MOLECULAR CHARACTERISATION OF

BENIGN AND MALIGNANT THYROID DYSFUNCTION

AMEEN D. Q. BAKHSH

Main Supervisor Professor Marian Ludgate

Co Supervisor Dr. Marian Hamshere

Institute of Molecular & Experimental Medicine

Institute of Psychological Medicine & Clinical Neurosciences

School of Medicine, Cardiff University

PhD 2014

Declaration of Originality

I hereby declare that the work embodied in this thesis is the result of my own independent investigations unless otherwise stated. This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any other degree.

Candidate:

Ameen Bakhsh

Director of studies

Prof. Marian Ludgate BSc (Hons), PhD

Library Loan and Photocopy

I hereby consent for my thesis to be made available for inter-library loan, and for the title and summary to be available to outside organizations.

Candidate:

Ameen Bakhsh

Summary

Nodular thyroid disease is common (prevalence 2-6%) and is a significant risk factor for the development of thyroid cancer.

My aim was to apply genome-wide- linkage-analysis to identify the defect in a large kindred with multi-nodular goitre (MNG) of adolescent onset progressing to papillary thyroid cancer (PTC).

Genomic DNA from 18 individuals (8 affected) was hybridized to Affymetrix GeneChip® Human Mapping 10K 2.0 Arrays. Results were analysed with Affymetrix GTYPE software to produce a call rate of ~92%. Extensive quality control steps were performed (PLINK, GRR) prior to linkage analysis using Merlin software in multipoint non-parametric and parametric (dominant) model.

A non-parametric LOD score of 3.01 was obtained on chromosome 20 across 20cM, the same region produced a dominant LOD score of 2.03. Haplotype analysis reduced the region of interest to 3.7 Mbp, (encodes 10 genes). Analysis of copy number variation in an affected individual (Illumina Human 660W-Quad) revealed an intronic deletion of ~1000 bp in one copy of Phospholipase-C β1 (*PLCβ1*), (the first in the 10 gene list), which is present in all affected family members and carriers. The deletion contained 'ATAA' at the junction site and this InDel was found in 1 of 105 healthy unrelated people, a similar variant was reported in the database of genomic variants in ~1% of Europeans . The deletion was not present in 70 unrelated PTC patients but was found in 4/81 with MNG (all European); the deletion frequency in the general population vs. MNG gives a X^2 value of 5.076 (p=0.024).

PLCβ1 expression was measured in thyroid tissues from affected family members and subjects free of the InDel and were significantly higher in the former ($p< 0.02$).

In conclusion, the InDel identified in familial MNG occurs in a proportion of sporadic MNG. It predisposes to goitre formation, possibly by increasing PLCβ1 transcription and activating the diacyl-glycerol, PKC and MAPK pathways.

Acknowledgements

I must begin to thank Professor Marian Ludgate who with endless enthusiasm and encouragement has guided me through my PhD studies. I also thank my co-supervisor who guided me when worked on data analysis.

Many Thanks to the collaborators; Dr. Roderick J Clifton-Bligh, Dr. Martyn Bullock, Dr. Lei Zhang, Dr. Fiona Grennan-Jones and Ziduo Li.

Thank you also to Professor John Gregory for introducing this family, to Mr. David Scott-Coombes for providing thyroid tissues, Professor Colin Dayan for his support as a head of the department and Professor Dillwyn Williams for his work on histopathological part and introducing the second family.

Many thanks also to the Departments of Genetics and CBS department, for DNA sequencing and hybridization of the chip, and to the Department of Psychological Medicine, especially Prof. George Kirov, Dr. Nigel Williams and Dr. Lyudmila Georgieva who helped with GWS.

Many thanks also to my friends and colleagues in the Department of Medicine (Dr. Shazli Mohd Draman, Dr. Peter Taylor, Lynn Tailor) for their support, friendship and amusement over the past years.

A special thank-you to my mum, dad and uncle (Sedi Abdunnaser) for his endless support throughout the whole of my studies, and to my valuable wife, my daughter Esraa, my sons Ahmad, Abdulkarim and Almubasher, for providing a suitable environment to complete this over the past years.

Abbreviations

ABI	Applied Biosystems
AE	Elution Buffer
AFTN	Autonomously Functioning Thyroid Nodule
AIT	Apical Iodide Transporter
AITD	Auto Immune Thyroid Disease
AL	Lysis Buffer
Ala	Poly-Alanine Tract
ASH	Allele-Specific Hybridization
ATC	Anaplastic Thyroid Cancer
AW	Washing Buffer
BAC	Bacterial Artificial Chromosomes
BAF	B Allele Frequency
BAT	Brown Adipose Tissue
BMR	Basal Metabolic Rate
bp	Base Pairs
BSA	Bovine Serum Albumin
cAMP	Cyclic Adenosine Monophosphate
CAPZB	Capping Protein, Beta
CBS	Central Biotechnology Services
CCDC6	Encodes Coiled-Coil Domain-Containing Protein 6
CDK	Cyclin Dependant Kinase
CDKN2A	Cyclin-Dependent Kinase Inhibitor 2a
cDNA	Complementary DNA
CEU	European or Caucasian Population
CGH	Comparative Genomic Hybridization
CH	Congenital Hypothyroidism
CHB	Chinese Population
CHGR	Centre for Human Genetic Research
cM	CentiMorgan
CNV	Copy Number Variation
DAG	Diacylglycerol
DEHAL1	Iodotyrosine Dehalogenase 1
DEPTH	DEPression and Thyroid
DI-1	Deiodinase 1
DIT	Di-Iodotyrosines
dsDNA	Double-Stranded DNA
DTC	Differentiated Thyroid Cancers
DTT	Data Transfer Tool
DUOX 2	Dual Oxidase 2
EGF	Epidermal Growth Factor
ENCODE	Encyclopaedia Of DNA Elements
ER	Estrogen Receptors

ERK	Extracellular signal-Regulated Kinase
FA	Follicular Adenomas
FGF7	Fibroblast Growth Factor 7
FISH	Fluorescence In Situ Hybridization
FMTC	Familial Medullary Thyroid Cancer
FNA	Fine-Needle Aspirates
FNMTC	Familial Non Medullary Thyroid Cancer
FOXE-1	Forkhead Box E1
FTA	Follicular Thyroid Adenoma
FTC	Follicular Thyroid Carcinoma
GCOS	GeneChip Operating Software
GD	Graves' Disease
GPCR	G Protein-Coupled Receptor
GRR	Graphical Representation of Relationships
gsp	G-Protein Mutation
GTYPE	GeneChip® Genotyping Analysis Software
H_2O_2	Hydrogen Peroxidase
HapMap	Haplotype Map
HCA	Hurthle Cell Adenoma
HGF	Hepatocellular Growth Factor
Hhex	Homeobox
HMA10K	GeneChip® Human Mapping 10K 2.0 Array
HMM	Hidden Markov Model
Hpttg	Human Pituitary Tumour Transforming Gene
HT	Hashimoto's Thyroiditis
HTA	Human Tissue Act
I-	Iodide
I^{131}	Iodine-131
IBS	Identical By State
IBT	Identity By Type
IGF-1	Insulin like Growth Factor-1
IGF-1R	Insulin like Growth Factor-1 Receptor
IP3	Inositol Triphosphate
ISVs	Intermediate-Sized Structural Variants
JPT	Japanese Population
Kbp	Kilo Base Pair
LCVs	Large-Scale Copy-Number Variations
LD	Linkage Disequilibrium
LOD	Logarithm Of odds
LREC	Local Research Ethics Committee
LRR	Log R Ratio
m.u.	Genetic Map Unit
MAF	Minor Allele Frequencies
MAPK	Mitogen Activated Protein Kinase
MAS	Microarray Suite

Mb	Mega Base
Mbp	Mega Base Pair
MCT8	Mono Carboxylate Channel
MDS	Multidimensional Scaling
MEN 2	Multiple Endocrine Neoplasia type 2
MGH	Massachusetts General Hospital
miRNA	microRNA
MIT	Mono-Iodotyrosines
MM	Mismatch
MM	Master Mix
MOI	Mode of Inheritance
MTC	Medullary Thyroid Carcinoma
NADP	Nicotinamide Adenine Dinucleotide Phosphate
NCOA4	Encodes Nuclear Receptor Co Activator 4
NGF	Nerve Growth Factor
NIBD	Non-Identical By Descent
NIS	Sodium Iodide Symporter
NKX2-1	NK2 Homeobox 1
NP	Non-Polymorphic
npl	Non Parametric Linkage Analysis
NT-MNG	Nontoxic-Multinodular Goitre
NTRK1	Neurotrophic Tyrosine Kinase, Receptor Type 1
OdT	Oligo dT
ORF	Open Reading Frame
PAC	P1-derived Artificial Chromosomes
PAX-8	Paired Box 8
PBq	Peta Becquerel
PCR	Polymerase Chain Reaction
PDS	Pendrin
PedFile	Pedigree File
PEG	Polyethylene Glycol
PG	PicoGreen®
PI3K	Phosphatidylinositol-3 kinase
PIP2	Phosphatidylinositol Bisphosphate
PKA	Protein Kinase A
PKC	Protein Kinase C
PLC	Phospholipase C
PLCβ1	Phospholipase-C β1
PM	Perfect Match
PPAR-γ	Peroxisome Proliferator Activated Receptor Gama
PTC	Papillary Thyroid Cancer
PTEN	Phosphatase and Tensin Homolog
QPCR	Quantitative PCR
RAI	Radioactive Iodine
RAS	Relative Allele Signal

RB	Retinoblastoma
RELP	Fragment Length Polymorphism
RI	RNAas Inhibitor
RT	Reverse Transcriptase
rT3	Reverse T3
RTK	Receptor Tyrosine Kinase
SDs	Segmental Duplications
SF	Scatter Factor
SLC16A2	Solute Carrier Family 16, Member 2
SLC26A4	Solute Carrier Family 26, Member 4
SLC5A5	Solute Carrier Family 5, Member 5
SNP	Single Nucleotide Polymorphism
ssDNA	Single-Stranded DNA
stDev	Standard Deviation
T3	Tri-iodothyronine
T4	Thyroxine
TA	Toxic Adenoma
TBG	Thyroxine Binding Globulin
TD	Thyroid gland Dysgenesis
TFC	Thyroid Follicular Cells
TG	Thyroglobulin
TG-Ab	Thyroglobulin Antibody
TGF-ß	Transforming Growth Factor ß
TMNG	Toxic Multinodular Goitre
TNBC	Triple-Negative Breast Cancer
TPO	Thyroid Peroxidase
TPO-Ab	Thyroid Peroxidase Antibody
TRH	Thyrotropin Releasing Hormone
TSAB	Thyroid Stimulating Antibodies
TSHR	Thyroid Stimulating Hormone Receptor
TSHR	Thyroid Stimulating Hormone
TTF-1	Thyroid Transcription Factor 1
TTF-2	Thyroid Transcription Factor 2
TTN	Toxic Thyroid Nodule
Tyr	Tyrosine
V600E	Valine to Glutamine
USF	Upstream Transcription Factor
GWLA	Genome Wide Linkage Analysis
GWAS	Genome Wide Association Study
WGS	Whole Genome Scan
WHO	World Health Organisation
WT	Wild Type
YAC	Yeast Artificial Chromosomes
YRI	Nigerian Population
θ	Recombination Fraction

Table of Contents

List of Figures

List of Tables

Chapter 1 General Introduction

1.1 The normal thyroid

1.1.1 Anatomy

The thyroid is a butterfly-shaped endocrine gland located in the neck region and consists of two encapsulated lobes. The gland synthesizes thyroid hormone and calcitonin through two different cells, the epithelial (or follicular) cells, and the parafollicular C cells, respectively. Each lobe of the gland is composed of follicles, each of which consists of a single layer of cuboidal thyroid follicular cells (TFC) that surrounds a lumen that contains colloid. TFCs are the functional units of the mature thyroid and are responsible for the biosynthesis and secretion of the thyroid hormones (T3 and T4) (reviewed in[3]).

Thyroid stimulating hormone receptor (TSHR) is located at the basal membrane of thyroid cells (Figure 1.1) and is the main regulator of the gland, when Thyroid stimulating hormone (TSH) binds with it. The interaction activates the sodium iodide symporter (NIS), also known as solute carrier family 5, member 5 (SLC5A5), which is also located at the basal membrane and works as an iodide (I^-) pump. Mono carboxylate channel MCT8 (SLC16A2), is also located on the basal membrane, through which thyroid hormones are released into the extracellular fluid. On the opposite side of the cell at the apical membrane, Pendrin (PDS/SLC26A4) is located, through which, I^- is transported from the TFC to the follicular lumen, where colloid is located (Figure 1.1). Blood supply of the thyroid is via the superior and inferior thyroid arteries (reviewed in [3, 4].

The colloid is composed mainly of thyroglobulin (TG), which works as a matrix for thyroid hormone biosynthesis as it is the source of tyrosine (Tyr), the main component for thyroid hormones. Iodination of tyrosine residues occurs within TG, after the I^- is oxidized by thyroid peroxidase (TPO).

The thyroid cell nucleus contains several transcription factors; Thyroid Transcription Factor 1 (TTF-1, also known as TITF-1/NKX2.1), Thyroid Transcription Factor 2 (TTF-2 also known

as FOXE-1), Paired box 8 (PAX-8), and homeobox (Hhex). These factors play a role in the development of the thyroid gland, as they regulate the transcription of the genes encoding NIS, TPO, TG, TSHR and other genes within the thyroid (Reviewed in [3, 5]).

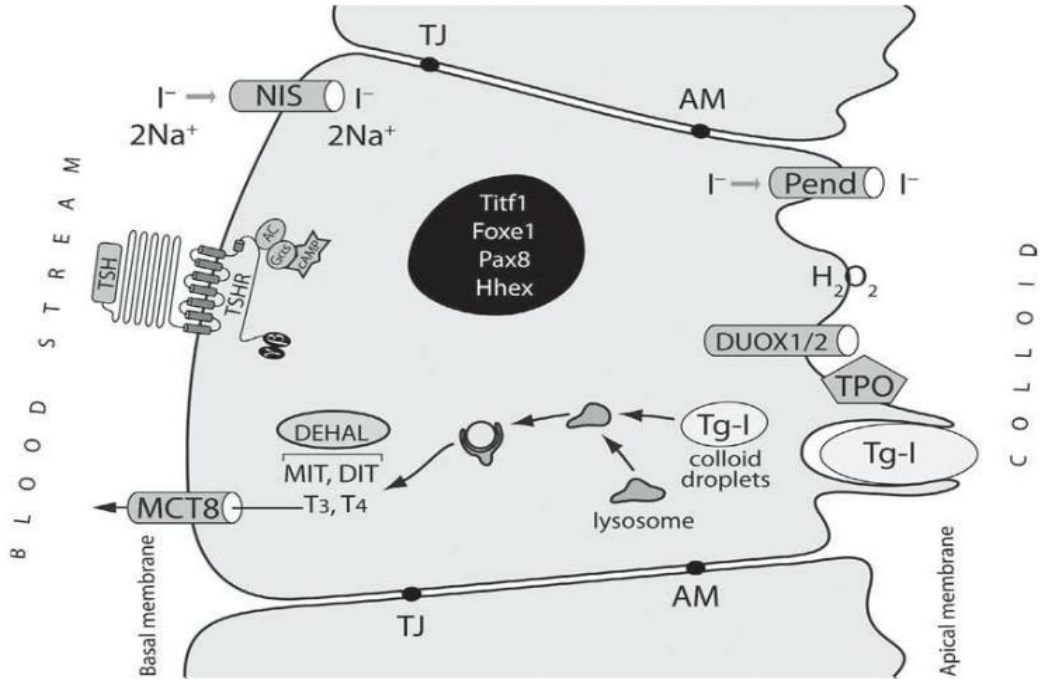


Figure 1.1: Schematic diagram showing the thyroid follicular cell with its major structures. Thyroid stimulating hormone receptor (TSHR), sodium iodide symporter (NIS) and mono carboxylate channel (MCT8) are located on the basal membrane. G protein s alpha (Gsα) activates the adenylyl cyclase (AC) and cyclic adenosine mono phosphate (cAMP) cascade. Iodotyrosine dehalogenase (DEHAL) removes iodine from mono-iodotyrosines (MIT) and di-iodotyrosines (DIT). The nucleus contains transcription factors [thyroid transcription factor 1 (TITF1), fork head box E-1 (FOXE-1), paired box 8 (PAX-8) & homeobox (Hhex]. Pendrin, thyroperoxidase (TPO) and other proteins involved in tri-iodothyronine (T3) & tetra-iodothyronine or thyroxine (T4) biosynthesis are located on the apical membrane. Using hydrogen peroxide (H_2O_2) generated by dual oxidase 1 & 2, (DUOX1 and DUOX 2), TPO oxidizes iodine;

the iodination of tyrosine residues within thyroglobulin (TG) then occurs. The end products are MIT & DIT which couple (catalysed by TPO) to form T4 and T3. Adopted from [3].

1.1.2 Physiological Thyroid Function

Hormone biosynthesis by the thyroid gland starts with active transport of I^- into thyrocytes. This process is known as iodide trapping and is performed by NIS, which is activated by TSH via cAMP. NIS increases thyroid iodine concentration to be 40 times higher than blood, and depends on the electrochemical gradient generated by the Na+, K+-ATPase. Then I^- is transported into the follicular lumen by Pendrin, in a process known as I^- efflux. The I^- is then oxidized by TPO, using H_2O_2 generated by (the calcium dependent flavo-protein NADPH) mainly, dual oxidase 2 (DUOX 2), (which is stimulated by PLC and PKC pathway) and probably (DUOX1), (which is stimulated via cAMP/PKA pathway) [6]. TPO iodinates selected tyrosyl residues on TG in a process known as organification or iodination, which leads to the formation of Mono-and Di-iodotyrosines (MIT, DIT). TPO then catalyses coupling of one MIT and one DIT to form Tri-iodothyronine (T3), while coupling of two DIT forms Tetra-iodothyronine or Thyroxine (T4) , (Figure 1.4). On the other hand iodotyrosine dehalogenase 1 (DEHAL1) enzyme, (also known as iodotyrosine deiodinase) removes iodine from iodinated tyrosine (MIT and DIT) and re-cycles iodide inside the thyrocytes (reviewed in [4]).

Iodinated thyroglobulin is stored as colloid in the follicular lumen and T4 (~80%) and T3 (~20%) are released (when required) into the blood stream, via MCT8 channel. T3 & T4 are bound to and carried by thyroid hormone binding proteins in the blood, such as thyroxine binding globulin (TBG) [7].

Although T3 is the more active form of thyroid hormone, only 20% is produced by the gland, while the rest is produced by exterior deiodination of T4 (at extra thyroidal sites), by deiodinase enzymes (will be described in 1.1.4), and the released iodide is recycled for hormone synthesis[8].

All reactions necessary for the production of T3 and T4 are positively regulated by TSH which is secreted by anterior pituitary thyrotropic cells, binds to the TSHR and activates the thyroid gland. Hypothalamus on the other hand enhances pituitary gland to secrete TSH via

synthesizing Thyrotropin Releasing Hormone (TRH). When serum concentration of T4 and T3 reach high levels, a negative feedback maintains the hypothalamic-pituitary-thyroid axis and inhibits TSH and TRH secretion (Figure 1.2). In addition an adequate supply of iodine is essential to maintain normal levels of thyroid hormone production. A daily intake of less than 50 micrograms is associated with goitre (will describe in 1.5.1). On the other hand iodide oxidation and organification are inhibited by higher iodine intake (reviewed in [3]).

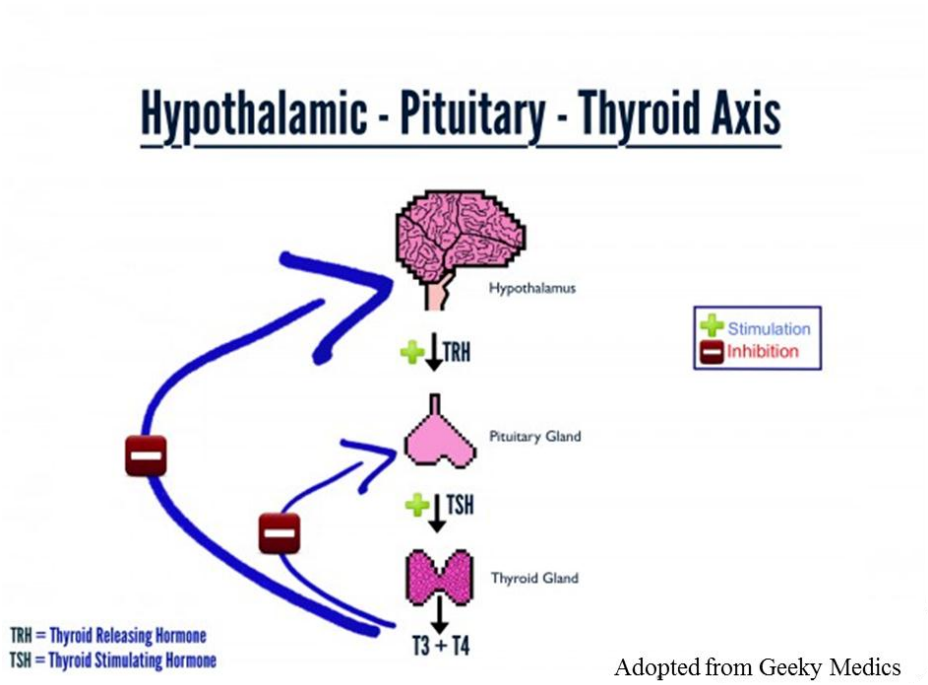


Figure 1.2; Cartoon showing the hypothalamic-pituitary-thyroid Axis. The hypothalamus secretes thyrotropin releasing hormone (TRH), which stimulates the pituitary to release thyroid stimulating hormone (TSH) to bind to thyroid stimulating hormone receptor (TSHR) and activate the thyroid gland. Negative feedback due to high levels of T3 and T4, which decreases the production of TRH and TSH, is indicated by the two blue arrows

TSHR plays the main role in the activation and growth of the thyroid gland. TSHR is a seven-transmembrane receptor and belongs to the G protein-coupled receptor family (GPCR). It couples a G protein, which contains α, β & γ subunits (Gα subunit has 4 families; G_i, G_s, G_q & $G_{12/13}$). When TSH binds to its receptor, it dissociates the trimeric G protein in to $G_s\alpha$ and Gβγ subunits. The main activation signal is mediated through $G_s\alpha$, which activates the Adenylyl Cyclase (AD) cascade that involves Cyclic Adenosine Monophosphate (cAMP) and Protein kinase A (PKA), (Figure 1.3). However, when TSHR is activated more e.g. by high TSH levels, the $G\alpha_q$ cascade is also involved and thus activates phospholipase C (PLC), which convert phosphatidylinositol bisphosphate (PIP_2) into diacylglycerol (DAG) and inositol triphosphate (IP3). The Gβγ subunits activate many signalling cascades including PLC, PIP2, DAG and IP3. The last two cascades activate mitogen activated protein kinase (MAPK) though BRAF and Protein Kinas C (PKC), (Figure 1.3) [9].

The main determining factor of the rate of I^- uptake is TSH (also known as Thyrotropin), which controls NIS activity. TSH also regulates thyroid hormone secretion and the number and size of TFCs, when it binds with its receptor TSHR [10-13]. Therefore any increase in TSH level, activates the thyroid gland, which produces more hormone, if excess hormone is produced it is known as hyperthyroidism (will be described in 1.3.1). That enhances the growth of the gland leading to its enlargement, a phenomena known as goitre (will be described in 1.5.1). On the other hand less production of TSH causes decreased activation of the gland, which leads to a low production of the hormones, if too little T4 is produced it is known as hypothyroidism (will be described in 1.3.2).

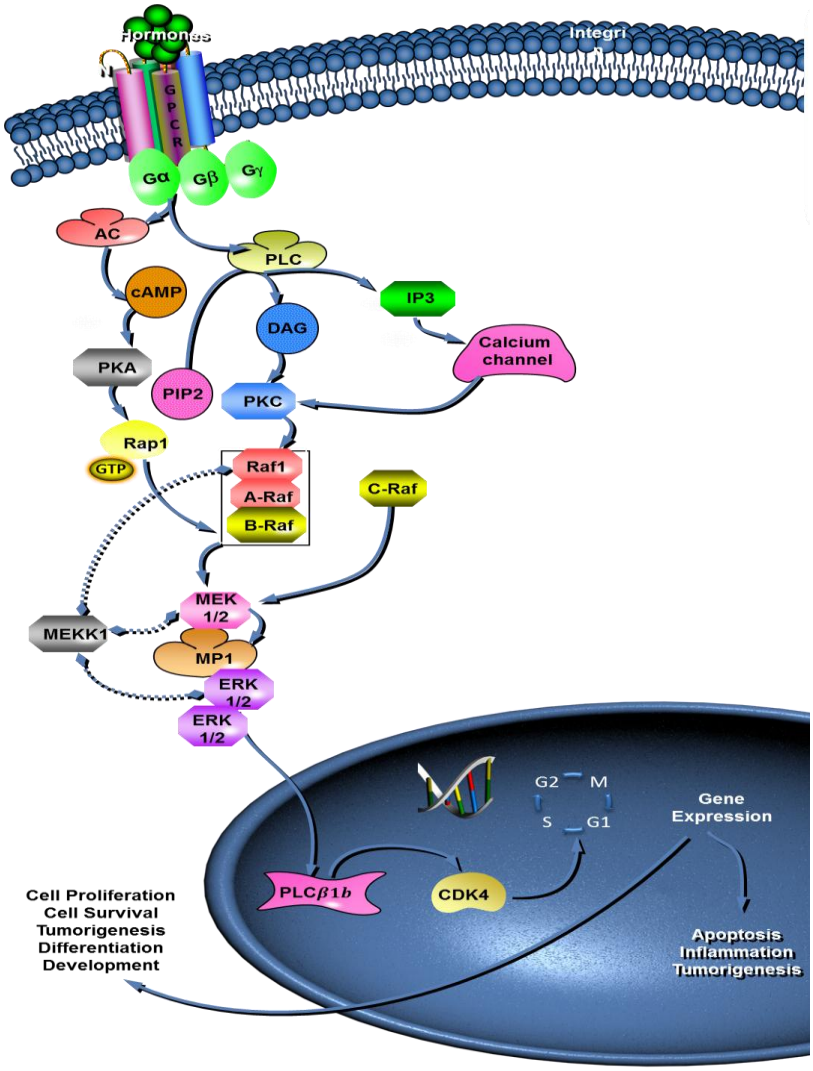


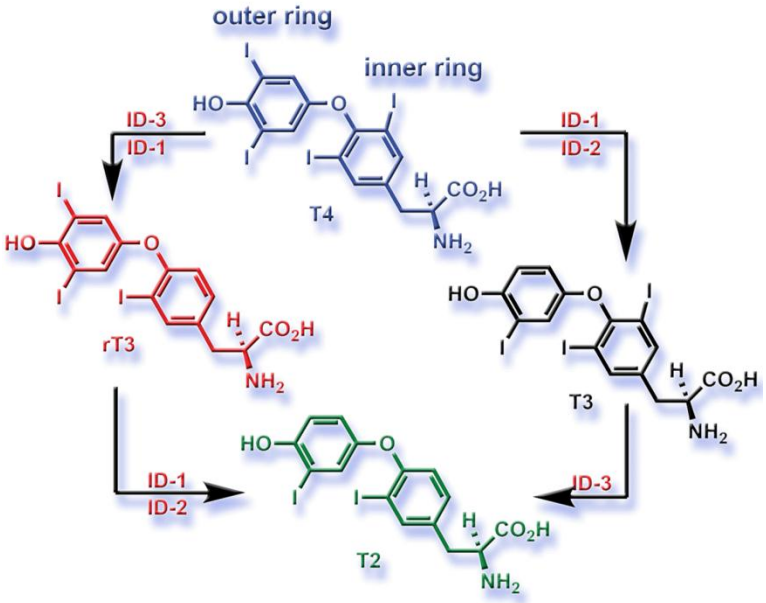
Figure 1.3; Cartoon showing a G protein-coupled receptor (GPCR) signalling cascade; 1) Ligands which signal through G protein α subunit (Gsα) activate cyclic adenosine monophosphate (cAMP) and protein kinase A (PKA). 2). Ligands which signal via Gαq subunits activate protein kinase C (PKC) and mitogen activated protein kinase (MAPK) via phospholipase C (PLC) and B- rapidly accelerated fibrosarcoma (B-RAF). A full description is in the text. Modified from; Ladas I et al 2013 (poster) and http://www.qiagen.com/products/genes

1.1.3 Functions of thyroid hormone

Thyroid hormones are important for normal growth and development of the newborn, mainly for brain development, thyroid defect in neonates can cause irreversible brain damage[14]. The World Health Organisation (WHO) considers iodine deficiency to be 'the single most important preventable cause of brain damage' worldwide[15]. Thyroid hormone plays a critical role in the maturation of other organs as well including skeletal, brown adipose tissue (BAT), liver, lungs, and heart[16]. T3 controls the basal metabolic rate (BMR) in adults by affecting the intermediary metabolism of fats, proteins, carbohydrates, and the rate of terminal oxidative metabolism. Thus the BMR will increase in cases of hyperthyroidism but decrease in hypothyroidism. Thyroid hormone effects the growth, development and function of most adult tissues including heart, muscle and brain. They have a role in maintaining metabolic energy balance, by increasing the number & size of mitochondria. Thyroid hormone increases bone turnover and osteopenia can occur in chronic hyperthyroidism [17]. Thyroid hormone stimulates hepatic gluconeogenesis and glycogenolysis as well as intestinal absorption of glucose, thus increases serum glucose (reviewed in [3, 18]).

1.1.4 Thyroid Hormones Metabolism

Although T3 is the active form of thyroid hormone, the thyroid gland produces more T4 (80%), but deiodinase enzymes convert T4 to T3 intracellularly. The deiodination is performed by three types of the enzyme which are distributed in the body; DI-1, DI-2, and DI-3. DI-1 is expresses in the liver and kidney, stimulated by T3, has low affinity for T4 and can deiodinase either inner or outer ring (Figure 1.4). DI-2 is founds in the pituitary gland, brain, brown fat, and skeletal muscle, and is more active when T4 concentration is low and is an obligate outer ring deiodiniser [19]. DI-3 is expressed in placenta, brain, and skin, deiodinases inner ring only and is the major T3 and T4 inactivating enzyme. DI-3 and DI-1 convert T4 to reverse T3 (rT3) by removing the outer ring iodine, which inactive the hormones, while rT3 and T3 can be further deiodinated to form Di-iodotyrosines in the liver (Figure 1.4). About 40% of the T4 and nearly all of theT3 that is produced each day is eventually deiodinated by inner ring deiodination, mostly via DI-3 (reviewed in [3, 20]).



Adopted from http://ipc.iisc.ernet.in/~dmanna/research.html

Figure 1.4; Cartoon showing the structure of thyroxine (T4) (blue) with inner ring and outer ring, and its conversion by deiodinase 1 or 2 (ID-1 or ID-2) (also known as Dio 1 or Dio 2) to T3 (active form), or by deiodinase 1 or 3 (ID-1 or ID-3) (also known as Dio 1 or Dio 3) to reverse tri-iodothyronine (rT3), (inactive form). The last structure is for di-iodotyrosines (DIT), (also called T2), which can be converted by ID-3 from tri-iodothyronine (T3) or by ID-1 & ID-2 from rT3.

1.2 Thyroid Proliferation

1.2.1 Development of the Thyroid

The thyroid gland has a dual embryonic origin, with the most abundant thyroid follicular cells, arising from the embryonic endoderm (thyroid anlage), which emerges as a visible bud. The thyroid develops from the anterior foregut endoderm in which progenitor cells expressing four critical transcription factors, NKX2.1, PAX-8, FOXE-1, and Hhex, assemble to form the thyroid bud [21]. Thyroid C cells (which secrete calcitonin) arise from the ultimobranchial bodies (originating from pharyngeal pouch of the embryo). Thyroid cell

specification occurs in parallel to morphological and biochemical changes that make these cells clearly different from their neighbouring cells. Later the thyroid lobes expand and the gland obtains its definitive form, with a narrow isthmus connecting the two lateral lobes. Fully differentiated follicular cells express different thyroid-specific transcription factors. The beginning of thyroid development is on the floor of the primitive pharynx, which is followed by a migration to reach its final position at the pharynx, where the two lobes expand and the follicles formed (reviewed in [3]).

After migration, thyroid cells express the essential proteins to start hormone biosynthesis. Expression of TG, TPO and TSHR starts first [22], then NIS, Duox1 & Duox2 [23, 24]. The biosynthesis of T4 starts before T3 [25]. Maturation of the hypothalamic–pituitary–thyroid system is also required for the hormone to be synthesised [26]. In addition to their function at the early embryonic stage, *NKX2.1*, *PAX-8, FOXE-1,* and *Hhex* have been shown to be playing an important role in the expression of these essential proteins of the gland. *FOXE-1* and *PAX-8* have binding domains in the promoter regions of the *TG, TPO, TSHR*, and *NIS* genes, however *NKX2.1* and *PAX-8* have overlapping binding domains in *TG* and *TPO* genes promoters. In addition *Hhex* has been shown to play a fundamental role not only in the formation of the thyroid but also in the functional differentiation of the gland [21, 27-29]. Although these factors are present in other tissues, they are only expressed together in the thyroid. NKX2.1 and FOXE-1 are required for differentiation of both follicular and C cells of the thyroid, while PAX-8 is important only in differentiation of follicular cells. On the other hand lack of *Hhex* causes absence of the above three factors while absence of *NKX2.1* and *PAX-8* causes absence of Hhex [3, 29].

Mutations in *FOXE-1, NKX2.1* or *PAX-8* have been shown to be the cause of severe congenital hypothyroidism (CH) in humans, while germ line deletions of these transcription factors in animal models have provided important information on thyroid dysgenesis as a genetic disease [30-32]. A recent study [33] has reported that a transient overexpression of NKX2.1 and PAX-8 is sufficient to direct mouse embryonic stem-cell differentiation into thyroid follicular cells that are capable for iodide organification in vitro. The cells have shown an ability to rescue thyroid hormone deficits in athyroid animals. That has opened a new field for treating hypothyroid patients using stem-cell technologies.

In addition insulin like growth factor-1 (IGF-1) and epidermal growth factor (EGF) can promote thyroid cell proliferation in culture [34, 35]. They are expressed during embryonic life and could be the primary regulators of thyroid growth at that time [36, 37], (will describe more in 1.1.2).

1.2.2 Physiological regulation of Thyroid growth

Human thyroid cells divide about five times in adulthood, which reveals that there is a slow constant turnover (division and death) of these cells. Adult thyroids maintain their size with a slow cell turnover, but keep the capacity to grow by cell hypertrophy and proliferation when stimulated. During adulthood TSH enhances thyroid growth at several stages as TSH signalling has been shown to be the best growth stimulus for adult thyroid cells. However, this signalling is not a global regulator of thyroid function during the embryonic stage [25, 27].

On the other hand, some non-thyroidal factors act as thyroid proliferators, such as IGF-1 [38], transforming growth factor ß (TGF-ß)[39], and EGF [40]. Increased expression of these factors has been reported in proliferating thyroid tissue. Although IGF-1 plays a role in regulating growth in children generally and has anabolic effects in adults, it may also effect thyroid function and growth [41]. Cooperation between TSH, the major goitrogen, and IGF-1 to regulate thyroid growth has been suggested, as studies showed both to be required for cell growth [42]. Moreover, enlarged thyroid and reduced TSH levels were also reported when IGF-1 and IGF-1 receptor (IGF-1R) were overexpressed in thyroid [43]. TGF-ß on the other hand, is a type of cytokine which has a role in the cell cycle (will be described in 1.6.2) and acts as an anti-proliferative factor [44]. In contrast, TGF-ß has been shown to down regulate NIS in *BRAF*^{V600E} mutant rat thyroid cells and its overexpression was observed in aggressive forms of human PTC in which absence of NIS was reported [45]. The growth factor, EGF binds to its receptor and stimulates cell growth, differentiation, proliferation and survival [46]. It has been reported to be increased in gene expression arrays of PTC tumours and shown to be a MAPK pathway activator [47].

Thyroid gland was shown to be more active in children and adolescents than in young adults, while elderly thyroid is less active[48]. Several studies [48-50]on T4 and T3 levels during adulthood have pointed out that T4 secretion is reduced by age, which affects T3 levels,

because most circulating T3 derives from the deiodination of T4. However some studies have shown decrease in T3 but still in the normal range. Others have not found a decrease in T3 even in very old individuals [51]. The fact that the mean calculated ratios of serum rT3, T4, free T3 and free T4 progressively increased with age suggest that the relative rate of T4 conversion to rT3 increases with age during childhood and adolescence. In the thyroids of both adults and children, TSH plays the main role in thyrocyte growth (reviewed in [3]).

1.3 Thyroid dysfunction

Thyroid dysfunction can cause hyper production of the T3 & T4 hormones (hyperthyroidism) or hypo production (hypothyroidism). In both conditions goitre (will be described in 1.5.1) can be formed. Goitre can be caused by Auto Immune Thyroid Diseases (AITDs), which cause either of the above conditions, depending on the disease type. They often affect several members of the same family, often over many generations. These disorders are polygenic multifactorial with environment also playing a role. Some families with hyperthyroidism and cases of CH, (will be described in 1.3.2) are the result of single gene defects.

1.3.1 Hyperthyroidism

A ‘continuous increase in thyroid hormone biosynthesis and secretion by the thyroid gland’ is known as hyperthyroidism. This leads to an increase in the T3 & T4, accompanied by TSH suppression due to negative feedback. The thyroid gland in hyperthyroid patients produces more hormone than the body requirements, mainly as a consequence of TSHR activation. This can be due to two common conditions, either thyroid self-regulation (autonomy), or a presence of autoantibodies, which bind to, and activate the receptor. In thyroid autonomy, the receptor is activated without ligation of TSH, often due to gain-of-function- mutations in the *TSHR*.[52]. Toxic multinodular goitre (TMNG), diffuse thyroid autonomy or toxic thyroid nodule (TTN) are manifestations of the condition. In the case of antibodies, the TSHR is stimulated by Thyroid Stimulating antibodies (TSAB), in a condition known as Graves’ disease (GD), which is an AITD. It is the main cause of hyperthyroidism as it is responsible for about 60% to 90% of the hyperthyroidism cases across the world. Both conditions lead to

an increase of cAMP and start the cycle of thyroid hormone production and growth of the gland (reviewed in [3]).

Toxic Multinodular Goitre (TMNG) is another form of Hyperthyroidism, in which autonomously functioning thyroid nodules (AFTNs) occur. These nodules also cause uni-nodular goitre, which is known as Toxic Adenoma (TA). Both TMNG and TA are more common in women over 60, while TMNG is more common (~50%) in areas with iodine deficiency than TA (~10%)[53]. However enough supply of iodine decreases thyroid autonomy. TMNG prevalence showed 73% reduced after doubling iodine content of salt in Switzerland for 15 year[53]. TA is characterised by increased I- uptake of nodular cell with suppression of uptake in the surrounding extra-nodular tissue (hot nodule), while TMNG can be seen with or without additional nodules and with normal or decreased uptake (cold nodules) (reviewed in [3]).

1.3.2 Hypothyroidism

Hypothyroidism is the most common clinical disorder of thyroid function, with multiple etiologies. It is a condition in which the circulating T3 & T4 level is decreased accompanied by increased TSH and TRH released as a consequence of suppression of negative feedback.

Nontoxic-Multinodular goitre (NT-MNG) is a form of hypothyroidism and is one of the most common endocrine diseases worldwide, affecting 500-600 million people in different areas of the globe [54]. It is related to iodine deficiency in many areas, which causes reduction in thyroid hormone synthesis. This reduction leads to an increase in TSH (by suppressing the negative feedback) and thus increase in both hormonal production and growth of the gland (endemic goitre) [54]. Nontoxic goitres can cause goitre in different types of thyroiditis (inflammation of the thyroid gland) such as Hashimoto's thyroiditis (HT), which is the most common form of primary hypothyroidism. HT is autoimmune disease in which patient's own immune system attacks the thyroid gland, making it unable to produce the hormones[55]. It is thought that Hashimoto's is the end result of lymphocytes becoming sensitized to thyroidal antigens, to produce autoantibodies. All forms of AITD are associated with the presence of autoantibodies to thyroglobulin antibody (TG-Ab) and thyroid peroxidase (TPO-Ab) [56]. Gender distribution in AITD is four females to one male, older patients have a much higher incidence [54].

Hypothyroidism can occur in childhood and is known as congenital hypothyroidism (CH). It is the commonest congenital endocrine disorder, affecting 1 in 3000 to 4000 newborns [57], which (if not treated on time) causes severe neurodevelopmental impairment, and has been shown to cause congenital heart disease as well [58]. CH is usually sporadic but up to 2% of thyroid dysgenesis, is familial [59]. The main causes of CH are genomic mutations in any genes related to thyroid function, development or growth and these mutations are often recessively inherited. These genes can be divided in two groups depending on their effect; those causing Thyroid gland Dysgenesis (TD), such as *TSHR* (causes non-syndromic CH), *Gsα, NKX2.1, FOXE-1* and *PAX-8* (cause syndromic CH)[60]. *FOXE-1* Homozygous mutations, within the fork head domain (100 amino acids long domain), have been reported as a cause of CH (agenesis, dysgenesis or ectopic)[61], e.g. A65V was reported in two Welsh male siblings with the CH due to malfunctioning of the thyroid gland and thyroid dysgenesis [62]. On the other hand, variations within *FOXE-1* poly-alanine tract may affect the susceptibility to TD. Studies on the poly-alanine tract suggested that FOXE-1 is more likely to be a TD susceptibility gene than a disease-causing gene[61]. Study of Italian subjects [63] reported 71% of the patients with 14 poly-alanine homozygous comparing with 85% healthy controls. Another Study of Italian subjects [64] showed that 14 poly-alanine homozygous present more in normal, while CH patient show more 16/16 and 14/16. Study on Japanese cohort[65] showed 14 poly-alanine homozygous in 96 normal and 43 TD cases.

The second group of genes causing CH are those associated with defects in the organification of iodide, leading to dyshormonogenesis such as *TPO*, *TG*, *NIS*, *PDS* (Pendred syndrome), and *DUOX2*. Furthermore, genes involved in iodothyronine transporter defects such as *MCT8*, are shown to be a third group causing CH (reviewed in[60]).

In 80 to 85 percent of CH cases, the thyroid gland is absent, abnormally located, or severely reduced in size (hypoplastic)[58].

Radioactive iodine (RAI) therapy and surgical treatment for hyperthyroidism can cause hypothyroidism, as well as severe iodine deficiency[66].

Secondary hypothyroidism occurs in cases of insufficiency of TRH secretion from the hypothalamus and/or insufficient secretion of TSH from the pituitary gland [67].

Since thyroid hormone has important action on organs and tissues throughout the body, the symptoms and signs of thyroid hormone deficiency are multi-systemic and variable in expression. Subclinical hypothyroidism, the mildest form of thyroid failure, may be asymptomatic or have non-specific clinical features (reviewed in [3]).

1.4 Nodular proliferation

Enlargement in the thyroid gland (goitre) can occur without any disturbance of its function. It is more common at puberty or during pregnancy. Goitre can be toxic (in hyperthyroidism) or nontoxic (in euthyroid or hypothyroidism). Various types of familial goitre could be due to inherited thyroid disorders. They appear during childhood and are often associated with signs of hypothyroidism.

Although many factors, such as age, gender, iodine intake and smoking are associated with nodularity of the thyroid, genetic factors also play a major role. A study has shown that children of parents with goitres have 2.7 fold higher risk of developing goitre [68, 69].

Thyroid nodules are very common, but few studies have been done to investigate their origins. One study [70], in iodine deficiency area was performed on a cohort with ages between 41 and 71 years and reported 46% of cold nodules (hypo functioning), 44% isofunction nodules (normal functioning) and 6% hot nodules (hyper functioning). In another two populations of 119 patients (from iodine deficient area) and 2537 controls (from iodine sufficient area), thyroid nodules were found in 5.1% of the patient and 1.9% of the control with cold nodules 2.5 times more frequent in patients [71]. On the other hand thyroid autonomy has been reported in 40% of 236 patients with euthyroid endemic goitres in an area of iodine deficiency[72].

Furthermore, DNA oxidation and damage can be caused by higher concentrations of H_2O_2, leading to mutagenesis and apoptosis [73]. In human thyrocytes this can cause frequent mutagenesis in the thyroid gland, which might provide the source for the frequent nodular transformation of endemic goitres [74]. In TSH receptor signalling, the cAMP cascade is shown to inhibit H_2O_2 generation, whereas the PLC and Ca^{2+}/diacylglycerol cascade activates H_2O_2 production by DUOX2 stimulation and controlling TPO, which consequently control

thyroid hormone synthesis [75]. Furthermore, Iodine deficiency or autoimmunity cause diffuse thyroid hyperplasia [76], which increases proliferation and can lead to DNA damage due to elevated H_2O_2 levels.

1.5 Benign thyroid diseases (adenoma)

Benign diseases of the thyroid are quite common mainly in women with 5 times higher incidence than in men. There are three different types of benign thyroid disease; nontoxic, toxic and inflammatory [54]. A survey in northern England has reported goitres in 5.9% with a female/male ratio of 13:1 and 5.3% of women with single and multiple thyroid nodules comparing with 0.8% of men, with higher frequency in women over 45 years [77].

The most common forms of benign thyroid disease are Goitre (an enlargement of the entire thyroid gland), Graves' disease (AITD) (described in1.3.11.3) and Thyroid nodules (a result of clonal proliferation of a single thyroid cell). It has been proved that a history of benign thyroid disease is a high risk factor for the development of thyroid cancer [78, 79].

1.5.1 Goitre

The presence of goitre in the absence of autoimmune thyroid disease, malignancy, or inflammation, is known as MNG or diffuse goitre. Both genetic and environmental factors are involved in the goitrogenic process, but iodine deficiency is the most important risk factor. MNG is common even in areas without iodine deficiency [80], suggesting that other factors such as genetic, may play a role in the development. Both, thyroid mass and gland function increase due to MNG. Normal hormone secretion rate makes the condition as eu-metabolic but goitrous. Usually the patient with MNG is unaware of any problem until the diagnosis is made during a routine physical exam or evaluation for another problem.

Radiation is the clearest risk factor for thyroid carcinoma in humans, but a history of goitre or benign nodules also appears to significantly increase the risk for developing thyroid carcinoma. The roles of iodine, diet, reproductive and hormonal factors still remain unclear. The understanding of predisposing genetic factors is increasing (reviewed in [3]).

MNG is a significant risk factor for thyroid cancer[81] mainly Papillary Thyroid Cancer (PTC), (will be described in 1.6), whose incidence globally has increased in recent years [82]. The terms nontoxic nodular goitre, adenomatous goitre, and colloid nodular goitre are used as descriptive when a MNG is found.

Familial thyroid cancers make up about 5% of cases, with earlier age of onset than sporadic disease. Several studies have reported candidate loci as factors causing nodular goitre, such as MNG1 locus on chromosome 14q13.3[83] and MNG2 locus on chromosome Xp22[84]. An association with 4 other susceptibility loci, three of which are associated with Familial Non Medullary Thyroid Cancer (FNMTC), has been reported on chromosomes 19p13.2, 2q21, 1q21 and 10q23 [85-88]. In addition, loci on chromosomes 3q26, 2q, 3p, 7q, 8p and 14q32, have been reported to be overlapping with familial goitre [89-91].

Genes implicated in familial goitre and cancers generally differ from those in sporadic disease, with the exception of *NKX2.1*[92] and *FOXE-1*[93].

Four genetic loci have been reported in genome-wide association studies, to be associated with thyroid volume: Two independent loci located upstream of and within capping protein, beta (CAPZB), one within fibroblast growth factor 7 (FGF7), and one on chromosome 16q23 [94, 95].

1.5.2 Thyroid Nodules

Nodules develop from a single cell having a growth advantage. They could be euthyroid, as in family #1 of this study, or hypo or hyper thyroid. Thyroid nodules may progress to neoplasia due to somatic mutation or over activated oncogenes (will be described in 1.6.1). Necrosis can lead to pseudocystic nodules due to an imbalance between angiogenesis and growth. Other mechanisms, including immune-toxicity and hyper-secretion of factors that increase permeability of the endothelium have been proposed in the pathogenesis of cyst formation (reviewed in [3, 96]).

In addition, genetic factors have been demonstrated to play a role in familial nodular goitre presenting at a young age [97]. Many CH cases are due to mutations in genes regulating thyroid physiology. Elevated TSH, which is a compensation mechanism to try to increase thyroid hormone levels and often present in CH, may produce a euthyroid goitre [3, 97].

1.6 Malignant Thyroid disease (Thyroid cancer)

Thyroid cancer occurs when the normal equilibrium of regulatory pathways is disrupted. That could be through enhancement of stimulatory pathways (e.g. growth hormones and proto-oncogenes) or deficiency in inhibitory pathways (e.g. tumour suppressors, normal cell-to-cell interactions, and cellular immortalization) [98]. Thyroid cancer is the commonest among all endocrine malignancies comprising 95% of such disorders [99, 100]. Its incidence is increasing with a global estimate of 0.5 million new cases in 2004[101].

Two main types of thyroid cancer are identified depending on the cellular origin of the cancer itself. The commonest type arises from Follicular thyrocytes (thyroxine producing cells) and is called Differentiated Thyroid Cancer (DTC). It can be either papillary carcinoma (the commonest, representing up to 80% of all thyroid malignancies), or follicular carcinoma (the most aggressive type, giving rise to metastases). Follicular carcinoma occurs in a slightly older age group than papillary (age is a very important factor in terms of prognosis), and is also less common in children. It occurs only rarely after radiation therapy in older people (Figure 1.5).

Hurthle cell adenoma (HCA) (Figure 1.5) can be classified as a very rare type of thyroid tumour and could be sub-classified from PTC and FTC. The Hurthle cells are enlarged epithelial cells which can be found in several other thyroid abnormalities such as HT, toxic and non-toxic goitre[102]. There is general agreement in considering Hurthle cell neoplasms as a subset of all differentiated thyroid cancer, regardless of the papillary or follicular growth pattern. Hurthle cell carcinoma should be included in a subset of thyroid neoplasms different from true follicular cancers [103, 104]

The second type is Medullary thyroid carcinoma (MTC) and is a neuro-endocrine tumour arising from calcitonin producing parafollicular thyroid cells (C-Cells)[105]. It is the third most common (5%) of all thyroid cancer, but more aggressive with high mortality rate if not treated. Approximately 25% of MTCs are classified as familial (FMTC) and are due to RET proto-oncogene mutation (will be described in 1.6.1) [106]. FMTC can occur alone or as part of Multiple Endocrine Neoplasia type 2 (MEN 2) and can be considered as a clinical variant of MEN 2A [107]. Clinically, MEN 2A can be characterized by MTC in combination with pheochromocytoma (neuroendocrine tumour of the medulla and the adrenal gland) and

hyperparathyroidism. MEN 2B can be characterised by MTC in combination with pheochromocytoma, Marfanoid habitus (changes in limbs, fingers & joints) and multiple mucosal neuromas [108].

However, sporadic MTC can coexist with tumours of the parathyroid gland and can be classified as MEN2. It can be defined by absence of family history and lack of germline RET mutation and other MEN 2A related tumours [109]. Risks of metastases are high in both sporadic and hereditary MTC [110]

Carcinoma cells in the above types can be identified under the microscope, but another type of thyroid carcinoma, Anaplastic Thyroid Cancer (ATC) is undifferentiated. This type has very poor prognosis and is resistant to cancer treatments and thus causes high mortality. ATC is very rare with prevalence of 1.3%. A personal history of a long-standing goitre, previous resection for PTC, and family history of other cancers were risk factors for ATC in a case–control study[111].

The early stages of thyroid tumour development appear to be due to an activation of several receptors, which are associated with the development of differentiated neoplasms [112]. Some cause benign toxic adenomas such as G-protein mutation (gsp) and TSHR, others, such as RAS, cause FTC or RET which causes PTC. PTC can also be caused due to gene rearrangements and chromosomal translocations such as RET/PTC, NTRK, and PAX-8/PPARγ. Tumour suppressor gene (*Tp53, RB*) alternations, on the other hand, are associated with ATC (reviewed in [98]). The presumed stages of non-medullary thyroid tumourigenesis are summarised in (Figure 1.5).

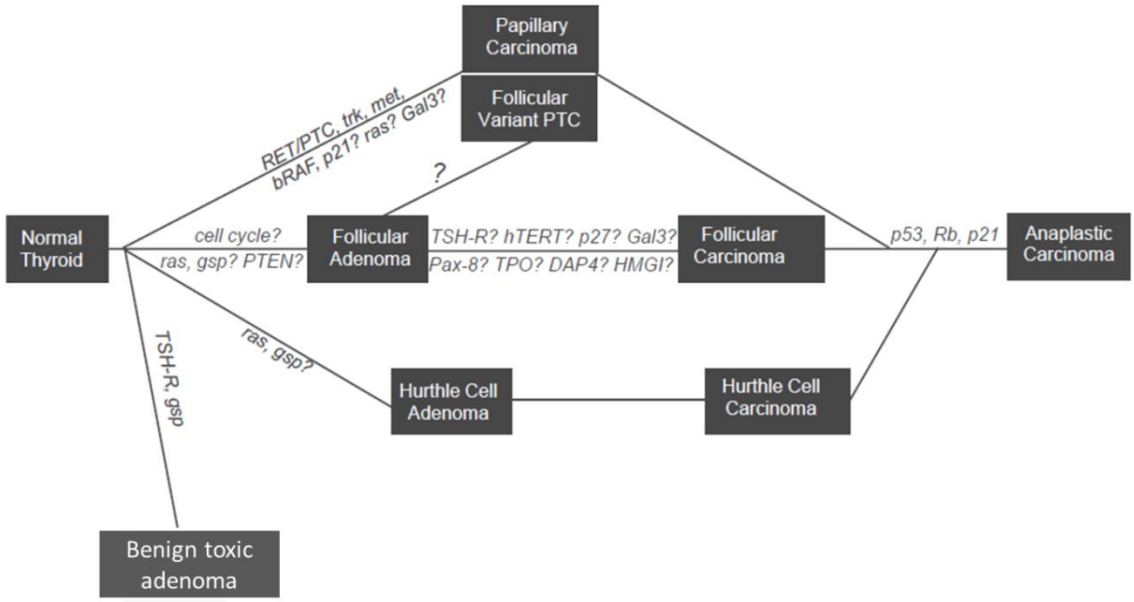


Figure 1.5; Cartoon showing presumed stages of non-medullary thyroid tumourigenesis, starting from a defect in a normal thyroid which leads to either benign toxic adenoma, papillary thyroid cancer (PTC), follicular thyroid cancer FTC, Follicular adenoma or Hurthle cell adenoma. These forms can be further developed to Follicular or Hurthle cell carcinomas. The later stage is anaplastic carcinoma. A full description is in the text. (adopted from[98]).

One of the main concerns of any thyroid group is the accurate distinction between benign and malignant nodules. That can mostly be achieved in papillary cases, when using the common test for initial thyroid nodule diagnosis, which is histology of fine-needle aspirates (FNA)[113]. However, this test cannot differentiate between thyroid tumours with a follicular pattern. This challenge is creating problems in distinguishing between (FTA) follicular thyroid adenoma (benign) and (FTC) follicular thyroid carcinoma (malignant), accurately. Therefore most guideline recommendations for follicular pattern are to be surgically removed (partial thyroidectomy), and examined histopathologically, looking for invasion through the tumour capsule or the blood vessels[101]. A study on approximately 6,300 FNAs of the thyroid was reviewed [114], only 8–17% of these cytologically suspicious nodules were

definitely malignant on histology, leading to the conclusion that distinguishing follicular or Hurthle-cell adenomas from follicular or Hurthle-cell carcinomas is almost impossible in FNA. It is often necessary to remove healthy thyroid to rule out carcinoma. Attempts to identify markers to improve diagnosis have compared gene expression profiles in FTC and FTA, and have identified 4 genes of potential value although these are still not widely used [101].

1.6.1 Proto-oncogenes & growth factors

Most proto-oncogenes (e.g. *RET, PPARγ, RAS & Met*) are genes that code proteins responsible for regulation of cell growth and differentiation. These proteins are activated through their ligand binding region (by growth factor) and involved in signal transduction and execution of mitogenic signals. Abnormal activation of proto-oncogenes converts them to tumour inducing agents, which are then called oncogenes. Although oncogenes are involved in the normal programed cell death (apoptosis), activated oncogenes can cause those cells that should die to survive and proliferate instead[115]. Most oncogenes require an additional step, such as mutations in another gene, chromosomal changes (re arrangements, translocation), or environmental factors to cause cancer. Activated oncogenes have been indicated in several types of thyroid cancers and include mutation, overexpression (gene amplification), or chromosomal rearrangement (chromosomal translocations mainly). Several reports have shown that genetic cases of sporadic thyroid cancers could be *RET* chromosomal re-arrangements [116], translocations between chromosome 2 and 3 generating a *PPARγ-PAX-8* fusion protein[117], mutations in *Ras* genes[118], and poly-alanine tract length variation in *FOXE-1*[91, 119]. Although *BRAF* mutations[120] causing constitutive activation, are the most frequent cause of PTC.

RAS belongs to a protein family encoding small GTPase proteins that function as binary molecular switches, control intracellular signalling and are involved in cell growth, differentiation and survival. Activating mutations in the *RAS* proto-oncogenes, have been found in a wide spectrum of benign and malignant thyroid pathologies, including MNG, follicular adenomas, FTC and PTC, suggesting it is an early event in thyroid tumourigenesis. Overall, around 40% of benign and malignant thyroid neoplasms display *RAS* activation, and the frequency of *RAS* mutations is higher in FTC (80%) than in PTC (20%) [118]. *RAS*

mutations at amino acids 12, 13 and 61 cause constitutive activation of RAS proteins and convert them into active oncogenes, generating a continuous flow of growth signals. All three *RAS* genes (*H-RAS, N-RAS* and *K-RAS4A* and *K-RAS4B*) can be activated in thyroid cancer (reviewed in [3]).

The *RET* proto-oncogene encodes a receptor tyrosine kinase (RTK), normally expressed in neuroendocrine cells, and the collecting duct of the kidney, but not normally present in thyroid follicular cells. Chromosomal re arrangements link the promoter and N-terminal domains of unrelated gene(s) to *RET* C-terminal, leading to an irregular expression of a chimeric form of the RTK in thyroid follicular cells which is constitutively active, without ligand binding requirement. This condition has been reported in PTC [116, 121]. There are at least 12 different types of *RET* rearrangements in PTC [122]. *RET/PTC1* and *RET/PTC3* are the most common, accounting for >90% of all rearrangements[121]. *RET/PTC1* is formed by fusion with the CCDC6 (H4) gene (encodes coiled-coil domain-containing protein 6), and *RET/PTC3* by fusion with the NCOA4 (ELE1) gene (encodes nuclear receptor co activator 4). NCOA4 interacts with PPAR γ and Androgen receptor [123, 124]. A study [125] on human PTC specimens has reported (*H4/PTEN*) intra-chromosomal rearrangements between *H4* and *PTEN*, a gene located on 10q22-23 (will be described in 1.6.2.1). In clinically detectable tumours, the presence of *RET/PTC* rearrangements varies widely, with frequencies between 0% and 59% reported in different series. The frequency is increased following exposure to radiation, and post-Chernobyl, the dramatic rise in paediatric PTC was associated with *RET/PTC* rearrangements. The mechanism is likely to be the increased incidence of chromosomal breaks following exposure of the thyroid gland to radiation [126].

Absence of any significant increase of thyroid cancers after the Fukushima accident in spite of screening (until now), would certify the high level of health care in Japan especially among children and adolescents [127]. The estimation of iodine-131 (I^{131}) release was 770 PBq (1 Peta Becquerel = 10^{15} Bq), which is about 15% of the estimated Chernobyl release of 5200 PBq. Additionally, the release of contaminated cooling water from the damaged reactor buildings resulted in a direct discharge of about 4.7 PBq of radioactivity into the Pacific Ocean. The ocean discharge of radioactivity was estimated to be about 11 PBq of I131. The total amount of radioactive material released into the air during the event was

revised downwards to 130 PBq for 131I, which represents about 11% of estimated Chernobyl emissions (reviewed in [128]).

Somatic rearrangements (on chromosome 1) of Neurotrophic tyrosine kinase, receptor type 1 (*NTRK1*), the proto-oncogene that encodes the receptor for nerve growth factor (NGF), are seen in PTCs [129]. They arise from the fusion of the 3 terminal sequences of the *NTRK1/NGF* receptor gene with 5 terminal sequences of various activating genes, such as *TPM3, TPR* and *TFG. TRK* onco-proteins display constitutive tyrosine-kinase activity [130]. *NTRK1* rearrangements have been found only in PTC, but with a lower prevalence than that reported for *RET/PTC*. Furthermore TRK receptors have been implicated in the onset and progression of MTC. Targeted over expression of one of the TRK family; TRK-T1 induces lesions resembling PTC in transgenic mice, suggesting its role in an alternative pathway for tumour initiation[131].

On the other hand germ-line *RET* mutations are shown to be more common in hereditary MTC (described in 1.6). These mutations are identified in more than 97% of cases of MEN 2A and 95% of FMTC. Mutations of the cysteine at codon 634 account for around 80% of germ-line *RET* mutations MEN 2A is an example[132]. A review in 2012 [108] has reported a total of 39 different *RET* germline mutations identified in FMTC. The most affected codons were 609, 611, 618, 620 (exon 10) and 634 (exon 11) in the extracellular domain, and codons 768 (exon 13) and 804 (exon 14) of the intracellular tyrosine kinase domain. On the other hand a high percentage (25-50%) of sporadic MTCs are caused by *RET* proto-oncogene activating mutations [133]. *RET* mutation met918thr is the most common somatic mutation, while met918thr germ-line mutation is less common and causes MEN 2B [134]. Mutations in codon 918 cause the most aggressive types of MTC[135]. Almost all patients with *RET* mutations will develop MTC. Although there is a clear genotype–phenotype relationship, aggressive disease can occur irrespective of the mutation involved. Somatic *RET* mutations that occur later in life and are limited to C cells are present in 40–50% of sporadic MTCs [136]. Although these tumours are heterogeneous, somatic *RET* mutations have been shown in up to 80% of sporadic MTCs, commonly 918ATG > ACG [137, 138], and less frequently at codon 634 and others (reviewed in [107].

Peroxisome proliferator activated receptor gamma (PPAR-γ) is member of the nuclear receptor superfamily, in common with the receptors for thyroid hormone. PPAR-γ is a regulator of adipocyte differentiation and plays a role in fatty acid storage [139]. Although the receptors for thyroid hormone and PPAR-γ have diverse effects on developmental and metabolic processes and have their own cellular functions, recent studies[140] have indicated crosstalk between the nuclear receptor family with hormones affecting the activity of other receptors leading to alterations in various biological functions including carcinogenesis.

