

Genetic Studies into Rare Diseases and Cancer Using Next Generation Sequencing Technologies

By

Naser M Ali

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Institute of Cancer and Genomic Sciences

School of Medicine & Dentistry

The University of Birmingham

September 2018

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

Rare genetic diseases (RGDs) and rare sporadic cancers are often considered as two separate groups of diseases. Nevertheless, both groups share the same burden: their rarity and the challenges in diagnosis and treatment, thus affecting the wellbeing of many patients and their families around the world. Although next generation sequencing (NGS) technologies have revolutionised the genetic landscape of RGDs and cancers, many patients with these diseases are still without a definitive molecular diagnosis. In this thesis, NGS was conducted on congenital hypothyroidism (CHT) families (an example of an RGD) and three rare bone cancers, aiming to expand the understanding of the genetic and pathogenic mechanisms of these diseases.

To identify known or novel disease-causing genes, WES was conducted on four families with CHT. In one family, a homozygous candidate variant in *SIX2* was identified, and subsequent functional characterisation experiments and family segregation analyses were performed. After more family members were included, the *SIX2* variant did not segregate with the disease in the family and, therefore, was classified as unlikely disease causing.

WES and RNA sequencing (RNA-Seq) were conducted on three rare bone tumours: undifferentiated pleomorphic sarcoma of bone (UPSb), adamantinoma and osteofibrous dysplasia (OFD)-like adamantinoma. In UPSb tumours, 31 genes were recurrently mutated, including *TP53* in 4/14 samples (29%), and chromatin remodelling genes (*ATRX*, *H3F3A*, *DOT1L*) in 5/14 samples (36%). In addition, two previously reported gene fusions (*CLTC-VMP1* and *FARP1-STK24*) were identified in these tumours. In adamantinoma tumours, *KMT2D*, a histone methyltransferase, was recurrently mutated in 2/8 adamantinomas (25%). In addition, a cancer-predisposing germline fusion (*KANSL1-ARL17A*) was identified in 4/6 adamantinoma (66.7%) and in 3/4 OFD-like adamantinoma (75%) tumours.

This thesis is a practical example demonstrating how rare diseases and cancers can be investigated using the same high-throughput techniques. Moreover, the three bone tumour studies represent the first comprehensive WES and RNA-Seq analyses conducted on these tumours, revealing novel molecular insights that can be translated into clinical practices to enhance the diagnosis, prognosis and the outcomes of patients with these diseases.

Acknowledgements

First, I would like to sincerely thank my co-supervisors Prof Farida Latif and Prof Timothy Barrett for all the help, guidance and support they offered during my PhD. None of this work would have been possible without their supervision. I would also like to thank my secondary supervisors, Dr Mark Morris and Dr Kristien Boelaert. A special thank you goes to Dr Mark Morris for being very supportive whenever needed and for his helpful comments on the thesis. Big thanks to Dr Vaiyapuri Sumathi for providing the tumour samples and their clinical information.

A big thank you goes to Dr Malgosia Zatyka and Dr Dewi Astuti for their assistance during my research and for teaching me various laboratory techniques. I also would like to thank Dr Masahiro Otso for sharing his laboratory expertise and for the wonderful chats about Japan.

I would like to thank everyone in the Medical and Molecular Genetics Department, past and present, for their help, chats and support during my PhD. Special thanks go to all my friends in the lab and office, including Alamin Mohammed, Stefania Niada, Ghazala Begum, Abdullah Alholle, Amy Slater, Uncaar Boora, Jan Hatfield and Kris Hepherd.

Very special thanks go to all my family members and friends in Kuwait. A special thank you goes to my friend Abdulmuhsen Alharbi. I would like to acknowledge the Kuwait Government (Ministry of Health) for providing generous funding for this research as well as the Kuwait Medical Genetics Center for nominating me to do this PhD.

Finally, I would like to extend my eternal gratitude to my mum, thanks a lot for your support and encouragement throughout my PhD.

Table of Contents

Chapter 1: Introduction 1

1.1 Thesis overview 1
1.2 Human disease: overview and Mendelian modes of inheritance 1
1.3 The complexity of human diseases: penetrance, variable expressivity, digenic and oligogenic inheritance 3
1.3.1 Consanguinity and increased risk of genetic malformations 5
1.4 Rare diseases: definition, complexity and current understanding 6
1.5 Cancer 10
1.5.1 Cancer incidence and mortality 11
1.5.2 Cancer aetiology: a genetic disease 12
1.5.2.1 Cancer driver and passenger mutations 16
1.5.2.2 Oncogenes 17
1.5.2.3 Tumour suppressor genes and the Knudson two-hit hypothesis 18
1.5.3 Primary tumours of bone 20
1.5.3.1 Incidence and survival of bone sarcomas 21
1.5.3.2 Clinical presentation, diagnosis and treatment of bone sarcomas 23
1.5.3.3 Overview of the genetics of bone sarcomas 24
1.6 Identification of the genetic basis of disease: disease-causing gene discovery 25
1.6.1 Classical cytogenetic, fluoresce *in situ* hybridisation and array CGH 26
1.6.2 Sanger sequencing of candidate genes and positional cloning approach 28
1.7 NGS technologies and human disease 32
1.7.1 Targeted sequencing and WGS 33
1.7.2 WES: definition, experimental process and details of Illumina sequencing platform 35
1.7.3 Disease gene identification strategies using WES: bioinformatics analyses and prioritisation of variants 40
1.7.3.1 Analytical frameworks that can be applied in disease-gene discovery of RGDs using WES 41
1.7.3.2 Identification of somatic cancer variants using WES 45
1.7.4 RNA-Sequencing: overview and technology workflow 48
1.8 Aims of the research presented in this thesis 52

Chapter 2: Material and Methods 53

2.1 Patient material samples 53
2.1.1 CHT samples 53
2.1.2 UPSb of the bone samples 54
2.1.3 Adamantinoma and OFD-like adamantinoma samples 54
2.2 Nucleic acid extraction, quantification and quality assessment 55
2.2.1 Tumour tissue disruption and homogenization 55
2.2.2 DNA extraction from fresh frozen sporadic cancer tissues 55
2.2.2.1 DNA extraction from FFPE cancer tissues 56
2.2.3 RNA extraction from sporadic cancer tissues 57
2.2.3.1 RNA extraction using the TRIzol-chloroform method 57
2.2.3.1.1 DNase treatment of RNA extracted using the TRIzol-chloroform method 58

2.2.3.2 Second method for RNA extraction from cancer tissues: RNAeasy Mini kit 58
2.2.4 Quantification and quallity assessment of DNA/RNA nucleic acids using Nanodrop 59
2.3 WES of CHT and three bone cancer projects 60
2.3.1 Sample preparation for WES 60
2.3.2 Exon Capture: selection of protein-coding regions....60
2.3.3 Bioinformatic tools used in WES data analyses....61
2.3.3.1 SIFT and PolyPhen2 missense prediction tools 61
2.3.3.2 IGV tool for visualising NGS data....62
2.3.4 Raw data analysis pipeline of WES: alignment and PCR duplicates ...62
2.3.5 Variant calling and annotation in CHT and bone cancer projects....63
2.3.5.1 CHT project 63
2.3.5.2 VarScan2 and MuTect tools for somatic variant calling in cancer projects 63
2.3.6 Filtering WES data to identify somatic candidate variants in the bone cancer projects....64
2.3.6.1 Overview and default variant allele frequency cut-off....64
2.3.6.2 Filtering and visualisation of variants in normal-paired tumours....65
2.3.6.3 Assessment of the established filtering criteria in UPSb, adamantinoma and OFD-like adamantinoma tumours using Sanger sequencing 66
2.3.6.4 Additional rigorous filtering steps applied to normal-unpaired UPSb tumours 67
2.4 RNA-Seq studies on bone tumours....68
2.4.1 Overview: RNA samples and sample quality check using Bioanalyzer 68
2.4.2 RNA-Seq library preparation and sequencing....68
2.4.3 Bioinformatic tools used in RNA-Seq data analyses....70
2.4.3.1 BaseSpace Sequence Hub and TopHat2 alignment tool for detection of gene fusions 70
2.4.3.1.1 The Tuxedo Suite by Illumina: TopHat2 Alignment....70
2.4.4 Data analysis of RNA-Seq data 71
2.4.4.1 Filtering and prioritisation of non-recurrent candidate gene fusions .71
2.5 Laboratory validation of WES data....72
2.5.1 Standard PCR for validation of variants: primer design criteria....72
2.5.1.1 Standard PCR reaction components and optimization 73
2.5.1.2 Agarose gel electrophoresis: visualisation of PCR products 74
2.5.2 Sanger sequencing of PCR amplified products 74
2.5.2.1 Preparation and clean-up of PCR products using microCLEAN....74
2.5.2.1.1 Sanger Sequencing (chain termination sequencing) reaction (Part 1) 75
2.5.2.1.2 Ethanol precipitation of Sanger sequencing reactions (Part 2) ...75
2.5.2.2 Analysis and visualisation of Sanger sequencing data....76
2.5.3 Laboratory validation of RNA-Seq data....77
2.5.3.1 RT-PCR primer design and method 77
2.5.3.1.1 cDNA synthesis of bone cancer samples....77
2.5.3.1.2 RT-PCR primer design and technique 77
2.5.3.2 Long range PCR as a method for genomic validation of gene fusions 78

2.6 Online sources, databases and browsers used in genome and variant information 79
2.6.1 Ensembl Genome Browser 79
2.6.2 UCSC Genome Browser by the University of California, Santa Cruz 79
2.6.3 GeneCards and the NCBI website 80
2.6.4 COSMIC and inTOgen databases 80
2.6.5 The Drug Gene Interaction database to identify drug targeted genes 81
2.6.6 ExAc browser 81
2.6.7 Ingenuity Pathway Analysis 81
2.7 Molecular gene cloning 82
2.7.1 Obtaining of *SIX2* IMAGE and extraction and purifying plasmid 82
2.7.2 Site-directed mutagenesis to introduce *SIX2* candidate variant 83
2.7.2.1 Site-directed mutagenesis kit and mutagenic primer design 83
2.7.2.2 SDM exponential PCR amplification 85
2.7.2.3 Kinase, Ligase & DpnI treatment and transformation 86
2.7.2.4 Verification of SDM reaction 86
2.7.3 PCR amplification of *SIX2* Wild-type and *SIX2*-Mutant ORFs using primers with restriction sites and high-fidelity polymerase 87
2.7.3.1 Purification of amplified PCR products 90
2.7.3.2 Restriction digestion and ligation of ORFs to plasmid vectors 91
2.7.3.3 Transformation of bacterial cells with plasmid vectors ligated with ORF 91
2.7.4 Small-scale plasmid DNA extraction and purification from bacterial cells using Miniprep kit 92
2.7.4.1 Verification of DNA plasmids after digestion and ligation 93
2.7.5 Large-scale plasmid DNA extraction and purification using the Plasmid Maxi Kit 94
2.8 Mammalian cell culture and manipulation 95
2.8.1 General tissue culture 95
2.8.2 Culture media preparation, cell maintenance and propagation 95
2.8.2.1 Culture media preparation 95
2.8.2.2 Passage of confluent cells 96
2.8.2.3 Preparation of LB agar plates and broth medium 96
2.8.2.4 Freezing cells for prolonged storage 97
2.8.2.5 Thawing and reviving of frozen cells 97
2.8.2.6 Cell counting using haemocytometer 97
2.9 Transfection of Cells 98
2.9.1 Harvesting of transfected cells for protein extraction 99
2.10 Protein analysis using Western blot 99
2.10.1 Protein quantification 99
2.10.2 Sodium dodecyl sulfate polyacrylamide gel electrophoresis for protein separation in WB 100
2.10.2.1 Loading of protein samples and running gel 101
2.10.3 Protein transfer to nitrocellulose membrane 101
2.10.4 Immunodetection of protein blotted to membrane 102
2.10.5 Developing western blots using the enhanced chemiluminescence method 103
2.10.6 Quantitative analysis of protein expression levels 103
2.11 Immunocytochemistry to assess protein cellular localisation 104
2.11.1 Cell fixation 104

2.11.2 Blocking and staining of cells to be visualized under fluorescent microscopy 105
Chapter 3: Genetic studies into congenital hypothyroidism families using whole exome sequencing 106
3.1 Introduction 106
3.1.1 CHT Definition, overview and neonatal screening 106
3.1.2 Epidemiology 107
3.1.3 Classification and subtypes: transient versus primary and secondary hypothyroidism 107
3.1.4 Complexity of the aetiology of CHT cases 108
3.1.5 Thyroid gland dysgenesis 109
3.1.5.1 Genetics of thyroid gland dysgenesis 109
3.1.6 Thyroid dyshormonogenesis 110
3.1.6.1 Genetics of thyroid dyshormonogenesis 110
3.2 Aims of the study and the clinical information of CHT families 114
3.3 WES data analysis and results 115
3.3.1 Quality check metrics and investigation into variants in CHT-associated genes 115
3.3.2 Filtering of WES data to identify potential disease-causing candidate variants 121
3.3.2.1 Ranking candidate genes based on expression data, gene biological function and animal model databases 123
3.3.3 The ranked 4 candidate variant: a homozygous missense change identified in *SIX2* in a TDH Family-1 124
3.4 Discussion 129
3.4.1 The identification of the p.Val287Met *SIX2* candidate variant in a male patient with posterior urethral valve 129
3.4.1.1 the increased risk of congenital renal anomalies in children with CHT. 130
3.4.2 Overview of development, differentiation and morphogenesis of the thyroid gland in vertebrates 130
3.4.3 *SIX2*: tissue expression, role in human disease, and biological function 131
3.4.3.1 Expression of *SIX2* in human and vertebrates 131
3.4.3.2 The molecular function of *SIX2* in thyroid-related tissues ` 132
3.4.3.3 Association of *SIX2* mutations with renal defects and assessment of renal function in Family-1 133
3.4.3.4 Possible interactions of *Six2* with *Eya1*, *Pax8* and *Nkx2.5* in mice. 134
3.5 Summary and conclusion 137
Chapter 4: Investigations into *SIX2* candidacy in thyroid gland dysgenesis using functional experiments and family segregation analysis 139
4.1 Introduction and aims 139
4.2 Cloning and site-directed mutagenesis of *SIX2* 140
4.3 Results 141
4.3.1 Experiment 1: protein expression assessment by western blot 141
4.3.1.1 Results of Experiment 1: western blotting 142
4.3.2 Experiment 2: subcellular localisation of the *SIX2* candidate variant versus *SIX2* wild type using immunocytochemistry 144

4.3.3 Availability of other unaffected family members of CHT Family-1 under study 145
4.3.3.1 Assessing the segregation of rank-3 candidate genes in CHT Family-1 147
4.4 Discussion....149
4.5 Chapter summary....152
4.6 Future work....153
Chapter 5: Genetic and transcriptomic profiling of undifferentiated pleomorphic sarcoma of bone using NGS technologies....154
5.1 Introduction....154
5.1.1 UPSb tumours....154
5.1.2 NGS technologies in the field of cancer....156
5.2 Aims of the study....156
5.3 WES results and data analysis....157
5.3.1 WES samples information and somatic variant calling tools....157
5.3.2 Quality check metrics of WES data using FastQC tool....158
5.3.3 *IDH1/2* hotspot mutations....159
5.3.4 Filtering and identification of somatic candidate variants in WES data 159
5.3.5 Identification of recurrent candidate genes in UPSb samples....162
5.3.5.1 Investigating recurrent genes in COSMIC cancer gene census, inTOgen and the drug interaction databases....167
5.4 RNA sequencing data analysis/processing and results....167
5.4.1 Overview of RNA-Seq experiment and processing of RNA-Seq raw reads 167
5.4.2 RNA-Seq data and identification of gene fusions....168
5.4.2.1 *CLTC-VMP1* gene fusion in UPSb-T13....173
5.4.2.2 *FARP1-STK24* gene fusion in UPSb-T6....173
5.4.3 Identification of fusion breakpoints at the genomic level using long-range PCR....174
5.4.3.1 Genomic mapping of *CLTC-VMP1* gene fusion breakpoints in UPSb-T13 178
5.4.3.2 Genomic mapping of *FARP1-STK24* fusion breakpoints in UPSb-T6 179
5.5 Summary of findings....182
5.6 Discussion....183
5.6.1 Discussion part one: WES findings....183
5.6.1.1 *TP53* cancer-associated gene....186
5.6.1.1.1 Overview of *TP53* mutation frequency in cancer....186
5.6.1.1.2 Four somatic *TP53* variants identified in four UPSb samples by WES 187
5.6.1.1.3 Localisation of recurrent *TP53* variants in p53 DNA-binding domain and their biological relevance in literature....189
5.6.1.2 *ATRX* cancer-associated gene....190
5.6.1.2.1 Overview....190
5.6.1.2.2 Potential biological relevance of recurrent *ATRX* variants identified in UPSb tumours....191
5.6.1.3 *H3F3A* gene....192
5.6.1.4 Limitations of part 1 WES study and WES technology....194

5.6.2 Discussion part two: RNA-Seq findings 196
5.6.2.1 *CLTC-VMP1* gene fusion in sample UPSb-T9....197
5.6.2.1.1 Potential tumour suppressor role of *CLTC-VMP1* 198
5.6.2.2 *FARP1-STK24* fusion identified in UPSb-T6 200
5.6.2.2.1 Potential apoptotic and oncogenic roles of *FARP1* and *STK24* genes 201
5.6.2.3 Potential targeted therapeutics for fusion genes composed of protein kinase genes 205
5.6.2.4 Limitations of RNA-Seq technology....205
5.7 Future work....207
5.8 Peer reviewed publications 208
Chapter 6: Genetic and transcriptomic analyses of adamantinoma and osteofibrous dysplasia-like adamantinoma bone tumours using NGS technologies 209
6.1 Introduction 209
6.2 Aim of the study 211
6.3 Part 1: WES data results and analysis....212
6.3.1 Details of WES samples, somatic variant callers and quality check metrics 212
6.3.2 Filtering of WES variants to identify somatic candidate variants....213
6.3.3 Recurrent candidate gene analysis in adamantinoma and OFD-like adamantinoma tumours....214
6.3.4 Ingenuity Pathway Analysis for somatic candidate genes identified in adamantinoma and OFD-like samples 216
6.4 Part 2: RNA sequencing data analyses and results 217
6.4.1 Overview of the RNA-Seq experiment and data analysis 217
6.4.2 Identification of a recurrent fusion, *KANSL1-ARL17A,* in adamantinoma and OFD-like adamantinoma tumours....220
6.4.3 Prioritising non-recurrent gene fusion candidates in adamantinoma and OFD-like adamantinoma samples 221
6.4.3.1 *EPHB4-MARCH10* gene fusion in ADA-T4 sample: RT-PCR and LR-PCR analyses....225
6.5 Summary of findings: WES and RNA-Seq 229
6.6 Discussion....231
6.6.1 *KMT2D* recurrent gene mutations identified in adamantinoma tumours by WES 231
6.6.1.1 Overview and association with cancer 231
6.6.1.2 *KMT2D* variants identified in adamantinoma and OFD-like adamantinoma tumours....232
6.6.2 *EPHB4-MARCH10* gene fusion identified by RNA-Seq 234
6.6.2.1 *EPHB4-MARCH10* effect on protein domains organisation and the dual functionality of EPHB4 in cancer....236
6.7 Remarks and future work....240
Chapter 7: General discussion....242
7.1 WES in studying RGDs, an example being CHT 243
7.1.1 Rare diseases and accelerating disease-gene discovery 245
7.2 WES for studying rare bone tumours 246
7.2.1 UPSb 246

7.2.2 Adamantinoma and OFD-like adamantinoma 249
7.3 Final conclusions 252
Chapter 8: Appendix and supplementary information 254
8.1 QC metrics for WES data of the CHT and bone tumours projects 254
8.1.1 WES FastQC QC scores 254
8.2 VarScan2 for the detection of somatic alterations in bone tumour projects 260
8.3 Identification of somatic point variations using MuTect 261
8.4 RNA sequencing 263
8.4.1 RNA-Seq data processing: overview of data QC of reads for bone tumour projects 263
8.4.2 RNA-Seq read alignment and identification of gene fusions using TopHat-Fusion and STAR-Fusion 264
8.5 Supplementary figures and tables 272
8.5.1 Chapter 3: CHT families study using WES 272
8.5.2 Chapter 5: UPSb study using WES and RNA-Seq 273
8.5.3 Chapter 6: Adamantinoma and OFD-like adamantinoma study using WES and RNA-Seq 283
8.6 List of primers used in confirmational experiments of NGS data 286
Chapter 9: Reference list 294

List of Figures

Figure 1–1: A diagrammatic representation showing the increased risk of autosomal recessive diseases in a consanguineous family 7
Figure 1–2: A recommended scheme for rare disease classifications and recommended steps for their evaluation 9
Figure 1–3: The proportion of different cancer cases, classified by cancer anatomical site in 1993 (observed; left), 2014 (observed; middle) and 2035 (projected; right), and split by males (top) and females (bottom) 13
Figure 1–4: The proportion of cancer deaths, classified by cancer anatomical site in 1993 (observed; left), 2014 (observed; middle) and 2035 (projected; right), and split by males (top) and females (bottom) 14
Figure 1–5: An example of the mutation distributions in oncogenes (*PIK3CA, IDH1*) and tumour suppressor genes (*RB1, VHL*) 19
Figure 1–6: The Knudson two-hit hypothesis in both hereditary (familial) and non-hereditary (sporadic) retinoblastoma cases 21
Figure 1–7: The incidence by age for the osteosarcoma, chondrosarcoma, Ewing sarcoma and chordoma in England between 1985–2009 22
Figure 1–8: Proportion of bone sarcoma cases diagnosed by age group and tumour site 23
Figure 1–9: Schematic representation of the main steps in the positional cloning approach to identify disease-causing genes 30
Figure 1–10: A schematic depiction of the multisteps involved in Sanger sequencing technology using capillary electrophoresis 31
Figure 1–11: Approximation of the number of genes discovered by conventional methods versus WES and WGS since 2010 34
Figure 1–12: The cost of genome sequencing in comparison to Moore's law 35
Figure 1–13: A schematic representation of the major steps in a WES experiment 37
Figure 1–14: The SureSelect target enrichment system to capture exons during WES library preparation 38
Figure 1–15: Diagrammatic representation of the Illumina massive-parallel sequencing platform 39
Figure 1–16: Common prioritisation scheme for WES variants 42
Figure 1–17: Analytical frameworks for disease-gene identification using WES 44
Figure 1–18: Two WES analytical frameworks for analysing autosomal dominant and *de novo* disorders 47
Figure 1–19: Identification of somatic variants using WES data from tumour and corresponding normal samples 48
Figure 1–20: Schematic representation of a standard RNA-Seq experiment. RNA molecules are first converted to a library of cDNA fragments 51
Figure 2–1: Examples of RIN value measurement in UPS-T2 and UPS-T14 samples using Agilent 2100 Bioanalyzer and Agilent Nano chips 69
Figure 2–2: Molecular gene cloning procedure 84
Figure 2–3: Flag-tagged CMV-4 (pFLAG-CMV-4) vector map including the multiple cloning sites 88
Figure 2–4: Untagged pCDNA3.1 (with CMV4 promoter) overexpression vector map showing detailed vector components and sequences 89
Figure 3–1: Schematic representation of key elements involved in the synthesis of thyroid hormone of a follicular thyroid cell 112

Figure 3–2: The pedigree of CHT families. All families were associated with TGD, except for Family-2 which was a TDH 118
Figure 3–3: Filtering and prioritisation scheme of WES data in CHT families to prioritise potential CHT candidate variants 122
Figure 3–4: *SIX2* candidate variant visualised using the Integrative Genomic Viewer (IGV) 125
Figure 3–5: Sanger sequencing confirmation of *SIX2* variant (c.859G>A; p.Val287Met) in Family-1 128
Figure 3–6: High conservation of the substituted valine amino acid in the *SIX2* candidate variant (c.859G>A; p.Val287Met) 128
Figure 3–7: *Pax8* expression in the metanephric mesenchyme of mice embryos at E10.5-11.5 136
Figure 3–8: Expression of *Nkx2.5* in the murine *Six*$^{-/-}$ stomach. (A) *Nkx2.5* expression in wild type pyloric sphincter territory 136
Figure 4–1: Protein expression of *SIX2*-wild type and *SIX2*-Mut plasmids 143
Figure 4–2: COS7 cells fixed and stained with anti-FLAG conjugated antibody (x40 mag), contrast adjusted 146
Figure 4–3: Sanger sequencing analysis of *SIX2* candidate variant in four unaffected members of Family-1 147
Figure 5–1: The prioritization scheme used to filter and prioritise WES variants identified in UPSb tumours 161
Figure 5–2: The total number of filtered WES alterations identified in 14 UPSb samples. UPSb-T12* is an FFPE tumour. INDELs: small insertions/deletions; SNVs: single nucleotide variants. 162
Figure 5–3: Filtering and prioritization scheme of gene fusions identified by RNA-Seq in eight UPSb tumours 171
Figure 5–4: The distribution of junction/supporting reads for (A) *CLTC-VMP1* and (B) *FARP1-STK24* gene fusions 175
Figure 5–5: Schematic representation and RT-PCR confirmation of *CLTC-VMP1* fusion in UPSb-T13 176
Figure 5–6: Schematic representation and RT-PCR confirmation of *FARP1-STK24* fusion in UPSb-T6 177
Figure 5–7: Genomic representation of *CLTC-VMP1* gene fusion and LR-PCR confirmation of the genomic breakpoints 180
Figure 5–8: Genomic representation of *FARP1-STK24* gene fusion and LR-PCR confirmation of genomic/DNA breakpoints 181
Figure 5–9: Distribution of *TP53* alterations across 20 top-mutated cancer projects 187
Figure 5–10: The location of the *TP53* somatic variants in relation to p53 functional domains 190
Figure 5–11: Multi domains of the ATRX protein 192
Figure 5–12: The location of the *H3F3A* variants in the N-terminal tail of H3.3 194
Figure 5–13: Canonical splicing of *CLTC-VMP1* gene fusion and the affected functional biological domains of the fusion gene partners 199
Figure 5–14: The impact of *FARP1-STK24* fusion on the functional domains of FARP1 and STK24 proteins. 203
Figure 6–1: The total number of filtered WES variants identified in adamantinoma and OFD-like adamantinoma tumours 215
Figure 6–2: RT-PCR confirmation of *KANSL1-ARL17A* fusion transcript in six adamantinoma and four OFD-like adamantinoma tumours 222

