT	WT
Family Member V.67	TC	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT	WT
Family Member IV.60	MNG	WT	WT	WT	-	WT	WT	WT	WT	MUT	WT	WT	WT
Family Member IV.63	MNG	WT	WT	WT	MUT	WT	WT	WT	WT	MUT	WT	WT	WT
Segregation Model		Thyroid Tumor + Colon (Nuclear Family)	Inconclusive	Thyroid Tumor + MNG + Colon (Nuclear Family)	Inconclusive	Thyroid Tumor + Colon (Nuclear Family)	Thyroid Tumor + Colon (Nuclear Family)	Thyroid Tumor + Colon (Nuclear Family)	Thyroid Tumor + Colon (Nuclear Family)	Thyroid Tumor + MNG + Colon (Nuclear Family) and MNG (complete family)	Thyroid Tumor + Colon (Nuclear Family)	Thyroid Tumor + Colon (Nuclear Family)	Thyroid Tumor + MNG + Colon (Nuclear Family)

TC, thyroid cancer; **MNG,** multinodular goiter; **CP,** colon polyps; **U**, unaffected; WT, wild type; **MUT,** carrier of the genetic variant; **-** ,the sample did not amplify in PCR.

For the great majority of the genes (*ATHL1, KCTD16, SPRY4, STK33, TARS2* and *TIGD4*) the variants in heterozygosity were observed only in family nuclear members with thyroid carcinomas (some with MNG), that also had colon lesions (**IV.2**, **IV.6**, **IV.8**, **V.3** and **V.8**). In two other genes (*CASP8AP2* and *TRIO*) the heterozygous variants were observed in the nuclear family members with thyroid carcinomas (some also with MNG) and in the nuclear family member that only presented MNG (**IV.4**). The heterozygous variant in *TBC1D14* was observed

in nuclear family members with thyroid tumors (some also with MNG), in the nuclear family member that only presented MNG and in two other members (**IV.60** and **IV.63**) with MNG from another branch of family 2. The variants in the remaining two genes (*BLMH* and *DNAH17*) were found in some members with thyroid tumors and/or with MNG and in unaffected members of apparently healthy branches (**III.12** and **III.14**), and were categorized with an inconclusive model of segregation, being excluded from the subsequent studies and analyses performed in this project.

4.4 STUDY OF GENOMIC VARIANTS IN TUMORS SAMPLES

The variants which presented a segregation model in family 2 that was not inconclusive (*ATHL1, CASP8AP2, KCTD16, SPRY4, STK33, TARS2, TBC1D14, TIGD4* and *TRIO*), were sequenced in the DNA from familial tumor samples to explore if any of these variants presented loss of heterozigoty (LOH) in those tumors, a genetic event very common in the mechanism of tumor suppressor gene inactivation.

At least, one tumor of the six tumors available was analysed for each variant and, in this explorative analysis, no pattern suggestive of LOH was found in any gene.

4.5 *IN SILICO* ANALYSES

The segregation studies in family 2 revealed very similar segregation patterns between the different variants. One of the strategies used to help in the selection of the most promising genetic variants was the accomplishment of several *in silico* analyses, using different softwares and databases.

4.5.1 IMPACT PREDICTION

Nowadays numerous softwares are available to virtually predict whether a particular genetic alteration has (or not) a damaging impact in the encoded protein function. In this work, four different tools (SIFT, PolyPhen, Mutation Taster and Mutation Assessor) were used to predict the impact of the variants for functional studies, detected in nine selected candidate genes *(ATHL1, CASP8AP2, KCTD16, SPRY4, STK33, TARS2, TBC1D14, TIGD4* and *TRIO*), in

the function of the proteins encoded by these genes. The results are indicated in Table XII.

The impact prediction for *TRIO* and *TIGD4* variants was benign with all tools, and was always damaging for *CASP8AP2, KCTD16*, *SPRY4* and *TBC1D14* variants. The results for the remaining genes differed between softwares. The *ATHL1* variant was predicted to benign by SIFT, PolyPhen and Mutation Taster, but damaging by Mutation Assessor. The *STK33* variant was predicted to be benign by SIFT, Mutation Taster and Mutation Assessor, but damaging by PolyPhen. The *TARS2* variant was predicted to be benign by PolyPhen and Mutation Taster, but damaging by SIFT. The information of Mutation Assessor for *TASR2* and *SPRY4* variants, as well as the information of Mutation Taster for *CASP8AP2* was not possible to obtain, since the tools for these variants presented an error message.

Table XII- Impact prediction of SIFT, PolyPhen, Mutation Taster and Mutation Assessor tools for the variants detected in the nine candidate genes.

Genetic Alteration	SIFT	PolyPhen	Mutation Taster	Mutation Assessor
ATHL1 p.Arg34Trp	tolerated (0.2)	benign (0.008)	polymorphism	medium
CASP8AP2 p.Thr211Ile	deleterious (0)	possibly damaging (0.499)	Information not available	medium
KCTD16 p.Ala249Thr	deleterious (0.01)	probably damaging (0.997)	disease causing	medium
SPRY4 p.Thr234Met	deleterious (0)	probably damaging (0.982)	disease causing	Information not available
STK33 p.Asp442His	tolerated (0.07)	possibly damaging (0.711)	polymorphism	neutral
TARS2 p.Pro257Leu	deleterious (0.05)	benign (0.001)	polymorphism	Information not available
TBC1D14 p.Glu106Gln	deleterious (0.04)	possibly damaging (0.653)	disease causing	medium
TIGD4 p.Ala432Thr p.Gln489Arg	tolerated (0.29) tolerated (0.61)	benign (0.193) benign (0.002)	polymorphism polymorphism	neutral low
TRIO p.Ala19Thr	tolerated (0,2)	benign (0,034)	polymorphism	neutral

SIFT score: 0 to 1 (≤ 0.05: deleterious and >0.05: tolerated). **PolyPhen score:** 0 to 1 (0-0.15: benign, 0.15-0.85: possibly damaging and 0.85-1: probably damaging). **Mutation Taster score**: polymorphism for benign and disease causing for damaging. **Mutation Assessor score**: low and neutral for benign; medium and high for damaging. **Information not available**: the software presented an error message and did not supply any prediction for the variant.

4.5.2 PROTEIN EXPRESSION

The protein expression profile of the candidate genes in normal thyroid tissue was evaluated using the information available in the Human Protein Atlas (HPA) database. The expression is categorized as: not detected, low, medium or high, and the results for the nine candidate genes are graphically represented in Figure 18.

The protein expression of KCTD16 gene was not detected by immunohistochemistry in thyroid tissues, and the protein expression of CASP8AP2 is not yet available in the database. The protein expression for the remaining genes is indicated as low for TIGD4, medium for ATHL1, SPRY4, STK33 and TBC1D14, and high for TARS2 and TRIO.

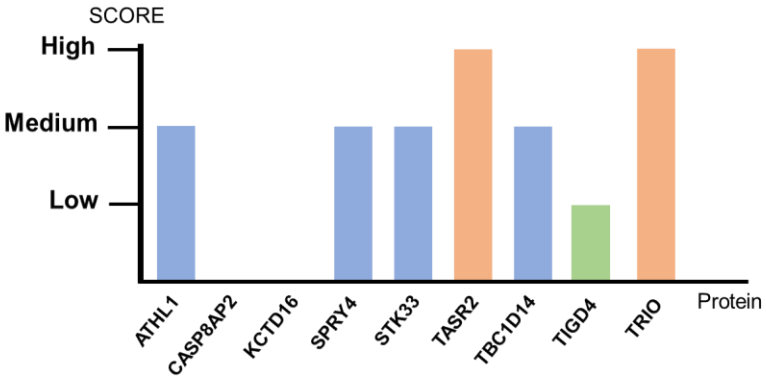


Figure 16- Protein expression level of candidate genes in normal thyroid tissue. The y axis represents the protein expression score, categorized as not detected, low, medium and high. The x axis indicates the candidate genes analysed. The green bar represents the gene with a low expression in thyroid gland; the blue bars represent the genes with a medium expression in thyroid tissue; the orange bars represent the genes with a high expression in thyroid tissue; and the genes which no protein expression was detected, or the information is not yet available for thyroid tissue, do not have any bars associated.

4.5.3 RNA EXPRESSION

The mRNA expression profile of the candidate genes in distinct normal tissues, particularly in thyroid tissue, was also evaluated using the information of HPA dataset. The mRNA expression levels for each gene are represented in Figures 19 to 27, according to the thresholds mentioned in Chapter 3, and the expression was categorized as: not detected, low, medium or high.

The *ATHL1* (Figure 19) and *CASP8AP2* (Figure 20) genes are expressed in almost all tissues and in both cases the expression in the thyroid gland is low.

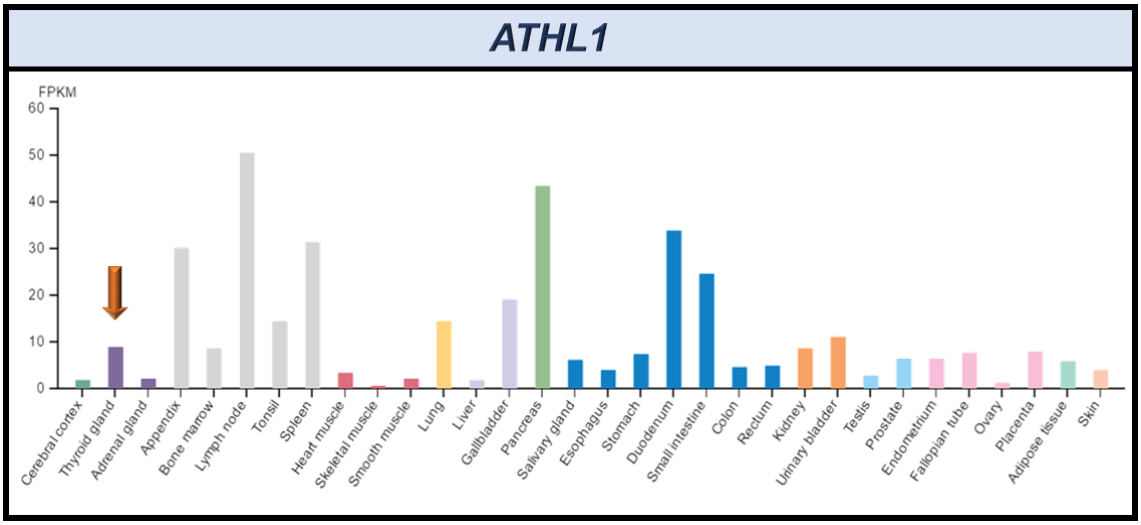


Figure 17- *ATHL1* mRNA expression level. The y axis represents the expression level in FPKM (fragments per kilobase gene model and million reads) and the x axis represents the tissues evaluated; The orange arrow points out the thyroid gland tissue. This Figure was adapted from[89].

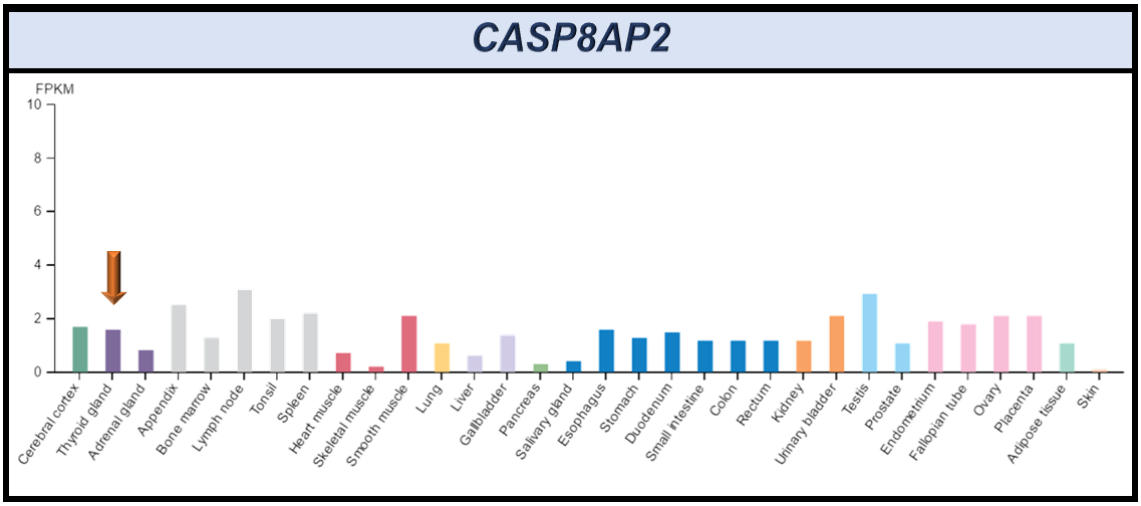


Figure 18- *CASP8AP2* mRNA expression level. The y axis represents the expression level in FPKM (fragments per kilobase gene model and million reads) and the x axis represents the tissues evaluated; The orange arrow points out the thyroid gland tissue. This Figure was adapted from[89].

The *KCTD16* gene (Figure 21), is principally expressed in the cerebral cortex. Regarding thyroid gland, the expression of *KCTD16* is categorized as not detected (FPKM<0.5).

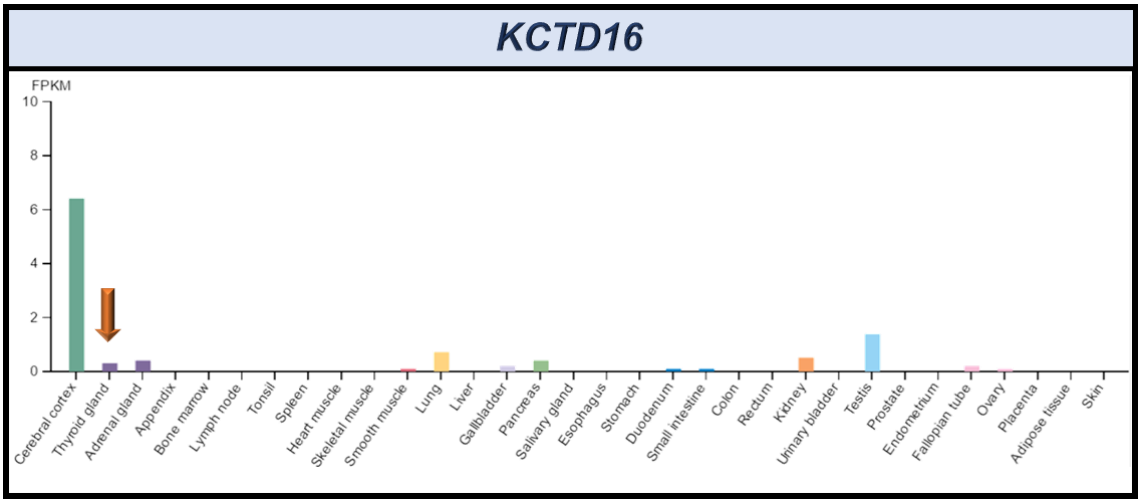


Figure 19- *KCTD16* mRNA expression level. The y axis represents the expression level in FPKM (fragments per kilobase gene model and million reads) and the x axis represents the tissues evaluated; The orange arrow points out the thyroid gland tissue. This Figure was adapted from[89].

The *SPRY4* gene (Figure 22) is expressed in all tissues assessed, having the highest expression in adipose tissue and a medium expression in the thyroid gland.

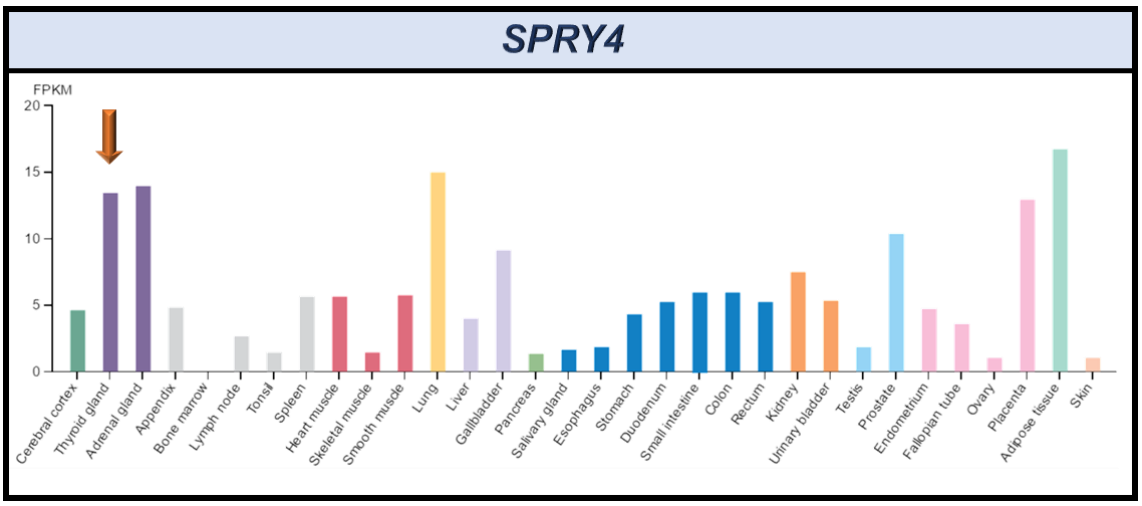


Figure 20- *SPRY4* mRNA expression level. The y axis represents the expression level in FPKM (fragments per kilobase gene model and million reads) and the x axis represents the tissues evaluated; The orange arrow points out the thyroid gland tissue. This Figure was adaptedfrom[89].

The *STK33* gene (Figure 23) is principally expressed in the testis and fallopian tube, and presents a medium mRNA expression in thyroid gland.