Translocations between *PAX-8* on chromosomes 2 and the member of the nuclear hormone receptor superfamily (*PPAR-γ*) on chromosome 3 have been detected in a series of mainly follicular cancers. The translocation causes an in-frame fusion *of PAX-8 to PPAR-γ* [117]. The predicted fusion protein has been detected in follicular cancer tissue by immune precipitation and has been postulated to have a role in the pathogenesis of thyroid follicular carcinogenesis. While PAX-8, a paired box containing transcription factor, is central to the expression of several thyroid-specific genes including *TG* and *TPO*, the precise role of *PPAR- γ* in this context has not been elucidated. *PAX-8/PPAR-γ* rearrangements have generally not been reported in PTC, follicular adenomas, or MNG [117]. Moreover, activation of PPAR-γ reduces plasma thyroid hormones and deiodination level [141].

Met, (*c-Met*) is a proto-oncogene which encodes c-met protein, a membrane spanning RTK that regulates cell growth, tissue repair and survival. Mutations in *RTK* have been implicated in the progression of many human cancers. c-Met is a receptor for Hepatic growth factor (HGF), a liver regeneration mediator which is also known as scatter factor (SF)[142]. These growth factors are multifunctional cytokines that bind to the c-Met receptors and activate tyrosine kinase signalling, which stimulate mitosis and cell motility and thus play a role in tumourigenesis and tissue regeneration [143]. *c-Met* has been shown to be overexpressed in approximately 90% of PTC and is significantly associated with an aggressive phenotype [144]. It was recently reported that the *c-Met* gene is overexpressed in 37% of PTCs in Saudi patients [145]. *c-Met* is a potent thyrocyte mitogen whose expression is thought to be induced by *RAS* and *RET* oncogene activation[146]. Dysregulation of c-Met signalling has been shown to contribute to tumourigenesis in a number of malignancies, including thyroid cancer[147].

Growth factors IGF-1, TGF-ß & EGF play role in thyroid proliferators (as described in 1.2.2). In summary expression of any of the three factors has been increased in proliferating thyroid. TGF-ß overexpression was observed in aggressive forms of human PTC. EGF has been reported in gene expression array of PTC tumours and shown as a MAPK pathway activator. IGF-1 showed association with thyroid growth and goitrogenesis. It is known to be implicated in tumour cell apoptosis, invasion, transformation and metastasis, and to play a role in thyroid nodules. A study [148] on 62 thyroid tissues (blocks), including18 follicular adenomas (FA), 17 nodular goitres, 13 PTC, 2 FTC, and 12 normal controls, has reported that IGF-1 plays an important role in the genesis and development of certain solid cold thyroid nodules, including PTC, nodular goitres, and FA.

On the other hand, these three growth factors (IGF-1, EGF & TGF-α) are reported to enhance both expression and phosphorylation of the Human Pituitary Tumour Transforming Gene (*hPTTG*). hPTTG is a proto-oncogene which encodes the PTTG protein and is overexpressed in pituitary, thyroid, breast, ovarian and other carcinomas [149]. Expression and phosphorylation of hPTTG in thyroid cells is associated with MAPK and PI3K activation. In addition hPTTG in thyroid carcinoma cells is reported to be activated independently of the original mediators by these growth factors [150]. Conversely, PTTG and its binding factor PBF are reported to supress expression of NIS mRNA, and this leads to inhibition of iodide uptake. This could have adverse effects on prognosis, by reducing the efficiency of radioiodine treatment of differentiated thyroid cancer, in which overexpression of PTTG and PBF has been reported [151].

1.6.1.1 **BRAF**

The BRAF oncogene is involved in the signalling pathway mediated by TSH, through Gβγ dimer which activates the PI3K pathway. It appears to be a signalling connector between PLC, DAG, PKC and ERK, in the MAPK pathway (Figure 1.3) [9]. TSH seems to be cooperating with BRAF to induce thyroid tumourigenesis. Thyroid hormone biosynthesis is suppressed by BRAF; the compensatory rise in TSH will then also contribute to increasing the proliferation of thyrocytes. This may explain the major role for TSH as a tumour promoter in this disease [152]. Studies of *BRAF*-mutant PTCs have shown changes in the expression of NIS, the apical iodide transporter and or TPO [153].

BRAF is a member of RAF family which has three isoforms (ARAF, BRAF and CRAF) and activate extracellular signal-regulated kinase (ERK) through MAPK (Figure 1.3). Although RAS mediates the activation of the three isoforms, there is no evidence of overlapping between PTC with RET/PTC, *BRAF* or *RAS* mutations. It has been proposed that BRAF is the main isoform coupling RAS to MAPK/ERK (MEK), and that C-RAF and A-RAF signal to ERK to fine-tune cellular responses (reviewed in [154]). Point mutations of *BRAF* gene (the human oncoprotein) are found in about 45% of PTC and in 70% of melanomas. Most PTC (90%) are due to $BRAF^{V600E}$ (valine to glutamine) mutation. The mutant amino acid is situated between residues T599 and/or S602, the phosphorylation either of which is sufficient for BRAF activity. The $BRAF^{V600E}$ mutation is thought to mimic T599/S602 phosphorylation, rendering BRAF constitutively active. $BRAF^{V600E}$ probably functions in the early stages of tumour development by increasing growth, and promoting cell survival. On the other hand $BRAF^{V600E}$ somatic mutations have been reported as a common genetic change in PTCs. Inducing $BRAF^{V600E}$ mutation in mouse thyroid cells formed aggressive papillary thyroid cancers with short latency. A study [155] has observed that thyrocyte specific $BRAF^{V600E}$ expression leads to hypothyroidism in mice which agrees with other reports about BRAF suppression of thyroid hormone biosynthesis. During the course of 1 year mice with the $BRAF^{V600E}$ developed PTC. However, thyroid size and tumour development decreased and thyroid function was restored when MEK1/2 were inhibited.

1.6.2 Tumour suppressor Genes

Tumour suppressor genes normally inhibit the neoplastic process and if lost, promote tumour growth (e.g. p53 & RB Gene Product). Environmental factors such as ultraviolet light or various chemicals can damage DNA leading to mutations in these genes, which are believed to be the main reasons for activating the cell death program in the normal cell cycle. (reviewed in [156]).

Cell cycle event is composed of two phases, interphase and mitotic phase (M phase). Interphase occurs in the cell nucleus as it is responsible for DNA copying. This phase proceeds in three stages, G1, S, and G2, during which the cell grows, copies its DNA, and prepares to divide. The second phase mitotic phase (M phase) comprises five stages or phases; Prophase, Metaphase, Anaphase, Telophase & Cytokinesis. In this phase Cell

division occurs and two daughter cells are formed. Progression through each phase of the cell cycle is monitored by sensor mechanisms known as checkpoints [157].

After mitosis or in new embryonic cells, the cell starts the cycle with G1 phase in which, more growth of the cell occurs and the cycle should pass pre-DNA synthesis checkpoint (G1/S) (Figure 1.6). In G1 phase the cell forms the required proteins and enzymes for DNA synthesis in the next phase and is under the control of the p53 gene (*Tp53*). Some cells stop growing (e.g. brain cells), and exit this process, and are described as being in G0 or rest phase. The next phase is for DNA replication (S phase, for synthesis), in which the amount of DNA in the cell doubles and all chromosomes replicate. Another cell growth phase (G2 Phase or gap 2) then follows. G2 is the gap between DNA synthesis and mitosis, and has a checkpoint (G2/M) to ensure that everything is ready for entering the next phase, which is the M phase. In this phase, cell division occurs through mitosis, which comprises 5 phases as described earlier. After cell division each of the daughter cells begin at interphase of the new cycle (reviewed in [158]). The progression from one cell cycle phase to another is regulated by a family of proteins known as cyclins (A, B, D, E), which are degraded at specific points during each cell division (Figure 1.7). Cyclin binds with cyclin dependant kinase (CDK), via its unique binding site, creating cyclin-CDK complexes. These complexes determine the cell's progress through different phases of the cell cycle, via chemical modification (phosphorylation) of other proteins, and are known as positive regulators [159]. Depending on the type of cyclin-CDK complex (Table 1.1), this phosphorylation can either activate or inactivate target proteins to arrange coordinated entry into the next phase of the cell cycle. For progression from G1 to S, phosphorylation of *RB* gene or its related proteins, p107 and p130 can play a role in the expression of genes required for entry into S phase [160]. The Nobel prize in Physiology or Medicine was won in 2001 for the discovery of the genes that control the cell cycle and checkpoints, CDK and cyclin molecules[161].

During these cycles, detection and repair of any genetic damage occurs, in addition to the regulation of uncontrolled cell division. Different cyclin-CDK combinations decide which of the downstream proteins are targeted. CDKs are expressed all the time whereas cyclins are synthesised at different cell-cycle stages, depending on different molecular signals [162]. Cyclin-CDK complexes switch on cell cycles events at the correct time and in the correct

order to prevent any mistakes.G1/S checkpoint and G2/M checkpoint control these events. These two check points are controlled by different proteins (Figure 1.8). These proteins work by inhibiting cyclin or CDK or the complex of both. The cell cannot proceed to the next phase until checkpoint requirements have been met. They ensure that damaged or incomplete DNA has been repaired and thus not passed on to daughter cells (reviewed in [162]).

Proteins controlling the checkpoint at G1 phase (also called restriction point), are $p15^{INK4B}$, $p16^{INK4A}$, $p18^{INK4C}$, $p19^{INK4D}$, $p21^{CIP}$, $p27^{Kip1}$ and $p57^{Kip2}$, while P57, p27, P21 are regulating S phase [163]. These proteins have the ability to bind to several different classes of cyclin and CDK molecules. Some have the ability to bind and prevent the activity of cyclin-CDK complex, which prevent the addition of phosphate residues to its substrate protein. They are often referred to as cell cycle inhibitor proteins, because their main function is to stop or slow down cell division. Elevated expression of these proteins causes cells to arrest in the G1 phase. Furthermore mutations of the genes encoding these proteins may lead to loss of control over the cell cycle and thus uncontrolled cell proliferation. These mutations have been observed in different tumour types [164], and shown to be associated with increased risk of cancers [165].

As an example, p21 protein, which is regulated by p53, binds directly to cyclin-CDK complexes and inhibits their kinase activity, causing cell cycle arrest, to allow DNA repair to take place. The p21 (*WAF1*) gene contains several p53 response elements that mediate direct binding of the p53 protein and resulting in transcriptional activation of the gene encoding the p21protein[160].

Damaged DNA can be repaired by the tumour protein (p53), the transcription factor that causes cell cycle arrest. p53 helps to prevent cancer through its role in activating programmed cell death or apoptosis, one of the main pathophysiological changes governing a malignant cell. Cells have a built-in system of DNA repair, created by p53, which halts the cell cycle to allow such repairs to be made. There is increasing evidence that chemotherapy and radiation work by activating programmed cell death; tumours that contain normal p53 seem to respond well to therapy, while those that lack normal p53 generally have a poorer prognosis.

In cases of severe irrecoverable damage, it directs the cell to the programmed cell death pathway resulting in apoptosis. Cells with impaired tumour protein p53 function are more likely to permit the accumulation of altered genes, and mutations in *Tp53* are the commonest genetic lesion in human cancer. *Tp53* mutations are associated with poorly differentiated and anaplastic thyroid tumours in the majority of cases (but not in well-differentiated papillary or follicular carcinomas) [166, 167]. In contrast to RAS and RET, *Tp53* mutations are not targeted to a discrete region of the gene. Loss of expression of TG, TPO and TSHR, as well as impairment of expression of PAX-8, have been observed when introducing mutant *Tp53* expression vectors into the well-differentiated thyroid cell line PCCL3 [168].

In addition, both copies of *TP53* must be mutated for the phenotype to manifest. Mice with homozygous disruption of both *TP53* alleles develop normally but get cancers at many sites after birth[169].

Proteins	**Types**
Cyclin	A (A1, A2) B (B1, B2, B3) D (D1, D2, D3) E (E1,E2)
CDK	1 2 3 4 5 6 7 8 9 10 CDK-activating kinase
CDK inhibitors	INK4a/ARF (p14arf/p16, p15, p18, p19) cip/kip (p21, p27, p57)
Tumour suppressor	p53 p63 p73

Table 1.1; List of types of cyclins, cyclin dependant kinases (CDK), CDK inhibitors and tumour suppressors.

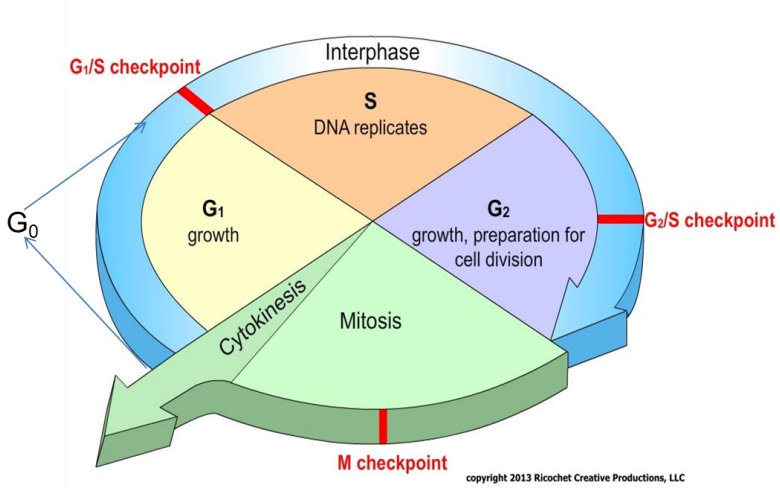


Figure 1.6; Diagram showing phases of the cell cycle starting with interphase which is itself comprised of three phases (G1 phase) for growth, (S phase) for DNA replication, then (G2 phase) for continued growth. This is followed by the mitotic phase (M phase), when cells divide by mitosis leading to cell duplication. G0 phase is when cells stop growing and exit the cell cycle. A full description is in the text. Modified from http://ricochetscience.com/category/diseases/cancer/.

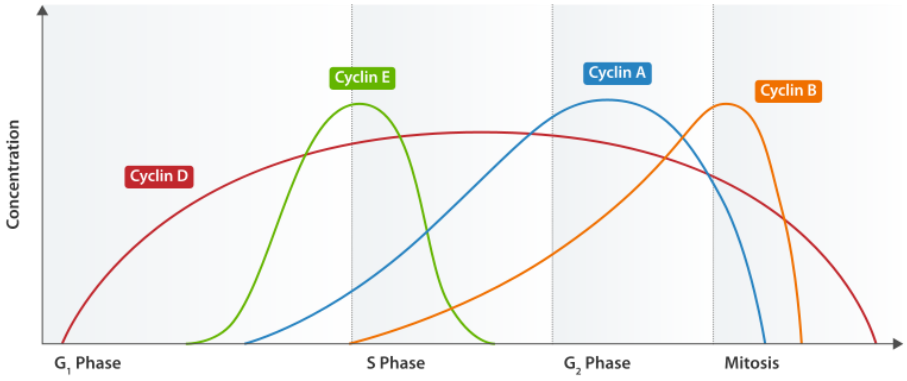


Figure 1.7; Graph showing the concentration of members of the cyclin family (D, E, A & B) during different phases of the cell cycle. Cyclins bind with cyclin dependant kinases (CDK) to create cyclin-CDK complexes, which promote the progression of one phase of the cell, cycle to the next. Modified from: http://maptest.rutgers.edu/drupal/?q=node/224

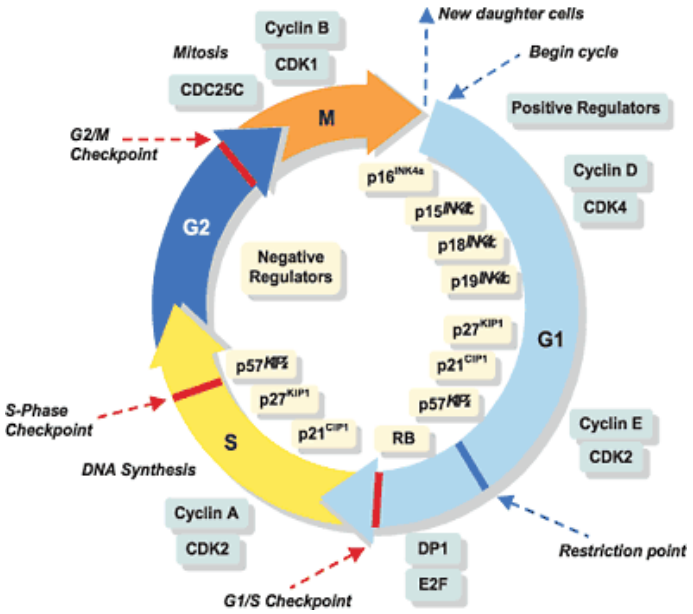


Figure 1.8: Cartoon showing checks points at various stages of the cell cycle and the main regulatory proteins present. These are $p15^{INK4B}$, $p16^{INK4A}$, $p18^{INK4C}$, $p19^{INK4D}$, $p21^{CIP}$, $p27^{Kip1}$ and $p57^{Kip2}$ in G1 phase, while P57, p27, P21 regulate S phase. Adopted from [163].

Retinoblastoma (*RB*), a tumour suppressor gene is inactivated when the cells complete the preparation to replicate their DNA. Inactivating mutations of both copies of *RB* gene, illegitimately release cell cycle block. Development of pituitary tumours was reported in heterozygous RB-null mice, which also develop MTC when crossed with P53-null heterozygotes [170]. Clonal loss of RB alleles has been reported in early stages of development of familial forms of MTC, while the role of RB in thyroid follicular tumours is controversial (reviewed in [144]).

1.6.2.1 **PTEN and Cowden Syndrome**

PTEN (Phosphatase and tensin homolog) is a tumour suppressor gene, coding for a protein phosphatase. Its absence (due to inactivating mutations) can cause Cowden Syndrome, which

is characterised by increased risk of developing thyroid and breast cancer. Approximately two-thirds of patients with Cowden's syndrome have thyroid abnormalities[171], half of which are PTC and FTC carcinomas [172]. Studies report thyroid neoplasia (mostly FTC) in 10% - 50% of affected individuals [173]. On the other hand, PTEN loss itself is unlikely to be playing a fundamental role early in PTC. PTEN low mRNA expression reported to occurs more in anaplastic thyroid cancers, but less so in FTC and PTC [174].

BRAF and ERK signalling have an important role in PTC tumourigenesis, while PTEN, which is also in the signal cascade with BRAF, may have indirect role. On the other hand loss of PTEN expression or function is assumed to be important both in the development of FTCs and more broadly in the progression and dedifferentiation of thyroid cancers[174].

PTEN requires loss of expression of both alleles to stop functioning as a tumour suppressor. However loss of one allele (LOH) can also occur in thyroid cancer, although LOH could not cause a complete loss of PTEN gene expression (reviewed in [174]).

PTEN is a dual phosphatase that negatively regulates phosphatidylinositol-3 kinase (PI3K) signalling and prevents AKT (also known as Protein Kinase B) activation. PI3K consists of two subunits, one of which is coded by the PIK3 CA gene (Figure 1.9). In thyroid tumours, this pathway can be activated by gain-of-function mutations of *RAS, PIK3 CA*, and *AKT* and loss-of-function mutations of *PTEN* [175]. Many genetic alterations activating the PI3K/AKT pathway have been previously identified in thyroid cancer [176, 177].

PTEN mutations have been reported in sporadic thyroid neoplasms, somatic mutations are found in 6% to 12% of follicular carcinomas and in 5% to 15% of anaplastic carcinomas[3].

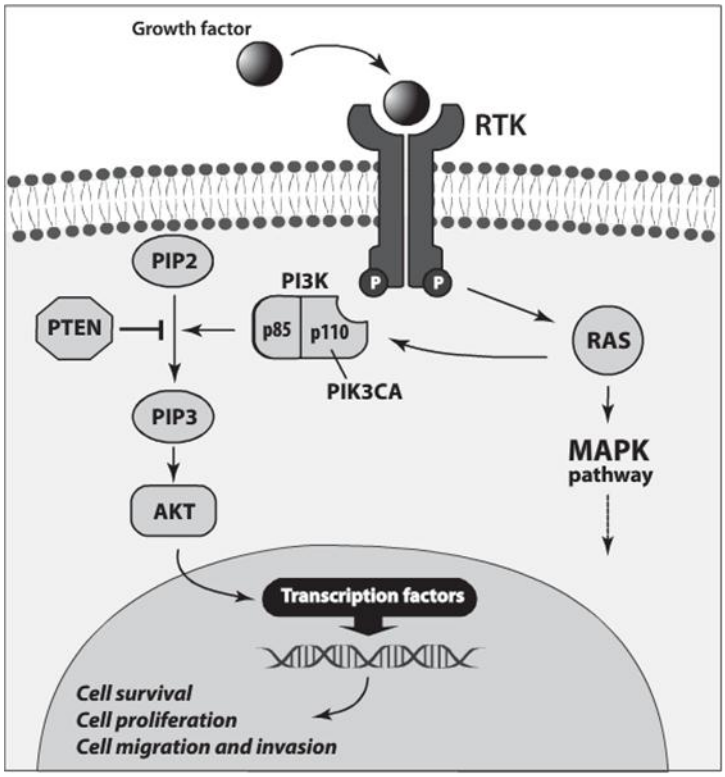


Figure 1.9; Cartoon showing the signalling pathway of phosphatidylinositol-3 kinase / phosphatase and tensin homolog / protein kinase B (PI3K/PTEN/PKB(AKT)), in which transmission occurs from surface growth factor receptors, which are receptor tyrosine kinases (RTK). This may activate PI3K directly or through a member of the GTPase protein family RAS, while PI3K is negatively regulated by PTEN. Downstream activation of PI3K regulates multiple genes that oppose apoptosis and promote cell survival, proliferation, and migration/invasion. In thyroid tumours, this pathway can be activated by loss-of-function mutations of PTEN, or by gain-of-function mutation of RAS or AKT. Adopted from [3].

1.6.2.2 **FOXE-1**

Expression of (forkhead box E1) *FOXE-1* also known as (TTF-2) has a potential role in follicular development (described in 1.2.1), as confirmed by genetic analysis. During embryonic development, *FOXE-1* controls thyroid cell migration from the primitive origin to its normal position at both sites of the trachea; *FOXE-1* expression is continued in adult life

(reviewed in [3]). Several studies as detailed below, reported FOXE-1 association with PTC and FTC.

A study in 2009, investigating thyroid cancer risk in Icelandic population [91], has performed genome-wide association study on (192 cases, of histopathologically confirmed thyroid cancer and 37,196 control). The study detected a strong association of two variants (rs965513 on 9q22.33 and rs944289 on 14q13.3), with high risk of PTC and FTC (5.7 fold greater than non- carrier). In the general European population, both risk alleles were shown to be associated with low concentrations of TSH, while allele on 9q22.33 was associated with low concentration of T4 and high concentration of T3.

In a second genome-wide association study on data from PTC in children and/or adult exposed to Chernobyl radioactive iodine [178], four SNPs have been identified on chromosome 9q22.33 showing strong associations with the disease. One of these SNPs (rs965513) located 57-kb upstream from FOXE-1. Although the study reported FOXE-1 as the strongest genetic risk marker of sporadic PTC in European populations, it is unlikely to be the only key player in radiation-related thyroid carcinogenesis. It has been found that SNP showing the second strongest association with sporadic PTC in Europeans (rs944289 on 14p13.3), is not associated with radiation-related PTC, that indicates how complex is the pathogenesis, suggesting a presence of other genetic factors in radiation-related PTC.

Another study of GWAS on large Spanish cohort reported that FOXE-1 gene has the strongest association with PTC susceptibility. The functional assay of rs1867277 on 9q22, (the only functional variant associated with sporadic PTC) showed effect on FOXE-1 transcription, while transfection studies showed an allele-dependent transcriptional regulation of FOXE-1. Furthermore, the risk allele (A) of this SNP increased FOXE-1 transcription by creation of a binding site for (Upstream Transcription Factor 1) USF1 and USF2 transcription factors [179]. rs1867277 was strongly associated with TC risk in Spanish and Italian cohorts[180].

1.7 MPhil summary (study on Malignant Thyroid Dysfunction)

I have studied a large family (Family #1) with euthyroid MNG of adolescent onset, progressing to PTC, affecting 7 individuals (at that time), in 4 generations (family tree is in Figure 2.1). In addition to MNG and PTC, individual (III-2) developed breast cancer in her fourth decade and individual (IV-3) had polycystic kidney disease in infancy. I have investigated the genes indicated on the basis of the accompanying breast cancer, (*NIS*) and kidney disorder, (*PAX-8*). Both genes were eliminated by Sanger sequencing [181]. *PTEN* gene has also been sequenced and eliminated. Furthermore I have investigated the known MNG loci on chromosomes 14q, Xp and 3q [84, 89, 182] and the known Familial Non-Medullary Thyroid Cancer (FNMTC) on chromosomes 19p, 2q and 1q [85-87], on 17 individuals of the family. That was performed by using microsatellite markers followed by linkage analysis, which have excluded all loci. Thus I hypothesize that this family represents a new form of inherited MNG with a significant risk of progression to papillary carcinoma.

1.8 Aims & Objectives

The aim of this study was to increase the understanding of the genetic defects causing MNG/PTC in family #1.

- Primary objective was performing whole genome scan using SNP markers on microarray chip, (chapter 2).
- The secondary objective was to perform the statistical analyses GWLA, to find the region(s) on chromosome(s) which has/have shown significant linkage with the disease allele, (shared common SNPs by the affected individual only), (chapter 3).
- The third objective was to perform GWLA on the second family (family #2), to seek any shared predisposing genetic factors (chapter 4)
- The fourth objective was to further investigate regions showing linkage and identify candidate gene(s) implicated in the thyroid defects and causing MNG/PTC, by any analyses indicated ,such as copy number variation (CNV) (chapter 5).

Chapter 2 GeneChip® Mapping Assay

In this chapter I will describe the first family studied and the Microarray used for the whole gene scan.

2.1 Case report and family history of family #1

2.1.1 Index patient

The family tree of family #1 is shown in (Figure 2.1).

The index patient (IV-6), a girl (DOB; 17/12/90) was referred (at age 12) to the Paediatric Endocrine service based at the University Hospital of Wales, Cardiff, with multinodular goitre.

She had been born at 39 week gestation by spontaneous vaginal delivery. The prenatal history was uneventful and the neurodevelopment had been within normal limits. The girl was of average height for her age.

Ultra-sound scan showed multiple hyper echoic thyroid nodules (three in the left and one in the right lobe). Fine needle aspirate (FNA) histology showed multiple papilloid adenomata with no features of papillary carcinoma. The kidney scan was normal.

Thyroid function tests (at age 12) showed normal TSH (1.4 mU/l; normal range 0.35-5.5 mU/l) and normal free T4 (16.2 pmol/l; normal range 9.8 - 23.1pmol/l), confirming a euthyroid condition. TSHR antibodies "TSAB" were negative and TPO antibodies were within the normal range (8.6 kU/l; normal range <32.0 kU/l). In addition, the TG antibodies were 22 kU/l (normal < 46.0 kU/l).

Since the girl had a large compressive goitre, with concerns for coexisting thyroid cancer, she underwent total thyroidectomy. Histopathological examination of the removed gland showed papillary hyperplasia, but no morphological evidence of malignancy. The nodules were multiple papilloid adenomata and resembled the histology of her mother's thyroid at age 17. Areas with ectopic thymus tissue were also observed.

The "photographs of mitotic chromosomes arranged in homologous pairs" showed a normal karyotype; 46 XX.

The girl is currently euthyroid, receiving daily replacement with thyroxine, 125mcgs.

2.1.2 Family History

There is a strong family history of MNG progressing to papillary thyroid cancer in some members of the family, accompanied by cystic kidney disease (in one male) and breast cancer (in a female).

Although physical examination of the thyroid of the girl's older brother (IV-3), (DOB; 31/01/84) was normal (at age 18), thyroid scan revealed 8 or 9 hyperechoic cysts. Furthermore, his FNA showed changes suggestive of papillary thyroid cancer. He also underwent total thyroidectomy. In infancy (5 months) he had undergone nephrectomy for polycystic kidney disease.

Her oldest sister (IV-1) (DOB; 30/8/80) has MNG but thyroid scan revealed small hypo-echoic cysts. She did not undergo thyroidectomy but will be investigated regularly for changes. In addition one of her 2 sons has been refereed for thyroid scan in 2011.

The girl's mother (III-2) (DOB; 9/12/61) underwent partial thyroidectomy at age 17 for MNG. The thyroid contained many benign lesions; some cystic, some follicular, and some papillary.

At the age of 27 she underwent a second partial thyroidectomy. The histology revealed multiple, mostly benign, papilloid adenomas. Some lesions again had a mixed follicular/papillary appearance. Papillary regions displayed very pale nuclei and other changes suggestive of papillary cancer and at age 28 she had a total thyroidectomy. She has developed breast carcinoma at age 41 which recurred at 47.

The girl's one brother (IV-5) (DOB 30/9/89) and one sister (IV-4) (DOB14/8/86), have no sign of thyroid or kidney lesions. The sister (IV-4) has undergone thyroidectomy for Graves' disease in 2011.

The girl's oldest brother (IV-2) (DOB 9/12/82), was not tested in the microsatellite analysis performed during my MPhil. However he attended UHW in 2008 and was confirmed to have a normal thyroid and kidneys and his blood sample was taken then. Her youngest sister (IV-7)

(DOB 27/6/94) has also been refereed for thyroidectomy at age 18 (in 2012). This has increased the affected individuals in the family to 8.

Her maternal uncle (III-3) and his family, have no sign of thyroid or kidney lesions. By contrast, her maternal grandmother (II-2) and great-grandmother (I-2) underwent surgery for presumed benign thyroid disease, which has not been possible to confirm.

Her mother's first cousin (III-5) underwent thyroid surgery for benign disease.

2.1.3 LREC

Cardiff University research ethics governance falls to the South East Wales Local Research Ethics Committee (LREC) which provided approval for the study. Informed consent was obtained from all family members and all samples were stored in accordance with Human Tissue Act (HTA) policy and procedure (under the guidance and direction of Cardiff University Human Tissue Act Compliance Team).

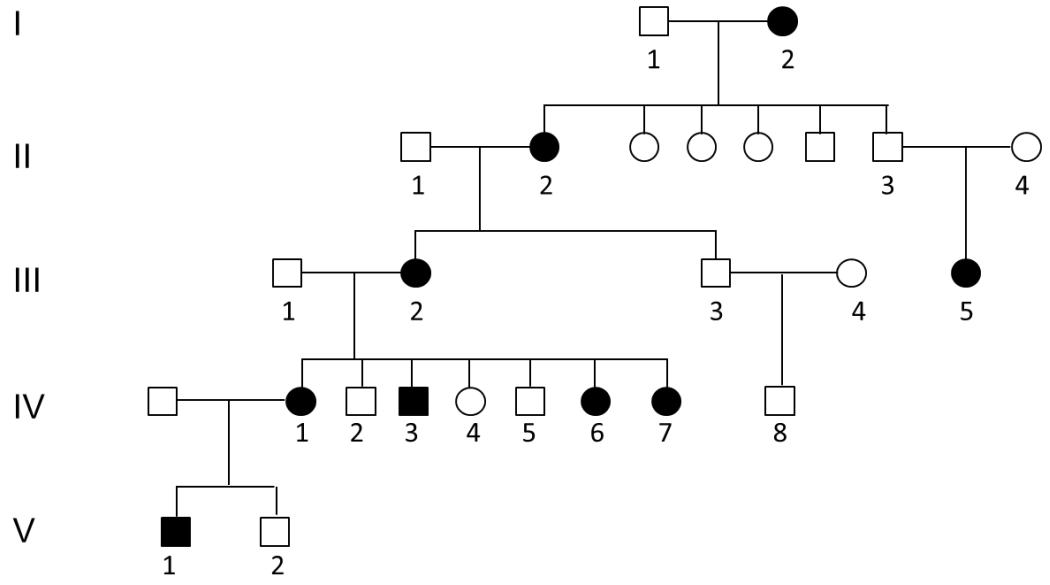


Figure 2.1: Tree of family #1. Affected members are in black boxes. The Index patient IV-6 was referred with multinodular goitre (MNG), her kidney scan was normal. Thyroid FNA of her 18 year old brother

IV-3 showed changes suggesting papillary thyroid cancer (PTC). In infancy he had a nephrectomy for polycystic kidney disease. Her sister IV-1 had MNG but thyroid scan revealed small hypo-echoic cysts, and her son V-1 recently developed MNG at age 5 (not genotyped). Her sister IV-7 was pre-pubertal in 2002 but has since developed MNG at age 18 whilst IV-4 developed Graves' disease. The brothers IV-2 and IV-8 had no thyroid or kidney lesions. The mother of the index patient III-2 underwent partial thyroidectomies at age 17 and 27 for MNG and a total thyroidectomy at 28 when PTC was present. She developed breast carcinoma, at 41. The maternal uncle III-3 and his family have no thyroid or kidney lesions but the maternal grandmother II-2, great-grandmother I-2 and mother's first cousin III-5 underwent surgery for presumed benign thyroid disease. The 2 boys in the last generation V-1 & V-2 have not been tested.

2.2 Introduction

2.2.1 Whole Genome Scan (WGS)

The traditional method of discovering disease genes begins with family based linkage gene-scans, looking for regions of the genome that are transmitted in parallel with the transmission of a trait [183]. WGS can be performed by using known markers across all chromosomes. Two different types of markers have been broadly used, microsatellite markers (will be described in 2.2.3) and single nucleotide polymorphisms (SNPs) markers (will be described in2.2.4).

WGS depends on the genetic differences between individuals and populations, or different individuals among the same family. This population differentiation can be ascertained by comparing (1) target individual vs. the entire population, (2) target individual vs. the subpopulation (to which that individual belongs) and (3) randomly chosen individual vs. the entire population[184]. Method (2) has been used in this study as there are specific individuals in the family known as affected (targets) and these have been compared with the rest of the family members (subpopulation). The best markers are those which show high levels of population differentiation (heterogeneous distribution).

2.2.2 Genetic Markers

The first genetic markers, measuring variations at the DNA sequencing level, were restriction fragment length polymorphisms (RFLP's), and were first used as a tool for genetic analysis in 1974. RFLP is a sequence of DNA that has a restriction site on each end with a "target" sequence in between. The target sequence binds to a probe labelled with radioactivity or an enzyme. The combination of restriction enzyme and probe sequence produces a series of bands in Southern blot, with lengths complementary to the probe. An RFLP occurs when the length of a detected fragment varies between individuals. Each fragment length is considered an allele, and can be used in genetic analysis (reviewed in [191]).

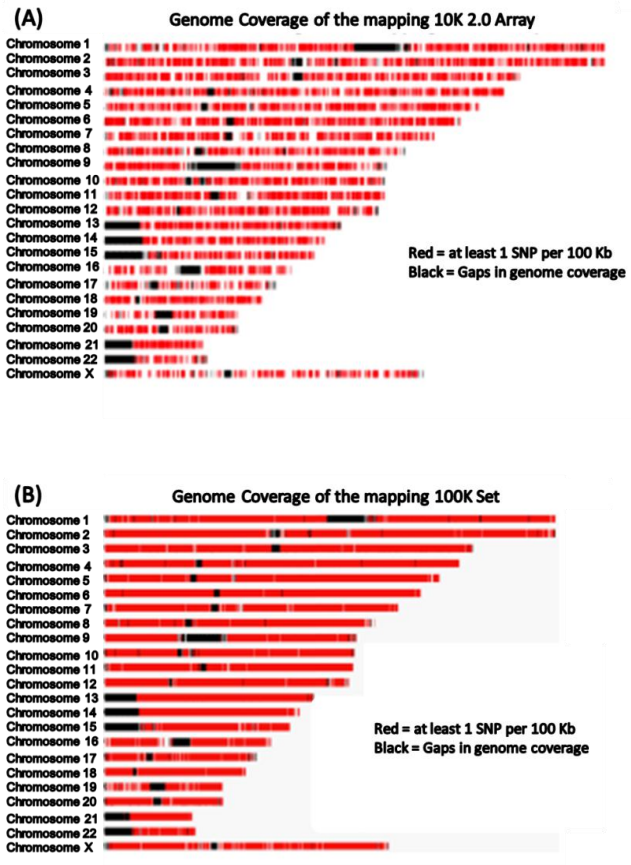
Microsatellites and SNPs (will be described in 2.2.4 respectively in more detail) are the most common markers to be used for genetic linkage mapping (will be described in 3.2.1). This mapping will estimate the distance between the disease locus and the genotyped markers and thus can be used to help identify the location of a disease gene [185].

Until recently it was common practice to use a panel of about 400 microsatellites at 10 cM average inter-marker distance from the well-defined human genetic map for linkage analysis (will be described in 3.2.1) [186-188]. However, recent progress in SNP discovery and genotyping provides the opportunity to use this marker type for linkage analysis as well [189, 190].

The 1000 Genomes project [192] examined a wider variety of populations to the original HapMap dataset [193]. One thousand genomes discovered additional SNPs and broadened the understanding of genetic diversity; it was started in 2008 and finished in 2010. It is an international effort in which at least 1000 unnamed members of different population have their genome sequenced using about 15 million SNPs. The project described the location, allele frequency and local haplotype structure of those SNPs, the majority of which were previously un-described [194]. This project aims to understand the genotype/phenotype relationship describing 1,092 individuals from 14 ethnic groups, showing that individuals from different populations carry different profiles of rare and common variants, and that low-frequency variants show substantial geographic differentiation, which is further increased by the action of purifying selection. It refined the human genome map and provided the data online in 2012 [194]. However, this study is about to be at least partially replaced by the

UK10K project [195], which aims to study the genetic code of 10,000 people in much finer detail than ever before.

One of the important challenges in the human genome scan is gaps in the centromeres and telomeres (Figure 2.4) [196].



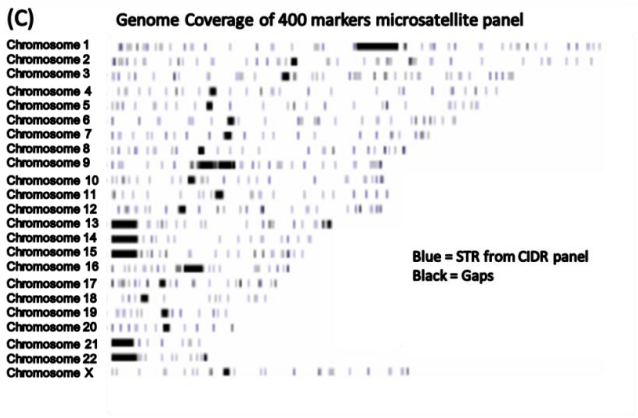


Figure 2.2; Comparing genome coverage when mapping using arrays containing 10,000 (A) or 100,000 (B) single nucleotide polymorphism (SNP) markers or 400 short tandem repeat (STR) markers, also known as microsatellites (C). Gaps (black dots) at centromeres and telomeres are shown, e.g. chromosomes 1, 9, 13, 14, 15, 21 & 22 have bigger gaps than others. Adopted from; http://www.docstoc.com/docs/119871841/AFFYMETRIX-SNP-chips

2.2.3 Microsatellite Markers

Microsatellites are a special class of tandem repeat loci (di-, tri-, or tetra nucleotide) of the DNA, which can be repeated up to 100 times. They can be distinguished by different numbers of repeats of these short sequences of nucleotides, and are widely distributed throughout the genomes of eukaryotes. The number of repeats is variable in populations of DNA and within the alleles of an individual (Figure 2.3). These alleles have been reported to have quite high mutation rates (10^{-3} per generation) [197].

If microsatellites are flanked with fluorescent PCR primers then amplification will give a pair of fluorescent allelic products, varying in size according to their repeat length (Figure 2.3).

According to the type of repeat sequence, microsatellites are classified as perfect (*e.g.* CACACACACACACACA), imperfect (*e.g.* CACACACATACACACACA), interrupted (*e.g.* CACACACTGCTACACACACA) or composite

(*e.g.* CACACACACACTCTCTCT) [198]. As an example, G to A mutation has changed the sequence ATGTGTGT to ATGTATGT, which creates "tetra" microsatellites (ATGT), with two repeats. These microsatellite were reported in African monkeys (ATGT), with 4 repeats and in human (ATGT) with 5 repeats[199].

A study [200] reported two hypotheses for the origin of the short DNA repeats; commonly substitutions, which is the main source for 2 nucleotides repeats and insertions (less common), but is the mean source of more than 70% of 2 to 4 nucleotide repeats. Both generate new repeats and thus new microsatellites originating from random mutations (Figure 2.4).

Microsatellite can be used in DNA fingerprinting[201]. Its application in linkage analysis can involve the examination of a large number of families or one large pedigree to observe the difference in the repeat-score of the same markers (allele) among each individual. If similar alleles were inherited together with a particular phenotype, then it could be assumed that this allele is linked to that particular phenotype. After PCR amplification, the alleles from each individual in a family are separated by size. Then linkage analysis can be performed between all markers [202].

Microsatellite markers generally are about 10Mbp (~10cM) apart, which is close enough to detect marker - marker linkage, and thus be capable of identifying the genetic regions associated or linked with the phenotype. Since microsatellites are spaced at approximately 10cM intervals, 300 or more microsatellites can be used to screen the entire genome[203].

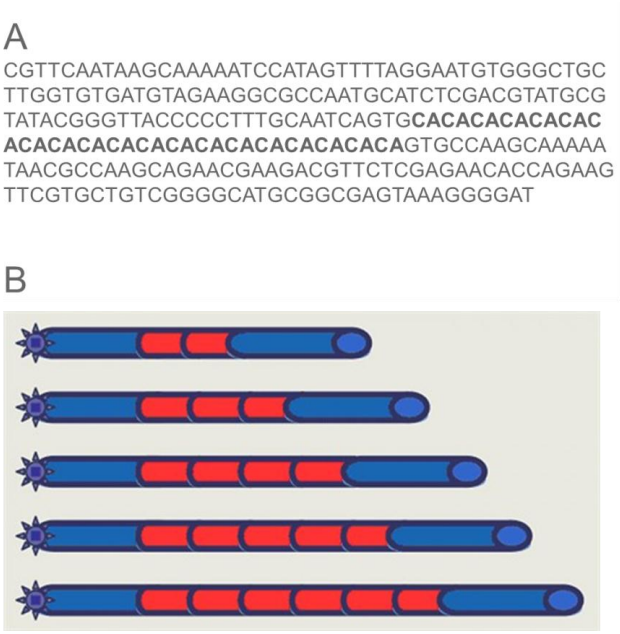


Figure 2.3; Microsatellite marker containing 20 dinucleotide repeats of CA (40bp), shown in bold (A). PCR products (labelled using fluorescent primers), vary in size according to their repeat length. (B) The example shows 5 different repeat lengths in a population. Adopted from; http://www.lifesciences.sourcebioscience.com

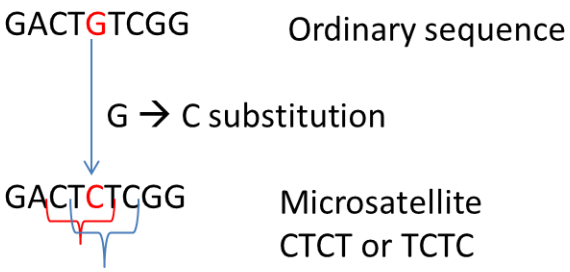


Figure 2.4; Cartoon showing point mutation Guanine → Cytosine (G→C), which has created a microsatellite. Modified from [200].

2.2.4 SNPs

A single nucleotide polymorphism (SNP) as its name suggests, is a small genetic variation in the DNA sequence in which a single nucleotide e.g. Adenine (A) is replaced by one of the other three nucleotides (T, G, C) (Figure 2.5). This creates 2 alleles at the same region (two sequences). These variations are more likely to occur in the non-coding region, and can only be called SNPs, when both sequences are present in at least 1% of the population. The term "polymorphism" is commonly used to refer to a normal variation, or to changes that do not directly cause disease. Furthermore, variation with lower than 1% frequency can be called mutation, even though, the 1% cut-off is not accurate enough [204]. Thus SNPs (or variants in general) can be classified depending on their allele frequency in genes (or population), which can be defined by Minor Allele Frequencies (MAFs); the frequency at which the less abundant (or minor) allele of a SNP is present in a population. There are three types of variants; rare with a frequency of less than 1%, uncommon with a frequency between 1-5%) and common SNPs with a frequency greater than 5% [205].

The main point behind using SNPs in linkage analysis is the frequency variation between minor and major alleles of the same SNP, across the population [206]. Population susceptibility to disease can be assessed by testing for any correlation between disease status and the SNP genetic variations. The simplest such study design used to test for association is the case–control study, in which cases affected with the disease of interest are compared with controls who are free of disease. The power of these studies is improved when the cases are related to the controls, as in family #1 [207], which will be described in more detail in Chapter 3.

Two areas of study commonly focus on; detection of rare mutations in highly selected cases, and common SNPs and haplotype of the general population. In WGS with rare variants, weak association between common tag SNPs (show linkage disequilibrium, will be described in 3.2.1) and rare causal variants has been reported [208]. In order for variants to be highly associated they must have similar allele frequencies and the two loci have equal minor allele frequencies. Therefore for associations with rare variants it is necessary to perform direct mapping and rare variants within a sample must first be identified [209]. Sequencing of candidate genes or entire genomes is the optimal way to identify rare variants. A number of studies have successfully used the approach of sequencing candidate genes [210, 211].

The main challenge in WGS is to develop statistical procedures that minimize false positives without greatly sacrificing true positives. That can be approached by selecting the SNPs with the right MAF, as the power to detect genetic effects mainly depends on the MAF of the risk allele tested. Loci with a low MAF (<10%) have significantly lower power to detect weak genotypic risk ratios than loci with a high MAF (>40%) [212]. Furthermore, rare genotypes are more likely to result in false findings and thus, many WGS remove SNPs with MAF<10% [213]. Many studies use common filter for SNPs which is based on MAF, with typical thresholds of 3, 5 or 10%. The latter approach (10%) is not sensitive enough to detect SNPs with a small genotype class because rare homozygous genotypes can be observed with minor alleles of moderate frequency [214].

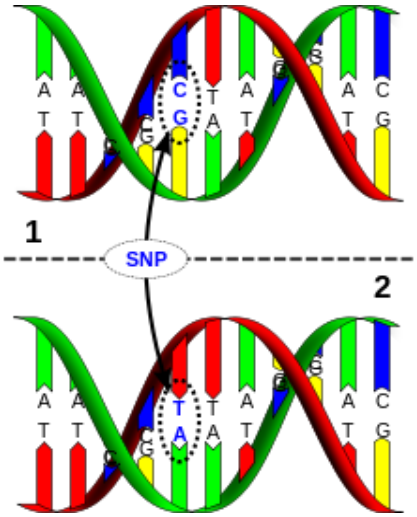


Figure 2.5; Example of a single nucleotide polymorphism (SNP) variant, in which a single nucleotide cytosine (C) is replaced by thymine (T) and thus the reverse sequence changes from guanine (G) to adenine (A). Adopted from; http://blog.biosyn.com/2012/09/genes-that-define-shape-of-our-face.html

2.2.5 The International HapMap Project

This project is the result of collaboration between researchers and companies in several countries, aiming to develop a haplotype map (HapMap) of the human genome. The Project describes the common patterns of human genetic variation. Samples from normal individuals of four populations were used for genotyping and Copy Number Variation Analysis as follow; 30 adult (and both their parents) from Nigeria (YRI), 30 adult (and both their parents) from U.S., residents of northern and western European ancestry, Caucasian (CEU), 44 unrelated individuals from Japan (JPT) and 45 unrelated individuals from China (CHB). The information produced by the project is made freely available to researchers around the world.

The goal of this project was to "create a public, genome-wide database of common human sequence variation, providing information needed as a guide to genetic studies of clinical phenotypes" [215]. The project has become a very good source of information about human genome sequence and used as databases of common SNPs genotypes of several populations (will be described in 3.1.6.2). In particular it also provides insights into human Linkage Disequilibrium (will be described in 3.2.1 in more detail) and has led to the development of inexpensive, accurate technologies for high-throughput SNP genotyping with imputation. The output of the project is made available online with web-based tools for storing and sharing data, and provides frameworks to address associated ethical and cultural issues[216].

In Phase I of the HapMap project, more than one million SNPs were genotyped, one SNP every 5,000 bases, using five different genotyping technologies. By 2006 the database included more than ten million SNPs; over 40% of them common (MAF >5%). During Phase II, more than 2.5 million additional SNPs have been genotyped. In phase I and phase II, the project has collected and studied 1,184 samples from 11 populations [217].

Although microsatellites were more frequently used in the past, SNPs are more popular nowadays because of their higher number "roughly 10 million" across the genome and thus the distance between is narrower than microsatellites (Figure 2.5). Avoiding bigger gaps between markers is important to avoid any inaccuracy due to the possibility of multiple recombination events (described in 3.1.4) therefore microsatellites are no longer the first choice of genetic marker. The greater number of alleles of microsatellite markers, along with a less automated method of genotyping leads to a higher error rate relative to SNPs [203].

2.2.6 DNA Microarrays

Microarray analyses are of two types, both uses reference nucleic acid on a solid surface (chip), which is known as the probe and hybridises to the target nucleic acid. One type of microarray is used for gene expression profiling, these usually involve the binding of cDNA (generated from RNA extracted from cells or tissues) to the probe sequence but as they have not been used in my work will not be described further. The second type of microarray is used for genotyping and will be summarised in the following sections.

DNA microarrays take advantage of the ability of complementary base pairs of nucleic acid strands to bind each other. For example, GCGCATAATT will bind to its complement (AATTATGCGC). On this basis the whole genome can be scanned (WGS). A DNA microarray is a collection of microscopic DNA spots attached to a solid surface (a silicon chip, e.g. Affy chip in case of Affymetrix Chip), and can be used in genotyping multiple regions of a genome. Each DNA spot (feature) contains specific sequences known as probes. These can be short sections of a gene that are used to hybridize target (the DNA sample), which is prepared in multiple copies (PCR amplified). Probe-target hybridization occurs and is detected and quantified using fluorescence-labelled targets which generate signals and thus the identity of the feature can be known [219].

In this study, DNA microarrays will be used to perform WGS, using SNP markers across the genome as features (probes) attached to affy chip.

2.2.7 Affymetrix 10K chip DNA Microarray

GeneChip® Human Mapping 10K 2.0 Array (HMA10K) is one of the arrays, (provided by Affymetrix), which utilizes SNPs to increase understanding of the genetics of complex human diseases.

The Gene Chip ® Mapping Assay performs genotyping of more than 10,000 human SNPs, distributed throughout the human genome at a median inter-marker distance of 210 kb [220]. It uses a single array and single PCR primer and allele-specific hybridization (will be described in2.2.8). The selection of SNPs is based on those which demonstrate 99.9% reproducibility, and average call rate of >95%. In addition to heterozygosity average of 0.37 and genome scan resolution of 0.31 cM (centi Morgan), (will be described in 3.1.5). The SNP

array is a viable alternative to panels of microsatellites [221]. It is a mapping tool for identifying regions of the genome that are linked to, or associated with, disease or a particular phenotype. Allele frequency in different populations can also be determined by this tool. Some studies have used it for mapping regions with chromosomal copy number variation (CNV), but others consider the 10K chip inadequate for CNV analyses, especially since the availability of arrays with substantially more SNPs, 100K as in [222], 500K and 1.5M (as will be described in chapter 3). Linkage studies can be conducted more quickly than microsatellite markers with higher accuracy and better results with 10,000 SNPs [223-227]. One study [225] indicated (on the basis of a direct comparison between the linkage peaks obtained using microsatellite markers and the HMA10K assay), that generally there is a high degree of correspondence between the two approaches but traditional microsatellite approaches are largely insufficient to detect (even) highly significant linkage peaks. The HMA10K assay provides a high level of accuracy due to the greater information content and coverage and the ability to rapidly perform and analyse the data obtained from genome wide scans with MERLIN and other software (will describe in 3.3.2 & 3.3.3).

2.2.8 Principles of Allele-Specific Hybridization on GeneChip® Probe Arrays

Allele-specific hybridization (ASH) is the method of studying allele variations in genotyping [228]. The Gene Chip® Mapping Assay contains probes corresponding to both of the two possible alleles at each SNP and labelled as A & B. the probes are labelled with fluorescent dyes and contain 25 bp with the single SNP at the central position (Figure 2.7 A). After adding the target DNA, hybridization to the array probe occurs. There are two types of probes: reference probes that match a target sequence exactly, called the perfect match (PM), and partner probes which differ from the reference probes only by a single base in the centre of the sequence. These are called the mismatch (MM) probes. Analysing the resulting signals from the allele-specific probes determines whether a SNP is AA, AB, or BB, depending on PM or MM with A allele, or B allele (Figure 2.6).To determine specificity in binding, for each SNP there are up to 40 different 25 bp oligonucleotides, each with a slight variation in perfect matches, mismatches, and flanking sequence around the SNP (Figure 2.7).

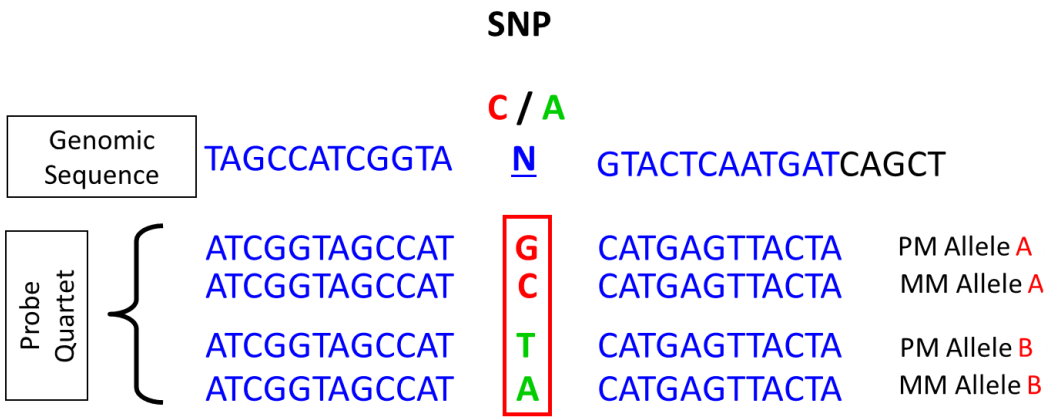


Figure 2.6; Cartoon showing the four probe sequences (probe quartet) for a target SNP (genomic sequence). Mismatch (MM) and perfect match (PM) for allele A and allele B is also shown. Adapted from: http://www.docstoc.com/docs/119871841/AFFYMETRIX-SNP-chips

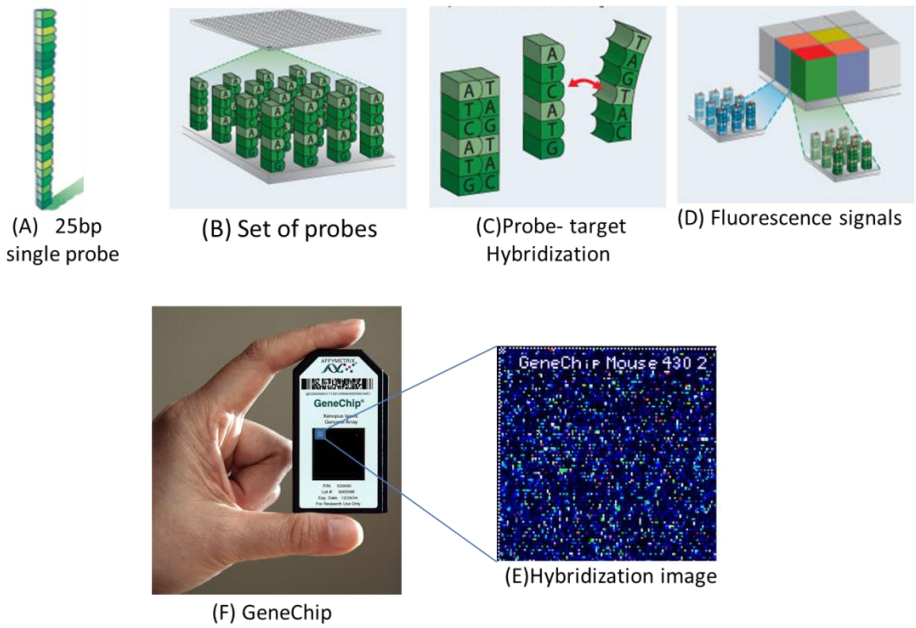


Figure 2.7; Cartoon showing (A) single probe with 25 nucleotides, (B) set of probes, (C) Probe-target Hybridization, (D) Fluorescence signals, (E) Hybridization image and (F) actual size of the chip. Adopted from; http://esd.lbl.gov/research/facilities/andersenlab/phylochip.html

2.2.9 GTYPE software and allele calls

GeneChip® Genotyping Analysis Software (GTYPE) from Affymetrix analyses the data from GeneChip Mapping 10K array, to determine the alleles of the SNPs represented in the DNA sample.

The fluorescence signals (from microarray probes, representing each SNP), which are known as Relative Allele Signal (RAS), can be detected and calculated, by a clustering process, on forward and reverse strands across many DNA samples. RAS is the ratio of the signal of the A allele to the sum of the A and B alleles A/(A+B). Depending on MM and PM of DNA samples and the probes (described in 2.2.8), two independent RAS values are derived for each SNP, from the forward (sense) strand (RAS1) and the reverse (anti-sense) strand (RAS2). Affymetrix GTYPE software plots RAS1 scores against RAS2 scores and tracheotomise (divide into 3 elements) these RAS scores, which forms the genotypes for DNA of an individual. RAS values near 0.0 are identified as a BB homozygote, 0.5 as an AB heterozygote, and 1.0 as an AA homozygote.

The call zone (known as the Distance to Radius ratio), for each allele type of the SNP can be calculated. GTYPE calculates the distance between "tested allele call" clusters and the RAS value (default 0.8). Adjustable call zones increase the stringency of the genotype assignment by filtering the data to mitigate nonspecific hybridization. RAS points that fall outside call zones are assigned as "No Calls (Figure 2.8) [221, 229]. For SNPs with sufficient minor allele frequencies, all three possible genotypes should be observed (AA, AB & BB). Only SNPs with all three observed clusters could be considered informative. Quantitative measure of clustering quality can be derived by the silhouette score (s), whose value ranges from 0 to 1. SNPs with s values closer to 1 show clusters that are tight and well separated, while poorly clustering SNPs show lower s values (Figure 2.8).

SNPs on the Chromosome X are first called using general model. Then the calls are evaluated to determine the gender of the sample (if has been specified by the operator). The SNPs are then called again using the appropriate gender-specific model. The call for Chromosome X SNPs displayed in the Scatter Plot are made using the gender-specific model; however, the gender model clusters and call zones are displayed in the Scatter Plot [221, 230]. By

following the genotyping calls of these SNPs markers, it would be known if they are identical by descent or only identical by state (will be described in 3.1.7).

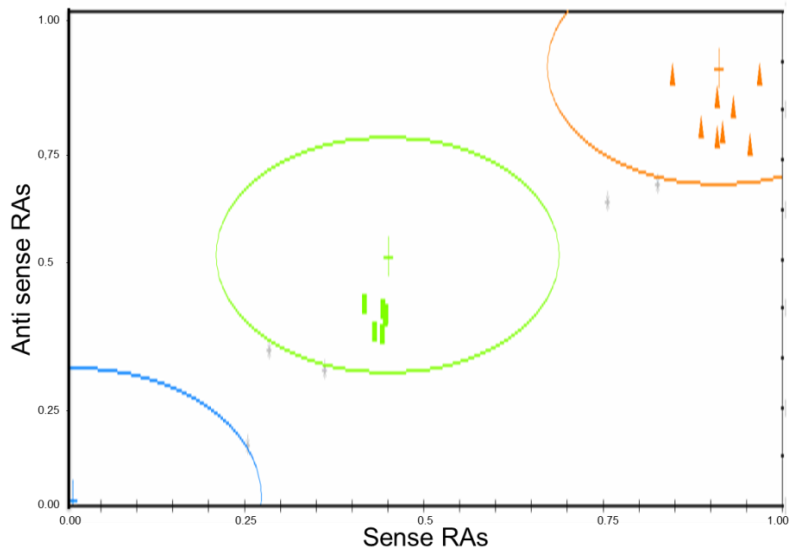


Figure 2.8; Screen shot showing clusters of genotypes obtained by comparing the ratio of the relative allele signal (RAS) of the A allele to the sum of the RAS for A and B alleles A/(A+B).

2.2.10 Protocol Summary

In summary genomic DNA from each family member was digested with restriction enzyme (Xba1), (will be described in 2.2.11) and ligated to adaptors recognizing the cohesive four base overhangs (Figure 2.10), (all fragments resulting from restriction enzyme digest are substrates for adaptor ligation). A generic primer, which recognizes the adaptor sequence, was used to amplify the ligated DNA fragments and PCR conditions were optimized to amplify fragments in the 250-1000 bp size range. The amplified DNA was further fragmented to smaller size (100-25 bp), then labelled and hybridized to GeneChip array. All arrays were washed and stained on a GeneChip fluidics station, then scanned on a GeneChip Scanner 3000 (Figure 2.9).

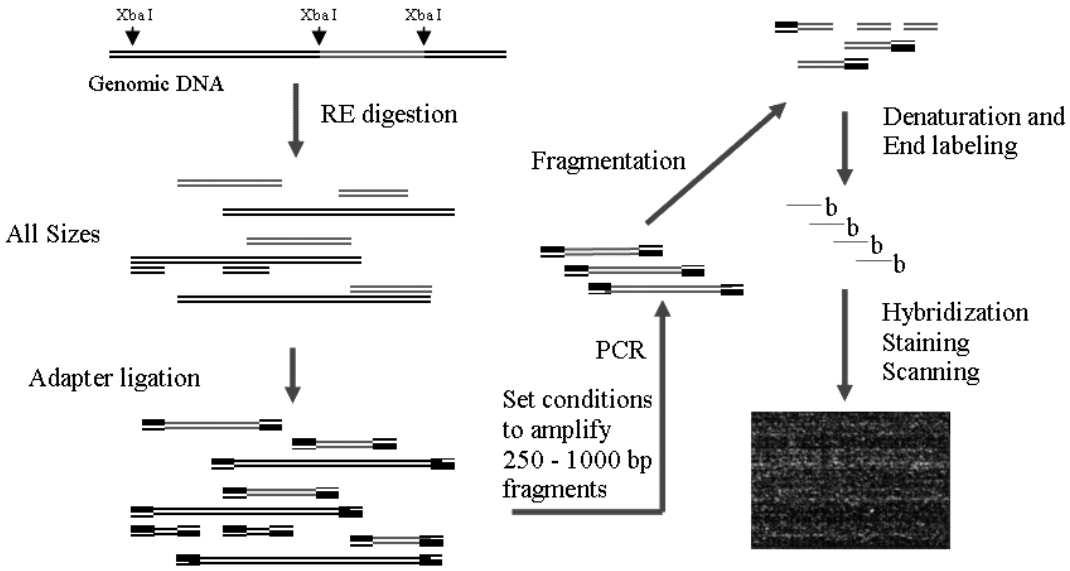


Figure 2.9; Protocol summary which starts with genomic DNA digestion to 250-1000 bp, then adapter ligation and single primer PCR amplification. This is followed by further fragmentation of the PCR amplicons to 100-25bp, then labelling and hybridization to GeneChip arrays. The arrays were washed and stained using GeneChip fluidics, then scanned on a GeneChip Scanner 3000. Modified from: GeneChip Mapping 10k 2.0 Assay Manual.

The genome-wide distribution of SNPs in the GeneChip Mapping 10K 2.0 Array is well distributed across the genome, but coverage is not absolutely uniform, as some regions contain less markers than others.