Figure 6–3: Schematic representation of the *KANSL1-ARL17A* fusion transcript ...223
Figure 6–4: Filtering and prioritisation steps used to identify genuine somatic gene fusions in adamantinoma and OFD-like adamantinoma tumours.....224
Figure 6–5: Schematic representation and RT-PCR confirmation of *EPHB4-MARCH10* fusion in ADA-T4 tumour.....227
Figure 6–6: Genomic characterisation and LR-PCR validation of *EPHB4-MARCH10* 228
Figure 6–7: The distribution *KMT2D* mutation across 20 cancer projects 233
Figure 6–8: Schematic of *KMT2D* alterations and their relative protein positions in adamantinoma and OFD-like adamantinoma tumours.....235
Figure 6–9: *EPHB4-MARCH10* gene fusion effect on the protein domain organization of the fusion gene partners.....237
Figure 6–10: A schematic representation of the domain structure of EPHB4 and ephrin ligand.....239
Figure 8–1: A subset of the QC metrics of the sequence quality part in USPb-T1 as an example.....258
Figure 8–2: Combined FastQC sequence quality graphs for all WES of eight normal-paired UPSb samples.....259
Figure 8–3: VarScan2 pipeline to identify somatic alterations in tumour/normal paired samples.....262
Figure 8–4: Overview of the MuTect analyses steps for somatic SNVs detection...265
Figure 8–5: TopHat Alignment bundle tool workflow used to analyse RNA-Seq data and identify gene fusions.....266
Figure 8–6: TopHat-Fusion algorithm pipeline for the identification of gene fusions270
Figure 8–7: Overview of STAR-Fusion workflow for gene fusion detection.....271
Figure 8–8: Visualising somatic variants using the Integrative Genomics Viewer (IGV) to identify genuine somatic calls from false positive and/or germline ones 274
Figure 8–9: RT-PCR somatic confirmation of six gene fusions identified in UPSb tumours at the cDNA level.....281
Figure 8–10: LR-PCR confirmation of three gene fusions identified in UPSb tumours 282

List of Tables

Table 2–1: Steps of standard PCR reaction using a thermocycler 73
Table 2–2: Sanger sequencing reaction protocol using a thermal cycler 75
Table 2–3: Thermocycler programme for cDNA synthesis from RNA template. 77
Table 2–4: The 3-step thermocycling protocol for LR-PCR reactions.....79
Table 2–5: Mutagenic primers used to introduce *SIX2* candidate variant during SDM 85
Table 2–6: SDM PCR amplification programme. 85
Table 2–7: Primers with added restriction sites to amplify *SIX2*-WT and *SIX2*-Mut ORFs.....90
Table 3–1: Genes that have been implicated in TGD.....111
Table 3–2: Clinical information of the four CHT-affected families.....117
Table 3–3: Exons of CHT-associated gene that were not adequately covered by WES.....119
Table 3–4: Variants detected in CHT-associated genes in affected members. 120
Table 3–5: The ranking scheme applied to prioritise AR candidate genes identified in CHT families.....126
Table 3–6: The number of potential AR candidates in the four families in each ranked category 126
Table 4–1: Mean of quantified protein expression of *SIX2*-Wildtype and *SIX2*-Mut plasmids from three independent western blot experiments 143
Table 4–2: Sanger sequencing analysis of rank-3 candidate genes in TGD-affected (Family-1-A) and an unaffected sister (Family-1-S2).....148
Table 5–1: Recurrent candidate genes identified in 14 UPSb samples using WES 164
Table 5–2: Sanger sequencing confirmation of WES variants in UPSb tumours.....166
Table 5–3: Literature and database search to prioritise validated UPSb somatic gene fusions.....172
Table 5–4: WES and RNA-Seq genetic landscape in 14 UPSb samples 185
Table 5–5: COSMIC findings of the recurrent *TP53* somatic variants identified in four UPSb samples 188
Table 6–1: Details and Sanger sequencing confirmation of WES variants detected in *KMT2D* 218
Table 6–2: Canonical and bio function pathways affected in adamantinoma and OFD-like adamantinoma tumours 219
Table 6–3: WES and RNA-Seq alterations landscape in adamantinoma and OFD-like adamantinoma tumours.....230
Table 8–1: FastQC read alignment QC metrics of the WES data of the four CHT families 255
Table 8–2: FastQC read alignment QC metrics of the WES data for UPSb samples 256
Table 8–3: FastQC read alignment QC metrics of the WES data for adamantinoma and OFD-like adamantinoma samples 257
Table 8–4: Details of the generated RNA-Seq reads in eight UPSb tumours.....267
Table 8–5 Details of the generated RNA-Seq reads in five adamantinoma and three OFD-like adamantinoma tumours 268
Table 8–6: List of the autosomal recessive candidate genes identified by WES in the four CHT families 273
Table 8–7: The total number of unfiltered variants identified in UPSb.....273

Table 8–8: Sanger sequencing confirmation of WES variants identified in UPSb tumours 279
Table 8–9: Somatic gene fusions identified by TopHat-Fusion and STAR-Fusion and validated by RT-PCR in UPSb tumours.....280
Table 8–10: The total number of unfiltered WES variants in adamantinoma and OFD-like adamantinoma 283
Table 8–11: The coding mutation rate of adamantinoma and OFD-like adamantinoma tumours 284
Table 8–12: Gene fusions identified in adamantinoma and OFD-like adamantinoma tumours by RNA-Seq 285
Table 8–13: List of the Sanger primers used in the WES study of four CHT families. 286
Table 8–14: Sanger sequencing primers used in the validation of WES data of UPSb samples.....290
Table 8–15: Sanger sequencing primers used in the validation of WES data of adamantinoma and OFD-like adamantinoma tumours. 291
Table 8–16: List of primers used in RT-PCR confirmation of gene fusions (RNA-Seq) detected in UPSb, adamantinoma and OFD-like adamantinoma tumours. 291
Table 8–17: List of primers used in LR-PCR confirmation of gene fusions (RNA-Seq) identified in UPSb and adamantinoma tumours 293

List of Abbreviations

aCGH	array comparative genomic hybridisation
AD	autosomal dominant
ADA	adamantinoma sample
ALT	alternative elongation of telomeres
AME	autosomal monoallelic expression
AR	autosomal recessive
BSA	Bovine Serum Albumin
CCGC	COSMIC Cancer Gene Census
cDNA	complementary DNA
CHT	congenital hypothyroidism
CNVs	copy number variations
COSMIC	The Catalogue of Somatic Mutations in Cancer
CR	chloramphenicol
DGIdb	The Drug Gene Interaction Database
DMEM	Dulbecco's Modified Eagles Media
DMSO	dimethyl sulphoxide
ds-DNA	double-stranded DNA
ECL	enhanced chemiluminescence
Eph	erythropoietin-producing hepatocellular carcinoma
ExAC	the Exome Aggregation Consortium
EZH2	enhancer zeste homologue 2
Family-1-A	affected member of Family 1
Family-1-B	unaffected brother in Family 1
Family-1-S1	unaffected sister 1 in Family 1
Family-1-S2	unaffected sister 2 in Family 1
Family-1-S3	unaffected sister 2 in Family 1
FBS	Fetal Bovine Serum
FFPE	formalin-fixed paraffin-embedded
FISH	fluorescence *in situ* hybridisation
H3K4	histone H3 at the lysine 4 residue
IGV	Integrative Genomics Viewer
INDELs	insertions/deletions

inTOgen	Integrative Onco Genomics
IPA	Ingenuity Pathway Analysis
KLD	Kinase, Ligase & DpnI
KMTs	lysine methyltransferases
LB	Luria Bertani
LR-PCR	long range PCR
MAF	minor allele frequency
MRI	magnetic resonance imaging
NCBI	the National Centre for Biotechnology Information
NF1	neurofibromatosis type 1
NGS	next generation sequencing
OFD	osteofibrous dysplasia-like
OFD-like-ADA	OFD-like adamantinoma sample
OGT	Oxford Gene Technology
ORF	open reading frame
PBS	phosphate buffered saline
PolyPhen2	Polymorphism Phenotyping v2
PRC2	Polycomb repressive complex 2
QC	quality check
RB1	retinoblastoma
RGDs	rare genetic diseases
Rho GTPase	Rho guanine nucleotide exchange factor
RIN	RNA Integrity Number
RIPA	radioimmunoprecipitation assay buffer
RNA-Seq	RNA sequencing
ROH	runs of homozygosity
RPM	revolutions per minute
RT-PCR	reverse transcription PCR
RTK	receptor tyrosine kinase
SDM	Site-Directed Mutagenesis
SDS	sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
SET/COMPASS	COMplex of Proteins ASsociated with Set)

SET1	enhancer of zeste, trithorax
SIFT	Sorting Tolerant From Intolerant
***SIX2*-Mut**	*SIX2* ORF containing *SIX2* candidate variant
***SIX2*-WT**	*SIX2* wild-type ORF
SNVs	single nucleotide variants
SOC	Catabolite repression
ss-DNA	single-stranded DNA
Tm	melting temperature
TBE	Tris-Borate-EDTA
TCGA	the Cancer Genome Atlas
TDH	thyroid dyshormonogenesis
TEMED	Tetramethylethylenediamine
TGD	thyroid gland dysgenesis
TSGs	tumour suppressor genes
TSH	thyroid stimulating hormone
UPSb	undifferentiated pleomorphic sarcoma of bone
UTRs	untranslated regions
VAF	variant allele frequency
VCF	variant call format
WB	western blot
WES	whole exome sequencing
WGS	whole genome sequencing

Chapter 1: Introduction

1.1 Thesis overview

The research presented in this thesis focuses on the application of next generation sequencing (NGS) technologies for studying rare inherited diseases and sporadic cancers. Whole exome sequencing (WES) was utilised to investigate the genetic profile of families with congenital hypothyroidism (CHT) (Chapter 3), an example of a rare genetic disease. WES and RNA sequencing (RNA-Seq) was used to characterise the genomic and transcriptomic profiles of three sporadic bone sarcomas, undifferentiated pleomorphic sarcoma of bone (UPSb) (Chapter 5), adamantinoma and osteofibrous dysplasia-like (OFD)-like adamantinoma (Chapter 6). Hence, this introduction is centred on three primary topics: inherited disorders, cancer and the application of NGS technologies in studying human genetic diseases.

1.2 Human disease: overview and Mendelian modes of inheritance

A genetic disease is defined as a disease caused by an abnormality in the genome of an individual. Among the human genome, protein-coding regions constitute ~1–1.5% (Biesecker and Green, 2014) of the human genome (20,000–25,000 genes) and are split across ~180,000 exons (Ng et al., 2009). Genetic diseases are generally grouped into three major categories: single gene (monogenic), chromosomal and multifactorial. Multifactorial diseases (e.g., congenital heart disease) can be caused by multiple defective genes and environmental factors, including air pollution, drug use and maternal exposure to harmful chemicals (Khan et al., 2015; Mahdieh and Rabbani,

2013). Monogenic diseases arise from defects in the DNA sequence of a single gene, producing an altered/faulty protein product. Monogenic diseases are classified based on Mendel's principles of hereditary into four primary groups: autosomal recessive (AR), autosomal dominant (AD), recessive or dominant X-linked diseases, and Y-linked diseases (Mahdieh and Rabbani, 2013). Mendel's experiments conducted on plants provided revolutionary insights into the patterns of inheritance, including AR and AD inheritance.

AR diseases are caused by mutations in both maternal and paternal alleles (i.e., loss of function). That is, heterozygous parents can transmit a defective allele to their offspring, in turn, making the affected offspring homozygous or compound heterozygous for the genetic change (Mahdieh and Rabbani, 2013). Heterozygous parents have a 25% chance of having an affected child affected with an AR disease. Cystic fibrosis, characterised by thickened mucus secretion in the lungs and other organs, is an example of an AR disease caused by homozygous or compound heterozygous mutations in the *CFTR* gene (Cutting, 2015).

In AD diseases, one mutated allele of a disease-causing gene is sufficient for disease manifestation. Affected parents have a 50% chance of transmitting the defective allele to their offspring. Achondroplasia, the most common inherited form of dwarfism, is an example of an AD disease and is caused by a heterozygous mutation in the *FGFR3* gene (Richette et al., 2008). Huntington's disorder, a progressive neurodegenerative disease caused by the defective expansion of CAG repeats in the *HTT* gene, also follows an AD inheritance pattern (Apolinario et al., 2017).

X-linked diseases are caused by mutations in genes located on the X chromosome. X-linked dominant and recessive diseases affect both females and males, while the recessive form is more commonly present in males, as males are obligatory

hemizygous for the X chromosome (Khan et al., 2015; Mahdieh and Rabbani, 2013). An example of X-linked recessive disorder is haemophilia B, which is caused by mutations in *F9* and is characterised by the inability of the body to form blood clots (Wang et al., 2016).

1.3 The complexity of human diseases: penetrance, variable expressivity, digenic and oligogenic inheritance

Although the established Mendel's inheritance principles are usually considered to be oversimplified, these principles still form the basis of inheritance patterns in human monogenic diseases. Since Mendel's time, the understanding of human disease has expanded, leading to the realisation that the genetics and inheritance patterns of human diseases are more complex than previously believed. There are other phenomena that can be associated with disease, including incomplete penetrance and variable expressivity, as well as digenic and oligogenic inheritance.

Disease penetrance is defined by the proportion of individuals carrying a genetic defect who show the defined disease phenotype. A disease shows complete penetrance when the clinical symptoms are present in 100% of affected individuals (Cooper et al., 2013). For instance, neurofibromatosis type 1 (NF1) is an AD multisystem disease with a penetrance approaching 100% by the age of 20 years; that is, virtually all people carrying a mutation in *NF1* will exhibit disease manifestation by the age of 20 years (Boyd et al., 2009). Conversely, if only a proportion of individuals harbouring a pathogenic genetic alteration exhibit clinical symptoms, the condition is said to have reduced/incomplete penetrance. Incomplete penetrance in generally observed in diseases following an AD inheritance pattern; however, it can also occur in AR diseases (Cooper et al., 2013). An example of incomplete penetrance is observed in

individuals carrying *BRCA1* and *BRCA2* mutations in which some carriers do not develop cancer during their lifetime, suggesting the involvement of both genetic and environmental factors (Gronwald et al., 2008).

Variable expressivity is defined as the variability of the clinical phenotype in individuals having the same genotype. Individuals with a specific genetic defect can show differences in disease severity and age of onset (Ahluwalia et al., 2009). Holoprosencephaly is an example of a genetic disease that shows variable expression in clinical manifestations, including different anomalies of the developing forebrain that can be accompanied with a broad spectrum of craniofacial abnormalities (Pineda-Alvarez et al., 2010).

In a typical Mendelian inherited disease, a mutation in a single gene gives rise to a clinical phenotype. By contrast, in digenic or oligogenic inheritance, the disease is caused by defects in more than one gene (Schaffer, 2013). One of the well-characterised oligogenic diseases is Bardet–Biedl syndrome, for which a minimum of 17 genes are associated with phenotypic manifestations (Cooper et al., 2013). In digenic-inherited diseases, a heterozygous mutation in each gene is generally required to manifest the full disease phenotype (component mutations). However, if one of the component mutations is absent, the other mutation can be non-penetrant or may cause a less severe form of the disease, a phenomenon that explains variable expressivity and penetrance in some diseases (Cooper et al., 2013). Moreover, it is sometimes unclear whether a given disease is caused by digenic inheritance or the coincidence of the two defective genes. In these cases, establishing a definitive genotype-phenotype correlation can be challenging. However, as the wealth of available genetic data increases, the understanding of digenic and oligogenic inheritance continues to improve (Gazzo et al., 2017).

1.3.1 Consanguinity and increased risk of genetic malformations

Consanguinity is a traditional practice that has been present since the early existence of modern humans. A consanguineous marriage refers to a marriage between couples who are related (e.g., first or second cousins) and share a common ancestor (Hamamy, 2012). The rates of consanguinity are variable across the world; nevertheless, higher consanguinity rates are seen in the communities of North Africa, the Middle East and West Asia, accounting for approximately 20–50% of all marriages (Hamamy, 2012). In addition, higher consanguinity rates are present in communities of North America and Europe that migrated from these regions (Teeuw et al., 2014).

Consanguineous marriages are associated with an increased risk of genetic defects in offspring due to the manifestation/expression of AR gene mutations inherited from a common ancestor (Hamamy, 2012). Conversely, when only one parent of a consanguineous marriage is affected, the risk for AD disorders is not increased (Hamamy, 2012). In theory, first cousins are expected to share 1/8 (12.5%) of their genes; hence, their offspring will be homozygous at 1/16 (6.25%) of gene loci (Hamamy, 2012). On average, a child of first-cousin marriage has a 1.7–2.8% additional risk of having an AR condition (Teeuw et al., 2014) (Figure 1–1). The increased expression of AR gene mutations in consanguineous families make these families favourable candidates for family-based genetic studies that aim to identify novel disease-causing genes. In Chapter 3, a study using WES was conducted in four consanguineous families with CHT.

In a consanguineous family with a suspected AR inheritance pattern, the disease-causing locus is assumed to occur in runs of homozygosity (ROH) (Figure 1–1). ROH are regions showing homozygous genotypes, which occur as a result of inheriting identical chromosome haplotypes from each parent (Ceballos et al., 2018). Single

nucleotide polymorphism microarrays technology can be used to identify these ROH in the entire genome, an approach known as autozygosity mapping (Carr et al., 2013; Sheridan et al., 2015). In consanguineous families, genes residing in ROH can be further prioritised to investigate the genetic mechanisms of diseases.

1.4 Rare diseases: definition, complexity and current understanding

By definition, a disease is classified as rare if it affects less than 1 in 2,000 individuals in Europe (Lopes and Oliveira, 2013) or fewer than 200,000 individuals in the United States (Whicher et al., 2018). Although rare diseases are individually uncommon, they collectively affect up to 10% of the worldwide population (Boycott et al., 2013; Pogue et al., 2018). The majority of rare diseases (~80%) are caused by genetic alterations, suggesting a strong phenotype-genotype involvement in these disorders (Lopes and Oliveira, 2013). There are approximately ~7000 rare genetic diseases (RGDs) (Boycott et al., 2013), also termed Mendelian or monogenic diseases, and 75% of these diseases affect children, impacting the wellbeing of these children and their families (Beaulieu et al., 2014). However, the exact number of RGDs is difficult to gauge due to phenotypes that have yet to be defined (Boycott et al., 2017). In the current thesis, CHT, an example of an RGD, was investigated using WES.

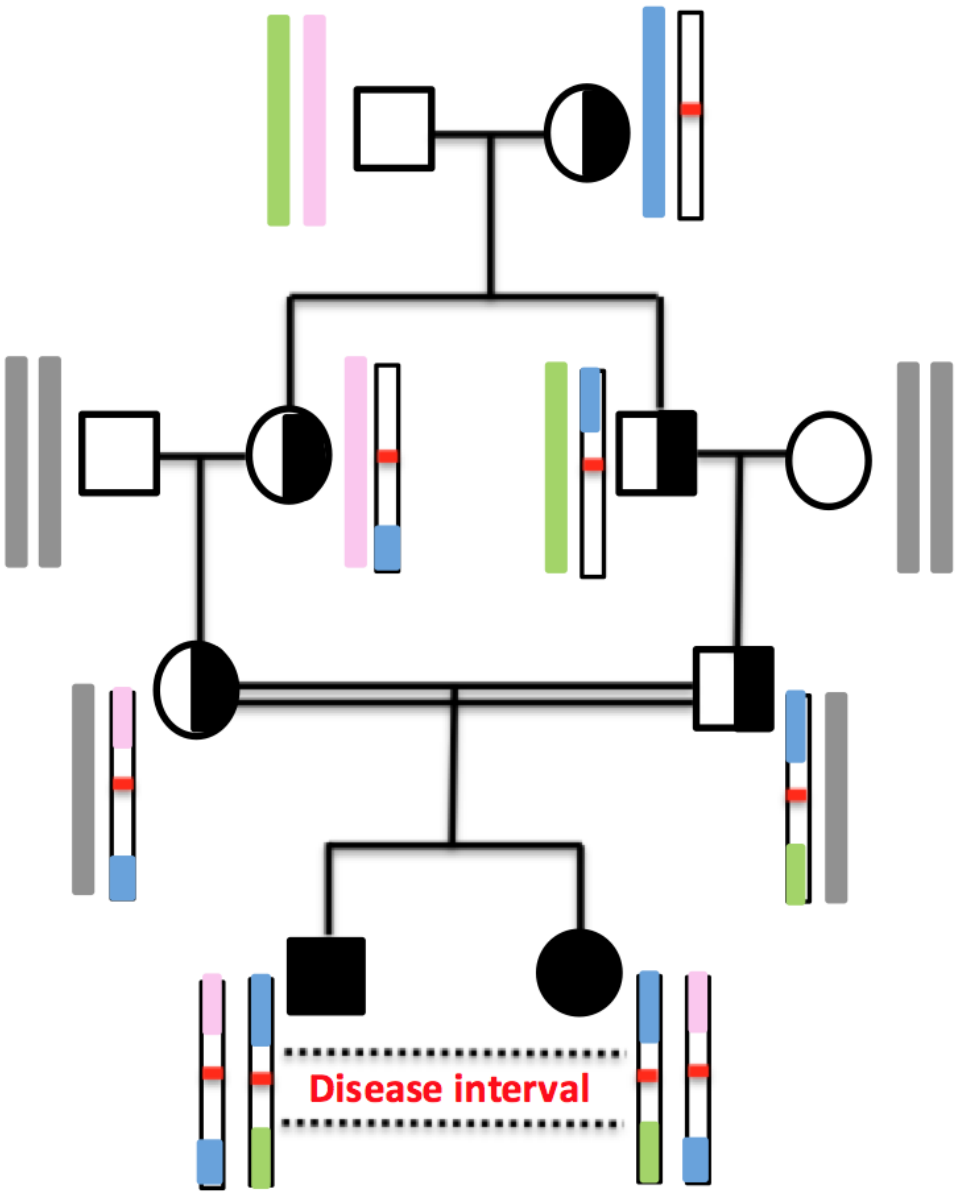


Figure 1–1: A diagrammatic representation showing the increased risk of autosomal recessive diseases in a consanguineous family. In consanguineous family with multiple affected individuals, the disease-causing genetic locus is inherited from a common ancestor and is likely to segregate in the family. The disease-causing interval is assumed to be residing in runs of homozygosity. Paternal and maternal chromosomes are in different colours, and the differently coloured blocks denote chromosomal crossing over between haplotypes. The red line represents a mutation. Original figure, compiled from information in (Becker et al., 2011; Carr et al., 2013).

Early diagnosis of patients with RGDs allows for optimal disease management, resulting in enhanced patient wellbeing and reduced long-term complications (Boycott et al., 2017). The diagnosis of RGDs can be challenging due to their rarity and clinical and genetic heterogeneity, leading to long and expensive diagnostic assessments (Boycott et al., 2013). Healthcare professionals are advised to acknowledge all available resources that can support the diagnosis, management and treatment of patients with RGDs (Pogue et al., 2018). Online Mendelian Inheritance in Man (OMIM) (https://www.omim.org) and Orphanet (https://www.orpha.net/consor/cgi-bin/index.php) are two internationally recognised databases that provide curated clinical information and genetic knowledge about RGDs. A rare disease classification and steps for evaluation are presented in Figure 1–2, which can serve as a starting point when evaluating RGDs.

RGDs are diverse in nature and are complicated by genetic and phenotypic heterogeneity (Boycott et al., 2013). For example, families affected with the same disease may have different alterations on the genetic level (genetic heterogeneity), a phenomenon that can be explained by incomplete penetrance and variable expressivity. This genetic heterogeneity can, in turn, lead to phenotypic variability, in which different clinical manifestations are observed in affected individuals harbouring the same genetic defect due to the possible involvement of other factors such as environmental factors (Boycott et al., 2013; Pogue et al., 2018).

The phenotypic and genetic diversity of RGDs is increasingly being acknowledged; nevertheless, a comprehensive knowledge about RGDs remains unestablished. Although significant progress has been made toward understanding the molecular aetiology of RGDs, approximately half of RGDs remain without a definitive molecular diagnosis (Boycott et al., 2017; Pabinger et al., 2014). Positive identification of a

disease-causing variant in a single gene establishes a definitive molecular diagnosis, which, in turn, allows for better disease management and provides better counselling options for family members, including prenatal diagnostic testing for future pregnancies (Bamshad et al., 2011; Pogue et al., 2018). The use of NGS technologies and multiple analytical frameworks has resulted in expanding our understanding of the genetic causes of RGDs.

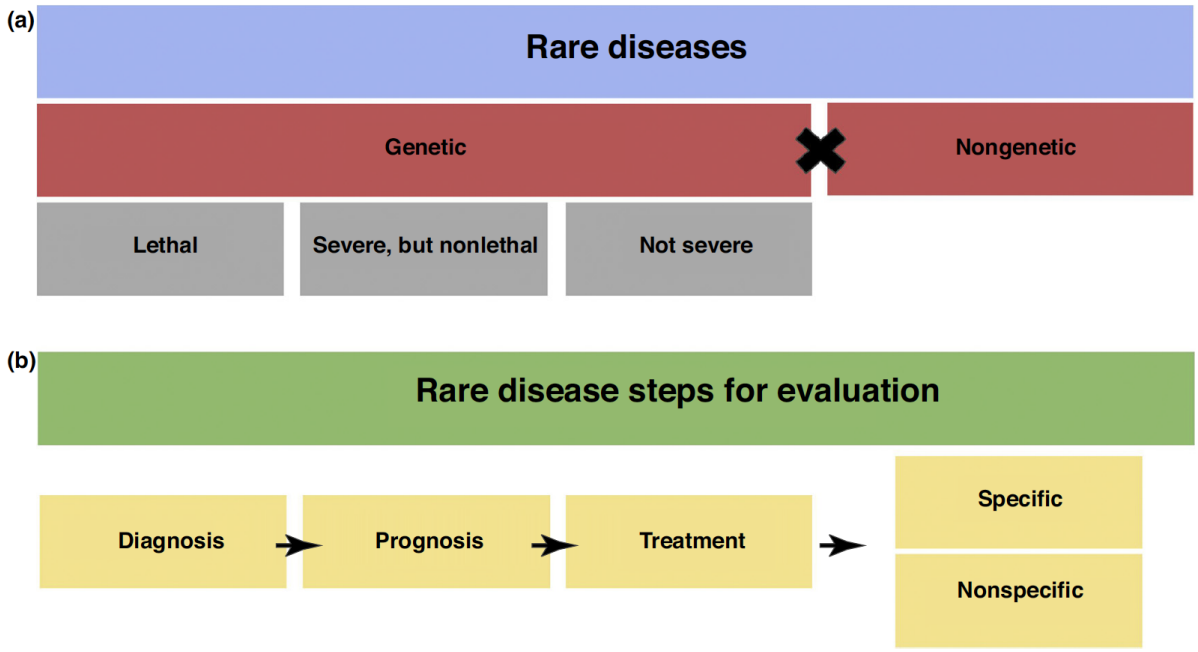


Figure 1–2: A recommended scheme for rare disease classifications and recommended steps for their evaluation. (A) Genetic and nongenetic factors can be associated with rare diseases. (B) Correct diagnosis and evaluation of rare diseases are important to provide better prognosis and treatment options, as well as offer appropriate counselling options for the entire family. Figure reproduced from (Pogue et al., 2018).