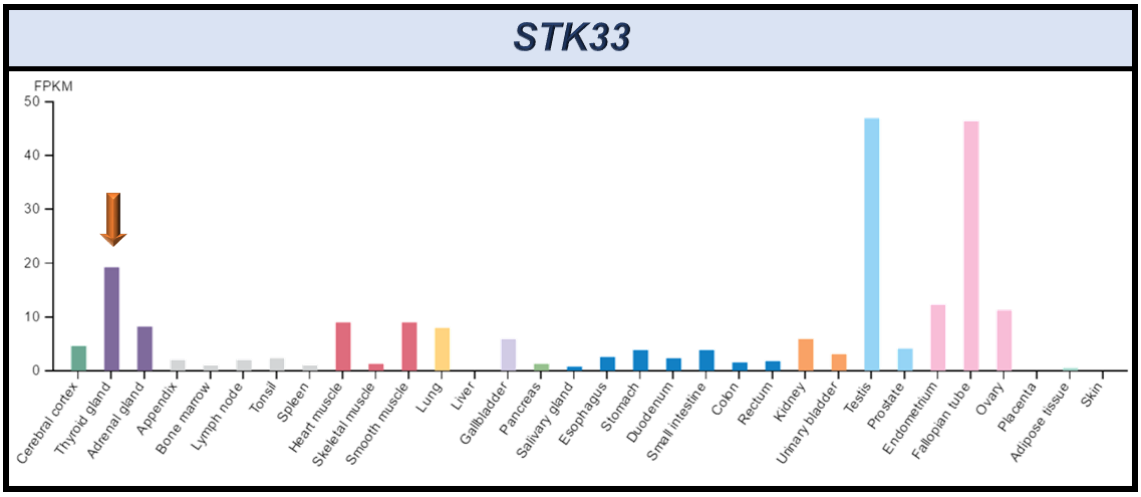


Figure 21- *STK33* mRNA expression level. The y axis represents the expression level in FPKM (fragments per kilobase gene model and million reads) and the x axis represents the tissues evaluated; The orange arrow points out the thyroid gland tissue. This Figure was adapted from[89].

The *TARS2* (Figure 24) and *TBC1D14* (Figure 25) genes are expressed in all tissues assessed, with levels very similar in numerous tissues. These genes present a medium expression in the thyroid gland.

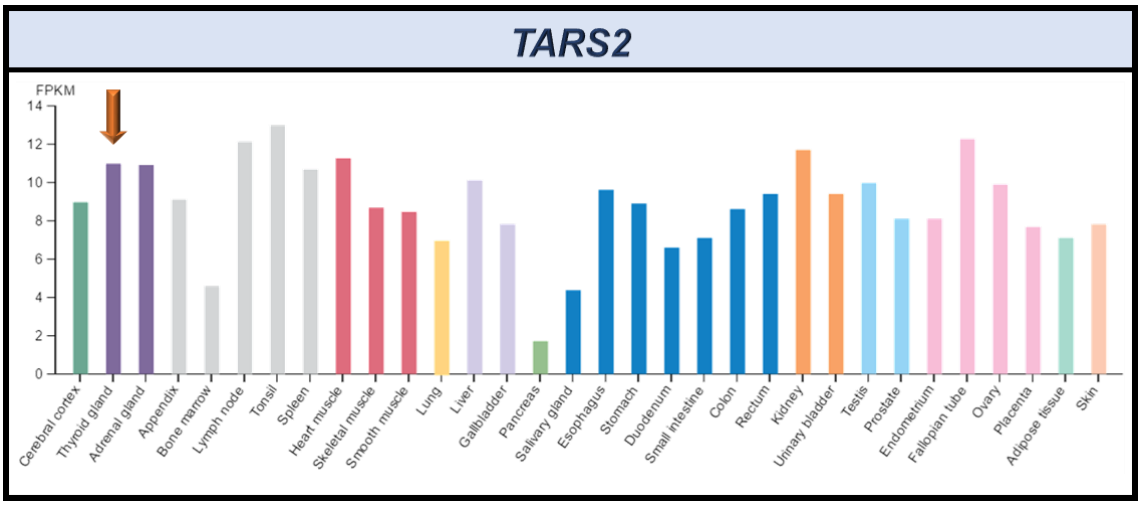


Figure 22- *TARS2* mRNA expression level. The y axis represents the expression level in FPKM (fragments per kilobase gene model and million reads) and the x axis represents the tissues evaluated; The orange arrow points out the thyroid gland tissue. This Figure was adapted from[89].

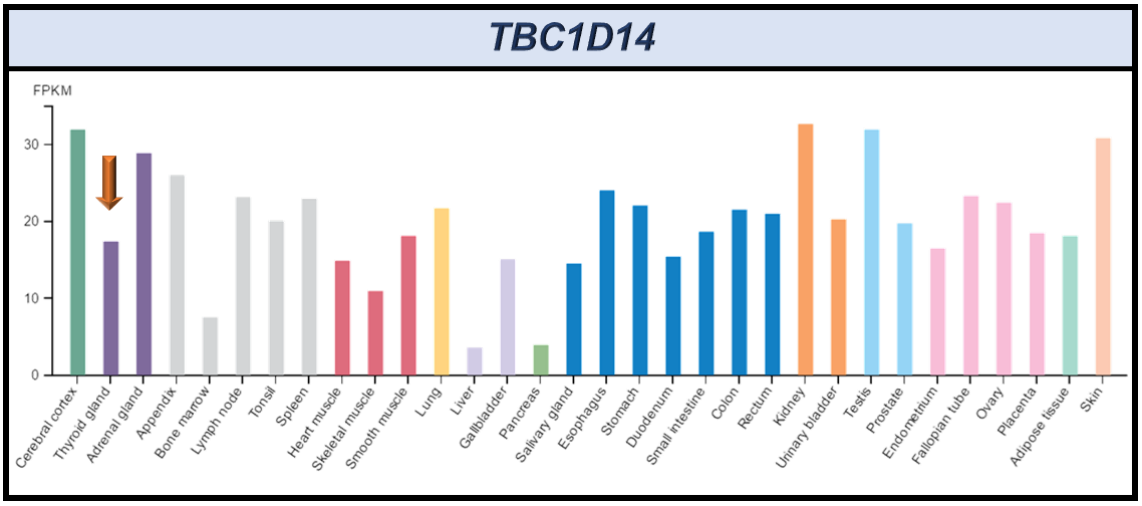


Figure 23- *TBC1D14* mRNA expression level. The y axis represents the expression level in FPKM (fragments per kilobase gene model and million reads) and the x axis represents the tissues evaluated; The orange arrow points out the thyroid gland tissue. This Figure was adapted from[89].

TIGD4 (Figure 26) mRNA expression is low in almost all tissues assessed, including thyroid gland (0.5 FPKM).

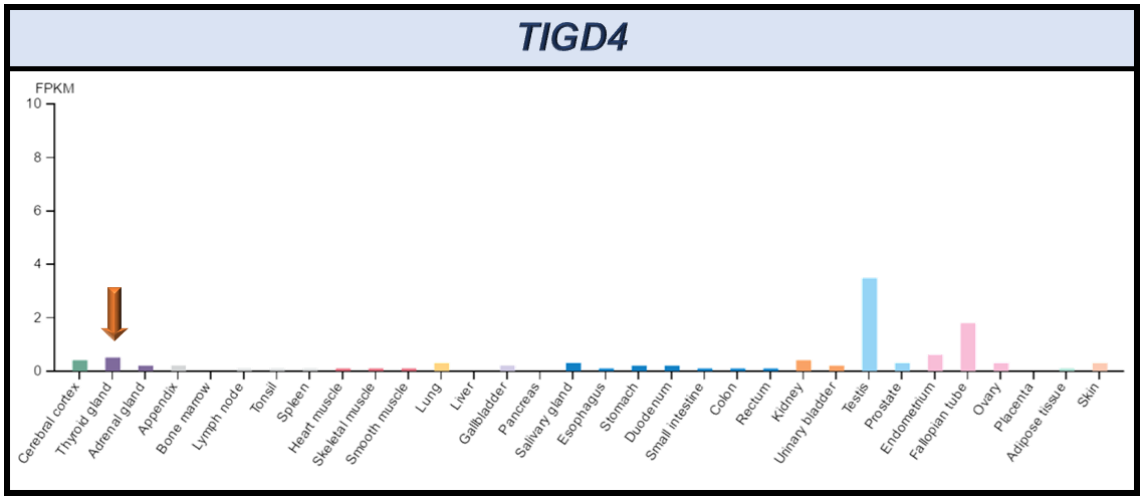


Figure 24- *TIGD4* mRNA expression level. The y axis represents the expression level in FPKM (fragments per kilobase gene model and million reads) and the x axis represents the tissues evaluated; The orange arrow points out the thyroid gland tissue. This Figure was adapted from[89].

Finally, the *TRIO* gene (Figure 27) is expressed in all tissues assessed, with levels very similar in several tissues. This gene presents a low expression in the thyroid gland.

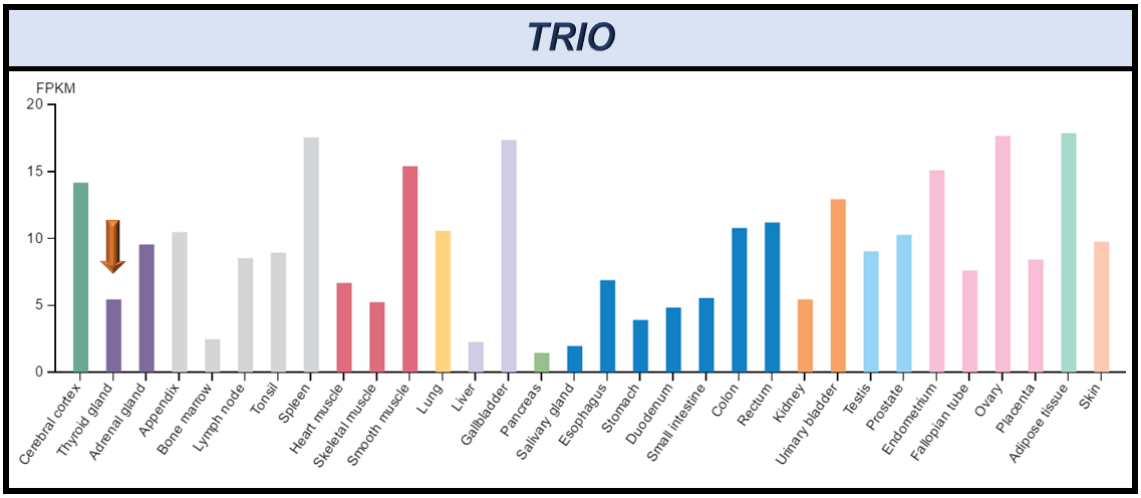


Figure 25- *TRIO* mRNA expression level. The y axis represents the expression level in FPKM (fragments per kilobase gene model and million reads) and the x axis represents the tissues evaluated; The orange arrow points out the thyroid gland tissue. This Figure was adapted from[89].

4.6 ADDITIONAL DATA ON THE CANDIDATE GENES

An examination of the literature information available for the nine candidate genes regarding the function and the pathways where these genes may be involved, was performed to evaluate which variants could be more relevant, that is, most likely to be involved in tumorigenesis of thyroid carcinomas in the family 2.

- ***ATHL1***

Regarding the *ATHL1* [Acid Trehalase-Like 1 (yeast)] gene very few is known. It seems that this gene encodes PGGHG (protein-glucosylgalactosylhydroxylysine glucosidase), an enzyme that specifically hydrolyzes glucose residues of collagen and collagen-like proteins[103]. Thus, *ATHL1* encoded protein has a hydrolase activity, acting on glycosyl bonds[90,92,93].

The predicted functional partners are *B4GALNT4* (*galactosaminyl transferase 4*), *TREH* (*Trehalase*) and *IFITM10* (Interferon induced transmembrane protein 10)[94]. To our knowledge, there are no clear evidences of the involvement of this gene in cancer.

- ***CASP8AP2***

CASP8AP2 (*Caspase 8 Associated Protein 2*) also known as *FLASH* (*FLICE-Associated Huge Protein*) is a gene that has been implicated in many different cellular processes[90,91,104].

Originally, the protein encoded by this gene was described as pro-apoptotic, needed for activation of caspase-8, via the extrinsic pathway linked to the Fas receptor and thereafter, was also linked to apoptosis[105,106]. Additionally, it has also been associated to TNF (Tumor Necrosis Factor) receptor-induced activation of the transcription factor *NFκB (nuclear factor kappa B)*, via the association with the signalling molecule TRAF2 (TNF receptor-associated factor 2) and Caspase-8[107], and as a co-activator of the c-Myb oncogene[108,109].

Subsequently, a number of studies described that *CASP8AP2* is involved in the control of cell cycle progression, being essential for histone gene pre-mRNA processing and histone gene expression[110–114]. It was showed that the depletion of *CASP8AP2* leads to accumulation of cells in S phase, probably through histone depletion[112].

This gene was also identified as a predictive marker for poor prognosis in acute lymphoblastic leukemia[115,116].

Finally, in recent studies, it was reported that the *CASP8AP2* gene plays an important role in initial embryogenesis, and the knockout of this gene in mice was lethal, in early embryogenesis[117,118].

- ***KCTD16***

The *KCTD16* (*Potassium Channel Tetramerization Domain Containing 16*) is a member of a subfamily that encodes KCTD proteins with voltage-gated ion channel activity[90,91,93]. This gene, along with other genes of KCTD family, were recently identified as auxiliary of GABA-B receptor subunits, which play an important role in regulating membrane excitability and synaptic transmission in the brain. It increases the agonist potency and alters the G-protein signalling of the receptors[119–121]. To our knowledge, there are no clear evidences of the involvement of this gene in cancer.

- ***SPRY4***

The *SPRY4* [*Sprouty Homolog 4 (Drosophila)*] gene encodes a member of a family of cysteine- and proline-rich proteins and is an inhibitor of EGFR (epidermal growth factor receptor)-transduced MAPK signalling pathways[90,91,122].
This gene seems to impair the formation of GTP-RAS[123], inhibiting the VEGF (vascular endothelial growth factor)-induced, RAS-independent (but not RAS-dependent) activation of Raf-1[122,124]. Additionally, it was reported that *SPRY4* may function as negative regulator for several types of PCL (phospholipase C)-dependent signalling pathways[125].
Moreover, several studies have described *SPRY4* as an important player in cancer [122].
This gene was reported to have a putative role in Wnt7A/Fzd9 tumor suppressing pathway, as a downstream effector through *PPARγ*, inhibiting transformed cell growth, migration and invasion in non-small lung cancer[122,126,127]. The influence of this gene in cell migration and cancer progression was also described for numerous other types of cancer such as melanoma, endometrial, ovarian, prostate and breast cancer[122,128–132].
Finally, genome wide studies have also suggested the involvement of *SPRY4* in the predisposition to testicular germ cell cancer[133].

- ***STK33***

STK33 (Serine/Threonine Kinase 33) encoded protein is a member of the CAMK (calcium/calmodulin dependent protein kinase) family, a group of kinases involved very often in fundamental cellular processes, such as DNA replication, metabolic pathways, cell growth, differentiation, proliferation and cell death[90,93,134]
It was reported that *STK33* may play an important role in the dynamics of the intermediate filament cytoskeleton by phosphorylating the intermediate filament protein vimentin[135,136].
In addition, recent studies described the involvement of this gene in the tumorigenesis and cell progression of hepatocellular carcinoma[137], hypopharyngeal squamous cell carcinoma[138] and lung cancer[139], possibly through MAPK signalling pathway activation. Overall, the *STK33* gene is

overexpressed in tumors, and is correlated with poor prognosis, and it has been hypothesized that this gene could be a new oncogene[137–139].

- ***TARS2***

The *TARS2 (Threonyl-TRNA Synthetase 2, Mitochondrial)* gene encodes mitochondrial aminoacyl-tRNA synthetase, a protein involved in mitochondrial translation[90,91].

Mutations in this gene have been associated with combined oxidative phosphorylation deficiency-21, a mitochondrial disorder characterized by a lethal encephalomyopathy. One study, identified by WES a heterozygous missense mutation (p.Pro282Leu) and a nucleotide change in position +3 of intron 6, and both segregated with the disorder in a family[140]. Analysis of patient cells showed decreased *TARS2* mRNA and protein compared to controls, and the defect in mitochondrial respiration in patient immortalized cells was rescued by expression of wildtype *TARS2*[140]. A recent study confirmed *in vitro* and *in vivo* the pathogenicity of p.Pro282Leu variant, which revealed that this mutation induces decrease of threonyl activation, aminoacylation and proofreading activities, and a change in the protein structure and/or stability that might cause reduced catalytic efficiency[141]. To our knowledge, there are no clear evidences of the involvement of this gene in cancer.

- ***TBC1D14***

TBC1D14 (*TBC1 Domain Family, Member 14*) gene is a member of TBC genes, which function as been related with protein kinase binding and GTPase activator activity[90,93]. This gene was described as a negative regulator of starvation – induced autophagosome formation, acting in the TFNR–Rab11 recycling pathway and activating Rab11[142] by binding to the TRAPP (transport protein particle) complex[143].

- ***TIGD4***

The *TIGD4 (tigger transposable element derived 4)* gene encodes a protein that belongs to the tigger subfamily of the pogo superfamily of DNA-mediated transposons in humans[90]. These proteins are related to DNA transposons found

in fungi and nematodes[90]. There is no information available related with this gene function.

The predicted functional partners are *INCENP* (*inner centromere protein antigens*), a component of the CPC (chromosomal passenger complex) that acts as a key regulator of mitosis; *CENPA* (*centromere protein A*) a protein required for recruitment and assembly of kinetochore proteins, mitotic progression and chromosome segregation; and *AURK* (*aurora kinases*), mitotic serine/threonine kinases that contributes to the regulation of cell cycle progression[94].

- ***TRIO***

TRIO (*Trio Rho Guanine Nucleotide Exchange Factor*), as the name suggests, is a gene that promotes the exchange of GDP by GTP and has been widely studied over the years[90,144]. It is considered a complex molecule with several protein-protein interaction domains and it seems to integrate numerous signalling pathways, mainly through Rac1 and RhoG activation[144].

TRIO seems to have a major role in cell motility, axon outgrowth and guidance, as well a prominent role in the development of the nervous system in mammals, since *TRIO* knockout is lethal in embryogenesis and the mice presented an aberrant organization in different brain regions and defects in secondary myogenesis[144,145].

Moreover, it has been suggested that *TRIO* plays an important role in coordinating cell-matrix and cytoskeletal rearrangements, necessary for cell migration and cell growth, interacting with adhesive receptors especially cadherins, through Rac1[144,146].