2.2.11 Xba1 enzyme

DNA can be reduced to smaller fragments by restriction endocnucleases such as Xba1, which can recognise specific sequence in the genome (T: CTAG A) and cut the DNA at that sequence. Since this sequence is repeated every 250-1000 bp, restriction by Xba1 enzyme breaks the DNA in small fragments ranging from 250 bp to 1000 bp, all with similar start of five base (CTAGA) in forward and (TGATC) in reverse sequence (Figure 2.10). These consistent four bases can be recognised by specific adaptors which ligate and overhang all

fragments. The sequence of these adaptors can be recognised by the generic primer which is used to amplify the ligated DNA fragments by PCR.

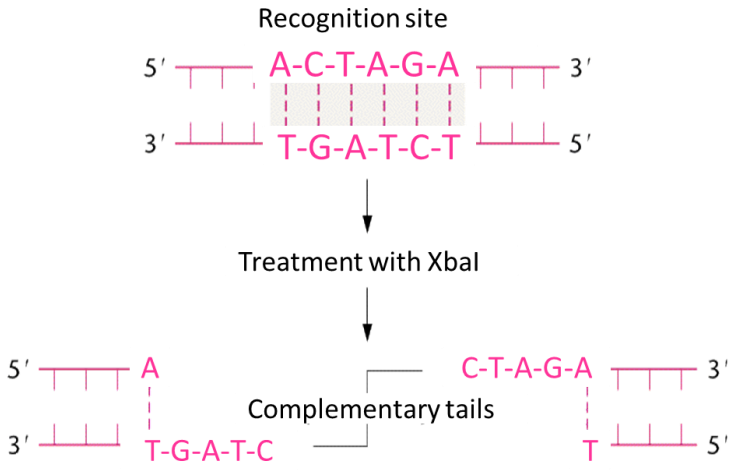


Figure 2.10: Cartoon showing the recognition site of Xba1 enzyme and the sequence of the complementary tails after restriction. A full description is in the text. Modified from; Klug & Cummings 1997.

2.2.12 Fragmentation

DNA can also be fragmented by DNAse enzymes which cut anywhere in the sequence. For DNA fragmentation, Affymetrix 10K GeneChip® procedure has used DNase I enzyme, which is thought to be responsible for fragmentation during apoptosis[231].

2.2.13 Hybridization

The main principle of this protocol is ASH (described in 2.2.8) in which Probe-target hybridization occurs and the fluorescence on the probes generates signals that identify the genotypes of the SNPs on the target sample (described in 2.2.6-2.2.9).

2.2.14 Image analysis

The strength of signals generated from probe-target hybridization (intensity of the features), (described in 2.2.6), can be measured by the GeneChip Scanner 3000, which creates (yields) a specific file (CEL file) that stores the result of the intensity calculated from the image of "probe-target hybridization" in the Affymetrix GeneChip® (Figure 2.7 E). Image analysis identifies markers with poor-quality and low-intensity features, which can be removed.

2.2.15 Software for data receiving

GeneChip® Operating Software (GCOS) can be used to acquire the data after scanning in the GeneChip® Scanner 3000. The software creates images and cell intensity files, by extracting the measured intensity of raw signal then prepares the file to be suitable for processing by GTYPE (described in 2.2.9).

2.3 Aim

To genotype 10,000 SNPs covering the entire genome in the single large family, for identifying the region(s) of the genome that harbour disease susceptibility loci.

2.4 Materials & Methods

2.4.1 DNA Extraction from peripheral lymphocytes

Approximately 5 ml of blood was collected into EDTA vacutainers (Beckton Dickenson). DNA was extracted from whole blood with the QIAamp Blood Kit (Qiagen), using the "large sample volume blood and body fluid protocol".

EDTA blood (1 ml) was incubated for 10 min at 70 C with 1 ml of lysis buffer (AL) and 125 μl of proteinase K, after which 1050 μl of 100% ethanol was added. The sample was then passed through a Qiagen spin column by centrifugation (at 8000 rpm for 1 min) in a Mikro 20 Patterson Scientific centrifuge. The column was washed twice with washing buffer (AW), prior to the elution of DNA by centrifugation (at 8000rpm for 1 min), with 200 μl of elution buffer (AE) preheated to 70 C.

2.4.2 DNA quantification by spectrophotometer

DNA was quantified by measurement of absorbance at 260 nm wavelength (λ) using a Gene Quant pro spectrophotometer (Amersham Pharmacia Biotech). DNA concentration was calculated based on a solution of 50 μg/ml having an optical density of 1 at 260 nmλ, and with a light path of 1cm. An absorbance ratio 260/280 of greater than 1.8 is indicative of a pure DNA sample.

2.4.3 Genomic DNA Preparation

Working stocks of all DNA samples were prepared to be 50ng/μl, as this concentration is required by the protocol for DNA digestion step, for high throughput assay, TE buffer (0.1 mM EDTA, 10 mM Tris HCl, pH 8.0) was used for the dilution process when required. Five μl (50ng/μl) of each diluted genomic DNA has been aliquoted into each well of the 96-well plate.

2.4.4 DNA control

The Affymetrix kit has provided DNA reference control (reference genomic DNA 103), to be processed with the samples as part of quality control check. Thus that DNA was also processed with DNA of family member.

2.4.5 DNA digestion with Xba1 restriction enzymes

A master mix of 10.5 μl H_2O, 2 μl of NE buffer 2 (10X), 2 μl of bovine serum albumin (BSA) (10X (1mg/ml)) and 0.5μl of Xba1 (20 u/μl) was prepared on ice. Fifteen μl of the master mix was then added to the 5 μl (50ng/μl) of the diluted genomic DNA in the 96-well plate. The plate has been covered with a plate cover and sealed tightly, mixed by vortex, and centrifuged briefly at 2,000 rpm for 1 minute. The plate was then placed in a thermal cycler for 120 minutes at 37 C and hold at 4 C.

2.4.6 DNA Ligation

A master mix of 1.25 μ of (5μM) Adaptor Xba and 2.5μ T4 DNA Ligase buffer (10X) was prepared. In each digested DNA sample, 3.75 μl of the Ligation Master Mix has been added followed by 1.25 μl of T4 DNA Ligase making the final reaction volume 25μl. The plate was then covered with a plate cover and sealed tightly, vortexed at medium speed for 2 seconds, and spun briefly at 2,000 rpm, for 1 minute. The plate was then placed in a thermal cycler (at 16 C for 120 minutes then 70 C for 20 minutes). Each DNA ligation reaction has been diluted

by adding 75 µl of molecular biology-grade H_2O to make the final volume of ligated DNA to be 100µl.

2.4.7 PCR

2.4.7.1 PCR Master Mix

A master mix was prepared in the pre-PCR clean room; it comprised 10 µl of PCR buffer (10X), 10µl of dNTP (2.5 mM each), 10µl of MgCl2 (25 mM), 7.5µl of PCR Primer Xba (10 µM), 2 µl AmpliTaq Gold® (5 U/µl) and 50.5µl of H_2O making the total of 90µl.

2.4.7.2 PCR staging

PCR master mix (90 µl) was added to each well of a new PCR plate, followed by diluted, ligated DNA (10 µl) with 4 wells being used per sample. The plate was then sealed with plate cover, vortexed at medium speed for 2 seconds and spun at 2,000 rpm for 1 minute. The plate was then put in a thermal cycler using specific programme for MJ DNA Engine machines (as shown in Table 2.1).

Temperature	Time	Cycles
95C	3 minutes	1X
95C 59C 72C	20 seconds 15 seconds 15 seconds	35X
72C	7 minutes	1X
4C	Hold	

Table 2.1; The PCR programme used described in terms of temperature (column 1), time of incubation (column 2) and the number of cycles performed (column 3). The programme is specific for MJ DNA Engine PCR machines as stated in the protocol manual.

2.4.7.3 **PCR Quality Control (agarose gel)**

Each PCR product was run on a 2% agarose gel, as a part of quality control. A 2% agarose (Gibco® BRL) gel containing 400 ng/ml ethidium bromide was prepared in TAE buffer, (2 M Tris base, 50 mM EDTA (pH 8) and 1 M glacial acetic acid); the same concentration of ethidium bromide was maintained in the TAE running buffer. Three µl of each PCR product mixed with 3µl 2X gel loading buffer (Promega) were loaded per well. Each PCR product was compared against a 1kb bp DNA ladder (Promega) to confirm the size of the amplicon. The gel then was run at 120V for 1 hour.

2.4.8 PCR purification and elution with QIAGEN Min-Elute 96 UF PCR purification plate

All 4 PCR products of each sample (400µl) were consolidated into a single well of a Min-Elute 96 UF PCR Purification Plate. After preparing a vacuum manifold and adjustment to create a vacuum source able to maintain ~800 mbar, the Min-Elute 96 UF PCR Purification Plate was placed on top of the manifold. Unused wells were covered with PCR plate cover and the used well left uncovered. Vacuum applied and the ~800 mbar vacuum maintained until the wells were completely dry. That has taken around 2 hours to dry 400 µl PCR samples.

The PCR products were washed by adding 50 µl molecular biology water and drying the wells completely (approximately 50 minutes). The washing steps were performed 3 times. Then the Min-Elute plate was removed from the vacuum manifold and gently tapped on a stack of clean absorbent paper to remove any liquid that might remain on the bottom of the plate. To elute the PCR product, 40 µl elution buffer (EB) (10mM Tris-HCl, pH 8.5) was added to each well, the plate covered with PCR plate cover film and been moderately shaken on a plate shaker at 960 rpm for 4 minutes. The purified PCR product was recovered by pipetting the eluate out of each well.

2.4.9 Purified PCR product's concentration adjustment

Purified PCR product needed to be at a concentration of 20µg per 45µl solution to be prepared for the next step (Fragmentation). EB (10mM Tris-HCl, pH 8.5) was used for any required dilutions.

2.4.9.1 **Quantification of purified PCR product**

Purified PCR products were measured (quantified) by Nano Drop (from Thermo Fisher Scientific Inc), and 20µg per 45µl of each Purified PCR product was prepared for Fragmentation.

2.4.10 Fragmentation

Fragmentation Buffer (5 µl 10X) was added, (using fragmentation plate), to each 45µl of purified PCR products on ice and vortexed at medium speed for 2 seconds. Fragmentation reagent was diluted (0.048 U/µl) on ice, by adding 5µl buffer to 1µl fragmentation reagent and the volume completed to 50µl by molecular biology water. Then 5µl of the mixture was added to the previous 50µl. The fragmentation plate was covered with a plate cover and sealed tightly. The fragmentation plate was vortexed at medium speed for 2 seconds, and briefly spun at 2,000 rpm for 1 minute. The plate was then placed in pre-heated thermal cycler (37 C) immediately for 30 minutes then at 95 C for 15 minutes and then cooled at 4 C.

2.4.10.1 **Fragmentation Quality Control (agarose gel)**

The fragmented PCR products (4µl) were mixed with 4µl 2X gel loading dye and run on a 4% agarose gel at 120V for 30 minutes (as quality control check), the 100 bp ladder was used as a size standard. After confirming the success of fragmentation reaction by analysing the gel picture, the samples then were handled to a third party, the Central Biotechnology Services (CBS) department, for labelling, hybridization, washing, staining and scanning.

2.4.11 Labelling

Master mix for labelling reaction was prepared (in ice) by mixing 14µl of 5X TdT buffer with 2 µl GeneChip DNA labelling reagent (5 mM) and 3.4 µl of TdT (30 U/µl). A total volume of 19.4 µl of this mixture was added to 50.6 fragmented PCR products making the total reaction of 70µl. (The remaining fragmented DNA was used in the agarose gel). The plate was then sealed tightly with a plate cover, vortexed at medium speed for 2 seconds, and briefly spun at 2,000 rpm for 1 minute. The plate was then placed in pre-heated thermal cycler (37 C) for 2 hours then at 95 C for 15 minutes and cooled down at 4 C. The plate was briefly spun at 2,000 rpm for 1 minute after the labelling reaction.

2.4.12 Target Hybridization

Hybridization mix was prepared according to manufacturer's instructions. Each labelled sample was transferred from the plate to a 1.5 ml Eppendorf tube and 190µl of the hybridization Cocktail Master aliquoted. The 260 µl of hybridization mix and the labelled DNA were heated at 95 C in a heat block for 10 minutes to denature, and then cooled down on crushed ice for 10 seconds. That was followed by a brief spin at 2,000 rpm for 1 minute in a microfuge to collect any condensate. The tubes were then placed at 48 C for 2 minutes and 80 µl of the denatured hybridization mix injected into the array. It was then hybridized at 48 C for 16 to 18 hours at 60 rpm.

2.4.13 Washing, Staining, and Scanning

Fluidics Station 450/250 automatic washing and staining probe arrays were used according to manufacturer's instructions for GeneChip® mapping. That was followed by scanning of the probe arrays using the GeneChip® Scanner 3000, according to manufacturer's instructions.

2.4.14 Data acquiring

GCOS software (described in 2.2.15) was used to acquire the data from the GeneChip® Scanner 3000. The software creates images and cell intensity files, by extracting the measured intensity of raw signal. Thus special files were created for analysis; "Data Transfer Tool (.DTT) Archive files". These files contain the genotyping data from all members of family #1. Such files are not compatible with Microarray Suite (MAS) and thus GCOS restores these files and made them suitable to be processed by GTYPE software (described in 2.2.9). GTYPE has then been used to expose the raw image data of cell fluorescence intensity (.CEL files), which contain information about the expression levels of the individual probes. GTYPE then generates CHP files (from CEL files), for each sample to display a Relative Allele Signal (described in 2.2.9) and thus the genotype of each DNA sample across the ~ 10,200 markers (Figure 2.15). All data have been saved in Excel format which will be analysed in chapter 3.

2.4.15 Data Check per Individual, per SNPs and per Batches

The data imported into an Excel worksheet has undergone a range of quality control steps. The first step was to check call rate for each SNP across the 19 DNA samples (18 family members and 1 duplicate, as described in the next section), and remove any SNP showing 'no

call' for more than 4 individuals. The second step was to check call rate for each individual across the 10208 SNPs and remove any individuals showing 'no call' for more than 10% of SNPs. The third step was to check for any common error in data from the same batch.

2.5 GeneChip Mapping Assay RESULT

2.5.1 DNA Extraction from peripheral lymphocytes

Genomic DNA samples from all family members (except IV-2), were available from my MPhil microsatellite studies. However, as these were several years old, even though they were stored at -20C they might not have been suitable for micro-array. I was able to obtain new samples from 6 family members which were extracted and quantified. The new DNA was extracted and quantified by measurement of absorbance at 260 nm wavelength (λ) using a Gene Quant pro spectrophotometer (Amersham Pharmacia Biotech). DNA concentrations were between 50 and 80 ng/µl and the ration was good enough to confirm purity.

2.5.2 Comparing microarray results from New and Old Genomic DNA

Samples of the same individual (IV-7) were used, new (N) and old (O), to perform all steps, then compare the output as quality control check of old DNA samples.

2.5.3 Genomic DNA Preparation

Working stocks of DNA samples need to be adjusted to 50ng/µl, if required. Concentrations of both new and old DNA were close to the recommended value and thus no adjustment was required. Reference DNA from Affymetrix was ready to use thus has been processed directly.

2.5.4 DNA digestion with Xba1 restriction enzymes

Three digestion runs have been performed for each sample to have enough quantity for subsequent steps

2.5.5 PCR Quality Control (agarose gel)

Two percent ethidium bromide stained agarose gel electrophoresis was prepared, which showed a smear of PCR products (Figure 2.11). This is expected since the PCR generates many products of differing lengths and is similar to the example provided in the manufacturer's protocol.

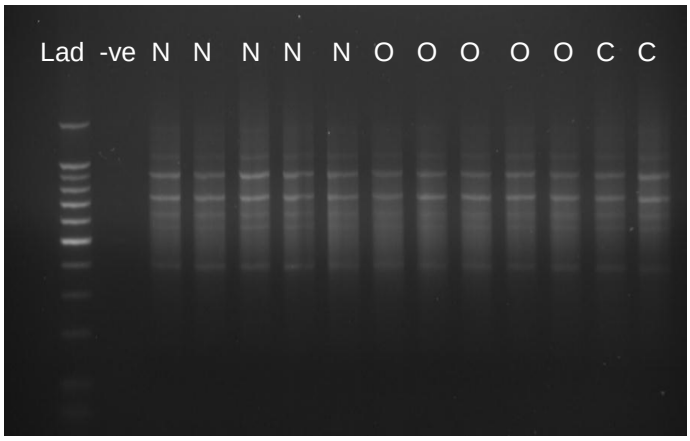


Figure 2.11; PCR products from individual IV-7 analysed by agarose gel electrophoresis and ethidium bromide staining. The first lane (Lad) contains the 100bp ladder, the second lane (-ve) contains the negative control (master mix without DNA template), lanes 3-7 (N) show the PCR products obtained using new DNA samples and lanes 8-12 (O) with old DNA sample. Lanes 13 & 14 (C) show the amplicons obtained using "reference genomic DNA control 103" provided with the kit. Gel picture shows a smear of PCR product similar to the example provided in the manufacturer's protocol

2.5.6 Quantification of purified PCR product

PCR products were purified and the DNA concentration was measured using Nano Drop, aiming to adjust the concentration for the following step (fragmentation), in which requires 20µg/45µl. Sample N shows concentration of 385.7 µg/µl = 17µg/45µl and sample O shows 266.8 µg/µl = 12µg/45µl, while reference control shows 233.9 µg/µl = 10.5/45µl. As the concentrations were less than the recommended value, more PCR products have been purified using an additional 400µl of PCR product for each sample. In this purification, previously purified PCR product was used at elution instead of EB, which increased DNA concentration when measured. Purified PCR product of N showed 490.5 µg/µl = 22µg/45µl,

which has reached to the recommended concentration, but O showed 400.3 µg/µl= 18µg/45µl, which is still below the recommended concentration. Therefore more PCR product of O was purified as above to yield a concentration of 466.3 µg/µl = 20.9 µg/45µl, same done with reference control.

2.5.7 Fragmentation quality control (agarose gel)

Agarose gel (after fragmentation) was performed as quality control step. Four µl of each fragmented DNA was mixed with 4µl 2X gel loading dye and run on 4% agarose gel at 120V for 30 minutes. One hundred bp DNA ladder was used as size standard. (Figure 2.12) shows the gel picture with a smear of PCR products, similar to the example in the manufacturer's protocol.

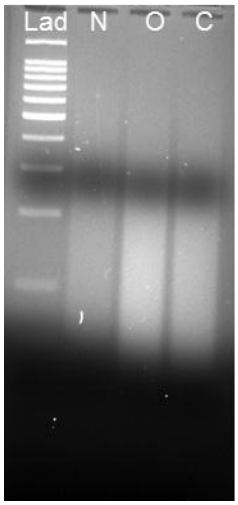


Figure 2.12; Fragmentation products of DNA for individual IV-7 analysed by agarose gel electrophoresis and ethidium bromide staining. The first lane (Lad) contains the 100bp ladder, the second lane (N) new sample, the third lane (O) old sample and fourth lane (C) for "reference genomic DNA control 103" provided with the kit. A smear of PCR products between 100-25 bp is shown, similar to the example in the manufacturer's protocol.

2.5.8 Processing DNA of the remaining family #1 members (in 5 batches)

All DNA samples underwent quantification, digestion with Xba1 enzyme, PCR amplification, purification of PCR products, fragmentation, labelling, staining, and scanning, (as described in 2.5.12.5.7). The DNA concentration of all samples was measured and found to vary from 32 to 160ng/µl. Different dilutions were performed to keep the concentration closer to 50ng/µl, for high DNA concentration sample. At lower DNA concentrations (less than 50ng/µl) calculated increases in DNA volume were used in the PCR master mix (instead of water) to increase the concentration to 50ng/µl. Figure 2.13 is showing PCR product for all family members, after DNA digestion and amplification.

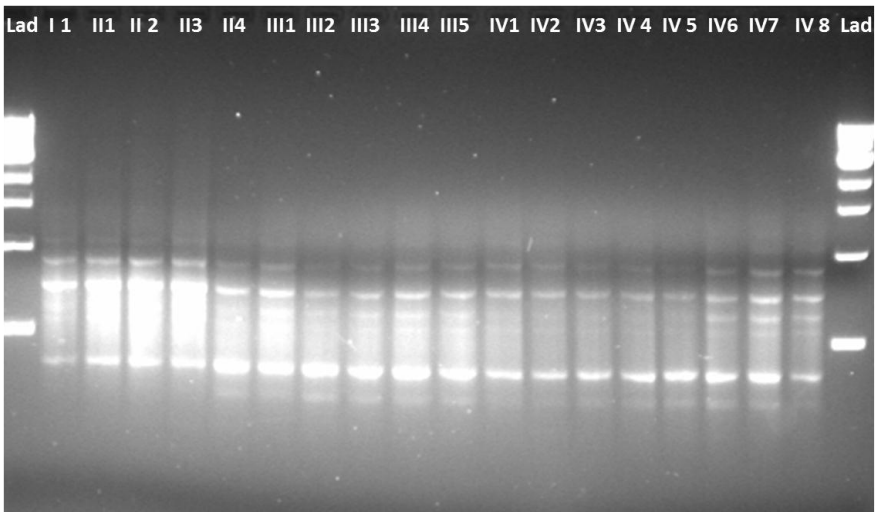


Figure 2.13; PCR products from all individuals of family #1, analysed by agarose gel electrophoresis and ethidium bromide staining. The first and last lanes (Lad) contain the 100bp ladder, lanes 2-18 (I 1, II 1, II 2, II 3…..etc.) show the PCR products of DNA samples from all members labelled as in the family tree. Gel picture shows a smear of PCR product similar to the example provided in the manufacturer's protocol.

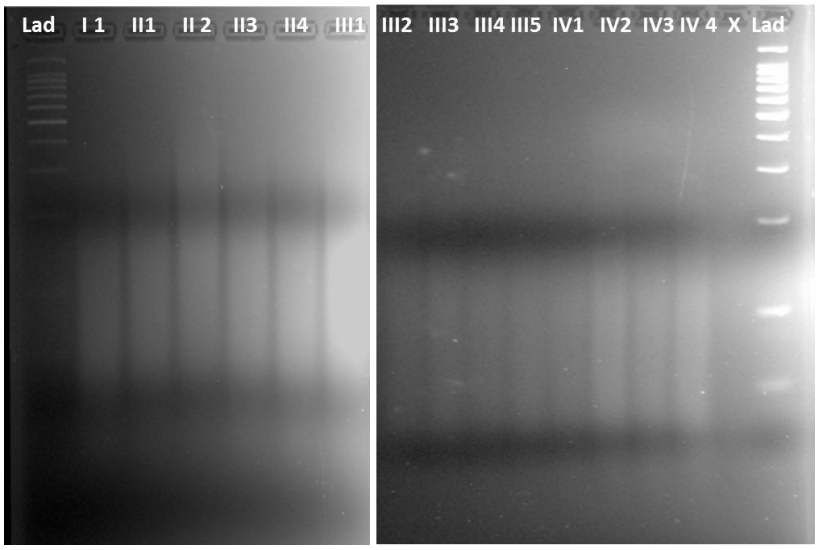


Figure 2.14 Fragmentation products of DNA for some members from family #1 analysed by agarose gel electrophoresis and ethidium bromide staining. The first and last lanes (Lad) contain the 100bp ladder, lanes 2-14 (I 1, II 1…..etc.) show fragmented PCR products of DNA samples from some members of the family, labelled as in the family tree. A smear of PCR products between 100-25 bp is shown, similar to the example in the manufacturer's protocol.

2.5.9 The reference genomic DNA control 103

Following fragmentation, labelling, target hybridization, washing, staining and scanning, the DNA control showed a call rate of > 92%. That is within the manufacturer's recommended specification, which indicates that the assay was performed correctly.

2.5.10 Scanned Data reception

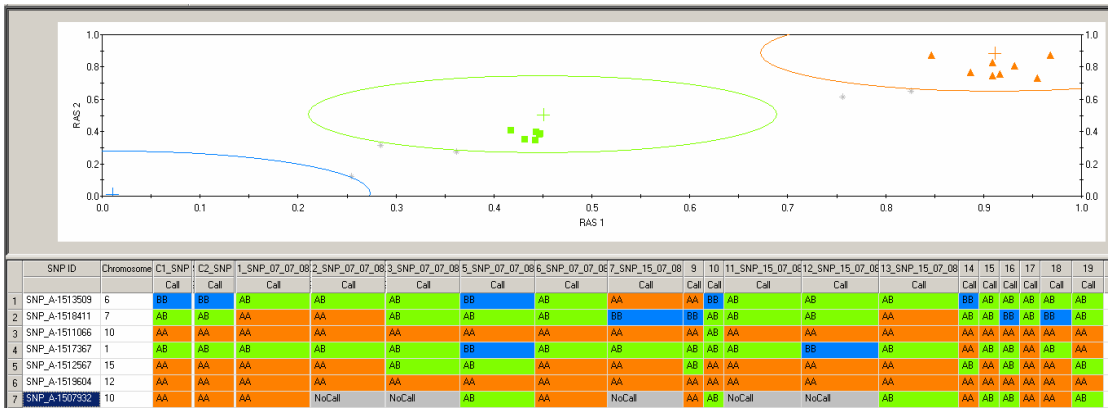
Following fragmentation, labelling, target hybridization, washing, staining and scanning, the data of the 10K GeneChip® Arrays for the 19 DNA samples (of family #1) were imported into GeneChip Operating Software (GCOS), which made it possible to be processed by GeneChip® Genotyping Analysis Software (GTYPE). Using GTYPE software, the genotypes of each DNA sample have been called (as described in in 2.2.9), across 10,208

markers (Figure 2.15). The table containing the genotype data was exported into a Microsoft Excel worksheet (Figure 2.16). The standard allele format for data output from GTYPE is A B representing the genetic calls for each SNP, (described in 2.2.9).

2.5.11 Check Data Drop Rates per Individual, SNPs and Batches

The three steps of data quality control (individual, SNPs and batches) were performed and all SNPs giving "no calls" for more than 4 individuals (5 or more) were removed. Since 19 DNA samples were processed, this rate (4 out of 19) is equivalent to 21%. In the second step, each individual was checked for "no call" rate across the 10208 SNPs and the rate was varying from 900 SNPs and 286 SNPs. That is equivalent to 8.8% and 2.8% respectively. This means none of the samples has a "no call" rate below the threshold of 10%, and thus all individuals remain included (Figure 2.16). In contrast the new and old DNA of (IV-7) demonstrated a "no call" rate of 286 SNPs and 410 SNPs respectively. In the third step, no common error was found in data coming from the same batch.

Following all quality control analyses, data from all individuals and 9,527 SNPs (93%) were retained, for further preparation and then analysis (as will be seen in chapter 3).



	SNP ID	Chromosome	C1_SNP	C2_SNP	1_SNP_07_07_08	2_SNP_07_07_08	3_SNP_07_07_08	5_SNP_07_07_08	6_SNP_07_07_08	7_SNP_15_07_08	9	10	11_SNP_15_07_08	12_SNP_15_07_08	13_SNP_15_07_08	14	15	16	17	18	19
			Call	Call	Call	Call	Call	Call	Call	Call	Call	Call	Call	Call	Call	Call	Call	Call	Call	Call	Call
1	SNP_A-1513509	6	BB	BB	AB	AB	AB	BB	AB	AA	AA	BB	AB	AB	AB	BB	AB	AB	AB	AB	AB
2	SNP_A-1518411	7	AB	AB	AA	AA	AB	AB	AB	BB	BB	AB	AB	AB	AA	AB	AB	BB	AB	BB	AB
3	SNP_A-1511066	10	AA	AA	AA	AA	AA	AA	AA	AA	AA	AB	AA	AA	AA	AA	AA	AA	AA	AA	AA
4	SNP_A-1517367	1	AB	AB	AB	AB	AB	BB	AB	AB	AB	AB	AB	BB	AB	AA	AB	AB	AA	AB	AA
5	SNP_A-1512567	15	AA	AA	AA	AA	AB	AB	AA	AA	AB	AA	AA	AA	AA	AB	AA	AB	AA	AA	AB
6	SNP_A-1519604	12	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA	AA
7	SNP_A-1507932	10	AA	AA	AA	NoCall	NoCall	AB	AA	NoCall	AA	AB	NoCall	NoCall	AB	AA	AB	AB	AA	AA	AB

Figure 2.15; Screen shot of Genotyping Analysis Software (GTYPE) output, showing clusters of genotypes obtained by comparing the ratio of the relative allele signal (RAS) of the A allele to the sum of the RAS for A and B alleles A/(A+B). The table below shows a portion of the calls for single nucleotide

polymorphisms (SNP) with the SNP identity (SNP ID) followed by the chromosomal location of each SNP marker (chromosome) and then 19 columns with the genotypes of the 19 DNA samples (18 family members, 1 in duplicate). Allele calls are highlighted in orange for AA, Green for AB, Blue for BB and gray for No Call. Row 6 shows SNP with monomorphic calls of AA for all individuals. A full description is provided in the text.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
			#NoCall	286	410	494	572	900	356	308	565	651	468	437	549	665	302	359	274	383	656	395
		Physical	%NoCall	0.03	0.04	0.05	0.06	0.09	0.03	0.03	0.06	0.06	0.05	0.04	0.05	0.07	0.03	0.04	0.03	0.04	0.06	0.04
SNP ID	Chr	Position	dbSNP RS ID	4	8	1	2	3	5	6	11	12	13	7	9	10	14	15	16	17	18	19
SNP_A-1513509	6	162491313	rs713055	BB	BB	AB	AB	AB	BB	AB	AB	AB	AB	AA	AA	BB	BB	AB	AB	AB	AB	AB
SNP_A-1518411	7	42608794	rs949459	AB	AB	AA	AA	AB	AB	AB	AB	AB	AA	BB	BB	AB	AB	AB	BB	AB	BB	AB

Figure 2.16; Excel worksheet showing a portion of the calls for single nucleotide polymorphisms (SNP), imported from the Genotyping Analysis Software (GTYPE). It encloses the SNP identity (SNP ID) (column A), followed by the chromosomal location (column B), and the physical distance (column C) of each SNP marker. That is followed by the (national centre for biotechnology information) "NCBI" SNP ID (dbSNP RS ID) [232] (column D), and then the genotyping of all family individuals (18 family members, 1 in duplicate), in A & B format (columns E-W). No call rate counts and percentages for each individual (across all SNPs) are shown in rows 1 and 2 respectively. A full description is provided in the text.

2.6 Conclusion

There are several conclusions to be drawn from the results in this chapter. At the time of performing the analysis using the Affymetrix GeneChip mapping 10K array, with mean distance between SNPs of 210Kb and the average heterozygosity 0.37, it was very popular and was validated in over 250 publications by the end of 2007, with a growing number of researchers getting better results for several applications including linkage analysis, whole-genome association, population genetics and chromosomal copy number changes [233]. Although 100K and 500K mapping arrays were starting to be available, with mean distance between SNPs of 23.6Kb and 5.8Kb respectively and an average heterozygosity of 0.30 for both [234], these mapping arrays use more than one chip (e.g. 2 chip of 250K in case of 500K), and more than one restriction enzyme (e.g. Xba1 and Hind III in case of 100K chip) [235]. Therefore, the 10K chip was the cheapest and least time consuming option and thus was the best choice at this stage as further investigation after Genome Wide Linkage Analysis (GWLA) was planned, to identify any region of interest and undertake more analyses if required.

As mentioned earlier, blood samples from various family members have been obtained at different times. To ensure that results obtained with 'old' and 'new' samples were identical, I have performed the WGS with 10K chip on an old and a new sample from family member (IV-7). After obtaining the genotyping data on Excel file, I have compared the genotype of both samples among 9266 SNPs; data from both samples were identical in all SNPs except 4. The similarity indicates that old samples can be used in the chip analysis when new samples are not available.

Since DNA quantity is one of the main affecting factors in this protocol, DNA quantification and concentration adjustment was required twice in the protocol; once at the start for preparation of Xba1 restriction process, second after PCR purification and before fragmentation. As a result, call rate of the DNA control and sample was within the manufacturer's recommended specification > 92%. That confirmed the accuracy of the assay which means the samples were prepared well and the DNA quality was good enough to perform this Genome scan. In addition, the three quality control steps on the data obtained

has also removed any SNPs not up to the standard and have confirmed that all DNA sample are good enough to proceed to further steps of data analysis .

Processing DNA samples of all family members and obtaining the genotyping data of all SNPs for each individual will enable us to compare the data of affected vs. un affected, and search for the region on the genome which contains common SNPs inherited by all affected individuals but not all unaffected, using GWLA. That region could be linked with the disease allele and one or more of the genes in that region could be playing a role in the pathology of family #1.

Chapter 3 Genome Wide Linkage Analysis of family #1

In chapter 2, I described the hybridization of DNA from all family members to the Affymetrix GeneChip 10K array, which was followed by scanning, to obtain the genotyping of 10,208 SNPs from 19 DNA samples (of family #1) and importing the data in Excel format, in addition to some quality control checks. In this chapter, I will perform further quality control steps on the data, followed by genome-wide linkage analysis (GWLA), (will describe in 3.2.1) with the aim of identifying loci implicated in the thyroid defects affecting the family.

3.1 Introduction

GWLA of a particular family detects the genetic differences between specific individuals in the family known as affected (targets) and the rest of the family members (subpopulation). These differences occur due to crossing over (will describe is 3.1.2) during cell division and meiosis across several generations.

3.1.1 Cells division and meiosis:

When cells in the human body divide the new cell carries a complete set of chromosomes (n=46) which is two copies of each (23), containing an identical copy of DNA (mitosis). However the reproductive cells carry (haploid) only one copy of the chromosome (23) after division (meiosis). Paternal and maternal reproductive cells (gametes) will each carry a single copy of chromosomes. When the gametes combine, that creates a new genetically variable offspring or zygote with 46 chromosomes (23 from father and 23 from mother) (reviewed in [236]). During meiosis chromosome pairs are in close proximity and in homologous regions meet at a point called a chiasma (Figure 3.1) and their genetic material can be exchanged in a process known as crossing over. Consequently, the gamete forms with any combination of paternal or maternal chromosomes.

3.1.2 Crossing over & recombination

Chromosomal crossing over is the exchange of genetic material between homologous chromosomes (carrying the same type of information) leading to a recombination (the outcome of recombinant chromosomes), (Figure 3.1 A & B). It occurs when correspondent

regions on homologous chromosomes break and then reconnect to the other homologous chromosome at meeting point known as chiasma (Figure 3.1 A). The number of chiasma across the chromosome could be non-random and multiple chiasmas can increase the probability of multiple crossovers. This phenomenon provides the basis for a whole genome scan (described in 2.2.1), using SNP markers (Figure 3.1 C). The new regions, due to multiple crossovers in any chromosome, provide a powerful and realistic means of accounting for genetic interference when applied to the problems of gene localization (reviewed in [237]). On the other hand if two crossing-overs occur between mapped loci A and B then recombination between them may be missed. That is why closer markers give stronger output as they can detect every single crossover[183].

Crossing over was first described, in theory by Thomas Hunt Morgan in 1916 who proposed that the crossover frequency might indicate the distance separating genes on the chromosome. Thus the unit of measurement for linkage is known as the Morgan [238].

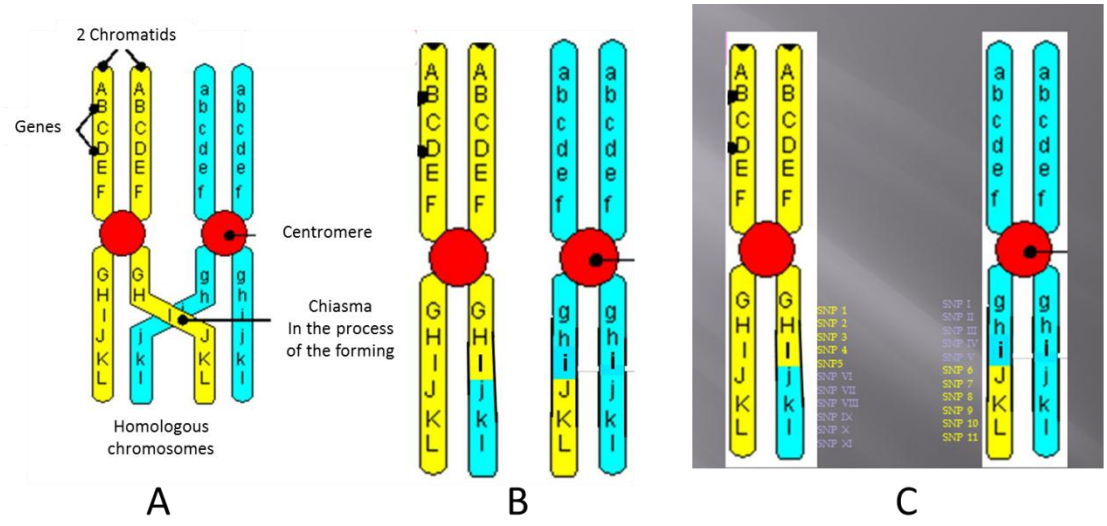


Figure 3.1; Cartoon showing; (A) crossing over between maternal and paternal homologous chromosomes during meiosis, at chiasma. (B) The new chromosome shows a recombination in three genes J, K & L which are replaced by j, k & l. (C) Example of single nucleotide polymorphism (SNP) marker and their transfer during crossover. Modified from; http://online.santarosa.edu

3.1.3 Mendel's law of independent assortment

It is also known as "Inheritance Law" and it states that; "alleles of different genes assort (separate) independently of one another during gamete formation", concluding that different characters (traits) are inherited independently of each other, when genes of these traits are not linked to each other. Along with crossing over, independent assortment increases genetic diversity by producing novel genetic combinations. Thus the possible combinations of gametes formed from maternal and paternal chromosomes will occur with equal frequency (reviewed in [236]).

3.1.4 Linkage

Genes (or loci) on the same chromosome are physically connected or linked. If they are closer they are more likely to segregate together non-randomly. On the other hand loci on different chromosomes are predicted to segregate independently during meiosis, following Mendel's law of independent assortment. Gene pairs that segregate independently have recombination frequency of 50%.

During meiosis, independent segregation occurs when genes are on different chromosomes. Genes on the same chromosome, but wide apart, can also segregate independently since the probability of being inherited together is low. The recombination frequencies of pairs of genes indicate how often 2 genes are transmitted together. For linked genes, the frequency is less than 50%. The greater the distance between linked genes, the higher the recombination frequency. Since crossing over can occur with approximately equal frequency at any point on a chromosome then the probability of any two loci lying on the same chromosome (syntenic) being inherited together will also be 0.5 (reviewed in [239]).

3.1.5 Recombination Fraction

The recombination fraction (θ) (also called recombination frequency) is a statistical measure calculating the probabilities of crossover occurrence between two loci along a chromosome. It ranges between 0 (no recombination) and 0.5 (unlinked loci). The recombination fraction is an indicator of the distance between the loci and thus the zero value of θ gives a clue of tight linkage while a value of 0.5 indicates independent assortment. That is because non-linked alleles have a 50% chance of recombination, due to independent assortment. Thus

recombinant fraction is equal to [R/(NR+R)]. When NR denotes the number of non-recombinant offspring, and R denotes the number of recombinant offspring. Recombination fractions define genetic distance rather than physical distance. To find markers near the disease locus, we should look for a recombination fraction of < 0.5 between them [239].

3.1.6 Physical & Genetic Distance

The length of DNA can be measured in two different ways, by counting the nucleotides between two locations in any DNA, which is known as physical distance and is measured in base pairs (bp). The second method counts the number of recombination events between two loci (on a single chromatid) per meiosis. This is known as genetic distance and is measured in Morgan or centiMorgans (cM), also known as the genetic map unit (m.u.). A centiMorgan is defined as the distance between genes for which one product of meiosis in 100 is recombinant. The correlation between physical and genetic distance can be summarised as one centiMorgan corresponds to about 1 million base pairs in humans on average.

3.1.7 Identity By Descent (IBD) & Identity By State (IBS)

SNPs are described as having identity by type (IBT) when they have the same DNA sequence (calls). SNPs that are identical by type either are identical by descent (IBD) because they inherited from the same member in an earlier generation or are non-identical by descent (NIBD). IBD is often used in genetic linkage and association analyses (will be described in 3.2.1), to identify alleles which are potential candidates for genetic disease [240]. NIBD can also be identical by state (IBS), if they share the same DNA sequence, but are not from the same member in an earlier generation (not origin). Parent-offspring pairs share 50% of their genes IBD, and monozygotic twins share 100% IBD [254].

3.2 Method for Genetic Mapping

3.2.1 Linkage Analysis

Linkage analysis is a strategy that aims to identify genes responsible for certain inherited diseases solely through their chromosomal location within the genome. In the case of scanning the whole genome, a more general approach can be used which is GWLA. Linkage Analysis is a statistical method based on the fact that if two loci are close enough to each

other on the same chromosome, they tend to be inherited together since the chance of recombination between them will be reduced. This analysis tests linkage between genetic loci and any human trait. By investigating the genetic loci it counts the number of recombinants and nonrecombinants, estimates the recombination fraction and tests this fraction to see if it is significantly smaller than 0.5. Linkage studies have been shown to have high power to detect loci that have alleles (or variants) with a large effect size, i.e. alleles that make large contributions to the risk of a disease or to the variation of a quantitative trait [254].

Although linkage studies have been known since 1955[241], the advent of low-cost genotyping platforms for interrogation of several hundred thousand to one million SNPs, coupled with the completion of the second generation haplotype map of the human genome (HapMap2) by the mid-2000s, made genome-wide association and linkage studies feasible. However, identifying the actual sequence variant(s) responsible for these linkage signals is challenging because of difficulties in sequencing the large regions implicated by each linkage peak [242]. This challenge could to be solved by "next generation" DNA sequencing techniques [243], which will be described in chapter 6 of this thesis.

3.2.2 LOD Score

When performing linkage analysis on extended multigenerational families, with known clinically characterized members (affected or unaffected), the parametric logarithm of odds (LOD) score is one of the most popular statistical tools. The LOD score method compares the probability of observing the data if two loci (the sites of genes) are close to each other on a chromosome and are therefore likely to be inherited together, to the probability of observing the same data purely by chance [239]. In this method the likelihood of linkage at different values of θ is tested against the null hypothesis of no linkage (i.e. θ=0.5).

$$\text{LOD score} = \log_{10} \frac{\text{(Likelihood if the loci are linked } \theta)}{\text{Likelihood if the loci are } \mathbf{un}\text{linked } \theta=0.5}$$

If the LOD score between markers is positive, it means that these markers may be close to each other. A LOD score of 3 or more is generally taken to indicate that two gene loci are linked [244]. A LOD score of 3 means the odds are a 1000 to 1 in favour of genetic linkage, (conventional p=0.05 threshold of significance). Negative LOD score, (mainly LOD <-2), on

the other hand, indicates the opposite i.e. rejects the hypothesis of linkage. furthermore, no conclusion can be made when LOD value is above -2 and less than 3 (-2 < LOD < 3) [245].

However, this is based on the LOD score being maximised over only a single parameter (recombination fraction), if the maximisation is over more than one parameter (such as recombination fraction and penetrance) then a more stringent threshold of significance is required as evidence for linkage, i.e. LOD score of 3 is no longer significant and it is necessarily to be larger [246].

It is less common to find a single family large enough (2 offspring at least) to reach statistical significance. However, it is recommended to use data of more than one independent family (with related phenotype), so LOD score can be added up across families. That was my intention in this thesis, to add family #1 data to family #2 data (as described in chapter 4).

3.2.3 Non Parametric Linkage Analysis (npl)

This is the approach by which examination of the pattern of allele-sharing by relatives, (such as siblings, parents and offspring) and transmission from parent to offspring can be made. Methods for this analysis do not make any assumptions about the disease model, such as mode of inheritance, penetrance or disease allele frequency (the parameters). This analysis makes use of the information for alleles identical by descent (IBD) between affected sib-pairs (or discordant sib-pairs) to conclude the location of genes that are linked to the trait of interest [245].

3.2.3.1 Linear and exponential npl

As described above, a popular tool in genetic mapping of complex traits is model free (also call non parametric). That is because no assumption about the underlying mode of inheritance is required. Accurate p-values and LOD scores calculation is very important in linkage analysis between a genomic region and a trait. Depending on the segregation of marker alleles at a genomic region linked to the trait, specification of a model for the trait-dependent segregation could be required. With several families and large number of member in each, the p value can be well approximated by applying normal approximation to the test statistics, which is used in traditional npl studies and known as npl linear model [247]. However, normal approximation may not work well when the data are for one single family, which

usually has a very high number of shared common alleles. In such cases, applying normal approximation can give a very conservative p value due to excessive sharing and thus normal approximation may not be reliable. An exponential model can provide more reliable output in such data and thus is recommended. The two models are much closer and the score statistic corresponding to the exponential model is exactly the same as that of the linear model (reviewed in [248]).

3.2.4 Parametric Analysis

After aggregation and/or segregation studies established a genetic component for a phenotype of interest, parametric linkage analysis has been the traditional approach used for Mendelian disease gene mapping since the 1970's [241]. It is the approach by which examination of the pattern of allele-transmission from parent to offspring can be made. In this analysis a genetic model for the disease must be specified including allele frequency and penetrance parameters. These parameters will be assumed as known without error in order to estimate recombination fractions for testing linkage [249]. The parametric analysis is more powerful than nonparametric if the genetic model specified for analysis is sufficiently close to the true mode of inheritance (MOI) that governs the defect. Thus the key issue in parametric linkage analysis is the specification of the correct genetic model [250].

3.2.5 Multipoint & Single point Linkage Analyses

Linkage analysis used to be described with respect to two loci; however it is now performed in multipoint form (multiple markers per chromosomal region) as this provides much better information on the origin of each chromosomal segment segregating through a family, and much better estimates of IBD sharing [251]. Multipoint analysis is commonly used to evaluate linkage of a disease to multiple markers in a small region and is highly powerful in the studies when the IBD between family members at the trait locus are not very clear, as it uses haplotype information from several markers to infer the IBD.

Single-point (also known as two-point) linkage analyses indicate the mapping using two loci, a single known genetic marker locus and one unmeasured disease loci (or another genetic marker), and that is why it is known as single point and two point. In single-point analysis each marker is analysed separately (reviewed in [252]).

The power of single-point is also decreased by the lack of using haplotype, thus Multipoint analysis is more accurate than (and preferable to) single-point analysis [239]. On the other hand Multipoint linkage analysis is quite sensitive to the correctness of the map on which it is based (due to misspecification of inter marker distances). Such misspecification has more effect when dealing with closely spaced markers. In addition single marker analyses may provide more reliable measures of the strength of support for linkage than multipoint statistics. Therefore comparing the output of both is highly recommended especially when the inter marker distances are misspecified or if there are doubts about the accuracy of the map (reviewed in[239]).

3.2.6 Allele Frequency

During linkage analysis and likelihood estimation, the population frequency of each marker and disease allele is required. It is important to choose the markers which are more frequent in the population, as the uniquely known genotype at a locus in any pedigree indicates that, the gene frequencies for that locus have no effect on the value of the LOD score [253].It should be noted that common variants will be non-informative as well as the monomorphic genotypes (reviewed in [254]).

3.2.7 Haplotype Analysis

A haplotype in SNP analysis refers to a group of SNPs statistically associated and located on a single chromosome of a chromosome pair. Analysing these associations can identify alleles of a haplotype sequence, which can clearly identify all other polymorphic sites in its region. This is able to provide information about gene flow in a pedigree which can be used to reconstruct likely haplotypes for families and individuals. Most of the knowledge of SNP haplotype comes from the international HapMap and 1000 genome projects [255].

With the advanced method of genome-wide SNP microarrays for whole-genome scan, a vast amount of information can be obtained through phased haplotypes, which identify the alleles that are co-located on the same chromosome. Determination of haplotype phase (which is maternal and which paternal), is becoming increasingly important when dealing with large-scale genotyping such as the 10K chip used in this study (reviewed in [256]). In general haplotypes have no defined size and can refer to anything from a few closely linked loci up to an entire chromosome.

3.2.8 Multidimensional scaling (MDS)

Multidimensional scaling (MDS) is a data analysis technique that reduces data into more manageable pieces of information from which conclusions can be drawn. MDS condenses large amounts of data into a relatively simple spatial map that transmits important relationships in the most economical manner [257]. Performing MDS analysis on the genome-wide IBS pair wise distances creates a file containing columns with largest variants across the X-axis (horizontal) in the MDS plot (C1), second largest variant across the Y-axis (vertical) in the MDS plot (C2), third largest variant across the Z-axis (C3) and the fourth largest variant across T-axis (C4). Using Excel scatter graphs, pair wise comparisons for each dimension (C1 to C4) can be performed. For example scatter graph for C1 against C2 will create clusters in which each point is an individual. The distance between these clusters demonstrates the ethnic background of an individual (which population he/she belongs to) [258].

3.3 Data Analysis "Methods"

A wide range of software can be downloaded online and used to check data and perform linkage analysis. Depending on the data and the aim of the study, particular software can be chosen depending on the function of that software and its enhanced performance. In the below section I will explain the software used in data streamlining and quality control.

3.3.1 Software for streamlining and Quality control

PLINK software is an "open-source toolset for whole genome association analysis, designed to perform a range of basic, large-scale analyses in a computationally efficient manner"[258]. "PLINK (one syllable) was developed by Shaun Purcell at the Centre for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the Broad Institute of Harvard & MIT" [258]. It is used regularly in the scientific community for data management, summary statistics, population stratification, association analysis, and identity-by-descent estimation. PLINK is an analysis package and as such supports steps for genotype/phenotype data analysis; it does not provide support for study design, planning, generating genotype or CNV calls from raw data.

PedCheck software reads the pedigree file and checks for any errors (mainly Mendelian) by analysing the entire family in one step [259]. It analyses Mendelian errors in 4 levels, each leads to the next. These levels are: - nuclear-family algorithm (level 1), genotype-elimination algorithm (level 2), critical-genotype algorithm (level 3) and odds-ratio algorithm (level 4). Thus PedCheck can indicate that the data do not have any errors on basis of nuclear family, by level 1 analysis. In level 2, PedCheck creates a new nuclear family, by eliminating founders and checks for errors. This step can identify any errors that have not been detected by level 1. However level 2 checks cannot identify errors created by an individual with genotypes causing LD (critical genotype), when treated as unknown the inconsistency will be eliminated. These errors can be identified by level 3 analysis in which individuals with critical genotypes will be un-typed (one at time), then re checked for consistency. In level 4 PedCheck will identify the valid typings of critical genotypes that eliminate the inconsistency and provide a list of the best genotypes (reviewed in [259]).

Graphical Representation of Relationships (GRR) software investigates the genotype of each marker for every individual and compares them to see how many allele are shared [identity by state (IBS)] at each locus. GRR estimates the mean and standard deviation (stDev) of these IBS counts for each pair of individuals across all markers. The highest percentage of alleles shared by IBS will be in identical twins. First-degree relationships such as full siblings or parent-offspring pairs will show the second highest percentage, followed by half siblings and finally unrelated individuals. The software then generates a graph with X axis (the mean) and Y axis (stDev). Different types of relationship pairs will have different overall mean and stDev IBS values, in first degree related, the value will be much higher than unrelated. If the software shows clear variation in the IBS mean within a category that could be an indication of labelling error or relationship errors (not real parent). GRR is able to show five relationships provided in the PedFile, in different colours (Figure 3.3). Pairs with the same relationship should cluster together. If not then that would be the second indication of labelling error or relationship errors (not real parent) [260].

3.3.2 Software available for linkage Analysis

LOD score calculation has been implemented using different algorithms in several public domain software packages. The first generally available computer program for linkage

analysis was introduced in 1974 [261]. The commonly used programs that perform linkage analysis (but have not been used in this study) are explained in this section. Only Merlin software was used in this study, and will be described in the next section (3.3.3).

FASTLINK software is flexible for the linkage analysis of a large pedigree with a small number of family members with a strongly Mendelian mode of inheritance [262]. FASTLINK is competent for pedigrees containing inbreeding loops [262]. It is more common to be used for parametric analysis than non-parametric [253].

VITESSE Software was shown to be faster (than FASTLINK) to run single- or multi-point analysis for most pedigree structures [262]. Thus it is more common to be used for parametric analysis than non-parametric as is FASTLINK [253].

GENEHUNTER [263] offers a wide package of linkage and association tests and has been used a lot for statistical genetics analysis. It has the advantage of analysing a considerable number of markers although the number of individuals it can handle is limited (approximately 12 non-founders). When the pedigree exceeds the maximum allowed number of individuals, the program discards individuals starting with the unaffected. Splitting the kindred into smaller pedigrees could sort this issue (reviewed in [264]). GENEHUNTER underwent several developmental processes when it was renamed as GENEHUNTER-PLUS (version 1.3) and GENEHUNTER-IMPRINTING. By the later version, Strauch and the team obtained a LOD score of 4.76 for a study on sensitization to mite allergen in an English population [262, 265].

ALLEGRO is a faster version of GENEHUNTER (20-100 fold faster), and has been designed to perform multipoint linkage analyses. It can be used with most pedigree structures, providing more accurate significance levels for non-parametric statistics. It allows 20−30% larger pedigrees and reduces memory requirements by a factor of 20−60 in comparison with GENEHUNTER. It offers more parameters such as family weighting schemes. It provides linear-model LOD score option [261].

3.3.3 Software used in this study for linkage Analysis (MERLIN)

Merlin software is efficient software that provides fast solutions to common problems such as allele-sharing analyses and haplotyping. It provides fast pedigree analyses, including non-

parametric linkage and error detection [266]. It offers a faster algorithm that is particularly useful in dense marker maps, for which the number of recombinations allowed between markers can be constrained [262]. It also provides the linear-model LOD score option as in ALLEGRO. Thus it is considered as a competitor to GENEHUNTER and the fastest programme mainly for non-parametric analysis [261]. The standard non-parametric linkage analysis in Merlin uses the Kong and Cox linear model to evaluate the evidence for linkage [248]. However Kong and Cox "Exponential" model (an alternative likelihood-based non-parametric linkage test), has been shown to be more powerful in samples which include large pedigrees [266], such as family #1 in this study, (--**exp** command). Merlin provides swap-file support for handling very large numbers of markers [266].

3.3.3.1 **Haplotype Analysis (by Merlin)**

Merlin software represents haplotypes estimation in three modes; haplotypes corresponding to the most likely form of gene flow (--**best** command), sample gene flow forms according to their likelihood (--**sample**) or all non-recombinant haplotypes (--**zero,** --**all**). For this study the first model was used (--**best**). The output file (*Merlin.chr*), lists the two haplotypes for each individual, maternal haplotype first then paternal, for non-founders. Between the parents haplotypes several signs gave specific indications (a **|** indicates no recombination event between the current locus and the previous informative locus, a **/** indicates a recombination event in the maternal haplotype, a **** indicates a recombination event in the paternal haplotype, a **+** indicates a recombination event in both the maternal and paternal chromosomes, and a **:** indicates information about recombination between the current marker and the previous marker is not available), (Table 3.5, page 104).

3.4 Aim

The Aim of this chapter was to analyse the SNP genotyping data obtained by Microarray Chip processed in chapter 2, to identify alleles in regions linked to disease. In other words, to identify linkage between the SNP genotypes and "thyroid growth & neoplasia" in the family, and find the region on the genome where all affected members are carrying the same SNPs

but not the unaffected. That would be achieved by performing linkage analysis, after several quality control steps and data streamlining, followed by Haplotype analysis.

3.5 Data Analysis

The SNPs data was acquired (as described in 2.5.10) with standard allele format for data output from GTYPE as (A & B) genetic calls for each SNP, (as described in 2.2.8). The data were imported into an Excel worksheet and some data quality checks were performed (as described in 2.5.11). Following those quality control steps, data from all individuals and 9,527 SNPs (93%) were retained.

3.5.1 Preparation of PedFile (Pedigree File):

The Excel file, imported from the 10K chip, was used to prepare the Pedigree File; a file with details of all family members (father ID, mother ID, sex) together with their genotype as described in (Zhao, J. H.) [267]. The genotype information was in the alphabetical format (A & B) and (No Call), (as described in2.2.8), while the "No Calls" have been converted to 0 0. Since individuals #II-2 and #II-3 have maternal but not paternal SNPs data, a dummy father (I-1) has been created with all details as other members, and a genotype of 0 0 for all SNP markers. The Excel file has then been transposed and saved as a text file (tab delimited). That was followed by separating the genotypes of each SNP with space (to be as; A B or 0 0). The file now can be called PedFile (Figure 3.2).

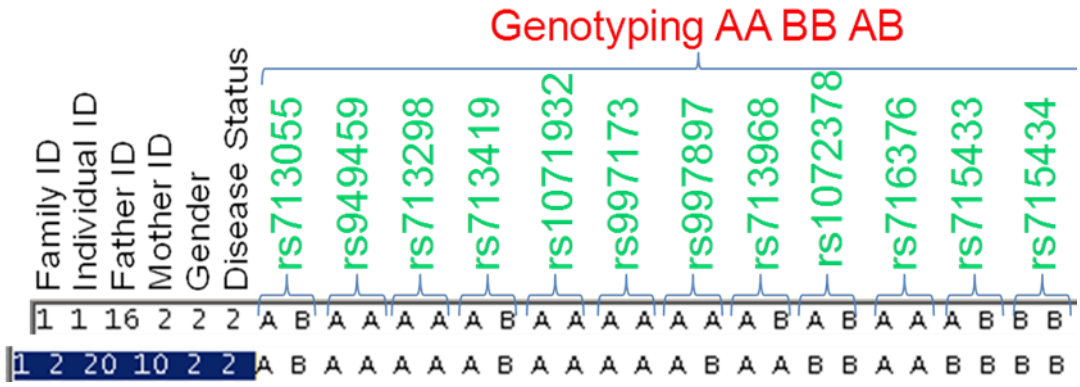


Figure 3.2; Example of pedigree file showing the first six columns containing information about each individual (family identity (ID), individual ID, father ID, mother ID, sex, disease status). That is followed by the genotype information of single nucleotide polymorphism (SNP) for that individual as A A, A B or B B. The SNP markers are shown (in green) above their calls. The first line shows the data of individual III-2 (marked as 1) and the second line that for individual II-2 (marked as 2).

3.5.2 Streamlining the data before analysis (Appendix A1a)

Some SNP markers are not informative and thus will not be analysed, such as 96 SNPs with unknown chromosomal location. Therefore these SNPs have been removed from the data, using PLINK (described in 3.3.1). That reduced the number of returned SNPs to 9,431 SNPs which is equivalent to 92% of the total SNPs at the start. Although monomorphic SNPs are non-informative in linkage analysis, they can be used in allele frequency estimation. Thus they have been conserved within the data.

3.5.3 Mendelian Error Analysis

3.5.3.1 Mendelian error analysis by PedCheck software (Appendix A1b)

PedCheck software (described in 3.3.1), has been used to check the accuracy of family members relationship and mainly to analyse Mendelian errors. Mendelian errors analysis was performed in 4 levels (as described in 3.3.1). PedCheck has found 459 inconsistencies, across 200 SNPs. These SNPs have been excluded from the family data using PLINK software. That reduced the number of returned SNPs to 9,231 SNPs which is equivalent to 91% of the total SNPs at the start. PedCheck was run again and no inconsistencies have been found.

3.5.3.2 **Mendelian error analysis by PLINK software. (Appendix A1c)**

The remaining 9,231 SNPs were then investigated (using PLINK) for Mendelian errors and no further error was detected. However 5 SNPs (on chromosome 1) have been detected in PLINK "HH" file indicating invalid genotypes and thus were removed. That has reduced the number of returned SNPs to 9,226 SNPs which is equivalent to 90% of the total SNPs at the start.

3.5.3.3 **Mendelian error analysis for chromosome X**

Of the 200 SNPs showing Mendelian error by PedCheck, 118 were on chromosome X. The data of the remaining SNPs on chromosome X with no Mendelian errors (122 SNPs), have been extracted (by PLINK) and Mendelian error was re analysed by both PedCheck and PLINK software; none were detected.

3.5.4 IBS Mean check by GRR among the family members

GRR software (described in 3.3.3) was used to confirm the assumed relationships between family members by analysing their genetic data. The three main within family relationships have been seen in this family (Figure 3.3). The 3 common relationship pairs have been shown in this family and the IBS Mean was predominantly more than 1.6 in the sib-pairs and parent-offspring, and less than 1.6 in the unrelated. In addition the duplicated samples in the family have shown an output similar to the *monozygotic* twins with very high IBS Mean (2.0), providing an indication of good data quality.

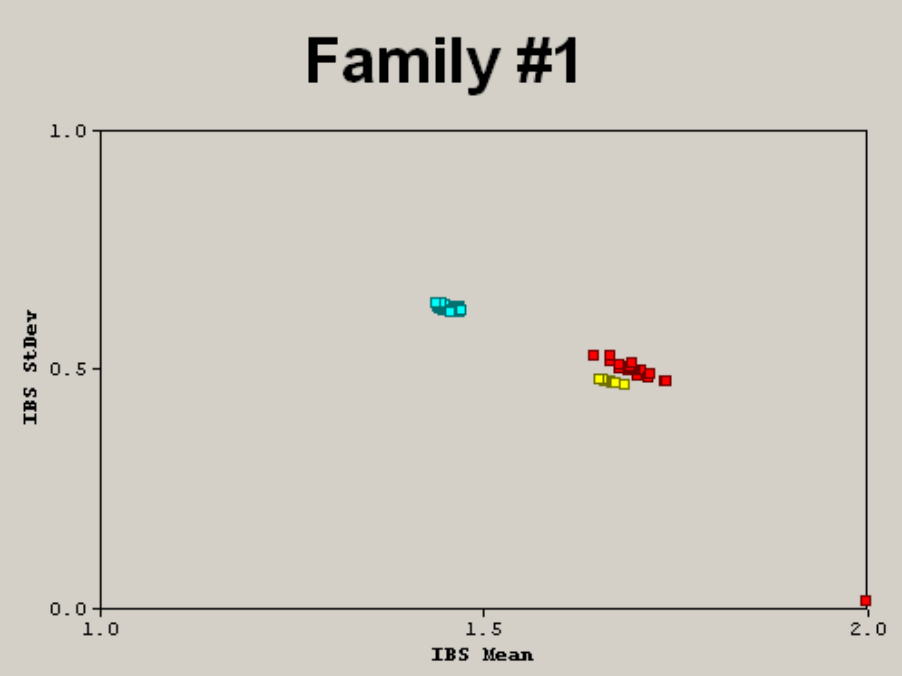


Figure 3.3; Screen shot of graphical representation of relationship (GRR) software's output for family #1. Three clusters were obtained by studying the inheritance of the single nucleotide polymorphisms (SNPs) between the relatives. The software calculates the mean on (x-axis) and the standard deviation (stDev) on (Y-axis) for all SNPs calls to create these clusters. Blue clusters indicate unrelated, red indicate sib-pairs and yellow for parent-offspring. IBS mean value for sib-pairs and parent-offspring is higher (more than 1.6) indicating inheritance of the same SNPs more than the unrelated who shows IBS mean (<1.6). The highest IBS mean value (of 2.0) was shown with the duplicate sample, which provides an indication of good quality of old and new DNA samples.

3.5.5 Data preparation for LINKAGE Analysis

3.5.5.1 **Change 1 2 to A C G T. (Appendix A1d)**

The data (in 1 & 2 genotype format) have been converted to A, C, G or T. To perform this conversion by PLINK, data (provided by Affymetrix Technical support), showing the allele codes of A, C, G or T for each SNP, have been used.