One of the main aims of diagnosis is to offer effective therapies that are specifically tailored for RGDs, when available. If specific treatments are not available, healthcare professionals can focus on managing disease symptoms (Pogue et al., 2018). In metabolic disorders, therapies that target pathophysiological features (e.g., enzyme

replacement therapeutics) can significantly improve patient quality of life (Tarailo-Graovac et al., 2016). For example, the use of asfotase alfa enzyme replacement therapy has shown clinical evidence for successful management of hypophosphatasia, an inborn error of metabolism disorder (Whyte et al., 2012). This drug is currently approved for use in Japan, Canada, and the USA (Whyte et al., 2012).

1.5 Cancer

Cancer is defined as a disease characterised by abnormal and uncontrolled growth of cells with the ability to spread and invade surrounding tissues. In normal cells, broad signalling and molecular networks are responsible for maintaining normal division, differentiation and apoptosis of cells. By contrast, in cancer cells, these signalling networks are impaired, resulting in abnormal growth and proliferation (Hejmadi, 2009). Histopathological classification of the different types cancers relies on identifying the cellular origin of the tumour (Idikio, 2011). For example, leukaemias and lymphomas originate from blood cell precursors. Carcinomas (the most common type of cancer) are derived from epithelial cells. Sarcomas, such as those of bone or soft tissue, are of mesenchymal origin (embryonic mesoderm) (Hui, 2016).

Differentiating between benign (usually harmless) and malignant tumours (potentially fatal) is crucial for providing accurate diagnosis, prognosis and treatment to the patient. Various well-established key features can differentiate between malignant and benign tumours, including rapid growth, increased cell turnover (proliferation and apoptosis), invasive growth and metastasis (Idikio, 2011). Moreover, microscopic examination of cells can determine their histopathological characteristics, and, in turn, whether a tumour sample is benign or malignant (Hejmadi, 2009).

1.5.1 Cancer incidence and mortality

Cancer in the wider perspective includes more than 277 different cancer subtypes and is considered the second leading cause of death worldwide (Hassanpour and Dehghani, 2017). Understanding cancer incidence and mortality rates is useful in anticipating the burden of this disease on health services (Smittenaar et al., 2016). In 2012, an estimated 14.1 million new cancer cases and 8.2 million cancer deaths occurred worldwide (Ferlay et al., 2015). The three most commonly diagnosed cancers worldwide are lung (1.82 million), breast (1.67 million), and colorectal (1.36 million) (Ferlay et al., 2015). Based on 50 registries selected to represent various regions of the world, males have a higher cancer incidence rate (400 cases per 100,000 males) than females (300 per 100,000 females) (Torre et al., 2016).

Due to the continuous population growth and increasing adoption of lifestyles that can contribute to increased risks, cancer incidences and deaths are expected to increase (Torre et al., 2016). These increases are more likely to affect low- and middle-income countries as they undergo economic transitions, adopting many of the lifestyle risk factors for cancer that are already present in high-income countries, including tobacco use, reduced physical activity and higher body weights (Torre et al., 2016).

A study by Smittenaar et al. (2016) aimed to determine the projections of cancer incidence and mortality rates in the UK up to 2035 based on the recorded cancer data between 1979–2014. Between 2015–2035, the authors projected a minor increase in cancer incidence rate for females (0.11%), compared with a very slight decrease for males (0.03%); moreover, mortality rates for both males and females were projected to decline. Figure 1–3 and Figure 1–4 illustrate the proportions of different cancer cases and deaths, respectively, in 1993, 2014 and 2035 (projected). Until 2035, the authors projected an increase in the overall number of cancer cases and deaths,

explained by the increasing size and ageing of the population. The fastest accelerating cancer subtypes include thyroid, liver, oral and kidney cancers. However, over the same period, the overall mortality rates are projected to decline for both males and females. The use of vaccines can reduce the incidence of some cancer subtypes, and the advancement of therapeutic strategies may reduce cancer mortality (Smittenaar et al., 2016). Conducting cancer projection studies at regular intervals is important to capture changes in cancer data.

1.5.2 Cancer aetiology: a genetic disease

Cancer is a complex disease that can be caused by factors within the cell (somatic or inherited mutations) and/or by external environmental factors (tobacco use, chemicals, and radiation). The combination of these factors can result in abnormal cell behaviour and proliferation, leading to cancer development (Hejmadi, 2009). Studies in the early 20th century identified multiple chromosomal aberrations in the cancer cell, providing early insights into the association of genomic changes with cancer development (Stratton et al., 2009). Afterwards, numerous cancer genomic studies identified genomic abnormalities in multiple cancer subtypes, including the Philadelphia translocation (see Section 1.6.1).

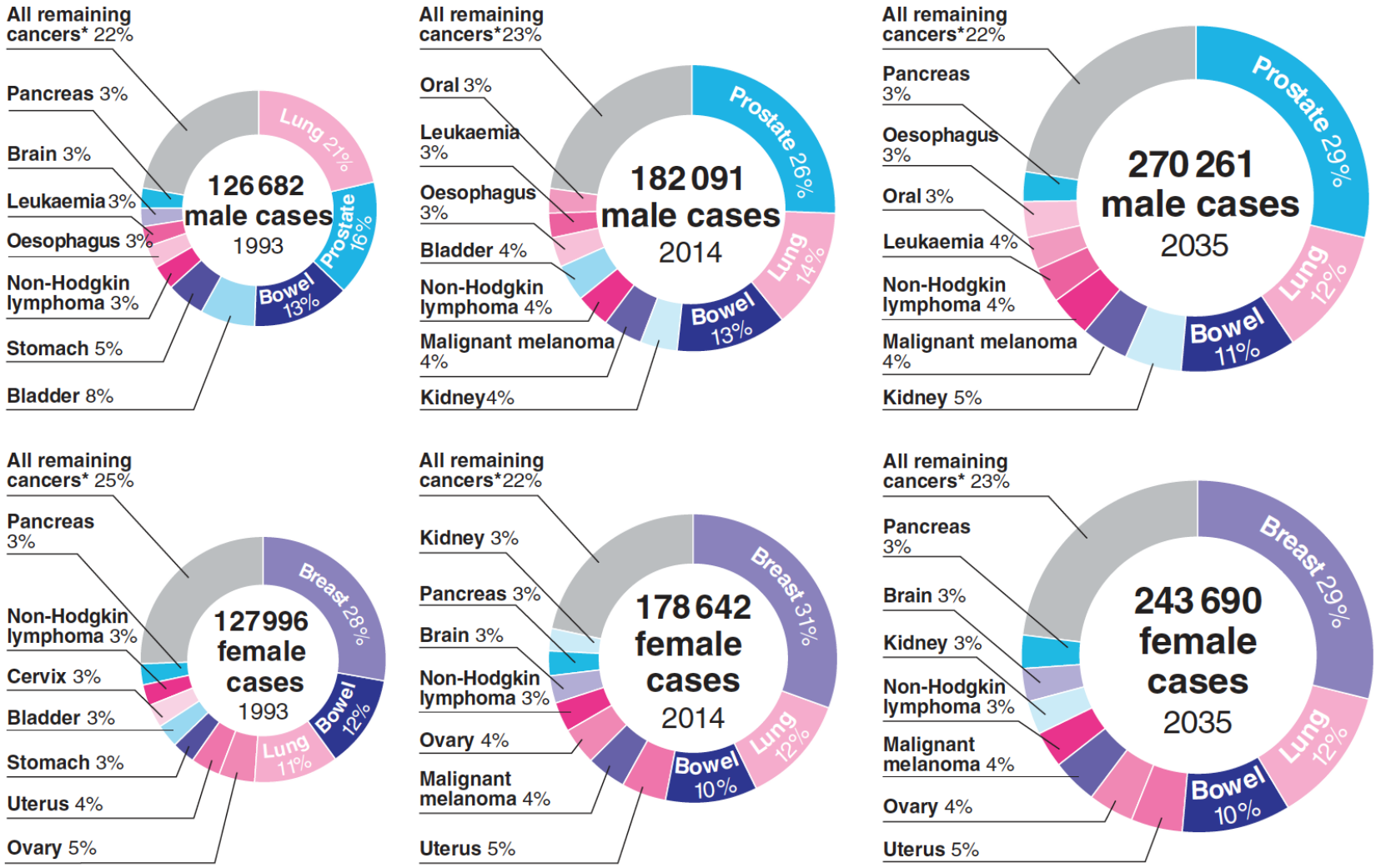


Figure 1–3: The proportion of different cancer cases, classified by cancer anatomical site in 1993 (observed; left), 2014 (observed; middle) and 2035 (projected; right), and split by males (top) and females (bottom). The size of each doughnut is representative of the total number of cases. Image reproduced from (Smittenaar et al., 2016).

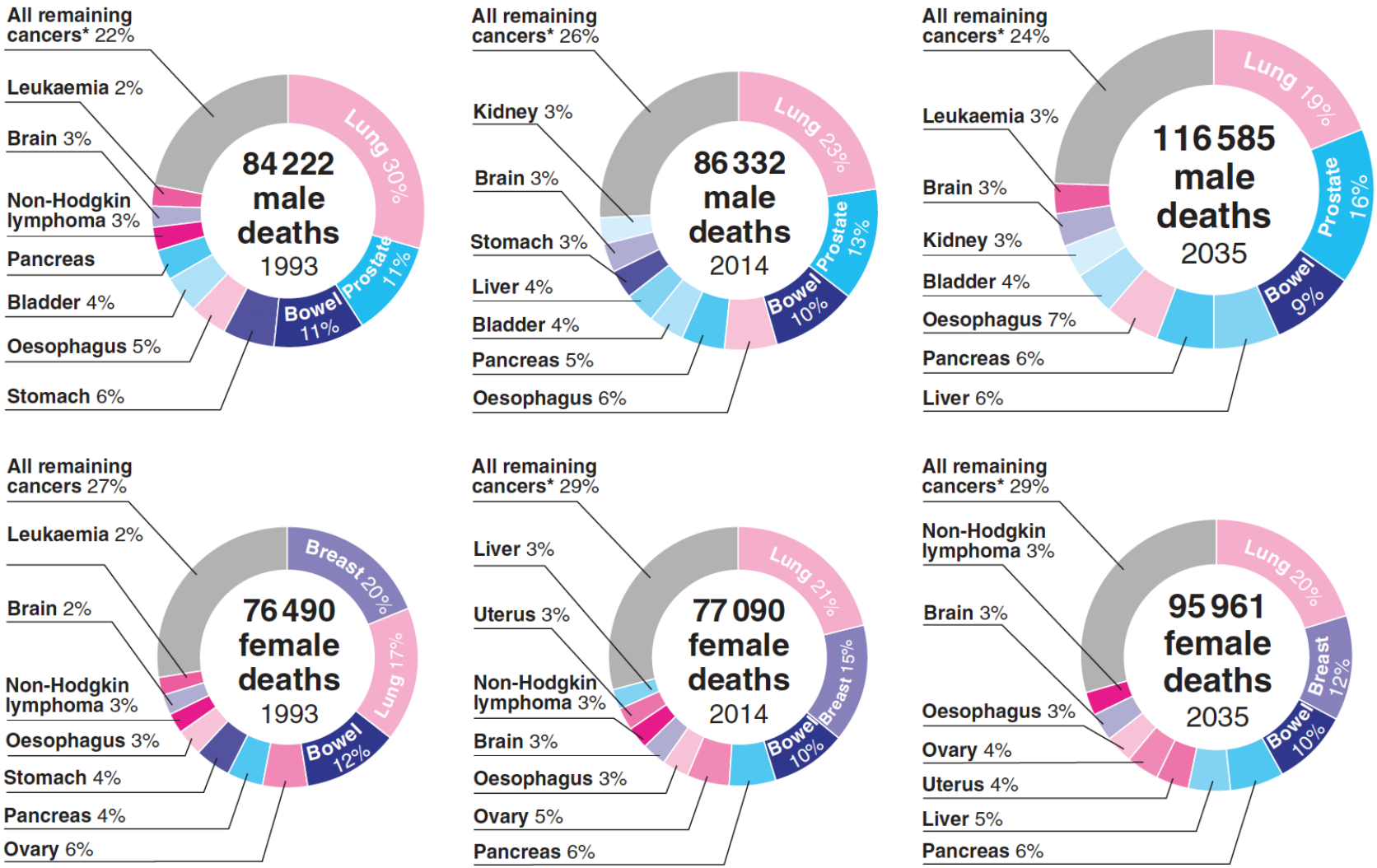


Figure 1–4: The proportion of cancer deaths, classified by cancer anatomical site in 1993 (observed; left), 2014 (observed; middle) and 2035 (projected; right), and split by males (top) and females (bottom). The size of each doughnut is representative of the total number of died cases. Image reproduced from (Smittenaar et al., 2016).

DNA mutations can affect the normal behaviour and integrity of a cell, leading to an uncontrolled proliferative outcome. However, the behaviour of an individual cell is not autonomous and is normally controlled by complex networks, which involve the surrounding cells and microenvironment (Hejmadi, 2009). The complexity and the diverse subtypes of cancer suggest the involvement of complex and distinct biological capabilities (known as the hallmarks of cancer) that can be responsible for cancer development. Six hallmarks of cancer can be partially or fully present in almost all cancers: sustaining proliferative signalling (immortality), evading growth suppressors (negative regulation of cell morality), resisting cell death (impaired programmed apoptosis), enabling replicative immortality (oncogenic growth signalling), angiogenesis (formation from new blood vessels) and metastasis (cancer spreading to other sites) (Hanahan and Weinberg, 2011).

Somatic mutations in a cancer genome occur from mistakes in DNA replication or from DNA damage (e.g., UV light, radiation) that are repaired erroneously or left unrepaired. The number of mutations in a cancer genome differs widely among cancer subtypes. However, most cancers can harbour 1,000–20,000 somatic changes, most of which are single nucleotide variants (SNVs) (Martincorena and Campbell, 2015). In addition, cancer genome can harbour a few to hundreds of insertions/deletions (INDELs) and genomic rearrangements (Martincorena and Campbell, 2015). The cancer genome can also acquire epigenetic changes, which alter its chromatic structure (packing of the DNA). Gains or losses in copy number variations (CNVs), which change the two copies of genetic loci in the normal diploid genome, can also be present in cancer cells (Stratton et al., 2009). This increasing number of acquired genomic alterations in cancer is explained by genomic instability, a hallmark associated with cancer in which cells acquire mutations as they grow and proliferate (Burrell et al., 2013).

1.5.2.1 Cancer driver and passenger mutations

A benign lesion can transform into a malignant one by acquiring a series of mutations over time (Vogelstein et al., 2013). Many of these mutations are neutral 'passenger' alterations, accumulating throughout the carcinogenesis process, and only a handful of changes will be responsible for 'driving' tumour development (Sakoparnig et al., 2015; Tamborero et al., 2013). Thus, these somatic mutations, regardless of their structural nature, are not equal in biological effect and can be classified based on their consequences in tumour growth and progression (Stratton et al., 2009; Vogelstein et al., 2013). For example, passenger mutations are changes that have an indifferent effect on the tumourigenesis process. By contrast, a driver mutation is causally implicated in tumourigenesis as it confers a selective growth advantage and promotes tumour progression (Stratton et al., 2009). Based on age-incidence statistics, it has been suggested that tumourigenesis requires ~7 driver mutations in epithelial cancer (e.g., breast, colon), fewer than those in paediatric tumours (Tamborero et al., 2013). Hence, the identification of genuine driver genes is critical in oncology research and in the development of personalised therapeutics.

Although extensive research has been conducted on cancers, accurate identification of true cancer driver genes can be a challenging task. One of the primary ways to classify driver genes is to look for somatic mutations in genes that are non-randomly altered in more than one cancer sample—recurrent mutated genes (Dees et al., 2012; Hofree et al., 2016). Cancer driver genes can be distinguished from passenger alterations by their increased degree of recurrence (mutation frequencies) among tumours (Hofree et al., 2016; Sakoparnig et al., 2015). However, it has been reported that cancer driver genes can also be present at lower frequencies (Sakoparnig et al., 2015). Another way to classify 'true' cancer driver genes relies on the enrichment of

recurrent candidate genes in cancer annotation databases (Sakoparnig et al., 2015). The Catalogue of Somatic Mutations in Cancer (COSMIC) Cancer Gene Census (CCGC) provides a comprehensive list of curated cancer-associated genes (http://cancer.sanger.ac.uk/cosmic/curation). In addition, the Integrative Onco Genomics (inTOgen) mutation database classifies cancer driver genes based on the analysis of 4,623 exomes from 13 cancer sites (https://www.intogen.org/).

1.5.2.2 Oncogenes

Proto-oncogenes are genes that are involved in promoting normal cellular growth, proliferation or apoptosis (Giordano and Lee, 2006). When mutated, these genes become activated (subsequently called oncogenes) and acquire a pro-tumourigenic effect. Oncogenes are constitutively activated, causing selective and uncontrolled growth advantage of the cell and, thereby, leading to cancer (Giordano and Lee, 2006). Generally, oncogenes encode signalling receptors and signalling transducing transcription factor proteins that have fundamental roles in cell homeostasis regulatory networks (Liu et al., 2017).

Genetic alterations that activate oncogenes are generally gain-of-function, including single base substitutions or structural alterations such as gene amplification or chromosomal rearrangements (inversion or translocations) (Croce, 2008). Base pair mutations in oncogenes can change the structure and conformation of oncogenes, thus enhancing their transforming activity. For instance, activating mutations in the *RAS* oncogene family (*KRAS, HRAS, RNAS*) lead to the production of altered proteins that remain in the active state, henceforth continuously transducing oncogenic signals (Croce, 2008). An example of oncogenic chromosomal rearrangements includes the *BCR-ABL* gene fusion identified in leukaemias (more in 1.6.1) (Vogelstein et al., 2013).

Generally, activating mutations in oncogenes are recurrent and occur at the same amino acid positions of the protein; by contrast, mutations in tumour suppressor genes (TSGs) are non-recurrent, protein-truncating and occur throughout the genes' length (Figure 1–5) (Vogelstein et al., 2013).

The important role of oncogenes in cancer pathogenesis sparked researchers to explore these genes, especially as targets for precision cancer therapies. Small molecules (e.g., inhibitors) or monoclonal antibodies can be used to target oncogenic proteins in cancer cells (Croce, 2008). For example, antibodies and kinase inhibitors targeting the *MET* oncogene are currently in early or advanced stages of clinical trials (Comoglio et al., 2018). *MET* encodes a receptor tyrosine kinase that, when genetically altered, can initiate and sustain neoplastic transformation. A recent study by Comoglio et al. (2018) suggested that patients with tumours (both with primary and metastatic) harbouring amplification or activating mutations in *MET* are likely to exhibit tumour regression when using MET inhibitors.

1.5.2.3 Tumour suppressor genes and the Knudson two-hit hypothesis

TSGs encode proteins that play key roles in cell division, cell proliferation, and initiating cell apoptosis or DNA damage pathways, thus maintaining the integrity and genomic stability of cells (Zhao et al., 2016). Inactivation of a TSG can interfere with its cellular protective role, resulting in uncontrolled cell growth and proliferation (Hanahan and Weinberg, 2011). In animal functional studies, the deletion of TSGs in mice usually results in spontaneous developments of tumours or the progression of existing tumours to more advanced stages (Acosta et al., 2018).

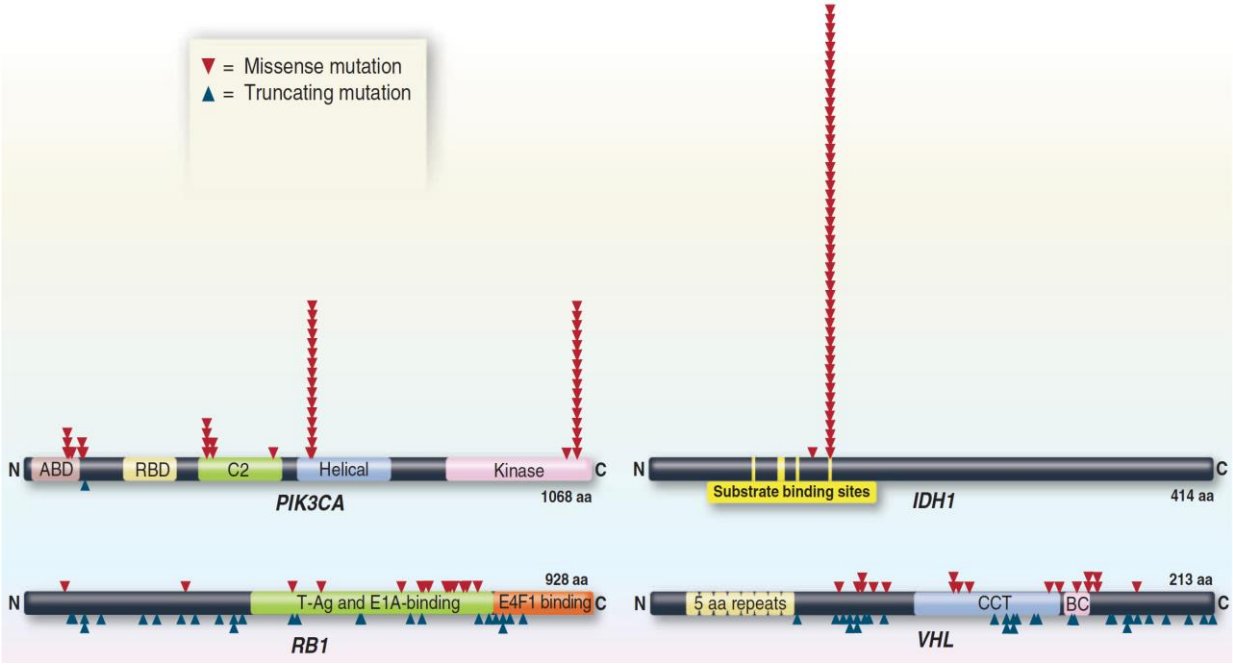


Figure 1–5: An example of the mutation distributions in oncogenes (*PIK3CA, IDH1*) and tumour suppressor genes (*RB1, VHL*). *PIK3CA* and *IDH1* are examples of oncogenes; *RB1* and *VHL* are tumour suppressor genes. The distribution of missense mutations (blue arrowheads) and truncating mutations (red arrowheads) are shown, as reported in COSMIC (release version 61) in these oncogenes and tumour suppressor genes. Figure reproduced from (Vogelstein et al., 2013).

Unlike oncogenes, TSGs are usually inactivated by two hits/mutations (e.g., homozygous deletions), producing a loss-of-function protein (Vogelstein et al., 2013). This phenomenon, termed the Knudson two-hit hypothesis, was initially described in a study by Knudson (1971) in familial and sporadic retinoblastoma (cancer of the retina) cases, in which two mutational events were required for disease manifestation. The Knudson two-hit hypothesis states that (1) in familial retinoblastoma cases (dominant inheritance), the first mutation is inherited in germline cells, followed by a second mutation acquired later in life in the somatic cells; and (2) in sporadic (nonhereditary) cases, both mutations are acquired in somatic cells (Figure 1–6) (Knudson, 1971; Knudson, 2001). Consistent with this hypothesis, subsequent studies reported that

familial retinoblastoma cases are caused by two mutations in the well-studied retinoblastoma (*RB1*) TSG, in which the majority of the cases harboured germline *RB1* mutations (first hit), and their tumours exhibited loss of heterozygosity (second hit) (Cavenee et al., 1983; Fung et al., 1987). The Knudson hypothesis was also evident in sporadic tumours, which harboured two hits, in the absence of germline *RB1* alterations (Burkhart and Sage, 2008; Knudson, 2001). This hypothesis explains the development of retinoblastoma (often bilaterally, affecting both eyes) at a younger age in children carrying an *RB1* germline mutation than the sporadic form, which usually occurs unilaterally and at an older age as somatic mutations take longer time to acquire (Knudson, 1971; Knudson, 2001). These findings enhanced the field of cancer genetics and led to the current understanding of the role of TSGs in cancer development.

1.5.3 Primary tumours of bone

Primary bone tumours comprise a heterogeneous group of tumours that can be benign or malignant. In the UK, the incidence of primary bone tumours is approximately six cases for every one million in the population (Plant and Cannon, 2016). Benign bone tumours include osteoma, osteoid osteoma, benign osteoblastoma and giant cell tumour of bone. Although benign bone tumours are non-cancerous and non-metastatic, these tumours can grow in size and affect healthy bone tissue (Hakim et al., 2015). By contrast, osteosarcoma, Ewing sarcoma and chondrosarcoma are examples of malignant bone tumours (bone sarcomas). By definition, bone sarcomas comprise a wide variety of primary, non-epithelial malignant lesions, arising from bone cells (or their precursors) with a potential to metastasise (Mavrogenis and Ruggieri, 2015). In the current thesis, three bone tumours were investigated using NGS: UPSb

(malignant high-grade), adamantinoma (malignant low-grade) and OFD-like adamantinoma (usually benign).