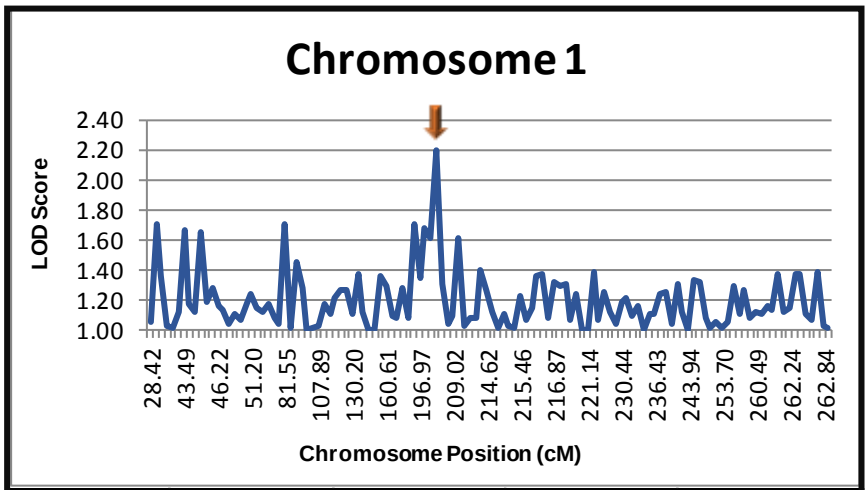
Additionally, it has also been described the involvement of this gene in the TNF-α inflammatory pathway and in the $G_{\alpha q}$ (G_q alpha protein)-elicited mitogenic response, by acting through Rho GTPase-mediated activation of p38 and Jun kinase[144].

Finally, *TRIO* seems to be upregulated in different types of tumors, including breast, urinary bladder, head and neck, glioblastoma and more recently, was associated with colorectal cancer progression, being indicated as one of the downstream effectors of Notch signalling activation, a pathway relevant in colorectal cancer[144,147].

4.7 LINKAGE ANALYSIS

Another strategy used to facilitate the selection of the most likely causative variant(s) for FNMTC in this family, was the combination of WES results with linkage analysis.

A two-point parametric linkage analysis was performed for all chromosomes (genome-wide) and LOD scores equal or superior to 2.0 were accepted as the threshold for suggestive linkage. LOD scores equal or superior to 1.5 were also considered suggestive when the corresponding marker was closely located to a candidate gene *locus*. The results of the genome-wide linkage analysis are only shown, for the chromosome *loci* where the candidate genes were located (chromosomes 1, 4, 5, 6 and 11), and when the LOD scores obtained were ≥ 1 (shown below as graphics). The results for the remaining chromosomes are shown in the Appendix.

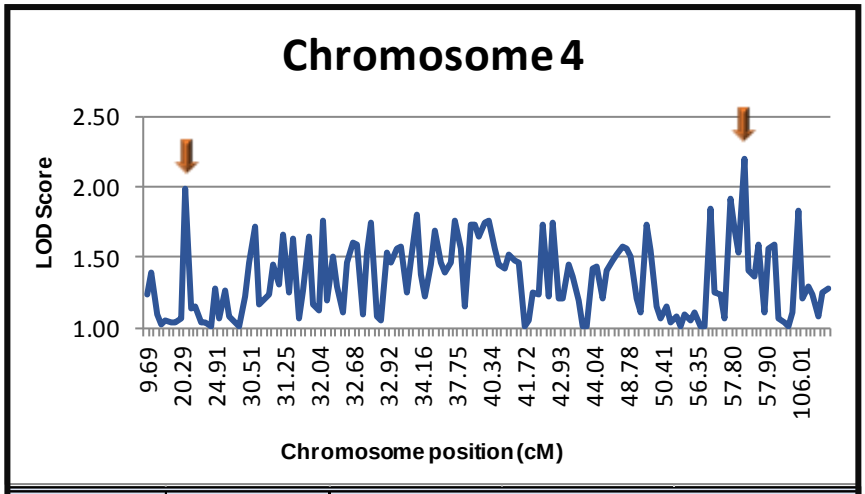


Max LOD_Score	Marker rs	Marker Position		
		Genetic (cM)	Physical (bp)	Cytogenetic (Chr)
2.19	rs1977265	203.24	207707288-207707388	**1q32.2**

Figure 26- Linkage analysis at chromosome 1. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM). LOD scores ≥ 2.0 are pointed out by orange arrow in the graphic. Markers for which LOD scores ≥ 2 were obtained are identified in the Table by the RefSNP (rs), together with the corresponding genetic position in centiMorgan (cM), the physical location on the chromosome in base pairs (bp), and the cytogenetic chromosomal position (Chr).

Table XIII- List of the candidates genes located at chromosome 1. For each candidate gene, the cytogenetic chromosomal position and the result of linkage analysis are indicated.

Chromosome 1		
Candidate Gene	**Chromosome position**	**Linkage Analysis**
TARS2	1q21.2	**No linkage**

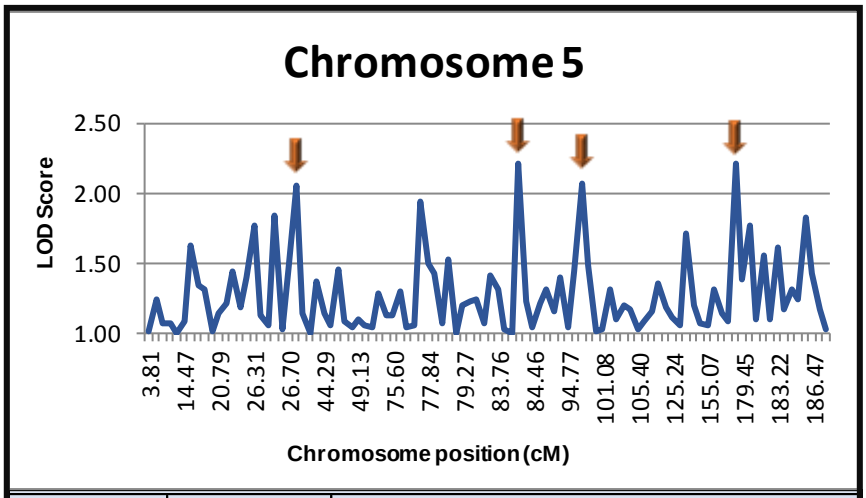


Max LOD_Score	Marker rs	Marker Position		
		Genetic (cM)	Physical (bp)	Cytogenetic (Chr)
1.98	rs7677806	20.83	10,381,331-10,381,431	**4p16.1**
2.20	rs12504525	57.83	44,573,677-44,573,777	**4p13**

Figure 27- Linkage analysis at chromosome 4. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM). LOD scores ≥ 2.0 or ≥ 1.5 (when closely located to a candidate gene) are pointed out by orange arrows in the graphic. Markers for which LOD scores ~ 2 were obtained are identified in the Table by the RefSNP (rs), together with the corresponding genetic position in centiMorgan (cM), the physical location on the chromosome in base pairs (bp), and the cytogenetic chromosomal position (Chr).

Table XIV- List of the candidates genes located at chromosome 4. For each candidate gene, the cytogenetic chromosomal position and the result of linkage analysis are indicated.

Chromosome 4		
Candidate Gene	**Chromosome position**	**Linkage Analysis**
TBC1D14	4p16.1	**Linkage**
TIGD4	4q31.3	**No linkage**

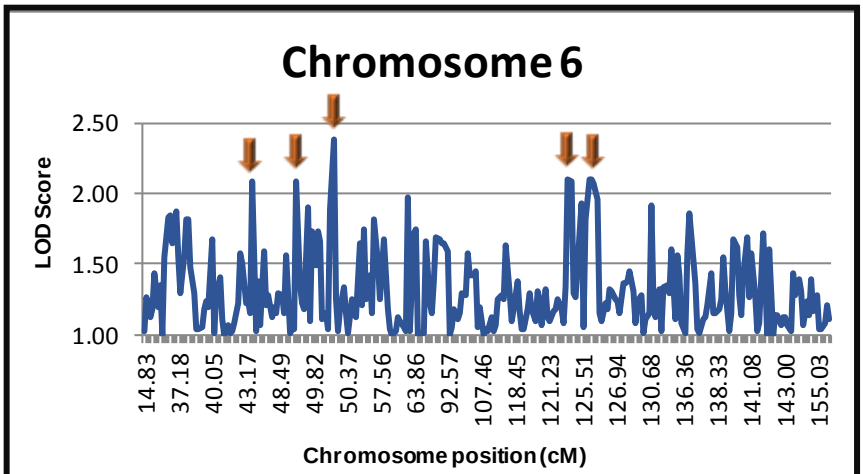


Max LOD_Score	Marker rs	Marker Position		
		Genetic (cM)	Physical (bp)	Cytogenetic (Chr)
2.05	rs7729515	26.71	12,309,439-12,309,539	**5p15.2**
2.22	rs2289599	83.95	77,912,275-77,912,375	**5q14.1**
2.07	rs2545612	96.92	89,311,721-89,311,821	**5q14.3**
2.21	rs1007775	178.44	174,496,837-174,496,937	**5q35.2**

Figure 28- Linkage analysis at chromosome 5. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM). LOD scores ≥ 2.0 are pointed out by orange arrows in the graphic. Markers for which LOD scores ≥ 2.0 were obtained are identified in the Table by the RefSNP (rs), together with the corresponding genetic position in centiMorgan (cM), the physical location on the chromosome in base pairs (bp), and the cytogenetic chromosomal position (Chr).

Table XV- List of the candidates genes located at chromosome 5. For each candidate gene, the cytogenetic chromosomal position and the result of linkage analysis are indicated.

Chromosome 5		
Candidate Gene	**Chromosome position**	**Linkage Analysis**
TRIO	5p15.2	**Linkage**
KCT16/SPRY4	5q31.3	**No linkage**

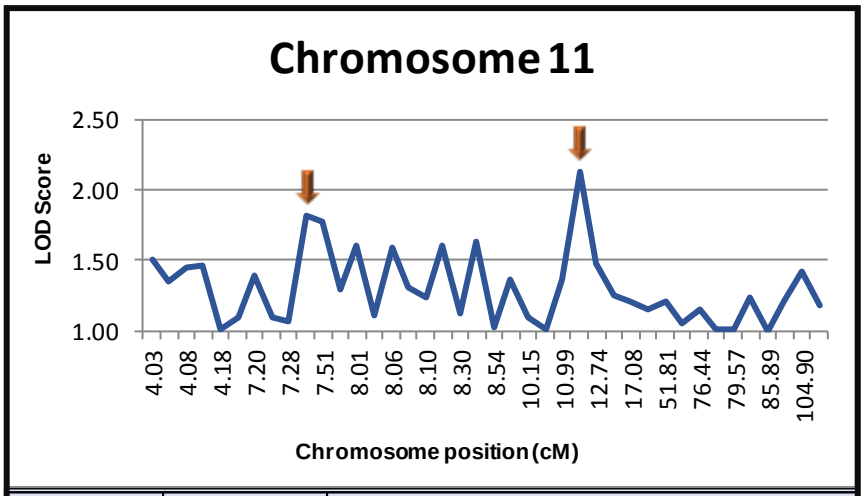


Max LOD_Score	Marker rs	Marker Position		
		Genetic (cM)	Physical (bp)	Cytogenetic (Chr)
2.08	rs7450844	43.85	22,451,541-22,451,641	**6p22.3**
2.09	rs3130404	49.76	30,264,423-30,264,523	**6p22.1**
2.38	rs3134762	50.08	31,243,039-31,243,139	**6p21.33**
2.10	rs11154221	125.37	124,354,050-124,354,150	**6q22.31**
2.10	rs642608	125.52	124,450,172-124,450,272	**6q22.31**

Figure 29- Linkage analysis at chromosome 6. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM). LOD scores ≥ 2.0 are pointed out by orange arrows in the graphic. Markers for which LOD scores ≥ 2.0 were obtained are identified in the Table by the RefSNP (rs), together with the corresponding genetic position in centiMorgan (cM), the physical location on the chromosome in base pairs (bp), and the cytogenetic chromosomal position (Chr).

Table XVI- List of the candidates genes located at chromosome 6. For each candidate gene, the cytogenetic chromosomal position and the result of linkage analysis are indicated.

Chromosome 6		
Candidate Gene	**Chromosome position**	**Linkage Analysis**
CASP8AP2	6q15	**No linkage**



Max LOD_Score	Marker rs	Marker Position		
		Genetic (cM)	Physical (bp)	Cytogenetic (Chr)
1.82	rs2723663	7.49	6,462,230-6,462,330	**11p15.4**
2.13	rs12418898	12.18	10,844,430-10,844,530	**11p15.4**

Figure 30- Linkage analysis at chromosome 11. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM). LOD scores ≥ 2.0 or ≥ 1.5 (when closely located to a candidate gene) are pointed out by orange arrows in the graphic. Markers for which LOD scores ≥ 1.5 were obtained are identified in the Table by the RefSNP (rs), together with the corresponding genetic position in centiMorgan (cM), the physical location on the chromosome in base pairs (bp), and the cytogenetic chromosomal position (Chr).

Table XVII- List of the candidates genes located at chromosome 11. For each candidate gene, the cytogenetic chromosomal position and the result of linkage analysis are indicated.

Chromosome 11		
Candidate Gene	**Chromosome position**	**Linkage Analysis**
ATHL1	11p15.5	**No linkage**
STK33	11p15.4	**Linkage**

Taken together, linkage analysis in family 2 revealed that the majority of the candidate genes identified by WES did not present evidences of being linked to the disease (*ATHL1*, *CASP8AP2*, *KCTD16*, *SPRY4*, *TIGD4* and *TARS 2*). Nonetheless, three of the nine genes were positioned in *loci* for which LOD scores close to 2 were obtained, thus suggesting linkage with the disease (*STK33*, *TBC1D14* and *TRIO*).

4.8 ANALYSIS OF SOMATIC *HRAS* MUTATIONS IN OTHER PORTUGUESE FAMILIES WITH FNMTC

The characterization of somatic mutations in the thyroid tumor samples of patients from family 2 revealed *HRAS* mutations in all the tumors from the nuclear family. Although mutations in *HRAS* can occur in thyroid carcinomas of follicular origin, mutations in *NRAS* are much more frequent and common in these type of tumors. The fact that almost all tumors presented mutations in the same isoform of *RAS*, except for patient's V.67 tumor, which, as already mentioned, had a *BRAF* mutation and could be a phenocopy, can be an indication of a specific molecular signature associated with tumor development, in this particular family. For this reason, the presence of somatic *HRAS* mutations was also evaluated in the tumor samples of other families with FNMTC and, a subsequent analysis of the candidate gene variants in the tumors with *HRAS* mutation was also undertaken.

The majority of the family thyroid tumor samples available had been previously characterized by our group for *BRAF* mutations, and some for *RAS* mutations. Since the mutations in *BRAF* and *RAS* have been described as mutually exclusive, in this work the search for mutations in *HRAS* was performed only in 22 tumor samples, which did not have any *BRAF* mutations or that had not been analysed before for *RAS*/*BRAF* mutations.

In the 22 tumors from 22 families with FNMTC sequenced, only one sample presented a mutation in *HRAS* (p.Gln61Arg).

4.9 ANALYSIS OF THE CANDIDATE GENE VARIANTS IN OTHER FNMTC FAMILIES

In this work, only one familial thyroid tumor, from a proband, presented a somatic mutation in *HRAS* (see point 4.8). However, previous studies by our group, identified a *HRAS* mutations (p.Gln61Lys) in a proband from another family with FNMTC. Thus, the germline alterations in the candidate genes were analysed in DNA samples from these two probands.

Additionally, the germline alterations in the candidate genes were also analysed in a proband from a FNMTC family with a clinical spectrum similar to family 2, that had patients with MNG, with NMTC and colorectal cancer, as well patients with other types of cancer. The family tree of this family is shown in the Appendix.

Sanger sequencing of the candidate variants found in the *ATHL1, BLMH, CASP8AP2, DNAH17, KCTD16, SPRY4, STK33, TARS2, TBC1D14, TlGD4*, and *TRIO* genes was performed in the three above mentioned probands. For one sample (with *HRAS* mutations p.Gln61Lys) the amplification by PCR of *TARS2* was not possible, thus this gene was not sequenced in this sample

No germline mutations were detected in the assessed gene regions analysed in the three samples.

4.10 COMPREHENSIVE OF ANALYSIS THE DIFFERENT RESULTS OBTAINED FOR CANDIDATE GENES

An overview of all the results achieved in the different analyses performed in the present thesis, for the 10 variants detected in the 9 candidate genes, is presented in Table XVIII.

Table XVIII- Summary of the results obtained in the different analyses performed for the nine final candidate genes.

Genetic Alteration	Chromosomal position	Gene Function	Segregation Model	Impact Prediction	Thyroid expression		Mapping Linkage Analysis
					Protein expression	mRNA expression	
ATHL1 p.Arg555Trp	11p15.5	Hydrolase activity	**Thyroid Tumor + Colon (Nuclear Family)**	Benign (not consensual)	Medium	Low	No linkage
CASP8AP2 p.Thr211Ile	6q15	Pro-Apoptotic protein; Histone gene expression; Control of cell cycle	**Thyroid Tumor + MNG + Colon (Nuclear Family)**	Damaging	No information available	Low	No linkage
KCTD16 p.Ala249Thr	5q31.3	Voltage-gated ion channel activity; Auxiliar of GABA-B receptor	**Thyroid Tumor + Colon (Nuclear Family)**	Damaging	Not detected	Low	No linkage
SPRY4 p.Thr234Met	5q31.3	Negative regulation of EGFR-transduced MAPK pathway; Downstream effector of Wnt7A/Fzd9 pathway; Inhibit cell growth and migration	**Thyroid Tumor + Colon (Nuclear Family)**	Damaging	Medium	Medium	No linkage
STK33 p.Asp442His	11p15.4	Protein serine/threonine kinase activity; Tumor cell progression; Cytoskeleton dynamics	**Thyroid Tumor + Colon (Nuclear Family)**	Benign (not consensual)	Medium	Low	Linkage
TARS2 p.Pro387Leu	1q21.2	Mitochondrial Translation	**Thyroid Tumor + Colon (Nuclear Family)**	Benign (not consensual)	High	Medium	No linkage
TBC1D14 p.Glu446Gln	4p16.1	Protein kinase binding; GTPase activator activity; Negative regulator of autophagy	**Thyroid Tumor + MNG + Colon (Nuclear Family) and MNG (complete family)**	Damaging	Medium	Medium	Linkage
TIGD4 p.Ala432Thr p.Gln489Arg	4q31.3	DNA binding	**Thyroid Tumor + Colon (Nuclear Family)**	Benign	Low	Low	No linkage
TRIO p.Ala19Thr	5p15.2	Protein serine/threonine kinase activity; ATP binding; Cell migration; Axon outgrowth; TNF-α inflammatory pathway	**Thyroid Tumor + MNG + Colon (Nuclear Family)**	Benign	High	Low	Linkage

Not consensual, the impact prediction differed between softwares.