3.5.5.2 **Change all negative strand SNPs to positive. (Appendix A1e)**

All data of negative strand SNPs have been converted to positive using PLINK and data (provided by Affymetrix Technical support), showing all negative and positive strand SNPs.

3.5.5.3 **Merging the data with HapMap data. (Appendix A1f)**

To obtain allele frequency data for each SNP and to investigate the ethnic background of family #1, family data have been merged (using PLINK) with 4 files of HapMap data (founders only), one file for each population, 60 Europeans individuals (CEU), 90 Chinese (CHB) & Japanese (JPT) & 60 Yoruba (YRI). A further 22 SNPs show errors in terms of strand and have been removed, leaving 9,204 SNPs (90%) of the total SNPs to be merged with HapMap data.

3.5.5.4 **Multidimensional scaling (MDS)**

The aim of using MDS (described in 3.2.8) was to study to which population family #1 data are most similar. The MDS analysis has been performed on the merged data of family #1 and HapMap. The ethnicity of family #1 then was determined by importing the data in Excel. Pair-wise comparisons for each dimension (C1 to C4) were plotted against each other for each two dimensions. Family data have been shown to be closer to the European cluster (when plotting C1 against C2), as shown in Figure 3.4). That indicates the family members are belonging to European population and thus European allele frequencies (for all SNPs) will be used for linkage analysis of family #1.

3.5.5.5 **Removing all HapMap data except European**

Since European allele frequency will be used in the Linkage Analysis, HapMap data of other populations have been removed from the PedFile. Thus the new PedFile contains family data and European data only and will direct linkage analysis software to use European allele frequencies. Thus the study will be more precise but loses generalizability.

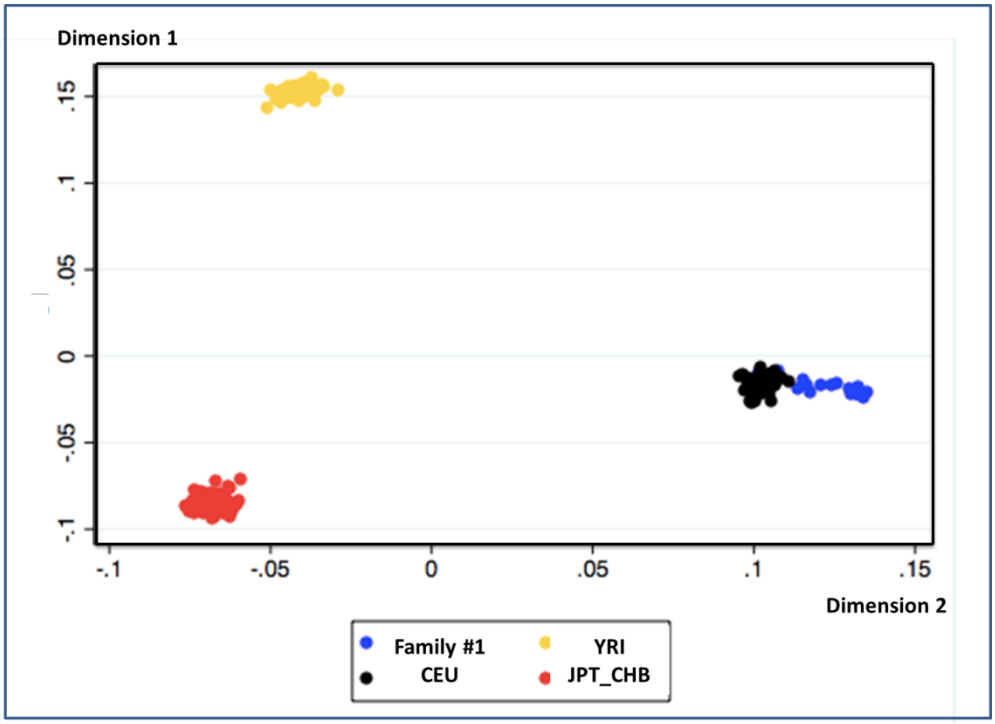


Figure 3.4; Screen shot of multidimensional scaling (MDS) output showing dimension1 (Y axis) plotted against dimension 2 (X-axis) for data from family #1 and HapMap. The scatter plot shows three clusters one for Chinese (CHB) & Japanese (JPT) population (red), one for Yoruba (YRI) population (yellow) and one for Caucasian (CEU) population (black). Family #1 data (blue) are shown to be closest to the cluster of the Caucasian (European) population.

3.5.5.6 Changing the chromosomal position (of all markers) from base pair to centiMorgan (cM) in the Map file (Appendix A1g)

The positions of the SNPs analysed were originally provided in physical distance in base pairs (as described in 3.1.6). Since linkage analysis will be performed using Merlin software, genetic distance will be measured in cM (as described in 3.1.6). Marshfield map, and deCODE map [188] were used to change bp into cM by using Excel sheet and commands. The cM positions of 9111 SNPs (out of the total 9,204 SNPs) were found in the Marshfield map while no position of the remaining 93 SNPs has found in Decode map either. Thus these 93 SNPs have been removed from the data leaving 9,111 SNPs (89%) for further studies.

3.5.6 Data preparation for Merlin software:

3.5.6.1 **Changing SNPs genotyping format from A C G T to 1 2 3 4 in PedFile**

Merlin software, (described in 3.3.3), was used for performing Linkage analysis. This software requires the genotype format to be in numbers rather than letters. Therefore the PedFile data have been changed from A C G T to 1 2 3 4 respectively by using --allele1234 command in PLINK.3.5.6.2 3.5.6.1

3.5.6.2 **Adding age in PedFile**

The age of individuals is one of the parameters required in the parametric linkage analysis (described in 3.2.4). Therefore age has been added in the Pedigree file for each individual, after his or her allele calls. For affected individuals, age at onset (AAO) has been added, while current age (on 2013) was added to others. Since no information is available for the age of European founders of HapMap, age of 55 year was added to each assuming they are old founders. Two different age ranges (44 & 50 years) were tried and no difference was observed in linkage analyses, giving an indication that this age has no effect on the analyses. 3.5.6

3.5.6.3 **Creating Data File**

In addition to the Map and Pedigree files, Merlin requires a third file known as the Data file. This file starts with disease title and it is (Thyroid Cancer) for this study, followed by a list of all SNPs markers (M) and ends with one line for AGE (Figure 3.5). The data file interprets what is in the PedFile. It says that the first 6 columns in the PedFile are for patients of (Thyroid Cancer), followed by genotypes of markers as listed in the data file (under title M) and after the genotypes of all markers there is the age of that patient.

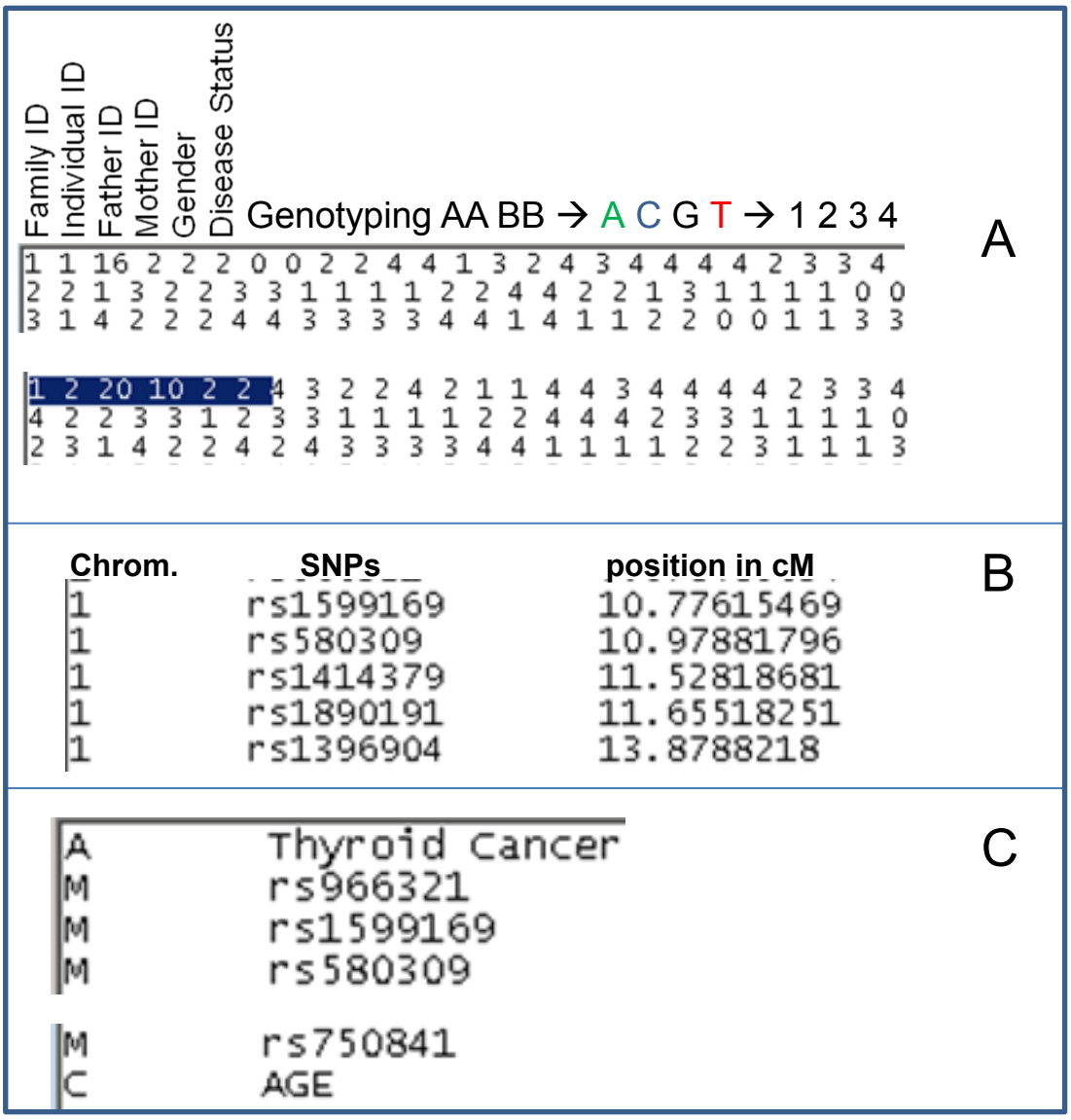


Figure 3.5; Type of files required to perform linkage analyses by Merlin software are shown above. Panel (A) shows the Pedigrees file, in numerical format. Panel (B) shows the map file which includes the list of chromosomes (chrom), single nucleotide polymorphism (SNP) on each chromosome (SNPs) and positions of each SNP in centiMorgans (position in cM). Panel (C) shows the data file which includes a title of the disease (A Thyroid Cancer) at the top, list of markers (M) and Age at the bottom. A full description is provided in the text.

3.6 LINKAGE Analysis by Merlin

When deciding which type of analysis to perform, several factors were taken into consideration; these included the lack of clarity regarding age of disease onset in some family members and the variable penetrance. Therefore non-parametric linkage analysis was selected as the primary analysis and parametric 'dominant model' was chosen for the secondary analysis.

3.6.1 Non -Parametric Linkage Analysis (Appendix A1h)

The first analysis performed was the (model free) Non-Parametric (Multipoint) Linkage analysis (as described in 3.2.5). The Kong and Cox "Exponential" npl analysis has been chosen (as described in 3.3.3). That was followed by performing Single point npl analysis (described in 3.2.5).

3.6.2 Parametric Linkage Analysis

The traditional (model based) Parametric Linkage Analysis (described in 3.2.4) was performed as a secondary analysis. Dominant model of inheritance was the main model to be run and Recessive model was the secondary choice as described above (in 3.6).

3.6.2.1 **Defining Disease Model (Dominant)**

A specific disease model was defined in which 'Thyroid cancer' was selected as disease title and Dominant as mode of inheritance. Since the index patient has been referred for MNG at age 13, thus the age of onset (of MNG for all affected members) was chosen to be above 12 years. Based on the disease history in the family described in (1.18.3), disease penetrance was chosen to be 90% in females (0.001, 0.9, 0.9 for aa Aa AA) and 50% in males (0.001, 0.5, 0.5 for aa Aa AA). SNP allele frequencies will be estimated based on the European population data, disease allele frequency was stated to be 0.01 (Table 3.1).

ThyroidCancer	0.01	* DOMINANT_DISEASE_MODEL
AGE < 12		0.001,0.0001,0.0001
SEX = 1		0.001,0.5,0.5
SEX = 2		0.001,0.9,0.9
OTHERWISE		0.001,0.0001,0.0001

Table 3.1: The inputs file specifying the disease model and other input parameters for Merlin software. First line shows the disease name, allele frequency and mode of inheritance (dominant in this case). Second line shows the disease penetrance before age 12 years (for the three types of inheritance; aa, aA & AA). Third line shows the penetrance in the female above or equal to 12 years. Fourth line shows the penetrance in the male above or equal to 12 years. Fifth line shows the penetrance in any other cases. A full description is in the text.

3.6.2.2 **Parametric Linkage Analysis (Dominant) (Appendix A1i)**

Multipoint linkage analysis, using Dominant mode, was performed by Merlin with the commands described in the appendix. That was followed by Single point Dominant analysis. Both "multi and single" point analyses (described in 3.2.5) have been performed.

3.6.2.3 **Defining Disease Model (Recessive)**

The disease model for both Dominant and Recessive were the same except for the mode of inheritance. Thus with disease penetrance of 90% in females, the mode was defined as (0.001, 0.001, 0.9 for aa Aa AA) and 50% in males as (0.001, 0.001, 0.5 for aa Aa AA), (Table 3.2). The age at onset has been adjusted to be over 12 years as all affected individuals are affected after that age. The reason behind adjusting disease penetrance in female to be 90% is that out of 8 females in the family, 7 were affected. The reason behind adjusting disease penetrance in male to be 50% is that only one male in the family is affected and one male is obligate carrier as he transferred the disease to his daughter.

3.6.2.4 **Parametric Linkage Analysis (Recessive) (Appendix A1j)**

Multipoint linkage analysis, using Recessive mode, was performed by Merlin with the commands described in the appendix. That was followed by Single point Recessive analysis. Both "multi and single" point analyses (described in 3.2.5) have been performed.

```
ThyroidCancer    0.01    * RECESSIVE_DISEASE_MODEL
AGE < 12                 0.001,0.0001,0.0001
SEX = 1                  0.001,0.001,0.5
SEX = 2                  0.001,0.001,0.9
OTHERWISE                0.001,0.0001,0.0001
```

Table 3.2: The inputs file specifying the disease model and other input parameters for Merlin software. First line shows the disease name, allele frequency and mode of inheritance (recessive in this case). Second line shows the disease penetrance before age 12 years (for the three types of inheritance; aa, aA & AA). Third line shows the penetrance in the female above or equal to 12 years. Fourth line shows the penetrance in the male above or equal to 12 years. Fifth line shows the penetrance in any other cases. A full description is in the text.

3.7 RESULTS

3.7.1 Non Parametric Linkage Analysis

The output of nonparametric multipoint linkage analysis on chromosomes 1 to 19 has shown that maximum LOD score values vary from -0.11 (on chromosome 8) and 1.00 (on chromosome 17) and minimum values varys from -0.2 (on chromosome 14) and -0.8 (on chromosomes 8 & 12) . A summary of all chromosomes is shown in (Figure 3.6) & (Table 3.3). Chromosome 20 showed the maximum LOD score of 3.01, while the minimum LOD score was -2.4.

Chromosomes 21, 22 and X showed a maximum values of 0, 0.01 and 0.67 respectively, while the minimum scores were -0.19, -0.06 and -0.03 respectively. Thus the LODs on chromosome 20 were the only score indicating significant evidence for linkage and were located between (20p12.3-20p11.23). In this region 52 SNPs showed LOD scores of 3.01. That was at position between 23.7 cM and 43.2 cM of chromosome 20, over a region of 19.5 cM, and was shown to contain 46 genes, (Figure 3.7).

The output of nonparametric single point analysis has supported the findings of multipoint in all chromosomes including chromosome 20, in which 33 SNPs (out of 53), have single point LODs > 1 in the region identified by multipoint analysis LODs, (over 19.5 cM), (Table 3.4) and (Figure 3.7). That has supported the region identified in multipoint Analysis.

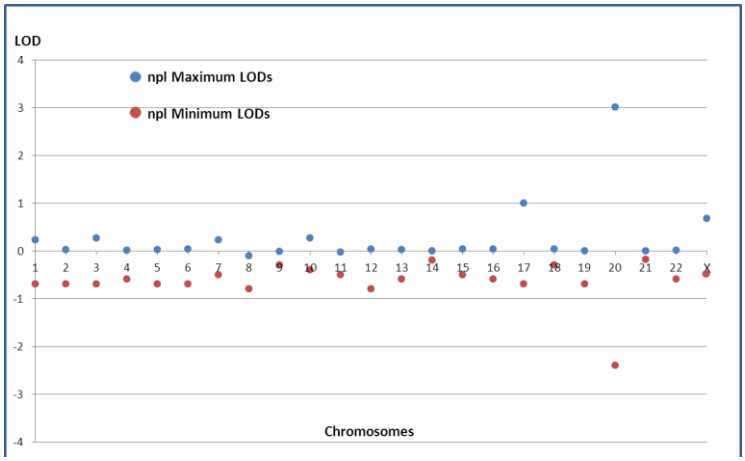


Figure 3.6; Graph showing the maximum and minimum LOD scores (of all chromosomes), obtained by non-parametric linkage analysis. All chromosomes showed maximum LOD score <1.00, except chromosome 20 which showed the maximum LOD score of 3.01, while the minimum LOD score was -2.4.

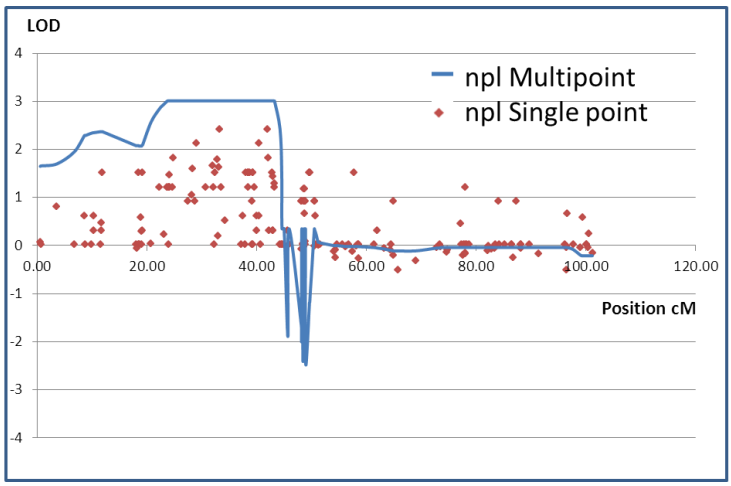


Figure 3.7; Graph showing a regional plot of chromosome 20 comparing LOD scores obtained using multi-point non-parametric linkage analysis (npl) (blue solid line) and single point npl (red dots). The region ($20p^{12.3}$-$p^{11.23}$) displayed the maximum npl LOD score of 3.01 and was supported by single-point npl in which the maximum npl LOD score was 2.41.

3.7.2 Parametric Linkage Analysis (Dominant)

The output of Parametric Dominant multipoint linkage analysis on chromosomes 1 to 19 has shown that maximum LOD score values vary from 0.96 (on chromosome 3) and -2.10 (on chromosome 8) and minimum values vary from -2.2 (on chromosome 10) and -4.1 (on chromosomes 8, 12 & 19) . A summary of all chromosomes is shown in (Figure 3.8) & (Table 3.3). Chromosome 20 showed the maximum LOD score of 2.03, while the minimum LOD score was -2.1. Chromosomes 21, 22 and X showed a maximum value of -2.03, 0 and 0.56 respectively, while the minimum score was -2.39, -2.6 and -1.4 respectively. Thus the LODs on chromosome 20 were the only scores that suggested linkage and were located between ($20p^{13}$-$20p^{11.23}$). In this region 69 SNPs showed LOD scores of 2.03. That was at position (11.69.7- 43.2 cM) over a region of 31.51 cM (Figure 3.9). This region is larger than the one which occurred in the npl, but both shared a common region (Figure 3.10).

The output of Dominant single point analysis supported the findings of multipoint in all chromosomes including chromosome 20, in which 21 SNPs (out of 69), have single point LODs > 1 in the region identified by multipoint, (over 1.51 cM), (Table 3.4) and (Figure 3.9).

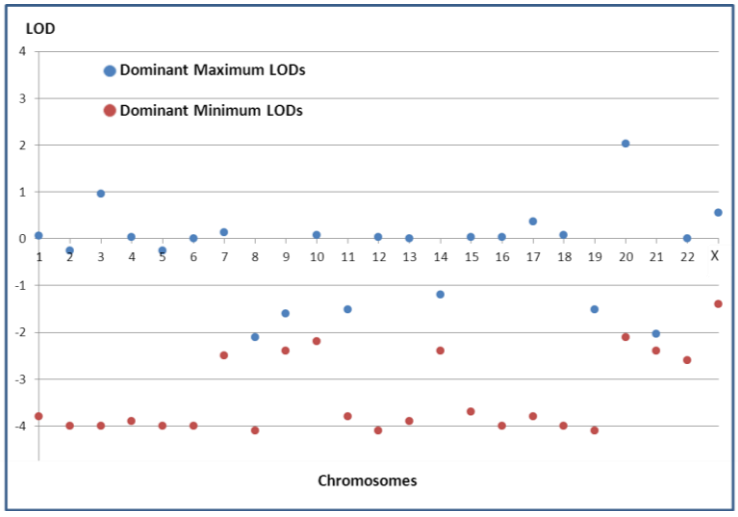


Figure 3.8; Graph showing the maximum and minimum LOD scores (of all chromosomes), obtained by parametric linkage analysis (dominant). All chromosomes showed maximum LOD score <1.00, except chromosome 20 which showed the maximum LOD score of 2.03, while the minimum LOD score was -2.1. Many chromosomes show very low minimum LODs (-4.00_ which is normal in parametric analyses

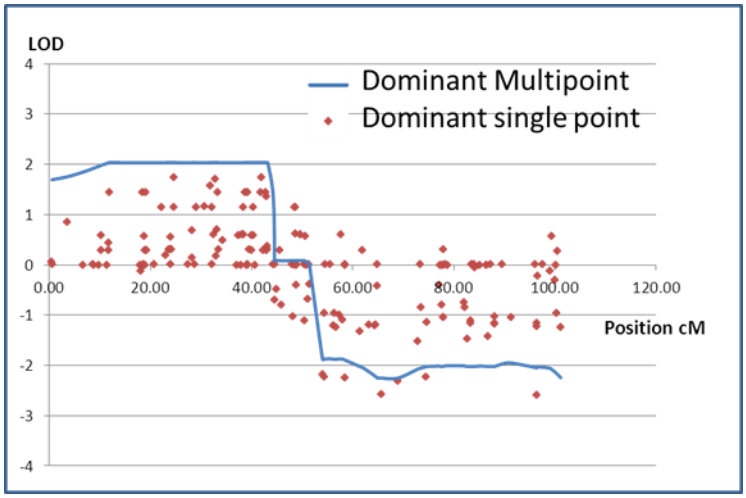


Figure 3.9; Graph showing a regional plot of chromosome 20, comparing LOD scores obtained using multi-point dominant (blue solid line) and single point dominant (red dots). The region ($20p^{13}$-$p^{11.23}$) displayed the maximum dominant LOD score of 2.03 and was supported by single-point dominant in which the maximum LOD score was > 1

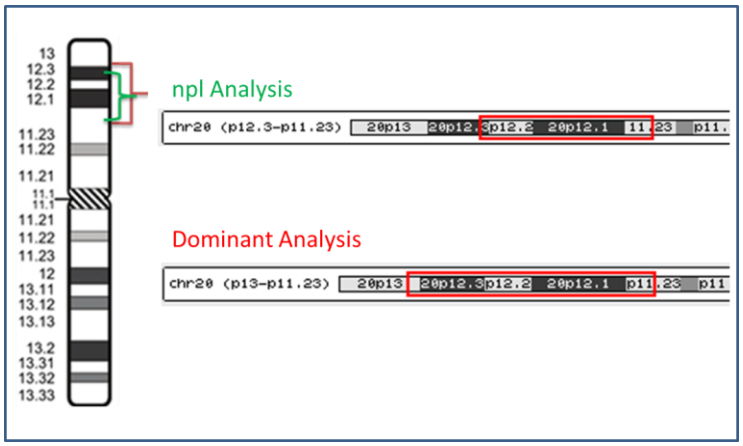


Figure 3.10; Schematic of chromosome 20 showing the region with the maximum LOD value in parametric dominant ($20P^{13}$-$P^{11.23}$, in red) with 57 genes, and non-parametric ($20p^{12.3}$-$P^{11.23}$ in green) with 46 genes. The over lapping region is shown (within the green and the red brackets).

3.7.3 Parametric Linkage Analysis (Recessive)

The maximum recessive multipoint LOD score obtained was less than 1.1 among all chromosomes, except chromosomes 2, 8, 9, 13, 15 & 19 which had no LOD scores above 0 and chromosome 20, on which the maximum LOD was 1.08 (over 149 SNPs). Part of this region was identified in Dominant (Figure 3.11).

The output of Recessive single point analysis has supported the findings of multipoint in all chromosomes including chromosome 20, in which 41 SNPs (out of 149), have single point LODs > 0.6 in the region of multipoint maximum LOD (Table 3.4). That has supported the region identified in multipoint analysis.

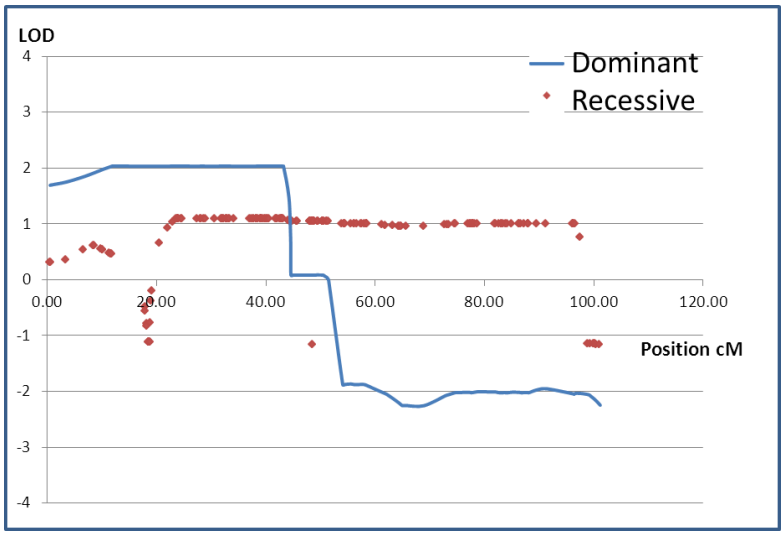


Figure 3.11; Graph showing a regional plot of chromosome 20 comparing LOD scores obtained using multi-point dominant (blue solid line) and recessive (red dots). The region ($20p^{13}$-$p^{11.23}$) displayed the maximum dominant LOD score of 2.03, while recessive LOD score is 1.09.

Chromosome	Maximum LODs, npl	Maximum LODs, Dominant	Minimum LODs, npl	Minimum LODs, Dominant
1	0.22	0.06	-0.7	-3.8
2	0.02	-0.26	-0.7	-4.0
3	0.27	0.96	-0.7	-4.0
4	0.01	0.039	-0.6	-3.9
5	0.02	-0.26	-0.7	-4.0
6	0.03	0	-0.7	-4.0
7	0.22	0.14	-0.5	-2.5
8	-0.11	-2.10	-0.8	-4.1
9	-0.02	-1.60	-0.3	-2.4
10	0.27	0.08	-0.4	-2.2
11	-0.03	-1.51	-0.5	-3.8
12	0.04	0.03	-0.8	-4.1
13	0.02	0.01	-0.6	-3.9
14	0	-1.2	-0.2	-2.4
15	0.04	0.03	-0.5	-3.7
16	0.03	0.04	-0.6	-4.0
17	1.00	0.36	-0.7	-3.8
18	0.04	0.07	-0.3	-4.0
19	0	-1.52	-0.7	-4.1
20	3.01	2.03	-2.4	-2.1
21	0	-2.03	-0.19	-2.39
22	0.01	0	-0.6	-2.6
X	0.67	0.56	-0.3	-1.4

Table 3.3: A summary of maximum and minimum LODs of nonparametric linkage analysis (npl) and dominant analyses for all chromosomes. The first column shows the chromosome number, the second the maximum npl LOD, the third the maximum dominant LOD, fourth the minimum npl LOD and sixth the

minimum dominant LOD. The maximum npl LOD varies from -0.11 and 1.00 (on chromosomes 8 and 17 respectively) and minimum LOD varies from -0.2 (on chromosome 14) and -0.8 (on chromosomes 8 & 12). Chromosome 20 showed the only significant npl LOD (of 3.01). Maximum dominant LOD vary from 0.96 and -2.10 (on chromosomes 3 and 8 respectively) and minimum LOD vary from -2.2 (on chromosome 10) and -4.1 (on chromosomes 8, 12 & 19). Chromosome 20 has showed the only significant dominant LOD (of 2.03).

SNPs ID	Position Mb	Position cM	npl Multipoint LOD	npl Single point LOD	Dominant Multipoint LOD	Dominant Single point LOD	Chromosomal location
rs2250711	8.20	22.19	2.840	1.200	2.034	1.152	
rs771943	8.36	23.02	2.940	0.210	2.035	0.196	$20p^{12.3}$
rs2223539	8.49	23.69	3.010	1.200	2.035	0.303	
rs2327088	8.71	23.93	3.010	0.000	2.035	0.001	
rs2143267	8.76	23.96	3.010	1.200	2.035	0.303	
rs3848831	8.80	23.98	3.010	1.200	2.035	0.302	
rs756374	8.89	24.04	3.010	1.460	2.035	0.553	
rs724089	9.00	24.11	3.010	1.200	2.035	0.302	
rs953021	9.90	24.65	3.010	1.200	2.035	1.142	
rs725565	9.95	24.68	3.010	1.810	2.035	1.739	
rs2327282	10.38	27.42	3.010	0.900	2.034	0.014	
rs973542	10.64	28.17	3.010	1.030	2.035	0.135	
rs2327302	10.66	28.23	3.010	1.590	2.035	0.686	
rs763383	10.85	28.71	3.010	0.900	2.035	0.015	
rs1028846	10.96	29.01	3.010	2.110	2.034	1.146	
rs726867	11.18	30.67	3.010	1.200	2.034	1.154	
rs3887413	11.75	31.90	3.010	1.640	2.035	1.574	
rs2206798	11.85	32.09	3.010	1.200	2.035	1.153	
rs720489	11.90	32.20	3.010	0.000	2.035	0.001	
rs729552	12.13	32.42	3.010	1.500	2.035	0.598	
rs1099620	12.74	32.80	3.010	1.770	2.035	1.704	
rs2327450	12.93	32.91	3.010	0.190	2.035	0.179	
rs721243	13.17	33.06	3.010	1.610	2.035	0.704	
rs2327790	13.46	33.25	3.010	2.410	2.035	1.444	

rs1358766	13.86	33.50	3.010	1.200	2.035	0.303	
rs724820	15.02	34.23	3.010	0.500	2.034	0.482	
rs763659	15.74	37.09	3.010	0.000	2.033	0.000	
rs726207	15.96	37.41	3.010	0.600	2.034	0.585	
rs718610	16.21	37.82	3.010	0.000	2.034	0.000	
rs2144883	16.27	37.97	3.010	1.500	2.034	0.598	
rs1080954	16.45	38.38	3.010	1.200	2.035	1.153	
rs720592	16.45	38.38	3.010	1.500	2.035	0.598	
rs720593	16.45	38.38	3.010	1.500	2.035	0.598	
rs2328024	16.60	38.72	3.010	1.510	2.035	1.441	
rs724053	16.93	39.08	3.010	0.000	2.035	0.000	
rs956110	17.02	39.18	3.010	0.900	2.035	0.013	
rs2010316	17.08	39.24	3.010	1.510	2.035	1.441	
rs2010307	17.08	39.24	3.010	0.000	2.035	0.000	
rs1078530	17.08	39.24	3.010	1.510	2.035	1.441	
rs726256	17.40	39.60	3.010	1.200	2.035	0.304	
rs16823	17.74	39.96	3.010	0.300	2.035	0.289	
rs1073052	17.79	40.02	3.010	0.600	2.035	0.585	
rs2024673	18.12	40.37	3.010	0.600	2.035	0.585	
rs2328245	18.14	40.40	3.010	2.110	2.035	1.153	
rs2328293	18.31	40.61	3.010	0.000	2.035	0.000	
rs1535487	18.94	41.86	3.010	2.410	2.035	1.442	
rs2328361	19.03	42.05	3.010	1.800	2.035	1.734	
rs1074615	19.13	42.25	3.010	0.300	2.035	0.289	
rs2328384	19.37	42.72	3.010	1.510	2.035	1.441	
rs2328410	19.49	42.98	3.010	0.300	2.035	0.289	
rs2328411	19.50	42.98	3.010	0.300	2.035	0.289	
rs2328412	19.50	42.98	3.010	1.420	2.035	1.353	
rs725862	19.58	43.14	3.010	1.200	2.035	0.304	
rs720436	19.60	43.19	3.010	1.280	2.035	0.376	$P^{11.23}$

Table 3.4; The output data of the linkage analyses on chromosome 20 (at the region with over lapping maximum LODs of both, non-parametric linkage (npl) and dominant). First column shows single nucleotide polymorphisms (SNPs) ID, second the position in mega base (Mb), and third the position in centiMorgan (cM). Fourth column for npl multipoint LOD scores, fifth for npl single point LOD, sixth for dominant multipoint LOD, seventh for dominant single point LOD and eighth for chromosomal location. The first and last markers showing maximum LODs (in the table) are highlighted in yellow. A full output data of chromosome 20 are in the Appendix C1a).

3.8 Haplotype Analysis (Appendix A1k)

A Haplotype Study (described in 3.2.7) was then undertaken among all individuals in the region of maximum LOD score on chromosome 20. The aim of the analysis was to identify any significant evidence for biased transmission of one of the haplotypes linked to the disease. The study has been performed using Merlin (--best) as described in (3.3.3.1). The haplotypes of each founder have been labelled in alphabetical formats; A & B for individual (II-4), C & D for individual (I-2), E & F for individual (II-1), G & H for individual (III-1), I & J for individual (I-12) and K & L for individual (III-4) (Figure 3.12).

Using the (*Merlin.chr*) output file (described in 3.3.3.1), a comparison between the haplotypes of affected and unaffected individuals was carried out by following the transmission of the (alphabetically labelled) haplotypes across generations (manually on Excel worksheet). To identify the disease haplotype, the inheritance of each alphabetic label was looked for, seeking for what has only been inherited by the affected individuals and assuming it as a disease haplotype. The study has shown several crossings over (defined in 3.1.2), between paternal and maternal haplotypes among the siblings and that have increased and/or decreased the LOD score in several regions (Table 3.5).

The genotype of affected individual (I-1) has been labelled as C D; the same C haplotype has been inherited by her affected daughter (II-2) and her (disease carrier) son (II-3) (who transferred it to his daughter (III-5). The same C haplotype was inherited by the next generation female, (III-2) who has transferred it to all her affected children (IV-3, IV-6, IV-1 and IV-7). Some of her unaffected children have also inherited this haplotype (IV-4 and IV-5), and are assumed to be carriers. Only individual (IV-2) of this nuclear family has not inherited the C haplotype (as well as his maternal uncle (III-3) and his son (IV-8). Therefore the data of these three individuals (IV-2, III-3& IV-8) have been used for any comparison between family members to identify linkage to the disease haplotype. Since C haplotype was shown to be inherited by the affected individuals (and carriers) only at a specific region (from 2.27Mb to 12.13Mb) which is about 10 mega bases (with 53 genes), it could be assumed that the disease haplotype is linked to this region[83].

3.8.1 Haplotype vs. Linkage Data

The haplotype analysis demonstrated that the assumed disease region began at position 2.27Mb on chromosome 20, and ends at 12.13 Mb. Since npl maximum LODs started at position 8.49 Mb, the region above will be ignored. In addition the maximum LODs in both Dominant and npl analyses were maintained until 19.60 Mb, but the second crossing over was at 12.13Mb, thus the region below will be ignored. That will make the region of interest to be "between 8.49Mb and 12.13Mb" because it is the region where the overlapping between haplotype and npl maximum LODs occurs. That region is 3.64 Mb long ($20p^{12.1}$ – $20p^{12.3}$) and comprises 10 genes. These genes are PLCβ1, PLCβ4, LAMP5, PAK7, ANKRD5, SNAP25, MKKS, C20orf94 (later known as LOC128710 or SLX4IP), JAG1 and BTBD3, (Figure 3.12) and (Table 3.5).

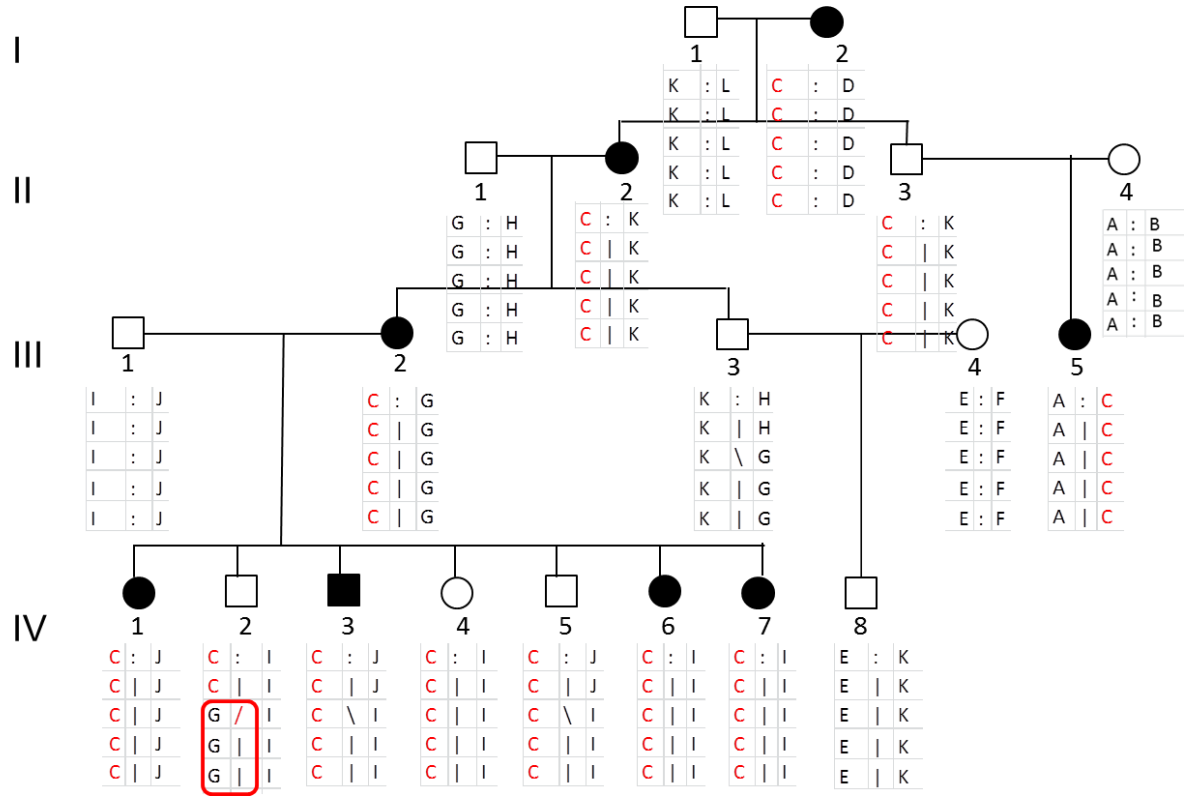


Figure 3.12; Family tree, showing a portion of the output of the haplotype study using Merlin software. The software has labelled the haplotypes of each founder in alphabetical format. A & B for individual II-4, C & D for individual I-2, E & F for individual III-4, G & H for individual II-1, I & J for individual III-1 and K & L for individual I-1 (estimated). The assumed disease allele C in (I-2) has been inherited by her daughter (II-2) and her son (II-3). Both children have transferred the same allele to their daughters (III-2 & III-5 respectively). Individual (III-2) has transferred the same allele to her 6 out of her 7 children (IV-1, IV-3, IV-4, IV-5, IV-6 and IV-7). Only (IV-2) has not inherited the assumed disease allele C, and has shown a clear "maternal" crossing over at third haplotype. Thus the region after this crossing over is assumed to be linked with the disease. More details are in (Table 3.5) below.

position	Linkage Analyses		Individuals haplotypes					
Mb	npl	Dom	II-4 (F)	I-1 (F)	II-2 (10,20)	III-2 (2,16)	IV-2 (1,19)	III-3 (2,16)
0.12	1.57	1.611	A:B	C:D	C:K	C:G	C:I	K:H
0.69	1.64	1.69	A:B	C:D	C|K	C|G	C|I	K|H
1.18	1.91	1.803	A:B	C:D	C:K	C:G	C:I	K:H
2.27	2.25	1.881	A:B	C:D	C|K	C|G	G/I	K\G
3.68	2.33	1.958	A:B	C:D	C|K	C|G	G|I	K|G
4.25	2.36	2.022	A:B	C:D	C|K	C|G	G|I	K|G
4.26	2.36	2.022	A:B	C:D	C|K	C|G	G|I	K|G
4.33	2.36	2.027	A:B	C:D	C|K	C|G	G|I	K|G
6.44	2.07	2.024	A:B	C:D	C|K	C|G	G|I	K|G
6.53	2.07	2.024	A:B	C:D	C|K	C|G	G|I	K|G
6.79	2.07	2.025	A:B	C:D	C|K	C|G	G|I	K|G
6.92	2.07	2.025	A:B	C:D	C|K	C|G	G|I	K|G
6.96	2.07	2.025	A:B	C:D	C|K	C|G	G|I	K|G
7.12	2.07	2.026	A:B	C:D	C|K	C|G	G|I	K|G
7.17	2.07	2.026	A:B	C:D	C|K	C|G	G|I	K|G
7.22	2.08	2.026	A:B	C:D	C|K	C|G	G|I	K|G
7.25	2.09	2.026	A:B	C:D	C|K	C|G	G|I	K|G
8.20	2.84	2.032	A:B	C:D	C|K	C|G	G|I	K|G
8.36	2.94	2.035	A:B	C:D	C|K	C|G	G|I	K|G
8.49	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
8.71	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
8.76	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
8.80	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
8.89	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
9.00	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
9.90	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
9.95	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
10.38	3.01	2.034	A:B	C:D	C|K	C|G	G|I	K|G
10.64	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
10.66	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
10.85	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
10.96	3.01	2.034	A:B	C:D	C|K	C|G	G|I	K|G
11.18	3.01	2.034	A:B	C:D	C|K	C|G	G|I	K|G
11.75	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
11.85	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
12.13	3.01	2.035	A:B	C:D	C|K	C|G	G|I	K|G
12.74	3.01	2.035	A:B	C:D	C|K	C|G	G|I	C/G
12.93	3.01	2.035	A:B	C:D	C|K	C|G	G|I	C|G

Table 3.5: Haplotype study of some members of family #1 (across part of chromosome 20). First column (position Mb) shows position in mega base (Mb), second (linkage analyses npl), non-parametric linkage

analysis (npl) LODs & third (linkage analyses dominant) dominant LODs. The next columns show the haplotypes of the individuals. Column four is for II-4 haplotypes, five II-2 haplotypes, six III-2 haplotypes and seven III-3 haplotype. Merlin software has marked founders with (F). For non-founders, parents ID added in brackets (instead of F). Haplotype pairs are separated by a | for no recombination, : for no information on recombination and /, \, + for recombination (in the maternal, paternal & both haplotype respectively). At position 8.49 Mb the npl showed maximum value (highlighted in yellow) and thus the region after that assumed as significant. Crossing over in individual III-3 occurred at 12.74 Mb (highlighted in blue), led him inherit C haplotype again and thus the region below was counted as unlinked. A full haplotype data of chromosome 20 are in the Appendix C1b)

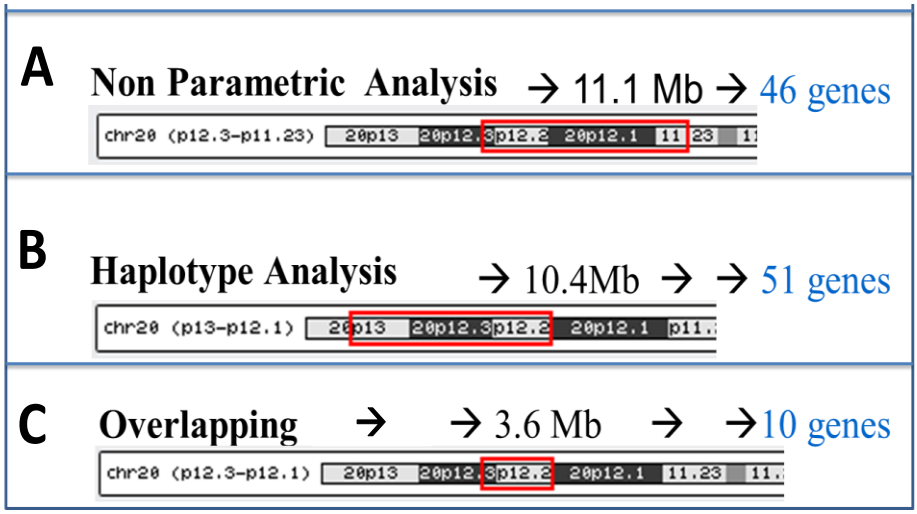


Figure 3.13: Summary of the regions of interest identified by non-parametric linkage analysis (npl) and Haplotype studies and number of genes located in each region. (A) Region of 11.1 Mega base (Mb) showed maximum LODs obtained by npl and contains 46 genes. (B) Region of 10.4 Mb showed linkage with disease haplotype and includes 51 genes. (C) Region of 3.6 Mb showed overlapping of both npl and Haplotype analyses and includes 10 genes.

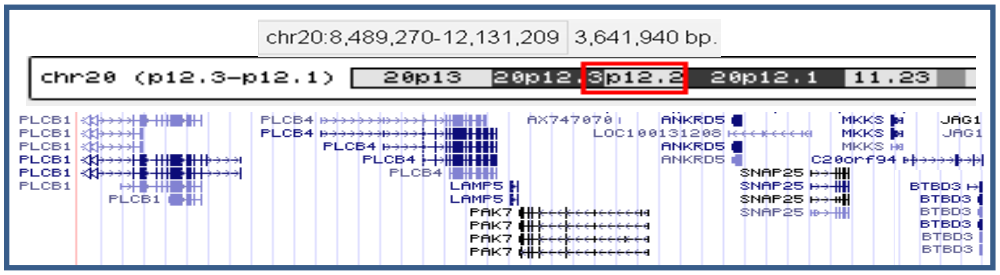


Figure 3.14: The 10 genes in the region of 3.6 Mega base (Mb) where non-parametric linkage analysis (npl) maximum LOD and Haplotype analyses overlapped. These genes are PLCβ1, PLCβ4, LAMP5, PAK7, ANKRD5, SNAP25, MKKS, C20orf94 (later known as LOC128710 or SLX4IP), JAG1 and BTBD3

3.9 Conclusion

There are several conclusions to be drawn from the results in this chapter. As mentioned in chapter 2, blood samples from various family members have been obtained at different times. Although the similarity of old and new sample has been indicated after obtaining the genotyping data of SNPs, further data check steps also have confirmed that finding. In GRR output the relationship detected by the software showed the IBS between old and new DNA to be 2.00, as if they are identical twins. Furthermore, after performing all quality control steps the quality of the data was good enough to be considerable for using in GWLA.

Although only one old DNA sample has been directly compared to a new one the finding that the output was identical for both indicates that reliable data can be obtained from old DNA samples, although it does not confirm that all old DNA samples will provide reliable data. Further direct comparisons of old and new samples would be needed to provide a good indication as to the reliability of data from old DNA samples.

In the linkage analyses, the half of 20p closest to the telomere has shown significant LOD score in npl from the beginning of the chromosome. Although the first 8 mega bases have not given the highest score, the next 19 mega bases showed the highest score of 3.01 (from 8.49Mb). After a complete study of the output all three linkage analyses (npl, dominant &

recessive), and performing haplotype analysis, I came to a conclusion that the linkage region between disease locus and markers could be the overlapping region between haplotype and npl highest LODs which is 3.64Mb long (between 8.49Mb and 12.13Mb) and comprised 10 genes at $20p^{13.3}$-$P^{12.1}$ (Figure 3.13) and (Table 3.5).

In contrast, the big region (34.7Mb) with a significant LOD (8.49Mb to 43.19Mb) could suggest more than just a single mutation in a candidate gene. This could indicate a major chromosomal re-arrangement rather than a point mutation in a single gene. However, the Karyotyping of the index patient reveal 46 chromosomes of normal shape and size; although such analyses would not detect small deletions, which might also explain the high LOD score observed over the long distance in chromosome 20p.

To conclude on the LOD score across all chromosomes (other than chromosome 20), the maximum npl was 1.0 on chromosome 17 and maximum dominant was 0.96 on chromosome 3. Minimum npl was -0.8 on chromosomes 8 and 12 and minimum dominant was -4.1 on chromosomes 8, 12 and 19. However npl LOD score on chromosome 20 showed the maximum (3.01) and minimum (-2.4), which are the highest and the lowest value across all chromosomes (Table 3.3). This could add more to our understanding of what exactly occurred in that region.

Chapter 4 Microarray and GWLA of Family #2

This chapter analyses the data of a second family (family #2) which has been introduced to my department after publishing the data obtained during my MPhil study [181]. Several members of this new family have shown similar signs and symptoms to that of family #1 (Euthyroid MNG of adolescent onset) and same histopathology picture in some members of both families. Linkage analyses will be performed on this family as in chapter 4. Then both families will be merged and linkage analyses will be run on the merged data.

4.1 Family #2 History

4.1.1 Index patient

The family tree of family #2 is shown in (Figure 4.1). The index patient (III-1, in the family tree), a boy (D.O.B 4/8/90) has been refereed (at age 18 years) to the Paediatric Endocrine service based at the University Hospital of Wales, Cardiff, with clinical evidence of MNG (bilateral thyroid nodules), which were noted previously at age 16 years, he was found to be euthyroid when his thyroid function was tested.

The boy had been born at 39 weeks gestation by normal vaginal delivery, weighing 8lb 8ozs. The prenatal history was uneventful and the neurodevelopment had been within normal limits. The boy was of average size at birth.

Ultrasound scan of the neck was used to identify thyroid nodules. The scan indicated that although cystic, the nodules were complex. Previous ultrasound reports (03/2007, 10/2007 & 03/2008) show similar findings with thyroid lobe (cystic containing solid elements), measuring 4.1 cm on the left and 3.4 cm on the right.

At age 18 he underwent thyroidectomy and the histopathological picture showed multiple papillary adenomatosis. Since that age he is euthyroid, receiving daily replacement of thyroxine.

4.1.2 Family history

The Index patient's sister (III-2) (D.O.B 1/7/97) has been referred (at age 11 years) to the Paediatric Endocrine service based at the University Hospital of Wales, Cardiff, with a thyroid swelling. Ultrasound scan (at age 8 years) confirmed the presence of a Multi-nodular thyroid

gland. Thyroid peroxidase antibodies were negative at that age. Subsequent follow up thyroid ultrasound scan at age 11 years continued to show heterogeneous nodularity of the thyroid gland, particularly in the right lobe. Some of the previous nodules have changed in appearance and reduced in size. There was no evidence of microcalcification (associated with PTC). She appeared to be clinically and biochemically euthyroid. At age 14 years a dramatic expansion in the right lobe of her thyroid recommended thyroidectomy be performed with some urgency.

The father (II-2) (DOB; 18/5/69) was found to have sudden onset of goitre at age 22 years. The thyroid histopathology revealed Multi-nodular goitre with numerous hypoplastic nodules. One of the nodules had infarcted either spontaneously, or following a Fine Needle Aspirate (FNA). The other nodules 10 to 20 in number, were well circumscribed, and are composed of follicles of varying size, with papillary infolding. Some were more cellular, with crowded oval nuclei, but none show the nuclear features of papillary carcinoma. The histopathological picture (multiple papilloid adenomas) strongly resembles that seen in family #1. He underwent thyroidectomy and is currently euthyroid, receiving daily replacement with thyroxine.

The boy's paternal grandmother (I-2), (DOB; 17/12/45) had non-toxic nodular goitre at age 10 and underwent thyroidectomy at age 17 years. Her niece (II-5), (DOB 5/2/53) underwent total thyroidectomy at age 27 years. Her 2 sisters (I-4 & I-6) underwent thyroidectomy for undefined reason. The boy's paternal uncle (II-4) is also affected (MNG) as well as his daughter (III-4), (DOB 2000) who had thyroid swelling (MNG) at age 5 years. His mother (II-1), (DOB 1/1/69) is unaffected.

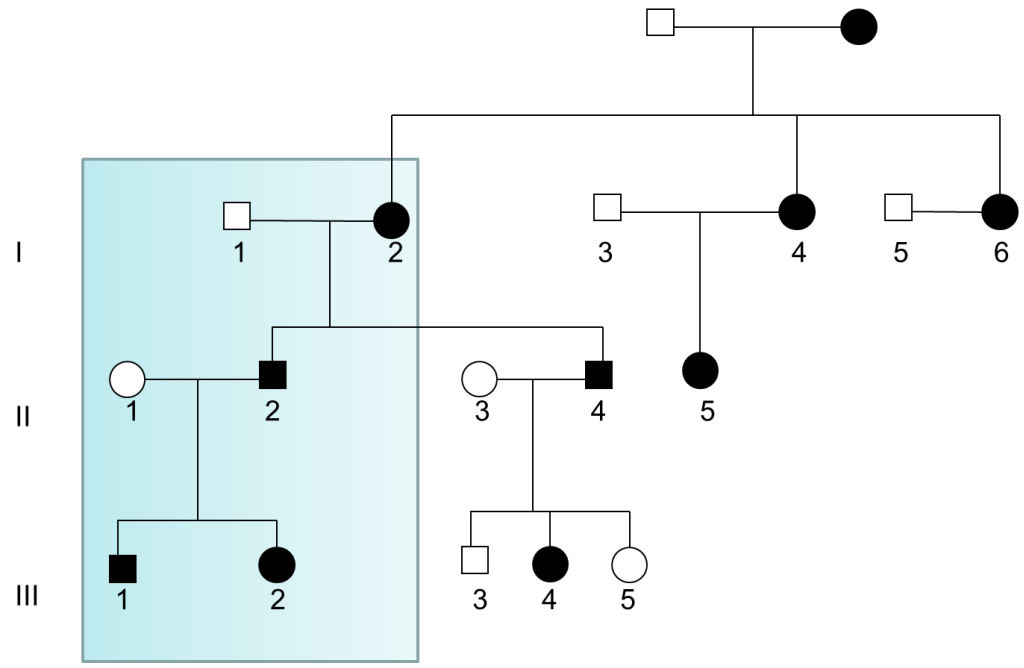


Figure 4.1; Tree of Family #2. Affected members are in black boxes. The Index patient III-1 was referred with multinodular goitre (MNG). Thyroid ultrasound of his 11 year old sister III-2 confirms MNG, but no evidence of PTC. His father II-2 has sudden goitre at age 22 years, with histopathological picture resembles to what seen in one member in family #1 (III-2). His paternal grandmother (I-2) had non-toxic MNG, her niece and 2 sisters underwent thyroidectomy. The boy's paternal uncle (II-4) is also affected (MNG) as well as his daughter (III-4), (MNG). The boy's mother (II-1) is unaffected.

4.2 Aim

To determine whether the genes predisposing to goitre and thyroid cancer have any similarity in these two families, displaying similar clinical and histological features.

4.3 GeneChip® Mapping Assay Materials, Methods & Results

4.3.1 DNA Extraction from peripheral lymphocytes

Blood samples from 6 members of family #2 (green box in family tree (Figure 4.1)), were collected with informed consent (EDTA vacutainers Beckton Dickenson). DNA was extracted, quantified, digested with Xba1 restriction enzymes, the linkers then were ligated (as described in sections 2.3.3 to 2.3.7). The same steps were performed in parallel on the reference genomic DNA 103 (described in 2.2.4).

4.3.2 PCR

Eight PCR reactions have been performed for each DNA sample, all were quality checked using 2% agarose gel electrophoresis (Figure 4.2), prior to purification and adjustment of the concentration of the purified PCR product (all as described in sections 2.3.8 - 2.3.10).

4.3.3 Fragmentation

Subsequent fragmentation of all samples was performed as previously described (sections 2.3. 12) and analysed by 4% agarose gel electrophoresis (Figure 4.3).

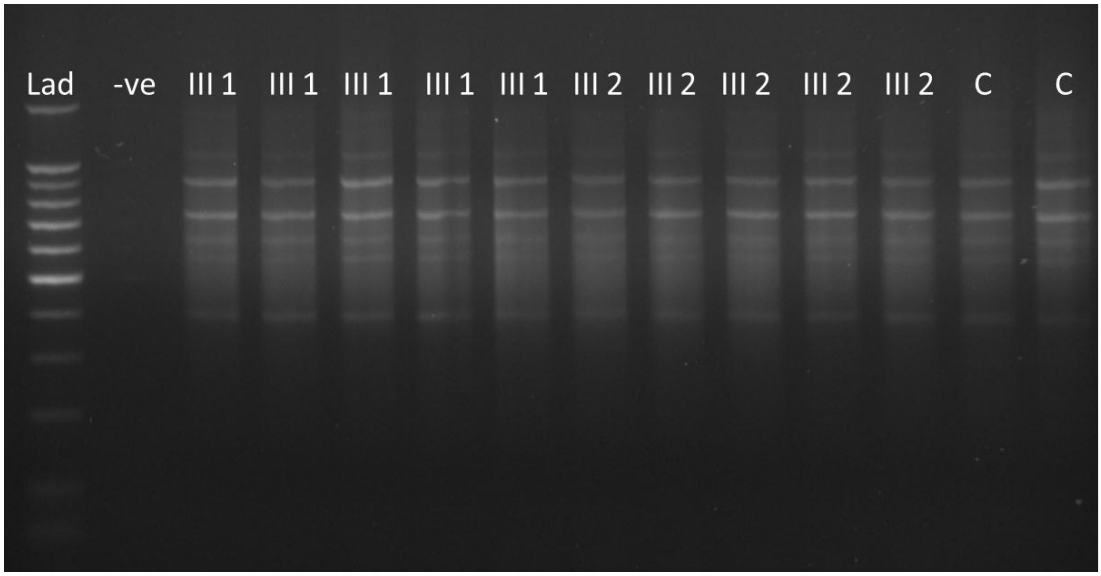


Figure 4.2; PCR products from individuals III-1 and III-2 analysed by agarose gel electrophoresis and ethidium bromide staining. The first lane (Lad) contains the 100bp ladder, the second lane (-ve) contains the negative control (master mix without DNA template), lanes 3-7 (III 1) show the PCR products of individual III-1 and lanes 8-12 (III 2) with DNA sample of individual III-2. Lanes 13 & 14 (C) show the amplicons obtained using "reference genomic DNA control 103" provided with the kit. Gel picture shows a smear of PCR product similar to the example provided in the manufacturer's protocol

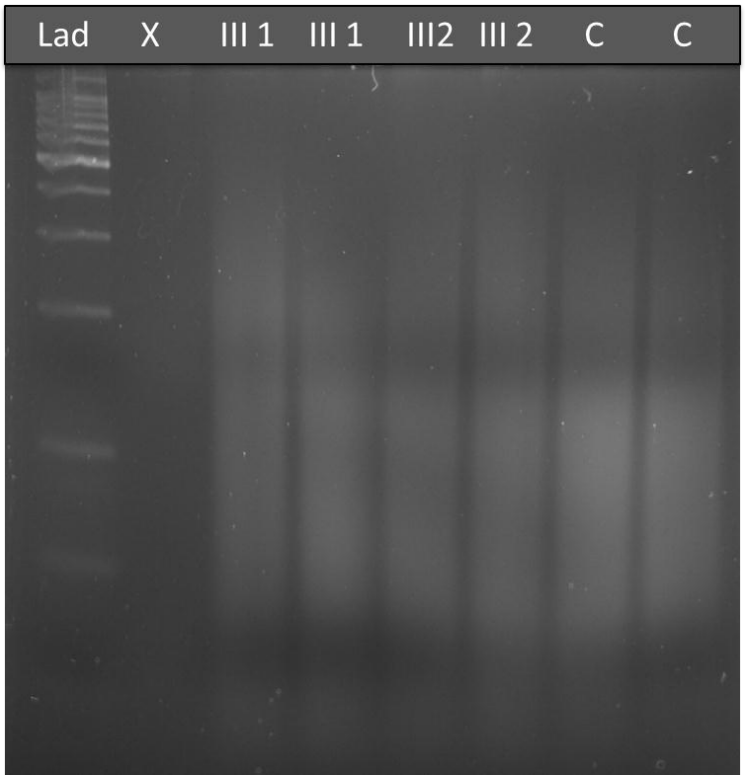


Figure 4.3; Fragmentation products of DNA for individuals III-1 and III-2 analysed by agarose gel electrophoresis and ethidium bromide staining. The first lane (Lad) contains the 100bp ladder, the second lane (X) empty, lanes 2 & 3 (III 1) for sample of individual III-1, lanes 5 & 6 (III 2) for sample of individual III-2 and lanes 7 & 8 (C) for "reference genomic DNA control 103" provided with the kit. A smear of PCR products between 100-25 bp is shown, similar to the example in the manufacturer's protocol.

4.4 Genome Wide Linkage Analysis, Family #2

4.4.1 Data Reception

The hybridization quality has been checked and then genotyping data was imported to Excel worksheet (as described in section 3.3.1, for family #1).

4.4.1.1 **Check Data Drop Rates per Individual, SNPs and Batches**

The quality control checks were performed as in section 3.3.2. The standard for family #1 regarding omitting SNPs was; "any SNP giving No Call for more than 4 individuals (out of 19) was omitted (i.e. SNPs show 21% No Call, have been omitted). To keep the same standard for family #2, any SNP was giving No Call for more than 1 individual, out of 6 was omitted, (i.e. SNPs show 17% No Call, have been omitted). Total number of SNPs after omission was 9,803 (96%). A further 100 SNPs with unknown chromosomal location were removed. Following all quality control analyses, all individuals were included and 9,703 SNPs (95%) remained for further study.

4.4.1.2 **Data Streamlining and Errors check**

The data then were prepared to be in PedFile format as described in section (3.5.1). That was followed by streamlining as described in (3.5.2).

Mendelian errors have been analysed using PedCheck software and 156 inconsistencies have been detected, among 122 SNPs, 105 of which were on chromosome X. These SNPs have been removed from the data (as in 3.5.2 & 3.5.3.1). That reduced the number of returned SNPs to 9,581 SNPs which is equivalent to 94% of the total SNPs at the start.
Mendelian errors have also been analysed on the remaining 9581 SNPs, no error was detected. However a further 2 SNPs (on chromosome 2) have shown invalid genotypes (in PLINK .hh file) and thus have been removed. That has reduced the number of returned SNPs to 9,579 SNPs which is equivalent to 94% of the total SNPs at the start.