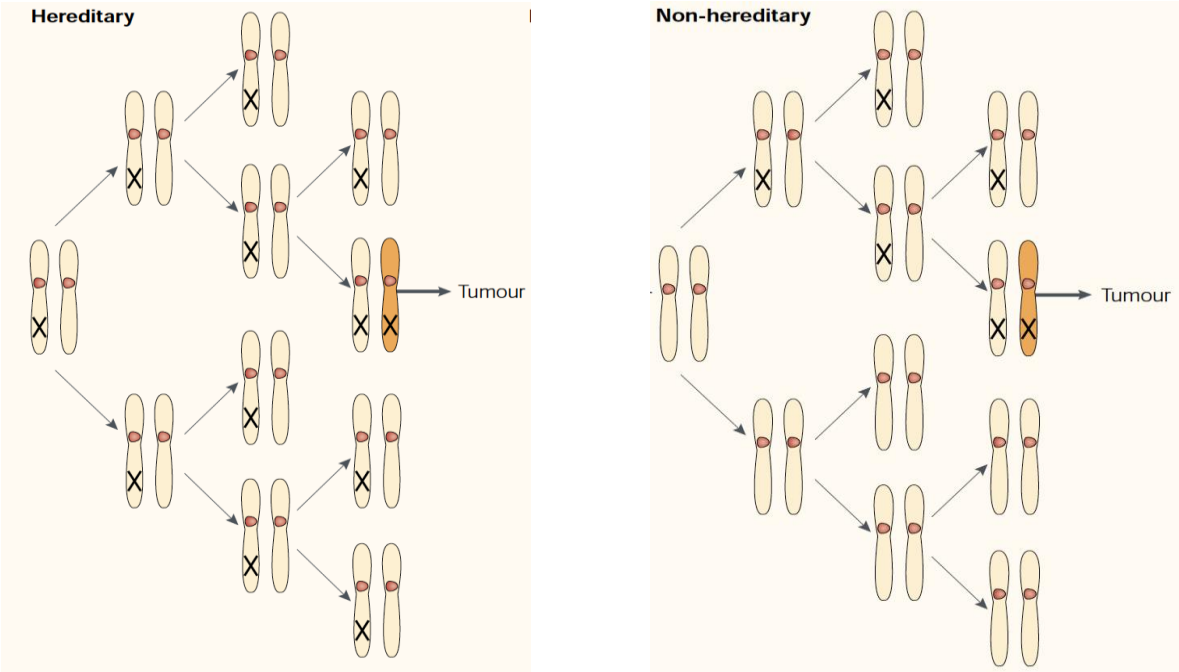


Figure 1–6: The Knudson two-hit hypothesis in both hereditary (familial) and non-hereditary (sporadic) retinoblastoma cases. In the hereditary form, the first hit (denoted by an X) is inherited and present in all the cells; the second hit is acquired, leading to tumour development. In the non-hereditary form, the two hits are acquired somatically, causing tumour development. Figure reproduced from (Knudson, 2001).

1.5.3.1 Incidence and survival of bone sarcomas

Primary bone sarcomas account for approximately 0.2% of all human neoplasms (Evola et al., 2017). Despite their rarity, bone sarcomas account for approximately 5% of paediatric cancers in Europe (Gerrand et al., 2016). Osteosarcoma and Ewing sarcoma are the most common paediatric primary malignant bone tumours, where the former accounts for approximately 50% of bone cancers in children (Interiano et al.,

2016). Chondrosarcoma is also a common primary bone sarcoma subtype that is usually diagnosed in older patients during their third to sixth decade of life (Mavrogenis and Ruggieri, 2015).

Between 1996 and 2010, a total of 7,229 patients were diagnosed with a primary bone sarcoma (Matthew et al., 2013). Generally, bone sarcomas affect males more than females. In 2011, 559 bone sarcoma cases were diagnosed in the UK; 58% of the cases were males and 42% were females (Beckingsale and Shaw, 2017). The majority of bone sarcomas affect children or young patients and occur in the extremities, most commonly in the lower limbs (38%), pelvis (16%) or upper limbs (14%) (Beckingsale and Shaw, 2017). The incidences by age for the most common sarcomas and the most affected anatomical sites are provided in Figure 1–7 and Figure 1–8, respectively.

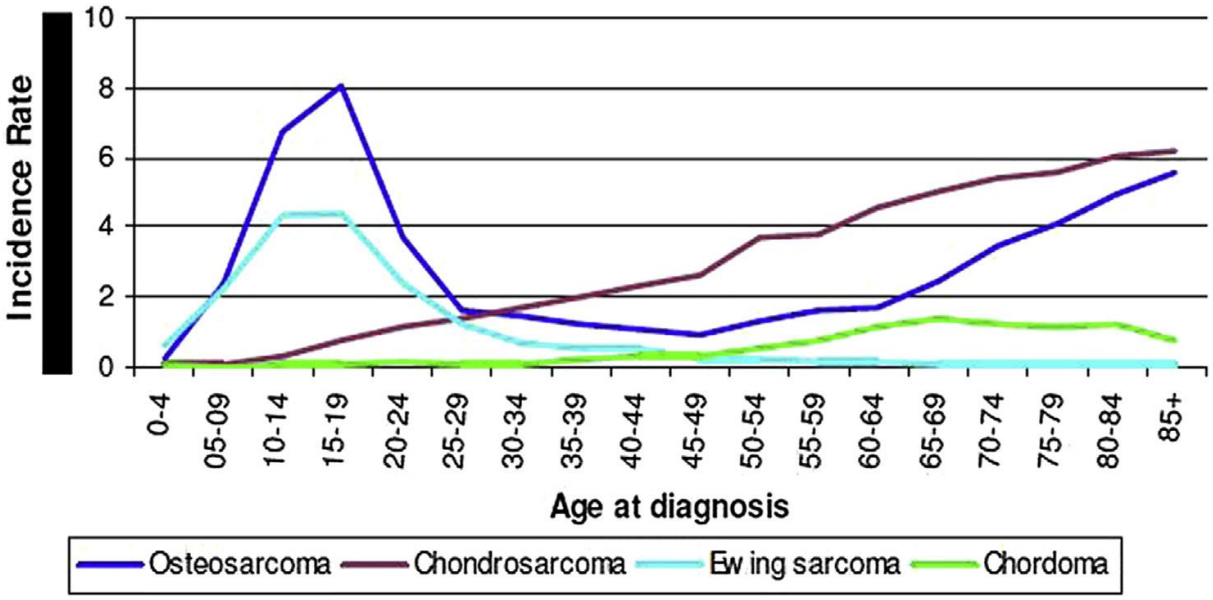


Figure 1–7: The incidence by age for the osteosarcoma, chondrosarcoma, Ewing sarcoma and chordoma in England between 1985–2009. Age-specific incidence of these common tumours are shown based on data collected from Public Health England. Figure reproduced from (Beckingsale and Shaw, 2017).

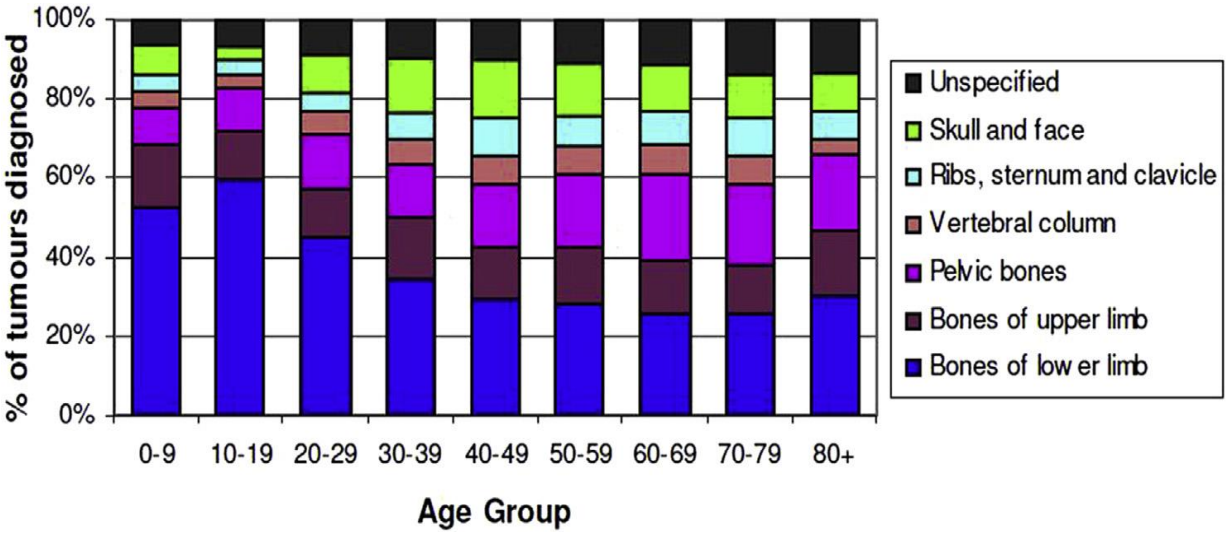


Figure 1–8: Proportion of bone sarcoma cases diagnosed by age group and tumour site. In young patients (age of 20 and below), 70% of bone sarcomas occur in the extremities. By contrast, in patients over the age of 40, 40% of tumours occur in the limbs. Data collected from Public Health England. Figure reproduced from (Beckingsale and Shaw, 2017).

1.5.3.2 Clinical presentation, diagnosis and treatment of bone sarcomas

Generally, the first symptom of bone sarcoma in patients is non-specific pain around a joint (typically at night), often accompanied by an unexplained lump. The pain usually increases with time and is present for more than a few weeks (Mavrogenis and Ruggieri, 2015).

Accurate diagnosis of bone sarcomas is essential to guide disease management and patient outcome. The diagnosis of the cases relies on clinical presentations and imaging findings. A detailed medical history and physical examination are required to provide the likely diagnosis of a suspected bone neoplasm (Mavrogenis and Ruggieri, 2015). A conventional x-ray test is simple, non-invasive, and generally the first diagnostic test of choice. However, a diagnosis of bone sarcoma cannot be ruled out based on ‘normal’ x-ray findings (Gerrand et al., 2016). If the diagnosis of malignancy cannot be excluded, the next imaging step is to perform a magnetic resonance imaging

(MRI) of the suspected area with the adjacent joints. MRI is considered the primary imaging test for detailed evaluation of bone neoplasms. In case of diagnostic doubts, computed tomography (CT) is requested to visualise and provide details about the suspected bone lesion (Mavrogenis and Ruggieri, 2015). A biopsy test is the gold standard diagnostic test for assessing the tumour histomorphology of bone sarcomas and confirming the provisional diagnosis (Plant and Cannon, 2016; Mavrogenis and Ruggieri, 2015). Moreover, immunohistochemical and molecular analyses (of specific molecular biomarkers) can be performed on the tumour specimen to provide a definitive molecular diagnosis (Gerrand et al., 2016).

The treatment of bone sarcomas is complex and usually requires a multidisciplinary input from various health professionals. No one specific therapeutic approach exists that can be followed across all bone sarcoma subtypes (Kreahling et al., 2013). However, surgical intervention (resection or amputation) is usually the primary therapeutic choice. The use of chemotherapy or radiation therapy is tailored according to the type of bone neoplasm. For example, in osteosarcoma and Ewing sarcoma, a chemotherapy course is completed prior to surgical intervention, which aims to target the potential micrometastatic disease and allows for easier resection (Mavrogenis and Ruggieri, 2015).

1.5.3.3 Overview of the genetics of bone sarcomas

The majority of primary bone sarcomas arise sporadically in the absence of underlying genetic predisposition. However, approximately 10% of osteosarcoma cases are caused by predisposing (inherited) genetic causes. For example, individuals with inherited *TP53* (associated with Li-Fraumeni syndrome) and *RB1* mutations are at an

increased risk of developing osteosarcoma or other malignancies (Siller and Lewis, 2010).

A comprehensive understanding of the genetic causes in sporadic bone sarcoma cases has not been achieved. However, the availability and the advancement of molecular and sequencing techniques have led to the identification of tumour-specific mutations, translocations and differentially altered genes, expanding the understanding of the genetic landscape of bone sarcomas. For instance, Ewing sarcoma tumours are characterised by somatic gene fusions involving *EWSR1* and *ETS* families (Pierron et al., 2012). Although the somatic genetic landscape of osteosarcoma is complex, somatic mutations in *TP53* (31-82% mutation rate) and *RB1* (19-64% mutation rate) cancer driver genes are commonly altered in osteosarcomas (Gianferante et al., 2017). With the continuous application of NGS technologies, the list of driver genes and mutational events in cancer is continuously increasing (more later).

1.6 Identification of the genetic basis of disease: disease-causing gene discovery

One of the primary aims of the medical genetics field is elucidating disease-causing genes underlying human disease and determining the associated phenotype of genetic alterations. There are various forms of genetic alterations that can be responsible for human disease, SNVs (missense, nonsense), small INDELs, CNVs (losses/gains), and interstitial deletions/amplifications (Mahdieh and Rabbani, 2013; Sathirapongsasuti et al., 2011).

Between the mid-1980s and mid-2000s, the primary method of gene discovery relied on a hypothesis-driven approach, consisting of a combination of linkage analysis,

positional cloning and Sanger sequencing of selected candidate genes (Boycott et al., 2017; Gilissen et al., 2012). Between 2001 and 2006, DNA microarrays provided the first genome-scale analysis method of DNA and RNA (Levy and Myers, 2016). The emergence of NGS technologies in 2009, primarily WES, has changed the landscape of RGDs and cancer research (Boycott et al., 2017). Conventional approaches and NGS technologies will be explained in detail in the subsequent sections.

1.6.1 Classical cytogenetic, fluorescent *in situ* hybridisation and array CGH

Classical cytogenetics is a conventional diagnostic tool that enables the detection of genomic aberrations such as genomic rearrangements within (e.g., inversions) or between chromosomes (e.g., translocations) and genomic segments harbouring gains and losses. This conventional method played a major role in identifying chromosomal aberrations related to inherited human disease and cancer (Riegel, 2014).

Chromosomal aberrations are often observed in leukaemias and lymphomas. The famous Philadelphia chromosome was the first chromosomal aberration discovered in cancer, specifically in chronic myeloid leukaemia, which results from a translocation between chromosomes 9 and 22, leading to the formation of *BCR-ABL1* oncogenic gene fusion (Wan, 2014). Conventional cytogenetic techniques can be used as the first diagnostic tool for newly diagnosed leukaemias, due to its simplicity and usefulness in establishing diagnosis (Wan, 2014). Cytogenetic studies are also used in suspected chromosome abnormalities (chromosomal trisomies), multiple congenital anomalies and multiple miscarriages (Mahdieh and Rabbani, 2013). Nevertheless, chromosomes banding techniques (e.g., karyotyping) have a limited resolution of detection genomic aberrations; that is, these techniques can only detect structural aberrations of at least 5–10 megabases in size (Riegel, 2014).

To overcome the limited resolution issue with chromosomal banding techniques, several molecular cytogenic methods, including fluorescent *in situ* hybridisation (FISH), were developed, upgrading the classical cytogenetics to molecular cytogenetics in clinical settings (Mahdieh and Rabbani, 2013). FISH uses fluorescently labelled probes that complementarily hybridise to specific genomic regions and are visualised using a fluorescent microscope. A wide range of FISH probes, ranging from whole-chromosome painting arms to subtelomeric and locus-specific, are available for detecting constitutional and acquired/somatic chromosomal aberrations (Riegel, 2014). FISH has a considerably higher resolution (100–200 kilobases) than conventional karyotyping, allowing for detection of genomic aberrations at submicroscopic level (Cui et al., 2016). FISH is used in diagnostic and research settings to delineate inversions, insertions, microdeletions as well as in defining chromosome breakpoints (such as in translocations) (Cui et al., 2016; Wan, 2014). The disadvantage of FISH is the inability to detect chromosomal aberrations smaller than 100 kilobases, as well as the false-negative results in cases with small microdeletions (Bishop, 2010; Cui et al., 2016).

Other molecular technologies have been developed to identify chromosomal structural aberrations. An example of this technology is array comparative genomic hybridisation (aCGH), which allows for a genome-wide screening for chromosomal CNVs (deletions, duplications) by comparing a patient's DNA sample to a control sample (Ferguson-Smith, 2015). An advantage of aCGH is that extracted DNA material is the input material, which does not require tissue culturing as in karyotyping and FISH (Riegel, 2014). Low or higher resolution arrays are available, ranging from 3000 to over one million markers that can be spaced at, for example, one megabase intervals (Ferguson-Smith, 2015). Multiple arrays are now available and are still being used in

clinics as a part of the genetic diagnosis of diseases such as developmental delays, intellectual disability and multiple congenital anomalies (Cui et al., 2016). A common limitation of array-based technologies is the incapability of detecting balanced translocations and inversions (Riegel, 2014).

1.6.2 Sanger sequencing of candidate genes and positional cloning approach

Positional cloning is one of the traditional approaches used to define the molecular origin of genetic diseases. The positional cloning approach consists of several steps to define causative genes, including the identification of a candidate gene based on chromosomal location, followed by mutational analysis using Sanger sequencing (Figure 1–9) (Puliti et al., 2007). From its introduction in 1977 until the invention of NGS, Sanger sequencing was the gold standard method in mutational analysis studies. Also known as chain termination sequencing, Sanger sequencing utilises dye-labelled deoxynucleotides (dNTPs) and dideoxy-modified deoxynucleotides (ddNTPs) to sequence previously amplified DNA fragments of 500–1000 basepairs in length (Figure 1–10) (Kircher and Kelso, 2010).

The first step in the positional cloning approach begins with linkage analysis, in which families with the same phenotype are analysed using sets of DNA polymorphisms to define genomic regions that segregate with the disease (Botstein and Risch, 2003). Candidates genes can be nominated if (1) they share similar functionality of genes associated with similar phenotypes; (2) the functionality of their produced protein products is related to or can explain the pathogenicity of the disease; or (3) a positional cloning approach has located these candidates in a genomic region (Gilissen et al., 2012). A positional mapping approach does not rely on prior biological knowledge of a disease beyond an accurate assessment of the phenotype, leading to disease gene

identification in an unbiased manner (Botstein and Risch, 2003; Gilissen et al., 2012). Nevertheless, in positional cloning, it is almost impossible to determine if the phenotype is caused by a SNV or by a larger genomic CNV. Other possible difficulties in positional cloning arise when the identified critical disease locus remains large, resulting in a large number of candidate genes that can complicate Sanger sequencing mutational analyses (Gilissen et al., 2012).

All the previously mentioned traditional cytogenetics and molecular techniques have led to the discovery of many genes associated with inherited diseases and cancer. Although the traditional positional mapping approach was successful in identifying many disease-causing genes, this approach has not been completely successful in studying RGDs, especially in very rare cases when performing genetic association studies is difficult or when the phenotype is sporadic and a family-based study is not feasible (Boycott et al., 2013; Gilissen et al., 2012). Therefore, more comprehensive disease-causing analyses are required to investigate rare phenotypes, especially those that are associated with locus heterogeneity or reduced reproductive fitness in the affected members.

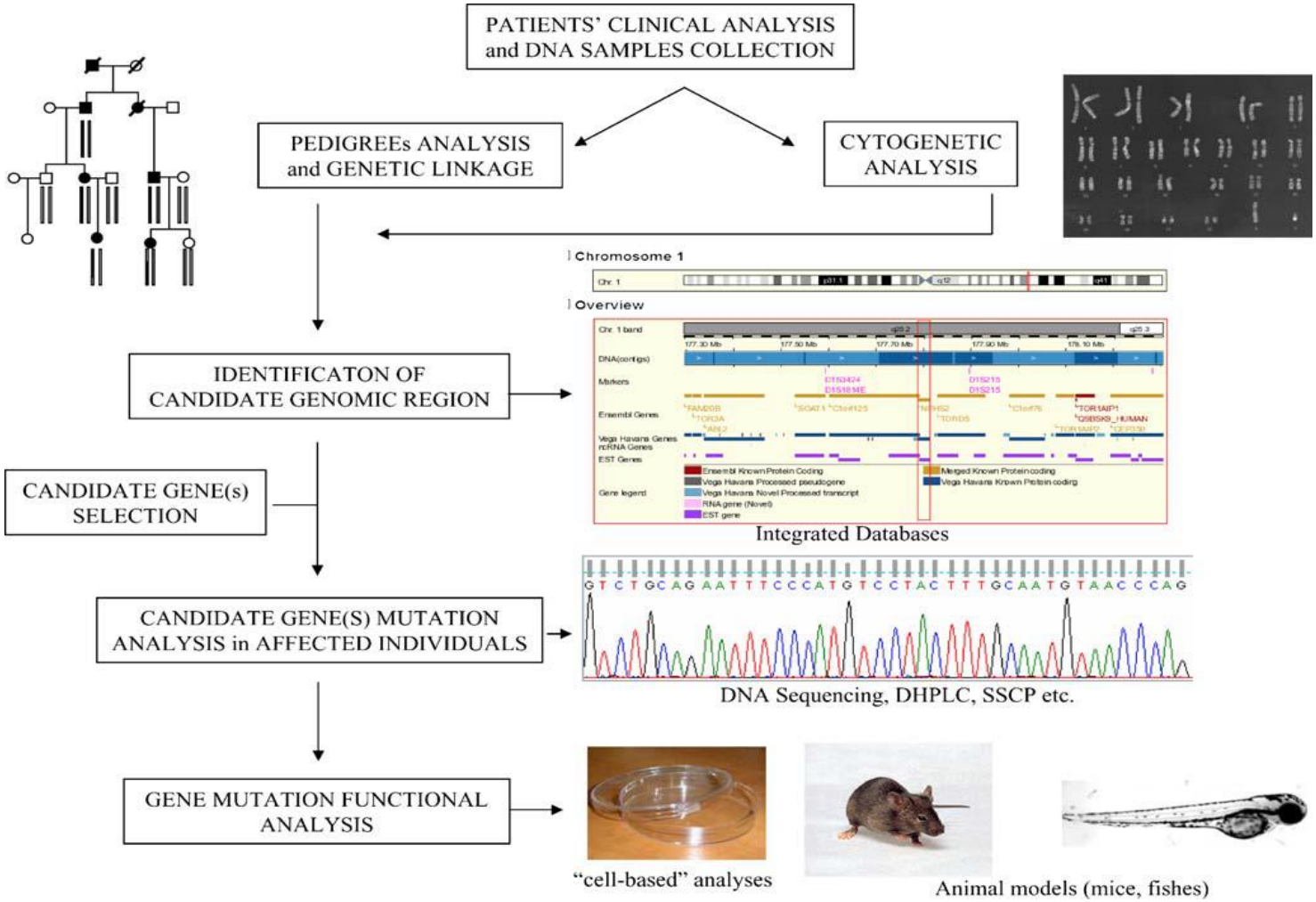


Figure 1–9: Schematic representation of the main steps in the positional cloning approach to identify disease-causing genes. First, linkage or cytogenetic analyses are performed to identify critical genomic regions that are segregating with the phenotype. Next, candidate gene(s) are selected based on their biological function. Mutational analysis is subsequently performed to identify relevant mutations, aiming to establish genotype-phenotype causative relationship. Finally, *in vitro* and/or experiments on animal model are performed to achieve phenotypic characterisation and understand the disease pathogenic mechanisms. Figure adapted from (Puliti et al., 2007).

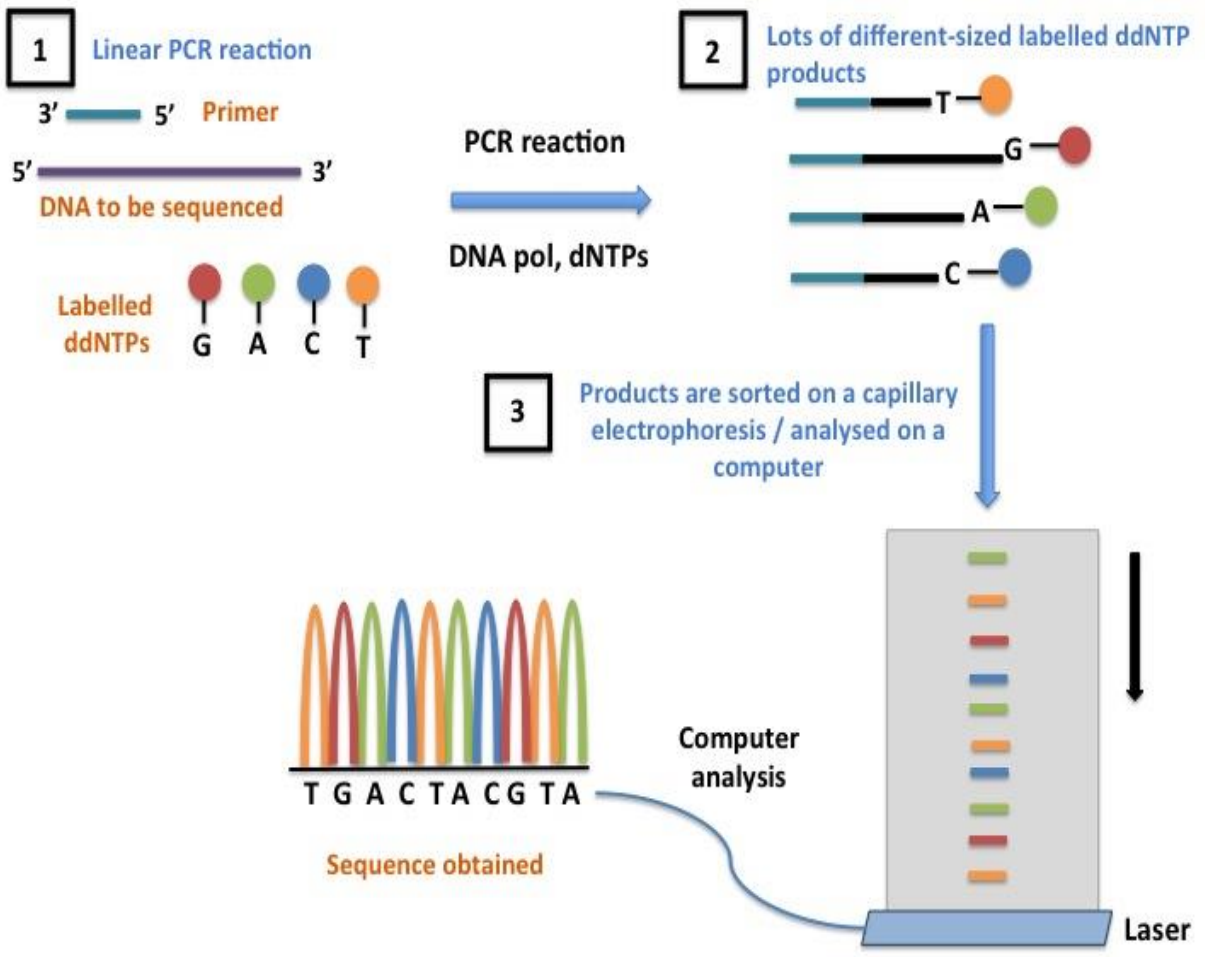


Figure 1–10: A schematic depiction of the multisteps involved in Sanger sequencing technology using capillary electrophoresis. Prior to sequencing reaction, DNA fragments are amplified using standard PCR. (1) A liner sequencing reaction, using either forward or reverse primer, is performed using fluorescently labelled dNTPs (dioxynucleotides) and ddNTPs (dideoxynucleotides). ddNTPs lacks the 3'-hydroxyl group that is required for elongation of the growing nucleotide chain. The strand synthesis of DNA fragments starts with the incorporation of dNTPs until a ddNTP is randomly incorporated into the synthesis reaction, causing the DNA synthesis process to terminate. (2) Labelled DNA fragments of different sizes are subsequently generated, (3) which are electrophoresed by capillary electrophoresis. A laser scanner positioned at the bottom of the capillary electrophoresis scans the labelled DNA and sends the sequences to a connected computer to interpret the data. Original figure, compiled from information in (Kircher and Kelso, 2010).