CHAPTER 5

Discussion and Perspectives

5.1 DISCUSSION

Thyroid cancer is the most common endocrine neoplasia, accounting for about 1-3% of all malignancies[15].

Non medullary thyroid carcinomas (NMTC) are originated from the thyroid follicular cells, and can present together with other pathologies in the context of familial cancer syndromes, however, in these cases, thyroid cancer is not usually the most frequent tumor[15,17]. Conversely, in some families, NMTC may be the main clinical manifestation or present itself in isolation[15,19].

It is thought that the inheritance pattern of FNMTC, where the main/only clinical manifestation is NMTC, is autosomal dominant with incomplete penetrance[15,17]. The familial forms of NMTC (FNMTC), represent approximately 5% of all NMTCs[23,33].

The aetiology of FNMTC is poorly understood, and the known susceptibility genes for this disease only account for a very small fraction of familial thyroid cancer genetic susceptibility, therefore, the early identification of mutation carriers and genetic counselling are still not conceivable in these families[15,17,24]. Hence, the elucidation of the molecular basis of FNMTC is crucial, as it may allow in the future presymptomatic diagnosis, genetic counselling and more effective treatments.

In the present work a highly informative family, family 2 (Figure 8, Chapter 3), with 7 cases of thyroid cancer, of whom 5 also had colon lesions (colorectal polyps), 7 cases of multinodular goiter (MNG), and other cases with different tumors, was studied. The main aim of this work was to identify potentially pathogenic genetic variant(s), that may confer susceptibility for the development of cancer in this family, and which could also confer susceptibility for other unrelated FNMTC cases.

Thus, whole exome sequencing (WES) was carried out in leucocyte DNAs from 6 members from family 2, affected with thyroid carcinomas, in order to identify genomic variants shared between these patients, and that could be involved in the aetiology of thyroid tumors in this family. WES was selected as the strategy to identify new candidate genes, because the exome, the coding sequences of the genome, harbours the great majority of pathogenic mutations (~ 85%), and few disease causing mutations are found in intronic or regulatory sequences[73–75]. Additionally, WES, when compared with whole genome

sequencing, is less expensive and, as it is focused on ~ 1% enriched subset of the genome, enables a coverage ideally suitable to detect mutations, and, generates less raw sequence data to analyse and interpret[73–75].

For the bioinformatic analysis of the WES data obtained, four different analyses were undertaken (Table IV, Chapter 3), under the assumption that two of the family members with thyroid cancer (III.10 and V.67) could be phenocopies.

The characterization of *RAS* and *BRAF* somatic mutations in 6 tumours samples from the patients of family 2, revealed that patient´s V.67 tumor had a *BRAF* mutation, whereas all of the other 5 tumors presented mutations in *HRAS*. These results suggested that the mechanism underlying tumor initiation and development in patient V.67 could be distinct from that in the remaining relatives of the nuclear family. This characterization could not be done in patient III.10, who also had thyroid cancer, because there was no tumor sample available. Thus, we hypothesised that the germline mutated gene(s) in this family, could account for this somatic *HRAS* signature in the mechanism involved in the progression of these familial tumors.

It is also interesting to observe that all patients of nuclear family had colon lesions together with thyroid cancer, whereas patients III.10 and V.67 (putative phenocopies), to our knowledge, did not present any colon lesion. This could represent, a syndromic aspect related with the disease in this family and, thus, could be also a clue for the molecular pathways involved in the disease development, which may not be the same for nuclear family and for the two patients with thyroid cancer from the other branches.

Nevertheless, all scenarios were taken in consideration regarding the possibility of two phenocopies in the family, and the same filters and criteria for variants selection were applied equally in all four analyses above mentioned.

All the criteria and filters (quality parameters, defined genotype, localization in protein coding transcripts, European allele frequency, synonymous variants, variants in families with other types of cancers, and homozygous variants) that were used to select genomic variants were defined taking into account the kind of variants that would more likely be pathogenic and causative of the disease under study. However, significant variants might have been excluded and, in the future, it may be necessary assess again the variants that where not chosen for the

subsequent studies performed in this project. Additionally, some of the variants that remained after the bioinformatic analysis, were not prioritized for the following studies, because, as it was explained in the Chapter 4.3, they were likely to be artefacts. Nonetheless, in the future, these potential artefacts also may have to be evaluated again.

After bioinformatic analysis of the WES data, the validation of selected variants by Sanger sequencing is a standard procedure, since errors can occur in the NGS, and the called variants may be false-positive findings. As it can be observed in the validation results, four variants in three different genes were not confirmed, probably being WES artefacts and coincidentally, or not, the percentage of reads for the alternative allele was less than 40% in all samples, for all these variants. Since these are heterozygous variants, the percentage of reads for the reference and alternative alleles are expected to be around 40-50%. For some samples, with worse quality, the percentage of the alternative allele can indeed be lower. However, if all samples present values < 40%, this could be a criteria to consider in the exclusion of false-positive findings.

For one of the genes, the called variants were not correct, as instead of two missense variants, a deletion covering the location of these variants was found in the validation process. This could be due to a misalignment with the genome reference, which lead to an incorrect calling of the alteration, or due to the fact that the pipeline used to call the variants was not appropriate to detect these deletions. This issue is being addressed, and a new analysis to specifically identify large indels (CNVs) is being conducted by the bioinformatic company. The deletion found was excluded afterwards, as it was found to be homozygous in some patients and healthy controls. Additionally, two other variants, in two different genes, were also excluded due to homozygosity.

Segregation studies were performed for the heterozygous variants validated by Sanger sequencing, and they were assigned to four different disease segregation models. The majority of the variants segregated according to the disease model defined by, thyroid tumours and colon lesions in the nuclear family only. Two variants were assigned to another defined disease model, which was similar to the previous one, but also included MNG in the nuclear family, as part of the phenotype. Only one variant segregated with a third disease model, characterized by thyroid cancer and MNG in the nuclear family, and also with

MNG in patients from another branch. Finally, the variants categorized as inconclusive were the ones that were detected in members of healthy branches. Since the exact disease model it not yet known for family 2, all models were taken in consideration in the subsequent studies performed in this work. Variants categorized as inconclusive were not considered high-priority at this stage. However, these variants should not be excluded from further studies, because the gene involved in the aetiology of this disease may present a lower penetrance in some family branches and, thus, some members that are presently unaffected may be asymptomatic carriers.

The variants that did not have an inconclusive segregation model were further analysed in the familial tumors samples, to evaluate if any of these alterations were associated with LOH. Although, the presence of these variants was confirmed in the tumors, no pattern suggestive of LOH was found for any of the assessed genes. These results, although preliminary, suggest that if some of these genes is a tumor suppressor gene, involved in the aetiology of thyroid cancer in family 2, the second hit is not likely to have occurred through LOH.

The principal goal of this project was the selection of the most promising genomic variants to conduct functional studies, using appropriate *in vitro* approaches. Following the segregation analyses, ten genomic variants in nine candidate genes, interestingly all from analysis 2, where the two possible phenocopies were excluded, remained for further analyses. However, the number of candidate genes continued to be high to proceed to functional studies. Thus, in order to further select the most relevant genes, several *in silico* analyses, combined with gene mapping using linkage analysis, were undertaken.

None of the results obtained with these analyses supported the exclusion of any candidate gene evaluated, as none should be totally reliable. For instance, some predictions of the variant effect in gene function differed between softwares and, as these results are mainly based in the conservation of the encoded protein throughout evolution and between species, there is no guarantee that a variant predicted to be damaging will indeed be damaging, or that a variant predicted to be benign will not have a damaging effect. Still, the most consensual results obtained with the distinct softwares were considered. Regarding protein expression and mRNA expression, the fact that some genes presented low expression in thyroid gland tissue does not mean that the genes may not have a

relevant role in the thyroid. For instance, the oncogenes generally present low expression in tissues. Although the involvement of tumour suppressor genes is more common in the aetiology of hereditary cancer diseases. Additionally, gene mapping through linkage analysis, a strategy widely used in the past in the identification of several susceptibility genes in hereditary cancer syndromes, was also used in this project. The majority of the candidate genes *loci* did not present evidence of linkage with the disease. However, careful should be taken in the interpretation of these results, because in the microarray selected for this analysis some chromosomal regions may not have a proper coverage by SNP markers, and, in addition, some chromosomal *loci* may have a high percentage of uninformative markers.

In the last stage of this work, an evaluation by Sanger sequencing, of the presence of the ten candidate variants in leucocyte DNAs from probands of three unrelated FNMTC families (two with *HRAS* mutations and one with clinical manifestations similar to family 2) was also performed. This study showed that none of these variants was present in the other families. These results are not totally surprising, since the series of families analysed was very small, and the sequencing just involved the specific gene variants, not including the remaining coding sequences of the candidate genes. Thus, finding exactly the same variant in a candidate gene, in another family, would uncover a founder mutation, which is not a common event. Therefore, the study of these candidate genes should be further addressed, sequencing the complete coding sequence and splice junctions, in the entire series of 87 FNMTC families.

Finally, a comprehensive analysis of all the results obtained was undertaken (Chapter 4; Table XIX). This broad analysis allowed the selection of the most likely pathogenic variants, in order to proceed for *in vitro* studies. In this selection process, a particular relevance was given to the information available on the function of the candidate genes, and on the pathways in which they were involved, and also to the evidence of *locus* linkage with the disease. Therefore, four of the nine candidate genes, *SPRY4*, *STK33*, *TBC1D14* and *TRIO* were prioritized to continue for functional characterization studies.

The *SPRY4* gene seems to have a critical role as modulator of MAPK/ERK pathways, one of the most important pathways associated with thyroid tumor development, as mentioned in Chapter 1[122]. Furthermore, several studies have

described *SPRY4* as an important player in cancer. Although this gene *locus* was not linked with the disease, its relevant function, together with the possibility of a false-negative linkage analysis result, supported its selection for functional characterization.

Regarding *STK33* gene, recent studies described the involvement of this gene in the tumorigenesis and progression of several types of cancer, possibly through MAPK signalling pathway activation, and this gene has been suggested to be a new oncogene. Interestingly this gene *locus* is possibly linked with the disease in this family[134]

Another gene chosen to be evaluated in the next studies was *TRIO*, which is also an interesting candidate gene since it controls a wide range of cellular processes through the activation of Rho GTPases, which are also important for the regulation of several pathways, such as the MAPK/ERK signalling pathway[144]. As for the previous gene, the *TRIO*'s *locus* also showed suggestive linkage with the disease in this family.

The last gene selected for the top four more interesting candidate genes was *TBC1D14*. Although, there is not many information about this gene, as there is for the last three genes, it has been reported that its function is related with protein kinase binding and GTPase activator activity. Additionally, this gene also plays a role as negative regulator of autophagy, all important functions and activities for the control of cellular processes[90,93]. Moreover, the data from the *in silico* and linkage analyses showed that *TBC1D14* may also be an important gene to study.

It is important to point out that the other genes that were not selected for the subsequent phases of this project can also be relevant candidates and, in future studies, they may be chosen for functional characterization.

Overall, an important conclusion from this work was that all the techniques, methodologies and approaches used were useful, but do need to be integrated with each other. Particularly, WES showed to be a valuable and a very fast technique, which allowed the identification of alterations in the exomes, nonetheless it still needs to be combined with other approaches to achieve the ultimate goal(s) with a certain reliability.

The variants identified in the four candidate genes selected, as well as that in the other five genes for which *in silico* and linkage analyses were performed, did

not segregate in the two possible phenocopies, further suggesting that these two patients may have developed sporadic thyroid cancer.

The co-existence of thyroid cancer and colon lesions in the patients from the nuclear family, may be due to the effect of a single germline mutated gene involved in the development of a thyroid and colon syndrome, or to the co-segregation in these cases of another mutated gene responsible for the colon polyps. The somatic *HRAS* signature found in all the thyroid carcinomas from the nuclear family supports the first hypothesis, as the majority of these tumors presented transversions in *HRAS*, that may result from defects in the DNA repair mechanisms (such as base excision repair and mismatch excision repair), which commonly account for colorectal tumor development.

In family 2, besides thyroid and colon lesions, ovarian carcinomas, were also present in two affected members from the nuclear family. Thus, the clarification of the underlying germline mutated gene will be crucial also to define the familial *versus* sporadic status of these tumors in this family.

The roles of the four genes selected as candidates have not been fully characterized, and certainly there are unknown aspects to unveil. Thus, their functional characterization will be crucial not only to assess the effect of the variants detected in these genes, but also to better understand the role of these genes in cancer aetiology.

5.2 PERSPECTIVES

Currently, the functional studies for the four candidate genes (*SPRY4*, *STK33*, *TBC1D14* and *TRIO*) are being developed. At the time that this thesis was written, the construction of vectors was being undertaken by our group in collaboration with Dr. José Ramalho from CEDOC (UNL), in order to study the genes function and evaluating the effect of the variants in normal thyroid cells and in cancer cell lines. Hopefully, the results obtained in these studies will lead to the identification of a novel susceptibility gene for the FNMTC in family 2 that, may also be relevant in the aetiology of this disease in other FNMTC families.

In the future, it is expected that these studies and hence, the identification of susceptibility genes for FNMTC, may allow families with FNMTC to undergo early diagnosis, improving the clinical management of these patients. Furthermore, understanding the biological pathways involved in thyroid cancer may also guide

the development of novel *in vivo* pre-clinical models and suggest possible targets for novel therapies.

CHAPTER 6

Bibliography

1. World Health Organization, Cancer. at <http://www.who.int/cancer/en/> (Accessed 30 April 2016)

2. Strachan, T., & Read, A. (2010). Cancer Genetics. In *Human Molecular Genetics* (4th ed., pp. 537–567). United States: Garland Science.

3. Stephens, F. O., & Aigner, K. R. (2009). What is Malignancy? In *Basics of Oncology* (1st ed., pp. 3–16). Berlin: Springer.

4. Lodish, H., Berk, A., Matsudaira, P., Kaiser, C. A., Krieger, M., Scott, M. P., Zipursky, L., & Darnell, J. (2003). Cancer. In *Molecular Cell Biology* (5th ed., pp. 935–973). New York: Freeman, W. H. & Company.

5. Hanahan, D. & Weinberg, R. a. The hallmarks of cancer. *Cell* **100,** 57–70 (2000).

6. Hanahan, D. & Weinberg, R. a. Hallmarks of cancer: The next generation. *Cell* **144,** 646–674 (2011).

7. Vogelstein, B. & Kinzler, K. W. Cancer genes and the pathways they control. *Nat. Med.* **10,** 789–799 (2004).

8. Croce, C. M. Oncogenes and cancer. *N. Engl. J. Med.* **358,** 502–511 (2008).

9. Payne, S. R. & Kemp, C. J. Tumor suppressor genetics. *Carcinogenesis* **26,** 2031–2045 (2005).

10. Shen, Z. Genomic instability and cancer. *J. Moecular Cell Biol.* **3,** 1–3 (2011).

11. Muro-Cacho, C. A. & Ku, N. N. K. Tumors of the thyroid gland: Histologic and cytologic features - Part 1. *Cancer Control* **7,** 276–287 (2000).

12. Erickson, L. A. (2011). Endocrine Pathology. (3rd ed., pp. 535–598). New York: Springer.

13. Nguyen, Q. T. *et al.* Diagnosis and treatment of patients with thyroid cancer. *Am. Heal. drug benefits* **8,** 30–40 (2015).

14. Thyroid Gland. at <http://www.cancer.gov/types/thyroid/patient/thyroid-treatment-pdq>(Accessed 1 May 2016)

15. Bonora, E., Tallini, G. & Romeo, G. Genetic predisposition to familial nonmedullary thyroid cancer: An update of molecular findings and state-of-the-art studies. *J. Oncol.* **2010,** 1–7 (2010).

16. Cabanillas, M. E. *et al.* Thyroid Gland Malignancies. *Hematol. Oncol. Clin. North Am.* **29,** 1123–1143 (2015).

17. Navas-Carrillo, D., Ríos, A., Rodríguez, J. M., Parrilla, P. & Orenes-Piñero, E. Familial nonmedullary thyroid cancer: Screening, clinical, molecular and genetic findings. *Biochim. Biophys. Acta - Rev. Cancer* **1846,** 468–476 (2014).

18. Nikiforov, Y. E. & Nikiforova, M. N. Molecular genetics and diagnosis of thyroid cancer. *Nat Rev Endocrinol* **7,** 569–580 (2011).

19. Rowland, K. J. & Moley, J. F. Hereditary thyroid cancer syndromes and genetic testing. *J. Surg. Oncol.* **111,** 51–60 (2015).

20. Xing, M. Molecular pathogenesis and mechanisms of thyroid cancer. *Nat Rev Cancer* **13,** 184–199 (2013).

21. Bible, K. C. & Ryder, M. Evolving molecularly targeted therapies for advanced-stage thyroid cancers. *Nat. Rev. Clin. Oncol.* 1–14 (2016).

22. Thyroid Cancer Epidemiology. at <https://www.thyrogen.com/healthcare/OLD-Thyroid-Cancer-Overview/Thyroid-Cancer-Epidemiology.aspx>(Accessed 1 May 2016)

23. Landa, I. & Robledo, M. Association studies in thyroid cancer susceptibility: are we on the right track? *J. Mol. Endocrinol.* **47,** (2011).

24. Malchoff, C. D. & Malchoff, D. M. Familial nonmedullary thyroid carcinoma. *Cancer Control* **13,** 106–110 (2006).

25. Moura, M. M., Cavaco, B. M. & Leite, V. RAS proto-oncogene in medullary thyroid carcinoma. *Endocr. Relat. Cancer* **22,** R235–R252 (2015).