The data of the remaining SNPs on chromosome X have been analysed for Mendelian error by both PedCheck and PLINK. No further errors were detected.

IBS Mean was checked using GRR software among the family members as described in section (3.5.4), no variation was obtained in the output indicating no labelling or relationship errors (Figure 4.4).

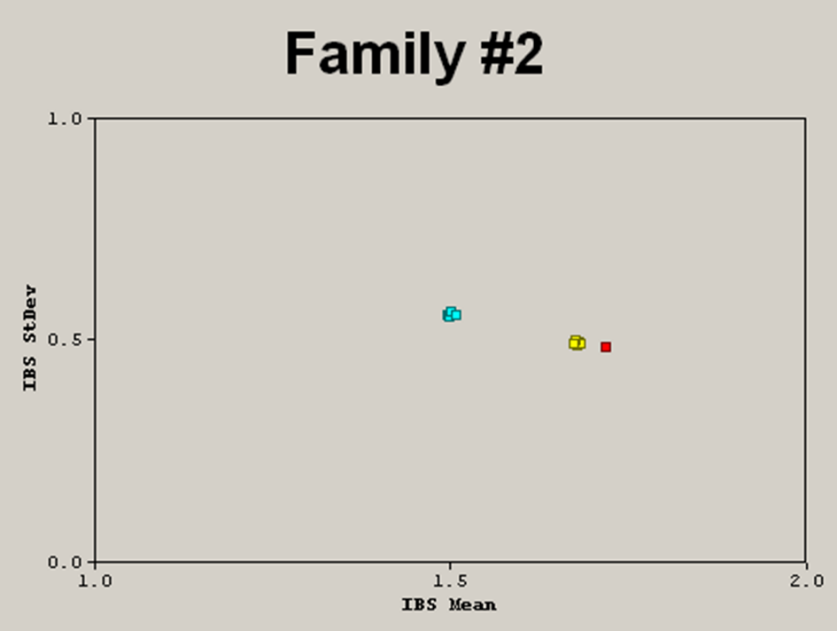


Figure 4.4; Screen shot of graphical representation of relationship (GRR) software output for family #2. Three clusters were obtained by studying the inheritance of the single nucleotide polymorphism (SNP) between the relatives. The software calculates the mean on (x-axis) and the standard deviation (stDev) on (Y-axis) for all SNP calls to create these clusters. Blue clusters indicate unrelated, red indicate sib-pairs and yellow for parent-offspring. IBS mean value for sib-pairs and parent-offspring is higher (more than 1.7) indicating inheritance of the same SNPs more than the unrelated who shows IBS mean (<1.6).

4.4.1.3 **Data preparation for LINKAGE Analysis**

PedFile format has been changed from numerical allele code to A C G T, as in family #1 (3.5.5). The data of all negative strand SNPs have been flipped (as in 3.5.5.2). Family #2 PedFile was then merged with HapMap data and a further 21 SNPs have been removed from the data due to merge error, leaving 9,558 SNPs, (equivalent to 94% of the total SNPs), to be merged with HapMap data. MDS analysis was then run on the merged data (as described in sections 3.5.5.4). Plotting the first two dimensions suggested that the family #2 belongs to European (CEU) population as does family #1 (Figure 4.5). Thus the data of other populations have been removed (as in 3.5.7.5), leaving the PedFile with family and CEU data only.

Chromosomal bp position has been changed to cM (as in 3.5.7.6), which has led to the loss of further 93 SNPs as no cM position was found. That has returned 9,465 SNPs, (equivalent to 93% of the total SNPs).

4.4.1.4 **Data preparation (for Merlin software)**

Further changes have been made to suit the data for Merlin software (as described in section 3.5.6). The allele calls in PedFile have been changed from A C G T to 1 2 3 4 respectively (as in 3.5.8.1). In addition ages for all individuals have been added in the PED file (as in 3.5.8.2) .Data file has been also created (as in 3.5.8.3). The files were then ready for Linkage analysis with 9,465 SNPs, (equivalent to 93% of the total SNPs).

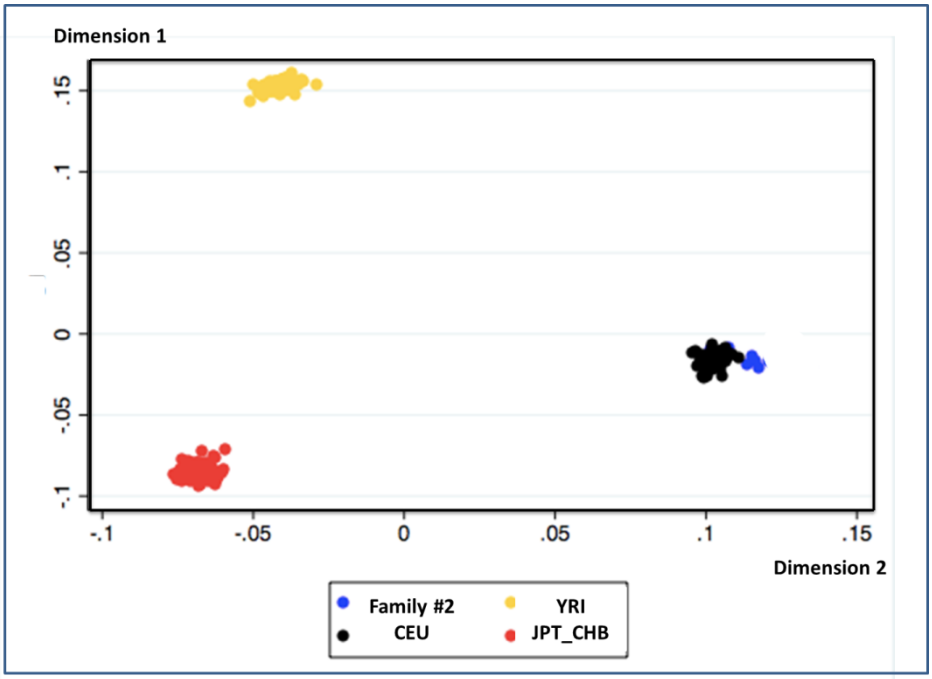


Figure 4.5; Screen shot of multidimensional scaling (MDS) output showing dimension1 (Y axis) plotted against dimension 2 (X-axis) for data from family #2 and HapMap. The scatter plot shows three clusters one for Chinese (CHB) & Japanese (JPT) population (red), one for Yoruba (YRI) population (yellow) and one for Caucasian (CEU) population (black). Family #2 data (blue) are shown to be closest to the cluster of the Caucasian (European) population.

4.5 LINKAGE Analysis by Merlin (Family #2, only)

As in family #1, some parameters in family #2 (described in family history,4.1) are not clear such as the age at onset, the mode of transmission and disease penetrance as described in family #1 (in 3.6). Thus non-parametric linkage analysis has been chosen to be the primary analysis in this study. Parametric linkage analysis "Dominant model" was chosen to be the secondary analysis and recessive model has also been performed (as described in 3.6).

Both non-parametric and parametric analyses have been performed in multi and single point (as described in sections 3.6.1 &3.6.2).

4.6 GWLA of merged data of (family #1 and family #2)

To confirm that no any information has been missed during examining the two GWLA data sets (of family #1 and family #2), both families have been analysed as one set. The data of both families have been merged and GWLA has been performed on them in one go. That will increase the power of the analysis by having larger number of individuals to be analysed.

4.6.1 Different between the SNPs in the data of family #1 and family #2

The data of family #1 contain 9,111 SNPs and of family #2 contain 9,465 SNPs with 8,935 SNPs common to both. One hundred and seventy six SNPs are in family #1 but not family #2 data, while 530 SNPs are in Family #2 but not in family #1 data.

4.6.2 Merge the data of family #1 with data of family #2

The data for all members of family #1 and family #2 have been merged after the step of preparing for Merlin software as seen in section (3.5.6). Thus both data have passed the steps of quality check and streamlining. A data of 9,641 SNPs have been merged in both family by PLINK. That is the data of all SNPs in family #2 plus the data of 176 SNPs in family #1 only. Positions have been changed to cM (as described in 3.5.5.6), and data file has been created (as described in 3.5.8.3)

4.6.3 Linkage Analysis by Merlin (on merged data of both families)

Both parametric and non-parametric analyses have been performed in multi and single point (as described in sections 3.4.1 & 3.4.2).

4.7 RESULTS (Family #2 only)

4.7.1 Non- Parametric Linkage Analysis

I have viewed the npl on chromosome 20 and all other chromosomes in general then focused on the region of interest in family #1 (described in 4.7.1.2).

4.7.1.1 **Npl for chromosome 20 and others**

The maximum (multipoint) npl LOD score for chromosome 20 is 0.41, at a region close to the P-Telomere of the chromosome, but upstream from the region of the interest in family #1 (described in 4.7.1.2).

Two SNPs have single point LODs of 0.6 & 0.15 at the same region of multipoint maximum LOD (Table 4.1) and (Figure 4.6). That has supported the region identified in multipoint Analysis.

In addition the maximum LOD scores on other chromosomes were 0.90 among chromosomes 2, 3, 4, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 21 & 22, while maximum LOD scores on chromosomes 1, 19 & X were 0.03, 0.57 & 0.29 respectively. In contrast no LOD score > 0.0 value was obtained for chromosomes 5, 6, 7 & 12. (Figure 4.7) showed the maximum and minimum LODs on each chromosome.

4.7.1.1.1 Comment on result

The small number of individuals analysed in family #2 (only 6 individuals) decreases the strength of any independent study of this family, as seen above with maximum LOD score of below 1 in all chromosomes (Figure 4.7). Therefore analysing family #2 data was only expected to add more to what has been found in family #1. Thus I aimed to focus more on the results of the analyses at the region of interest in family #1.

4.7.1.2 **The region of interest in Family #1**

The region in family #1 which has shown overlapping between haplotype and npl maximum LODs is between 8.49Mb and 12.13Mb and is 3.64 Mb long ($20p^{12.1}$–$p^{12.3}$) and comprises 10 genes (as described in 3.7.13.8.1).

4.7.1.3 **Npl of family #2 at the region of interest in family #1**

The region of interest in family #1 (described in 4.7.1.2) was focused on and the npl Multipoint analysis of family #2 at that region showed LOD score of -0.03 among the whole region (Figure 4.6).

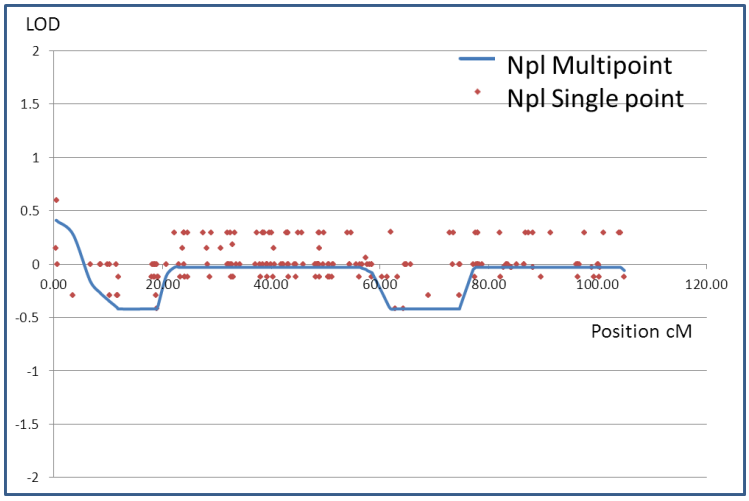


Figure 4.6; Graph showing a regional plot of chromosome 20 comparing LOD scores obtained using multi-point non-parametric linkage analysis (npl) (blue solid line) and single point npl (red dots). The maximum LOD score was 0.41, at a region close to the P-Telomere of the chromosome, but upstream from ($20p^{12.3}$-$p^{11.23}$, where maximum LOD obtained in family #1), and was supported by single-point npl in which the maximum npl LOD score was 0.6.

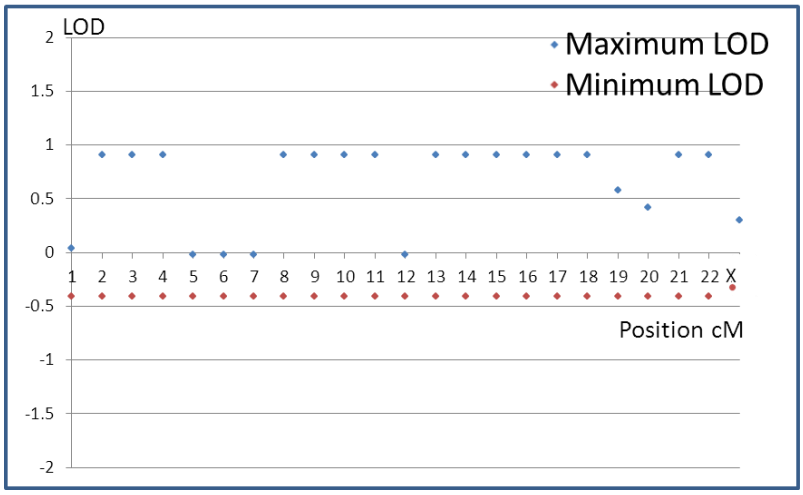


Figure 4.7; Graph showing the maximum and minimum LOD scores (of all chromosomes), obtained by non-parametric linkage analysis on family #2 data. All chromosomes showed maximum LOD score <1.00, including chromosome 20 where the maximum LOD score in family #1 was obtained

SNPs	Position cM	Multipoint	Single Point
rs1858597	0.40	0.41	0.15
rs722829	0.49	0.41	0.6
rs1342841	0.68	0.4	0
rs2317024	3.52	0.28	-0.29
rs1923876	6.69	-0.16	0
rs2013961	8.55	-0.27	0
rs1109010	8.71	-0.28	0
rs3848810	9.84	-0.33	0
rs910652	10.28	-0.35	-0.29
rs674110	10.28	-0.35	0
rs910952	11.47	-0.4	0
rs2422925	11.69	-0.41	-0.29
rs3904872	11.71	-0.41	-0.29
rs3904864	11.84	-0.42	-0.12
rs1411296	17.96	-0.42	-0.12

Table 4.1; LOD scores of part of chromosome 20 of multi vs. Single point non-parametric analyses (npl). First column with a list of single nucleotide polymorphism (SNPs), second (position cM), their position in centiMorgan (cM), third (multipoint), multipoint npl LOD score and fourth (single point), single point npl LOD scores. Two peaks of single point LODs (first 2 rows), supporting Multipoint npl (family #2 only).

4.7.2 Parametric Linkage Analysis (Dominant)

I have viewed the npl on chromosome 20 and all other chromosomes in general then focused on the region of interest in family #1 (described in 4.7.1.2).

4.7.2.1 Dominant analysis for chromosome 20 and others

The maximum dominant multipoint LOD for chromosome 20 is 0.59, at a region closed to the P-Telomere, but upstream from the region of the interest in family #1 (described in 4.7.1.2).

Two SNPs have single point LODs of 0.29 & 0.59 at the same region of multipoint high LOD (Figure 4.8). That has supported the region identified in multipoint Analysis.

In addition the maximum LOD scores on other chromosomes were 0.89 among chromosome 5, while chromosomes 2, 3, 4, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 20, 21 & 22 showed maximum LOD scores of 0.59. In addition chromosome 19 showed Maximum LODs of 0.30. In contrast no LOD > 0.0 value was obtained for chromosomes 1, 6, 7 & 12. Chromosome X showed 0.0 values of both maximum and minimum LODS. (Figure 4.9) and (Table 4.2) showed a summary of LODs on all chromosomes.

4.7.2.2 **Analysis of family #2 at the region of interest in family #1**

The region of interest in family #1 (described in 4.7.1.2) was focused on and the dominant multipoint analysis of family #2 at that region showed LOD score of -1.55 among the whole region (Figure 4.8).

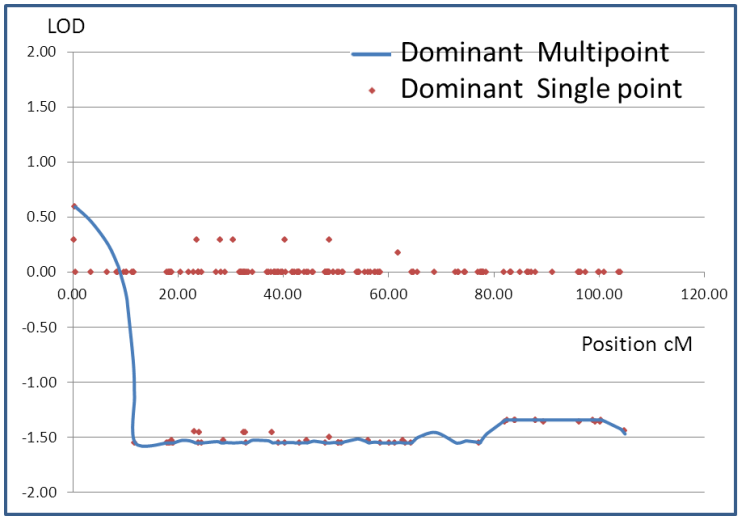


Figure 4.8; Graph showing a regional plot of chromosome 20, comparing LOD scores obtained using multi-point dominant (blue solid line) and single point dominant (red dots), linkage analysis. The maximum LOD score was 0.59, at a region close to the P-Telomere of the chromosome, but upstream from ($20p^{12.3}$-$p^{11.23}$, where maximum LOD obtained in family #1), and was supported by LOD score of 2 SNPs of single point analysis single (0.29 & 0.59).

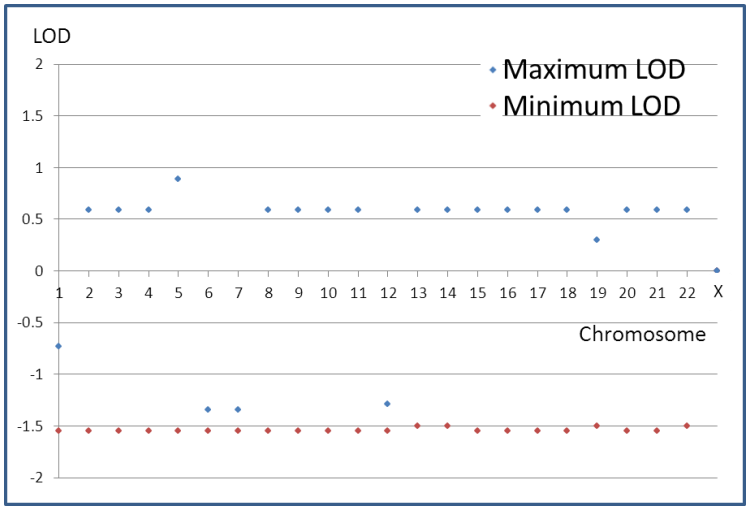


Figure 4.9; Graph showing the maximum and minimum LOD scores (of all chromosomes), obtained by dominant linkage analysis on family #2 data. All chromosomes showed maximum LOD score <1.00, including chromosome 20 where the maximum LOD score in family #1 was obtained.

Chromosome	Maximum LODs	Minimum LODs
1	-0.73	-1.55
2	0.59	-1.55
3	0.59	-1.55
4	0.59	-1.55
5	0.89	-1.55
6	-1.34	-1.55
7	-1.34	-1.55
8	0.59	-1.55
9	0.59	-1.55
10	0.59	-1.55
11	0.59	-1.55
12	-1.29	-1.55
13	0.59	-1.5
14	0.59	-1.5
15	0.59	-1.55
16	0.59	-1.55
17	0.59	-1.55
18	0.59	-1.55
19	0.3	-1.5
20	0.59	-1.55
21	0.59	-1.55
22	0.59	-1.5
X	0	0

Table 4.2; Maximum and minimum LOD scores of dominant analysis for all Chromosomes. First column (chromosome) shows chromosome ID, second maximum LOD and third minimum LOD. Maximum LOD score is 0.59 across most chromosomes

4.7.3 Parametric Linkage Analysis (recessive, multipoint)

I have viewed the npl on all chromosomes in general then focused on the region of interest in family #1, on chromosome 20, (described in 4.7.1.2).

4.7.3.1 **Recessive analysis for all chromosomes**

The maximum LOD scores on all chromosomes were; 0.28 among chromosomes 1, 5, 6, 7, 12 & 20, while chromosomes 2, 3, 4, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 21 & 22 showed maximum LODs of 0.35. In addition chromosomes 19 & X showed maximum LODs of 0.30. The minimum LOD scores were <-2.0 in all chromosomes, as summarised in (Figure 4.9).

4.7.3.2 **Analysis of family #2 at the region of interest in family #1**

The region of interest in family #1 (described in 4.7.1.2) was focused on and the recessive multipoint analysis of family #2 at that region showed LOD score of 0.28 among the whole region (Figure 4.10). Different from the results in npl and Dominant of this family, the maximum recessive multipoint LOD for chromosome 20 (LOD = 0.28), was away from the P-Telomere and closer to the region of the interest in family #1.

Twenty six SNPs have single point LODs > 0.28 at the same region of multipoint high LOD (Figure 4.10). That has supported the region identified in multipoint Analysis.

The maximum recessive LOD score across all chromosomes was <1 while minimum was ~ -2 (Figure 4.11).

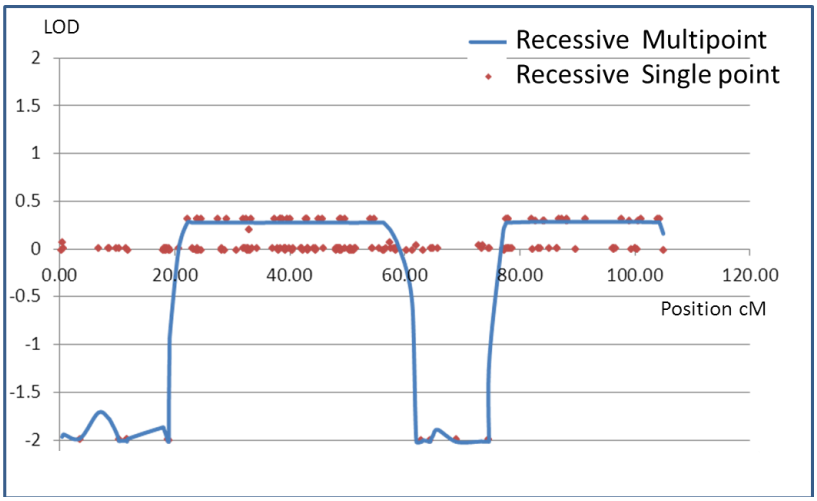


Figure 4.10; Graph showing regional plot of chromosome 20 comparing LOD scores obtained using multi-point recessive (blue solid line) and single point recessive (red dots). The maximum LOD score was 0.28 obtained at region away from the P-Telomere and closer to the region of the interest in family #1. Single point LODs were > 0.28 and have supported the Multipoint.

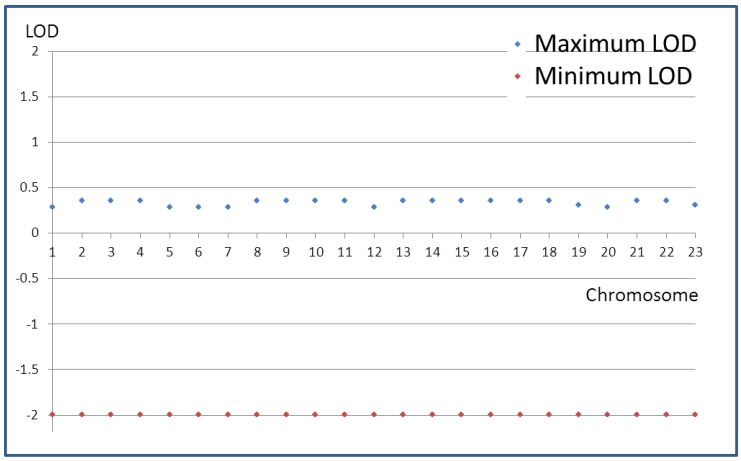


Figure 4.11; Graph showing the maximum and minimum LOD scores (of all chromosomes), obtained by recessive linkage analysis on family #2 data. All chromosomes showed maximum LOD score <1.00, including chromosome 20 where the maximum LOD score in family #1 was obtained.

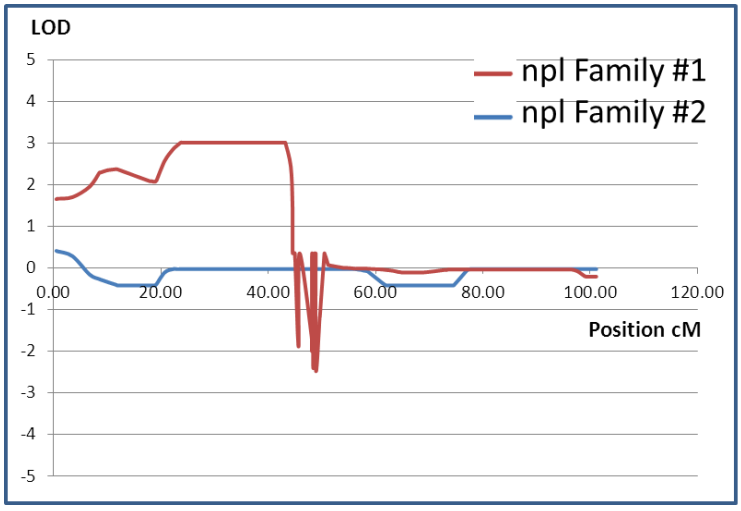


Figure 4.12: Graph showing non-parametric linkage (npl) LODs of the data of Family #1 (red line) vs. Family #2 (blue line) on chromosome 20. Family #2 data show maximum LOD score at a region close to the P-Telomere of the chromosome, but the this region is not where maximum LOD obtained in family #1. Thus family #2 data do not support the findings in family #1

4.7.4 RESULTS (merged data of family #1 & #2)

4.7.4.1 Non-parametric Linkage Analysis

I have viewed the npl on chromosome 20 and all other chromosomes in general then focused on the region of interest in family #1 (described in 4.7.1.2).

4.7.4.1.1 Npl for chromosome 20 and others

The maximum multipoint npl LOD score for chromosome 20 is 2.04, at a region closed to the P-Telomere, but upstream from the region of the interest in family #1. Another region with LOD score of 2.00 downstream from the region of interest in family #1.

Three SNPs have single point LODs >2.0 and they are; rs2327790 at 33.25 cM with LOD score of 2.7, rs1028846 at 29.01cM and rs1535487 at 41.86cM, both provide LOD score of and 2.4. Further 23 SNPs have LODs > 1.0 (out of 58 SNPs) at the same region of maximum multipoint LOD score (Figure 4.13). That has supported the region identified in multipoint analysis.

In addition the maximum LOD scores on other chromosomes were between 0.12 & 0.17 among chromosomes 1, 9, 10, 11, 13, 15 & 22, while maximum LOD scores on chromosomes 2, 3, 8 & X were 0.67, 0.59, 0.84 & 0.97 respectively. In addition maximum LOD scores between 0.21 and 0.31were obtained on chromosomes 4, 14, 16, 18 & 21. In contrast only chromosome 5 showed maximum LOD score of <0.0, while maximum LOD scores of < 0.1 were obtained for chromosomes 6, 7, & 12 & 19, as seen in (Figure 4.14).

4.7.4.1.2 Analysis at the region of interest in family #1

The region of interest in family #1 (described in 4.7.1.2) was focused on and the npl multipoint analysis of the merge data of family #1 & #2 showed LOD score of 2.0 among the whole region (Figure 4.13).

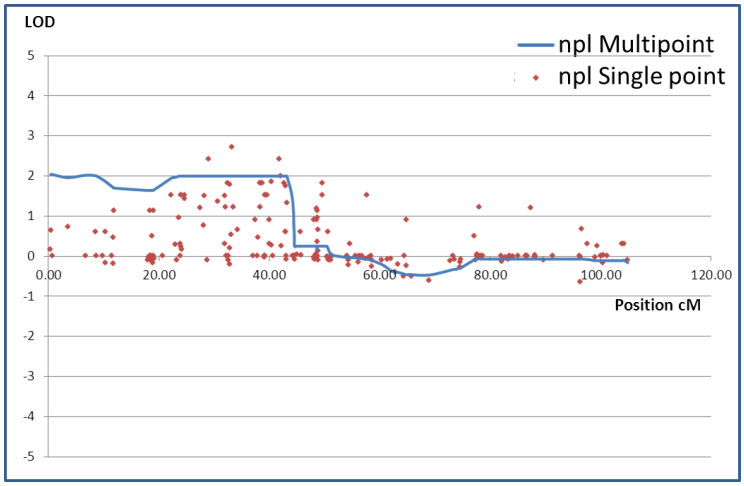


Figure 4.13; Graph showing regional plot of chromosome 20 comparing LOD scores obtained using multi-point non-parametric linkage (npl), (blue solid line) and single point npl (red dots), for the merged data of family #1 & #2. The maximum LOD score was 2.04, at a region close to the P-Telomere of the chromosome, but upstream from ($20p^{12.3}$-$p^{11.23}$, where family #1 showed maximum LOD). Another region with LOD score of 2 was obtained at a region downstream from ($20p^{12.3}$-$p^{11.23}$). Single point analysis supported multipoint with LODs >2.0 and higher (2.7 at 33.3cM and 2.4 at 29.01cM and 41.86cM).

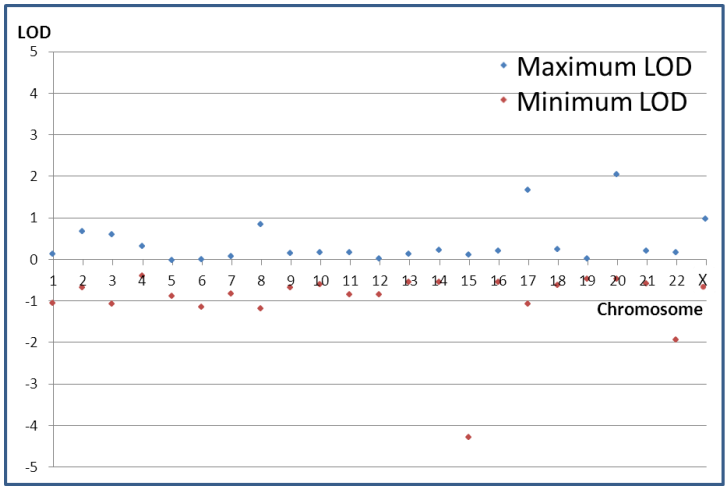


Figure 4.14; Graph showing the maximum and minimum LOD scores (of all chromosomes), obtained by non-parametric linkage analysis on merge data of family #1 and family #2. All chromosomes showed maximum LOD score <1.00, except chromosomes 17 & 20 where the maximum LOD scores are <2 and >2 respectively.

4.7.4.2 **Parametric Linkage Analysis (Dominant)**

I have observed the npl on chromosome 20 and all other chromosomes in general then focused on the region of interest in family #1 (described in 4.7.1.2).

4.7.4.2.1 Dominant analysis for chromosome 20 and others

The maximum dominant (multipoint) LOD score for chromosome 20 is 2.3, at a region closed to the P-Telomere, but upstream the region of the interest in family #1.

Sixteen SNPs have single point LODs > 1.1 (one of which was 1.9), and further 7 SNPs have LODs 6 (out of 58 SNPs) at the same region of multipoint high LOD (Figure 4.13). That has supported the region identified in multipoint Analysis.

In addition only 7 chromosomes showed maximum LODs > 0.0. These were 2, 3, 4, 16, 17, 20 & X, with chromosome 20 the highest in all with LOD 2.3. (Figure 4.16) Show the maximum and minimum Dominant LODs on each chromosome.

4.7.4.2.2 Analysis at the region of interest in family #1

I have focused more on the region of interest in family #1 (described in 4.7.1.2), and the dominant multipoint analysis of the merged data of family #1 & #2, showed maximum LOD scores of 0.5 among the whole region as seen in (Figure 4.15).

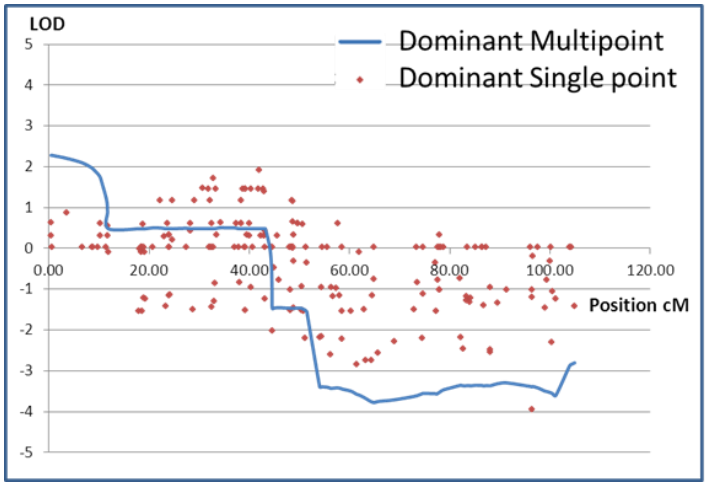


Figure 4.15; Graph showing regional plot of chromosome 20 comparing LOD scores obtained using multi-point dominant (blue solid line) and single point dominant (red dots), for the merged data of family #1 & #2. The maximum LOD score was 2.3, at a region close to the P-Telomere of the chromosome, but upstream from ($20p^{12.3}$-$p^{11.23}$, where maximum LOD obtained in family #1), and was supported by single-point npl in which the maximum npl LOD score was 1.9 & below.

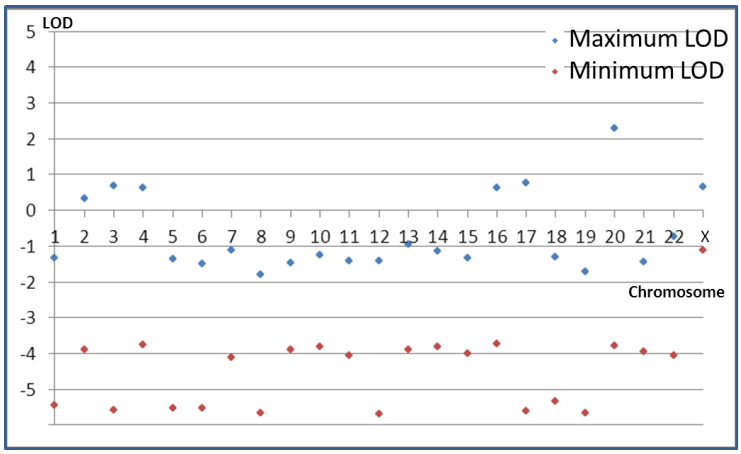


Figure 4.16; Graph showing the maximum and minimum LOD scores (of all chromosomes), obtained by dominant linkage analysis on merge data of family #1 and family #2. All chromosomes showed maximum LOD score <1.00, except chromosome 20 where the maximum LOD scores is 2.3 which is higher than what obtained by the data of family #1 only

4.8 Conclusion

In view of the family history and phenotypic similarity in family #1 and family #2, I hypothesized that genetic variants in both families would be similar and thus analysed the only available DNA sample of family #2 (from 6 individuals). By analysing these few individuals, the main target was to look at the region of interest in family #1 (between $20p^{12.3}$ and $20p^{11.23}$). My intention was to add the LOD scores for the two families; however the analysis of family #2 at that region did not support those for family #1. In fact the findings had a negative impact since adding the LOD scores reduced that obtained in family #1 alone. Examining the two GWLA data sets (of family #1 and family #2) across chromosomes on other than the target region have shown no significant findings with maximum LODs of 0.90 in npl, 0.89 in Dominant analysis and < 0.5 in recessive analyses.

Specifically, performing GWLA on the merged data of both families has decreased the maximum npl LODs (from 3.01 in family #1 only) to 2.0, at that region of interest. However, single point npl has shown a maximum LOD score of 2.7 at rs2327790 (33.25 cM), and 2.4 at 2 NSPs; rs1028846 (29.01) and rs1535487 (41.86). The gene at the region with LOD score of 2.7 (rs2327790 at 33.25 cM) is TASP1 (Taspase Threonine Aspartase, 1), encodes a protein required for the maintenance of HOX gene expression. The gene at the region with LOD score of 2.4 (rs1028846 at 29.01cM) is BTBD3, which is the last gene in the list of the 10 genes in family #1. The second gene in the region with LOD score of 2.4 (rs1535487 at 41.86 cM) is SLC24A3, described as solute carrier family 24 and encodes sodium/calcium transporter, a member of the family of K(+)-dependent Na(+)/Ca(2+) exchangers [268]. Same findings obtained in dominant analysis on chromosome 20, as the maximum LOD for family #1 only was 2.03, but for the data of both families it was 2.3, at a region closed to the P-Telomere, but upstream the region of the interest in family #1.

4.9 Discussion (chapter 3 & 4)

The Study in this thesis is the first study to my knowledge that has used SNPs and performed GWLA on a single family with MNG progressing to Thyroid cancer or Familial Non Medullary Thyroid Cancer (FNMTC). A similar study [269] used GeneChip Human Mapping 10K Array Xba 131, and identified a locus on 8p23.1-p22 in a large Portuguese kindred with a more classic FNMTC pedigree, but when the authors investigated additional families the locus was not found to be linked. This is a common finding in studies of familial MNG and/or PTC, and indeed was my experience when testing two families with apparently similar disease (clinical and histological features) but which did not share predisposing genetic variants.

Although several studies on thyroid cancer have performed linkage analyses on a single family, they have used microsatellite markers rather than SNPs. Canzian et al [85] studied a French pedigree with three generations and identified a locus on chromosome 19p13 (FNMTC 1). FNMTC 2 was localized to 2q21 in a large Tasmanian family [87]. Malchoff et al [86] studied an American family and identified a locus on 1q21 (FNMTC 3).

Identifying a region with 10 genes encouraged me to do some in silico analysis looking for any relationship between these genes and thyroid cancer or MNG or GWLA, but I did not find any link. Furthermore, I have searched for estrogen response elements in the promoters of the various genes in the areas of high LOD score and several were identified.

There has been increasing interest in exploration using linkage analyses for investigating cancers in general [270]. Some have focused on a specific region on a chromosome (related to thyroid cancer) [271], others have performed GWAS on thyroid cancer but on a group of families [91]. Some are more focused on specific pathways and pathological data such as Auto-Immune Thyroid Disease [272]. Many studies focused more on cancer in general, but some on thyroid cancer in particular. Suh and colleagues [273] have performed linkage analysis on 38 families with FNMTC with 113 affected and identified susceptibility loci on chromosomes 6q22 and 1q2. As mentioned in chapter 1, recently a study of GWAS using SNPs was performed on the Icelandic population [91] and has shown association of two common variants, located on 9q22.33 and 14q13.3 with thyroid cancer.

As mentioned in the summary of my MPhil (1.7), I have studied the known loci of MNG on chromosomes 14q, Xp and 3q, by linkage analysis using microsatellite markers, and also for the known loci for FNMTC, on chromosomes 19p, 2q and 1q. After GWLA using SNPs, I have re-checked the same region and have not found any linkage in family #1 data, as shown in (Figure 3.6) with the maximum and the minimum LOD scores of npl across all chromosomes. The same observation was made with the data of family #1 and #2 as shown in (Figure 4.14). However, npl single point for family #1 has shown a LOD score of 1.5 at $14q^{12}$ (rs1952966 at 31.7 cM), a similar LOD score was obtained at $1p^{36.21}$ (rs763821 at 34.3 cM), while dominant single point produced LOD scores of 0.6 and 1.4 for the same two SNP markers respectively.

Chapter 5 Studies on the region of interest in family #1; (CNV)

All subsequent studies (other than chapter 4) were performed only on family #1. In this chapter further investigations have been made on the region of interest on chromosome 20 ($20p^{12.3}$-$20p^{11.23}$), which showed the maximum LOD scores in GWLA (as described in 3.7.1). The region with high LOD score of 3.01 spans 19.5 cM and is large when compared with most GWLA reported in the literature [85-87, 269, 274, 275], I hypothesised that chromosomal copy number variants (CNV) such as deletion or duplication might provide an explanation.

In preliminary studies, using the limited CNV data which can be extracted from the 10K chip, some evidence for CNV in the region with the highest LOD score was apparent. Thus specific analysis for CNV has been performed using Illumina Human 660-Quad BeadChip. The chip contains 660,000 Markers (will be described in 5.1.4).

5.1 Introduction

5.1.1 Type of chromosomal variation

Human cells have 23 pairs of chromosomes with two copies of each, giving a total of 46 per cell. Several variations can be detected in abnormal cases. These include deletion of both copies (homozygous), deletion of one copy (hemizygous), (also known as loss of heterozygosity (LOH)), deletion of one copy and duplication in the available copy, deletion of one copy and several duplications of the available copy (Figure 5.1) (reviewed in [276]).

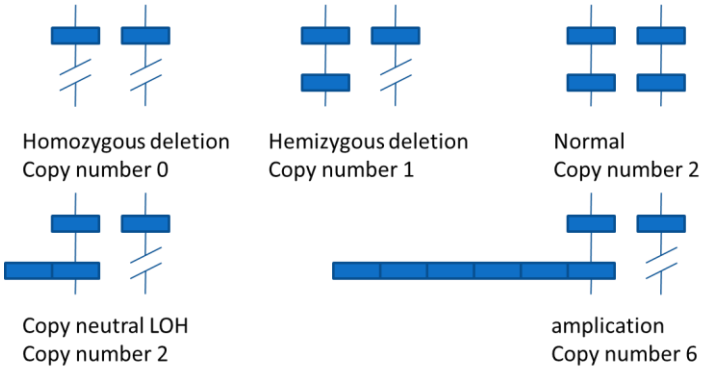


Figure 5.1; Schematic representation of chromosome pairs illustrating changes in copy number of distinct regions detected by chromosome copy number variation studies. A full description is in the text. Adopted from **http://link.springer.com/chapter/10.1007%2F978-3-540-71681-5_10#page-1**

5.1.2 Copy Number Variation

The variability of human populations is determined in a large part by environmental factors and/or genetic information. It has been confirmed that genetic variations are not caused by SNPs only, but there are also polymorphisms that extend over hundreds of thousands of DNA base pairs across the whole genome. Such alternations called copy number variation (CNV) often include genes and other functional genetic elements [277]. CNV refers to alterations in numbers of copies in segments greater than 1kb, (up to 1 mega base). Changes between 1 and 1000 bp are referred to as insertions or deletions, depending on whether nucleotides are added or lost respectively. Changes in a whole-chromosome (34 ~230 Mbp), whether gains or losses, are known as aneuploidies [278]. Depending on their frequency among population, CNVs can be divided according to their frequency. Those occurring in >1% of the population (usually not pathogenic) [279] and those in <1%, which can be considered pathogenic, and usually not observed in the unaffected siblings or ethnic matching controls [280]. On the other hand there is a counterproductive relation between sizes and frequencies of variants [280]. CNV analysis usually refers to the process of analysing data to test the variation of chromosomal copy numbers in DNA samples of patients vs. their unaffected relatives. Such analysis helps detect chromosomal abnormalities

that may cause or increase risks of any particular disorder. Data analysis for an array-based DNA copy number test can be very challenging because of the very high volume of the output data of any array platform [281].

In 2001 the human genome sequencing project was completed and provided a road map which has helped in discovering more CNVs. Around 5% of the genetic code was found to consist of redundant segments represented many times in different locations across the genome. These segments were shown to predispose to copy number variation and have been named as "segmental duplications" (SDs). SDs size varies from 1 kb to hundreds and have high sequence identity (>99%). They may include genes or noncoding regions but cannot be distinguished easily [282]. Since 2004, researches have revealed CNVs as a major source of human genetic variation [283-285]. A study has used large-scale copy-number variations (LCVs), with gains or losses involving regions up to 2 Mb in size. It demonstrated coincidence of markers with comparative gains or losses among samples having chromosomal imbalances [284]. Other studies have identified new sites of structural variation and named the insertions, deletions and inversions found as intermediate-sized structural variants (ISVs). The size of ISVs ranges from 6.3 to 24.7 kb [285, 286]. Several studies [287-289] demonstrated that small-scale CNVs (<10 kb) are common in the human genome. The unique level of genetic diversity identified by CNVs has led to a new stage of discovering sources of phenotypic variation, human development and disease susceptibility.

Considerable evidence indicates that CNV can have a significant biological impact and CNVs have been reported to play a role in severe developmental syndromes and familial diseases. CNVs can disturb gene expression within and flanking the CNV and CNVs can confer susceptibility to infectious and complex diseases [290]. Even small changes in DNA copy number are shown to be associated with multiple phenotypes, related to immune or environmental response. Many studies have shown that these changes in DNA copy number are associated with different types of disorders including Asthma [291], heart [292] and psychiatric defects [293], and at least 15% of human neurodevelopmental disorders are due to rare and large copy number changes [280]

5.1.3 Copy Number Variation History

The traditional method for detecting CNV was by cell culture and karyotyping analysis [294]. In this method the chromosomes are stained (during cell culturing), with certain dyes that show a pattern of dark and light bands. The bands are specific for each chromosome and thus lead to the identification of each of the 24 types of chromosomes (22 autosomal plus X and Y)[295]. Fluorescence in situ hybridization (FISH) was another traditional method for CNV. Further improvements in this method made FISH applicable for use in Comparative Genomic Hybridization (CGH), which can detect chromosomal copy number changes without the need for cell culturing [296]. That is by labelling the abnormal DNA (from tumour tissues) with a green fluorescent probe and reference DNA (with normal karyotypes) with red, and hybridising to normal human metaphase preparations (reviewed in [296]). As mentioned below, CNV studies have been performed on thyroid. For CGH labelling, cell lines of known BRAF mutation, can be used as positive control [297]. Consequently it was possible to identify regions of relative loss and gain in the test sample, by assuming the reference DNA to be diploid in copy number. The green to red fluorescence ratio at any locus represents loss or gain of genetic material in the abnormal DNA at that specific locus. The use of CGH has been applied mainly in the analysis of genetic changes in DNA extracted from solid tumours, and has greatly increased knowledge of chromosomal rearrangements in tumour biology [298]. Several studies on thyroid have reported the use of CGH, to detect chromosomal instability in PTC. Hemmer et al. [299] reported that DNA copy number changes were rare (3 of 26, 12%) in PTC. In contrast, Singh et al. [300] reported genetic abnormalities in 10 of 21 PTC cases (48%), including losses at 1p and 9q and complete loss of chromosomes 17, 19 and 22, and gains at chromosome 4 and at 5q, 6q and 13q. Bauer et al. [301] reported gain of chromosome 20 and loss of chromosome 13 in 4 of 15 PTC samples.

CGH was used first in 1992 [294] and despite its value, is hampered by low resolution (~20 Mb). The resources generated for the public-domain Human Genome Project led to improvements in the resolution (1 Mb) of CGH [302, 303]. A particularly important study was the HapMap project, which catalogued SNPs in four of the major ethnic human populations [217]. With these resources it became possible to replace metaphase chromosomes for CGH with arrays of clones accurately mapped onto the human genome. This increased CGH resolution and is known as matrix-CGH [304], and then as array-CGH,

which has increased resolution to < 100 kb [305] making it a standard method for clinical cytogenetic investigations [306]. Many studies [307-311] have analysed CNV using array-CGH and have confirmed the importance of such analyses in normal human variation. Many further improvements in array technology have been made aiming to increase numbers of features (spots) and decrease the length of DNA sequences as hybridization targets [312]. Cloned arrays were first to be used, in which clones such as BAC (Bacterial Artificial Chromosomes), YAC (Yeast Artificial Chromosomes) or PAC (P1-derived Artificial Chromosomes) have been used [312]. Array CGH uses oligonucleotides to provide higher resolution (than old CGH arrays with clones) [313]. Four different oligonucleotide platforms have been used in different studies; Affymetrix commercial CGH platform was the first, and contains short probes of 25 oligonucleotides [314]. The second CGH platform was introduced by Agilent Technologies [315, 316]. The Agilent platform has already been proven highly valuable in a lung cancer study [317]. A third CGH oligonucleotide platform offered commercially is by NimbleGen, in which the CGH oligonucleotides are designed to be isothermal and vary between 45 and 85 bp in length [318]. A fourth CGH platform is non-commercial and makes use of oligonucleotide libraries that are spotted as elements on the arrays [313, 319].

High-throughput array technologies for SNP genotyping from commercial companies such as Affymetrix and Illumina, have received considerable attention as a source to identify human variation [312]. Illumina Bead Array platforms represent a SNP by a number of beads for each allele. Several unique bead types containing different probe sequences are represented in each array, with an average 30-fold redundancy of each bead type. Independent of the array format, each bead in every array contains hundreds of thousands of covalently attached oligonucleotide probes. Affymetrix arrays represent a SNP by probes for each allele. Depending on the DNA sequence they target, the beads and or probes have different affinities, and thus produce signals of different strengths [320]. Affymetrix and Illumina arrays have started focusing more on intensity information, to determine genomic copy number, beside their original function of allelic content across the genome. Thus these arrays can serve a dual role for SNP- and CNV-based association studies.

Furthermore several developments have improved the output of the oligonucleotide array, signal-to-noise ratio was improved by digesting the DNA by restriction enzymes then ligation with adapters. Hybridization complexity was reduced by amplifying the smaller fragments of DNA by universal primers [312]. The reduction of complexity in hybridization (by PCR amplification of smaller restriction fragments, ~1.2 kb), brings the possibility of amplification bias of different regions of the genome [306]. Noise reductions can be achieved by taking length and GC content of the probes into account [321].

Although most CNV markers are SNPs, non-polymorphic (NP) markers (also known as CNV markers), have also been used in probes for better coverage of the gaps that are not covered by SNPs. That has helped in avoiding the variable resolution of the array across the genome and has increased the limit of CNV detection to >10kb (reviewed in [312]). Nomenclature of the markers allows them to be distinguished, SNP markers start with *rs* in almost all arrays, while NP markers start with *cnvi* in some arrays.

Although changes in the copy number start and end at a specific base pair, only simple duplications and/or deletions can be detected by the experimental platforms, as they measure the resolution rather than the size of the CNV. That includes high-density genotyping platforms such as SNP arrays with NP markers and ultrahigh-density whole genome tiling arrays. These technical platforms cannot detect complex structural variations, such as inversions or translocations [276].

5.1.4 Illumina Bead Array Generation of SNP genotyping array

Illumina Bead Array is very popular and has several types depending on the number of markers used varying from 100K to 650k and a 1 Million chip has also been released (human 1M). Furthermore Illumina Omni family of microarrays provides chips with over 4.3 million fixed markers per sample (Omni 5), with the ability to add 500K extra custom markers [322]. Affymetrix SNP array is another option with the number of markers varying from 10K which was released in 2003, 100K and 500K are also available. Similar to Illumina, a chip with 1 Million SNPs has also been released by Affymetrix (Genome-wide Human SNP 6.0) (reviewed in [320]).

Illumina Human660W-Quad BeadChip contains over 657,000 genetic markers for the whole-genome genotyping and CNV analysis. It builds from 550,000 SNPs markers from the Human Hap550 BeadChip, plus an additional (more than) 100,000 markers on regions known to be associated with copy number variations. Thus the assay utilizes Infinium platform, "an extremely high throughout genotyping system that relies on direct hybridization of genomic targets to array-bound sequences". Locus discrimination or copy number determination in the Infinium is provided by a combination of sequence-specific hybridization capture and array-based, single-base primer extension. If there is (sample/probe) perfect match, extension occurs and signal is generated, otherwise extension does not occur and no signal is generated. Signal amplification of the combined label further improves the overall signal-to-noise ratio of the assay. The Infinium platform was used first in December 2008(reviewed in [323]).

A study comparing different arrays [324] has shown that detection of variants between 1kb to 50 kb in size for Illumina 660W platform was >95%. On the other hand, studying array resolution and distribution of probes, showed that Illumina 660W array specifically targets CNVs and has a very uniform distribution of number of probes per CNV call. The study has also shown that the ratio of copy number gains to losses of certain arrays to be very biased toward detection of deletions, while Illumina 660W array has shown highest ratio of deletions to duplications. In terms of reproducibility, Illumina 660W was similar to the Affymetrix 6.0 and the other two Illumina arrays (1M and Omni) with a value around 80%, when analysed with the best performing algorithms. Furthermore the ability to detect variants by Illumina 660W and Illumina Omni was shown to be higher than Affymetrix, when used in the 1000 Genomes project (reviewed in [324]).

5.1.5 BeadStudio software and calling and normalising

The main function of this software is normalisation of array intensities in order to make comparisons across arrays and background correction for correct interpretation of the data (i.e. to reduce the number of false and missing calls) [320]. In the Illumina SNP genotyping assay two probes are used to detect the presence of the two different alleles at each SNP. There is one intensity channel for each of the fluorescent dyes associated with the two alleles of the SNP. The alleles measured by the X channel (Cy5 dye) are called the A alleles, whereas the alleles measured by the Y channel (Cy3 dye) are called the B alleles [325]. Two

values are automatically generated from Illumina BeadStudio software from the X and Y values, the Log R Ratio and the B allele frequency. The Log R Ratio (LRR) is a measure proportional to the total signal intensity (and therefore the copy number). The value log R ratio is the log (base 2) ratio of the observed normalized R value for the SNP divided by the expected normalized R value for the SNPs theta value. Expected R is calculated from the values theta and R, where R is the intensity of dye-labelled molecules that have hybridized to the beads on the array and theta is the ratio of signal at each polymorphism for beads recognizing an A allele to beads recognizing a B allele, (and is different from the theta "Recombination Fraction" described in3.1.2) [326]. Therefore, the ratio of observed R to expected R in any individual at any SNP gives an indirect measure of the binding efficiency of detected alleles for each polymorphism and thus of genomic copy number. LRR above 0 is indicative of an increase in copy number. If it is between 0 and 1 then it is heterozygous duplication, (of one copy), if it is above 1 so it is homozygous duplication, (of 2 copies). Values below 0 suggest a deletion, if LRR is below 0 but above -1 then it is heterozygous deletion, (of one copy), if it is below -1 so it is homozygous deletion, (of 2 copies). LRR is the base-2 log-ratio of the sample's direct intensity (R) at the given locus, versus a reference sample [327]. B Allele Frequency (BAF) is a measure of the relative contribution to the total signal from the B allele (and therefore reflects the allelic composition). In other words BAF is the "theta" value for an individual SNP corrected for cluster position, this gives an estimate of the proportion of times an individual allele at each polymorphism was called A or B; thus an individual homozygous for the B allele (BB) would have a score close to 1, an individual homozygous for the A allele (AA) would have a score close to 0 and a heterozygous individual (AB) have a score of ~0.5 (Figure 5.2) [325],[1]. BeadStudio software calculates the LRR and BAF from the X and Y values generated from Illumina as follow;

$$\text{Log R Ratio} = \log_2 (X + Y)$$

$$\text{B Allele Frequency} = Y/(X + Y).$$

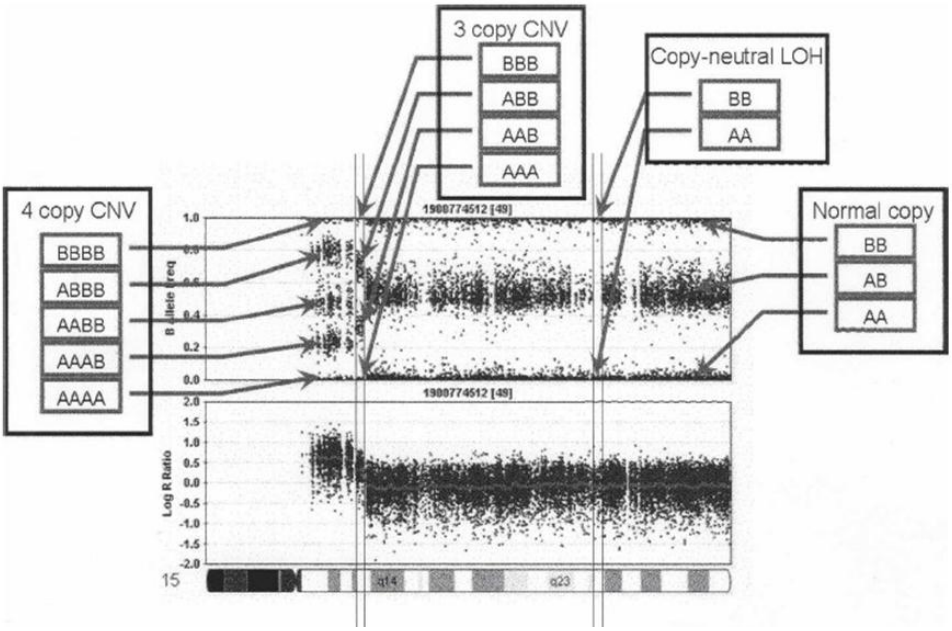


Figure 5.2; Screen shot of an example for Log R Ration (LRR), lower panel and B Allele Frequency (BAF), upper panel, in CNV (duplication) cases. BAF has three values; high for BB call, middle for AB and low for AA. Normal LRR value is close to zero and high or low values indicate duplication or deletion respectively. A full description is in the text. Adopted from [2].

5.1.6 PennCNV Software and CNV analysis

The signal intensity of genotyping data (pre-computed Log R Ratio and B Allele Frequency data), in BeadStudio project file format, can be subsequently analysed by "PennCNV software". It is an integrated hidden Markov model (HMM) designed software for high-resolution CNV detection in Illumina high-density SNP genotyping data. PennCNV software can be used to detect CNV through algorithm [2]. The PennCNV algorithm uses the two measures of signal intensity at each SNP, the LRR and the BAF (Figure 5.3). While BeadStudio software calls the genotypes and normalises the total signal intensity of the two alleles (A & B) of SNPs, PennCNV software analyses the CNV for each region across the genome (Figure 5.2). To derive the LRR and BAF values, a signal pre-processing procedure is necessary for SNP genotyping platforms. In the case of NP markers, the R-value is

calculated as the signal intensity rather than the sum of two alleles. Also BAF values cannot be derived for NP markers. Thus these markers are shown to be non-informative in the likelihood calculation, but used for CNV only. PennCNV software summarises the variants in each chromosome at the end of the analysis (reviewed in [1]).

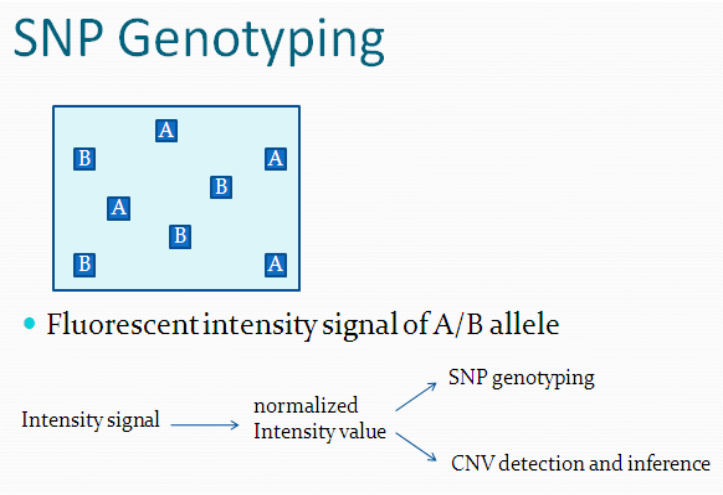


Figure 5.3: Diagram showing that the intensity signal created by the match between the probes and the single nucleotide polymorphisms (SNPs) in patient DNA, can provide genotyping of these SNPs as well as detecting any variation in the copy number of that region. A full description is in the text. Adopted from [2].

5.1.7 DNA quantification & PicoGreen

For optimal Chip results, it is highly critical that the DNA quantity is accurate. DNA quantification plays an important role is success or failure of a library preparation. Higher DNA concentration may lead to increased background signal and lower resolution, whereas low DNA concentration may result in reduced signals. The ordinary methods of DNA quantification are UV-spec based and measure any nucleotides present in the sample

including RNA, double-stranded DNA (dsDNA), single-stranded DNA (ssDNA), and free nucleotides, which can give an inaccurate measurement of gDNA. For highly sensitive protocols such as arrays analysis, fluorometric-based methods (such as PicoGreen) are recommended as they provide accurate quantification for dsDNA [328].

PicoGreen® (PG) is a reagent used for quantitation of dsDNA in molecular biology, due to the dramatic increase in its fluorescent emission upon interaction with dsDNA. It is an ultra-sensitive fluorescent nucleic acid stain that can interact with nucleic acids and help in its biophysical studies. When binding DNA, PG fluorescence increases >1000-fold in proportion to the quantity of DNA present. An interesting feature of PG is its ability to strongly bind not only to highly polymeric DNA but also to short duplexes <20 bp, likewise exhibiting a sensitivity in the picogram range (reviewed in [329]).

5.2 Aim

To identify any variation in the chromosomal copy number (loss or gain) in the region with maximum LOD score identified by GWLA. Any CNV detected, will be investigated further, by standard PCR and Sanger sequencing, on all family members, to determine whether it is found in affected and/or unaffected members.

This chapter will explain the CNV analysis performed using Human 660-Quad BeadChip, the analysis of the data obtained and the further investigations required after data analysis.

Since GWLA has identified a specific region with significant LODS, CNV in this study will not follow the more usual approach to data analysis so rather than comparing the CNV output between affected and non-affected members, the analysis will be performed on a DNA sample of one member of family #1, the index patient (IV-6). Thus the interpretation will focus on the region of interest ($20p^{12.3}$-$20p^{11.23}$) and not study the whole genome for CNV.

After Quantitation of the DNA sample by PicoGreen® (described in 5.1.7), The DNA sample was transferred to a Colleague in Psychological Medicine, who has considerable experience in CNV analysis. The sample was processed via several steps (as summarised in Figure 5.4), which were followed by data collection, pre-processing and signal intensity extraction by

Illumina BeadStudio software (described in 5.1.5). PennCNV software (described in 0), was used next for CNV detection and annotation (as summarised in Figure 5.5).