1.7 NGS technologies and human disease

The introduction of NGS technologies has changed the landscape of RGDs and cancer research. NGS encompasses multiple sequencing techniques that allow for massive parallel sequencing of the entire genome using whole genome sequencing (WGS) or targeted regions of the genome, such as WES or targeted sequencing (Goodwin et al., 2016). Since the first WES proof-of-concept study in 2009 (Ng et al., 2009), the rate of disease-causing genes discovery has increased using NGS technologies, primarily WES (Figure 1–11) (Boycott et al., 2017). Compared with positional cloning, NGS provides a more comprehensive analysis of the entire genome or the exome without the need to perform linkage analysis and prioritise candidate genes, reducing the disease-causing investigation process to a single step (Gilissen et al., 2012).

Since 2008, the cost of NGS technologies has dramatically decreased, making these technologies more accessible to the scientific community (Figure 1–12). This cost reduction facilitated the use of these techniques in research; moreover, it allowed for launching of large genomic projects that aim to analyse larger cohorts of samples, such as the NHS UK's 100,000 Genomes Project launched in 2012 (Siva, 2015). This project aspires to implement genomics testing to transform the way patients are diagnosed and treated by sequencing 100,000 whole genomes from patients who have rare diseases, infectious diseases or cancer (Samuel and Farsides, 2017). Characterising genetic variants and transforming this knowledge into clinical practice will pave the way for personalised medicine.

1.7.1 Targeted sequencing and WGS

The targeted NGS method analyses specific regions of the genome that have been linked to disease pathological mechanisms or that are related to established phenotypes. Based on the research setting and user preference, this approach can target intronic, intergenic or exonic regions (Dilliott et al., 2018). Targeted NGS is an efficient method to study a phenotype for which a foundation of candidate genes has been established, eliminating the generation and analysis of superfluous and irrelevant genetic data. Another advantage of using targeted NGS is that a minimal amount of starting material (as little as 100 cells) is required to obtain ultra-deep DNA sequencing data (Sibinga Mulder et al., 2018). Although the cost of WGS and WES is decreasing, targeted NGS can be a cost-effective cost approach, especially when studying a large cohort of samples (Dilliott et al., 2018). Targeted NGS panels have been successfully utilised in clinics, particularly in cancer diagnostics (Anis et al., 2018). Nevertheless, a major limitation of targeted NGS is missing potentially disease-related variants in regions that are not captured by the designed sequencing panel.

WGS enables massive-parallel sequencing of the entire genome (~99%) that consists of 3 billion nucleotides, analysing exons, non-coding introns, promoter, and enhancer regions (Nakagawa et al., 2015). WGS can identify structural aberrations, including translocations that can lead to gene fusions (Bertier et al., 2016). Compared with WES, WGS provides a more uniform data coverage across the genome, which therefore reduces the number of false positive calls (Belkadi et al., 2015). In 2007, the cost of WGS was approximately $10 million; however, this price has fallen to below $1,500 in 2015 and is expected to become cheaper over time, making WGS more attractive for research and large sequencing projects (Payne et al., 2018). However, WGS results in a high data output for which analyses and biological interpretations remain

challenging. With the emergence of nationwide sequencing projects and the continuous development of bioinformatic analysis tools, comprehensive analytical frameworks will improve and, thus, facilitate WGS data analysis, particularly for the interpretation of non-coding variants (Boycott et al., 2017; Nakagawa et al., 2015).

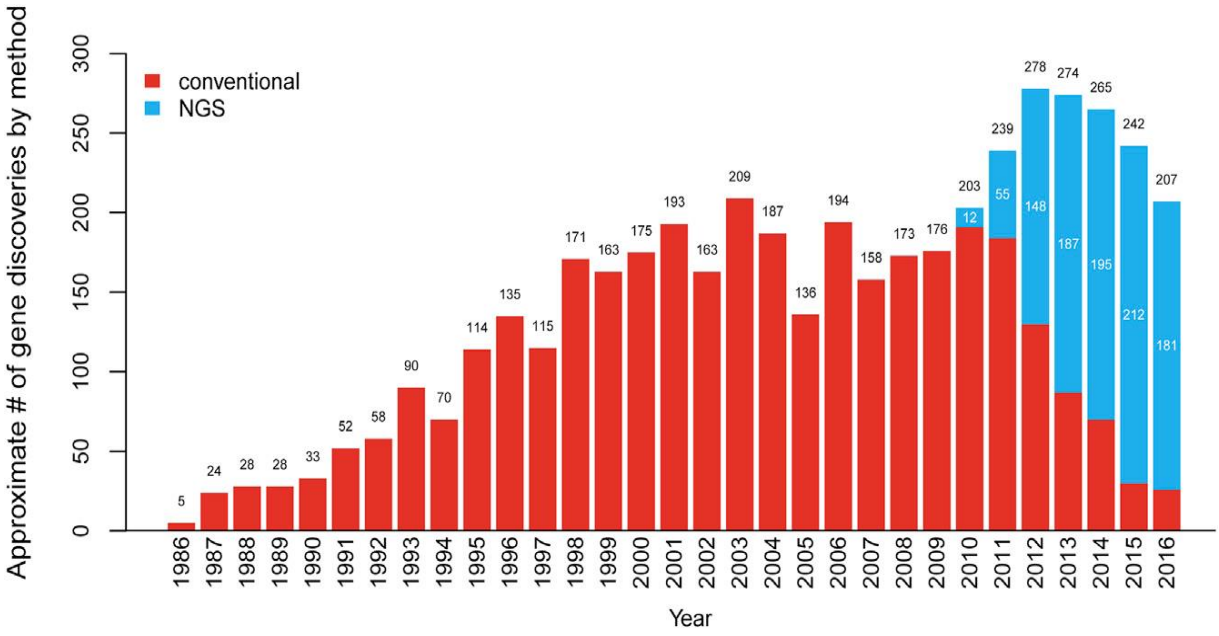


Figure 1–11: Approximation of the number of genes discovered by conventional methods versus WES and WGS since 2010. The introduction of NGS technologies (including WES and WGS) has accelerated the pace of disease-causing genes discovery. Since 2012, a steadily increase of gene discovery by conventional approaches (red) or NGS technologies (blue) has been observed. A preference shift of using NGS technologies over conventional methods is evident since this time. Graph adapted from (Boycott et al., 2017).

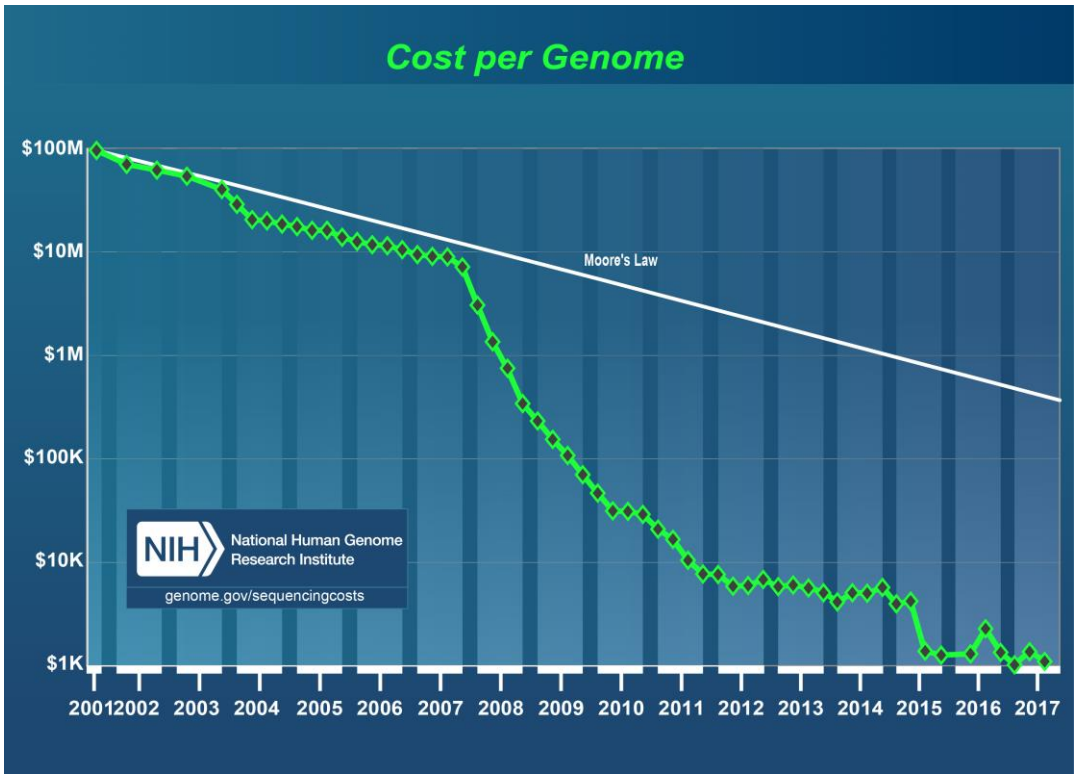


Figure 1–12: The cost of genome sequencing in comparison to Moore's law. A dramatic decrease of genome sequencing has been observed since 2008. The Moore's law is originally documented within the computer science field, which states that computer power doubles every two years at the same cost (linear line). Graph adapted from National Human Genome Research Institute (https://www.genome.gov/27541954/dna-sequencing-costs-data/), with added information from (Waldrop, 2016).

1.7.2 WES: definition, experimental process and details of Illumina sequencing platform

Unlike WGS, WES analyses the protein-coding regions of the genome (the exome), allowing for the simultaneous analysis of exonic and splice site genetic changes (Petersen et al., 2017). Although WES focuses on only ~1–1.5% of the genome, it has been suggested that the exome harbours 85% of disease-related mutations that can have functional impact on the protein (Rabbani et al., 2014). Moreover, the majority of monogenic diseases are caused by mutations in the exome (Kuhlenbaumer et al.,

2011). Hence, the exome represents a functional subset of the genome to identify causal mutations in RGDs and cancer.

The entire process of a WES experiment can be grouped into three major steps: (1) sample preparation (including library preparation) and massive parallel sequencing; (2) raw data analysis; and (3) interpretation of data (Figure 1–13). Multiple exome capturing methods and sequencing platforms are commercially available to users. Exome capturing kits have been estimated to target 95% of all the exons in the genome (Rabbani et al., 2014). Of these kits, the Agilent's SureSelect exon capture kit was used in all the WES studies presented in this thesis. This kit captures exonic regions using streptavidin beads, followed by amplification (enrichment) of genomic DNA attached to sequence-specific adaptors (Figure 1–14). Illumina sequencers such as HiSeq2500 are one of the most popular NGS platforms on the market (Sims et al., 2014). These sequencers use the clonally amplified template method to produce randomly distributed and clonally amplified clusters of DNA molecules. Repaired with sequence-specific adaptors, DNA attaches to the sequencing slide through the binding of the adaptors to the primers already attached to the slide. Bound DNA is subsequently amplified by the clonally amplification method, also known as bridge amplification, in which the immobilised template amplifies with the immediately adjacent primers to form 'bridge clusters' (Figure 1–15A). Each generated cluster is derived from one DNA fragment. These generated clusters are then massively sequenced using modified 3′-blocked reversible nucleotides that are coloured differently (Figure 1–15B). A fluorescent colour is emitted every time a labelled nucleotide is incorporated into the chain being synthesised, allowing for recording and identification of the incorporated bases (Metzker, 2010).

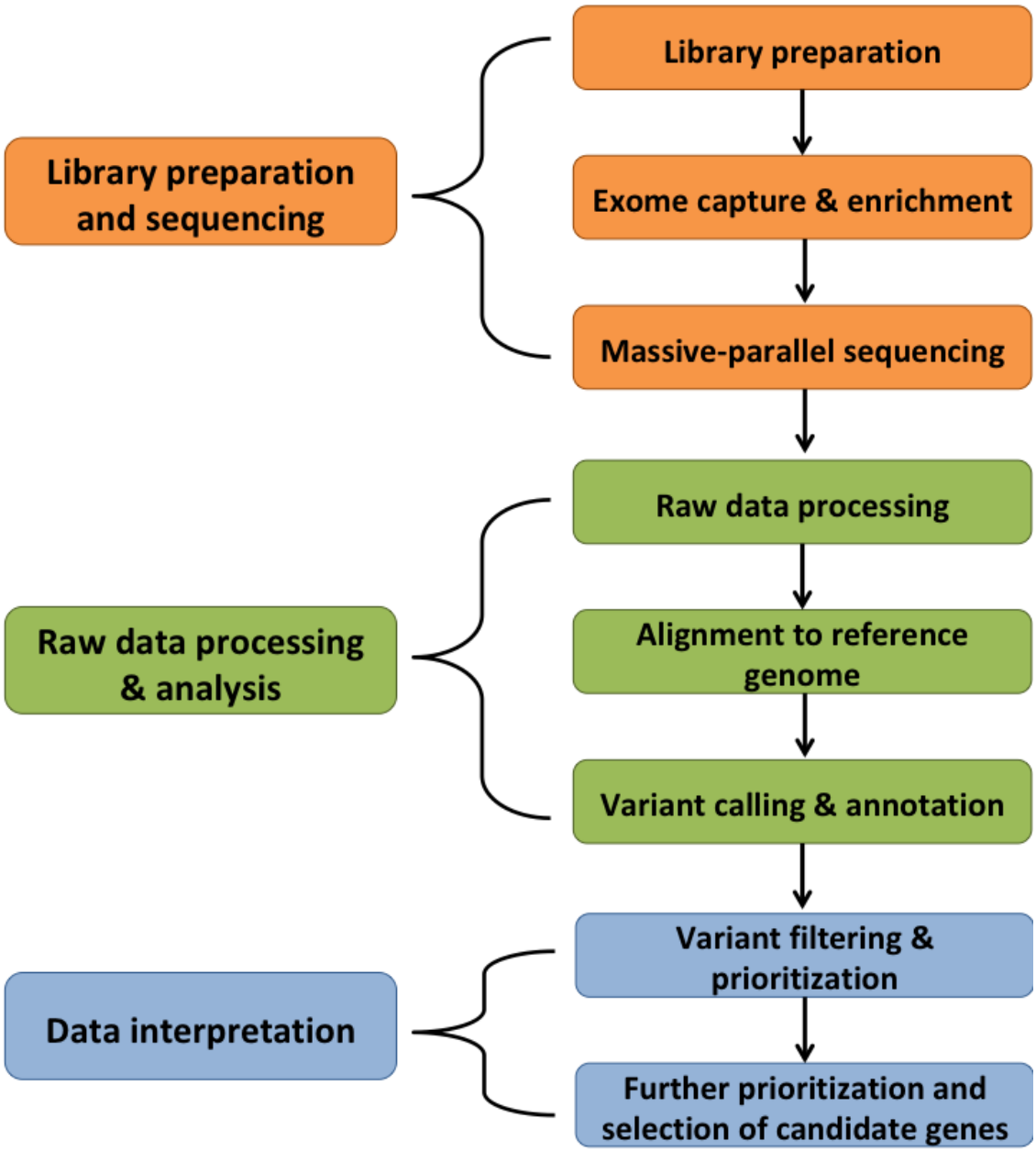


Figure 1–13: A schematic representation of the major steps in a WES experiment. First, library preparation and exon capture are performed as in Figure 1–14. Enriched library undergoes massive-parallel sequencing, for example using HiSeq2500 Illumina platform (see Figure 1–15). Quality checks of the generated raw data are performed to evaluate the quality of raw reads and to remove PCR generated duplicates. Processed reads are then aligned to a reference genome (e.g., hg19), followed by identification and annotation of variants. Annotated variants are subsequently filtered and prioritised to identify variants of biological relevance. For example, variants that fit an autosomal inheritance pattern are prioritised (as in Chapter 3). In cancer studies (Chapter 5 & 6), somatic variants that are present in tumour tissue but absent from corresponding normal samples are prioritised. Biological interpretation and further prioritization of candidate variants is performed based on the design and goal of the study. Original figure, compiled from information in (Bamshad et al., 2011; Dolled-Filhart et al., 2012; Pabinger et al., 2014).

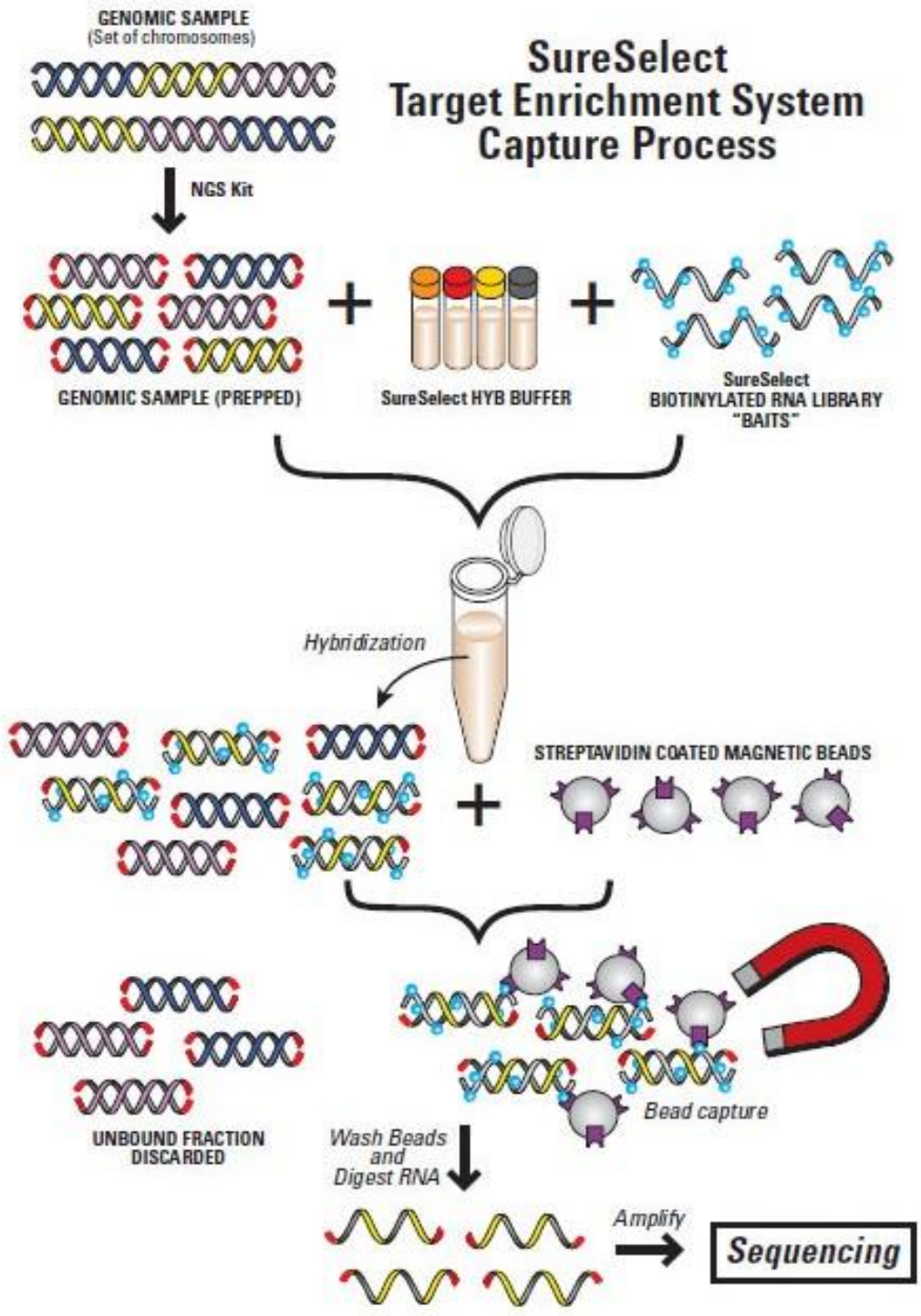


Figure 1–14: The SureSelect target enrichment system to capture exons during WES library preparation. First, the genomic DNA is sheared. The fragmented DNA is attached to sequence-specific adaptors (red ends), then are hybridised with exon-specific biotinylated RNA library baits to select for the exonic regions. Targeted regions are pulled down using bead capture; by contrast, intronic regions are washed away. Afterward, DNA is washed to remove unbound DNA fragments. The selected fragments are enriched by PCR amplification, and then loaded onto a massive-parallel sequencer (Figure 1–15). Figure obtained from (http://www.genomics.agilent.com/files/Media/SS_Halo/Magnet584.jpg).

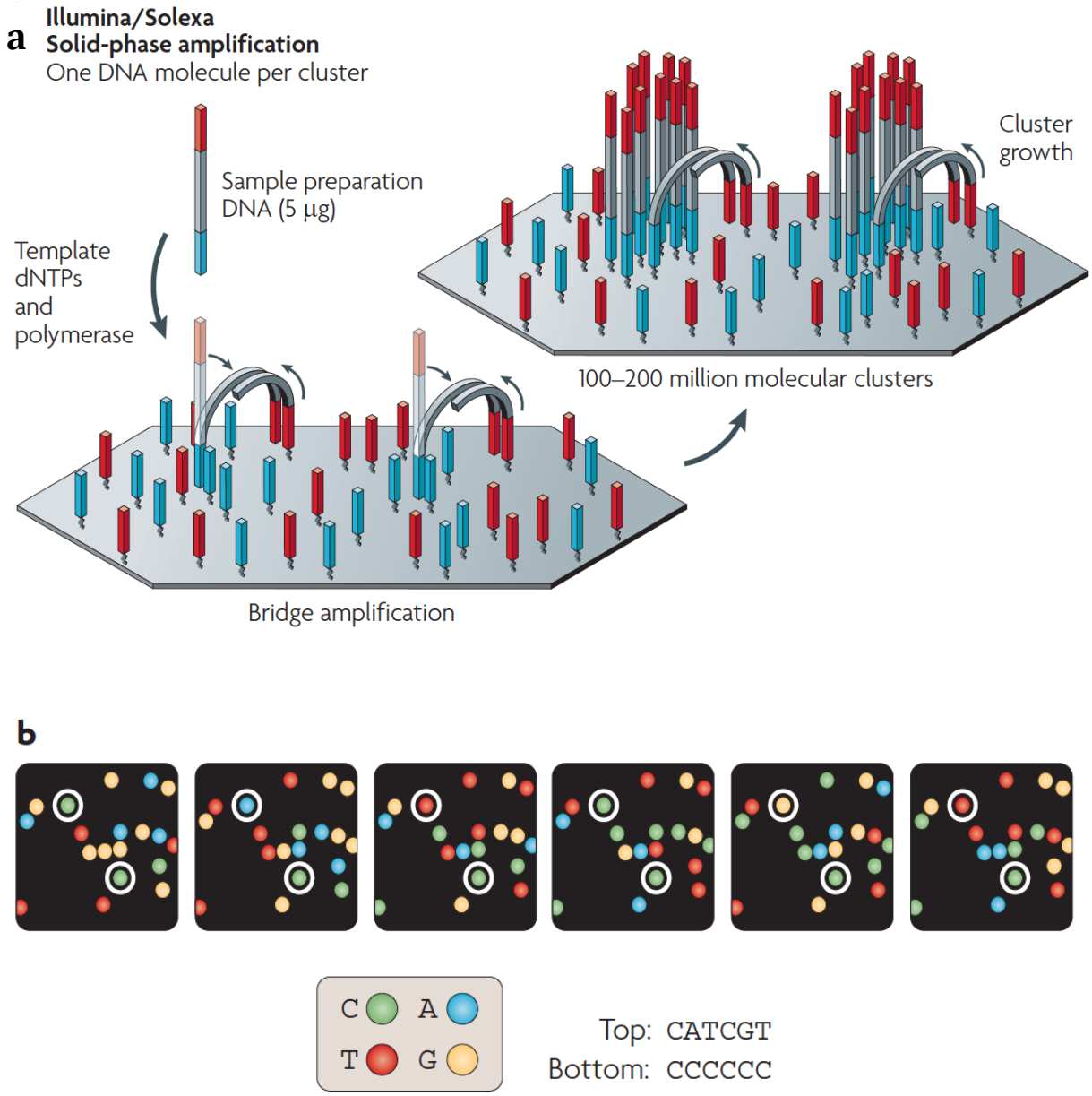


Figure 1–15: Diagrammatic representation of the Illumina massive-parallel sequencing platform. (A) Sequence-specific adapter are attached to the end of the DNA. DNA molecules attaches (via complementary pairing) to the primers already adhered to the sequencing slide. These attached DNA fragments amplifies and form cluster bridges using the adjacent primer already annealed to the slide. (B) To sequence DNA clusters, DNA fragments are synthesized using differently coloured nucleotides (C, A, T, G). A fluorescence colour is emitted every time a base pair is incorporated into a cluster and each emitted colour is recorded. Figure reproduced with minor modifications from (Metzker, 2010).

1.7.3 Disease gene identification strategies using WES: bioinformatics analyses and prioritisation of variants

NGS technologies result in a large data output that requires substantial bioinformatics resources and data analysis. Robust bioinformatics algorithms have been continuously growing to support NGS data analysis, including mapping, variant calling and annotation (Pabinger et al., 2014). Moreover, the wealth of NGS-generated data has encouraged the development of comprehensive databases, which can store information on genetic changes within a general population and make this information publicly available to users (Boycott et al., 2013; Pabinger et al., 2014). Examples of these databases include the Single Nucleotide Polymorphism database (dbSNP) (https://www.ncbi.nlm.nih.gov/projects/SNP/), the Exome Aggregation Consortium (ExAC) (dbsnhttp://exac.broadinstitute.org) and COSMIC (https://cancer.sanger.ac.uk/cosmic). These databases are routinely being used in the research and clinical settings for data analysis and prioritising variants (Boycott et al., 2017; Forbes et al., 2017).