26. Raue, F. & Frank-Raue, K. Epidemiology and Clinical Presentation of Medullary Thyroid Carcinoma. *Recent results cancer Res.* 61–90 (2015).

27. Chernock, R. D. & Hagemann, I. S. Molecular pathology of hereditary and sporadic medullary thyroid carcinomas. *Am. J. Clin. Pathol.* **143,** 768–777 (2015).

28. Moura, M. M., Cavaco, B. M., Pinto, A. E. & Leite, V. High prevalence of RAS mutations in RET-negative sporadic medullary thyroid carcinomas. *J. Clin. Endocrinol. Metab.* **96,** 863–868 (2011).

29. Buffet, C. & Groussin, L. Molecular perspectives in differentiated thyroid cancer. *Ann. Endocrinol. (Paris).* **76,** 1S8–1S15 (2015).

30. Zolotov, S. Genetic testing in differentiated thyroid carcinoma: indications and clinical implications. *Rambam Maimonides Med J* **7,** 1–10 (2016).

31. Pita, J. M., Figueiredo, I. F., Moura, M. M., Leite, V. & Cavaco, B. M. Cell cycle deregulation and TP53 and RAS mutations are major events in poorly

differentiated and undifferentiated thyroid carcinomas. *J. Clin. Endocrinol. Metab.* **99,** 497–507 (2014).

32. LiVolsi, V. a. & Baloch, Z. W. Familial thyroid carcinoma: the road less travelled in thyroid pathology. *Diagnostic Histopathol.* **15,** 87–94 (2009).

33. Moses, W., Weng, J. & Kebebew, E. Prevalence, clinicopathologic features, and somatic genetic mutation profile in familial versus sporadic nonmedullary thyroid cancer. *Thyroid* **21,** 367–371 (2011).

34. Pinto, A. E. *et al.* Familial vs sporadic papillary thyroid carcinoma: a matched-case comparative study showing similar clinical/prognostic behaviour. *Eur. J. Endocrinol.* **170,** 321–327 (2014).

35. Loh, K. C. Familial nonmedullary thyroid carcinoma: a meta-review of case series. *Thyroid* **7,** 107–113 (1997).

36. Capezzone, M. *et al.* Familial non-medullary thyroid carcinoma displays the features of clinical anticipation suggestive of a distinct biological entity. *Endocr. Relat. Cancer* **15,** 1075–1081 (2008).

37. Uchino, S. *et al.* Familial nonmedullary thyroid carcinoma characterized by multifocality and a high recurrence rate in a large study population. *World J. Surg.* **26,** 897–902 (2002).

38. Pitoia, F., Cross, G., Salvai, M. E., Abelleira, E. & Niepomniszcze, H. Patients with familial non-medullary thyroid cancer have an outcome similar to that of patients with sporadic papillary thyroid tumors. *Arq. Bras. Endocrinol. Metabol.* **55,** 219–223 (2011).

39. Bignell, G. R. *et al.* Familial nontoxic multinodular thyroid goiter locus maps to chromosome 14q but does not account for familial nonmedullary thyroid cancer. *Am. J. Hum. Genet.* **61,** 1123–1130 (1997).

40. Rio Frio, T. *et al.* DICER1 mutations in familial multinodular goiter with and without ovarian Sertoli-Leydig cell tumors. *JAMA* **305,** 68–77 (2011).

41. Rutter, M. M. *et al.* DICER1 mutations and differentiated thyroid carcinoma: Evidence of a direct association. *J. Clin. Endocrinol. Metab.* **101,** 1–5 (2016).

42. Frezzetti, D. *et al.* The microRNA-processing enzyme dicer is essential for thyroid function. *PLoS One* **6,** 1–10 (2011).

43. He, H. *et al.* SRGAP1 is a candidate gene for papillary thyroid carcinoma susceptibility. *J. Clin. Endocrinol. Metab.* **98,** 973–980 (2013).

44. Castanet, M. *et al.* Linkage and mutational analysis of familial thyroid dysgenesis demonstrate genetic heterogeneity implicating novel genes. *Eur. J. Hum. Genet.* **13,** 232–239 (2005).

45. De Felice, M. & Di Lauro, R. Minireview: Intrinsic and extrinsic factors in thyroid gland development: An update. *Endocrinology* **152,** 2948–2956 (2011).

46. Felice, M. De & Lauro, R. Di. The development of the thyroid gland: what we know and what we would like to know. *Curr. Opin. Endocrinol. Diabetes* **12,** 4–9 (2005).

47. Ngan, E. S. W. *et al.* A germline mutation (A339V) in thyroid transcription factor-1 (TITF-1/NKX2.1) in patients with multinodular goiter and papillary thyroid carcinoma. *J. Natl. Cancer Inst.* **101,** 162–175 (2009).

48. Pereira, J. S. *et al.* Identification of a novel germline FOXE1 variant in patients with familial non-medullary thyroid carcinoma (FNMTC). *Endocrine* **49,** 204–214 (2014).

49. Canzian, F. *et al.* A gene predisposing to familial thyroid tumors with cell oxyphilia maps to chromosome 19p13.2. *Am. J. Hum. Genet.* **63,** 1743–1748 (1998).

50. Bevan, S. *et al.* A comprehensive analysis of MNG1, TCO1, fPTC, PTEN, TSHR, and TRKA in familial nonmedullary thyroid cancer: Confirmation of linkage to TCO1. *J. Clin. Endocrinol. Metab.* **86,** 3701–3704 (2001).

51. Malchoff, C. D. *et al.* Papillary thyroid carcinoma associated with papillary renal neoplasia: Genetic linkage analysis of a distinct heritable tumor syndrome. *J. Clin. Endocrinol. Metab.* **85,** 1758–1764 (2000).

52. McKay, J. D. *et al.* Localization of a susceptibility gene for familial nonmedullary thyroid carcinoma to chromosome 2q21. *Am. J. Hum. Genet.* **69,** 440–446 (2001).

53. Cavaco, B. M., Batista, P. F., Sobrinho, L. G. & Leite, V. Mapping a new familial thyroid epithelial neoplasia susceptibility locus to chromosome 8p23.1-p22 by high-density single-nucleotide polymorphism genome-wide linkage analysis. *J. Clin. Endocrinol. Metab.* **93,** 4426–4430 (2008).

54. Capezzone, M. *et al.* Short telomeres, telomerase reverse transcriptase gene amplification, and increased telomerase activity in the blood of familial papillary thyroid cancer patients. *J. Clin. Endocrinol. Metab.* **93,** 3950–3957 (2008).

55. Suh, I. *et al.* Distinct loci on chromosome 1q21 and 6q22 predispose to familial nonmedullary thyroid cancer: A SNP array-based linkage analysis of 38 families. *Surgery* **146,** 1073–1080 (2009).

56. He, H. *et al.* A susceptibility locus for papillary thyroid carcinoma on chromosome 8q24. *Cancer Res.* **69,** 625–631 (2009).

57. Xiong, Y. *et al.* MiR-886-3p regulates cell proliferation and migration, and is dysregulated in familial non-medullary thyroid cancer. *PLoS One* **6,** 1–11 (2011).

58. Tomsic, J. *et al.* A germline mutation in SRRM2, a splicing factor gene, is implicated in papillary thyroid carcinoma predisposition. *Sci. Rep.* **5,** 10566 (2015).

59. Gara, S. K. *et al.* Germline HABP2 Mutation Causing Familial Nonmedullary Thyroid Cancer. *N. Engl. J. Med.* **8,** 1699–1712 (2015).

60. Tomsic, J. *et al.* HABP2 G534E Variant in Papillary Thyroid Carcinoma. *PLoS One* **11,** 1–8 (2016).

61. Alzahrani, A. S., Murugan, A. K., Qasem, E. & Al-Hindi, H. HABP2 gene mutations do not cause familial or sporadic non-medullary thyroid cancer in a highly inbred Middle Eastern population. *Thyroid* **26,** 1–17 (2016).

62. Sahasrabudhe, R. *et al.* The HABP2 G534E variant is an unlikely cause of familial non-medullary thyroid cancer. *J. Clin. Endocrinol. Metab.* 1–7 (2015).

63. Bailey-Wilson, J. E. & Wilson, A. F. Linkage analysis in the next-generation sequencing era. *Hum. Hered.* **72,** 228–236 (2011).

64. Teare, M. D. & Santibañez Koref, M. F. Linkage analysis and the study of Mendelian disease in the era of whole exome and genome sequencing. *Brief. Funct. Genomics* **13,** 378–383 (2014).

65. Ott, J., Wang, J. & Leal, S. M. Genetic linkage analysis in the age of whole-genome sequencing. *Nat. Rev. Genet.* **16,** 1–10 (2015).

66. Terwilliger, J. D. & Göring, H. H. H. Gene Mapping in the 20th and 21st Centuries: Statistical Methods, Data Analysis, and Experimental Design. *Hum. Biol.* **72,** 663–728 (2009).

67. Barecki, I. B. & Suarez, B. K. in *Adv. Genet.* **42,** 45–66 (Academic Press, 2001).

68. Shyr, D. & Liu, Q. Next generation sequencing in cancer research and clinical application. *Biol. Proced. Online* **15,** 1–4 (2013).

69. Luthra, R., Chen, H., Roy-Chowdhuri, S. & Singh, R. R. Next-generation sequencing in clinical molecular diagnostics of cancer: Advantages and challenges. *Cancers (Basel).* **7,** 2023–2036 (2015).

70. De Magalhães, J. P., Finch, C. E. & Janssens, G. Next-generation sequencing in aging research: Emerging applications, problems, pitfalls and possible solutions. *Ageing Res. Rev.* **9,** 315–323 (2010).

71. Mardis, E. R. & Wilson, R. K. Cancer genome sequencing: A review. *Hum. Mol. Genet.* **18,** 163–168 (2009).

72. Stadler, Z. K., Schrader, K. a., Vijai, J., Robson, M. E. & Offit, K. Cancer genomics and inherited risk. *J. Clin. Oncol.* **32,** 687–698 (2014).

73. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 19096–19101 (2009).

74. Tetreault, M., Bareke, E., Nadaf, J., Alirezaie, N. & Majewski, J. Whole-exome sequencing as a diagnostic tool: current challenges and future opportunities. *Expert Rev. Mol. Diagn.* **15,** 749–60 (2015).

75. Stranneheim, H. & Wedell, a. Exome and genome sequencing: A revolution for the discovery and diagnosis of monogenic disorders. *J. Intern. Med.* **279,** 3–15 (2016).

76. Haimovich, A. D. Methods, challenges, and promise of next-generation sequencing in cancer biology. *Yale J. Biol. Med.* **84,** 439–446 (2011).

77. Puregene DNA procedure. at <https://www.qiagen.com/pt/shop/sample-technologies/dna/dna-preparation/gentra-puregene-blood-kit#productdetails>(Accessed 13 June 2016)

78. Qiagen. GeneRead DNA FFPE Handbook. at <file:///C:/Users/ASUS/Downloads/1082041_SPOS_FFPE_DNA_GeneRead_0414_WW_chr.pdf>(Accessed 13 June 2016)

79. Illumina. *An Introduction to Next-Generation Sequencing Technology*. (2016).

80. Liu, L. *et al.* Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* **2012,** 1–11 (2012).

81. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat. Biotechnol.* **26,** 1135–1145 (2008).

82. 1000 Genomes | A Deep Catalog of Human Genetic Variation. at <http://www.1000genomes.org/>(Accessed between December 2015 to March 2016)

83. ExAC Browser (Beta) | Exome Aggregation Consortium. at <http://exac.broadinstitute.org/>(Accessed between December 2015 to March 2016)

84. Sauna, Z. E. & Kimchi-Sarfaty, C. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* **12,** 683–91 (2011).

85. SIFT. at <http://sift.jcvi.org/>(Accessed between December 2015 to May 2016)

86. PolyPhen-2 prediction of functional effects of human nsSNPs. at <http://genetics.bwh.harvard.edu/pph2/>(Accessed between December 2015 to May 2016)

87. Mutation Taster. at <http://www.mutationtaster.org/>(Accessed between December 2015 to May 2016)

88. Mutation Assessor. at <http://mutationassessor.org/r3/>(Accessed between December 2015 to May 2016)

89. The Human Protein Atlas. at <http://www.proteinatlas.org/>(Accessed between December 2015 to July 2016)

90. GeneCards - Human Genes | Gene Database | Gene Search. at <http://www.genecards.org/>(Accessed between September 2015 to June 2016)

91. OMIM - Online Mendelian Inheritance in Man. at <http://www.omim.org/> (Accessed between December 2015 to June 2016)

92. Human Protein Reference Database. at <http://www.hprd.org/>(Accessed between December 2015 to June 2016)

93. UniProt (Universal Protein Resource). at <http://www.uniprot.org/> (Accessed between December 2015 to June 2016)

94. STRING: functional protein association networks. at <http://string-db.org/> (Accessed between December 2015 to June 2016)

95. Garibyan, L. & Avashia, N. Polymerase Chain Reaction. *J. Invest. Dermatol.* **133,** 1–4 (2013).

96. Lui, C., Cady, N. C. & Batt, C. a. *Nucleic acid-based detection of bacterial pathogens using integrated microfluidic platform systems*. *Sensors* **9,** (2009).

97. *DNA Sequencing Protocols*. (Humana Press, 2001).

98. Nelson, D. L. & Cox, M. M. *Lehninger Principles of Biochemistry*. *Statew. Agric. L. Use Baseline 2015* **1,** (W.H.Freeman and Company, 2008).

99. Applied Biosystems. *DNA sequencing by capillary electrophoresis. Applied Biosystems Chemistry Guide*. (2009).

100. Applied Biosystems. *Application note - Advances in Fast PCR Contribute to a Fast Resequencing Workflow*. (2008).

101. Agarose gel electrophoresis. at <https://en.wikiversity.org/wiki/Agarose_gel_electrophoresis>(Accessed 21 June 2016)

102. GFXTM PCR DNA and Gel Band Purification protocol. at <http://www.gelifesciences.co.jp/catalog/0543.html>(Accessed 21 June 2016)

103. Hamazaki, H. & Hamazaki, M. H. Catalytic site of human protein-glucosylgalactosylhydroxylysine glucosidase: Three crucial carboxyl residues were determined by cloning and site-directed mutagenesis. *Biochem. Biophys. Res. Commun.* **469,** 357–362 (2016).

104. Xu, X. *et al.* Solution NMR Structures of Homeodomains from Human Proteins ALX4, ZHX1, and CASP8AP2 Contribute to the Structural Coverage of the Human Cancer Protein Interaction Network. *J Struct Funct Genomics* **15,** 201–207 (2014).

105. Imai, Y. *et al.* The CED-4-homologous protein FLASH is involved in Fas-mediated activation of caspase-8 during apoptosis. *Nature* **398,** 777–85 (1999).

106. Milovic-Holm, K., Krieghoff, E., Jensen, K., Will, H. & Hofmann, T. G. FLASH links the CD95 signaling pathway to the cell nucleus and nuclear bodies. *EMBO J.* **26,** 391–401 (2007).

107. Choi, Y. H. *et al.* FLASH Coordinates N-kB Activity via TRAF2. *J. Biol. Chem.* **276,** 25073–25077 (2001).

108. Alm-Kristiansen, A. *et al.* PIAS1 interacts with FLASH and enhances its co-activation of c-Myb. *Mol Cancer* **10,** 21 (2011).

109. Alm-Kristiansen, A. *et al.* FLASH acts as a co-activator of the transcription factor c-Myb and localizes to active RNA polymerase II foci. *Oncogene* **27,** 4644–4656 (2008).

110. Yang, X., Burch, B. D., Yan, Y., Marzluff, W. F. & Dominski, Z. FLASH, a pro-apoptotic protein involved in activation of caspase-8 is essential for 3' end processing of histone pre-mRNAs. *Mol. Cell* **36,** 1–24 (2009).

111. Kiriyama, M., Kobayashi, Y., Saito, M., Ishikawa, F. & Yonehara, S. Interaction of FLASH with arsenite resistance protein 2 is involved in cell cycle progression at S phase. *Mol Cell Biol* **29,** 4729–4741 (2009).

112. Barcaroli, D. *et al.* FLASH is required for histone transcription and S-phase progression. *Proc. Natl. Acad. Sci. U. S. A.* **103,** 14808–12 (2006).

113. Yang, X. *et al.* FLASH is required for the endonucleolytic cleavage of histone pre-mRNAs but is dispensable for the 5' exonucleolytic degradation of the downstream cleavage product. *Mol. Cell. Biol.* **31,** 1492–502 (2011).

114. Hummon, A. B. *et al.* Systems-wide RNAi analysis of CASP8AP2/FLASH shows transcriptional deregulation of the replication-dependent histone genes and extensive effects on the transcriptome of colorectal cancer cells. *Mol. Cancer* **11,** 1 (2012).

115. Flotho, C. *et al.* Genes contributing to minimal residual disease in childhood acute lymphoblastic leukemia : prognostic significance of CASP8AP2. *Gene Expr.* **108,** 1050–1057 (2006).

116. Callens, C. *et al.* Clinical impact of NOTCH1 and/or FBXW7 mutations, FLASH deletion, and TCR status in pediatric T-cell lymphoblastic lymphoma. *J. Clin. Oncol.* **30,** 1966–1973 (2012).

117. Minamida, Y., Someda, M. & Yonehara, S. FLASH/casp8ap2 Is indispensable for early embryogenesis but dispensable for proliferation and differentiation of ES Cells. *PLoS One* **9,** 1–11 (2014).

118. De Cola, a *et al.* FLASH is essential during early embryogenesis and cooperates with p73 to regulate histone gene transcription. *Oncogene* **31,** 573–582 (2012).

119. Schwenk, J. *et al.* Native GABA(B) receptors are heteromultimers with a family of auxiliary subunits. *Nature* **465,** 231–235 (2010).

120. Seddik, R. *et al.* Opposite effects of KCTD subunit domains on GABA(B) receptor-mediated desensitization. *J. Biol. Chem.* **287,** 39869–39877 (2012).