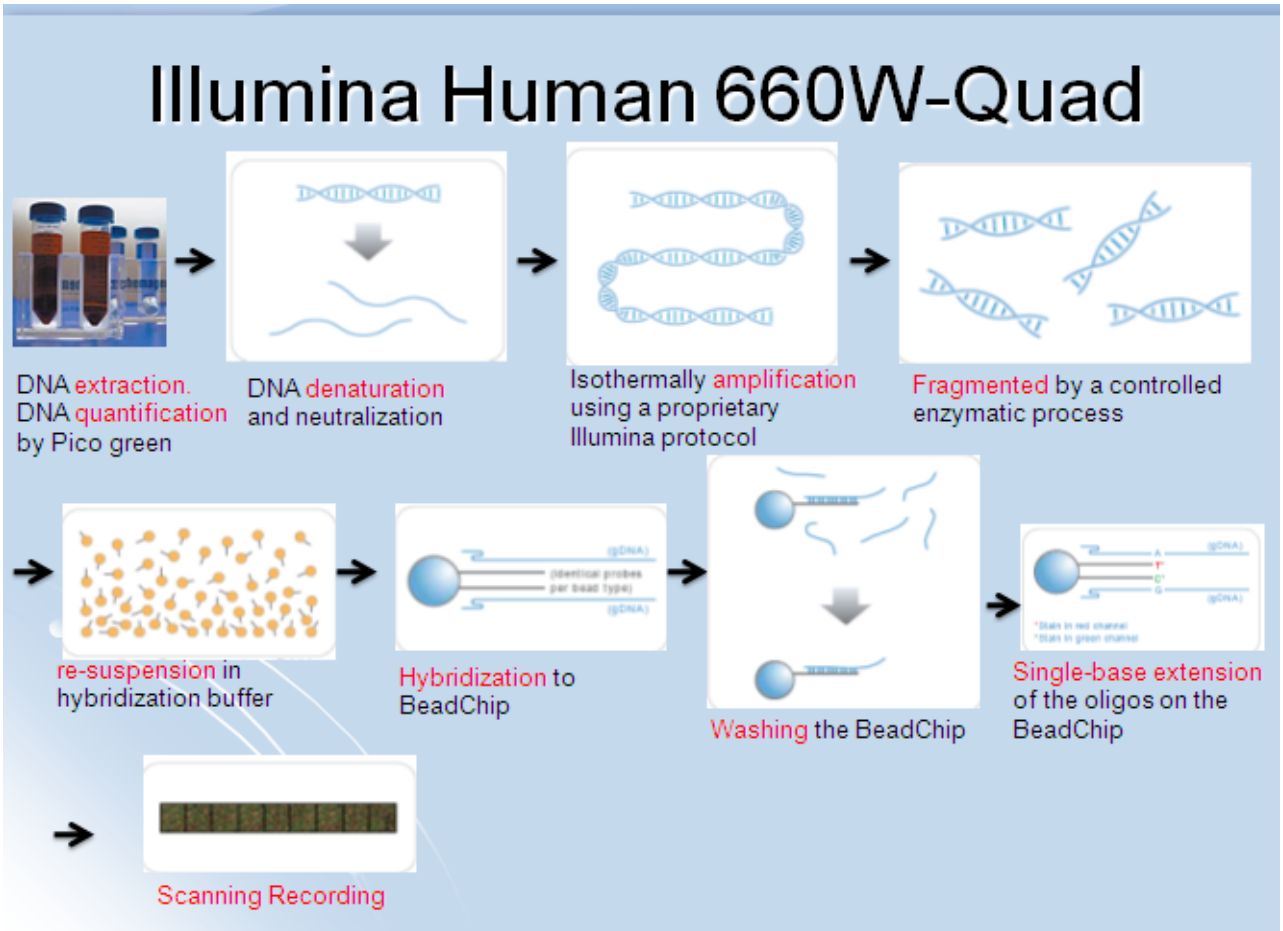


Figure 5.4: Sample processing steps in Illumina Human 660W-Quad kit. DNA sample underwent denaturation and isothermal amplification. The amplified product was then fragmented by a controlled enzymatic process and re-suspended in hybridization buffer. This was followed by Hybridization to BeadChip then washing. Single-base extension of the oligos on the BeadChip was then performed, followed by scanning and recording the output. The figure is adopted from Illumina Human 660W-Quad Protocol Manual.

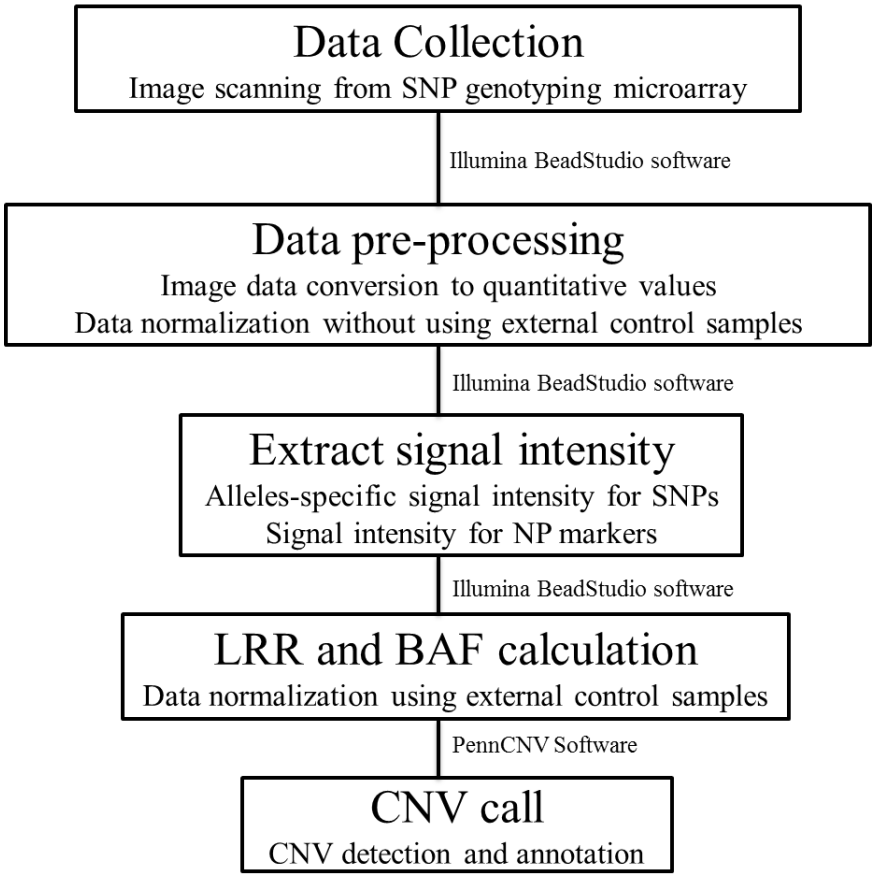


Figure 5.5: Flow diagram summarising the steps after Scanning and Recording the output in Illumina chip until performing CNV by PennCNV software. Adopted from in [1].

5.3 Data Analysis

The data have been analysed by PennCNV software, which has summarised the variations among each chromosome (Table 5.2, as an example). Several variations have been detected in each chromosome (Figure 5.1). Since chromosome 20 contains the region of interest of GWLA, only that chromosome and particularly that region have been investigated further.

5.3.1 Variation in Chromosome 20

A total of 9 variants have been detected in Chromosome 20 (Table 5.2), three of which were in the P arm of the chromosome ($20P^{12.3}$, $20P^{12.1}$ & $20P^{11.21}$). One variant out of these three was in the region of interest, at $P^{12.3}$ (8358121-8359033 bp) as seen in (Table 5.2).

The variant was a deletion of ~903 bp of one copy and was detected by 14 markers from the Illumina Chip (Figure 5.6) & (Table 5.3), four of which were SNP markers, while the other 10 were NP markers (described in 5.1.3). The one copy deletion is clearly observed by the LRR ratio of the 14 markers as almost all of them obtained LRR below 0 but above -1, which indicate a one copy deletion (as described in 5.1.5)

Obtainable copies / Chromosome	0 copy	1 copy	3 copies	4 copies
1	10	15	6	-
2	7	15	7	1
3	9	15	-	-
4	11	13	2	-
5	4	13	6	-
6	6	5	2	-
7	3	12	3	-
8	7	6	-	-
9	12	11	3	-
10	9	9	5	-
11	5	8	2	-
12	-	13	3	-
13	4	7	-	-
14	2	9	3	-
15	2	10	2	-
16	1	7	5	-
17	2	5	3	-
18	1	6	-	-
19	1	10	1	-
20	1	8	-	-
21	2	1	-	-
22	3	4	-	-

Table 5.1: Summary of the variations detected by PennCNV software on all chromosomes. First column is chromosome number. Column 2 is the number of deletions of both copies (0 copies) and column 3 is deletions of one copy (1 copy). While column 4 is duplication of one copy (3 copies) and column 5 is duplications in 2 copied (4 copies). First row represents copy number variation (CNV) in chromosome 1

and shows that a deletion of both copies has been detected in 10 different regions. One copy deletion has been detected in 15 regions and a duplication of one copy in 6 regions.

Chromosome and location	No of markers	Length of Variation (bp)	No of copies	Start Marker	End Marker
chr20:8358121-8359023	14	903	1	cnvi0102516	rs6055812
chr20:15249783-15251701	18	1,919	0	cnvi0056722	rs6110571
chr20:24359854-24370253	15	10,400	1	cnvi0110453	rs6036792
chr20:50085800-50087807	18	2,008	1	rs6068046	rs6068047
chr20:61236393-61237260	15	868	1	cnvi0059879	cnvi0059893
chr20:61708979-61711285	15	2,307	1	rs28664002	cnvi0053235
chr20:62066683-62068138	19	1,456	1	cnvi0050179	rs817329
chr20:62195699-62197068	14	1,370	1	cnvi0106384	cnvi0106397
chr20:62233717-62256817	21	23,101	1	cnvi0054409	rs4809258

Table 5.2: List of 9 variants on Chromosome 20, detected using PennCNV software. First column reports the location of each variant, column 2 the number of markers detecting that variant. Column 3 provides the length of each variant, column 4 shows number of copies identified. The last two columns (5 & 6) show the markers flanking each variant. The first variant is (loss of 1 copy) in chromosome 20 at 8358121-8359033 bp, detected by 14 markers, its length is 903 and is in the region of interest in family #1.

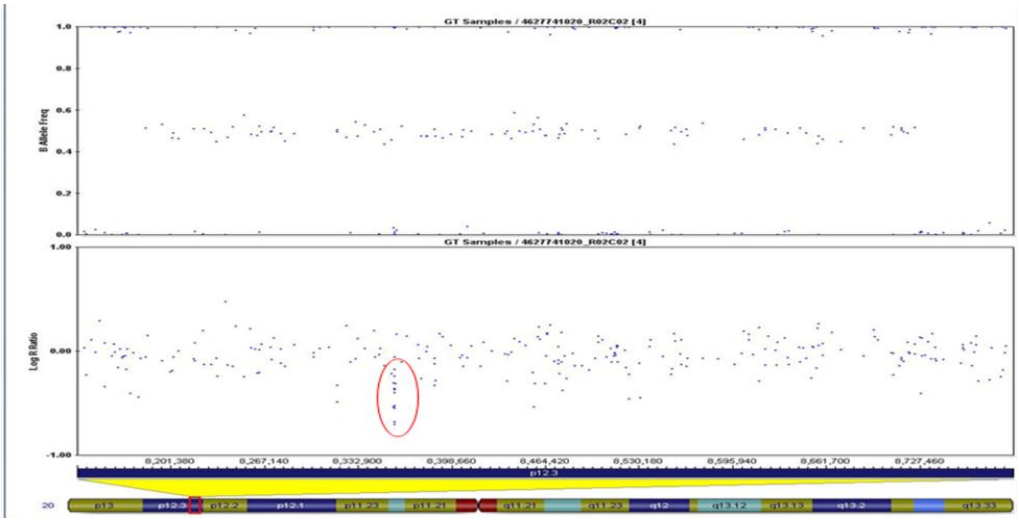


Figure 5.6: The two values generated by BeadStudio software i.e. the B allele frequency (BAF), (upper panel) and the log R ration (LRR), (lower panel). PennCNV software used these values to calculate the copy number variation (CNV). A decrease in the LRR (0> LRR > -1) indicates a single copy deletion which has been identified by 14 markers (circled in red).

Name	Position	Log R Ratio	B Allele Freq
cnvi0102516	8,358,121	-0,3111986	0,9980593
cnvi0102517	8358185	-1,052522	0,03195321
cnvi0112672	8358204	-0,5388463	0,9947988
cnvi0102518	8358334	-0,366732	0
cnvi0102519	8358434	-0,6839929	0,0108272
cnvi0102520	8358539	-0,4034659	0
cnvi0102521	8358590	-0,553011	0
rs708911	8358641	-0,7085187	0
rs771945	8358722	-0,1780442	0,001772681
cnvi0102523	8358825	-0,3786347	1
cnvi0102524	8358880	-0,2455926	0
cnvi0102525	8358931	-0,5293732	0,991891
rs6055811	8358955	-0,05986575	0,995051
rs6055812	8359023	-0,3161553	0,9949234

Table 5.3: A list of the 14 markers from Illumina Chip that detected the deletion. The four single nucleotide polymorphism (SNP) markers start with rs; (rs708911, rs771945, rs6055811 & rs6055812). The other 10 non polymorphic (NP) markers start with cnvi.

5.3.2 Significance of the Chromosome 20 deletion, at the region of interest

This deletion is in the region of chromosome 20 with highest LOD scores and also identified as carrying a possible disease allele by haplotype analysis. It is located at the start of the region of interest ($20p^{12.3}$-$20p^{11.23}$). It is flanked by rs2250711 at 8,203,369 bp and rs771943 at 8,362,876 bp (of 10K chip). The LOD score of both Dominant and npl analyses have dramatically increased at this region. In npl it has increased from 2.6 before the deletion to 2.94 after the deletion (very close to the maximum peak of 3.01). In Dominant analysis the LOD score has increased from 2.03 before deletion to the maximum peak 2.04 after deletion. The haplotype analysis at this region showed that all affected (and the obligate carrier II-3) are carrying the assumed disease allele C, as well as two of the unknowns (IV-4 & IV-5). The other three unknowns (IV-2, III-3 & IV-8) are not carrying the C allele, in addition to the 4 founders (II-4, III-4, II-1 and III-1). That could indicate that this deletion might be associated with the thyroid defect in the family.

Out of the 4 Illumina SNPs markers, only rs771945 is present in HapMap data (with 46% frequency of A and 54% of G in CEU), no further information is available for the other three SNPs which indicates that, they are commonly used in CNV only.

5.3.3 Using Marker rs6055812 to identify the exact location of the deletion

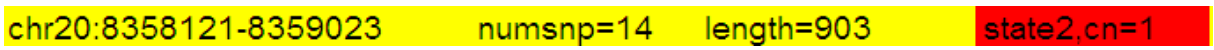
To find out the exact position and sequence of the deleted region, data of the last SNP (rs6055812), in the Illumina list of 14 markers (identified the deletion) have been used. In summary, NCBI software (http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?rs=6055812) was used to search for "probe sequence" of re6055812, (25 nucleotides). That sequence was matched with data base and has revealed the exact end of the deleted region (Figure 5.15). Thus 903 bp were counted upward and that region has been assumed as the deleted region.

5.3.4 Genes in the deleted region

Genome Browse software (http://genome-euro.ucsc.edu/cgi-bin/hgGateway) has been used to detect the gene(s) located in the deleted region (between 8,358,121 and 8,359,023 bp). The deletion was found to be in intron 3 of a huge gene known as PLCβ1 gene. It is the first gene

in the list of the 10 genes identified by GWLA (described in 3.8.1). The gene is located on 20p$^{12.3}$. Its DNA is of 749.21 kb with mRNA size of 3663bp encoding 32 exons. It starts at position 8,061,296 bp on chromosome 20 and ends at position 8,813,547. Exon 1 is 102 bp and followed by Intron one (of 17,524 bp length). Exon 2 is 78 bp and followed by Intron 2 (of 221,007 bp). Exon 3 is 69 bp and followed by "Intron 3 (256,843 bp)", where the deleted 903bp located (after 58,027 bp). Exon 4 is 138 bp (Figure 5.7).

A



B

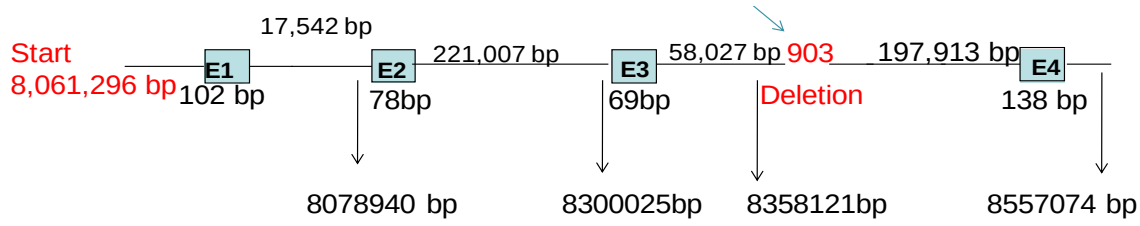


Figure 5.7: (A): PennCNV software output showing the deletion on chromosome 20 with its location in base pair (bp), the number of markers detecting this deletion and the type of deletion (one copy). (B): Phospholipase C Beta 1 (PLCβ1) gene structure showing the first 4 exons and 3 introns. The deletion is also shown at Intron 3.

After data analysis and identification of a deletion at the region of interest on chromosome 20, the sections below will describe further investigations performed on the deleted region as summarised in the workflow (Figure 5.8)

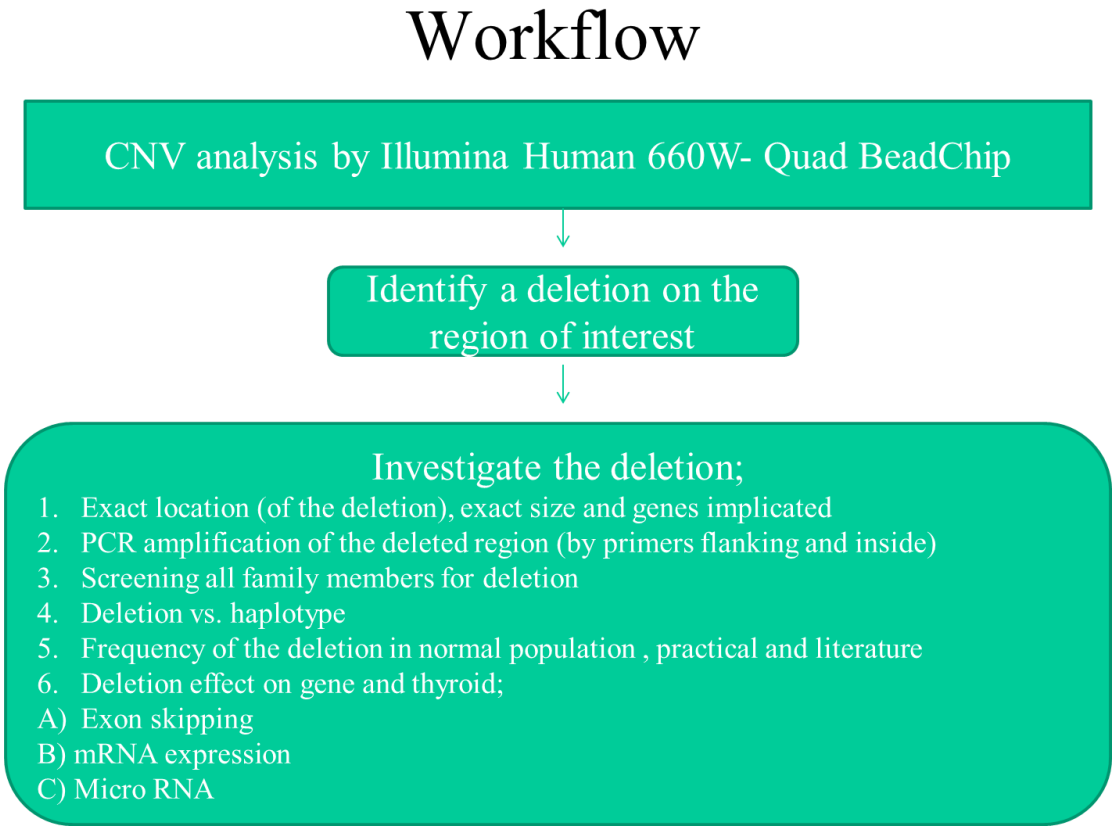


Figure 5.8: Flow diagram summarising the investigations undertaken on the identified deletion.

5.4 Materials & Methods

In this section further investigation of the deleted region will take place, including standard PCR and direct sequencing of DNA from all family members, to determine whether the deletion is in affected and/or unaffected individuals. Studies will be performed to assess the frequency of the deletion in additional cohorts and I will investigate the impact of the deletion on the PLCβ1 gene.

5.4.1 PCR1 to amplify the deleted region

5.4.1.1 **Designing primers**

Forward and Reverse Primers have been designed to amplify the deleted region (Table 5.5) and (Figure 5.9). The forward primer was 441 bp before the deleted region while the reverse primer was 448 bp after the deleted region (Figure 5.15). This PCR set will be referred to as *PCR1*. The primers were designed using Primer3 software and specificity increased by elongating the primer sequence taking care to maintain the GC content.

5.4.1.2 **PCR1 preparation, reaction and agarose gel electrophoresis**

A 50 μl reaction volume contained 3 μl of DNA (100 ng/μl), 5 μl PCR buffer (10x) (Applied Biosystems), with no $MgCl_2$, 1.25 μl dNTP (10 mM) (Promega), 5 μl $MgCl_2$ (25mM) (Applied Biosystems), 1.5 μl forward primer (10 pmol/μl) (Invitrogen), 1.5 μl reverse primer (10 pmol/μl) and 1 μl AmpliTaq Gold® polymerase (5U/μl) (Applied Biosystems). A master mix containing all components, except the DNA template, was prepared. This was aliquoted and DNA samples were added. (In the negative control H_2O was added instead of DNA).

PCR1 reaction was performed with 45 cycles (Table 5.4), using DNA from all family members. That was followed by agarose gel electrophoresis, (described in 2.4.7.3).

5.4.2 Purification of PCR1 products by PEG precipitation

All PCR products have been purified by Polyethylene glycol (PEG) precipitation. PEG solution for DNA extraction was prepared as follows, 8.156 g of sodium acetate were dissolved in 50 ml dH_2O and the pH adjusted to 5.2 with acetic acid. Twenty six grams of PEG (mol. wt. 8000) and 134 mg of $MgCl_2$ were then added and made up to a final volume of 100 ml before autoclaving.

Equal volumes of PCR product (35 µl) and PEG solution were vortexed and incubated for 10 min at RT, prior to centrifugation at 13,000 g, for 30 min. The supernatant was discarded and the pellet washed in 500 µl 70 % ethanol, centrifuged for 10 min at 13,000g and air dried before re-suspension in 20 - 40 µl of Milli-Q pure dH_2O in proportion to the brightness of the bands on the gel. Ten µl from the purified PCR products have then been run on Agarose gel again as quality control.

Temp	Time
95C	5 minutes
	45 cycles of:
95	30 seconds
60	30 seconds
72	30 seconds
72C	6 minutes

Table 5.4: PCR programme (PCR 1 & PCR 2) used to amplify the deleted region using AmpliTaq Gold® DNA polymerase enzyme. Temperature (column 1, Temp), time of incubation and number of cycles performed (column 2, Time).

5.4.3 Direct Sequencing

Each sequencing reaction contained 2 µl "Big Dye Terminator Cycle Sequencing Ready Reaction", (ABI Prism, PE Biosystems), 1 µl primer (10 pmol/L), approximately 25 ng of PEG precipitated PCR product (as quantified by comparison with the 100 bp DNA ladder on an agarose gel), and was made up to 10 µl with H_2O. Sequencing reactions were performed on a Genius (Techne) PCR amplifier, and the sequencing program used was exactly similar to the PCR1.

5.4.4 Sodium Acetate Precipitation

One µl of 3 M sodium acetate (pH 5.2) and 30 µl of absolute ethanol was added to each 10 µl sequencing reaction before vortexing and incubating at RT for 10 min. Samples were then centrifuged for 30 min at 13,000 g, the supernatant discarded and the pellet washed with 500 µl of 70 % ethanol. The supernatant was discarded and samples were air dried, before re-suspension and analysis on an ABI 3100 Genetic Analyser.

5.4.5 PCR2 preparation, reaction and agarose gel electrophoresis

Since primers in PCR1 were flanking the deleted region with ~400 bp apart from both side, a second set of primers was designed in which the reverse primer was located inside the deleted segment (PCR2) (Figure 5.9), (Figure 5.15) & (Table 5.5). PCR2 forward primer designed to be 292 bp upstream from the deletion. Thus the expected gel band for full length is ~1400 bp. PCR2 was run on all family members (as in 5.4.1). That was followed by purification and direct sequencing (as in 5.4.2 & 5.4.3). Hence PCR2 was employed to amplify the full length copy of the region of interest.

5.4.6 PCR1 to check deletion frequency in general population

To check how frequent is the deletion in the general population , PCR 1 (with the primers flanking the deleted region) has been run on (unrelated) 105 DNA samples; from a selected cohort of GP patients whose thyroid function was assessed because of depression DEPTH (DEPression and Thyroid) [330], (5.4.1).

F1	TATTATCCCC AGAAAGAATG AGAGGAGG
R1	CATAGCCACC ATGCCAGGCT AATTTTTTAA
F2	TTTAGGTGTT AACCACCAAA GTGGCTGACA
R2	TTACCTTGCC TGGGAACAGG ATGGGCCCCA A

Table 5.5: Forward and reverse primers, (F1 & R1) for PCR1 & (F2 & R2) for PCR2. The primers were designed using Primer3 software and specificity increased by elongating the primer sequence taking care to maintain the GC content.

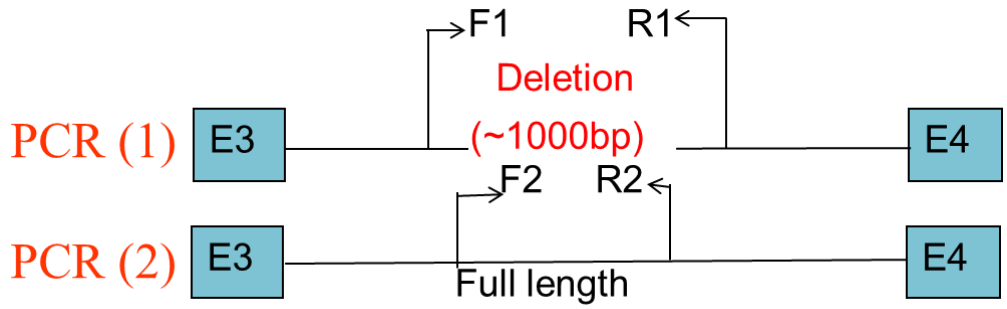


Figure 5.9; Cartoon showing the location of the forward and reverse primers for PCR1 & PCR2. PCR1 primers (F1 & R1) are at a ~400 bp distance outside the deletion. In PCR2, a forward primer (F2) was 292 bp upstream from the deletion and a reverse primer (R2) was inside the deleted region.

5.4.7 The deletion impact on the PLCβ1 gene, Exon Skipping investigation

I hypothesised that the deletion in intron 3 of PLCβ1 gene could lead to skipping of one Exon or more, either the Exon before deletion; Exon 3, or the Exon after deletion; Exon 4. Samples of cDNA from thyroid tissues of 2 affected members (IV-6 and IV-3) and one presumed unaffected member (IV-4) were prepared, (as described in 5.4.7.2).

5.4.7.1 **Designing exonic primers to amplify exons 3 and 4 of PLCβ1 in Thyroid cDNA**

To investigate Exon skipping, PCR primers have been designed to amplify both exons (in cDNA) using primer 3 software, and PLCβ1 mRNA sequence. Forward primers have been designed in exons 2 and 3 (F2 & F3) and reverse primers have been designed in Exon 4 and 5 (R4 & R5) (Figure 5.10) & (Table 5.6). A combination of F2 & R4 amplifies Exon 3 and a combination of F3 & R5 amplifies Exon 4. A combination of F2 & R5 has also been used to confirm any findings.

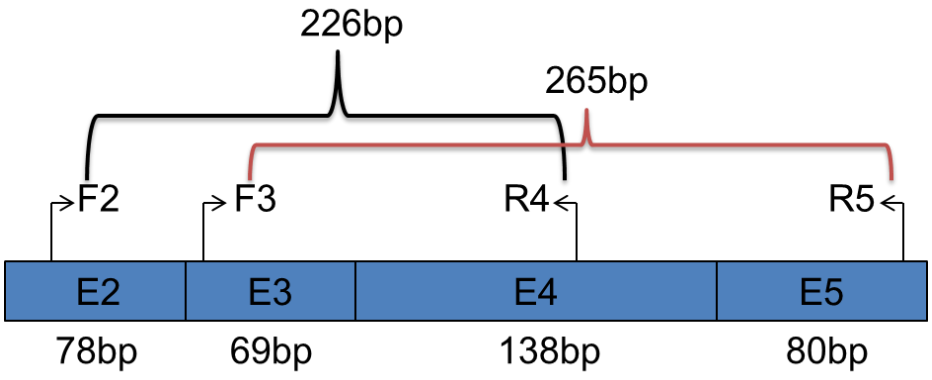


Figure 5.10: Cartoon showing part of phospholipase C beta 1 (PLCβ1) mRNA with exons 2, 3, 4 & 5 and locations of forward and reverse primers to amplify exons 3 & 4 in cDNA from thyroid tissues. Size of each exon is shown underneath in base pair (bp).

F2	TTTTGAGGAC TGACCCTCAG GGATTTTTCT TTTAC
F3	TCAGCCTTGT CAAAGATGCC AGATGTGGGA GAC
R4	GGTGTATGGG CCTGACCTCG TGAACATCTC CCATTT
R5	CCCAAAACAT GTCCAGGGAT GCATTTCTGG AAAA

Table 5.6: Forward and reverse primers used to amplify exons 3 & 4 in cDNA, of family #1 thyroid tissues. F2 & F3 are forward primers in exons 2 & 3 respectively. R4& R5 are reverse primers in exons 4 & 5 respectively, as seen in (Figure 5.10) above.

5.4.7.2 **RNA extraction from normal thyroid tissue for amplifying exons 3 and 4 of PLCβ1 mRNA**

Messenger RNA was extracted from the thyroid tissues of family members (IV-5, IV-6 and IV-4) using TRizol Reagent (Invitrogen) and reverse transcribed in 20-µL reactions using oligo dT (OdT) primers. In summary a snap frozen thyroid tissue was put in liquid nitrogen, ground to a powder and put in a tube. Five hundred µl of TRizol added to the tube, followed by addition of 100ul of chlorophorm then centrifugation. The upper layer then has been separated and 500 ul of isopropanol was added, and then centrifuged again. The pellet has been washed using 70% ethanol, then centrifuged. The final precipitate was reconstituted with RNAse free water. The optical density of the RNA has been measured at 260 and 280 nm wave length to obtain concentration and ratio.

5.4.7.3 **Reverse transcription**

Six µl of RNA (1ug/6 µl) has been added to 14 µl of reverse transcription Master Mix (contains dNTP, OdT primers, buffer, RNAas Inhibitor (RI) and Reverse Transcriptase enzyme), then placed in the Thermo cycler on reverse transcription programme (37C for 1 hour followed by 95C for 5 minutes). The cDNA produced has been tested by PCR of PGK (housekeeping gene; expressed in all tissues, to maintain cellular function [331]), using

master mix and PCR program (will be described in 5.4.7.4). The PCR product was then separated by electrophoresis on an agarose gel (described in 2.3.8.3).

5.4.7.4 **PCR amplification & gel**

After confirming the quality of the cDNA, PCR was performed to amplify PLCβ1 exons 3 & 4, using the cDNA of both affected family members (IV-6 and IV-5) as well as their unaffected sister (IV-4). Seven unrelated thyroid cDNAs were included as controls. Three sets of primer were used (F2 with R4, F3 with R5 and F2 with R5). PCR master mix was prepared as follow:

A 50 μl reaction volume contained 2 μl of cDNA (100 ng/μl), 5 μl PCR buffer (10x) (1.5mM $MgCl_2$) (Promega), 1 μl dNTP (10 mM) (Promega), 1 μl forward primer (10 pmol/μl) (Invitrogen), 1 μl reverse primer (10 pmol/μl) and 2 μl Taq polymerase (1U/μl) (Promega). A master mix containing all components, except the cDNA template, was prepared for each primer combination. This was aliquoted and cDNA samples were added. (In the negative control H_2O was added instead of cDNA).

PCR reaction programme used was similar to PCR1 in genomic DNA. All PCR products have been run on agarose gel (described in 2.4.7.3).

5.4.7.5 **QPCR for amplifying exons 3 and 4 of PLCβ1 mRNA**

To investigate the possibility of exon skipping further, Quantitative PCR (QPCR) was performed (by a colleague in the department; Dr Fiona Grennan-Jones), using the same primers to confirm the presence of both bands (short and long). QPCR was conducted using SYBR Green incorporation measured on a Stratagene MX 3000 light cycler. The cDNA samples from 3 family members (III-2, IV-6 & IV-4) were investigated along with thyroid cDNA samples from 5 unrelated controls. The experiment was performed in 3 sets of duplicates.

5.4.8 The deletion impact on the PLCβ1 gene, mRNA expression

Several PLCβ1 gene isoforms have been reported in the literature [332], the main two (PLCβ1a & PLCβ1b) have been investigated for family #1. Both isoforms are varying in their C terminal, with an extra exon (of 118 bp) in PLCβ1b.

PCR primers have been designed (using primer 3 software by a colleague in the department; Dr Lei Zhang) to amplify both isoforms (Figure 5.11). QPCR has been performed (by colleagues in the department; Ziduo Li & Dr Fiona Grennan-Jones), on 3 cDNA samples from family members (III-2, IV-6 & IV4) and 7 cDNA from unrelated controls.

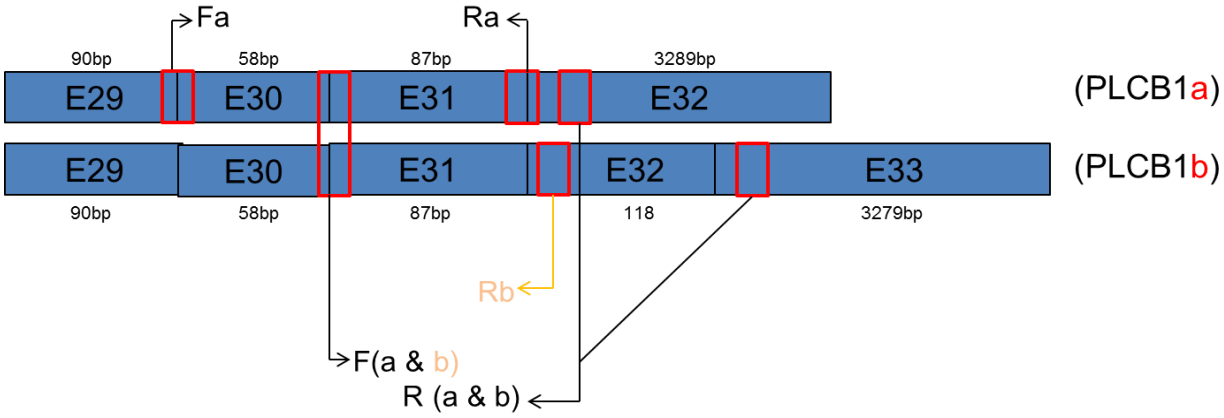


Figure 5.11; Cartoon showing the C- terminal of Phospholipase C beta 1 (PLCβ1) isoforms (PLCβ1a & PLCβ1b). PCR primers designed to amply each isoform are shown. (Fa & Ra) for PLCB1a, F (a & b) & Rb) for PLCβ1b, and (F (a & b) and R (a & b)) for both isoforms. The extra exon (118 bp) in PLCβ1b is shown and R (a & b) primer is located in exon 32 of PLCβ1a which is exon 33 in PLCβ1b, because of the extra exon.

Fa	TCAGATGGAA GAGGAGAAGA CA
Ra	ACCTGCAGCT TGGGCTTT
F(a&b)	GAGGCTAGAA GAAGCGCAAA
Rb	CCTGCACGAA TTCCAAAATC
R(a&b)	CTTGAGAGCT TGAGGGTTGG

Table 5.7; Forward and reverse primers used to amplify phospholipase C beta 1 isoforms. PLCβ1a alone amplified by (Fa Ra), PLCβ1b by (F a&b R b) and both isoforms by (Fa&b Ra&b)

5.5 Results

5.5.1 PCR1, gel

The expected size of the PCR product in PCR1 is 889 bp for the deleted allele but ~1700bp for the full length. (Figure 5.12) shows that all affected and carriers possess the short band (of ~800 bp) which comprises the regions before and after the deletion. Five unaffected (IV-2, III-3, IV-8, III-4 & III-1) showed no bands. The remaining two unaffected (II-4 & II-1) showed full length (of ~ 1700 bp) bands as does the unrelated control.

From gel images of PCR1 in all members of family #1 (e.g. Figure 5.12), it appears that members carrying the "assumed disease allele C" (in haplotype analysis, described in 3.8), displayed the short band in this PCR (Figure 3.12).

5.5.2 PCR1, direct sequencing result

When the ~800bp product was sequenced and then compared with the NCBI data base, I found the first ~350 bp corresponded to the expected sequence before the deletion. This was followed by the insertion of (ATAA) and then another ~268bp after the intervening ~1000 bp deleted sequence. Thus only the regions before and after the deletion were amplified (Figure 5.13) & (Figure 3.13). Thus the variation in the PLCβ1 is insertion and deletion (InDel).

When the ~1700 bp product was sequenced then compared with the NCBI data base I found an exact match with PLCβ1 (the region with deletion), indicating the presence of the full length DNA (Figure 5.13).

Since not all unaffected have shown the full length bands, PCR2 primers were designed with reverse primers inside the deleted region.

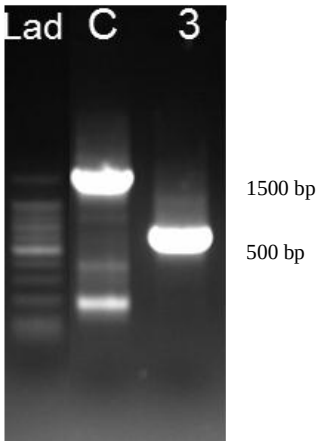


Figure 5.12: PCR1 products from individuals IV-6 genomic DNA, analysed by agarose gel electrophoresis and ethidium bromide staining. The first lane (Lad) contains the 100bp ladder, the second lane is the unrelated control (C) and the third lane subject IV-6 (labelled as 3). A clear band of ~800 bp in the patient demonstrates the deletion. The control shows a ~1700 bp of full length DNA.

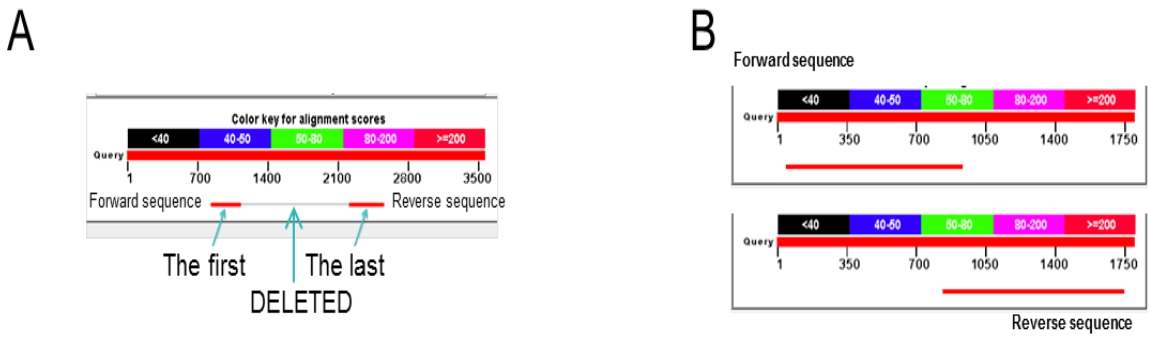


Figure 5.13: Screen shot of National Centre for Biotechnology Information (NCBI) output; (A) Alignment of the direct sequencing of PCR1 product for the affected family member (IV-3) and NCBI data base. Only a match with the first ~350 bp and the last 268bp can be seen. (B) Alignment of the direct sequencing of PCR1 product for the unrelated control and NCBI data base. A clear match with DNA without the deletion (full DNA) can be seen with forward and reverse sequences.

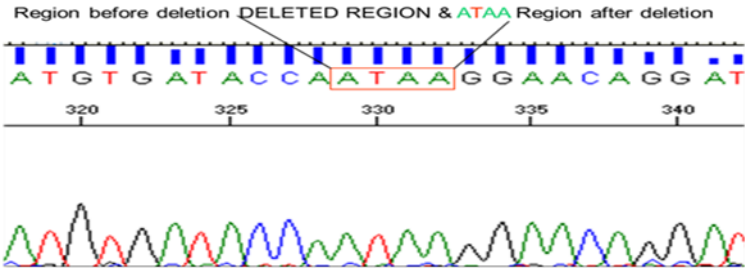


Figure 5.14; Screen shot of the direct sequencing electropherogram showing the deleted region and the junction of the regions before and after with the inserted nucleotides ATAA (red box).

5.5.3 Exact size of the deleted region

The sequencing result revealed that a further 77 bp are missing (before the 903 bp deletion, identified by Illumina chip), and further 101 bp are missing (after the 903 bp deletion, identified by Illumina chip), (Figure 5.15). By adding these 2 further missing regions to the deleted region and subtracting the 4 nucleotides detected at the junction that will make the exact deleted region to be 1076 bp. Thus the deleted region starts at 8358048 bp and ends at 8359124 bp, at ($20p^{12.3}$).

5.5.4 PCR2, gel and direct Sequencing

All family members including the founders and the control showed a single band of ~ 1300 bp on the gel (Figure 5.15). Sequencing data of PCR2 (~1300 bp) were BLASTed using NCBI data base. An exact match within the PLCβ1 sequence (containing the deleted region) was observed. Thus PCR2 was able to amplify the full length DNA copy, as the reverse primer was in the deleted region.

ATCCCATGGCAGTTAATATGGAATCTGTTGTGGTTTGTGGATTAGTTGTGTGATTCAGGTGCCATTGGAG
GTTTCTGGAAGAGAAGGGCCTTTTAAAAGCCTCTGAGAGGCCAGTATTATCCCCAGAAAGAATGAGAGGA
GGTTTATTGTATGCACGCAAGCAATATCTATGAATCCCTTTACTTGCTTGTATCTTACCCTCTTCTGCCC
TAGGAACAGGCTCGTGGGAGAAGTGACCCAACTACACCTTCCTGCCAGTCTCGAGCATTTTAGGTGTTAA
CCACCAAAGTGGCTGACACATGTACTTACTCACTTGGTTTTAATTTGCAGAGTAGGAACAATGGACTTAG
ATTTTTTTAAACTTCTCTAACTCTTTAATCACTGAATGAAATAACTCATGTTGGAAAGACAAAGAATTTC
CTGGTAGACCTTTTGAACAAGAGCTCATAGATTTCAGTGCCATTGGTAGCCAATGTGATACCATGGCATT
GTTGTCTATTGTGTTTTTTAGGTTGTGACCTATTTGTGGGTTATGAATTCAATTTTGTGGATTGGGACCT
GCAATTTAAAAAAAAGATGAAGGAAATAGAACATAATAGAGTCAAGCGCCATGCAATACAAAATGTCAGA
GGTATCCCAACTAACAAGGGTGAATGATGGTTTGTAAAACTTCAGTTACGTATGTATACGTGCATGTATG
AGTGTTCTGAGTGATAATATAAAATGCATTTTTTACTGAGGAGCATGATCAGAACAATATCAAACCATTG
ACCAACATCAACATATTTCTATTAGAAAGATATTTAAATGAAGTTTGTATAGTAAAAGGAAGAAAGAACA
CTAAAATATTTGCAAATCTACCCTGCACAGTCACTTTAATCTTTTCAGCTTCTAATTGAGATAGATGCTC
TATATTACCCTGTTTATTAATAAACCCTCAAAGATCTAAAAAGGTGAATTAACTCTCAGGGGCTTTAACA
GATAAATTAACTTGAACCTGCCTAGTTGTCTGTTTCAGAGCCAGAATTCGAACATGGATTCATTTTAGCC
GCAGTAAGCACCTGTGTTAGTCAGAATTCAGTGTGGGTAACAGAAGGGCAGGCCAGATTGGAGAGAGCTG
GAAACATATTCCAGGTAGAGACAATGGTGCCTGATGAAGCCTAAGGGTCTGAATGCTGGAGTAAAGCCAG
GGGTTATTTACACTGTGACTGGGTGTGAATCTACAGATAGATCTGAAACTAGAAGATAAGCCCATGAAGG
AAATGGGAAAATGTGCCCTGATTTTGAGTTACATTAGCAAAAAAAATCAAAGCTGAGTCCAAGATGGAAG
AACATCTACATAATTAGAGAGGACAGAAGTGATCACGAAGACCTAGGGGACAGGGAGAGGCCTCAGGTAA
TGCAAAATCTGTACATAATGATTGAGGTCAGGAGAACAGATGCTCTTCCCGGGTCAGTTTTAAGGCTGAA
ATGTTTCAGGAAAACTGGGAACAAAGCATGGTTGGAAATAGCAGGCTAAGCTGATGCATCAGGAGACTGC

AGATCTCAGTG**TTACCTTGCCTG**GGAACAGGATGGGCCCCAAGTATGGCTGAGAATGCAAGTCTCATGGG
TATATGTTTAATGAACCTAGTTGGGAGATGTTAGGAGGAAGACTGGTTCTGATGTGAAAGTTTCCTTCCA
AAGAATAGGCCAACCTAACTGCAGCAGTGCTGATTATGATGGAAATTTCCTATTTTGATTCTTCATCAAA
AGAAAGAAGAGCCGGTACCATGGTTCTTGCCTGTAATCCCAGCACTTTGGGAGGCCAAGGTGGGAGGATC
ACTTGAGGCCAGGAGTTTGATACCAGCCCCTATAGGAAACATAACAAGACCCCATCTCTACTAAAAATTA
AAAAATTAGCCTGGCATGGTGGCTATGTGTCTTAACAAGACCCCATCTCTACTAAAAATTAAAAAATTAG
CCTGGCATGGTGGCTATGTGTCTTTAGTCCTAGCTACTCAGGAGGCTGAGGCAGGAGGCTCTCTTGAGCC

Figure 5.15; The region of direct sequencing output showing position of the deletion in phospholipase C beta 1 (PLCβ1) gene. The deleted region identified by Illumina chip is highlighted in yellow. The exact deletion, identified by direct sequencing is highlighted in both yellow and green. Primer pair for PCR1, flanking the deleted region (highlighted in blue). Primers pair for PCR2 (typed in bold). Thus the exact deleted region is 1076 bp. The single nucleotide polymorphism (SNP) (re6055812) used to identify the exact location of the deletion (A) is in red.

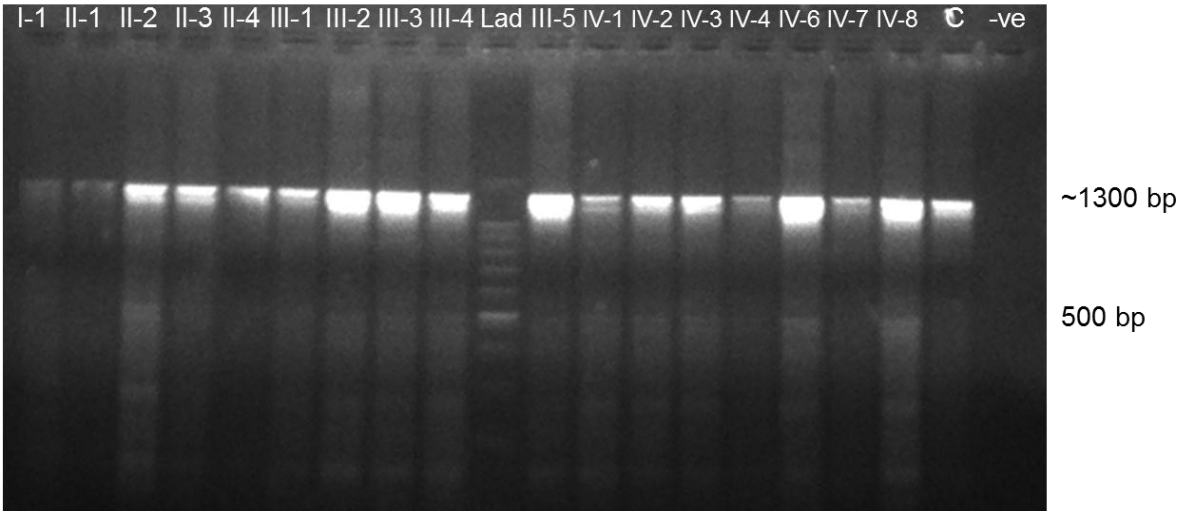


Figure 5.16; PCR2 products from all individuals genomic DNA, analysed by agarose gel electrophoresis and ethidium bromide staining. The first 9 lanes contain DNA from family individuals (labelled as in family tree). Lane 10 (Lad) contains the 100bp ladder and lanes 11-18 contain DNA from family individuals. Lane 19 (C) contains unrelated control and lane 20 (-ve) contains the negative control (master mix without DNA template). A ~1300 bp amplicons in all family members and unrelated control are shown, which indicates a presence of full length DNA (with PCR2).

5.5.5 Screening DNA samples from patient with MNG of PTC

Through collaboration with Dr Clifton Bligh in Sydney Australia, genomic DNA was extracted from thyroid tissue of 70 patients (all of self-reported white European ancestry) undergoing surgery for non-familial PTC and an additional 81 operated for non-familial MNG. PCR1 has been performed on both groups of samples, as screening for "InDel" in PLCβ1. The deletion was not detected in any of the PTC patients but 4 of the 81 MNG were heterozygous for the deletion. The direct sequencing showed the same picture as in family #1, including the insertion of (ATAA). The 4 MNG patients (3 women, 1 man) known to be unrelated and with no apparent family history of MNG or PTC at the time of their surgery. The age at thyroidectomy was from 27-59 years and the pathology is variously described as 'oncocytic neoplasm with variable patterns of growth' to 'cystic degeneration with calcification'. The histology in the family comprises 'multiple papilloid adenomata'.

5.5.6 Rarity of the "InDel" in PLCβ1 gene.

To confirm that this deletion is not just a normal variant in the general population, it should not have frequency of > 1%. In addition if deletion frequency was >5% in a disease cohort then it could be considered as pathogenic (as described in 5.1.2). To investigate that practically, PCR1 has been performed on 105 DNA samples (from unrelated individuals). A literature search has also been done looking for the same deletion and its frequency among CEU.

5.5.6.1 **Practical investigation; PCR1 on 105 DNA of unrelated people**

Out of 105 DNA samples from unrelated individuals (Figure 5.16), 104 showed no deletion, while the deletion has been detected in one sample (17D), making the frequency of 1% in the normal population.

The only sample with the deletion was a 40 years old euthyroid female (limited information of disease status is available). Direct sequencing has matched the regions before and after deletion, including the ATAA insertion at the junction, as in the affected members in family

#1. PCR2 has also been run on this sample and a full length amplicon was obtained (as in the members of family #1).

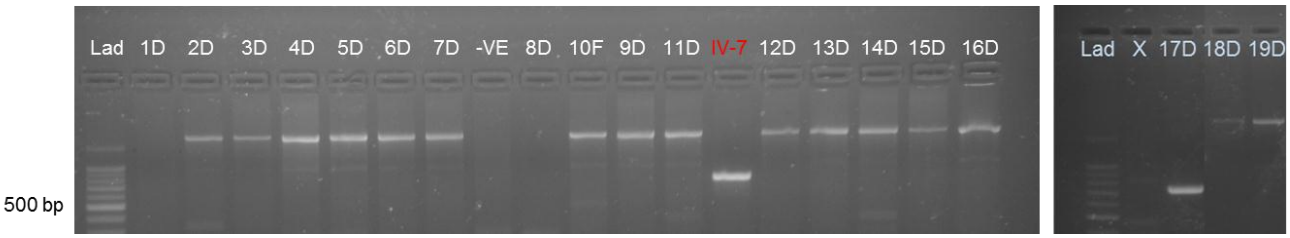


Figure 5.17; PCR1 products from some unrelated DNA samples (1D-19D), analysed by agarose gel electrophoresis and ethidium bromide staining. Lane 1 & lane 20 (Lad) contain the 100bp ladder. Lanes 2-8, 10-13, 15-19 & 22--24 contain DNA from 1D to 19D respectively. Lane 9 (-VE) contains the negative control (master mix without DNA template). Lane 14 (IV-7) contains DNA from affected family member and lane 21 (X) is empty. The products are showing the expected ~1700 bp amplicons in most of them (and no band in the other). Only (17D) has shown the short band (~800 bp), in addition to the affected family member IV-7 which was used as a positive control.

5.5.6.2 **Literature search for frequency of the deletion in CEU**

By looking in "Databases of Genomic Variant" website (**dgv**.tcag.ca), (Figure 5.18), for the deleted region, 7 studies were found which reported CNV (deletion or duplications) in the PLCβ1 gene, of which 4 were deletions [308, 311, 333] & [334] (Table 5.8). First study "a comprehensive, population-wide CNV map of the human genome" [311], has reported a deletion of one copy (LRR = -0.645), at the same region (8,358,097-8,359,094), and labelled

as “CNVR7782.1”. The deletion was found in 2 out of 180 CEU, but not in 180 YRI, 45 JPT neither in 45 CHB, of the same study. The second study [333], has used the same samples as the first study [311], including 60 individuals from CEU population. Both studies (1st & 2nd) have detected “CNVR7782.1” deletion in the same 2 individuals (NA12154 & NA07037). Thus the second study was not added to the total. The third study [308] has investigated CNV in 270 individuals 90 of which were CEU. Fourth study [334] investigated a large cohort (2026 individuals), comprising 1320 Caucasians (65.2%), 694 African-Americans (34.2%), and 12 Asian-Americans (0.6%). The last two studies (3rd & 4th) have reported intronic deletion (of > 0.35 Mb) at $20P^{12.3}$, a huge region includes CNVR7782.1 region (Table 5.8), but were unable to detect small deletions.

Since only the first study was able to detect the CNVR7782.1 deletion, it will be taken in account in investigating the rarity of the PLCB1 deletion. Thus 180 CEU and 450 other have been investigated and two “CNVR7782.1” deletions have been detected in CEU, but none in the others, giving deletion frequency of ~1%.

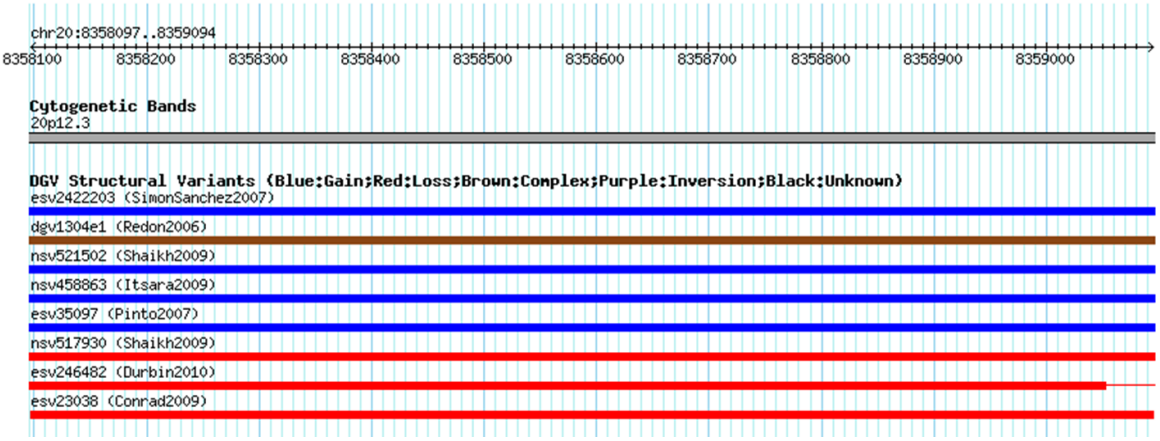


Figure 5.18: Screen shot of the output from databases of genomic variant (DGV)” websites for the deleted region, showing 7 studies at that region. One of the studies reported deletion [311] (red line), one reported deletion and duplication [308] (brown line) and one reported duplications (blur line) [335]. A full description is in the text and (Table 5.8).

Study	CNV	Starts (bp)	Ends (bp)	Size (bp)	CEU	Other
Family #1	inDel	8,358,048	8,359,124	1,076	18	
Unrelated control	inDel	8,358,048	8,359,124	1,076	105	
Conrad et al 2009 [311]	Deletion	8,358,097	8,359,094	997	180	270
Redon et al 2006 [308]	Deletion	8,002,182	8,595,665	593,483	90	180
Redon et al 2006 [308]	Duplication	8,022,116	8,536,318	514,202	==	==
Pinto et al 2007 [335]	Duplication	8,050,867	8,517,610	466,743	506	0

Table 5.8: Summary of the studies reporting deletion at the region of interest or closer. First column (study) reports the name of the study in which Family #1 and samples of the unrelated control are also included. Second column (CNV) reports the type of the Copy number variation (CNV). Third (starts (bp)) and fourth (ends (bp)) columns report start and end of each CNV. Fifth column (size (bp)) reports the size of CNV. The size is precise in family and unrelated control, due to PCR and direct sequencing, but presumed in the other studied as depends on the chip analysis only).

5.5.6.3 **Investigation of exon 3 or 4, skipping, by PCR of cDNA**

I hypothesised that the deletion might induce exon skipping, assuming that PCR product of F2 & R4 primers to be 226bp in normal case and 157bp in case of exon 3 (69bp) skipping. PCR product of F3 & R5 primers was assumed to be 265bp in normal cases and 127bp in case of exon 4 (138bp) skipping. The gel pictures have shown a clear single band of the correct expected sizes, which has excluded exon skipping (Figure 5.19). However the band obtained could be the product of the full length copy while the deleted copy (with missing exon 3 or 4) may not be visible. Therefore further investigation was made by colleagues and confirmed that no exon skipping had taken place (as described in 5.4.7.5).

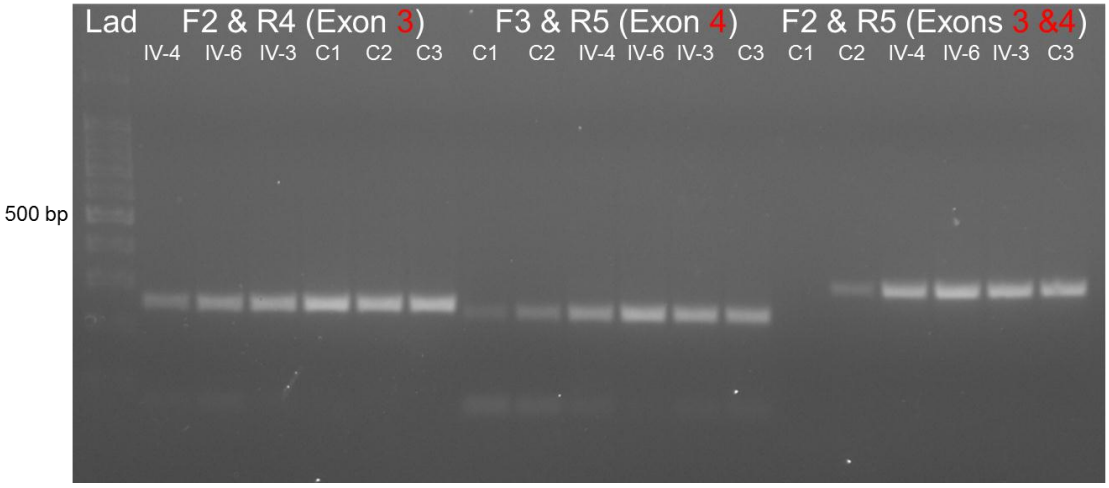


Figure 5.19; PCR products from cDNA of three members of the family IV-4, IV-6 & IV-3 and three unrelated cDNA controls (C1, C2 & C3), analysed by agarose gel electrophoresis and ethidium bromide staining. Three combinations of exonic primers used to amplify exon 3, exon 4 and both exons 3 & 4. Lane 1 (Lad) contains the 100bp ladder. Lanes 2-7 contain exon 3 of cDNA from IV-4, IV-6, IV-3, C1, C2 & C3 respectively. Lanes 8-13 contain exon 4 of cDNA from C1, C2, IV-4, IV-6, IV-3 & C3 respectively. Lanes 14-19 contain exons 3 & 4 of cDNA from C1, C2, IV-4, IV-6, IV-3 & C3 respectively. The gel shows a clear single band of both exons (3 & 4), indicating no sign of exon skipping.

5.5.7 mRNA expression investigation by QPCR

Expression of PLCβ1 isoforms (PLCβ1a & PLCβ1b) by QPCR was performed using the 3 cDNA samples from family members (III-2, IV-6 & IV-4). Comparison between mRNA expressions of both isoforms in the three members vs. normal controls indicated high expression of both isoforms in the affected and carriers family members compared with the unaffected unrelated controls. However, expression of the PLCβ1a isoform was higher than the PLCβ1b isoform in all family members as well as the unrelated controls.

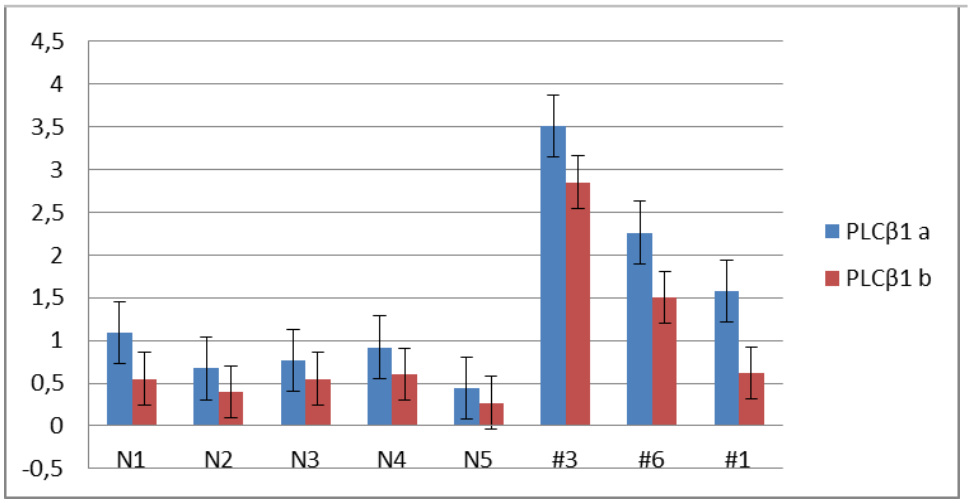


Figure 5.20; QPCR output chart showing mRNA expression level of phospholipase C beta 1 a (PLCβ1a) & PLCβ1b, on 3 cDNA samples from three affected members IV-6, IV-4 & III-2 of family #1, (labelled as #3, #6 & #1 respectively) and 5 normal controls (N1-N5). The data show higher transcript levels (of both PLCβ1a and PLCβ1b) in family members compared to normal controls. The PLCβ1a expression is higher than the PLCβ1b in all family members as well as the unrelated control

5.6 Conclusion

The PCR amplification and direct sequencing of the deleted region of 903 bp identified by Illumina chip, by PCR1 with primers flanking the deletion and ~440 bp apart from both side, was expected to show a short band ~880 bp in the individuals having the deletion reproducing the ~ 440 bp region before and ~440 bp after the deletion. However, only ~ 270 bp were seen after the deleted region (Figure 5.13). That is because the deletion was not only 903 as shown by the chip, but 1076 (as described in 5.5.3). Several different PCR primers were tried before identifying the right pairs, which were able to amplify the region. Even though, it was difficult to amplify the undeleted copy in affected family members which could suggest that, there may still be more to understand about the variations in chromosome 20 in the family, especially in light of the negative LOD score on the downstream region of 20p, at $20P^{11.23}$ (20.8 Mb, 45.6cM) and below.

The 2 sets of PCR (PCR1 & PCR2), with primers flanking or within the deletion, have shown that both deleted and undeleted regions can be amplified.

The PCR products of some affected family members show faint bands as seen in (Figure 5.15), which either indicates less copies of the product or poor quality of DNA samples, possibly because some samples were collected many years ago (over 10 years).

The deleted region identified by Conrad et al, (CNVR7782.1) was shown to be shorter than that found in this study. Although not many details are available in the article, it seems that the identifications in all CNV studies (shown in Table 5.8) depend only on the chip analysis and do not include direct sequencing. Even in my study the deletion identified by Illumina chip was of 903bp size, but after PCR and direct sequencing it was found to be 1076bp.