The number of identified variants in WES studies varies considerably due to exome enrichment methods, sequencing platforms and the bioinformatics tools used in variant calling (Boycott et al., 2013). It has been estimated that 20,000–50,000 variants can be detected per sequenced exome (Gilissen et al., 2012). Successful studies usually apply stepwise filtering approaches to reduce this large number of variants and pinpoint candidate variant(s) of biological significance (Biesecker and Green, 2014; Ng et al., 2009) (Figure 1–16). The choice of the filtering scheme is not mutually exclusive; however, genetic studies, including RGD studies, will usually retain variants that have direct impact on the protein produced. That is, intronic and synonymous variants are discarded because these changes are likely to have minimal effect on the protein; by

contrast, splice site, nonsynonymous and INDEL alterations are prioritised. Afterwards, variants that are common in public databases are excluded from the list as they are likely to be benign, resulting in an average of 150–500 variants that can be classified as potentially pathogenic (Gilissen et al., 2012). Even applying this common filtering scheme, the remaining number of private variants is still high in family-based WES studies and, therefore, requires additional tailored analytical frameworks to pinpoint the most likely disease-causing variant.

1.7.3.1 Analytical frameworks that can be applied in disease-gene discovery of RGDs using WES

The common prioritisation scheme presented in Figure 1–16 is only a preliminary data analysis step and is clearly not enough to determine the disease-causing variant(s). A study by MacArthur et al. (2012) demonstrated that a human genome can harbour ~100 genuine loss-of-function variants, affecting approximately 20 genes. Thus, an additional tailored framework is needed to reduce this number of private variants and, in turn, establish a disease-gene relationship. As mentioned earlier, the choice of the analytical framework is not mutually exclusive; nevertheless, it can be chosen based on the expected disease mode of inheritance, accessibility of additional family members (family-based approaches), or availability of unrelated individuals with the same disease phenotype (Beaulieu et al., 2014; Gilissen et al., 2012). Choosing the best analytical framework will increase the likelihood of identifying disease-related variants, resulting in an optimal and cost-effective experimental design.

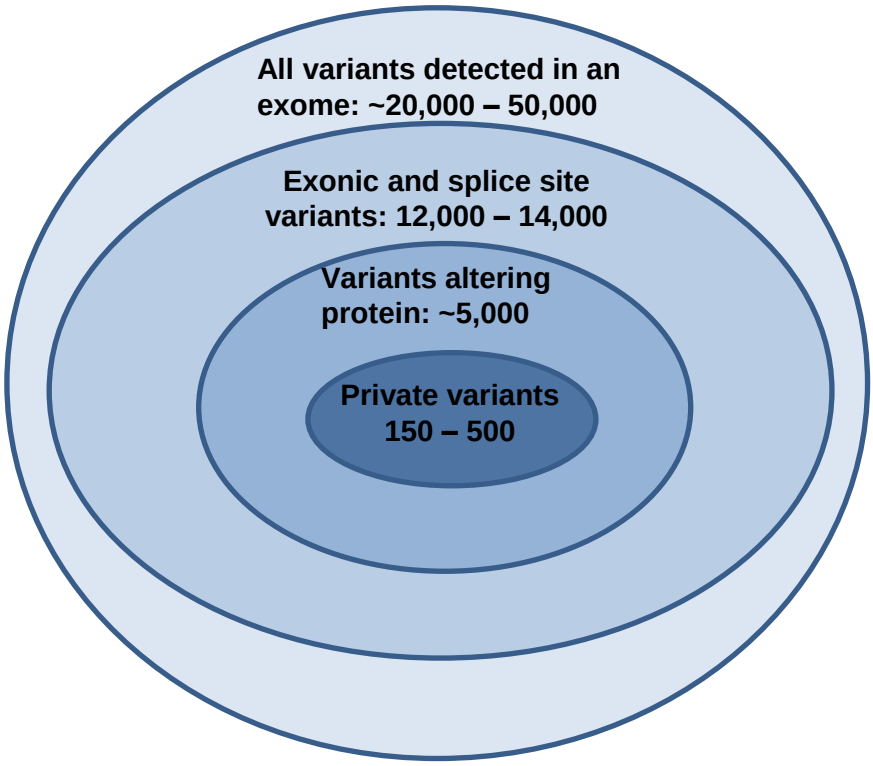


Figure 1–16: Common prioritisation scheme for WES variants. WES studies of rare diseases usually follow a filtering/prioritisation approach to reduce the number of variants that are likely non-pathogenic, while retaining variants that can alter the protein sequence. The size of the circles is indicative of the number of the variants retained after each filtering step. Figure redrawn with modifications from (Gilissen et al., 2012).

When a defined rare phenotype is present in multiple affected family members, the chance of the disease being monogenic is high (Boycott et al., 2013). Understanding the mode of inheritance of the disease under study is essential for the selection of additional family members which can be useful in excluding private benign variants (Gilissen et al., 2012). In the absence of additional family members, knowledge about the disease inheritance pattern can also help in choosing the most appropriate WES analytical framework (Boycott et al., 2013; Gilissen et al., 2012). For example, in some cases, WES performed on one affected sibpair (two siblings with both parents in common) can lead to the identification of the disease-causing gene (Figure 1–17A). A study by Schuurs-Hoeijmakers et al. (2012) performed WES on a sibpair presenting

with a complex form of hereditary spastic paraplegia and successfully identified compound heterozygous mutations in *DDHD2*.

In consanguineous families with a surmised AR inheritance, the disease-causing gene is expected to be homozygous in the affected member and heterozygous in unaffected parents (Figure 1–17B). Hence, sequencing the DNA of unaffected parents along with that of the affected individual, an approach known as trio-analysis, will be effective in reducing the number of genes to a reasonable number. Alternatively, in the absence of one unaffected parent, the DNA of unaffected siblings can be sequenced (Boycott et al., 2013). Both of these approaches were utilised in the four CHT families investigated in Chapter 3.

In families with an expected X-linked recessive inheritance, the disease-causing variant is expected to be heterozygous in unaffected females and hemizygous in affected males on the X-chromosome (Figure 1–17C). Hence, variants in autosomal chromosomes can be safely discarded in X-linked studies. However, differentiation between X-linked and AR inheritance patterns using pedigree information can be challenging, especially in families with only male offspring (Boycott et al., 2013). This challenge was evident in a study by Sankaran et al. (2012) that was conducted in two male siblings affected with Diamond-Blackfan anaemia, a bone marrow disorder. Diamond-Blackfan anaemia usually follows an AD inheritance; however, both parents studied showed normal haematological analyses, suggesting an AR or X-linked inheritance. WES conducted on two affected male siblings did not reveal any variants fitting an AR inheritance pattern; nevertheless, a splice site mutation in *GATA1* on the X-chromosome appropriately segregated in the family. *GATA1* was successively validated in an additional cohort of male patients affected with the same disorder.

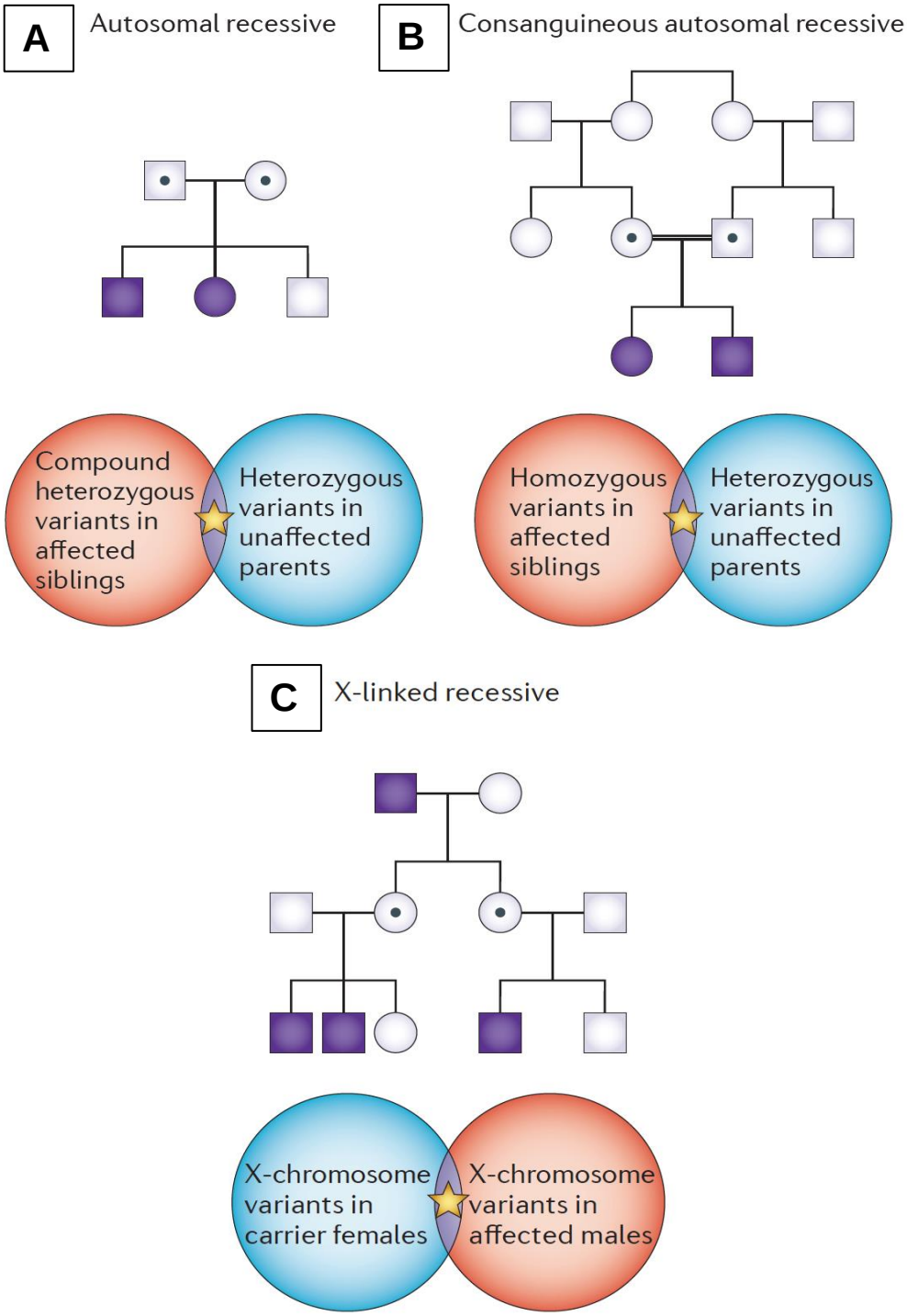


Figure 1–17: Analytical frameworks for disease-gene identification using WES. Males are represented by squares; females by circles. Purple coloured symbols represent affected members with a rare genetic disease (RGD). A black dot inside a symbol indicates an unaffected disease carrier (heterozygous). The star represents the area in which the disease-causing gene is expected to be present. Mode of inheritance is an important factor for selection of family members. (A) In an autosomal recessive RGD, WES of a sibpair can be approached, or (B) in consanguineous families, a trio approach (mother-father-affected), or alternatively, an unaffected parent, affected and unaffected sibling can be exome-sequenced. (C) In X-linked recessive diseases, two related affected males or an affected male and unaffected mother could be analysed. Images adapted from (Boycott et al., 2013).

The identification of AD disease-causing genes using WES has been more challenging due to the larger number of private heterozygous changes than private homozygous variants (Beaulieu et al., 2014; Boycott et al., 2013). In families with multiple members affected with AD disorders, two affected and one unaffected family members can be exome-sequenced to identify heterozygous variants that are shared between the affected patients and absent from the unaffected member (Figure 1–18A). In the absence of additional family members (small families) with an expected AD inheritance, an alternative framework can be approached in which data from multiple unrelated patients (same phenotype) are combined to identify variants that overlap between the patients, resulting in small number of candidate genes for subsequent analysis (Gilissen et al., 2012) (Figure 1–18A). The combination of WES data from multiple unrelated patients can also be applied to diseases associated with *de novo* dominant mutations. For instance, a study by Hoischen et al. (2011) combined the WES data of three clinically defined unrelated patients with Bohring-Optiz syndrome and successfully identified nonsense *de novo* mutations in *AXSL1*.

1.7.3.2 Identification of somatic cancer variants using WES

Somatic mutations are hallmarks of cancer. NGS technologies are very useful tools to compressively characterise acquired somatic mutations in cancer samples. As mentioned earlier, before the advancement of NGS technologies, cancer studies relied on traditional Sanger sequencing, which is limited by throughput. In conventional approaches, the selection of candidate genes primarily relies on a previous knowledge about the association of these genes in cancer (e.g., cancer drivers). Sanger sequencing is labour intensive and time consuming, especially when analysing

multiple candidate genes. Therefore, performing a genome-wide analysis using the Sanger-based method is almost impossible (Watson et al., 2013).

WES has accelerated the rate of discovering cancer-associated genes and helped to decipher the mutational landscape of various cancer subtypes (Wang et al., 2013). WES can be performed on matched normal-tumour pairs to identify genome-wide somatic variants. Using somatic caller tools such as VarScan2 (more in Chapter 2), somatic variants are expected to be present in tumour samples and absent from the matched normal material (Figure 1–19) (Koboldt et al., 2012). Somatic variants may occur at low frequencies and can also be influenced by tumour purity and heterogeneity. Depending of the number sequenced samples, deep sequencing (sequencing to a higher depth) is advised to detect low frequency variations (Ku et al., 2016).

Various ongoing international collaborative efforts have helped to characterise and coordinate comprehensive catalogues of genomic alterations generated using NGS data, accelerating the understanding of the pathogenic mechanisms of cancer (Ku et al., 2016). For example, The Cancer Genome Atlas (TCGA) (https://cancergenome.nih.gov) established in the USA is a comprehensive effort to molecularly characterise cancer. In April 2018, TCGA finished the Pan-Cancer Atlas projects, which provided a comprehensive molecular knowledge of over 11,000 tumours from 33 of the most common subtypes of cancer. In addition to deciphering the pathogenic mechanism of the tumour, identification of cancer somatic mutations also allows for the development of personalised therapeutics (Ryu et al., 2016).

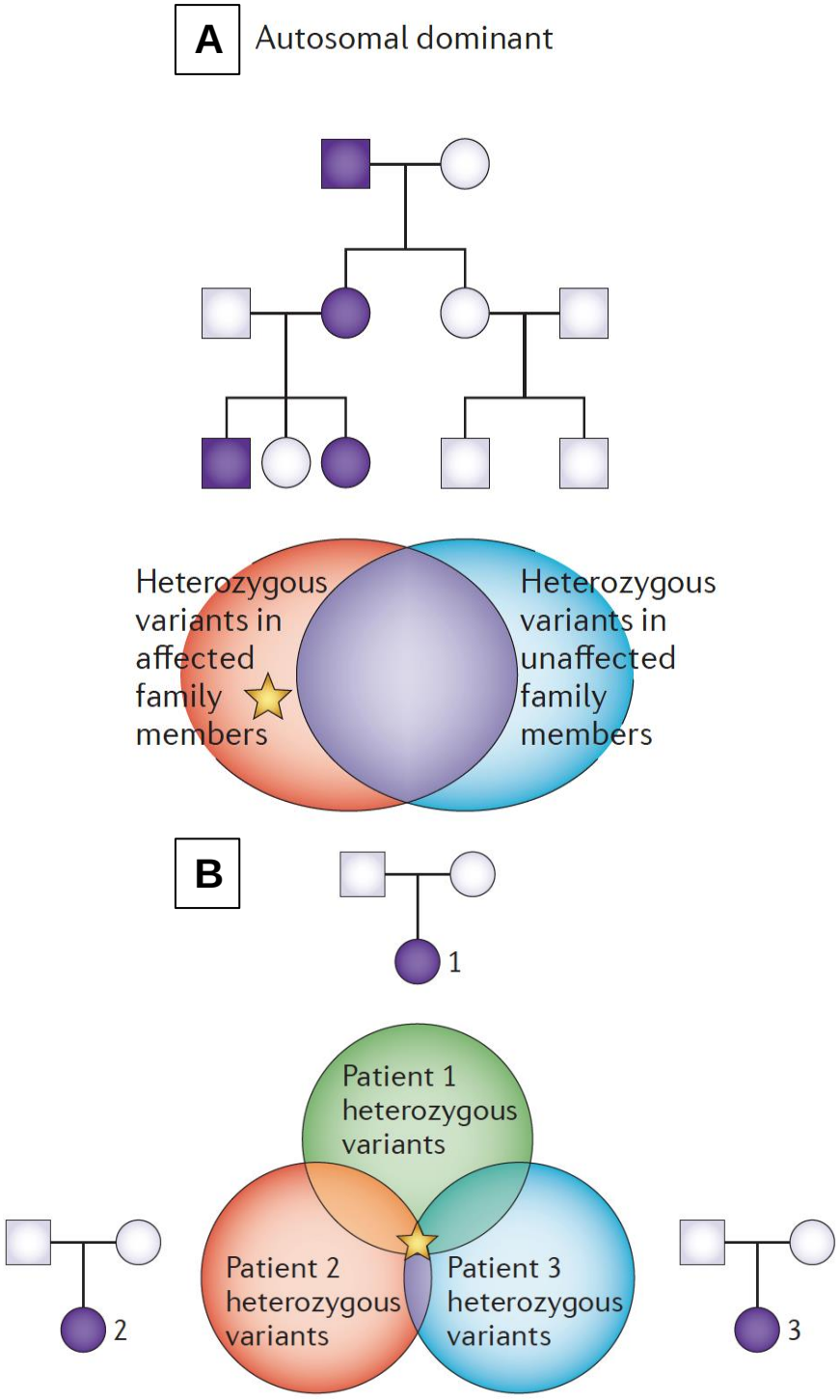


Figure 1–18: Two WES analytical frameworks for analysing autosomal dominant and *de novo* disorders. Note: Refer to Figure 1–17 for information about the symbols. (A) In autosomal dominant disorders, the affected and unaffected family members can be analysed; however, this approach is considered challenging (B) Combing the WES data from multiple unrelated patients, affected with the same phenotype, can be approached to identify *de novo* dominant mutations that overlap between the sequenced patients, in turn, identifying the disease-causing gene. In addition to affected patients, unaffected parents can also be sequenced (by WES or Sanger sequencing) to confirm the *de novo* occurrence of the variant. This overlapping approach may also be used in multiple patients affected with an autosomal dominant disorder. Images adapted from (Boycott et al., 2013).

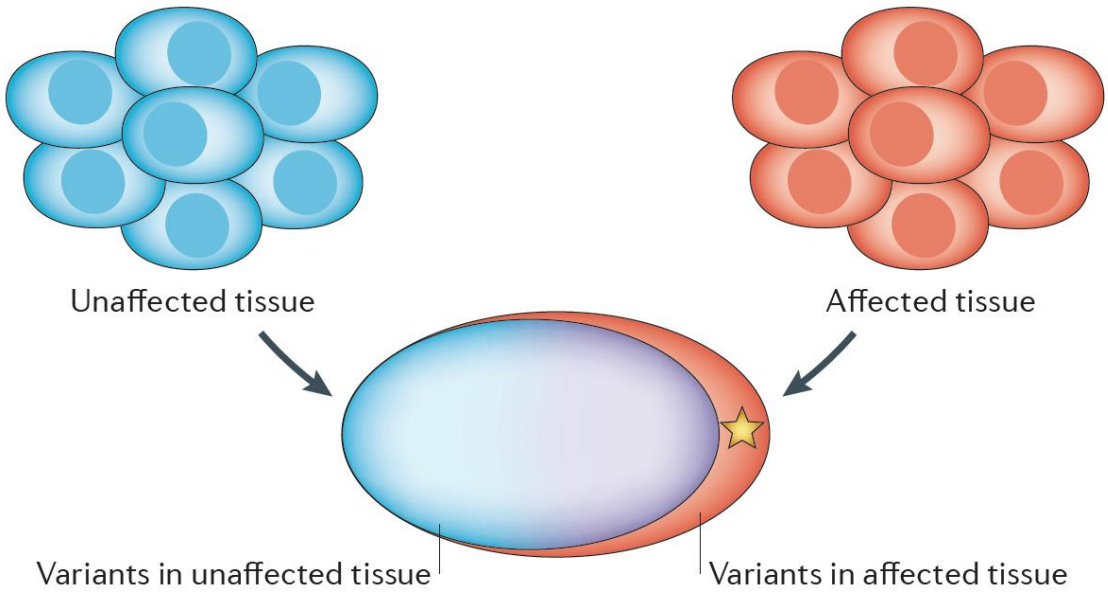


Figure 1–19: Identification of somatic variants using WES data from tumour and corresponding normal samples. Bioinformatic somatic calling tools analyse the WES from the tumour (affected) tissue and corresponding normal samples (unaffected) to identify somatic/acquired variants. The star represents the area in which somatic variants are expected to be identified. Figure adapted from (Boycott et al., 2013).

1.7.4 RNA-Sequencing: overview and technology workflow

Gene expression is a fundamental cellular process that regulates the expression of gene products and, in turn, directs protein synthesis (Garcia-Sanchez and Marques-Garcia, 2016). The expression levels of genomic loci are time- and cell-type-dependent and are tightly controlled by specific cellular conditions. Although only 2–3% of the genome encodes protein-coding genes, it has been estimated that 85% of the human genome can be transcribed (Hangauer et al., 2013). Therefore, investigations into the transcriptome can lead to a better interpretation of the functional elements of the genome and a wider understanding of pathogenic mechanisms of diseases (Wang et al., 2009).

Initial studies of the transcriptome primarily relied on hybridisation-based array approaches. These techniques involve incubating fluorescently labelled complementary DNA (cDNA) with custom-made microarrays. Identification and quantification of spliced isoforms were performed using specifically designed arrays that have probes spanning exon-intron junctions (Wang et al., 2009). Although hybridisation-based techniques are generally considered inexpensive, these technologies have several limitations. One limitation is the high background levels of results due to the cross-hybridisation nature of the technique (Ozsolak and Milos, 2011).

RNA-Seq, also known as transcriptomic sequencing, is an NGS platform that sequences RNA molecules, allowing for explorations of the entire transcriptome simultaneously (Byron et al., 2016; Han et al., 2015). RNA-Seq provides an analysis of gene expression levels by quantifying the abundance of gene transcripts. Compared with DNA microarrays, RNA-Seq has very minimal background levels (almost negligible) because DNA sequences can unambiguously map to unique regions of the genome (Wang et al., 2009). Moreover, hybridisation-based techniques require a prior knowledge of the genomic structure to choose the most appropriate platform, for example, a platform that targets CNVs, promoter regions or exonic boundary regions (Curtis et al., 2009). Unlike microarrays, RNA-Seq provides exploratory information about novel transcripts species, including non-coding RNA and novel transcripts arising from genomic structural aberrations (gene fusions) (Han et al., 2015). In addition, RNA-Seq provides information about the transcriptome at single base pair resolution (Han et al., 2015). Altogether, RNA-Seq revolutionised transcriptomic studies by overcoming the limitations of hybridisation-based array technologies.

Similar to the WES workflow, the RNA-Seq experiment workflow consists of three primary sections: experimental biology (sample preparation and sequencing), computational biology (alignment and data processing), and data interpretation (differential gene expression or gene fusion discovery). In the experimental biology section, cDNA sequences derived from mRNA are used to construct the RNA-Seq library, followed by massive-parallel sequencing (Han et al., 2015). Successful RNA-Seq experiments require high quality and minimally degraded mRNA with minimum ribosomal RNA content (Conesa et al., 2016).

In general, construction of a cDNA library consists of converting the total or fractioned RNA molecules to cDNA fragments, followed by attaching the adaptors to both ends (Figure 1–20). Subsequently, the cDNA library is massively sequenced (Han et al., 2015). Several commercially available NGS platforms, the same as the ones used in WES, can be utilised for RNA-Seq, including Illumina, Rosche 454 Life Science and Applied Biosystems SOLID (Wang et al., 2009). An Illumina high-throughput sequencing platform was used in the RNA-Seq studies conducted on bone tumours in Chapters 5 & 6 (see Section 2.4.1). Reads can be sequenced from one end (single-end) or both ends (paired-end), of which the former is generally suitable for gene expression profiling in well-annotated organisms. By contrast, paired-end sequencing is the preferred sequencing input for the detection of novel or alternatively spliced transcripts and gene fusions (Conesa et al., 2016; Ozsolak and Milos, 2011).

In the computational biology part, raw sequenced reads are aligned to a reference genome or transcriptome genome after the removal of technical artefacts. The qualified map reads are quantified to determine differentially expressed genes (Wang et al., 2009; Han et al., 2015). Other bioinformatic tools can be used, such as splice junction mapping tools (e.g., TopHat2-Fusion, Section 2.4.3.1) to identify gene fusions.

Interpretation of the data is last step to gain biological insight into the generated RNA-Seq data (Han et al., 2015).