121. Metz, M., Gassmann, M., Fakler, B., Schaeren-Wiemers, N. & Bettler, B. Distribution of the auxiliary GABAB receptor subunits KCTD8, 12, 12b, and 16 in the mouse brain. *J. Comp. Neurol.* **519,** 1435–1454 (2011).

122. Masoumi-Moghaddam, S., Amini, A. & Morris, D. L. The developing story of Sprouty and cancer. *Cancer Metastasis Rev.* **33,** 695–720 (2014).

123. Leeksma, O. C. *et al.* Human sprouty 4, a new ras antagonist on 5q31, interacts with the dual specificity kinase TESK1. *Eur. J. Biochem.* **269,** 2546–2556 (2002).

124. Sasaki, A. *et al.* Mammalian Sprouty4 suppresses Ras-independent ERK activation by binding to Raf1. *Cell Cycle* **2,** 281–282 (2003).

125. Ayada, T. *et al.* Sprouty4 negatively regulates protein kinase C activation by inhibiting phosphatidylinositol 4,5-biphosphate hydrolysis. *Oncogene* **28,** 1076–88 (2009).

126. Winn, R. a. *et al.* Antitumorigenic effect of Wnt 7a and Fzd 9 in non-small cell lung cancer cells is mediated through ERK-5-dependent activation of peroxisome proliferator-activated receptor γ. *J. Biol. Chem.* **281,** 26943–26950 (2006).

127. Tennis, M. a *et al.* Sprouty-4 Inhibits Transformed Cell Growth, Migration and Invasion, and Epithelial-Mesenchymal Transition, and Is Regulated by Wnt7A through PPAR(Y) in Non-Small Cell Lung Cancer. *Mol. Cancer Res.* **8,** 833–843 (2010).

128. Li, M. *et al.* SPRY4-mediated ERK1/2 signaling inhibition abolishes 17β-estradiol-induced cell growth in endometrial adenocarcinoma cell. *Gynecol. Endocrinol.* **30,** 600–604 (2014).

129. Jing, H. *et al.* Suppression of Spry4 enhances cancer stem cell properties of human MDA-MB-231 breast carcinoma cells. *Cancer Cell Int.* **16,** 19 (2016).

130. Wang, J., Benjamim, T., Ren, C., Ittmann, M. & Kwabi-Addo, B. Sprouty4, a Suppressor of Tumor Cell Motility, is Downregulated by DNA Methylationin Human Prostate Cancer. *Prostate* **66,** 613–624 (2006).

131. So, W.-K. *et al.* Sprouty4 mediates amphiregulin-induced down-regulation of E-cadherin and cell invasion in human ovarian cancer cells. *Tumor Biol.* (2016)

132. Shaverdashvili, K. *et al.* MT1-MMP dependent repression of the tumor suppressor SPRY4 contributes to MT1-MMP driven melanoma cell motility. *Oncotarget; Vol 6, No 32* **6,** (2015).

133. Karlsson, R. *et al.* Investigation of six testicular germ cell tumor susceptibility genes suggests a parent-of-origin effect in SPRY4. *Hum. Mol. Genet.* **22,** 3373–3380 (2013).

134. Mujica, A. O., Hankeln, T. & Schmidt, E. R. A novel serine/threonine kinase gene, STK33, on human chromosome 11p15.3. *Gene* **280,** 175–181 (2001).

135. Brauksiepe, B., Mujica, A. O., Herrmann, H. & Schmidt, E. R. The Serine/threonine kinase Stk33 exhibits autophosphorylation and phosphorylates the intermediate filament protein Vimentin. *BMC Biochem.* **9,** 25 (2008).

136. Brauksiepe, B., Baumgarten, L., Reuss, S. & Schmidt, E. R. Co-localization of serine/threonine kinase 33 (Stk33) and vimentin in the hypothalamus. *Cell Tissue Res.* **355,** 189–199 (2014).

137. Yang, T. *et al.* STK33 promotes hepatocellular carcinoma through binding to c-Myc. *Gut* **65,** 124–133 (2016).

138. Huang, L. *et al.* STK33 overexpression in hypopharyngeal squamous cell carcinoma: possible role in tumorigenesis. *BMC Cancer* **15,** 1–12 (2015).

139. Wang, P., Cheng, H., Wu, J., Yan, A. & Zhang, L. STK33 plays an important positive role in the development of human large cell lung cancers

with variable metastatic potential. *Acta Biochim. Biophys. Sin. (Shanghai).* **47,** 214–223 (2015).

140. Diodato, D. *et al.* VARS2 and TARS2 Mutations in Patients with Mitochondrial Encephalomyopathies. *Hum. Mutat.* **35,** 983–989 (2014).

141. Wang, Y. *et al.* A Human Disease-causing Point Mutation in Mitochondrial Threonyl-tRNA Synthetase Induces Both Structural and Functional Defects. *J. Biol. Chem.* **291,** 6507–20 (2016).

142. Longatti, A. *et al.* TBC1D14 regulates autophagosome formation via Rab11- and ULK1-positive recycling endosomes. *J. Cell Biol.* **197,** 659–675 (2012).

143. Lamb, C. a *et al.* TBC1D14 regulates autophagy via the TRAPP complex and ATG9 traffic. *EMBO J.* **35,** 281–301 (2016).

144. Schmidt, S. & Debant, A. Function and regulation of the Rho guanine nucleotide exchange factor Trio. *Small GTPases* **5,** 1–10 (2014).

145. DeGeer, J. *et al.* Hsc70 chaperone activity underlies Trio GEF function in axon growth and guidance induced by netrin-1. *J. Cell Biol.* **210,** 817–832 (2015).

146. Timmerman, I. *et al.* A local VE-cadherin and Trio-based signaling complex stabilizes endothelial junctions through Rac1. *J. Cell Sci.* **128,** 3041–54 (2015).

147. Sonoshita, M. *et al.* Promotion of colorectal cancer invasion and metastasis through activation of NOTCH–DAB1–ABL–RHOGEF protein TRIO. *Cancer Discov.* **5,** 198–211 (2015).

APPENDIX

- **LIST OF PRIMERS AND PCR REACTIONS CONDITIONS USED IN THIS WORK**

Table XIX- List of the primers used in this work

Primer	Primer Sequence (5'→3')	Amplicon Size	$MgCl_2$ [ ]	Desnaturation (time)	Annealing		Extension (time)	Cycles (number)
					Time	Temperature		
HRAS-1F	CAGGAGACCCTGTAGGAGGA	148 bp	1.8 mM	1 min	1 min	59 ºC	1 min 15 s	37
HRAS-1polR	GGGTCGTATTCGTCCACAA							
HRAS-2F	GATTCCTACCGGAAGCAGGT	140 bp	1.8 mM	1 min	1 min	59 ºC	1 min 15 s	37
HRAS-2R	ATGGCAAACACACACAGGAA							
KRAS-2F	TTAACCTTATGTGTGACATGTTCTAAT	175 bp	2.15 mM	1 min	1 min	57 ºC	1min 15 s	37
KRAS-2R	TTGTTGGATCATATTCGTCCAC							
KRAS-3F	AGGATTCCTACAGGAAGCAAG	141 bp	1.8 mM	1 min	1 min	59 ºC	1 min 15 s	37
KRAS-3R	TGGCAAATACACAAAGAAAGC							
NRAS-2F	GGTTTCCAACAGGTTCTTGC	145 bp	1.8 mM	1 min	1 min	59 ºC	1 min 15 s	37
NRAS-2R	CTCACCTCTATGGTGGGATCA							
NRAS-3F	CACCCCCAGGATTCTTACAG	148 bp	2.15 mM	1 min	1 min	59 ºC	1 min 15 s	37
NRAS-3R	TGGCAAATACACAGAGGAAGC							
BRAF-15F	AAACTCTTCATAATGCTTGCTCTG	182 bp	1.5 mM	1 min	1 min	57 ºC	1 min 10 s	36
BRAF-15R	GGCCAAAAATTTAATCAGTAGGA							
BLMH-F	TTCTTGTTTCAGGGCGATGC	231 bp	1.5 mM	1 min	1 min	58 ºC	1 min	35
BLMH-R	CACCTATTATGAAAGGGCAGGC							
CASP8AP2-F	CTGCAAGGACATTTGATACAGTT	246 bp	1.5 mM	1 min	1 min	58 ºC	1 min	35
CASP8AP2-R	CCTGATAATGAGAACGTGACCA							
CDKN2AIP-F	AGGATCTGGGAAAGCTTCTGA	217 bp	1.5 mM	1 min	1 min	59 ºC	1 min	35
CDKN2AIP-R	TCCTGAGGAGCCCAACAAG							
KCTD16-F	CCGATTTTGGTTTGTGGA	247 bp	1.5 mM	1 min	1 min	58 ºC	1 min	35
KCTD16-R	GTGTAGCTTGACCAGATCTTGT							
LFNG-F	CCTCACCTACTACCTTCCCC	204 bp	1.5 mM	1 min	1 min	59 ºC	1 min	35
LFNG-R	CGATCACACCTGCTGAGAGA							
SPRY4-F	TGTAAATGCAAGGAGTGTGCA	235 bp	1.5 mM	1 min	1 min	58 ºC	1 min	35
SPRY4-R	CGGAGAGAGCACCCATGAA							
TBC1D14-F	CCCAGAGGCTTTTCTTCCAG	236 bp	1.5 mM	1 min	1 min	58 ºC	1 min	35
TBC1D14-R	ACCCTTTTCTCGTGGAGCTT							
ATHL1-F	CCATTGGCCTCAGTCTTCTC	172 bp	1.7 mM	1 min	1 min	59 ºC	1 min	35
ATHL1-R	CTGACCTTGAAGGGTTCAGC							
DNAH17-F	GGGCTGAGTCTGTGGAGAAC	216 bp	1.2 mM	1 min	1 min	59 ºC	1 min	35
DNAH17-R	CAAAGTGCTGGCATTACAGG							
ERICH6B-F	CATTTTCTCCAGAGGGAGAGT	375 bp	1.5 mM	1 min	1 min	57 ºC	1 min	35
ERICH6B-aR	TCTTCCCCAGATACTCCTC							
GXYLT1-F	CGACCAGTTGATGATATTTGG	184 bp	1.7 mM	1 min	1 min	57 ºC	1 min	35
GXYLT1-R	TCCTTCTCATTCGAGTCATGT							
LILRB1-F	GAGTCCTGTGACCCTCAGGT	192 bp	1.5 mM	1 min	1 min	58 ºC	1 min	35
LILRB1-R	CCTGCAGTGTCGCTACCATA							
MTCH2-F	TTTTCCAGGGGAATATGAGC	237 bp	1.7 mM	1 min	1 min	58 ºC	1 min	35
MTCH2-R	AAGGCAGACACCAAACACCT							
STK33-F	TAAGCCGTCCACTGAAGAAA	158 bp	1.7 mM	1 min	1 min	59 ºC	1 min	35
STK33-R	AAATGTTCCAGGTGTGGTGA							
TARS2-F	CACTGGGAGCATTATCAGGA	154 bp	1 mM	1 min	1 min	57 ºC	1 min	35
TARS2-R	AGCTCCCAGCTTACCAGTGT							
TIGD4-F	TCTGGATTTGGTTGCTGATG	468 bp	1.7 mM	1 min	1 min	57 ºC	1 min	35
TIGD4-R	CCAAGGATTAGCAGTTTTCCA							
TRIO-F	CTCGTTTGGCCCATCACT	156 bp	1.5 mM	1 min	1 min	58 ºC	1 min	35
TRIO-R	TCACTCACGCATGCCTTC							

Bp, base pairs; mM, millimolar; min, minutes ; s, seconds; ^{O}C, temperature in Celsius

- **SEGREGATION STUDIES OF GENOMIC VARIANTS VALIDATED BY SANGER SEQUENCING**

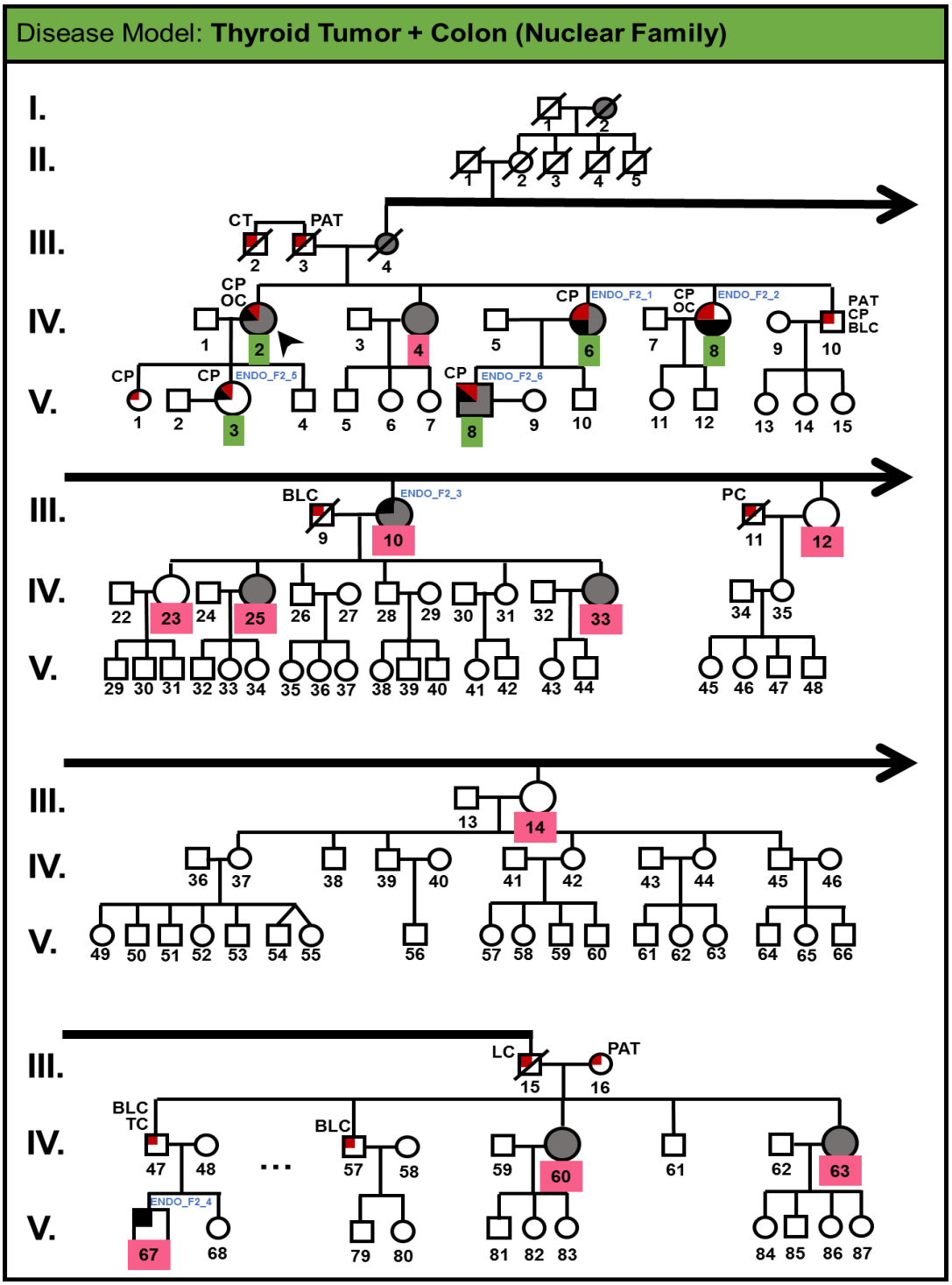


Figure 31- Disease model "Thyroid tumor + Colon (Nuclear Family)". Variants segregating according with this model were detected in the *ATHL1, KCTD16, SPRY4, STK33, TARS2* and *TIGD4* genes in family 2. The numbers highlighted in green indicate the family members that presented the variant(s) under study, and the numbers highlighted in pink indicate the members that were wild type. The symbol "…" in the last branch, indicates that some members of that branch are not represented in this pedigree. The remaining details are described in Figure 8 (Chapter 3)**.**

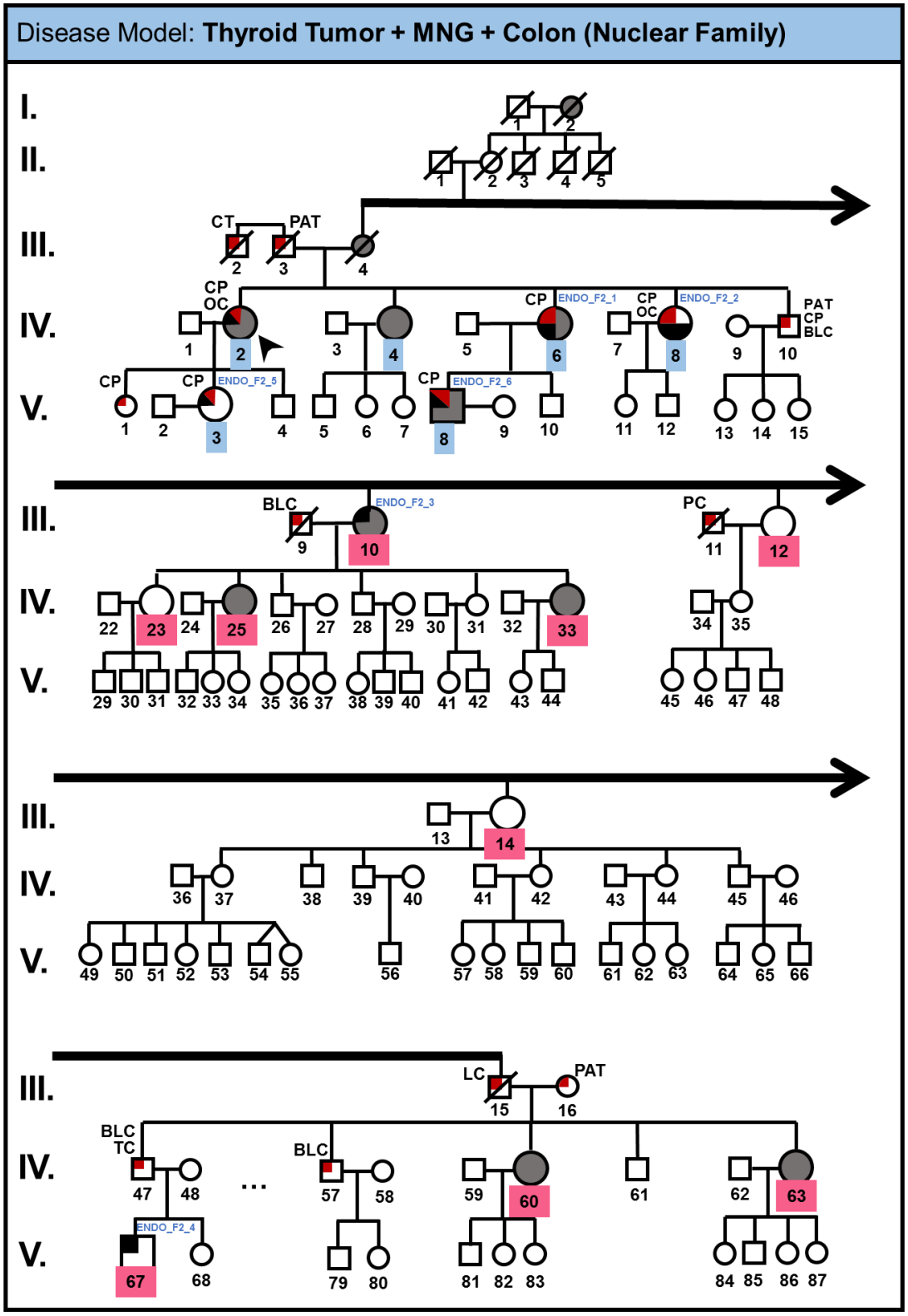


Figure 32- Disease model "Thyroid tumor + MNG + Colon (Nuclear Family)". Variants segregating according with this model were detected in the *CASP8AP2* and *TRIO* genes in family 2. The numbers highlighted in green indicate the family members that presented the variant(s) under study, and the numbers highlighted in pink indicate the members that were wild type. The symbol "…" in the last branch, indicates that some members of that branch are not represented in this pedigree. The remaining details are described in Figure 8 (Chapter 3)**.**