Since the revolution in Oligonucleotide arrays, plenty of studies have been performed on the whole genome, one of which [336] has linked CNVs with recurrent alteration involving a region on chromosome 20 ($20p^{13}$) in 55% of primary myelo-fibrosis patients. Another study [337] has evaluated the resulting effect of LOH on the gene expression in triple-negative breast cancer (TNBC), showing that the majority of mono-allelic expression of TNBC can be explained by genomic regions of LOH.

My study has used a highly informative array with 660,000 markers including both SNP and NP markers, which has given great strength to the output. Identifying a deletion at the same region where maximum LOD was obtained is a strong evidence of correlation between the two. Furthermore having the deletion in the affected or carriers of the C allele was another significant finding supporting that this study is going in the right direction and encouraging for further investigation.

The presence of the deletion in one samples out of 105 DNA investigated in the lab and in two samples out of 108 CEU investigated for CNV by Conrad et al, has concluded that the deletion is rare in CEU population with 1% frequency. Furthermore, it was not possible to obtain more information about the woman, identified with the same deletion out of the 105 DNA from unrelated individuals screened by PCR1. I cannot exclude the possibility that she has MNG, especially as she was in a cohort of people in whom testing thyroid function was indicated when visiting their GP (as described in 5.4.6).

Furthermore, I have searched for any open reading frames in the deleted region, (with and without ATAA), but none of any great length were found. The first ORF was 110 bp (36 amino acids) before the deletion (200 bp upstream from the ATAA insertion), and contains one start codon ATG and one stop codon TAA. The second was 125bp (41 amino acids) in the deletion (40bp before the end of the deleted region), with 4 ATG start codon and one TGA stop codon. The third ORF was 110 bp (36 amino acids), located 245bp after the deleted region, with 3 ATG start codon and one TGA stop codon.

In addition an experiment to investigate exon (3 & 4) skipping in cDNA of some affected in family #1 was done using standard end-point PCR and also QPCR. Both confirmed that no exon skipping had taken place. In standard PCR, I expected to see large (full-length) and small amplicons but only the latter were routinely observed. Although the strength of the band could give clue of being for two copies or a single copy, but this is not quantitative method. Therefore QPCR was necessary at this stage to calculate the number of DNA copies and thus identify any deletion of one copy.

PLCβ1 occurs in several isoforms, the main two are PLCβ1a and PLCβ1b. Expression of both isoforms has been investigated in thyroid tissues from family vs. controls (tested for the

PLCβ1 deletion), using QPCR measurements, which indicated significantly higher PLCβ1 transcript levels in thyroids from family members with the deletion, compared with controls who do not harbour the deletion. This could give a clue that MNG/PTC in the affected family member could be due to over-expression of PLCβ1, which is due to the deletion. Increased PLCβ1 expression has been reported in lung carcinoma[338] and high activity of PLC enzyme has been reported in thyroid neoplasms[339]. Furthermore CNVs in chromosome 20 have been reported in breast cancer [340], although these were shown to be associated with deletion of the entire PLCβ1 gene [341, 342].

PLCβ1 generates IP3 and DAG leading to PKC activation via G protein-coupled receptor signalling to the MAPK cascade[343]. BRAF is downstream of the Gq/ PLCβ1/MAPK signalling route, and activating mutations of *BRAF*, mainly $BRAF^{V600E}$, which is the most common cause of PTC and play role in thyroid proliferation [176].

Chapter 6 General Discussion

The aim of this work was to characterise the gene-defect behind the novel cause of euthyroid MNG of adolescent onset, progressing to PTC in 8 individuals across 4 generations of a large European family. The same family has been previously investigated during my MPhil study, when I have eliminated the candidate genes on the basis of thyroid/breast and thyroid/kidney disorders (*PAX-8, NIS*), by Sanger sequencing, through which *PTEN* was eliminated as well. I have also excluded the known loci for MNG and FNMTC on chromosomes 14q, Xp, 3q 9p, 2q and 1q, by Microsatellites analysis [181].

Thus in this study I have started with mapping the gene disposing to MNG/PTC in this family, I performed a dense GWLA, using high density Affymetrix 10k SNPs array. In chapter 2, I have processed the microarray protocol and obtained the genotyping of over 10,000 SNPs for all 18 family members. My study was performed some years ago, when 10K chips were more widely used. Several studies have reported the use of this chip, one of which reported that the chip was quicker than others [2]. Another study has used the same chip to scan the whole genome chromosomal copy number for renal epithelial tumours classification, by analysing 20 paraffin-embedded tissues and reported that SNP array can detect characteristic chromosomal abnormalities in such samples, thus recommended as a supplementary study to classify and test prognosis of such tumours [3]. Bigger chips such as 500K, which are more preferable these days, started to be available at that time, but obviously were more expensive but may have saved me having to use two chips (10K & 660K) in my study. The second and larger was for CNV and this data would have been available if I had used a chip with a larger SNP set for the GWLA.

In chapter 3, I undertook a genome-wide linkage analysis after an extensive quality control check using several softwares (PLINK, GRR). A Genome-wide significant evidence of linkage to chromosome $20P^{12.3}$-$P^{11.23}$ was obtained, with a maximum non-parametric (multipoint) LOD score of 3.01, across 20cM, using Merlin software. The same region showed a maximum (multipoint) dominant LOD score of 2.03, based on a disease model with 0.01 allele frequency, 50% penetrance for males and 90% for females, age of onset for both >12. LOD scores on the remaining 21 autosomes and X chromosome were all below 1. The software used to perform these analyses (Merlin) is very popular, reliable and has been used

for estimating IBD (identity-by-descent) allele sharing probabilities in pedigrees in many studies [4], one of which [5], has used Merlin in linkage analysis and reported Pendrin as a new susceptibility gene for autoimmune thyroid disease. In contrast to other complicated software, I have found Merlin straightforward and easy to use, especially for a nonmathematical person like me.

Although haplotype analysis is taken into consideration by Merlin while performing multipoint npl, analysing it separately has given me visible data which I can see and judge. Thus I found one haplotype, from rs1923876 to rs1099620 co-segregated with the affected phenotype in all 8 patients and 4 out of the 6 unknowns, one of which is an obligate carrier and has the haplotype, while the 4 founders in the family were not having the haplotype. One of the 2 unknowns not having the haplotype (IV-2) has presented telomeric recombination events at rs1923876, while the other (III-3) has presented centromeric recombination events at rs1099620. Combination of the haplotype region with maximum npl LOD score region (starts at rs2223539), has delimited the susceptibility region to an 8.73 cM (3.7Mbp physical position) span on chromosome ($20p^{12.3}$-$p^{12.1}$), which encodes 10 genes. Haplotype analysis has been used in many studies to represent the assumed disease or risk allele [344]. The usefulness of the method has also been increased by the haplotype data released from samples in the 1000 genome project [192] or international HapMap project [193].

The 10 genes found in the candidate region are PLCβ1, PLCβ4, LAMP5, PAK7, ANKEF1, SNAP25, SLX4IP, MKKS, JAG1 and BTBD3. I started with an in silico analysis, searching for previous reports of any of the 10 genes in either MNG or PTC cases or any GWLA reported linkage with disease. Since thyroid diseases are more common in women than men [345] and in vitro studies reported that estrogen can promote thyrocyte proliferation[346], my in silico analysis has included, searching for binding sites for estrogen receptors in the promoters of these genes. In addition, my in silico analysis included expression of these genes in different tissues including thyroid. No solid evidence was observed which could point out a candidate gene or help in decreasing the number of genes for further investigation.

After my first publication [181], a second family (family #2) has been introduced to our department showing similar signs and symptoms of MNG even similar histopathological picture for malignant thyroid cell (multiple papilloid adenomata). GWLA was performed on

6 individuals of that family, 4 of which are affected. In case of any positive LOD score in family #2 at the candidate region in family #1, both LOD scores would be added to each other increasing the maximum LOD score and giving more strength to the study. No common genetic variants were found in the families, as the region of interest in family #1 on chromosome 20 showed negative LOD scores in family #2 and no shared region with high LOD score in both families was obtained. Thus the data from the second family were not taken into account for any further analysis.

I also investigated FOXE-1 gene in family #1. This gene has been reported as the strongest genetic risk marker of sporadic PTC in European populations [178]. Subsequent studies focussed on the length of poly-alanine tract in the gene, which varies between (9Ala) and (16Ala), with 14Ala being the most common and considered as wild type (WT), while 16Ala is the second most common. The tract length was shown to be in linkage disequilibrium with SNP rs1867277 with FOXE-1 (16Ala) being associated with the allele for PTC. Functional studies in vitro showed that FOXE-1 (16Ala) was transcriptionally reduced compared with FOXE-1(14Ala) [347]. I have screened all affected family members and some unaffected for the poly-alanine tract, but not found any communal picture among the affected only. The results of affected were as follow: (I-2) with 14/14, (II-2) with 14/14, (III-2) with 14/14, (III-5) with 14/16, (IV-1) with 14/14, (IV-3) with 14/14 & (IV-6) with 14/14, while (IV-7) has not been tested as she was unaffected at that time. The unaffected have shown (III-1) with 14/14, (III-3) with 14/16 & (III-4) with 14/14, (Figure App.6.1, in Appendix D1a).

Furthermore rs965513 (located at ~100,556,109 bp) was also reported to be associated with thyroid cancer. I have investigated both SNPs, rs965513 which is located close to FOXE-1 gene and shown to be associated with increased risk of both PTC and FTC [91], and rs1867277 (located at ~100,615,914 bp) which is located in the FOXE-1 gene upstream to the fork-head domain and showed strong evidence of association with PTC susceptibility and linkage disequilibrium with FOXE-1 (16Ala) as mentioned above [347]118. The functional study of the latter has shown effect on FOXE-1 transcription [179]. I have studied both SNPs region in the GWLA data of family #1, by looking at the SNPs in 10K chip, flanking that region and found rs953199 at 99,522,797 bp and rs1407501 at 100,638,262 bp, (Figure App.6.2, in Appendix D1b). No linkage with either SNPs has been observed as npl LOD

score was -0.16 and dominant LOD score was -2.4, the latter suggests rejection of the hypothesis of linkage as LOD < -2 (described in 3.2.2). The same investigation was performed on the combined data for families #1 & #2; npl LOD score was -0.1, while dominant LOD score was -1.3, suggesting no linkage.

I have also genotyped the above 2 published SNPs, by designing flanking primers and direct sequencing the region, using DNA samples from all affected members in family #1 and some unaffected as control. The risk allele [A] of rs1867277, which show high frequency in PTC and was associated with FOXE-1 (16Ala) [347], was not found in the affected members of family #1. Same results obtained with risk allele A of rs965513 which showed strong association with thyroid cancer in GWAS [91].

The large region identified with a high LOD score in family #1 on chromosome 20 encouraged me to perform copy number variation (CNV) analysis, but only on the DNA of the index patient (IV-6) due to expense. I have used Illumina chip (Human 660W-Quad), which contains 660K markers, of both SNPs & NP types. That has identified an intronic deletion of ~900 bp in one copy of phospholipase-C β1 (*PLCβ1*) gene, (the first in the 10 gene list at the region of interest), by 14 markers of this chip. I have used the probe sequences of one of these markers to identify the exact region of the deletion and design PCR primers flanking and within the deletion. A standard PCR and Sanger sequencing have shown one copy of full-length amplicon and one deleted allele in all affected and obligate carrier (II-3), but only full-length allele was shown in the unaffected members of family #1 (mainly the founders). The output of haplotype analysis in family data was in accordance with this finding, i.e. all individuals carrying the assumed disease allele C have the one copy deletion, but the other haven't. Thus I came to a conclusion that there is a possibility of a chance finding of a segregation of this CNV (deletion) in family #1. The sequence result (with primers flanking the deletion) was matching that upstream and downstream of the deleted region but with an additional 'ATAA' inserted at the junction. That has indicated a possible 'cut-and-paste' event arising from activity of a transposable element (jumping genes). Remarkably, 11kb transposon cluster has been reported in PLCβ1 gene recently [348], at $20p^{12.3}$ (between 8,305,141-8,316,644 bp), which is immediately upstream of the 3.7Mbp

section on chromosome 20 displaying the non-parametric LOD score of 3.01 in the current study. The PLCβ1 gene is located between 8,061,299bp and 8,730,801bp.

Furthermore, I and Professor Marian have thought to screen known cases of MNG and PTC for the same deletion, by the same standard PCR, so collaborated with Thyroid team in Royal North Shore Hospital, Sydney (Dr. Martyn Bullock & Dr. Roderick J Clifton-Bligh). The screen was performed on genomic DNA from thyroid tissue of 70 patients (European) who underwent surgery for non-familial PTC, and on an additional 81 from non-familial MNG (European). Samples from PTC patient have not shown any deletion; however, 4 of the 81 MNG (all unrelated) were heterozygous for the deletion with the same ATAA insertion in the direct sequencing at the junction. Thus this variant could be described as an InDel, since it includes a combination of deletion and insertion.

To confirm that the one copy deletion is rare in the population, I have screened 105 DNA samples from unrelated individuals who needed their thyroid hormone levels checked for psychiatric reasons and thus considered as normal Caucasians. One out of the 105 showed the same deletion (heterozygous) and it was an euthyroid woman in her forties. That has given a 1% frequency of the deletion in a normal population. Realising that 105 DNA samples are not enough to confirm normal population frequency, I referred to the database for genomic variants webpage: (http://projects.tcag.ca/variation/), in which several deletions (> 5Kbp) were reported at that region. Only one was similar (~1Kbp) as in the family #1. That deletion was reported on 2 out of 180 Caucasian giving a 1% frequency and not found in more than 450 people of other ethnicities[311]. That has revealed 3 in 285 Caucasians harbouring the deletion, suggesting that it is rare. Furthermore, putting in mind the prevalence of goitre the 3 identified with the deletion may also have MNG. Evaluation of the frequency of the deletion in general population vs. in MNG cases, gives a X^2 value of 5.076 (1 degree of freedom), p= 0.024 (two-tailed). Since the same deletion was found in other MNG patients that has re-assured us that the right choice had been made. Very few analyses have performed study on a large population after testing a single family. One such [83], has tested a large family with MNG/PTC and identified a locus on chromosome 14q (MNG1), but study of 37 small pedigrees revealed a very small proportion was attributable to MNG1. Another study [87], tested a large family with recurrent PTC and obtained a LOD score of 1.28 on 2q21, then

studied 80 pedigrees and obtained a LOD score of 3.19. Out of the 80 pedigrees 17 family were selected (based on the presence of at least one case of follicular variant of PTC), and studied again, a LOD score of 4.17 obtained.

Furthermore, MNG/PTC investigation has been performed on a big cohort including 20 unrelated PTC patients with a history of MNG (MNG/PTC), other 284 PTC patient but without MNG history and 349 healthy controls. TITF-1 (NKX2.1) germline mutation (A339V) was detected in 4 out of the 20 PTC patients with MNG history, but not found among the 349 healthy controls nor among the 284 PTC patients with no MNG history. These 4 mutant patients showed more advanced tumours than all other patient (300) in the same study. Overexpression of A339V TTF-1 in rat cells was associated with increased cell proliferation compared with WT, in addition to TSH independent growth and impaired transcription of the TG, TSHR, and PAX-8. This germline mutation was dominantly inherited in two families, with some members bearing the mutation affected with MNG, associated with either PTC or colon cancer [349].

I also investigated how the deletion, or the novel junction sequence it generates, might affect the thyroid gland so I searched for micro RNA (miRNA), in both sequences against miRBase software [350], but did not find any significant hits. I have collaborated with the creators of miRBase software; Faculty of Life Sciences, University of Manchester where Dr Antonio Marco searched for hairpin structure in the deleted region (with and without ATAA) using RNAL fold program, but didn't see any hairpin likely to be a microRNA [351].

Goitre prevalence in women is 2 to 10 fold higher than in men in common with MNG, PTC and all other thyroid disease [345, 352]. That made me to think about a relation between the deletion and any female sex hormones; estrogens. In addition to their effect on reproductive cycles, they have shown to activate G-protein coupled receptors [353] besides, enhancing normal thyroid cell growth and thyroid cancer cells [346]. Estrogens like other steroid hormones, after entering the cell bind (via specific binding sites) and activate estrogen receptors (ER) which control several genes expression and are of two types; ERα and ERβ. ERα agonist is shown to be promoting thyroid cancer cells proliferation, while ERβ agonist reduces this proliferation. Thus thyroid proliferation increases when ERβ is down-regulated [354]. A study [355] mapped ERα binding sites in human breast cancer cell line and

identified > 1,200 ERα binding sites in the human genome. Only 5% of the mapped ERα binding sites were located within promoters and the majority were mapped to intronic or distal locations. This information has encouraged me to collaborate with Dr Davy Kavanagh, who has been introduced by my co supervisor Dr Marian Hamshere. He explored the deleted region in PLCβ1 and identified a presence of binding site for the ERα within the deletion, using data from the Encyclopaedia of DNA elements (ENCODE [348]).

I also conducted experiments on thyroid tissue from three affected family members who underwent thyroidectomy. I have extracted RNA from their tissues aiming to study any possible mechanisms by which the deletion might affect PLCβ1 function. I hypothesised that the deletion in intron 3 may leads to exon skipping, before the deletion (exon 3) or after (exon 4). To investigate that, I have designed exonic primers to amplify exon 3 & 4 using cDNA from the above tissues and standard PCR and direct sequencing; both exons were present, with no additional bands that could indicate transcripts missing exons. However this method was not enough to calculate the transcription level. Thus two of my colleague in my department; (Dr Fiona Grennan-Jones & Dr Lei Zhang), have repeated the exon skipping investigation, using QPCR analysis, and calculated the transcription levels, but did not detect skipping of exons 3 or 4 (flanking the deletion). They have also investigated mRNA expression of two main PLCβ1 isoforms by QPCR using the same cDNA and found higher PLCβ1 transcript levels ($p< 0.02$) in thyroids from the three affected family members, compared with normal controls. This could give a clue that the germ-line deletion in PLCβ1 may lead to over expression and that leads to MNG/PTC in the family. Higher expression of PLCβ1 has been reported in small cell lung carcinoma[341] and increased PLC enzyme activity has also been reported in thyroid neoplasms[338].

Furthermore the deletion is in the middle of intron 3 (with 58kb before and 197Kb after the deletion), of PLCβ1 gene which encodes phosphoinositide-specific enzyme, that play a main role in G protein-coupled receptor signalling to the MAPK cascade, via IP3, DAG, PKC and BRAF [343]. BRAF is a downstream element of the Gq/ PLCβ1/MAPK signalling pathway. This pathway was shown to be essential in the growth of the gland as seen by Kero et al [356], who studied Gqα/Gq11deficient mice and reported absence of TSH proliferative effects and resistance to goitre formation when on a goitrogenic diet. The cascade

involvement has also been reported in thyrocyte proliferation[176]. Several recent studies have reported BRAF activating mutations mainly BRAFV600E in many PTC [120]. I have screened genomic DNA of all affected members in family #1 for BRAFV600E mutation, and was excluded in this family.

PLCβ1 is not thyroid specific but may have an impact on some feature of thyroid biology which is, e.g. H_2O_2 generation. Production of H_2O_2 has been shown to be activated by PLC and Ca^{2+}/DAG, whilst the cAMP cascade was shown to inhibit H_2O_2 generation [75]. I^- oxidation by TPO, requires H_2O_2, which is generated by the calcium dependent DUOX family, that contains calcium-binding sites [6]. Higher concentrations of H_2O_2 can cause DNA oxidation and damage, leading to mutagenesis and apoptosis [73] and has been suggested as one mechanism contributing to the formation of endemic goitres in human [74]. Thus if PLCβ1 activity is increased due to the deletion, it may lead to higher levels of H_2O_2 and could explain why the PLCβ1 deletion has an effect only on thyroid, despite PLCβ1 being widely expressed in all cells/tissues of the body.

In some studies [269], after GWLA using SNPs, a further scan used microsatellite markers to narrow the candidate region, then sequenced all genes in the region and checked for loss of heterozygosity. In my study further scanning using microsatellites was not required since the candidate region was small enough to eliminate that screen; however in-depth sequencing of the 10 genes could be a future option. This could be performed using next generation sequencing or whole exome sequencing which provides data on both exons and introns. Sanger sequencing of all exons using standard PCR could be an option, but with recent studies confirming several roles of noncoding DNA, data of exons only could miss important information.

The limitation of this study is that it is not showing the percentage of the European population, this deletion represents. However, this will be clarified by investigating more affected individuals and controls. The work in this part is already started via screening more European samples by collaborating with a thyroid groups in German (Prof. Dagmar Führer) and in Italy (Prof. Rossella Elisei). Around 400 known MNG cases from areas with varying levels of iodine intake are targeted to be screened for the same deletion. That will provide an

idea of deletion frequency in patients with MNG and give a chance to compare with healthy population (controls) which will also be screened in the future.

What makes my study unique, is that I have found the deletion in one family then in another 4 MNG unrelated cases, while almost all previous studies have investigated one family then worked on other families which make their finding more family based, while this study started with familial MNG/PTC but was found to be relevant to sporadic disease too.

In conclusion, the deletion I identified in familial MNG also occurs in sporadic MNG, and may provide a biomarker to identify MNG patients most likely to progress to PTC. It predisposes to goitre formation, possibly by increasing PLCβ1 transcription and thereby activating the PKC and MAPK pathways.

Appendices

(Appendix A)

(Appendix A1a)

Using PLINK software:

In order to use PLINK a Map File has been created with 3 columns of data; the list of all markers, chromosomal ID of each marker and position in base pair (bp). In addition disease status column is required in the PedFile for all individuals, ending with 6 columns defining the family structure for each individual.

PLINK to remove SNPs with unknown location:

A list of all these 96 SNPs has been created and removed using --exclude command as follow:

--ped PedFile_NoCall.txt

--map MapFile_AffySNPsID.txt

--map3

--exclude List_96Unk.txt (described in 3.3.1)

--recode (described in 3.3.1)

--out Data_NoCall_No96Unk (described in 3.3.1)

(Appendix A1b)

PedCheck Software:

In order to use this software, disease status column has been removed from the PedFile for all individuals, ending with 5 columns defining the family structure for each individual instead of 6. The genotype format has been changed from A & B to be as two consecutive integers (1

and 2). For Mendelian error check, a map file with SNPs ID only was used, so the output will be as SNPs list with Mendelian error.

Commands: -p PedFile_NoCall_No96Unk_1_2.txt –n MapFile_96out_ID_ONLY.txt -4

Using PLINK to remove the 200 SNPs showing Mendelian errors in PedCheck:

A list of the 200 SNPs has made and removed from the data as follow:

--ped PedFile_NoCall_No96Unk.txt

--map MapFile_AffySNPsID_96out.txt

--map3

--exclude 200SNPs_PedcheckError.txt

--recode

--out Data_PedcheckOK200

(Appendix A1c)

PLINK for removing the 5 SNPs in HH file

--ped Data_PedcheckOK200.ped

--map Data_PedcheckOK200.map

--map3

--exclude Data_PedcheckOK200.HH

--recode

--out Data_PedcheckOK200_Nohh

PLINK for Mendelian Error check:

--ped Data_PedcheckOK200_Nohh.ped

--map Data_PedcheckOK200_Nohh.map

--map3

--mendel

(Appendix A1d)

PLINK for Convert 1 2 to A C G T:

--ped Data_PedcheckOK200_Nohh.ped

--map Data_PedcheckOK200_Nohh.map

--map3

--update-alleles Convert 1 2 to A C T G 2013.txt

--recode

--out data_ACGT

The output has been confirmed by comparing the data of 5 SNPs (extracted) before and after conversion, as part of Quality Control

(Appendix A1e)

PLINKs for convert –ve starnds to +ve:

--ped data_ACGT.ped

--map data_ACGT.map

--flip All -ve strand SNP ID.txt

--recode

--out Data_converted_flipped

The output has been confirmed by comparing the data of 5 SNPs (extracted) before and after flipping, as part of Quality Control

(Appendix A1f)

PLINK for Merging with HapMap:

Removing the 22 SNPs which showed errors during data merge with HapMap: A list of the 22 SNPs was made and removed by PLINK using --exclude command (as described in appendix A1c).

Merging data:

A standard PLINK-style binary genotype file of HapMap, has been used (.bim is the SNP map, .bed is the binary PedFile, and .fam describes the family relationships among your

samples). A Notepad file with the name of all HapMap file (.bed .bim .fam of each population in one line) has been created and saved as (MergeList.txt). Family data merged and the SNPs listed in the map file have been extracted as follow:

--ped dataStrand_22out.ped

--map dataStrand_22out.map

--map3

--merge-list MergeList.txt

--extract SNPsList_Strand_22out.txt

--recode

--out Data_HapMap

The output has been confirmed by comparing the data of 5 SNPs (extracted) before and after merge with HapMap, as part of Quality Control.

(Appendix A1g)

Changing SNPs chromosomal position:

Base pair positions changed to cM using Excel work sheets and the below command: [=IF(VLOOKUP(B1,D:E,1)=B1,VLOOKUP(B1,D:E,2),"")]

Column B is for SNPs ID from map file. Columns D & E are for SNPs ID and cM position (from Marshfield map) respectively. The above command was typed in F column. If the cM position of SNPs in B1 column was available anywhere in column E that value of cM will be typed in F column, otherwise will be kept empty. Same will occur with all SNPs in B column, ending with cM position for all SNPs in map file. That is only for the SNPs available in Marshfield map.

(Appendix A1h)

Merlin for npl Analysis:

-d DataFile_9111.txt -p Data_9111_1234Formate_No-9_4Affected_No8.txt

-m MapFile_Data_9111_cM.txt

--npl --exp (described in Appendix B1b)

--swap --smallSwap --quiet (described in Appendix B1b)

--markerNames --pdf --prefix npl_2 > utput_npl_2 (described in Appendix B1b)

(Appendix A1i)

Merlin commands for Dominant Analysis:

-d DataFile_9111_AGE.txt

-p Data_9111_1234Formate_No-9_4Affected_No8_AGE.txt

-m MapFile_Data_9111_cM.txt

--model DOMINANT_Age_12.txt (described in Appendix B1b)

--swap --smallSwap --quiet (described in Appendix B1b)

--markerNames --pdf --prefix Dom > utput_Dom (described in Appendix B1b)

(Appendix A1j)

Merlin commands for Recessive Analysis:

-d DataFile_9111_AGE.txt

-p Data_9111_1234Formate_No-9_4Affected_No8_AGE.txt

-m MapFile_Data_9111_cM.txt

--model RECESSIVE_140509_12.txt (described in Appendix B1b)

--swap --smallSwap --quiet (described in Appendix B1b)

--markerNames --pdf --prefix Recess > utput_Recessi (described in Appendix B1b)

(Appendix A1k)

Merlin commands for Haplotype Analysis:

-d DataFile_9111.txt -p Data_9111_1234Formate_No-9_4Affected_No8.txt

-m MapFile_Data_9111_cM.txt

--npl (described in Appendix B1b)

--best (described in 3.3.3.1)

--swap --smallSwap --quiet (described in Appendix B1b)

--markerNames --pdf --prefix Haplo > utput_Haplo (described in Appendix B1b)

(Appendix B)

(Appendix B1a)

Data Preparation:

Change disease status (Phenotype) format for HapMap data: HapMap data use a different phenotype labelling format (-9 rather than 0 for unknown), which was not accepted by Merlin. Therefore phenotype format in HapMap data has been changed from -9 to 0.

(Appendix B1b)

Merlin Commands

A command “--pdf", can be used to obtain graphical output data for each chromosome in a PDF file. A command “--prefix label > output_label.out” can be used to create output files, a PDF for graphical output data for each chromosome and a note file with lists of markers or positions and the LOD score for each marker. To perform the analysis on the basis of markers (not positions), “--markerNames” command can be used. A command “--quiet" can be used to avoid inclusion of unnecessary details on the screen or in the output file.

(Appendix C)

(Appendix C1a);

The output data of the linkage analyses on chromosome 20 (the whole chromosome).

chromosome	SNP	Position (Mb)	Position (cM)	npl	Dominant
20	rs722829	0.12	0.49	1.65	1.69
20	rs1342841	0.16	0.68	1.66	1.70
20	rs2317024	0.69	3.52	1.70	1.75
20	rs1923876	1.18	6.69	1.95	1.85
20	rs2013961	2.27	8.55	2.29	1.91
20	rs1109010	2.41	8.71	2.29	1.92
20	rs3848810	3.52	9.84	2.34	1.96
20	rs910652	3.68	10.28	2.35	1.98
20	rs674110	3.68	10.28	2.35	1.98
20	rs910952	4.14	11.47	2.37	2.02
20	rs2422925	4.25	11.69	2.37	2.03
20	rs3904872	4.26	11.71	2.37	2.03
20	rs3904864	4.33	11.84	2.37	2.04
20	rs1411296	6.44	17.96	2.08	2.03
20	rs723131	6.53	18.07	2.08	2.03
20	rs2224191	6.76	18.35	2.08	2.03
20	rs723511	6.79	18.39	2.08	2.03
20	rs952793	6.92	18.55	2.07	2.03
20	rs2149642	6.96	18.61	2.07	2.03
20	rs717074	7.12	18.82	2.07	2.03

20	rs720010	7.17	18.96	2.07	2.03
20	rs2009752	7.19	19.00	2.08	2.03
20	rs724084	7.22	19.09	2.09	2.03
20	rs2224053	7.25	19.17	2.10	2.03
20	rs2223276	7.83	20.69	2.57	2.03
20	rs2250711	8.20	22.19	2.84	2.03
20	rs771943	8.36	23.02	2.94	2.04
20	rs2223539	8.49	23.69	3.01	2.04
20	rs2327088	8.71	23.93	3.01	2.04
20	rs2143267	8.76	23.96	3.01	2.04
20	rs3848831	8.80	23.98	3.01	2.04
20	rs756374	8.89	24.04	3.01	2.04
20	rs724089	9.00	24.11	3.01	2.04
20	rs953021	9.90	24.65	3.01	2.04
20	rs725565	9.95	24.68	3.01	2.04
20	rs2327282	10.38	27.42	3.01	2.03
20	rs973542	10.64	28.17	3.01	2.04
20	rs2327302	10.66	28.23	3.01	2.04
20	rs763383	10.85	28.71	3.01	2.04
20	rs1028846	10.96	29.01	3.01	2.03
20	rs726867	11.18	30.67	3.01	2.03
20	rs3887413	11.75	31.90	3.01	2.04
20	rs2206798	11.85	32.09	3.01	2.04
20	rs720489	11.90	32.20	3.01	2.04
20	rs729552	12.13	32.42	3.01	2.04
20	rs1099620	12.74	32.80	3.01	2.04

20	rs2327450	12.93	32.91	3.01	2.04
20	rs721243	13.17	33.06	3.01	2.04
20	rs2327790	13.46	33.25	3.01	2.04
20	rs1358766	13.86	33.50	3.01	2.04
20	rs724820	15.02	34.23	3.01	2.03
20	rs763659	15.74	37.09	3.01	2.03
20	rs726207	15.96	37.41	3.01	2.03
20	rs718610	16.21	37.82	3.01	2.03
20	rs2144883	16.27	37.97	3.01	2.03
20	rs1080954	16.45	38.38	3.01	2.04
20	rs720592	16.45	38.38	3.01	2.04
20	rs720593	16.45	38.38	3.01	2.04
20	rs2328024	16.60	38.72	3.01	2.04
20	rs724053	16.93	39.08	3.01	2.04
20	rs956110	17.02	39.18	3.01	2.04
20	rs2010316	17.08	39.24	3.01	2.04
20	rs2010307	17.08	39.24	3.01	2.04
20	rs1078530	17.08	39.24	3.01	2.04
20	rs726256	17.40	39.60	3.01	2.04
20	rs16823	17.74	39.96	3.01	2.04
20	rs1073052	17.79	40.02	3.01	2.04
20	rs2024673	18.12	40.37	3.01	2.04
20	rs2328245	18.14	40.40	3.01	2.04
20	rs2328293	18.31	40.61	3.01	2.04
20	rs1535487	18.94	41.86	3.01	2.04
20	rs2328361	19.03	42.05	3.01	2.04

20	rs1074615	19.13	42.25	3.01	2.04
20	rs2328384	19.37	42.72	3.01	2.04
20	rs2328410	19.49	42.98	3.01	2.04
20	rs2328411	19.50	42.98	3.01	2.04
20	rs2328412	19.50	42.98	3.01	2.04
20	rs725862	19.58	43.14	3.01	2.04
20	rs720436	19.60	43.19	3.01	2.04
20	rs1028434	20.11	44.20	2.43	1.47
20	rs928066	20.26	44.51	1.54	0.68
20	rs721424	20.29	44.56	0.35	0.08
20	rs1074440	20.48	44.95	0.35	0.08
20	rs2038383	20.83	45.64	-1.89	0.08
20	rs755963	20.93	45.82	0.35	0.08
20	rs1888610	22.85	48.08	-1.66	0.08
20	rs717756	22.92	48.13	-1.97	0.08
20	rs2007743	22.99	48.18	0.35	0.08
20	rs1112819	23.27	48.38	-2.41	0.08
20	rs999072	23.30	48.40	-2.41	0.08
20	rs726217	23.53	48.57	0.35	0.08
20	rs2254635	23.57	48.60	0.35	0.08
20	rs2145231	23.57	48.60	0.27	0.08
20	rs3843776	23.94	48.67	0.35	0.08
20	rs3843777	23.94	48.67	0.35	0.08
20	rs3848799	23.94	48.67	0.35	0.08
20	rs761863	24.03	48.69	0.35	0.08
20	rs487665	24.12	48.70	0.35	0.08

20	rs722834	24.37	48.74	0.35	0.08
20	rs2387577	25.00	48.85	0.35	0.08
20	rs2207631	25.00	48.85	0.35	0.08
20	rs2387733	25.13	48.87	-2.45	0.08
20	rs1474945	29.31	49.58	-1.18	0.09
20	rs721220	29.47	49.61	-1.18	0.09
20	rs725478	31.98	50.46	0.35	0.08
20	rs2378132	32.31	50.51	0.33	0.08
20	rs819145	32.33	50.51	0.33	0.08
20	rs725908	33.43	50.68	0.27	0.07
20	rs3850528	34.17	50.79	0.23	0.06
20	rs1073768	35.31	51.16	0.07	0.04
20	rs724782	35.68	51.29	0.07	0.02
20	rs910760	35.93	51.46	0.06	-0.01
20	rs1569608	37.45	54.03	0.00	-1.88
20	rs718093	37.66	54.37	0.00	-1.88
20	rs718092	37.66	54.37	0.00	-1.88
20	rs1408871	37.80	54.61	0.00	-1.87
20	rs728331	38.42	55.60	-0.01	-1.87
20	rs723080	38.79	56.20	-0.02	-1.88
20	rs722728	38.87	56.30	-0.02	-1.88
20	rs2143233	39.66	56.76	-0.02	-1.88
20	rs2067084	40.17	57.42	-0.02	-1.87
20	rs727336	40.32	57.64	-0.02	-1.87
20	rs986831	40.59	58.05	-0.02	-1.88
20	rs208179	40.87	58.46	-0.02	-1.89

20	rs1397854	40.90	58.51	-0.02	-1.90
20	rs1397853	40.90	58.51	-0.02	-1.90
20	rs2868198	42.80	61.35	-0.04	-2.02
20	rs2157361	43.19	61.98	-0.05	-2.05
20	rs382515	43.81	63.25	-0.07	-2.13
20	rs3848731	44.35	64.35	-0.10	-2.21
20	rs3908612	44.42	64.49	-0.10	-2.22
20	rs1405567	44.62	64.89	-0.11	-2.25
20	rs430114	44.62	64.89	-0.11	-2.25
20	rs928486	44.89	65.69	-0.11	-2.25
20	rs1815969	45.97	68.90	-0.11	-2.25
20	rs756529	47.44	72.80	-0.05	-2.08
20	rs951497	47.76	73.31	-0.05	-2.06
20	rs967305	47.89	73.51	-0.04	-2.05
20	rs941798	48.60	74.57	-0.04	-2.02
20	rs956298	48.73	74.71	-0.04	-2.02
20	rs2208006	50.14	77.12	-0.04	-2.01
20	rs726248	50.27	77.35	-0.04	-2.02
20	rs2904155	50.40	77.59	-0.04	-2.02
20	rs2078577	50.44	77.67	-0.04	-2.02
20	rs727662	50.71	77.92	-0.04	-2.02
20	rs715434	50.71	77.93	-0.04	-2.02
20	rs715433	50.71	77.93	-0.04	-2.02
20	rs241792	50.80	77.99	-0.04	-2.02
20	rs1936973	51.17	78.28	-0.04	-2.01
20	rs916954	51.29	78.71	-0.04	-2.01

20	rs465402	52.10	82.02	-0.04	-2.01
20	rs290409	52.14	82.18	-0.04	-2.01
20	rs1418926	52.26	82.72	-0.04	-2.02
20	rs1293153	52.38	83.27	-0.04	-2.02
20	rs1293151	52.38	83.27	-0.04	-2.02
20	rs1407043	52.44	83.51	-0.04	-2.02
20	rs2870357	52.75	84.01	-0.04	-2.02
20	rs2065042	52.88	84.21	-0.04	-2.02
20	rs721437	53.75	85.11	-0.04	-2.01
20	rs869138	55.06	86.45	-0.04	-2.02
20	rs869136	55.06	86.45	-0.04	-2.02
20	rs911901	55.39	86.78	-0.04	-2.02
20	rs718307	55.64	87.32	-0.04	-2.02
20	rs728504	55.77	88.09	-0.04	-2.02
20	rs728506	55.77	88.09	-0.04	-2.02
20	rs967990	55.77	88.09	-0.04	-2.02
20	rs1889331	56.03	89.63	-0.04	-1.97
20	rs717671	56.36	91.34	-0.04	-1.95
20	rs271950	57.45	96.16	-0.04	-2.04
20	rs1857285	57.54	96.37	-0.04	-2.05
20	rs3905228	57.54	96.37	-0.04	-2.05
20	rs3906935	57.54	96.37	-0.04	-2.05
20	rs1116089	57.65	96.61	-0.04	-2.03
20	rs1412197	58.11	97.64	-0.08	-2.04
20	rs2144880	58.72	99.00	-0.20	-2.06
20	rs2017286	58.90	99.41	-0.21	-2.08

20	rs817710	59.18	100.03	-0.21	-2.13
20	rs947737	59.22	100.11	-0.21	-2.14
20	rs2427077	59.33	100.36	-0.21	-2.16
20	rs1892326	59.38	100.48	-0.21	-2.18
20	rs944257	59.68	101.15	-0.21	-2.25

Table Appen 6.1; The output data of the linkage analyses on chromosome 20 (the whole chromosome). First column (chromosome) shows chromosome name, second (SNP) single nucleotide polymorphisms (SNPs) ID, third position of SNPs in mega bases (Mb), and fourth the position in centiMorgan (cM). Fifth column is npl multipoint maximum LOD scores and sixth for dominant multipoint maximum LOD scores.

(Appendix C1b);

Haplotype study of some members of family #1, across chromosome 20 (whole chromosome).

				Haplotype			Haplotype			Haplotype		
Chromosome	SNP	Position Mb	Position cM	III-2 (II-2, II-1)			IV-2 (III-2, III-1)			III-3 (II-2, II-1)		
20	rs722829	0.12	0.49	C	:	G	C	:	I	K	:	H
20	rs1342841	0.16	0.68	C	:	G	C	:	I	K	:	H
20	rs2317024	0.69	3.52	C	|	G	C	|	I	K	|	H
20	rs1923876	1.18	6.69	C	|	G	G	/	I	K	\	G
20	rs2013961	2.27	8.55	C	|	G	G	|	I	K	|	G
20	rs1109010	2.41	8.71	C	|	G	G	|	I	K	|	G
20	rs3848810	3.52	9.84	C	:	G	G	:	I	K	:	G
20	rs910652	3.68	10.28	C	|	G	G	|	I	K	|	G
20	rs674110	3.68	10.28	C	|	G	G	|	I	K	|	G
20	rs910952	4.14	11.47	C	:	G	G	:	I	K	:	G
20	rs2422925	4.25	11.69	C	|	G	G	|	I	K	|	G
20	rs3904872	4.26	11.71	C	|	G	G	|	I	K	|	G
20	rs3904864	4.33	11.84	C	|	G	G	|	I	K	|	G
20	rs1411296	6.44	17.96	C	|	G	G	|	I	K	|	G
20	rs723131	6.53	18.07	C	|	G	G	|	I	K	|	G
20	rs2224191	6.76	18.35	C	|	G	G	|	I	K	|	G
20	rs723511	6.79	18.39	C	|	G	G	|	I	K	|	G
20	rs952793	6.92	18.55	C	|	G	G	|	I	K	|	G
20	rs2149642	6.96	18.61	C	|	G	G	|	I	K	|	G
20	rs717074	7.12	18.82	C	|	G	G	|	I	K	|	G

20	rs720010	7.17	18.96	C	|	G	G	|	I	K	|	G
20	rs2009752	7.19	19.00	C	:	G	G	:	I	K	:	G
20	rs724084	7.22	19.09	C	|	G	G	|	I	K	|	G
20	rs2224053	7.25	19.17	C	|	G	G	|	I	K	|	G
20	rs2223276	7.83	20.69	C	:	G	G	:	I	K	:	G
20	rs2250711	8.20	22.19	C	|	G	G	|	I	K	|	G
20	rs771943	8.36	23.02	C	|	G	G	|	I	K	|	G
20	rs2223539	8.49	23.69	C	|	G	G	|	I	K	|	G
20	rs2327088	8.71	23.93	C	|	G	G	|	I	K	|	G
20	rs2143267	8.76	23.96	C	|	G	G	|	I	K	|	G
20	rs3848831	8.80	23.98	C	|	G	G	|	I	K	|	G
20	rs756374	8.89	24.04	C	|	G	G	|	I	K	|	G
20	rs724089	9.00	24.11	C	|	G	G	|	I	K	|	G
20	rs953021	9.90	24.65	C	|	G	G	|	I	K	|	G
20	rs725565	9.95	24.68	C	|	G	G	|	I	K	|	G
20	rs2327282	10.38	27.42	C	|	G	G	|	I	K	|	G
20	rs973542	10.64	28.17	C	|	G	G	|	I	K	|	G
20	rs2327302	10.66	28.23	C	|	G	G	|	I	K	|	G
20	rs763383	10.85	28.71	C	|	G	G	|	I	K	|	G
20	rs1028846	10.96	29.01	C	|	G	G	|	I	K	|	G
20	rs726867	11.18	30.67	C	|	G	G	|	I	K	|	G
20	rs3887413	11.75	31.90	C	|	G	G	|	I	K	|	G
20	rs2206798	11.85	32.09	C	|	G	G	|	I	K	|	G
20	rs720489	11.90	32.20	C	|	G	G	|	I	K	|	G
20	rs729552	12.13	32.42	C	|	G	G	|	I	K	|	G
20	rs1099620	12.74	32.80	C	|	G	G	|	I	C	/	G

20	rs2327450	12.93	32.91	C	|	G	G	|	I	C	|	G
20	rs721243	13.17	33.06	C	|	G	C	/	I	C	|	G
20	rs2327790	13.46	33.25	C	|	G	C	|	I	C	|	G
20	rs1358766	13.86	33.50	C	|	G	C	|	I	C	|	G
20	rs724820	15.02	34.23	C	|	G	C	|	I	C	|	G
20	rs763659	15.74	37.09	C	|	G	C	|	I	C	|	G
20	rs726207	15.96	37.41	C	|	G	C	|	I	C	|	G
20	rs718610	16.21	37.82	C	|	G	C	|	I	C	|	G
20	rs2144883	16.27	37.97	C	|	G	C	|	I	C	|	G
20	rs1080954	16.45	38.38	C	|	G	C	|	I	C	|	G
20	rs720592	16.45	38.38	C	|	G	C	|	I	C	|	G
20	rs720593	16.45	38.38	C	|	G	C	|	I	C	|	G
20	rs2328024	16.60	38.72	C	|	G	C	|	I	C	|	G
20	rs724053	16.93	39.08	C	:	G	C	:	I	C	:	G
20	rs956110	17.02	39.18	C	|	G	C	|	I	C	|	G
20	rs2010316	17.08	39.24	C	|	G	C	|	I	C	|	G
20	rs2010307	17.08	39.24	C	:	G	C	:	I	C	:	G
20	rs1078530	17.08	39.24	C	|	G	C	|	I	C	|	G
20	rs726256	17.40	39.60	C	|	G	C	|	I	C	|	G
20	rs16823	17.74	39.96	C	|	G	C	|	I	C	|	G
20	rs1073052	17.79	40.02	C	|	G	C	|	I	C	|	G
20	rs2024673	18.12	40.37	C	|	G	C	|	I	C	|	G
20	rs2328245	18.14	40.40	C	|	G	C	|	I	C	|	G
20	rs2328293	18.31	40.61	C	|	G	C	|	I	C	|	G
20	rs1535487	18.94	41.86	C	|	G	C	|	I	C	|	G
20	rs2328361	19.03	42.05	C	|	G	C	|	I	C	|	G

20	rs1074615	19.13	42.25	C	|	G	C	|	I	C	|	G
20	rs2328384	19.37	42.72	C	|	G	C	|	I	C	|	G
20	rs2328410	19.49	42.98	C	|	G	C	|	I	C	|	G
20	rs2328411	19.50	42.98	C	|	G	C	|	I	C	|	G
20	rs2328412	19.50	42.98	C	|	G	C	|	I	C	|	G
20	rs725862	19.58	43.14	C	|	G	C	|	I	C	|	G
20	rs720436	19.60	43.19	C	|	G	C	|	I	C	|	G
20	rs1028434	20.11	44.20	C	|	G	C	|	I	C	|	G
20	rs928066	20.26	44.51	C	:	G	C	:	I	C	:	G
20	rs721424	20.29	44.56	C	|	G	C	|	I	C	|	G
20	rs1074440	20.48	44.95	C	|	G	C	|	I	C	|	G
20	rs2038383	20.83	45.64	C	|	G	C	|	I	C	|	G
20	rs755963	20.93	45.82	C	|	G	C	|	I	C	|	G
20	rs1888610	22.85	48.08	C	|	G	C	|	I	C	|	G
20	rs717756	22.92	48.13	C	|	G	C	|	I	C	|	G
20	rs2007743	22.99	48.18	C	|	G	C	|	I	C	|	G
20	rs1112819	23.27	48.38	C	|	G	C	|	I	C	|	G
20	rs999072	23.30	48.40	C	|	G	C	|	I	C	|	G
20	rs726217	23.53	48.57	C	|	G	C	|	I	C	|	G
20	rs2254635	23.57	48.60	C	|	G	C	|	I	C	|	G
20	rs2145231	23.57	48.60	C	|	G	C	|	I	C	|	G
20	rs3843776	23.94	48.67	C	|	G	C	|	I	C	|	G
20	rs3843777	23.94	48.67	C	|	G	C	|	I	C	|	G
20	rs3848799	23.94	48.67	C	|	G	C	|	I	C	|	G
20	rs761863	24.03	48.69	C	|	G	C	|	I	C	|	G
20	rs487665	24.12	48.70	C	:	G	C	:	I	C	:	G

20	rs722834	24.37	48.74	C	|	G	C	|	I	C	|	G
20	rs2387577	25.00	48.85	C	|	G	C	|	I	C	|	G
20	rs2207631	25.00	48.85	C	|	G	C	|	I	C	|	G
20	rs2387733	25.13	48.87	C	|	G	C	|	I	C	|	G
20	rs1474945	29.31	49.58	C	|	G	C	|	I	C	|	G
20	rs721220	29.47	49.61	C	|	G	C	|	I	C	|	G
20	rs725478	31.98	50.46	C	|	G	C	|	I	C	|	G
20	rs2378132	32.31	50.51	C	|	G	C	|	I	C	|	G
20	rs819145	32.33	50.51	C	|	G	C	|	I	C	|	G
20	rs725908	33.43	50.68	C	|	G	C	|	I	C	|	G
20	rs3850528	34.17	50.79	C	|	G	C	|	I	C	|	G
20	rs1073768	35.31	51.16	C	|	G	C	|	I	C	|	G
20	rs724782	35.68	51.29	C	:	G	C	:	I	C	:	G
20	rs910760	35.93	51.46	C	|	G	C	|	I	C	|	G
20	rs1569608	37.45	54.03	C	|	G	C	|	I	C	|	G
20	rs718093	37.66	54.37	C	|	G	C	|	I	C	|	G
20	rs718092	37.66	54.37	C	|	G	C	|	I	C	|	G
20	rs1408871	37.80	54.61	C	|	G	C	|	I	C	|	G
20	rs728331	38.42	55.60	C	|	G	C	|	I	C	|	G
20	rs723080	38.79	56.20	C	|	G	C	|	I	C	|	G
20	rs722728	38.87	56.30	C	|	G	C	|	I	C	|	G
20	rs2143233	39.66	56.76	C	|	G	C	|	I	C	|	G
20	rs2067084	40.17	57.42	C	|	G	C	|	I	C	|	G
20	rs727336	40.32	57.64	C	|	G	C	|	I	C	|	G
20	rs986831	40.59	58.05	C	|	G	C	|	I	C	|	G
20	rs208179	40.87	58.46	C	|	G	C	|	I	C	|	G

20	rs1397854	40.90	58.51	C	|	G	C	|	I	C	|	G
20	rs1397853	40.90	58.51	C	|	G	C	|	I	C	|	G
20	rs2868198	42.80	61.35	K	/	G	K	\	J	C	|	G
20	rs2157361	43.19	61.98	K	|	G	K	|	J	C	|	G
20	rs382515	43.81	63.25	K	|	G	K	|	J	C	|	G
20	rs3848731	44.35	64.35	K	|	G	K	|	J	C	|	G
20	rs3908612	44.42	64.49	K	|	G	K	|	J	C	|	G
20	rs1405567	44.62	64.89	K	|	G	K	|	J	C	|	G
20	rs430114	44.62	64.89	K	|	G	K	|	J	C	|	G
20	rs928486	44.89	65.69	K	|	G	K	|	J	C	|	G
20	rs1815969	45.97	68.90	K	|	G	K	|	J	C	|	G
20	rs756529	47.44	72.80	K	|	G	K	|	J	C	|	G
20	rs951497	47.76	73.31	K	|	G	K	|	J	C	|	G
20	rs967305	47.89	73.51	K	|	G	K	|	J	C	|	G
20	rs941798	48.60	74.57	K	|	G	K	|	J	C	|	G
20	rs956298	48.73	74.71	K	|	G	K	|	J	C	|	G
20	rs2208006	50.14	77.12	K	|	G	K	|	J	K	/	G
20	rs726248	50.27	77.35	K	|	G	K	|	J	K	|	G
20	rs2904155	50.40	77.59	K	|	G	K	|	J	K	|	G
20	rs2078577	50.44	77.67	K	|	G	K	|	J	K	|	G
20	rs727662	50.71	77.92	K	|	G	K	|	J	K	|	G
20	rs715434	50.71	77.93	K	|	G	K	|	J	K	|	G
20	rs715433	50.71	77.93	K	|	G	K	|	J	K	|	G
20	rs241792	50.80	77.99	K	|	G	K	|	J	K	|	G
20	rs1936973	51.17	78.28	K	:	G	K	:	J	K	:	G
20	rs916954	51.29	78.71	K	|	G	K	|	J	K	|	G

20	rs465402	52.10	82.02	K	|	G	K	|	J	K	|	G
20	rs290409	52.14	82.18	K	|	G	K	|	J	K	|	G
20	rs1418926	52.26	82.72	K	|	G	K	|	J	K	|	G
20	rs1293153	52.38	83.27	K	|	G	K	|	J	K	|	G
20	rs1293151	52.38	83.27	K	|	G	K	|	J	K	|	G
20	rs1407043	52.44	83.51	K	|	G	K	|	J	K	|	G
20	rs2870357	52.75	84.01	K	|	G	K	|	J	K	|	G
20	rs2065042	52.88	84.21	K	|	G	K	|	J	K	|	G
20	rs721437	53.75	85.11	K	:	G	K	:	J	K	:	G
20	rs869138	55.06	86.45	K	|	G	K	|	J	K	|	G
20	rs869136	55.06	86.45	K	|	G	K	|	J	K	|	G
20	rs911901	55.39	86.78	K	|	G	K	|	J	K	|	G
20	rs718307	55.64	87.32	K	|	G	K	|	J	K	|	G
20	rs728504	55.77	88.09	K	|	G	K	|	J	K	|	G
20	rs728506	55.77	88.09	K	|	G	K	|	J	K	|	G
20	rs967990	55.77	88.09	K	|	G	K	|	J	K	|	G
20	rs1889331	56.03	89.63	K	|	G	K	|	J	K	|	G
20	rs717671	56.36	91.34	K	|	G	K	|	J	K	|	G
20	rs271950	57.45	96.16	K	|	G	K	|	J	K	|	G
20	rs1857285	57.54	96.37	K	|	G	K	|	J	K	|	G
20	rs3905228	57.54	96.37	K	|	G	K	|	J	K	|	G
20	rs3906935	57.54	96.37	K	|	G	K	|	J	K	|	G
20	rs1116089	57.65	96.61	K	|	G	K	|	J	K	|	G
20	rs1412197	58.11	97.64	K	|	G	K	|	J	K	|	G
20	rs2144880	58.72	99.00	K	|	G	K	|	J	K	|	G
20	rs2017286	58.90	99.41	K	|	G	K	|	J	K	|	G

20	rs817710	59.18	100.03	K	|	G	K	|	J	K	|	G
20	rs947737	59.22	100.11	K	:	G	K	:	J	K	:	G
20	rs2427077	59.33	100.36	K	|	G	K	|	J	K	|	G
20	rs1892326	59.38	100.48	K	|	G	K	|	J	K	|	G
20	rs944257	59.68	101.15	K	|	G	K	|	J	K	|	G

Table Appen 6.2; Haplotype study of some members of family #1, across chromosome 20 (whole chromosome). The first 4 columns show; chromosome ID, position in mega base (Mb) and position in centi Morgan (cM). The next 3 big columns are for the haplotype of individuals; II-2 & IV-2 and III-3. Each of the three big columns contain 3 small columns showing haplotypes inherited from mother, information about recombination (as detailed below) and haplotype inherited from father, respectively. Merlin software has marked founders with (F) and for non-founders, added parent ID in brackets (instead of F). Haplotype pairs are separated by a | for no recombination, : for no information on recombination and /, \, + for recombination (in the maternal, paternal & both haplotype respectively).

(Appendix D)

(Appendix D1a)

FOXE-1 Poly-Alanine tract investigation

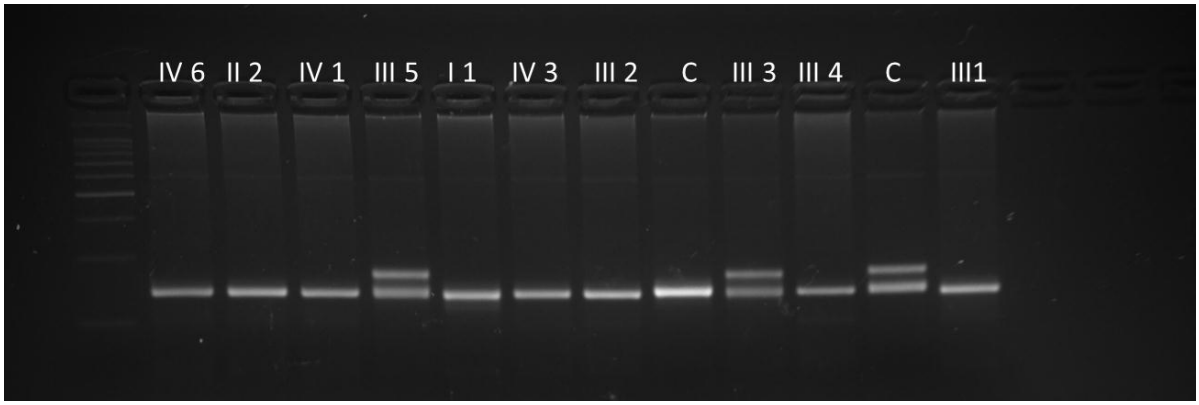


Figure App.6.1; Ethidium bromide stained agarose gel of FOXE-1 poly-Alanine tract screening for all affected family #1 members labelled as I the family tree and some unaffected in addition to unrelated controls. Double band represent heterozygous 14/16, while single band represents homozygous 16/16 14/14.

(Appendix D1b)

Investigation of the SNPs close to FOXE-1 and associated with PTC and FTC

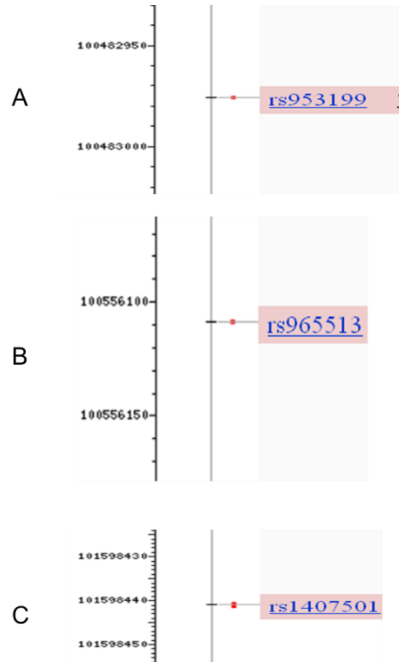


Figure App.6.2; Image of the SNPs (A & C) flanking the reported SNP (B) to be located close to FOXE-1 gene and associated with increased risk of both PTC and FTC

References

1. Wang, K. and M. Bucan, *Copy Number Variation Detection via High-Density SNP Genotyping.* CSH Protoc, 2008. **2008**: p. pdb top46.
2. Wang, K., et al., *PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.* Genome Research, 2007. **17**(11): p. 1665-1674.
3. Ingbar's, W., *The Thyroid: A Fundamental and Clinical Text* S. Seigafuse, Editor 2012, Lippincott Williams and Wilkins.
4. Bizhanova, A. and P. Kopp, *Minireview: The Sodium-Iodide Symporter NIS and Pendrin in Iodide Homeostasis of the Thyroid.* Endocrinology, 2009. **150**(3): p. 1084-1090.
5. Riesco-Eizaguirre, G. and P. Santisteban, *A perspective view of sodium iodide symporter research and its clinical implications.* European Journal of Endocrinology, 2006. **155**(4): p. 495-512.
6. Meitzler, J.L. and P.R.O. de Montellano, *Caenorhabditis elegans and Human Dual Oxidase 1 (DUOX1) "Peroxidase" Domains INSIGHTS INTO HEME BINDING AND CATALYTIC ACTIVITY.* Journal of Biological Chemistry, 2009. **284**(28): p. 18634-18643.
7. Benvenga, S., D. Lapa, and F. Trimarchi, *Thyroxine binding to members and non-members of the serine protease inhibitor family.* Journal of Endocrinological Investigation, 2002. **25**(1): p. 32-38.
8. Kopp, P.A., *Reduce, recycle, reuse - Iodotyrosine deiodinase in thyroid iodide metabolism.* New England Journal of Medicine, 2008. **358**(17): p. 1856-1859.
9. Sawyer, S.T. and S. Cohen, *ENHANCEMENT OF CALCIUM-UPTAKE AND PHOSPHATIDYLINOSITOL TURNOVER BY EPIDERMAL GROWTH-FACTOR IN A-431 CELLS.* Biochemistry, 1981. **20**(21): p. 6280-6286.
10. Carrasco, N., Werner, and Ingbar`s, *Thyroid Synthesis and Secretion*, in *Werner and Ingbar's The Thyroid*, L. Braverman and R. Utiger, Editors. 2000, Lippincott Williams & Wilkins: Philadelphia. p. 52-60.
11. Fernandez-Soto, L., et al., *Increased risk of autoimmune thyroid disease in hepatitis C vs hepatitis B before, during, and after discontinuing interferon therapy.* Archives of Internal Medicine, 1998. **158**(13): p. 1445-1448.
12. Vassart, G. and J. Dumont, *The thyrotropin receptor and the regulation of thyrocyte function and growth.* Endocr Rev., 1992. **13**(3): p. 596-611.
13. Thrush, A.B., A. Gagnon, and A. Sorisky, *PKC Activation is Required for TSH-mediated Lipolysis via Perilipin Activation.* Hormone and Metabolic Research, 2012. **44**(11): p. 825-831.
14. Vulsma, T., M.H. Gons, and J.J.M. Devijlder, *MATERNAL FETAL TRANSFER OF THYROXINE IN CONGENITAL HYPOTHYROIDISM DUE TO A TOTAL ORGANIFICATION DEFECT OR THYROID AGENESIS.* New England Journal of Medicine, 1989. **321**(1): p. 13-16.
15. Bath, S.C. and M.P. Rayman, *Iodine deficiency in the UK: an overlooked cause of impaired neurodevelopment?* Proceedings of the Nutrition Society, 2013. **72**(2): p. 226-235.
16. Fisher, D.A., *Thyroid System Immaturities in Very Low Birth Weight Premature Infants.* Seminars in Perinatology, 2008. **32**(6): p. 387-397.
17. Roef, G., et al., *Thyroid hormone status within the physiological range affects bone mass and density in healthy men at the age of peak bone mass.* European Journal of Endocrinology, 2011. **164**(6): p. 1027-1034.
18. Sawin, C.T., et al., *History, Ontogen, and Anatomy*, in *Werner and Ingbar's The Thyroid*, L. Braverman and R. Utiger, Editors. 2000, Lippincott Williams & Wilkins: Philadelphia. p. pp.3-51.