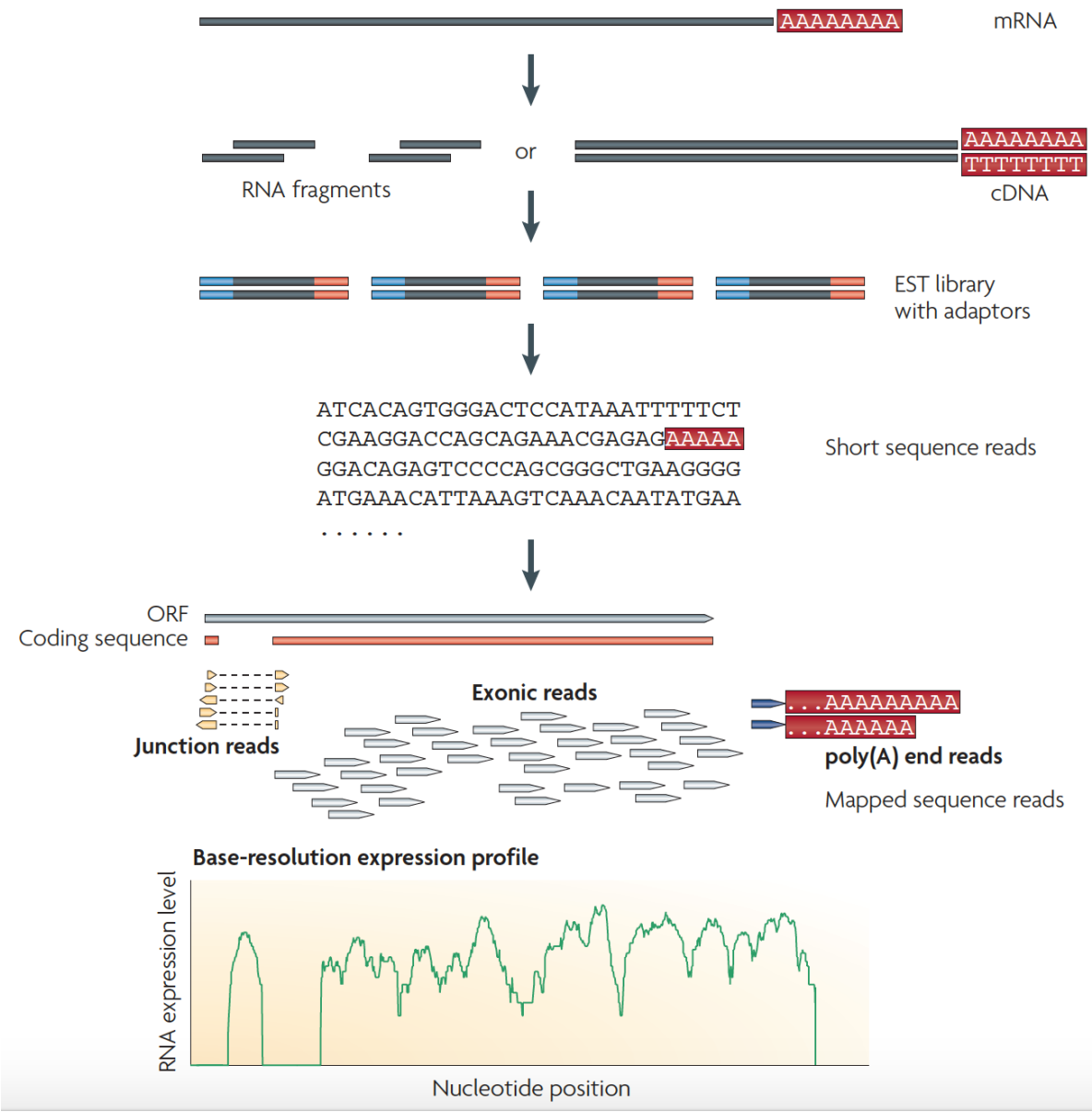


Figure 1–20: Schematic representation of a standard RNA-Seq experiment. RNA molecules are first converted to a library of cDNA fragments. Sequencing adaptors (blue) are added to each cDNA fragment, followed by massive parallel sequencing. Generated sequence reads are then aligned to a reference genome. Aligned reads can be classified as: exonic reads, junction reads and poly-A end-reads. A gene expression profile is generated using the three previously classified reads (bottom diagram). Figure adapted from (Wang et al., 2009).

1.8 Aims of the research presented in this thesis

As explained earlier, genetic studies into RGDs and cancer can provide useful insights into disease pathogenesis and diagnosis, improve or develop therapeutic options and enhance patient outcomes. A comprehensive genetic understanding of CHT has yet to be achieved (more in Chapter 3). To our knowledge, no NGS-based studies have been conducted on UPSb, adamantinoma and OFD-like adamantinoma tumours. A comprehensive understanding of the genetic and transcriptomic landscape of these rare sarcomas remains poorly understood.

NGS high-throughput technologies were utilised to investigate and expand the genetic landscape of CHT and the three rare bone sarcomas. Specifically, the aims of each study were as following:

1) In Chapter 3, WES was conducted on the DNA of four consanguineous families with CHT, aiming to identify genetic changes in previously known CHT genes or novel candidate gene(s). After a novel CHT candidate gene was identified, functional characterisation and variant segregation analyses were performed to establish a causal gene relationship (Chapter 4).
2) WES and RNA-Seq were conducted on UPSb (Chapter 5), adamantinoma and OFD-like adamantinoma tumours (Chapter 6) to investigate the genetic and transcriptomic alterations in these bone sarcomas. After identification of candidate genes, further bioinfomatic analyses were conducted to elucidate their potential involvement in tumourigenesis and determine if these candidates can be of therapeutic relevance.

Chapter 2: Material and Methods

2.1 Patient material samples

The research conducted in all the projects was in accordance with the ethical principles in the Declaration of Helsinki. Prior to starting the projects, ethical approvals were obtained from the responsible local research and institutional committees for all the projects presented in this thesis. Ethical approval reference for the tumour samples used in the current thesis is REC 12/EM/0048. Samples of all patients were pseudonymised.

The Regional Genetics laboratory at Birmingham's Women Hospital provided all the CHT samples. The Royal Orthopaedic Foundation Trust Birmingham Tumour Bank provided all the UPSb, adamantinoma and OFD-like adamantinoma bone tumour samples, courtesy of Dr Vaiyapuri Sumathi (Royal Orthopaedic Hospital Foundation Trust, Robert Aitken Institute of Clinical Research, University of Birmingham, Birmingham, United Kingdom).

2.1.1 CHT samples

The DNA samples used in the CHT project were from children with clinical and biochemical diagnoses of CHT. The CHT clinical diagnosis was confirmed based on heel prick elevated thyroid stimulating hormone (TSH) levels within seven days of birth. Samples (n=16) were collected from participants from the West Midlands as a part of a study named the Molecular Pathology of Human Genetic Disease and was approved by South Birmingham Ethics Committee. Patient clinical information was provided by Prof Timothy Barrett, Diabetes Unit, Birmingham Children's Hospital. Genetic nurse

counsellors obtained a family history and consented all the families willing to participate. DNA was extracted by the NHS genetics diagnostic laboratory from blood samples which were collected between 2008–2015. DNA was stored (at -80 ^{o} C) in the Regional Genetics laboratory, Birmingham's Women Hospital until used in the CHT project. The concentration of the DNA samples, were checked using the Nanodrop machine (Geneflow), ranged between 400-800 ng/µl and stored at -20^{o} C upon arrival.

2.1.2 UPSb of the bone samples

Dr Vaiyapuri Sumathi confirmed the pathological diagnoses of the UPSb tumours. All bone tumours were surgically removed (not microdissected) and estimated to contain approximately 80% tumour material. Fresh frozen tumour (n=13) and corresponding normal (n=10) tissue samples were obtained. In addition, one tumour and 4 corresponding normal formalin-fixed paraffin-embedded (FFPE) samples were provided. All samples were collected from previously mentioned tumour bank under the appropriate ethical approval. The age of the patients ranged from 24–88 years. Five patients were male and nine were female. Al tumours were of high-grade and ranged from 3 to 12.5 cm in size. UPSb tumours were from the following anatomical sites: ten femurs, two tibias, one humerus and one fibula.

2.1.3 Adamantinoma and OFD-like adamantinoma samples

Dr V. Sumathi reviewed the diagnosis and the histopathological classification of all adamantinoma and OFD-like adamantinoma tumours. Fresh frozen adamantinoma (n=8) and OFD-like adamantinoma (n=4) tumour-tissue samples were provided. In addition, corresponding normal tissues samples for six adamantinoma and three OFD-

like adamantinoma tumours were provided. The age of the adamantinoma patients ranged from 7–55 years. All adamantinoma tumours were of a low-grade except for two samples that were intermediate-grade. Adamantinoma tumours were from five male and three female patients and ranged from 3–12.5 cm in size. All adamantinoma tumours occurred in the tibia, except for one tumour that developed above the knee. All OFD-like adamantinoma tumours were of low-grade and from three female and one male patients. The age of OFD-like adamantinoma patients ranged from 6-15. These tumours ranged from 2–8.5 cm in size and all occurred in the tibia.

2.2 Nucleic acid extraction, quantification and quality assessment

2.2.1 Tumour tissue disruption and homogenization

Mortar and pestle were used to disrupt and homogenize fresh frozen tumour tissues (stored at -80^{o} C) for subsequent nucleic acid extraction (DNA/RNA). Prior to tissue disruption and homogenization procedure, the mortar and pestle were chilled to -80^{o} C. Tissue samples were mechanically ground using a pre-chilled mortar and pestle filled with liquid nitrogen. After ensuring proper grinding and homogenization, the pulverized tissue was placed, using a cold spatula, into pre-chilled cryovial tubes and were stored at -80^{o}C until needed.

2.2.2 DNA extraction from fresh frozen sporadic cancer tissues

Following the manufacturer's instructions, DNA was isolated and purified from the pulverized sporadic tissue, using the DNA Isolation Kit (Roche Diagnostic Ltd, Cat. No.11814770001). Ground tissue was resuspended in a pre-warmed (37^{o} C) lysis

buffer, and 2µl proteinase K solution was added. The mixture was vortexed for 5 seconds and the cell lysates were incubated for 2 hours at 65ºC. After incubation, 100µl RNase solution (concentration mg/ml) was added to the cell lysates, ensuring proper mixing using a vortex, followed by an incubation period of 20 minutes at 37º C. Afterwards, the cell lysates were centrifuged at maximum speed (26900 x g) at room temperature for 25 minutes; subsequently, the supernatants were collected and transferred to sterile microcentrifuge tubes. Finally, 0.7 of the lysate volume of isopropanol was added to each supernatant and mixed by pipetting. The sample lysates were then kept overnight at -20º C to continue the extraction procedure the following day.

After the overnight incubation, the samples were spun at 26900 x g for 20 minutes and the supernatant was carefully decanted. The pellets were then washed with 1ml of ice-cold 70% ethanol and centrifuged at 26900 x g for 20 minutes. After centrifugation, the supernatants were carefully discarded and pellets were left to air dry for approximately 20 minutes. Lastly, 50µl of TE buffer was used to resuspend the pellets, followed by a two-hour incubation period at 50º C. The samples were gently mixed (by pipetting) every 30 minutes to ensure proper DNA resuspension. The resuspended DNA was stored at -20 º C for routine laboratory work or at -80 º C for long-term storage.

2.2.2.1 DNA extraction from FFPE cancer tissues

DNA from the FFPE sample was extracted using the Arcturus PicoPure DNA Extraction kit (Thermo Fisher Scientific, KIT0103). Sections of 10x10µm for the sample were provided. Following the manufacturers' protocol, two treatments of 10-15 ml of fresh xylene were used to remove paraffin from tissue sections, followed by air-drying of the tissues for five minutes. 2µg of the dried tissue was carefully scraped into a clean

1.5 microcentrifuge tube, followed by adding 150µ of Extraction Solution and vortexing gently. The mixture was then incubated for approximately 24 hours at 65ºC. To finish the procedure, the tubes were incubated at 95ºC for ten minutes to inactivate Proteinase K and were subsequently stored at -20ºC.

2.2.3 RNA extraction from sporadic cancer tissues

2.2.3.1 RNA extraction using the TRIzol-chloroform method

RNA was extracted from fresh frozen cancer tissues using the standard TRIzol-chloroform extraction method. 1 ml of TRIzol (Invitrogen, 15596-026) was added to the pulverized tissue (tissue kept on dry ice prior to the addition of Trizol). The tissue lysates were mixed by pipetting up and down and then incubated for 10 minutes at room temperature. 200µl of chloroform was added to the mixture, the microcentrifuge tubes were inverted for 20 seconds and then incubated for 15 minutes at room temperature. Following this incubation period, the sample lysates were centrifuged at 12,000 x g for 25 minutes at 4ºC. Two solution layers were subsequently formed; the upper colourless phase (containing the RNA) was carefully transferred to new RNase-free sterile tubes, followed by addition of 500µl isopropanol to precipitate the RNA. The samples were incubated at room temperature for 30 minutes and, lastly, lysate mixture was kept overnight at -20ºC.

On the following day, the samples were inverted for 15 seconds and centrifuged at 8000 RPM (revolutions per minute) for 20 minutes at 4°C. the supernatant was carefully discarded, and the pellet was washed with ice-cold 70% ethanol and centrifuged at 8000 RPM for 20 minutes at 4°C. After centrifugation, all ethanol was carefully removed, and the pellets were left to air dry for 30 minutes. Lastly, the pellets were resuspended in 30-50µl RNase-free water and stored at -80 º C until needed.

2.2.3.1.1 DNase treatment of RNA extracted using the TRIzol-chloroform method

To eliminate DNA contamination from extracted RNA samples, RNA was treated with DNase enzyme using the DNA-free DNA Removal Kit (Ambion, AM1906). The DNase treatment mixture consisted of a volume of 10X DNase buffer equivalent to 0.1 X RNA volume and 1µl of rDNase I enzyme. The prepared DNase treatment mixture was directly added to the RNA samples and incubated for 30 minutes at 37°C. To inactive DNAse enzyme, 0.1 X volume of DNase inactivation reagent was added to the RNA mixtures, followed by a two-minute incubation at room temperature with frequent mixing. Aftewards, the samples were centrifuged for 10,000 x g for 3 minutes, and the RNA was cautiously transferred to an RNase-free clean tube and kept at -80 ^{o}C until needed.

2.2.3.2 Second method for RNA extraction from cancer tissues: RNAeasy Mini kit

As an alternative to the TRIzol-chloroform method, RNA from fresh frozen cancer tissues was extracted using the Qiagen RNeasy Mini kit (Qiagen, 74104). Following the manufacturer's protocol, 600µl of RLT buffer was added to previously pulverized cancer tissue (approximately 25g, kept on dry ice until used) followed by homogenisation of the tissue lysate by pipetting up and down using a sterile syringe and needle. 600µl of ice-cold 70% ethanol was added to the mixture, the tubes were inverted for 10–15 seconds and, subsequently, the complete cell lysate mixtures were transferred into assembled RNAeasy spin columns. All subsequent steps were performed at room temperate. The spin columns were centrifuged at 8000 x g at for 15 seconds and flow-through was cautiously discarded. To wash spin columns, 700µl of

the RW1 solution was added to the columns and then centrifuged at 8000 x g for 15 seconds. The flow-through was decanted.

After the washing step, the samples were DNAse treated with the Qiagen RNase-Free DNase Set kit (Qiagen, 79254) to remove DNA contamination. To prepare the DNase I stock solution, 550µl RNase-free water was pipetted to the DNase I vial. 10µl of the previously prepared DNase I stock solution was added to 70µl RDD buffer to prepare a total of 80µl incubation mix. The DNase treatment mix was then added to the RNeasy spin column and incubated for 20 minutes. Subsequently, 350µl of the RW1 stringent washing solution was added to the spin columns and centrifuged at 8000 x g for 15 seconds, flow-through was discarded. To wash membrane-bound RNA, 500µl of the RPE mild washing buffer was added to the spin column and centrifuged, and the flow-through was discarded for the second time. The columns were centrifuged again for 60 seconds at 8000 x g to remove excess liquid impurities. To elute RNA, 30-50µl RNase-free water was added to the spin columns (placed into a clean RNase-free microcentrifuge) and were incubated for 3 minutes at room temperature to increase RNA yield. Finally, the columns were centrifuged at ≥8000 x g for 3 minutes and eluted RNA was then stored at -80°C.

2.2.4 Quantification and quality assessment of DNA/RNA nucleic acids using Nanodrop

The nucleic acid concentration of extracted RNA and DNA was measured using Nanodrop 1000 Spectrophotometer (ND-1000, NanoDrop Technology, USA). The nanodrop's wavelengths were set at 260nm and 280nm were equilibrated at zero absorbance with a blank reference using DNA/RNA eluted solution (e.g., TE buffer or water). This instrument passes a 260 nm ultraviolet light though the loaded samples

and measures the concentration based on the amount of light that passes through the sample. 2µl of the nucleic acid was loaded onto the spectrophotometer; thereafter, concentration and wavelength absorbance were measured. A nucleic acid is assumed to be pure if the 260/280 measured ratio is ≥2.0 for RNA; ≥1.8 for DNA.

2.3 WES of CHT and three bone cancer projects

2.3.1 Sample preparation for WES

A total of 2-5 µg of genomic DNA of CHT (n=12), UPSb (9 normal-paired samples; 3 tumours-only), and adamantinoma (6 normal-paired samples; 2 tumours-only) and OFD-like adamantinoma (2 normal-paired; 1 tumour-only) samples were sent for WES to Oxford Gene Technology (OGT) Company (Oxfordshire, UK). In addition, two UPSb (1 normal-paired; 1 tumour-only) and one OFD-like adamantinoma (normal-paired) were exome-sequenced at Prof Eamonn Maher's laboratory (Department of Medical Genetics, the Stratified Medicine Core Laboratory, University of Cambridge). As stated earlier, all DNA from CHT samples were extracted from peripheral blood. All DNA from UPSb and adamantinoma samples were extracted from fresh frozen tissues, except for one FFPE UPSb specimen.

2.3.2 Exon Capture: selection of protein-coding regions

For samples sequenced at OGT, the exons were captured using the Agilent SureSelect Human All Exon V5 kit following manufacturer's instructions (Agilent, Santa Clara, USA) (for details see Figure 1–14). Briefly, DNA was sheared by fragmentation (Covaris, Woburn, USA), purified using Agencourt AMPure XP beads (Beckman

Coulter, USA), followed by repairing and ligation of adaptors to fragments. The purified libraries were then hybridised to biotinylated RNA baits. Magnetic Dynabeads (Invitrogen, USA) were used to select bound genomic DNA and unbound fragments were discarded. The hybridised DNA fragments were amplified before sequencing. The captured exons were then massively sequenced on either the Illumina HiSeq2000 sequencer to generate 100 basepair paired-end reads, or using NextSeq to generate 150 basepair paired-end reads to achieving a minimum of 50X average coverage,

For the samples sequenced at Prof Maher's laboratory, the exon capturing was performed using Illumina Nextera Rapid Capture Exome kit, and then sequenced on the Illumina HiSeq 4000, generating 150 basepair paired-end reads.

2.3.3 Bioinformatic tools used in WES data analyses

2.3.3.1 SIFT and PolyPhen2 missense prediction tools

Sorting Tolerant From Intolerant (SIFT) (Kumar et al., 2009b) and Polymorphism Phenotyping v2 (PolyPhen-2) and (Adzhubei et al., 2013) tools were used to predict the deleterious impact of missense variants on the protein function and/or structure. These tools were used to determine the potential biological effect (e.g., benign or deleterious) of the missense changes identified in all WES projects presented in this thesis (Chapter 3, 5 and 6).

For missense variants, SIFT calculates a prediction score based on the evolutionary conservation of the amino acid substituted within protein families, for which altering a highly conserved amino acid is predicted to be damaging. A score ≥0.05 is considered tolerated whereas <0.05 deleterious.

Unlike SIFT, PolyPhen-2 follows a different prediction approach by assessing eight sequence-based and three structure-based predictive criteria of the altered amino acid versus the wild-type amino acid. For example, one of these criteria assesses whether the substituted amino acid occurs at a site of biological importance (e.g., binding site region). After evaluating the 11 factors, PolyPhen-2 provides prediction scores of: (1) 0.0–0.15, predicted to be benign; (2) 0.15–0.85, possibly damaging; and (3) 0.85–1, probably damaging.

2.3.3.2 IGV tool for visualising NGS data

Integrative Genomics Viewer (IGV) is a powerful tool that can be used to visualise and explore large genomic datasets, such as WES and RNA-Seq data (Robinson et al., 2011b). BAM/BAM.BAI files are the required file input for the IGV. The IGV tool was used to visualise reads mapped to exons/exonic boundaries to assess the WES coverage of genes of interest. In addition, IGV was utilised to visualise WES variants identified in CHT (Chapter 3), UPSb (Chapter 5) and adamantinoma and OFD-like adamantinoma tumours (Chapter 6), as well as gene fusions identified in bone tumour projects using RNA-Seq data.

2.3.4 Raw data analysis pipeline of WES: alignment and PCR duplicates

Generated sequencing data require a series of processing steps. The sequenced reads were mapped to the GRCh37/hg19 human genome reference assembly using the Burrows-Wheeler Aligner tool (BWA 0.6.2) (Li and Durbin, 2009). Genome analysis and realignment were performed using Toolkit (GATK v1.6, Broad Institute). PCR

duplicates were removed using the Picard tool v1.107 (http://picard.sourceforge.net). The raw data were provided in BAM and BAM.BAI files (BAM index file).

2.3.5 Variant calling and annotation in CHT and bone cancer projects

2.3.5.1 CHT project

WES variants were called and annotated following the OGT's exome pipeline. Briefly, INDELs and SNVs were identified using the GATK Unified Genotype (Broad Institute). Variants with Phred score (sequencing quality score) of less than <20 were discarded to reduce false positive calls.

To perform the familial analysis (trio analysis) in the CHT families (Chapter 3), the GATK tool was used to generate multisample variant call format (VCF) files to identify recessively inherited variants. These VCF files identify recessively inherited variants if they are homozygous in affected individuals and (1) heterozygous in the unaffected parents; or (2) heterozygous (but not homozygous) in the unaffected family member (e.g., an unaffected parent or sibling). All variants were annotated using the Ensembl Variant Effect Predictor (McLaren et al., 2010). Population allele frequencies were obtained from dbSNP release 137, ExAC and or Ensembl population frequency data (see section 2.6.6).

2.3.5.2 VarScan2 and MuTect tools for somatic variant calling in cancer projects

VarScan2 (Koboldt et al., 2012) and MuTect (version1) (Cibulskis et al., 2013) are computational algorithmic tools used to identify somatic changes using WES data of tumour and corresponding normal samples (details in Appendix Section 8.2 & 8.3).

VarScan2 detects small INDELs and SNVs; by contrast, MuTect version 1 is only able to identify SNVs. Although MuTect version 1 is unable to detect INDELs, MuTect enhances the detection of low-allelic SNVs and, therefore, can lead to better mutational landscape analysis (Cibulskis et al., 2013). In cancer samples, genetic alterations can be present at low mutation frequency and, therefore, their detection can be challenging. That is, these changes can occur in a subpopulation of the tumour cells or if the tumour tissue is contaminated with normal cells (Cibulskis et al., 2013). To maximize the sensitivity of WES data analysis, both of these variant calling tools were utilised to detect somatic INDELs and SNVs in UPSb, adamantinoma and OFD-like adamantinoma tumours.

2.3.6 Filtering WES data to identify somatic candidate variants in the bone cancer projects

2.3.6.1 Overview and default variant allele frequency cut-off

WES requires a multi-step filtering scheme to identify somatic candidate variants. These filtering steps were applied to UPSb, adamantinoma and OFD-like adamantinoma tumours, unless stated otherwise.

The somatic variant analysis for normal-matched tumour pairs was performed using corresponding normal samples; by contrast, tumours without exome-sequenced corresponding normal samples were paired with a non-corresponding normal to perform the variant calling analysis. Variants were annotated using the Ensembl Variant Effect Predictor (McLaren et al., 2010). The default 10% variant allele frequency (VAF) cut-off for somatic calling in both VarScan2 and MuTect (Kang et al., 2017; Yadav et al., 2016) was applied to all bone tumour samples, except for one FFPE UPSb tumour in which the %VAF was adjusted to ≥20% to reduce false positive

artefacts. VAF is the proportion of reads at a particular site which contains/supports the variant allele. This increased number of false positive calls is expected in FFPE samples due to the formalin fixation process and degradation of nucleic acid material (Van Allen et al., 2014; Spencer et al., 2013).

2.3.6.2 Filtering and visualisation of variants in normal-paired tumours

WES technique focuses on coding regions; however, some variants in introns or regulatory regions (e.g., 5' or 3' untranslated regions [UTRs]) can be identified by off-targeted capturing during WES library preparation (Agnihotri et al., 2016). Although UTR variants may be involved in regulation or gene expression, these variants do not directly alter the translated protein sequence and, therefore, were excluded. Likewise, synonymous variants are silent changes and were also discarded. By contrast, nonsynonymous, nonsense, splice site and INDEL changes were prioritised as they directly alter the translated protein. Nonsense variants introduce a premature stop codon and, therefore, result in a truncated protein product. Splice site variants alter the nucleotides at the splice consensus sequence which can subsequently change the exon-intron splicing during the processing of mRNA (e.g., exon skipping or retention). Splice site changes that are within two base pairs of intron/exon boundary were prioritized (Agnihotri et al., 2016; Makinen et al., 2016).

SNVs and INDELs that are common in the general population are likely to be germline, and, therefore were discarded. Variants with minor allele frequency (MAF) >0.1% in ExAc, dbSNP (Build 137) or 1000 Genome project were excluded (Agnihotri et al., 2016; Dai et al., 2016; Makinen et al., 2016).

As mentioned earlier, SIFT and PolyPhen2 *in silico* prediction tools were used to bioinformatically assess the impact of missense alterations (see Section 2.3.3.1) Missense changes predicted to be deleterious by at least one tool were retained;

otherwise, changes classified as benign by both tools were discarded (Agnihotri et al., 2016; Makinen et al., 2016). Called variants were individually visualised by IGV tool (see Section 2.3.3.2) to identify genuine calls and discard potential false positive or germline calls. Variants that were present in any other exome-sequenced normal samples with ≥5% VAF were discarded as they are likely false positive artefacts or germline variants (Singh et al., 2014; Stachler et al., 2015) (example in Appendix Figure 8–8).

2.3.6.3 Assessment of the established filtering criteria in UPSb, adamantinoma and OFD-like adamantinoma tumours using Sanger sequencing

The high throughput nature of WES can identify false positive artefacts. Hence, it was essential to test the reliability of the above-mentioned filtering scheme at the laboratory level using Sanger sequencing. In ten UPSb, eight adamantinoma and two OFD-like adamantinoma tumours (all paired with corresponding normal samples), a total of 133 variants were tested by Sanger sequenced to establish a true positive variant criterion. A total of 21 variants with 10–20% VAF for which each variant was supported by two reads (i.e., two independent sequencing reads showing the variant) were not detected in Sanger sequencing. Consequently, the true positive threshold of called variants was adjusted and required a minimum of three reads supporting the variant for >20% VAF variants; five supporting reads for variants with 10-20% VAF (Eckstein et al., 2016; Martelotto et al., 2015). Following these criteria, the true positive Sanger sequenced variants rate was 98.4%. That is, 98.4% of the tested variants were confirmed at Sanger sequencing level (examples in Appendix Table 8–8). To test and confirm the somatic status of WES variants, 65 variants were Sanger sequenced using tumours

and corresponding normal samples. All tested variants were present in tumours but were absent in corresponding normal tissue, achieving a 100% true somatic rate.