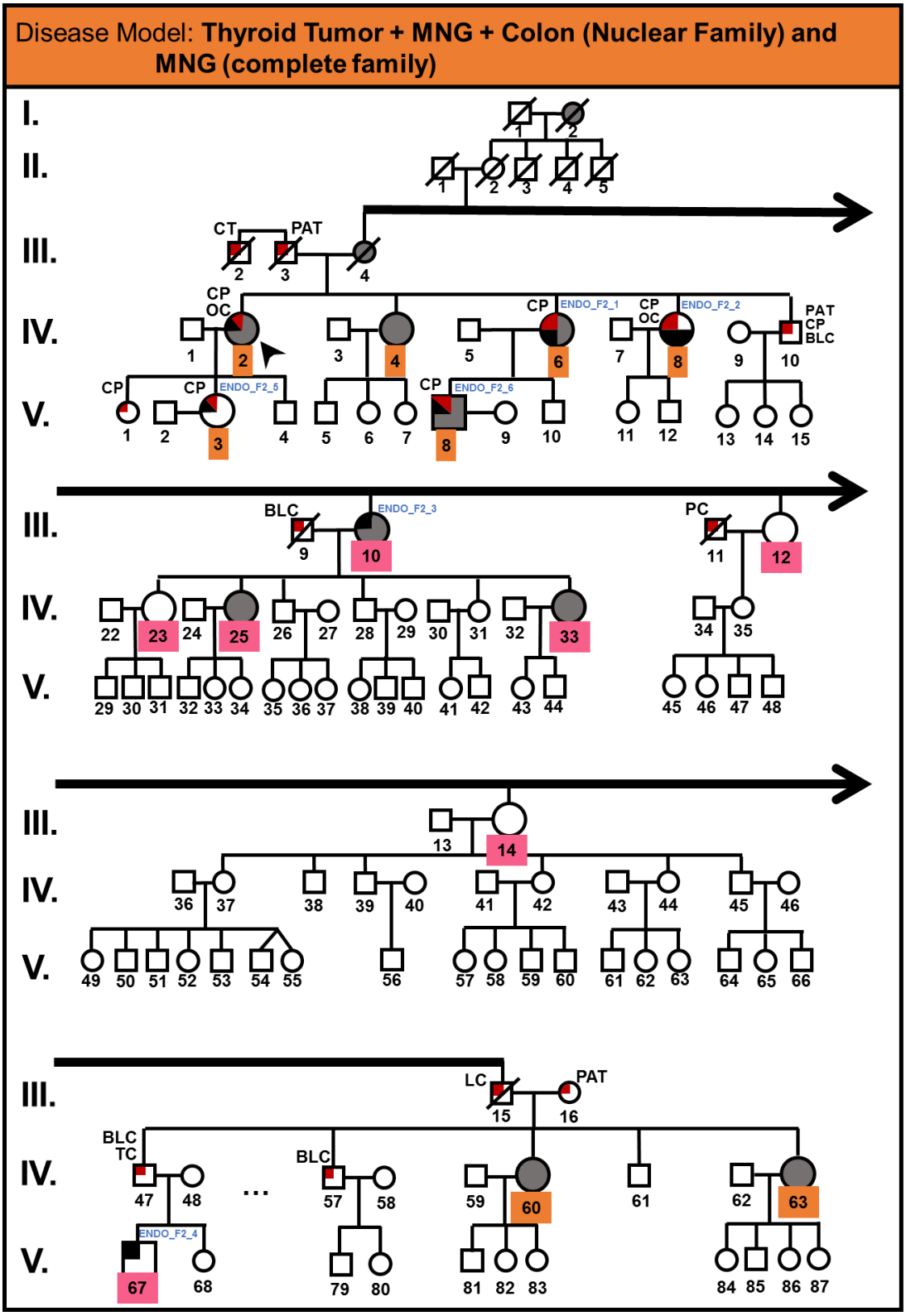


Figure 33- Disease model "Thyroid tumor + MNG+ Colon (Nuclear Family) and MNG (complete family)". Variants segregating according with this model were detected in the *TBC1D14* gene in family 2. The numbers highlighted in green indicate the family members that presented the variant(s) under study, and the numbers highlighted in pink indicate the members that were wild type. The symbol "…" in the last branch, indicates that some members of that branch are not represented in this pedigree. The remaining details are described in Figure 8 (Chapter 3)**.**

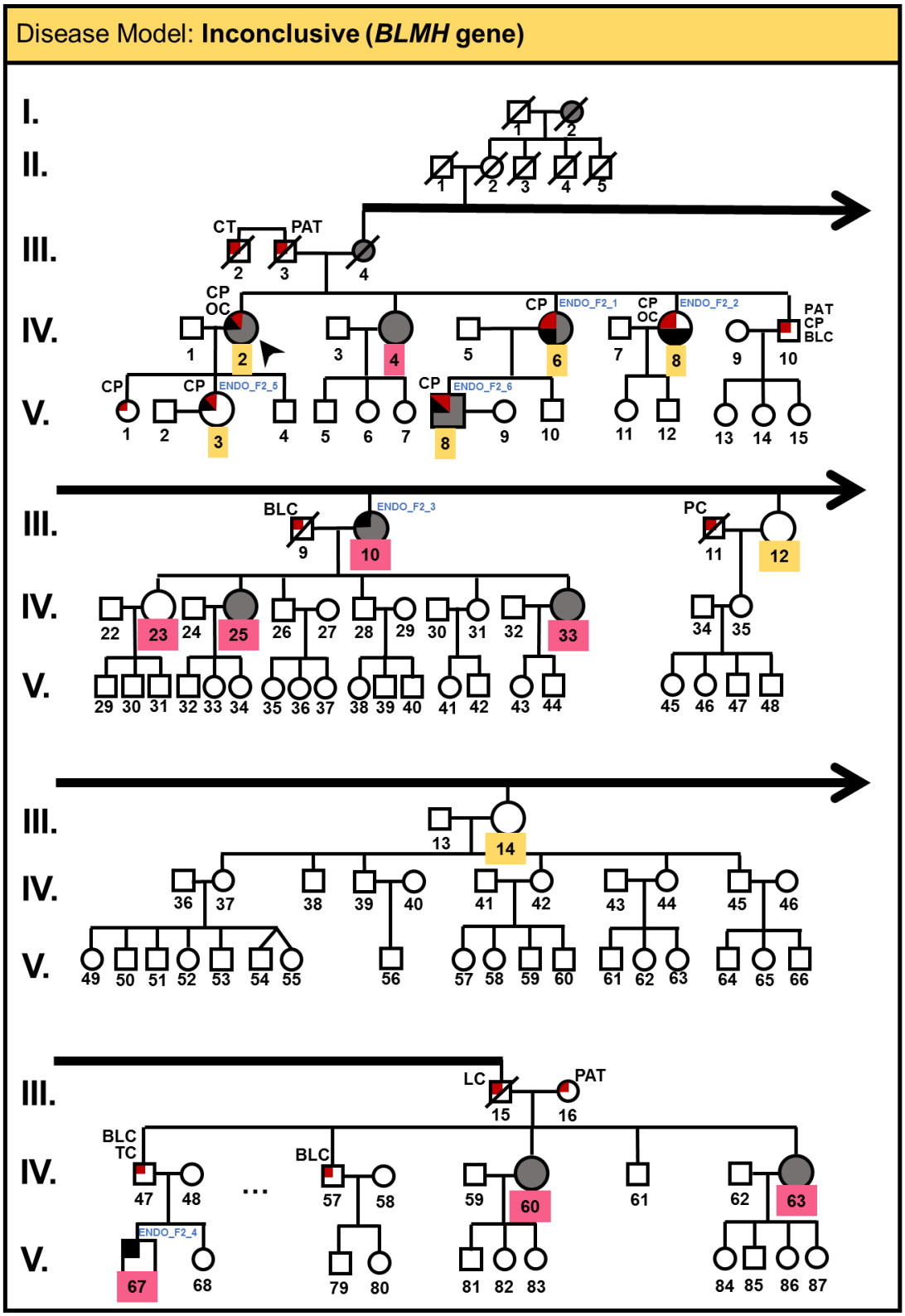


Figure 34- Disease model "Inconclusive 1". Variants segregating according with this model were detected in the *BLMH* gene in family 2. The numbers highlighted in green indicate the family members that presented the variant(s) under study, and the numbers highlighted in pink indicate the members that were wild type. The symbol "…" in the last branch, indicates that some members of that branch are not represented in this pedigree. The remaining details are described in Figure 8 (Chapter 3)**.**

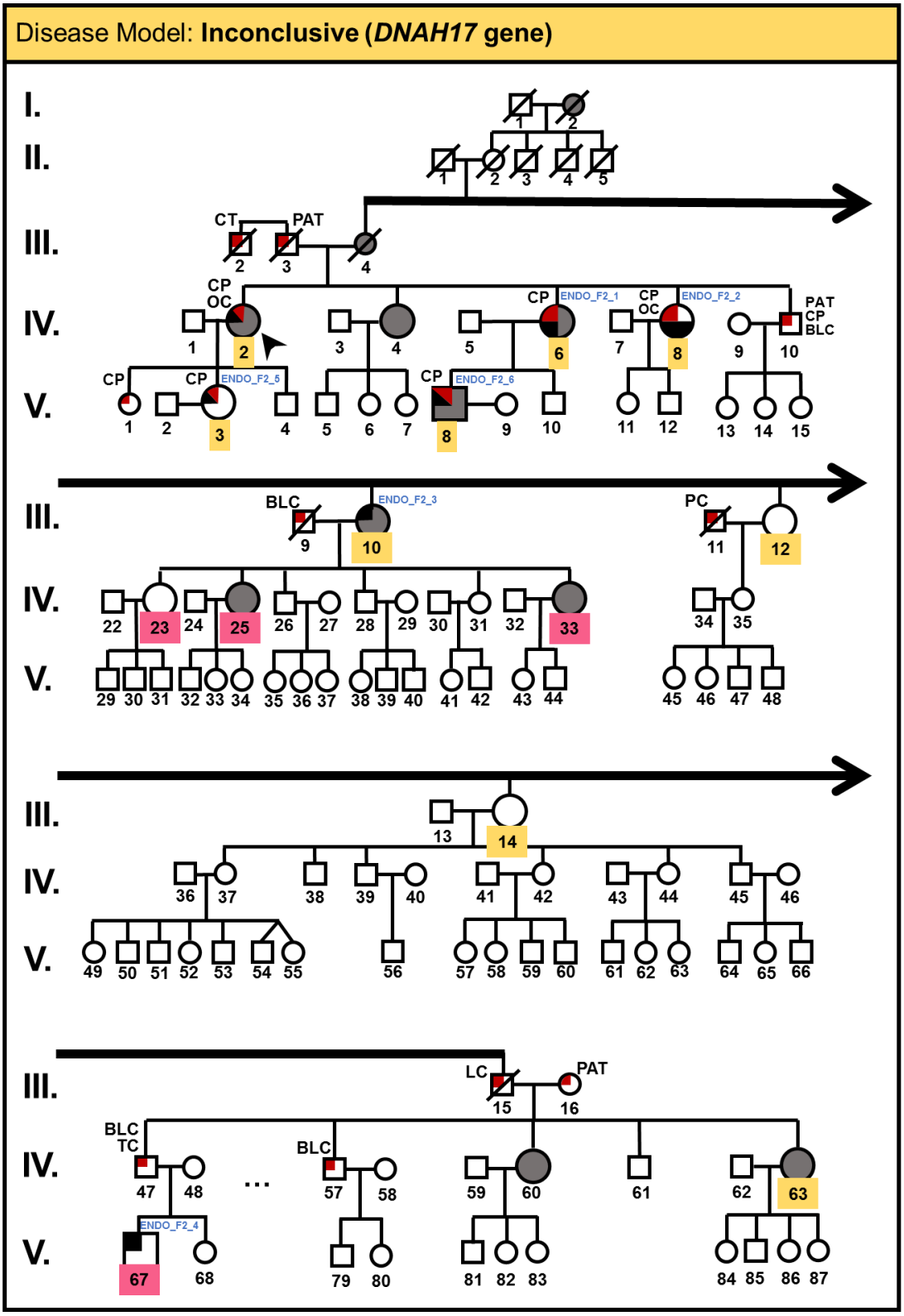


Figure 35- Disease model "Inconclusive 2". Variants segregating according with this model were detected in the *DNAH17* gene in family 2. The numbers highlighted in green indicate the family members that presented the variant(s) under study, and the numbers highlighted in pink indicate the members that were wild type. The symbol "…" in the last branch, indicates that some members of that branch are not represented in this pedigree. The remaining details are described in Figure 8 (Chapter 3).

- **LINKAGE ANALYSIS**

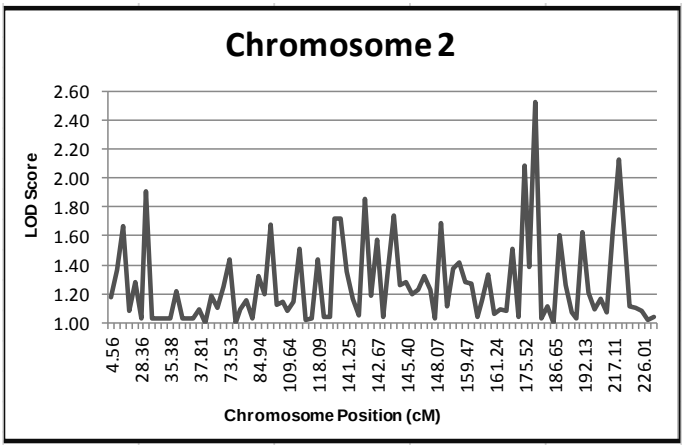


Figure 36-Linkage analysis at Chromosome 2. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM)

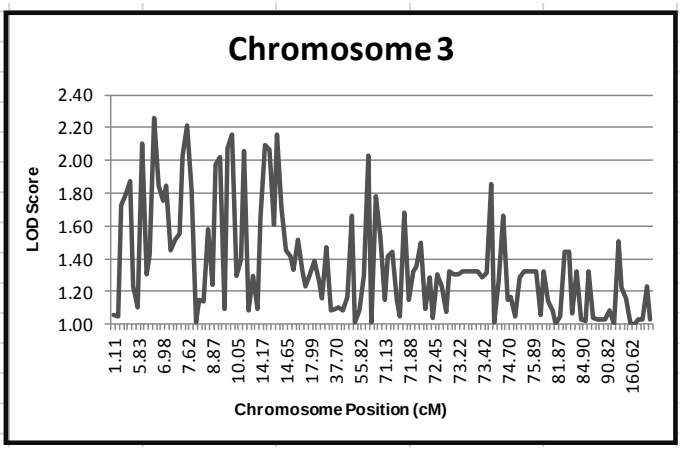


Figure 37- Linkage analysis at Chromosome 3. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM)

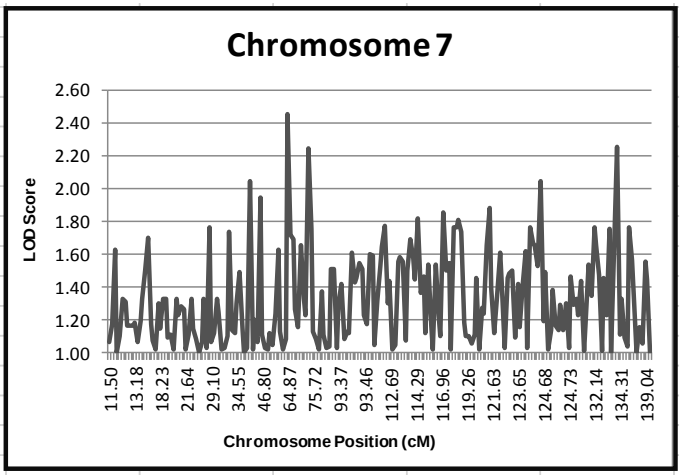


Figure 38- Linkage analysis at Chromosome 7. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM)