19. Arrojo E Drigo, R., et al., *Role of the type 2 iodothyronine deiodinase (D2) in the control of thyroid hormone signaling.* Biochimica et biophysica acta, 2013. **1830**(7): p. 3956-64.
20. Bianco, A.C., et al., *Biochemistry, cellular and molecular biology, and physiological roles of the iodothyronine selenodeiodinases.* Endocrine Reviews, 2002. **23**(1): p. 38-89.
21. Fagman, H. and M. Nilsson, *Morphogenesis of the thyroid gland.* Molecular and Cellular Endocrinology, 2010. **323**(1): p. 35-54.
22. Lazzaro, D., et al., *THE TRANSCRIPTION FACTOR-TTF-1 IS EXPRESSED AT THE ONSET OF THYROID AND LUNG MORPHOGENESIS AND IN RESTRICTED REGIONS OF THE FETAL BRAIN.* Development, 1991. **113**(4): p. 1093-&.
23. Dohan, O., et al., *The sodium/iodide symporter (NIS): Characterization, regulation, and medical significance.* Endocrine Reviews, 2003. **24**(1): p. 48-77.
24. Milenkovic, M., et al., *Duox expression and related H2O2 measurement in mouse thyroid: onset in embryonic development and regulation by TSH in adult.* Journal of Endocrinology, 2007. **192**(3): p. 615-626.
25. Postiglione, M.P., et al., *Role of the thyroid-stimulating hormone receptor signaling in development and differentiation of the thyroid gland.* Proceedings of the National Academy of Sciences of the United States of America, 2002. **99**(24): p. 15462-15467.
26. Fisher, D.A., et al., *Maturation of human hypothalamic-pituitary-thyroid function and control.* Thyroid, 2000. **10**(3): p. 229-234.
27. De Felice, M. and R. Di Lauro, *Thyroid development and its disorders: Genetics and molecular mechanisms.* Endocrine Reviews, 2004. **25**(5): p. 722-746.
28. Santisteban, P. and J. Bernal, *Thyroid development and effect on the nervous system.* Reviews in Endocrine & Metabolic Disorders, 2005. **6**(3): p. 217-228.
29. Fagman, H. and M. Nilsson, *Morphogenetics of early thyroid development.* Journal of Molecular Endocrinology, 2011. **46**(1): p. R33-R42.
30. Damante, G. and R. Dilauro, *THYROID-SPECIFIC GENE-EXPRESSION.* Biochimica Et Biophysica Acta-Gene Structure and Expression, 1994. **1218**(3): p. 255-266.
31. Damante, G., G. Tell, and R. Di Lauro, *A unique combination of transcription factors controls differentiation of thyroid cells.* Progress in Nucleic Acid Research and Molecular Biology, Vol 66, 2001. **66**: p. 307-356.
32. Ohno, M., et al., *The paired-domain transcription factor Pax8 binds to the upstream enhancer of the rat sodium/iodide symporter gene and participates in both thyroid-specific and cyclic-AMP-dependermt transcription.* Molecular and Cellular Biology, 1999. **19**(3): p. 2051-2060.
33. Antonica, F., et al., *Generation of functional thyroid from embryonic stem cells.* Nature, 2012. **491**(7422): p. 66-U170.
34. Medina, D.L. and P. Santisteban, *Thyrotropin-dependent proliferation of in vitro rat thyroid cell systems.* European Journal of Endocrinology, 2000. **143**(2): p. 161-178.
35. Kimura, T., et al., *Regulation of thyroid cell proliferation by TSH and other factors: A critical evaluation of in vitro models.* Endocrine Reviews, 2001. **22**(5): p. 631-656.
36. Bondy, C., et al., *CELLULAR-PATTERN OF TYPE-I INSULIN-LIKE GROWTH-FACTOR RECEPTOR GENE-EXPRESSION DURING MATURATION OF THE RAT-BRAIN - COMPARISON WITH INSULIN-LIKE GROWTH FACTOR-I AND FACTOR-II.* Neuroscience, 1992. **46**(4): p. 909-923.
37. Partanen, A.M., *EPIDERMAL GROWTH FACTOR AND TRANSFORMING GROWTH FACTOR-ALPHA IN THE DEVELOPMENT OF EPITHELIAL-MESENCHYMAL ORGANS OF THE MOUSE*, in *Nilsen-Hamilton, M.* 1990. p. 31-56.
38. Maciel, R.M., et al., *Demonstration of the production and physiological role of insulin-like growth factor II in rat thyroid follicular cells in culture.* J Clin Invest, 1988. **82**(5): p. 1546-53.

39. Bidey, S.P., D.J. Hill, and M.C. Eggo, *Growth factors and goitrogenesis.* J Endocrinol, 1999. **160**(3): p. 321-32.
40. Sugenoya, A., et al., *Adenomatous goitre: therapeutic strategy, postoperative outcome, and study of epidermal growth factor receptor.* Br J Surg, 1992. **79**(5): p. 404-6.
41. Ock, S., et al., *IGF-1 receptor deficiency in thyrocytes impairs thyroid hormone secretion and completely inhibits TSH-stimulated goitre.* FASEB J, 2013.
42. Eggo, M.C., L.K. Bachrach, and G.N. Burrow, *Interaction of TSH, insulin and insulin-like growth factors in regulating thyroid growth and function.* Growth Factors, 1990. **2**(2-3): p. 99-109.
43. Clement, S., et al., *Low TSH requirement and goitre in transgenic mice overexpressing IGF-I and IGF-Ir receptor in the thyroid gland.* Endocrinology, 2001. **142**(12): p. 5131-9.
44. Hill, J.J., et al., *Glycoproteomic analysis of two mouse mammary cell lines during transforming growth factor (TGF)-beta induced epithelial to mesenchymal transition.* Proteome Sci, 2009. **7**: p. 2.
45. Riesco-Eizaguirre, G., et al., *The BRAFV600E oncogene induces transforming growth factor beta secretion leading to sodium iodide symporter repression and increased malignancy in thyroid cancer.* Cancer Res, 2009. **69**(21): p. 8317-25.
46. Herbst, R.S., *Review of epidermal growth factor receptor biology.* Int J Radiat Oncol Biol Phys, 2004. **59**(2 Suppl): p. 21-6.
47. Hebrant, A., et al., *Long-term EGF/serum-treated human thyrocytes mimic papillary thyroid carcinomas with regard to gene expression.* Exp Cell Res, 2007. **313**(15): p. 3276-84.
48. Braverma.Le, N.A. Dawber, and S.H. Ingbar, *OBSERVATIONS CONCERNING BINDING OF THYROID HORMONES IN SERA OF NORMAL SUBJECTS OF VARYING AGES.* Journal of Clinical Investigation, 1966. **45**(8): p. 1273-&.
49. Ravaglia, G., et al., *Blood micronutrient and thyroid hormone concentrations in the oldest-old.* Journal of Clinical Endocrinology & Metabolism, 2000. **85**(6): p. 2260-2265.
50. Boucai, L. and M.I. Surks, *Reference limits of serum TSH and free T4 are significantly influenced by race and age in an urban outpatient medical practice.* Clinical Endocrinology, 2009. **70**(5): p. 788-793.
51. Maugeri, D., et al., *Thyroid function in healthy centenarians.* Archives of Gerontology and Geriatrics, 1997. **25**(2): p. 211-217.
52. Krohn, K. and R. Paschke, *Progress in understanding the etiology of thyroid autonomy.* Journal of Clinical Endocrinology & Metabolism, 2001. **86**(7): p. 3336-3345.
53. Delange, F., et al., *Iodine deficiency in the world: Where do we stand at the turn of the century?* Thyroid, 2001. **11**(5): p. 437-447.
54. Francis, B., et al., *Benign Thyroid Disease.* Grand Rounds Presentation, UTMB, Dept. of Otolaryngology, 2003.
55. Roitt, I.M., P.N. Campbell, and D. Doniach, *NATURE OF THE THYROID AUTO-ANTIBODIES PRESENT IN PATIENTS WITH HASHIMOTOS THYROIDITIS (LYMPHADENOID GOITRE).* Biochemical Journal, 1958. **69**: p. 248-&.
56. McLeod, D.S.A. and D.S. Cooper, *The incidence and prevalence of thyroid autoimmunity (vol 42, pg 252, 2012).* Endocrine, 2013. **43**(1): p. 244-244.
57. Toublanc, J.E., *COMPARISON OF EPIDEMIOLOGIC DATA ON CONGENITAL HYPOTHYROIDISM IN EUROPE WITH THOSE OF OTHER PARTS IN THE WORLD.* Hormone Research, 1992. **38**(5-6): p. 230-235.
58. Dentice, M., et al., *Missense mutation in the transcription factor NKX2-5: A novel molecular event in the pathogenesis of thyroid dysgenesis.* Journal of Clinical Endocrinology & Metabolism, 2006. **91**(4): p. 1428-1433.
59. Castanet, M., et al., *Nineteen years of national screening for congenital hypothyroidism: Familial cases with thyroid dysgenesis suggest the involvement of genetic factors.* Journal of Clinical Endocrinology & Metabolism, 2001. **86**(5): p. 2009-2014.

60. Park, S.M. and V.K.K. Chatterjee, *Genetics of congenital hypothyroidism.* Journal of Medical Genetics, 2005. **42**(5): p. 379-389.
61. Castanet, M. and M. Polak, *Spectrum of Human Foxe1/TTF2 Mutations.* Horm Res Paediatr, 2010. **73**(6): p. 423-9.
62. Clifton-Bligh, R.J., et al., *Mutation of the gene encoding human TTF-2 associated with thyroid agenesis, cleft palate and choanal atresia.* Nature Genetics, 1998. **19**(4): p. 399-401.
63. Tonacchera, M., et al., *Genetic analysis of TTF-2 gene in children with congenital hypothyroidism and cleft palate, congenital hypothyroidism, or isolated cleft palate.* Thyroid, 2004. **14**(8): p. 584-588.
64. Santarpia, L., et al., *TTF-2/FOXE1 gene polymorphisms in Sicilian patients with permanent primary congenital hypothyroidism.* J Endocrinol Invest, 2007. **30**(1): p. 13-9.
65. Hishinuma, A., et al., *Polymorphism of the polyalanine tract of thyroid transcription factor-2 gene in patients with thyroid dysgenesis.* Eur J Endocrinol, 2001. **145**(4): p. 385-9.
66. Tajiri, J., et al., *STUDIES OF HYPOTHYROIDISM IN PATIENTS WITH HIGH IODINE INTAKE.* Journal of Clinical Endocrinology & Metabolism, 1986. **63**(2): p. 412-417.
67. Larsen, P.R., T. Davies, and I. Hay, *Thyroid Physiology and Diagnositic Evaluation of Patients with Thyroid Disorders*, in *Williams textbook of endocrinology* . J. Wilson, et al., Editors. 2002, W.B. Saunders Company: Philadelphia. p. pp 331-455.
68. Singer, J., et al., *Evidence for a more Pronounced Effect of Genetic Predisposition than Environmental Factors on Goitrogenesis by a Case Control Study in an Area with Low Normal Iodine Supply.* Hormone and Metabolic Research, 2011. **43**(5): p. 349-354.
69. Hansen, P.S., et al., *The relative importance of genetic and environmental factors in the aetiology of thyroid nodularity: a study of healthy Danish twins.* Clinical Endocrinology, 2005. **62**(3): p. 380-386.
70. Knudsen, N., et al., *Thyroid structure and size and two-year follow-up of solitary cold thyroid nodules in an unselected population with borderline iodine deficiency.* European Journal of Endocrinology, 2000. **142**(3): p. 224-230.
71. Belfiore, A., et al., *THE FREQUENCY OF COLD THYROID-NODULES AND THYROID MALIGNANCIES IN PATIENTS FROM AN IODINE-DEFICIENT AREA.* Cancer, 1987. **60**(12): p. 3096-3102.
72. Bahre, M., et al., *THYROID AUTONOMY - SENSITIVE DETECTION INVIVO AND ESTIMATION OF ITS FUNCTIONAL RELEVANCE USING QUANTIFIED HIGH-RESOLUTION SCINTIGRAPHY.* Acta Endocrinologica, 1988. **117**(2): p. 145-153.
73. Stone, J.R., *An assessment of proposed mechanisms for sensing hydrogen peroxide in mammalian systems.* Arch Biochem Biophys, 2004. **422**(2): p. 119-24.
74. Krohn, K., J. Maier, and R. Paschke, *Mechanisms of disease: hydrogen peroxide, DNA damage and mutagenesis in the development of thyroid tumors.* Nat Clin Pract Endocrinol Metab, 2007. **3**(10): p. 713-20.
75. Song, Y., et al., *Roles of hydrogen peroxide in thyroid physiology and disease.* J Clin Endocrinol Metab, 2007. **92**(10): p. 3764-73.
76. E, G., *Environmental natural goitrogens. in Peter F, Wiersinga WM, Hostalek U, eds* The thyroid and environment, 2000: p. 69–78.
77. Tunbridge, W.M.G., et al., *SPECTRUM OF THYROID DISEASE IN A COMMUNITY - WHICKHAM SURVEY.* Clinical Endocrinology, 1977. **7**(6): p. 481-493.
78. Ron, E., et al., *A population-based case-control study of thyroid cancer.* Journal Of The National Cancer Institute, 1987. **79**(1): p. 1-12.
79. Preston-Martin, S., et al., *Thyroid cancer among young women related to prior thyroid disease and pregnancy history.* Br J Cancer., 1987. **55**(2): p. 191-5.
80. Vander, J., E. Gaston, and T. Dawber, *Significance of solitary nontoxic thyroid nodules; preliminary report.* N Engl J Med., 1954. **251**(24): p. 970-3.

81. Fiore, E., et al., *Lower levels of TSH are associated with a lower risk of papillary thyroid cancer in patients with thyroid nodular disease: thyroid autonomy may play a protective role.* Endocrine-Related Cancer, 2009. **16**(4): p. 1251-1260.
82. Kilfoy, B.A., et al., *International patterns and trends in thyroid cancer incidence, 1973-2002.* Cancer Causes & Control, 2009. **20**(5): p. 525-531.
83. Bignell, G.R., et al., *Familial nontoxic multinodular thyroid goitre locus maps to chromosome 149 but does not account for familial nonmedullary thyroid cancer.* American Journal of Human Genetics, 1997. **61**(5): p. 1123-1130.
84. Capon, F., et al., *Mapping a dominant form of multinodular goitre to chromosome Xp22.* American Journal of Human Genetics, 2000. **67**(4): p. 308-308.
85. Canzian, F., et al., *A gene predisposing to familial thyroid tumors with cell oxyphilia maps to chromosome 19p13.2.* American Journal of Human Genetics, 1998. **63**(6): p. 1743-1748.
86. Malchoff, C.D., et al., *Papillary thyroid carcinoma associated with papillary renal neoplasia: genetic linkage analysis of a distinct heritable tumor syndrome.* Journal of Clinical Endocrinology & Metabolism, 2000. **85**(5): p. 1758-1764.
87. McKay, J.D., et al., *Localization of a susceptibility gene for familial nonmedullary thyroid carcinoma to chromosome 2q21.* American Journal of Human Genetics, 2001. **69**(2): p. 440-446.
88. Frisk, T., et al., *Silencing of the PTEN tumor-suppressor gene in anaplastic thyroid cancer.* Genes Chromosomes & Cancer, 2002. **35**(1): p. 74-80.
89. Takahashi, T., et al., *A new locus for a dominant form of multinodular goitre on 3q126.1-q26.3.* Biochemical and Biophysical Research Communications, 2001. **284**(3): p. 648-654.
90. Bayer, Y., et al., *Genome-wide linkage analysis reveals evidence for four new susceptibility loci for familial euthyroid goitre.* Journal of Clinical Endocrinology & Metabolism, 2004. **89**(8): p. 4044-4052.
91. Gudmundsson, J., et al., *Common variants on 9q22.33 and 14q13.3 predispose to thyroid cancer in European populations.* Nat Genet, 2009. **41**(4): p. 460-4.
92. Barnett, C.P., et al., *Choreoathetosis, congenital hypothyroidism and neonatal respiratory distress syndrome with intact NKX2-1.* American Journal of Medical Genetics Part A, 2012. **158A**(12): p. 3168-3173.
93. Tomaz, R.A., et al., *FOXE1 polymorphisms are associated with familial and sporadic nonmedullary thyroid cancer susceptibility.* Clinical Endocrinology, 2012. **77**(6): p. 926-933.
94. Teumer, A., et al., *Genome-wide Association Study Identifies Four Genetic Loci Associated with Thyroid Volume and Goitre Risk.* American Journal of Human Genetics, 2011. **88**(5): p. 664-673.
95. Wichmann, H.E., et al., *KORA-gen - Resource for population genetics, controls and a broad spectrum of disease phenotypes.* Gesundheitswesen, 2005. **67**: p. S26-S30.
96. Salabe, G., *Pathogenesis of thyroid nodules: histological classification?* Biomed Pharmacother., 2001. **55**(1): p. 39-53.
97. Bottcher, Y., et al., *The genetics of euthyroid familial goitre.* Trends in Endocrinology and Metabolism, 2005. **16**(7): p. 314-319.
98. Segev, D.L., C. Umbricht, and M.A. Zeiger, *Molecular pathogenesis of thyroid cancer.* Surgical Oncology-Oxford, 2003. **12**(2): p. 69-90.
99. Hodgson, N.C., J. Button, and C.C. Solorzano, *Thyroid cancer: Is the incidence still increasing?* Annals of Surgical Oncology, 2004. **11**(12): p. 1093-1097.
100. Jemal, A., et al., *Cancer Statistics, 2009.* Ca-a Cancer Journal for Clinicians, 2009. **59**(4): p. 225-249.
101. Cerutti, J.M., et al., *A preoperative diagnostic test that distinguishes benign from malignant thyroid carcinoma based on gene expression.* Journal of Clinical Investigation, 2004. **113**(8): p. 1234-1242.

102. Barnabei, A., et al., *Hurthle cell tumours of the thyroid. Personal experience and review of the literature.* Acta Otorhinolaryngol Ital, 2009. **29**(6): p. 305-11.
103. Franssila, K.O., et al., *Follicular carcinoma.* Semin Diagn Pathol, 1985. **2**(2): p. 101-22.
104. Mai, K.T., et al., *Pathologic study and clinical significance of Hurthle cell papillary thyroid carcinoma.* Appl Immunohistochem Mol Morphol, 2004. **12**(4): p. 329-37.
105. Wong, K.K., et al., *Molecular imaging in the management of thyroid cancer.* Quarterly Journal of Nuclear Medicine and Molecular Imaging, 2011. **55**(5): p. 541-559.
106. Griebeler, M.L., H. Gharib, and G.B. Thompson, *Medullary thyroid carcinoma.* Endocrine practice : official journal of the American College of Endocrinology and the American Association of Clinical Endocrinologists, 2013. **19**(4): p. 703-11.
107. American Thyroid Association Guidelines Task, F., et al., *Medullary thyroid cancer: management guidelines of the American Thyroid Association.* Thyroid, 2009. **19**(6): p. 565-612.
108. Figlioli, G., et al., *Medullary thyroid carcinoma (MTC) and RET proto-oncogene: Mutation spectrum in the familial cases and a meta-analysis of studies on the sporadic form.* Mutation Research-Reviews in Mutation Research, 2013. **752**(1): p. 36-44.
109. Correia-Deur, J.E., et al., *Sporadic medullary thyroid carcinoma: clinical data from a university hospital.* Clinics (Sao Paulo), 2009. **64**(5): p. 379-86.
110. Machens, A., et al., *Prospects of remission in medullary thyroid carcinoma according to basal calcitonin level.* J Clin Endocrinol Metab, 2005. **90**(4): p. 2029-34.
111. Zivaljevic, V.R., et al., *Case-control study of anaplastic thyroid cancer: goitre patients as controls.* European journal of cancer prevention : the official journal of the European Cancer Prevention Organisation (ECP), 2008. **17**(2): p. 111-5.
112. Moretti, F., S. Nanni, and A. Pontecorvi, *Molecular pathogenesis of thyroid nodules and cancer.* Baillieres Best Pract Res Clin Endocrinol Metab, 2000. **14**(4): p. 517-39.
113. Gharib, H., *SUBSPECIALTY CLINICS - ENDOCRINOLOGY/METABOLISM - FINE-NEEDLE ASPIRATION BIOPSY OF THYROID-NODULES - ADVANTAGES, LIMITATIONS, AND EFFECT.* Mayo Clinic Proceedings, 1994. **69**(1): p. 44-49.
114. Goellner, J.R., et al., *FINE NEEDLE ASPIRATION CYTOLOGY OF THE THYROID, 1980 TO 1986.* Acta Cytologica, 1987. **31**(5): p. 587-590.
115. Dijkstra, C.D., *Nobel Prize for Physiology or Medicine 2002 is awarded for research into the genetic regulation of organ development and programmed cell death.* Nederlands tijdschrift voor geneeskunde, 2002. **146**(52): p. 2525-7.
116. Grieco, M., et al., *PTC is a novel rearranged form of the ret proto-oncogene and is frequently detected in vivo in human thyroid papillary carcinomas.* Cell, 1990. **60**(4): p. 557-63.
117. Kroll, T.G., et al., *PAX8-PPAR gamma 1 fusion in oncogene human thyroid carcinoma.* Science, 2000. **289**(5483): p. 1357-1360.
118. Lemoine, N.R., et al., *Activated ras oncogenes in human thyroid cancers.* Cancer Res, 1988. **48**(16): p. 4459-63.
119. Bullock, M., et al., *Association of FOXE1 Polyalanine Repeat Region with Papillary Thyroid Cancer*. Journal of Clinical Endocrinology & Metabolism, 2012. **97**(9): p. E1814-E1819.
120. Kimura, E.T., et al., *High prevalence of BRAF mutations in thyroid cancer: Genetic evidence for constitutive activation of the RET/PTC-RAS-BRAF signaling pathway in papillary thyroid carcinoma.* Cancer Research, 2003. **63**(7): p. 1454-1457.
121. Gandhi, M., V. Evdokimova, and Y.E. Nikiforov, *Mechanisms of chromosomal rearrangements in solid tumors: the model of papillary thyroid carcinoma.* Mol Cell Endocrinol, 2010. **321**(1): p. 36-43.
122. Pierotti, M.A., et al., *Characterization of an inversion on the long arm of chromosome 10 juxtaposing D10S170 and RET and creating the oncogenic sequence RET/PTC.* Proc Natl Acad Sci U S A, 1992. **89**(5): p. 1616-20.

123. Nikiforov, Y.E., *Molecular diagnostics of thyroid tumors.* Arch Pathol Lab Med, 2011. **135**(5): p. 569-77.
124. Schwaller, J., et al., *H4(D10S170), a gene frequently rearranged in papillary thyroid carcinoma, is fused to the platelet-derived growth factor receptor beta gene in atypical chronic myeloid leukemia with t(5;10)(q33;q22).* Blood, 2001. **97**(12): p. 3910-8.
125. Puxeddu, E., et al., *Characterization of novel non-chemical intrachromosomal rearrangements between the H4 and PTEN genes (H4/PTEN) in human thyroid cell lines and papillary thyroid cancer specimens.* Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis, 2005. **570**(1): p. 17-32.
126. Williams, E.D., et al., *Thyroid carcinoma after Chernobyl latent period, morphology and aggressiveness.* Br J Cancer, 2004. **90**(11): p. 2219-24.
127. Jargin, S.V., *On the RET Rearrangements in Chernobyl-Related Thyroid Cancer.* J Thyroid Res, 2012. **2012**: p. 373879.
128. Thakur, P., S. Ballard, and R. Nelson, *An overview of Fukushima radionuclides measured in the northern hemisphere.* Science of the Total Environment, 2013. **458**: p. 577-613.
129. Bangarzone, I., et al., *HIGH FREQUENCY OF ACTIVATION OF TYROSINE KINASE ONCOGENES IN HUMAN PAPILLARY THYROID CARCINOMA.* Oncogene, 1989. **4**(12): p. 1457-1462.
130. Greco, A., C. Miranda, and M.A. Pierotti, *Rearrangements of NTRK1 gene in papillary thyroid carcinoma.* Mol Cell Endocrinol, 2010. **321**(1): p. 44-9.
131. Russell, J.P., et al., *The TRK-T1 fusion protein induces neoplastic transformation of thyroid epithelium.* Oncogene, 2000. **19**(50): p. 5729-5735.
132. Eng, C., et al., *The relationship between specific RET proto-oncogene mutations and disease phenotype in multiple endocrine neoplasia type 2 - International RET mutation consortium analysis.* Jama-Journal of the American Medical Association, 1996. **276**(19): p. 1575-1579.
133. Mulligan, L.M., et al., *Germ-line mutations of the RET proto-oncogene in multiple endocrine neoplasia type 2A.* Nature, 1993. **363**(6428): p. 458-60.
134. Uchino, S., et al., *Novel point mutations and allele loss at the RET locus in sporadic medullary thyroid carcinomas.* Japanese Journal of Cancer Research, 1998. **89**(4): p. 411-418.
135. Schilling, T., et al., *Prognostic value of codon 918 (ATG -> ACG) RET proto-oncogene mutations in sporadic medullary thyroid carcinoma.* International Journal of Cancer, 2001. **95**(1): p. 62-66.
136. Elisei, R., et al., *Prognostic significance of somatic RET oncogene mutations in sporadic medullary thyroid cancer: a 10-year follow-up study.* J Clin Endocrinol Metab, 2008. **93**(3): p. 682-7.
137. Eng, C., *RET proto-oncogene in the development of human cancer.* J Clin Oncol, 1999. **17**(1): p. 380-93.
138. Marsh, D.J., et al., *Somatic mutations in the RET proto-oncogene in sporadic medullary thyroid carcinoma.* Clin Endocrinol (Oxf), 1996. **44**(3): p. 249-57.
139. Jones, J.R., et al., *Deletion of PPARgamma in adipose tissues of mice protects against high fat diet-induced obesity and insulin resistance.* Proc Natl Acad Sci U S A, 2005. **102**(17): p. 6207-12.
140. Lu, C. and S.-Y. Cheng, *Thyroid hormone receptors regulate adipogenesis and carcinogenesis via crosstalk signaling with peroxisome proliferator-activated receptors.* Journal of Molecular Endocrinology, 2010. **44**(3): p. 143-154.
141. Festuccia, W.T., et al., *PPARgamma activation attenuates cold-induced upregulation of thyroid status and brown adipose tissue PGC-1alpha and D2.* Am J Physiol Regul Integr Comp Physiol, 2012. **303**(12): p. R1277-85.
142. Chattopadhyay, C., et al., *Small molecule c-MET inhibitor PHA665752: effect on cell growth and motility in papillary thyroid carcinoma.* Head Neck, 2008. **30**(8): p. 991-1000.

143. Bottaro, D.P., et al., *IDENTIFICATION OF THE HEPATOCYTE GROWTH-FACTOR RECEPTOR AS THE C-MET PROTOONCOGENE PRODUCT.* Science, 1991. **251**(4995): p. 802-804.
144. Ludgate, M. and C. Evans, *Thyroid cance: molecular genetics.* Review, 2001.
145. Siraj, A.K., et al., *Genome-wide expression analysis of Middle Eastern papillary thyroid cancer reveals c-MET as a novel target for cancer therapy.* Journal of Pathology, 2007. **213**(2): p. 190-199.
146. Ivan, M., et al., *Activated ras and ret oncogenes induce over-expression of c-met (hepatocyte growth factor receptor) in human thyroid epithelial cells.* Oncogene, 1997. **14**(20): p. 2417-2423.
147. Tulasne, D. and B. Foveau, *The shadow of death on the MET tyrosine kinase receptor.* Cell Death and Differentiation, 2008. **15**(3): p. 427-434.
148. Liu, Y.J., et al., *Expression and significance of IGF-1 and IGF-1R in thyroid nodules.* Endocrine, 2013. **44**(1): p. 158-64.
149. Smith, V.E., J.A. Franklyn, and C.J. McCabe, *Pituitary tumor-transforming gene and its binding factor in endocrine cancer.* Expert Rev Mol Med, 2010. **12**: p. e38.
150. Lewy, G.D.R., G. A. Read, M. L. Fong, J. C. Poole, V. Seed, R. I. Sharma, N. Smith, V. E. Kwan, P. P. Stewart, S. L. Bacon, A. Warfield, A. Franklyn, J. A. McCabe, C. J. Boelaert, K., *Regulation of Pituitary Tumor Transforming Gene (PTTG) expression and phosphorylation in thyroid cells.* Endocrinology, 2013.
151. Boelaert, K.S., V. E. Stratford, A. L. Kogai, T. Tannahill, L. A. Watkinson, J. C. Eggo, M. C. Franklyn, J. A. McCabe, C. J., *PTTG and PBF repress the human sodium iodide symporter.* Oncogene, 2007. **26**(30): p. 4344-56.
152. Franco, A.T., et al., *Thyrotrophin receptor signaling dependence of Braf-induced thyroid tumor initiation in mice.* Proc Natl Acad Sci U S A, 2011. **108**(4): p. 1615-20.
153. Durante, C., et al., *BRAF mutations in papillary thyroid carcinomas inhibit genes involved in iodine metabolism.* J Clin Endocrinol Metab, 2007. **92**(7): p. 2840-3.
154. Knauf, J.A. and J.A. Fagin, *Role of MAPK pathway oncoproteins in thyroid cancer pathogenesis and as drug targets.* Curr Opin Cell Biol, 2009. **21**(2): p. 296-303.
155. Charles, R.-P., et al., *Mutationally Activated BRAF(V600E) Elicits Papillary Thyroid Cancer in the Adult Mouse.* Cancer Research, 2011. **71**(11): p. 3863-3871.
156. Attardi, L., *The role of p53-mediated apoptosis as a crucial anti-tumor response to genomic instability: lessons from mouse models.* Mutat Res, 2005. **569**(1-2): p. 145-57.
157. Hartwell, L.H. and T.A. Weinert, *Checkpoints: controls that ensure the order of cell cycle events.* Science, 1989. **246**(4930): p. 629-34.
158. Robinson, T.R., *Genetics For Dummies*, ed. n. Edition. 2010: Wiley Publishing, Inc.
159. Nigg, E.A., *Cyclin-dependent protein kinases: key regulators of the eukaryotic cell cycle.* Bioessays, 1995. **17**(6): p. 471-80.
160. Neganova, I. and M. Lako, *G1 to S phase cell cycle transition in somatic and embryonic stem cells.* J Anat, 2008. **213**(1): p. 30-44.
161. Leland H. Hartwell, R.T.H., and Paul M. Nurse. *Nobel Prize in Physiology or Medicine*. 2001;Availablefrom:http://www.nobelprize.org/nobel_prizes/medicine/laureates/2001/press.html.
162. Robbins and Cotran; Kumar, A., Fausto, *Pathological Basis of Disease.* . 8th ed, ed. R. Pathology. 2009: Saunders.
163. Petersen, L., et al., *p53-dependent G(1) arrest in 1st or 2nd cell cycle may protect human cancer cells from cell death after treatment with ionizing radiation and Chk1 inhibitors.* Cell Prolif, 2010. **43**(4): p. 365-71.
164. Fero, M.L., et al., *A syndrome of multiorgan hyperplasia with features of gigantism, tumorigenesis, and female sterility in p27(Kip1)-deficient mice.* Cell, 1996. **85**(5): p. 733-44.

165. Liggett, W.H., Jr. and D. Sidransky, *Role of the p16 tumor suppressor gene in cancer.* J Clin Oncol, 1998. **16**(3): p. 1197-206.
166. Fagin, J.A., et al., *HIGH PREVALENCE OF MUTATIONS OF THE P53 GENE IN POORLY DIFFERENTIATED HUMAN THYROID CARCINOMAS.* Journal of Clinical Investigation, 1993. **91**(1): p. 179-184.
167. Donghi, R., et al., *GENE P53 MUTATIONS ARE RESTRICTED TO POORLY DIFFERENTIATED AND UNDIFFERENTIATED CARCINOMAS OF THE THYROID-GLAND.* Journal of Clinical Investigation, 1993. **91**(4): p. 1753-1760.
168. Battista, S., et al., *A MUTATED P53 GENE ALTERS THYROID-CELL DIFFERENTIATION.* Oncogene, 1995. **11**(10): p. 2029-2037.
169. Donehower, L.A., et al., *MICE DEFICIENT FOR P53 ARE DEVELOPMENTALLY NORMAL BUT SUSCEPTIBLE TO SPONTANEOUS TUMORS.* Nature, 1992. **356**(6366): p. 215-221.
170. Harvey, M., et al., *MICE DEFICIENT IN BOTH P53 AND RB DEVELOP TUMORS PRIMARILY OF ENDOCRINE ORIGIN.* Cancer Research, 1995. **55**(5): p. 1146-1151.
171. Nose, V., *Familial Non-Medullary Thyroid Carcinoma: An Update.* Endocrine Pathology, 2008. **19**(4): p. 226-240.
172. Laury, A.R., et al., *Thyroid Pathology in PTEN-Hamartoma Tumor Syndrome: Characteristic Findings of a Distinct Entity.* Thyroid, 2011. **21**(2): p. 135-144.
173. Pilarski, R. and C. Eng, *Will the real Cowden syndrome please stand up (again)? Expanding mutational and clinical spectra of the PTEN hamartoma tumour syndrome.* J Med Genet, 2004. **41**(5): p. 323-6.
174. Paes, J.E. and M.D. Ringel, *Dysregulation of the phosphatidylinositol 3-kinase pathway in thyroid neoplasia.* Endocrinol Metab Clin North Am, 2008. **37**(2): p. 375-87, viii-ix.
175. Khan, A., et al., *Familial Nonmedullary Thyroid Cancer: A Review of the Genetics.* Thyroid, 2010. **20**(7): p. 795-801.
176. Xing, M., *Genetic alterations in the phosphatidylinositol-3 kinase/Akt pathway in thyroid cancer.* Thyroid, 2010. **20**(7): p. 697-706.
177. Saji, M. and M.D. Ringel, *The PI3K-Akt-mTOR pathway in initiation and progression of thyroid tumors.* Mol Cell Endocrinol, 2010. **321**(1): p. 20-8.
178. Takahashi, M., et al., *The FOXE1 locus is a major genetic determinant for radiation-related thyroid carcinoma in Chernobyl.* Hum Mol Genet, 2010. **19**(12): p. 2516-23.
179. Landa, I., et al., *The variant rs1867277 in FOXE1 gene confers thyroid cancer susceptibility through the recruitment of USF1/USF2 transcription factors.* PLoS Genet, 2009. **5**(9): p. e1000637.
180. Ruiz-Llorente, S., et al., *Association study of 69 genes in the ret pathway identifies low-penetrance loci in sporadic medullary thyroid carcinoma.* Cancer Research, 2007. **67**(19): p. 9561-9567.
181. Bakhsh, A., et al., *A new form of familial multi-nodular goitre with progression to differentiated thyroid cancer.* Endocrine-Related Cancer, 2006. **13**(2): p. 475-483.
182. Bignell, G.R., et al., *A familial non-toxic multinodular thyroid goitre locus maps to chromosome 14q but does not account for familial non-medullary thyroid cancer.* American Journal of Human Genetics, 1997. **61**(4): p. A268-A268.
183. Boehnke, M., *LIMITS OF RESOLUTION OF GENETIC-LINKAGE STUDIES - IMPLICATIONS FOR THE POSITIONAL CLONING OF HUMAN-DISEASE GENES.* American Journal of Human Genetics, 1994. **55**(2): p. 379-390.
184. Holsinger, K.E. and B.S. Weir, *FUNDAMENTAL CONCEPTS IN GENETICS Genetics in geographically structured populations: defining, estimating and interpreting F-ST.* Nature Reviews Genetics, 2009. **10**(9): p. 639-650.
185. Gulcher, J., *Microsatellite markers for linkage and association studies.* Cold Spring Harbor protocols, 2012. **2012**(4): p. 425-32.

186. Murray, J.C., et al., *A comprehensive human linkage map with centimorgan density. Cooperative Human Linkage Center (CHLC).* Science, 1994. **265**(5181): p. 2049-54.
187. Dib, C., et al., *A comprehensive genetic map of the human genome based on 5,264 microsatellites.* Nature, 1996. **380**(6570): p. 152-4.
188. Kong, A., et al., *A high-resolution recombination map of the human genome.* Nat Genet, 2002. **31**(3): p. 241-7.
189. Collins, F.S., M.S. Guyer, and A. Charkravarti, *Variations on a theme: cataloging human DNA sequence variation.* Science, 1997. **278**(5343): p. 1580-1.
190. Matise, T.C., et al., *A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set.* Am J Hum Genet, 2003. **73**(2): p. 271-84.
191. Botstein, D., et al., *CONSTRUCTION OF A GENETIC-LINKAGE MAP IN MAN USING RESTRICTION FRAGMENT LENGTH POLYMORPHISMS.* American Journal of Human Genetics, 1980. **32**(3): p. 314-331.
192. Keller, M.F., et al., *Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinsons disease (vol 22, pg 1696, 2013).* Human Molecular Genetics, 2013. **22**(14): p. 2973-2973.
193. Hollingworth, P., et al., *Genome-wide association study of Alzheimer's disease with psychotic symptoms.* Molecular Psychiatry, 2012. **17**(12): p. 1316-1327.
194. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
195. Muddyman, D., et al., *Implementing a successful data-management framework: the UK10K managed access model.* Genome Medicine, 2013. **5**.
196. Eichler, E.E., R.A. Clark, and X. She, *An assessment of the sequence gaps: unfinished business in a finished human genome.* Nat Rev Genet, 2004. **5**(5): p. 345-54.
197. Tautz, D., *NOTES ON THE DEFINITION AND NOMENCLATURE OF TANDEMLY REPETITIVE DNA-SEQUENCES.* DNA Fingerprinting : State of the Science, ed. S.D.J. Pena, et al. Vol. 67. 1993. 21-28.
198. Oliveira, E.J., et al., *Origin, evolution and genome distribution of microsatellites.* Genetics and Molecular Biology, 2006. **29**(2): p. 294-307.
199. Messier, W., S.H. Li, and C.B. Stewart, *The birth of microsatellites.* Nature, 1996. **381**(6582): p. 483-483.
200. Zhu, Y., J.E. Strassmann, and D.C. Queller, *Insertions, substitutions, and the origin of microsatellites.* Genetical Research, 2000. **76**(3): p. 227-236.
201. Jeffreys, A.J. and S.D.J. Pena, *BRIEF INTRODUCTION TO HUMAN DNA-FINGERPRINTING.* DNA Fingerprinting : State of the Science, ed. S.D.J. Pena, et al. Vol. 67. 1993. 1-20.
202. Todd, J.A., et al., *GENETIC-ANALYSIS OF AUTOIMMUNE TYPE-1 DIABETES-MELLITUS IN MICE.* Nature, 1991. **351**(6327): p. 542-547.
203. Evans, D.M. and L.R. Cardon, *Guidelines for genotyping in genomewide linkage studies: Single-nucleotide-polymorphism maps versus microsatellite maps.* American Journal of Human Genetics, 2004. **75**(4): p. 687-692.
204. Clancy, S., *Genetic Mutation.* Nature Education 1 (1) 2008.
205. Ian N.M. Day, K.K.A., Matt Smith, Mohammed A. Aldahmesh, Xiao-He Chen, Andrew J. Lotery, Gabriella Pante-de-Sousa, Guangwei Hou, Shu Ye, Diana Eccles, Nicholas C. P. Cross, Keith R. Fox and Santiago Rodriguez *Paucimorphic Alleles versus Polymorphic Alleles and Rare Mutations in Disease Causation: Theory, Observation and Detection.* Current Genomics, 2004. **5**(5): p. 431-438 (8).
206. Varela, M.A. and W. Amos, *Heterogeneous distribution of SNPs in the human genome: Microsatellites as predictors of nucleotide diversity and divergence.* Genomics, 2010. **95**(3): p. 151-159.

207. Lewis, C.M. and J. Knight, *Introduction to genetic association studies.* Cold Spring Harbor protocols, 2012. **2012**(3): p. 297-306.
208. VanLiere, J.M. and N.A. Rosenberg, *Mathematical properties of the r(2) measure of linkage disequilibrium.* Theoretical Population Biology, 2008. **74**(1): p. 130-137.
209. Li, B. and S.M. Leal, *Discovery of Rare Variants via Sequencing: Implications for the Design of Complex Trait Association Studies.* Plos Genetics, 2009. **5**(5).
210. Cohen, J.C., et al., *Multiple rare alleles contribute to low plasma levels of HDL cholesterol.* Science, 2004. **305**(5685): p. 869-72.
211. Alharbi, K.K., et al., *Population Mutation Scanning of Human GHR by meltMADGE and Identification of a Paucimorphic Variant.* Genetic Testing and Molecular Biomarkers, 2011. **15**(12): p. 855-860.
212. Ardlie, K.G., K.L. Lunetta, and M. Seielstad, *Testing for population subdivision and association in four case-control studies.* Am J Hum Genet, 2002. **71**(2): p. 304-11.
213. Tabangin, M.E., J.G. Woo, and L.J. Martin, *The effect of minor allele frequency on the likelihood of obtaining false positives.* BMC Proc, 2009. **3 Suppl 7**: p. S41.
214. Lam, A.C., et al., *Rapid and robust association mapping of expression quantitative trait loci.* BMC Proc, 2007. **1 Suppl 1**: p. S144.
215. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
216. International HapMap, C., *Integrating ethics and science in the International HapMap Project.* Nat Rev Genet, 2004. **5**(6): p. 467-75.
217. International HapMap, C., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
218. Piao, S., et al., *The impacts of climate change on water resources and agriculture in China.* Nature, 2010. **467**(7311): p. 43-51.
219. Maskos, U. and E.M. Southern, *OLIGONUCLEOTIDE HYBRIDIZATIONS ON GLASS SUPPORTS - A NOVEL LINKER FOR OLIGONUCLEOTIDE SYNTHESIS AND HYBRIDIZATION PROPERTIES OF OLIGONUCLEOTIDES SYNTHESIZED INSITU.* Nucleic Acids Research, 1992. **20**(7): p. 1679-1684.
220. Kennedy, G.C., et al., *Large-scale genotyping of complex DNA.* Nat Biotechnol, 2003. **21**(10): p. 1233-7.
221. Matsuzaki, H., et al., *Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array (vol 14, pg 414, 2004).* Genome Research, 2004. **14**(7): p. 1444-1444.
222. Segurado, R., et al., *Combining linkage data sets for meta-analysis and mega-analysis: the GAW15 rheumatoid arthritis data set.* BMC Proc, 2007. **1 Suppl 1**: p. S104.
223. Kaindl, A.M., et al., *Missense mutations of ACTA1 cause dominant congenital myopathy with cores.* J Med Genet, 2004. **41**(11): p. 842-8.
224. Ruschendorf, F. and P. Nurnberg, *ALOHOMORA: a tool for linkage analysis using 10K SNP array data.* Bioinformatics, 2005. **21**(9): p. 2123-5.
225. Middleton, F.A., et al., *Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22.* Am J Hum Genet, 2004. **74**(5): p. 886-97.
226. Puffenberger, E.G., et al., *Mapping of sudden infant death with dysgenesis of the testes syndrome (SIDDT) by a SNP genome scan and identification of TSPYL loss of function.* Proc Natl Acad Sci U S A, 2004. **101**(32): p. 11689-94.
227. Shrimpton, A.E., et al., *A HOX gene mutation in a family with isolated congenital vertical talus and Charcot-Marie-Tooth disease.* Am J Hum Genet, 2004. **75**(1): p. 92-6.
228. Wang, D.G., et al., *Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome.* Science, 1998. **280**(5366): p. 1077-82.

229. Meaburn, E., et al., *Genotyping DNA pools on microarrays: tackling the QTL problem of large samples and large numbers of SNPs.* BMC Genomics, 2005. **6**: p. 52.
230. Liu, W.M., et al., *Algorithms for large-scale genotyping microarrays.* Bioinformatics, 2003. **19**(18): p. 2397-2403.
231. Samejima, K. and W.C. Earnshaw, *Trashing the genome: the role of nucleases during apoptosis.* Nat Rev Mol Cell Biol, 2005. **6**(9): p. 677-88.
232. Sherry, S.T., M. Ward, and K. Sirotkin, *The NCBI dbSNP database for Single Nucleotide Polymorphisms and other classes of minor genetic variation.* American Journal of Human Genetics, 1999. **65**(4): p. A101-A101.
233. Affymetrix, I., *GeneChip® Mapping 10K Array Publications.* 2007.
234. Trent, R.J., *Clinical Bioinformatics*, ed. J.M. Walker. 2008, Herts, UK: Humana Press.
235. Affymetrix, I. *GeneChip Mapping 100K Assay*. 2013 Available from: http://www.affymetrix.com/estore/browse/products.jsp;jsessionid=1446DBEB1C5E034AE890AAECF40D98EB?productId=131469#1_1.
236. Sigelman C K, R.E.A., ed. *Life-Span Human Development*. 7 ed. 2011, Wadsworth Publishing: Belmond, California.
237. Goldgar, D.E. and P.R. Fain, *Models of multilocus recombination: nonrandomness in chiasma number and crossover positions.* Am J Hum Genet, 1988. **43**(1): p. 38-45.
238. Creighton, H.B. and B. McClintock, *A Correlation of Cytological and Genetical Crossing-Over in Zea Mays.* Proc Natl Acad Sci U S A, 1931. **17**(8): p. 492-7.
239. Strachan T, R.A., ed. *Human Molecular Genetics*. 4 ed. 2010, Garland Science.
240. Kong, A., et al., *Detection of sharing by descent, long-range phasing and haplotype imputation.* Nature Genetics, 2008. **40**(9): p. 1068-1075.
241. Morton, N.E., *Sequential tests for the detection of linkage1.* Am J Hum Genet, 1955. **7**(3): p. 277-318.
242. Badner, J.A., et al., *Genome-wide linkage analysis of 972 bipolar pedigrees using single-nucleotide polymorphisms.* Mol Psychiatry, 2011.
243. Bailey-Wilson, J.E. and A.F. Wilson, *Linkage analysis in the next-generation sequencing era.* Hum Hered, 2011. **72**(4): p. 228-36.
244. E, P., ed. *Color Atlas Of Genetics*. 2006, Non Basic Stock Line.
245. Dawn Teare, M. and J.H. Barrett, *Genetic linkage studies.* Lancet, 2005. **366**(9490): p. 1036-44.
246. Elston, R.C., *Methods of linkage analysis--and the assumptions underlying them [see comment]*. Am J Hum Genet, 1998. **63**(4): p. 931-4.
247. Basu, S., et al., *A likelihood-based trait-model-free approach for linkage detection of binary trait.* Biometrics, 2010. **66**(1): p. 205-13.
248. Kong, A. and N.J. Cox, *Allele-sharing models: LOD scores and accurate linkage tests.* American Journal of Human Genetics, 1997. **61**(5): p. 1179-1188.
249. Taylor, E.W., et al., *Linkage analysis of genetic disorders.* Methods Mol Biol, 1997. **68**: p. 11-25.
250. P., S., *Statistics in human genetics.* 1998, New York: Oxford University Press.
251. Halpern, J.W., A S, *Multipoint Linkage Analysis.* Hum Hered, 1999. **49**: p. 194-196.
252. Strachan T, R.A., ed. *Human Molecular Genetics.* . Genetic mapping of complex characters, ed. R.A. Strachan T. Vol. Chapter 12. 1999, Wiley-Liss: New York.
253. Li, C.C., *Genetic equilibrium under selection.* Biometrics, 1967. **23**(3): p. 397-484.
254. Terwilliger J, O.J., ed. *Handbook of Human Genetic Linkage*. 1994, The Johns Hopkins University Press (1 April 1994): Baltimore, Maryland. 320.
255. Consortium, T.I.H., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
256. Browning, S.R. and B.L. Browning, *Haplotype phasing: existing methods and new developments.* Nature Reviews Genetics, 2011. **12**(10): p. 703-714.

257. Mugavin, M.E., *Multidimensional scaling: a brief overview.* Nurs Res, 2008. **57**(1): p. 64-8.
258. Purcell S, N.B., Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC., *PLINK: a tool set for whole-genome association and population-based linkage analyses*, in *Am J Hum Genet.*2007. p. 559–575.
259. O'Connell, J.R. and D.E. Weeks, *PedCheck: a program for identification of genotype incompatibilities in linkage analysis.* Am J Hum Genet, 1998. **63**(1): p. 259-66.
260. Abecasis, G.R., et al., *GRR: graphical representation of relationship errors.* Bioinformatics, 2001. **17**(8): p. 742-3.
261. Gudbjartsson, D.F., et al., *Allegro, a new computer program for multipoint linkage analysis.* Nature Genetics, 2000. **25**(1): p. 12-13.
262. Dudbridge, F., *A survey of current software for linkage analysis.* Hum Genomics, 2003. **1**(1): p. 63-5.
263. Kruglyak, L., et al., *Parametric and nonparametric linkage analysis: A unified multipoint approach.* American Journal of Human Genetics, 1996. **58**(6): p. 1347-1363.
264. Romero-Hidalgo, S., et al., *GENEHUNTER versus SimWalk2 in the context of an extended kindred and a qualitative trait locus.* Genetica, 2005. **123**(3): p. 235-244.
265. Strauch, K., et al., *Parametric and nonparametric multipoint linkage analysis with imprinting and two-locus-trait models: Application to mite sensitization.* American Journal of Human Genetics, 2000. **66**(6): p. 1945-1957.
266. Abecasis, G.R., et al., *Merlin--rapid analysis of dense genetic maps using sparse gene flow trees.* Nat Genet, 2002. **30**(1): p. 97-101.
267. Zhao, J.H., *Pedigree-drawing with R and graphviz.* Bioinformatics, 2006. **22**(8): p. 1013-4.
268. Kraev, A., et al., *Molecular cloning of a third member of the potassium-dependent sodium-calcium exchanger gene family, NCKX3.* J Biol Chem, 2001. **276**(25): p. 23161-72.
269. Cavaco, B.M., et al., *Mapping a New Familial Thyroid Epithelial Neoplasia Susceptibility Locus to Chromosome 8p23.1-p22 by High-Density Single-Nucleotide Polymorphism Genome-Wide Linkage Analysis.* Journal of Clinical Endocrinology & Metabolism, 2008. **93**(11): p. 4426-4430.
270. Chung, C.C. and S.J. Chanock, *Current status of genome-wide association studies in cancer.* Human Genetics, 2011. **130**(1): p. 59-78.
271. Jones, A.M., et al., *Thyroid cancer susceptibility polymorphisms: confirmation of loci on chromosomes 9q22 and 14q13, validation of a recessive 8q24 locus and failure to replicate a locus on 5q24.* Journal of Medical Genetics, 2012. **49**(3): p. 158-163.
272. Brand, O.J. and S.C.L. Gough, *Immunogenetic Mechanisms Leading to Thyroid Autoimmunity: Recent Advances in Identifying Susceptibility Genes and Regions.* Current Genomics, 2011. **12**(8): p. 526-541.
273. 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, 2009. **146**(6): p. 1073-1080.
274. Reutter, H., et al., *Evidence for Linkage of the Bladder Exstrophy-Epispadias Complex on Chromosome 4q31.21-22 and 19q13.31-41 from a Consanguineous Iranian Family.* Birth Defects Research Part a-Clinical and Molecular Teratology, 2010. **88**(9): p. 757-761.
275. Greenwood, T.A., et al., *Genome-Wide Linkage Analyses of 12 Endophenotypes for Schizophrenia From the Consortium on the Genetics of Schizophrenia.* American Journal of Psychiatry, 2013. **170**(5): p. 521-532.
276. Hegele, R.A., *Copy-number variations and human disease.* American Journal of Human Genetics, 2007. **81**(2): p. 414-415.
277. Marcinkowska, M. and P. Kozlowski, *[The influence of copy number polymorphism on the human phenotype]*. Postepy Biochem, 2011. **57**(3): p. 240-8.
278. Tang, Y.-C. and A. Amon, *Gene Copy-Number Alterations: A Cost-Benefit Analysis.* Cell, 2013. **152**(3): p. 394-405.

279. Sudmant, P.H., et al., *Diversity of Human Copy Number Variation and Multicopy Genes.* Science, 2010. **330**(6004): p. 641-646.
280. Girirajan, S., C.D. Campbell, and E.E. Eichler, *Human Copy Number Variation and Complex Genetic Disease*, in *Annual Review Genetics, Vol 45*, B.L. Bassler, M. Lichten, and G. Schupbach, Editors. 2011. p. 203-226.
281. Bejjani, B.A. and L.G. Shaffer, *Application of array-based comparative genomic hybridization to clinical diagnostics.* J Mol Diagn, 2006. **8**(5): p. 528-33.
282. Bailey, J.A., et al., *Recent segmental duplications in the human genome.* Science, 2002. **297**(5583): p. 1003-1007.
283. Sebat, J., et al., *Large-scale copy number polymorphism in the human genome.* Science, 2004. **305**(5683): p. 525-8.
284. Iafrate, A.J., et al., *Detection of large-scale variation in the human genome.* Nat Genet, 2004. **36**(9): p. 949-51.
285. Tuzun, E., et al., *Fine-scale structural variation of the human genome.* Nat Genet, 2005. **37**(7): p. 727-32.
286. Newman, T.L., et al., *High-throughput genotyping of intermediate-size structural variation.* Human Molecular Genetics, 2006. **15**(7): p. 1159-1167.
287. Eichler, E., et al., *Fine-scale structural variation of the human genome.* Journal of Medical Genetics, 2005. **42**: p. S34-S34.
288. Conrad, D.F., et al., *A high-resolution survey of deletion polymorphism in the human genome.* Nature Genetics, 2006. **38**(1): p. 75-81.
289. Khaja, R., et al., *Genome assembly comparison identifies structural variants in the human genome.* Nature Genetics, 2006. **38**(12): p. 1413-1418.
290. Henrichsen, C.N., E. Chaignat, and A. Reymond, *Copy number variants, diseases and gene expression.* Human Molecular Genetics, 2009. **18**: p. R1-R8.
291. Itsara, A., et al., *De novo rates and selection of large copy number variation.* Genome Research, 2010. **20**(11): p. 1469-1481.
292. Greenway, S.C., et al., *De novo copy number variants identify new genes and loci in isolated sporadic tetralogy of Fallot.* Nature Genetics, 2009. **41**(8): p. 931-U98.
293. Craddock, N., et al., *Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls.* Nature, 2010. **464**(7289): p. 713-U86.
294. Kallioniemi, O.P., et al., *COMPARATIVE GENOMIC HYBRIDIZATION - A RAPID NEW METHOD FOR DETECTING AND MAPPING DNA AMPLIFICATION IN TUMORS.* Seminars in Cancer Biology, 1993. **4**(1): p. 41-46.
295. Miller, O.J. and E. Therman, *Human Chromosomes*. 2001: Springer Verlag.
296. Weiss, M.M., et al., *Comparative genomic hybridisation.* Molecular Pathology, 1999. **52**(5): p. 243-251.
297. Finn, S., et al., *Low-level genomic instability is a feature of papillary thyroid carcinoma - An array comparative genomic hybridization study of laser capture microdissected papillary thyroid carcinoma tumors and clonal cell lines.* Archives of Pathology & Laboratory Medicine, 2007. **131**(1): p. 65-73.
298. Kallioniemi, A., et al., *Gene copy number analysis by fluorescence in situ hybridization and comparative genomic hybridization.* Methods (Orlando), 1996. **9**(1): p. 113-121.
299. Hemmer, S., et al., *DNA copy number changes in thyroid carcinoma.* American Journal of Pathology, 1999. **154**(5): p. 1539-1547.
300. Singh, B., et al., *Screening for genetic aberrations in papillary thyroid cancer by using comparative genomic hybridization.* Surgery, 2000. **128**(6): p. 888-893.
301. Bauer, A.J., et al., *Evaluation of adult papillary thyroid carcinomas by comparative genomic hybridization and microsatellite instability analysis.* Cancer Genetics and Cytogenetics, 2002. **135**(2): p. 182-186.

302. Bentley, D.R., et al., *The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X.* Nature, 2001. **409**(6822): p. 942-943.
303. Cheung, V.G., et al., *Integration of cytogenetic landmarks into the draft sequence of the human genome.* Nature, 2001. **409**(6822): p. 953-958.
304. SolinasToldo, S., et al., *Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances.* Genes Chromosomes & Cancer, 1997. **20**(4): p. 399-407.
305. Ishkanian, A.S., et al., *A tiling resolution DNA microarray with complete coverage of the human genome.* Nature Genetics, 2004. **36**(3): p. 299-303.
306. Carter, N.P., *Methods and strategies for analyzing copy number variation using DNA microarrays.* Nature Genetics, 2007. **39**: p. S16-S21.
307. Locke, D.P., et al., *Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome.* American Journal of Human Genetics, 2006. **79**(2): p. 275-290.
308. Redon, R., et al., *Global variation in copy number in the human genome.* Nature, 2006. **444**(7118): p. 444-454.
309. Wong, K.K., et al., *A comprehensive analysis of common copy-number variations in the human genome.* American Journal of Human Genetics, 2007. **80**(1): p. 91-104.
310. Sharp, A.J., et al., *Segmental duplications and copy-number variation in the human genome.* American Journal of Human Genetics, 2005. **77**(1): p. 78-88.
311. Conrad, D.F., et al., *Origins and functional impact of copy number variation in the human genome.* Nature, 2010. **464**(7289): p. 704-712.
312. Ylstra, B., et al., *BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH).* Nucleic Acids Research, 2006. **34**(2): p. 445-450.
313. Carvalho, B., et al., *High resolution microarray-CGH analysis using spotted oligonucleotides.* Journal of Pathology, 2004. **204**: p. 11A-11A.
314. Zhao, X.J., et al., *An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays.* Cancer Research, 2004. **64**(9): p. 3060-3071.
315. Barrett, M.T., et al., *Comparative genomic hybridization using oligonucleotide microarrays and total genomic DNA.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**(51): p. 17765-17770.
316. Brennan, C., et al., *High-resolution global profiling of genomic alterations with long oligonucleotide microarray.* Cancer Research, 2004. **64**(14): p. 4744-4748.
317. Tonon, G., et al., *High-resolution genomic profiles of human lung cancer.* Proceedings of the National Academy of Sciences of the United States of America, 2005. **102**(27): p. 9625-9630.
318. Selzer, R.R., et al., *Analysis of chromosome breakpoints in neuroblastoma at sub-kilobase resolution using fine-tiling oligonucleotide array CGH.* Genes Chromosomes & Cancer, 2005. **44**(3): p. 305-319.
319. van den Ijssel, P., et al., *Human and mouse oligonucleotide-based array CGH.* Nucleic Acids Research, 2005. **33**(22).
320. Lamy, P., J. Grove, and C. Wiuf, *A review of software for microarray genotyping.* Human Genomics, 2011. **5**(4): p. 304-309.
321. Nannya, Y., et al., *A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays.* Cancer Research, 2005. **65**(14): p. 6071-6079.
322. Illumina. *Datasheet_omni_whole-genome_arrays*. 2013 [cited 2013 31-01]; Available from: http://res.illumina.com/documents/products/datasheets/datasheet_omni_whole-genome_arrays.pdf.
323. Illumina. *Genome-Wide DNA Analysis BeadChips*. 2010 28/07/2010]; Available from: http://www.illumina.com/Documents/products/datasheets/datasheet_infiniumhd.pdf.

324. Pinto, D., et al., *Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants.* Nature Biotechnology, 2011. **29**(6): p. 512-U76.
325. P.K. Janicki, J.L. *Accuracy of allele frequency estimates in pool DNA analyzed by high-density Illumina Human 610-Quad microarray*. 2009. **5**
326. Simon-Sanchez, J., et al., *Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals.* Human Molecular Genetics, 2007. **16**(1): p. 1-14.
327. Erickson, S.W., S.L. MacLeod, and C.A. Hobbs, *Cheek swabs, SNP chips, and CNVs: Assessing the quality of copy number variant calls generated with subject-collected mail-in buccal brush DNA samples on a high-density genotyping microarray*. Bmc Medical Genetics, 2012. **13**.
328. Illumina. *Input Requirements*. 2013; Available from: http://support.illumina.com/sequencing/sequencing_kits/truseq_dna_sample_prep_kit_v2/input_req.ilmn.
329. Dragan, A.I., et al., *Characterization of PicoGreen Interaction with dsDNA and the Origin of Its Fluorescence Enhancement upon Binding*. Biophysical Journal, 2010. **99**(9): p. 3010-3019.
330. Bould, H., et al., *Investigation of thyroid dysfunction is more likely in patients with high psychological morbidity*. Family Practice, 2012. **29**(2): p. 163-167.
331. Butte, A.J., V.J. Dzau, and S.B. Glueck, *Further defining housekeeping, or "maintenance," genes Focus on "A compendium of gene expression in normal human tissues"*. Physiological Genomics, 2001. **7**(2): p. 95-96.
332. NCBI. *Homo sapiens complex locus PLCB1, encoding phospholipase C, beta 1 (phosphoinositide-specific).* 2010; Available from: http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?db=human&c=Gene&l=PLCB1.
333. Altshuler, D., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-1073.
334. Shaikh, T.H., et al., *High-resolution mapping and analysis of copy number variations in the human genome: A data resource for clinical and research applications.* Genome Research, 2009. **19**(9): p. 1682-1690.
335. Pinto, D., et al., *Copy-number variation in control population cohorts.* Human Molecular Genetics, 2007. **16**: p. R168-R173.
336. Visani, G., et al., *SNPs Array Karyotyping Reveals a Novel Recurrent 20p13 Amplification in Primary Myelofibrosis.* Plos One, 2011. **6**(11).
337. Ha, G., et al., *Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer*. Genome Research, 2012. **22**(10): p. 1995-2007.
338. Kobayashi, K., et al., *INCREASED PHOSPHOLIPASE-C ACTIVITY IN NEOPLASTIC THYROID MEMBRANE*. Thyroid, 1993. **3**(1): p. 25-29.
339. Bertagnolo, V., et al., *PLC-beta 2 is highly expressed in breast cancer and is associated with a poor outcome: A study on tissue microarrays.* International Journal of Oncology, 2006. **28**(4): p. 863-872.
340. Pinkel, D., et al., *High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.* Nature Genetics, 1998. **20**(2): p. 207-211.
341. Strassheim, D., et al., *Small cell lung carcinoma exhibits greater phospholipase C-beta 1 expression and edelfosine resistance compared with non-small cell lung carcinoma.* Cancer Research, 2000. **60**(10): p. 2730-2736.
342. Lo Vasco, V.R., et al., *Inositide-specific phospholipase c beta 1 gene deletion in the progression of myelodysplastic syndrome to acute myeloid leukemia*. Leukemia, 2004. **18**(6): p. 1122-1126.
343. Kadamur, G. and E.M. Ross, *Mammalian Phospholipase C*, in *Annual Review of Physiology, Vol 75*, D. Julius, Editor. 2013. p. 127-154.

344. Williams, H.J., et al., *Fine mapping of ZNF804A and genome-wide significant evidence for its involvement in schizophrenia and bipolar disorder.* Mol Psychiatry, 2011. **16**(4): p. 429-41.
345. Vanderpump, M.P.J., et al., *The incidence of thyroid disease in the community: A twenty year follow up to the Whickham survey.* Journal of Endocrinology, 1994. **140**(SUPPL.): p. OC16-ABSTRACT OC16.
346. Manole, D., et al., *Estrogen promotes growth of human thyroid tumor cells by different molecular mechanisms.* Journal of Clinical Endocrinology & Metabolism, 2001. **86**(3): p. 1072-1077.
347. Bullock, M., et al., *Association of FOXE1 polyalanine repeat region with papillary thyroid cancer.* J Clin Endocrinol Metab, 2012. **97**(9): p. E1814-9.
348. Dunham, I., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
349. Ngan, E.S., et al., *A germline mutation (A339V) in thyroid transcription factor-1 (TITF-1/NKX2.1) in patients with multinodular goitre and papillary thyroid carcinoma.* J Natl Cancer Inst, 2009. **101**(3): p. 162-75.
350. Kozomara, A. and S. Griffiths-Jones, *miRBase: integrating microRNA annotation and deep-sequencing data.* Nucleic Acids Research, 2011. **39**: p. D152-D157.
351. Hofacker, I.L., B. Priwitzer, and P.F. Stadler, *Prediction of locally stable RNA secondary structures for genome-wide surveys.* Bioinformatics, 2004. **20**(2): p. 186-190.
352. Knudsen, N., et al., *Risk factors for goitre and thyroid nodules.* Thyroid, 2002. **12**(10): p. 879-888.
353. Prossnitz, E.R., J.B. Arterburn, and L.A. Sklar, *GPR30: A G protein-coupled receptor for estrogen.* Mol Cell Endocrinol, 2007. **265-266**: p. 138-42.
354. Chen, G.G., et al., *Regulation of cell growth by estrogen signaling and potential targets in thyroid cancer.* Current Cancer Drug Targets, 2008. **8**(5): p. 367-377.
355. Lin, C.Y., et al., *Whole-genome cartography of estrogen receptor alpha binding sites.* Plos Genetics, 2007. **3**(6): p. 867-885.
356. Kero, J., et al., *Thyrocyte-specific G(q)/G(11) deficiency impairs thyroid function and prevents goitre development.* Journal of Clinical Investigation, 2007. **117**(9): p. 2399-2407.