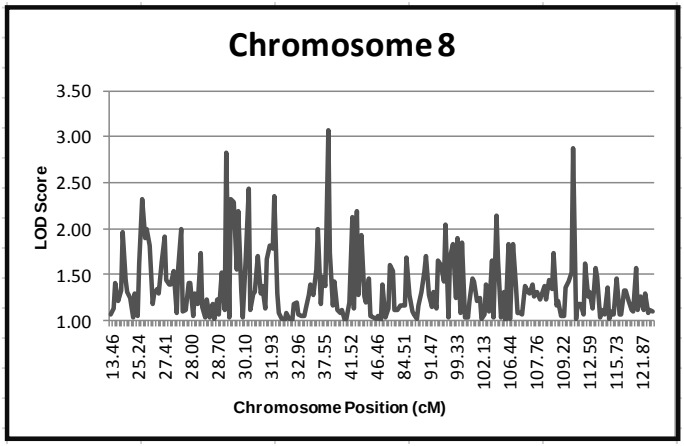


Figure 39- Linkage analysis at Chromosome 8. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

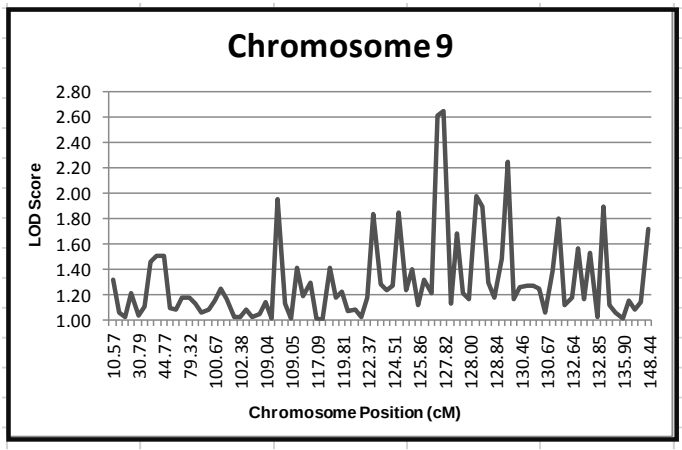


Figure 40- Linkage analysis at Chromosome 9. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM)

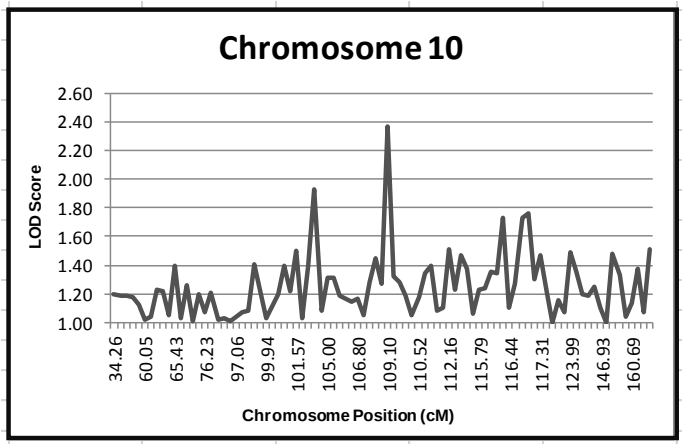


Figure 41- Linkage analysis at Chromosome 10. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

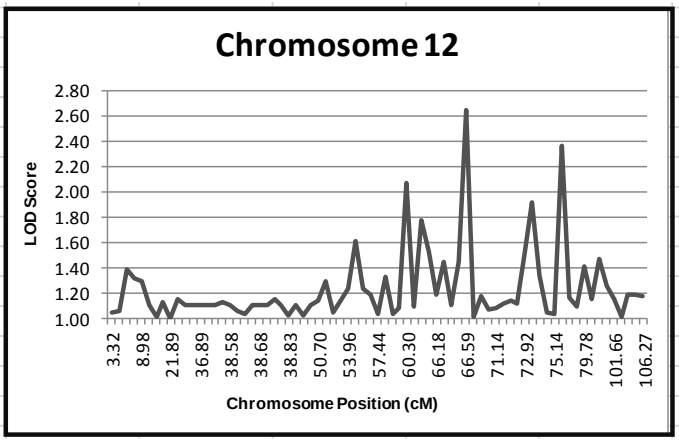


Figure 42- Linkage analysis at Chromosome 12. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

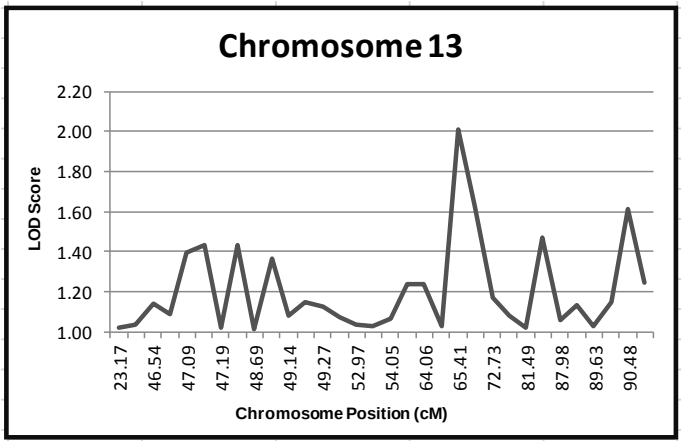


Figure 43- Linkage analysis at Chromosome 13. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

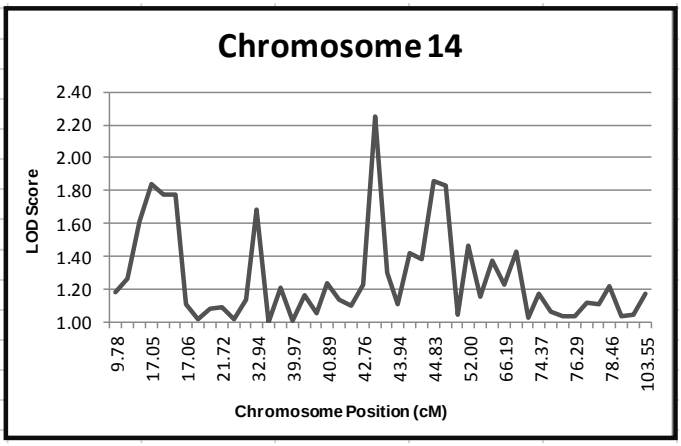


Figure 44- Linkage analysis at Chromosome 14. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

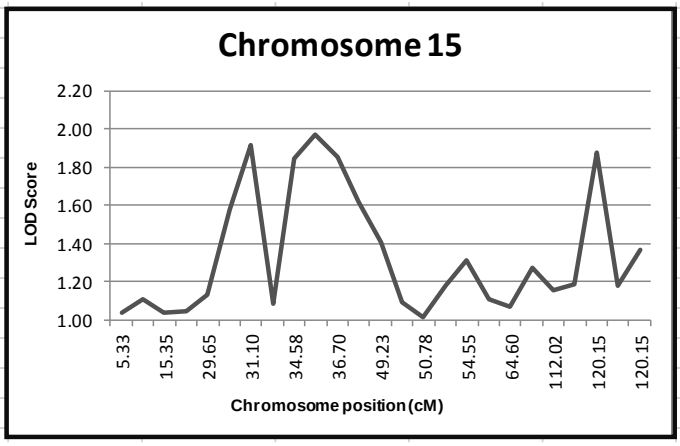


Figure 45- Linkage analysis at Chromosome 15. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

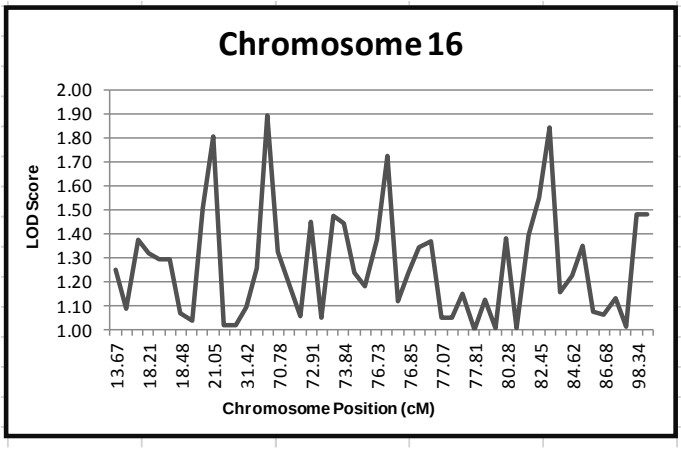


Figure 46- Linkage analysis at Chromosome 16. T The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

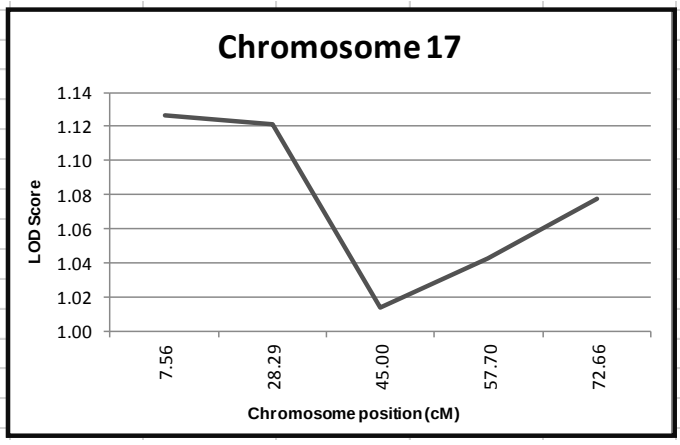


Figure 47- Linkage analysis at Chromosome 17. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

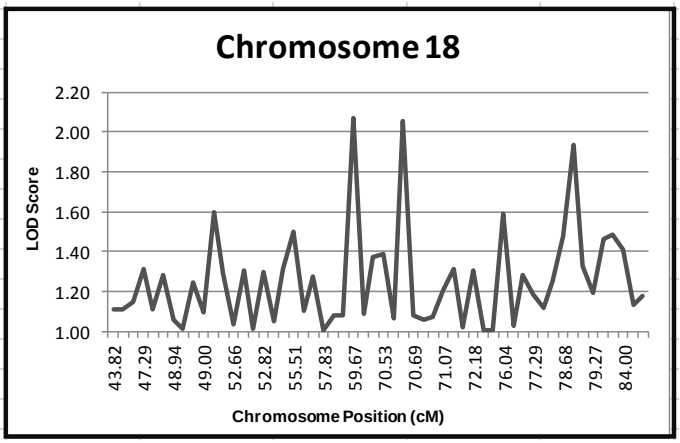


Figure 48- Linkage analysis at Chromosome 18. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

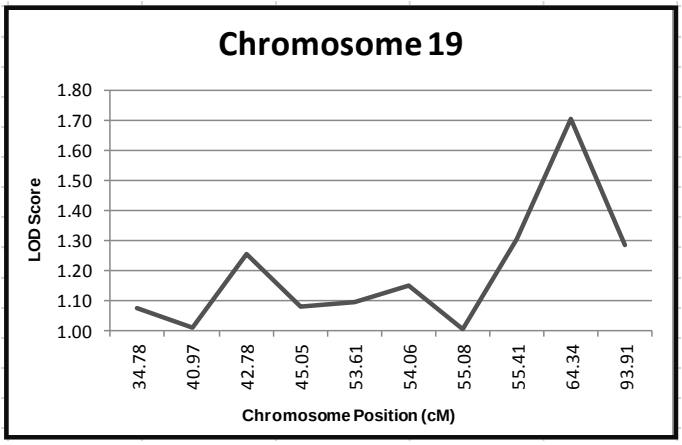


Figure 49- Linkage analysis at Chromosome 19. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

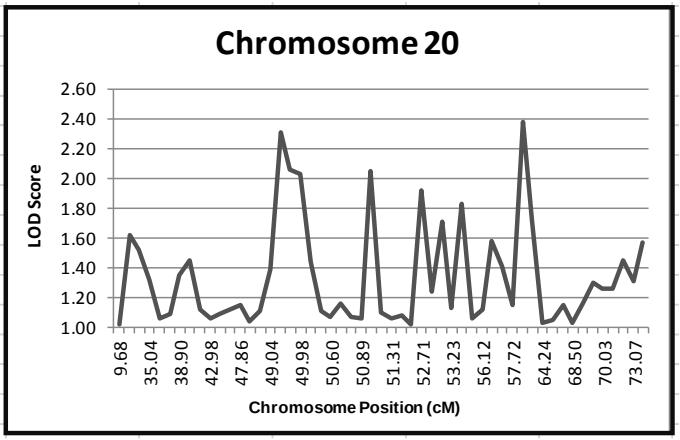


Figure 50- Linkage analysis at Chromosome 20. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

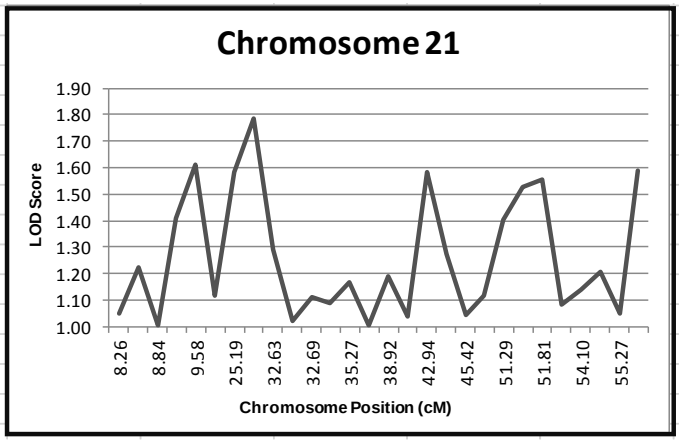


Figure 51- Linkage analysis at Chromosome 21. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

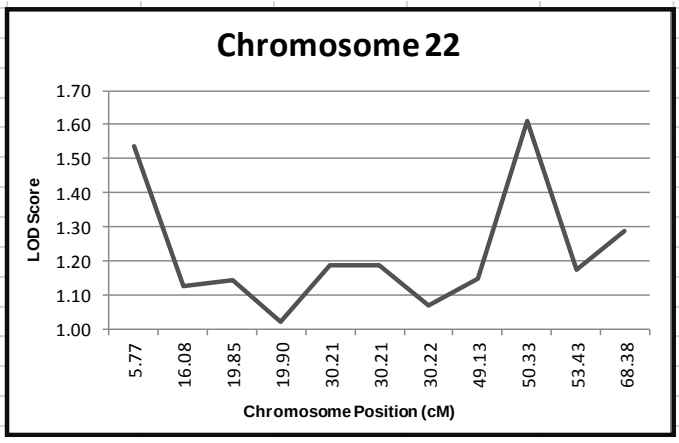


Figure 52- Linkage analysis at Chromosome 22. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

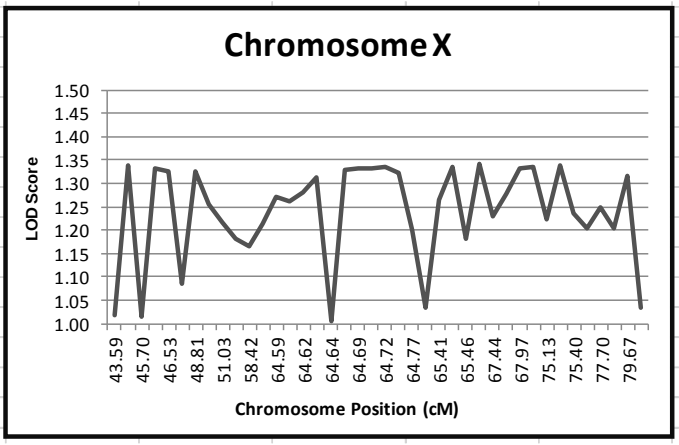


Figure 53- Linkage analysis at Chromosome X. The y axis indicates the LOD scores and the x axis indicates chromosome positions in centiMorgan (cM).

- **ANALYSIS OF CANDIDATE VARIANTS IN OTHER FNMTC FAMILIES**

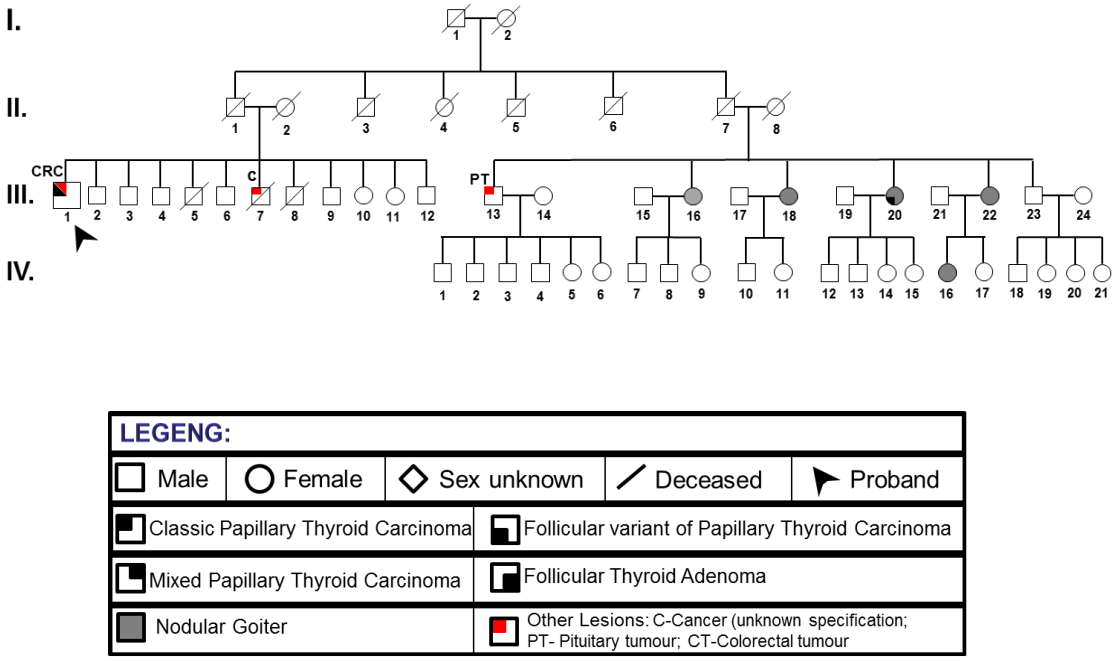


Figure 54- Pedigree of a Portuguese FNMTC family with a clinical spectrum similar to family 2. In this pedigree are represented 4 generations of this family.