2.3.6.4 Additional rigorous filtering steps applied to normal-unpaired UPSb tumours

Four UPSb, two adamantinoma and one OFD-like adamantinoma tumours were exome-sequenced without their corresponding normal samples (tumour samples only). However, DNA from corresponding normal tissues was only available for the four UPSb tumours. In the four normal-unpaired UPSb tumours, 36 randomly selected variants were confirmed germline by Sanger sequencing (i.e., variant present in tumours and corresponding normal) (examples in Appendix Table 8–8). Therefore, additional filtering steps were applied to these tumours to identify genuine somatic calls. SNV and INDEL changes that are reported in ExAc or dbSNP population frequency databases, regardless of MAF, were excluded, except for variants that are reported somatic in the COSMIC database and have <0.01% MAF in population frequency databases (Garofalo et al., 2016; Jiang et al., 2016; Kumar et al., 2011). In addition, changes with >50% VAF were eliminated to rule out the possibility of private germline events as the majority of somatic alterations are likely to be present at <50% VAF (Shain et al., 2015). To test the newly adjusted filtering criteria for normal-unpaired UPSb tumours, 38 variants were Sanger sequenced in both tumours and corresponding normal samples achieving 94.7% true positive somatic rate (examples in Appendix Table 8–8).

2.4 RNA-Seq studies on bone tumours

2.4.1 Overview: RNA samples and sample quality check using Bioanalyzer

A total of 16 RNA samples (eight UPSb, five adamantinoma and three OFD-like adamantinoma tumours) were RNA-Sequenced by the Genomics Birmingham facility (Institute of Cancer & Genomic Sciences, University of Birmingham). A total of 100ng of extracted and DNase-treated RNA (Section 2.2.3) was supplied and the sample quality and concentration were checked by Tapestation (Agilent, USA) and Qubit (Thermo Scientific, UK).

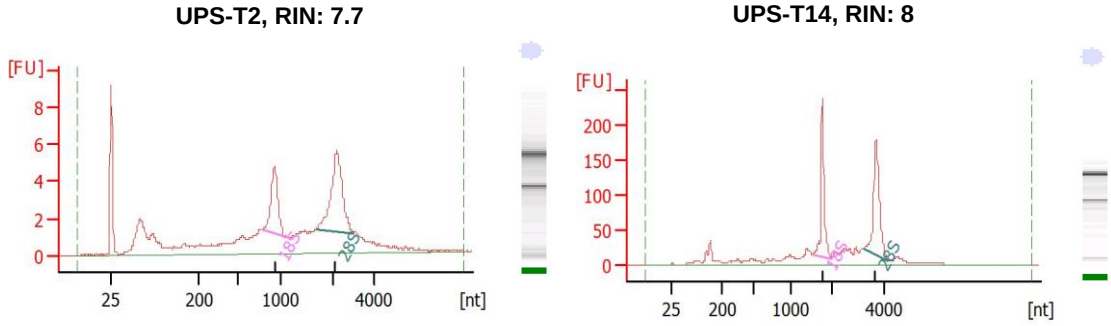
The quality of the RNA sample is an important factor for successful RNA-Seq experiment. The quality of the tumour RNA was assessed by the RNA Integrity Number (RIN) which was measured using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) (Courtesy of Functional Genomics Laboratory (University of Birmingham, UK). 2µl of the samples were separately loaded onto an RNA 6000 Nano Chip I (Agilent, 5067-1511). Minimally degraded RNA samples are expected to show RIN values of ≥7 (maximum value is 10). The majority of UPSb, adamantinoma and OFD-like adamantinoma samples had RIN measurement of >7 (range: 5.4-8.1); (examples in Figure 2–1). The lowest RIN value in these tumours was 5.4. Due to the unavailability of additional tissue material, the samples with lower RIN values (below 7) were included in the studies.

2.4.2 RNA-Seq library preparation and sequencing

For each sample, RNA-Seq libraries were generated using the Neoprep stranded mRNA library prep (Illumina, NP-202-1001), following the manufacturer's protocol. The RNA-Seq library preparation involves converting total/fractional RNA into cDNA

libraries in a series of steps (see Figure 1–20). In short, oligo-dT coated magnetic beads were used to capture mRNA, followed by mRNA fragmentation and random addition of primers. Random primers were used to synthesize the first strand cDNA and, subsequently, amplification of the second strand. cDNA amplified fragments were end-repaired, followed by adaptor ligation and PCR amplification prior to massive parallel sequencing.

Prepared libraries were sequenced on NextSeq 500/550 High Output (Illumina) to produce a minimum of 50 million paired-end reads per sample (75bp in length). Generated data were uploaded and analysed using BaseSpace Sequence Hub (details in Section 2.4.3.1)



Overall Results for UPS-T2 sample:	Overall Results for sample UPSb-T14:
RNA Area: 91.1	RNA Area: 1,398.8
RNA Concentration: 134 ng/µl	RNA Concentration: 1,694 ng/µl
rRNA Ratio [28s / 18s]: 1.7	rRNA Ratio [28s / 18s]: 1.1
RNA Integrity Number (RIN): 7.7 (B.02.08)	RNA Integrity Number (RIN): 8 (B.02.08)
Result Flagging Color:	Result Flagging Color:
Result Flagging Label: RIN: 7.70	Result Flagging Label: RIN: 8

Figure 2–1: Examples of RIN value measurement in UPS-T2 and UPS-T14 samples using Agilent 2100 Bioanalyzer and Agilent Nano chips. The Agilent 2100 Bioanalyzer measures the ratio of 18S (red arrow) and 28S (blue arrow) ribosomal RNA to identify RNA degradation. A degraded RNA shows a decrease in the 18S and 28S peaks with an increase in the baseline signal between the two peaks (not present here). Electrophoresed gel showing shorter fragment sizes can also be indicative of RNA degradation (not present here; two black prominent bands are observed). Graphs courtesy of the Functional Genomics Laboratory, University of Birmingham.

2.4.3 Bioinformatic tools used in RNA-Seq data analyses

2.4.3.1 BaseSpace Sequence Hub and TopHat2 alignment tool for detection of gene fusions

The BaseSpace online platform was used to analyse the RNA-Seq data of bone cancer projects (https://basespace.illumina.com/home/index). BaseSpace is a user-friendly tool that can be utilised by researchers to simplify bioinformatic data analyses and store NGS data. Raw sequence data were first uploaded to the storage platform, followed by choosing the desired analysis. The hub was used to generate FastQ (text-based format of nucleotide sequence) and BAM/BAM.BAI files, and perform subsequent data analyses (additional details in Appendix Section 8.4).

2.4.3.1.1 The Tuxedo Suite by Illumina: TopHat2 Alignment

Raw RNA-Seq data require a series of processing steps. The Tuxedo suite, provided by BaseSpace, offers various bioinformatics tools for RNA-Seq data analyses. The TopHat2 Alignment package was used to analyse RNA-Seq data of UPSb (Chapter 5), adamantinoma and OFD-like adamantinoma (Chapter 6) and identify gene fusions The TopHat2 Alignment (version 2.1.1) is a splice junction mapper that aligns and analyses RNA-Seq data to detect splice junctions between exons (Trapnell et al., 2012). Frist, Bowtie aligner tool, a part of TopHat2, maps the RNA-Seq reads to the hg19 human reference genome assembly. Aligned reads are then analysed by the TopHat2-Fusion algorithm (Kim et al., 2013) to identify gene fusions (additional details in Appendix Section 8.4, Figure 8–6).

In addition to TopHat2-Fusion, STAR-Fusion (Haas et al., 2017) was the second gene fusion caller tool used in these tumours (Courtesy of Yun Shao Sung, MS and Cristina

R. Antonescu, MD, Memorial Sloan-Kettering Cancer Center, New York) (See 2.4.3). Both tools maps RNA-Seq reads to the human genome reference (hg19 assembly) and identifies discordantly mapping reads indicative of possible genomic rearrangements (details about the fusion calling tool in Appendix Figure 8–7).

2.4.4 Data analysis of RNA-Seq data

2.4.4.1 Filtering and prioritisation of non-recurrent candidate gene fusions

A total of 109 gene fusions were identified in UPSb and 42 non-recurrent fusions in both adamantinoma and OFD-like adamantinoma tumours. This relatively large number of candidate gene fusions required filtering steps to prioritize genuine candidates rather than sequencing artefacts. First, candidate gene fusions identified in intragenic and intronic regions were excluded as they are likely to be DNA contaminants or unspliced mRNA precursors rather than authentic gene fusions with exonic fusion junctions (Lee et al., 2015; Veeraraghavan et al., 2014). Thereafter, gene fusions with breakpoints within exonic or intronic/exonic boundaries were prioritised (Peters et al., 2015). To establish a minimum threshold for junction/supporting fusions reads, four fusions with 1-3 junction/supporting reads were tested in the laboratory using reverse transcription PCR (RT-PCR) (more in Section 2.5.3.1). Fusion junction/supporting reads are reads that support the gene fusion and map across the fusion breakpoint on both sides of the partner genes. All four fusions were not amplified by RT-PCR using two different sets of primers and were therefore classified as false positive calls. Subsequently, the established cut-off required ≥3 junction/supporting fusion reads for candidate gene fusions to be considered as true positive. Gene fusions were visualised using IGV tool. To enrich for potentially real gene fusions, candidate

gene fusions with even distribution and uniform depth of mapped fusion junction/supporting reads were prioritised (Peters et al., 2015; Kim and Salzberg, 2011). Candidate fusions with junction/supporting fusion reads of equal length were excluded, as these reads are likely to be amplification artefacts during RNA-Seq sample preparation(Delespaul et al., 2017; Pflueger et al., 2011).

2.5 Laboratory validation of WES data

2.5.1 Standard PCR for validation of variants: primer design criteria

Standard PCR amplification of desired genomic regions followed by Sanger sequencing (Section 2.5.2) were utilised to confirm variants identified by WES in CHT and bone cancer projects A set of two oligonucleotide primers was used to amplify a genomic region of interest (e.g., flanking a variant site). The primers were designed using the Primer3 software (http://primer3.ut.ee) The primer design criteria consisted of: (1) primers being 18–24 nucleotides in length with a melting temperature (Tm) between 53–62°C; and (2) a G-C content of 35–65%. Moreover, long strings of nucleotides (e.g., >5 A/T/G/C bases) were avoided to eliminate the possibility of hairpin loop formation. The amplified products aimed to be between 150–600 base pairs to ensure successful amplification. Designed primers were subsequently BLAST-searched against the human genome using the National Centre for Biotechnology Information (NCBI) BLAST programme (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to ensure high annealing specificity of primer. The designed primers used in the CHT WES project (Chapter 3) are in Table 8–13; UPSb (Chapter 5) in Table 8–14; adamantinoma and OFD-like adamantinoma (Chapter 6) in Table 8–15.

2.5.1.1 Standard PCR reaction components and optimization

Each PCR reaction consisted of 2.5µL 10x buffer (containing 1.5mM $MgCl_2$) (Roche,12158264001), 2.5µl (2.5 mM) dNTPs (Thermo Fisher Scientific, R0141), 1µl of each forward and reverse Primer (20 pmole), 0.1µl (0.5 unit) Fast Start Taq DNA*taq* polymerase (Roche,12158264001), 25–50 ng DNA template and distilled water to make a 25µl final volume reaction. When amplifying genomic regions with high G-C makeup, 5µl 5X GC rich solution (Roche, 12158264001) was added to optimize PCR reaction. To ensure proper experiment settings, both positive (non-degraded DNA) and negative (DNA-free distilled water) controls were included in each PCR experiment. The PCR amplification reactions were performed using a tetrad GeneAmp 9700 thermocycler.

Before amplifying patients' DNA, a set of optimization conditions were performed using commercially purchased DNA to avoid wasting limited patients' material. If needed, a multiple of thermal cycles with different annealing temperature or product extension time were performed to achieve optimal product amplification. Table 2–1 shows the typical steps of a standard PCR programme using a thermocycler.

Step	Temperature	Time	Purpose
1	95°C	5 minutes	Initial denaturation
2	95°C	30 seconds	Denaturation
3	Ta°C	30 seconds	Annealing
4	72°C	40-60 seconds	Extension
5	72°C	5 minutes	Final extension

Table 2–1: Steps of standard PCR reaction using a thermocycler. Ta refers to the theoretical annealing temperate of the primers. Steps 2-4 are repeated for 30–35 cycles.

2.5.1.2 Agarose gel electrophoresis: visualisation of PCR products

PCR amplified products were visualised using electrophoresis of agarose gels (Bioline, 41025) stained with ethidium bromide or Midori Green Advanced DNA Stain (Geneflow, Cat. No. MG 04). PCR products were run on 1-3% (w/v) agarose gels, where high percentage gel was used for visualising small PCR products (e.g., <300 bp). An example of a 2% agarose gel composition consists of 200ml of 1X Tris-Borate-EDTA (TBE) diluted from 10X TBE (National Diagnostics, EC-860) and four grams of agarose powder. Agarose gel mixture was heated until agarose powder was completely dissolved, then the mixture was left to cool down (but not solidify) for 10–15 minutes. Lastly, 2µl 10mg/ml of ethidium bromide or 4–6µl of Midori Green was added to the mixture before pouring into a gel casting tray.

Before loading on the gel, PCR products were mixed with 5% (v/v) loading dye (48.5% dH20, 50% glycerol, 0.5% Orange G and 1% EDTA). PCR products were electrophoresed at 100V for approximately 30–45 minutes, and subsequently visualised using a UV transilluminator.

2.5.2 Sanger sequencing of PCR amplified products

2.5.2.1 Preparation and clean-up of PCR products using microCLEAN

Prior to Sanger sequencing, PCR products require removal of the PCR reaction impurities (e.g., unincorporated dNTPs, excess primers and enzyme). In a 96 well plate, 2.5 µl PCR products are loaded with an equivalent volume of microCLEAN (Microzone, 2MCL-10) and centrifuged at 4000 rpm for 40–45 minutes at 4°C. The plate was then flipped (without tapping) on a piece of tissue paper and centrifuged at

500 RPM (4°C) for 30 seconds to remove supernatant. The plate was left to air dry for 3-5 minutes prior to Sanger sequencing reaction.

2.5.2.1.1 Sanger Sequencing (chain termination sequencing) reaction (Part 1)

The Big Dye Terminator V3.1 (Applied Biosystems, Cat. No. 4336917) was the kit of choice for Sanger sequencing reactions. After preparation of the plate containing cleaned PCR products, each sequencing reaction consisted of 0.5µl of Big Dye, 2µl of 5X Big dye sequencing buffer, 20 pmol of either the forward or reverse primer and nuclease-free water to make a 10 µl final reaction volume. The plate was then loaded onto a thermocycler following sequencing reaction protocol described in Table 2–2.

Temperature and time	Number of cycles
94° C for 4 minutes	1 cycle
94° C for 25 seconds	30 cycles
50° C for 25 seconds	
60° C for 4 minutes	

Table 2–2: Sanger sequencing reaction protocol using a thermal cycler.

2.5.2.1.2 Ethanol precipitation of Sanger sequencing reactions (Part 2)

After the sequencing thermal cycling protocol, each reaction was precipitated using: 2µl of 0.125 EDTA (containing NaAC, Sigma-Aldrich, E7889) which supports washing

of unincorporated dyed bases, and 30µl of 100% ethanol, ensuring proper mixing by pipetting up and down. It is advisable to minimize the amount of light on the plate to preserve dye intensities. Therefore, the plate was covered with aluminium foil whenever possible. The plate was centrifuged at 2000 RPM for 25 minutes at 4°C and subsequently inverted (without tapping) onto a piece of paper towel. The plate was spun at 450 RPM for 30 seconds to remove supernatant. Next, each well was resuspended in 90 µl of 70% (v/v) ethanol (ensuring proper mixed), centrifuged at 2000 RPM for 15 minutes at 4°C, and lastly the supernatant was removed as mentioned previously. The plate was left to air dry for 5 minutes.

To rehydrate the pellets, 10µl HiDi Formamide solution (Applied Biosystems, 4311-329) was added to each well and allowed to stand for 5 minutes. The suspended pellets were thereafter denatured at 95°C for 5 minutes, snap chilled on ice for 5 minutes and finally loaded onto an ABI 3730 DNA analyser (Applied Biosystems).

2.5.2.2 Analysis and visualisation of Sanger sequencing data

The Mutation Surveyor DNA variant analysis software was used to analyse the sequencing data (Softgenetics.com, 2015). Multiple sequencing traces could be simultaneously analysed using this software. The software compared the sequencing traces to the provided reference sequence and, subsequently, highlighted bases that deviate from the reference (e.g., variants). Highlighted bases were inspected manually to confirm true variant calls from background noise or sequencing artefacts. Chromas (Softpedia, 2015, Windows PC) or 4Peaks (Nucleobytes.com, 2015, Mac) software were also used to view and manually analyse sequencing electropherograms.

2.5.3 Laboratory validation of RNA-Seq data

2.5.3.1 RT-PCR primer design and method

2.5.3.1.1 cDNA synthesis of bone cancer samples

cDNA was synthesized from RNA samples using the SensiFAST cDNA Synthesis Kit (Bioline, BIO-65053). Following the manufacturers' instruction, each cDNA reaction was prepared on ice and contained: 4µl 5x TransAmp buffer, 1µl reverse transcriptase enzyme, 1µg of total RNA and DNase/RNase free water to make a 20µl total volume reaction. The reaction components were gently mixed by pipetting and loaded on a thermal cycle using the programme described in Table 2–3.

Step	Temperature	Time	Purpose
1	25°C	10 minutes	Primer annealing
2	42°C	15 minutes	Reverse transcription
3	48°C	15 minutes	Additional RT step
4	85°C	5 minutes	Inactivation
5	4°C	Hold	Hold

Table 2–3: Thermocycler programme for cDNA synthesis from RNA template.

2.5.3.1.2 RT-PCR primer design and technique

RT-PCR was used to validate gene fusions identified by RNA-Seq in UPSb and adamantinoma and OFD-like adamantinoma tumours (Chapter 5 and 6). Primers flanking a gene fusion junction were designed, checked parametrically and blasted against human RNA reference assembly as explained earlier. Using 50–100 ng of

cDNA as a PCR template, the RT-PCR reaction makeup consisted of Fast Start DNA polymerase (Roche, 12158264001), 2.5µl 10X buffer, 2.5µl dNTPs (2.5 mM), 5µl 5X GC rich solution, 1µl (20 pmole) of each forward and reverse primers and 0.1µl (0.5u) and water to make a final volume of 25µl, including proper experimental controls. The RT-PCR product size aimed to be between 120-400bp to ensure a successful PCR amplification. PCR products were visualized using agarose gels (see Section 2.5.1.2). (cDNA primers used in gene fusion confirmation are listed in Table 8–16).

2.5.3.2 Long range PCR as a method for genomic validation of gene fusions

Long range PCR (LR-PCR) was used to validate gene fusions identified in the bone tumours projects at the DNA level (i.e., identify genomic breakpoints) (Chapter 5 and 6). Similar to standard PCR, the primers were designed using the Primer3 online tool (http://primer3.ut.ee/). Each primer was designed on each gene fusion partner and was 25–35bp in length (Ta: 58–64°C) with 35-55% GC content (list of LR-PCR primers in Table 8–17).

The genomic breakpoint analysis required designing multiple primers to map genomic breakpoint, a process known as genome walking. During genome walking process, the size of amplified products involuntarily ranged between 0.5–12 kilobase (depending on fusion genomic breakpoints). The high fidelity PrimeStar GXL DNA polymerase (TAKARA, R050Q) was used to amplify these large PCR products. Following the manufacturer's instructions, the LR-PCR reaction was prepared as follows: 10µl of 5X PrimeSTAR GXL buffer (1X final concentration), 200µM dNTPs, 0.2–0.3µM forward and reverse primers, 50 ng genomic DNA template, 1.25U PrimeSTAR GXL DNA polymerase and nuclease-free distilled water to make up 50µl reaction volume (thermocycler programme in Table 2–4).

Step	Temperature	Time	Purpose
1	98°C	10 seconds	Denaturation
2	Ta	15 seconds	Primer annealing
3	68°C	60 seconds per kb	Extension

Table 2–4: The 3-step thermocycling protocol for LR-PCR reactions. Steps 1–3 were repeated for 30 cycles. Ta represented the theoretical primer annealing temperature. For larger products (>8 kilobase) a two-step programme eliminating step 2 was used.

2.6 Online sources, databases and browsers used in genome and variant information

2.6.1 Ensembl Genome Browser

The Ensembl genome database browser was established in response to the completion of the human genome project. This browser hosts vertebrate (e.g., human) genomes and is publicly available at (http://www.ensembl.org). Using the hg19 human genome assembly reference, this comprehensive genome was utilised to obtain information regarding genes such as gene names, number/ID of validated/processed transcripts, translation ID's, number of exons, cDNA sequence of genomic translated regions, protein sequence, and biological domains and features of a transcript. Ensembl can also be used to obtain population genetic frequency of reported human polymorphisms (e.g., minor allele frequency).

2.6.2 UCSC Genome Browser by the University of California, Santa Cruz

The UCSC Genome Browser is a freely available human genome assembly browser hosted by the University of California, Santa Cruz (https://genome.ucsc.edu/). This

browser was used to obtain information about gene structure and sequence data as well as to provide a graphical representation of the genome.

2.6.3 GeneCards and the NCBI website

The GeneCards database is openly available to users (www.genecards.org/). This website provides comprehensive information about gene annotation, genetic and clinical information, proteomic, transcriptomic, RNA and protein expression data, protein localization and drug-targeted genes.

The NCBI website also provides gene annotation and sequence information of curated genes (http://www.ncbi.nlm.nih.gov). It is also an excellent source of published scientific papers.

2.6.4 COSMIC and inTOgen databases

The COSMIC database is a comprehensive repository of somatic alterations and their biological significance in human cancer (http://cancer.sanger.ac.uk). This database was utilised to assess the somatic status of the variants identified by WES in bone cancer projects (Chapter 5 and 6). It was also used to identify previously reported gene fusions, cancer drug-targeted genes, and cancer driving genes and the COSMIC CCGC list.

The inTOgen database (https://www.intogen.org/search, release 2014.12) uses multiple data resources to identify cancer driver genes. These resources include the International Cancer Genome Consortium, the TCGA and independent cancer research projects. Genes are classified as cancer driver genes if: (1) Genes show a high functional mutation in OncodriveFM tool and show statically significant cancer

mutations using MutSigCV; and (2) Gene are categorised depending on the mode of action as gain or loss of function using OncodriveROLE.

2.6.5 The Drug Gene Interaction database to identify drug targeted genes

The Drug Gene Interaction Database (DGIdb) (v3.0) was used to find genes that can be potentially drug targeted in UPSb, adamantinoma and OFD-like adamantinoma tumours. In a user-friendly manner, DGIdb provides gene 'druggability' information and drug-gene interactions by grouping and normalizing information from papers, databases and web resources (Cotto et al., 2018). A detailed expert-curated summary about the input gene list can be obtained by the user to identify potential drug-gene interactions.

2.6.6 ExAc browser

The ExAC browser contains the sequencing data of 60,706 unrelated individuals that participated in various genetic studies (http://exac.broadinstitute.org). With such a large number of individuals sequenced, this resource was used to obtain ethnic-specific (Caucasian or South Asian) population frequencies (e.g., MAF) for WES variants identified in the CHT and bone cancers projects.

2.6.7 Ingenuity Pathway Analysis

QIAGEN Ingenuity Pathway Analysis (IPA) software (http://www.ingenuity.com/products/ipa) was utilised in Chapter 6 to identify known biological pathways/networks that were significantly altered in adamantinoma and

OFD-like adamantinoma samples. It was also used to group the recurrent genes identified in UPSb tumours (Chapter 5) into functional pathways. IPA also provided a comprehensive data analysis of function and the interaction partners of genes and whether these genes are associated with known diseases.

2.7 Molecular gene cloning

Gene cloning techniques were utilised to produce copious amount of a defined DNA region (gene open reading frame [ORF]). Gene cloning was used to perform the functional experiments needed for the assessment of the candidacy of the *SIX2* homozygous variant in CHT (Chapter 4). The molecular cloning steps are summarized in Figure 2–2 and will be explained in details.

2.7.1 Obtaining of *SIX2* IMAGE and extraction and purifying plasmid

Purified and sequence-verified *SIX2* gene (*Homo sapiens*) IMAGE (The Integrated Molecular Analysis of Genomes and their Expression, ID: 3027992) clone was purchased from Source BioScience Company (Nottingham, UK). The clone was hosted in *E. coli* cells and grown on a nutrient agar slope containing chloramphenicol (CR) antibiotic selection.

To extract plasmid DNA, the *SIX2* IMAGE clone was grown on Luria Bertani (LB)-agar plate containing CR antibiotic (Section 2.8.2.3). A sterile inoculation loop, lightly immersed in agar slope, was streaked over the surface of an agar plate, followed by incubating overnight at 37°C. Two single colonies were aseptically selected, individually grown in 5ml LB broth (with CR) and incubated >16 hours at 37°C on a shaker (250 RPM). On the following day, the cells were pelleted by centrifuging the LB

broth at 4,000 RPM for 5 minutes, followed by pouring off the supernatant. The Qiagen Spin Maxiprep Kit (see Section 2.7.5) was used to purify and extract plasmids. Although the *SIX2* ORF is sequence-verified, the ORF sequence was checked again by Sanger sequencing.

2.7.2 Site-directed mutagenesis to introduce *SIX2* candidate variant

2.7.2.1 Site-directed mutagenesis kit and mutagenic primer design

Q5 Site-Directed Mutagenesis (SDM) Kit (NEB, E0554S) was used to introduce the *SIX2* homozygous candidate variant (c.859C>T; p.Val287Met) into the *SIX2* wild-type ORF. The NEBase changer tool (http://nebasechanger.neb.com) was used to design the mutagenic primer sequences. To design primers, the entire *SIX2* ORF sequence was inserted into the tool. The base pair to be substituted (e.g., Guanine) was selected, followed by selection of desired nucleotide (i.e., candidate variant, Adenine) A base pair substitution was generated by designing a mismatch nucleotide in the centre of the mutagenic primer (SDM_Val-Met_F). The 5’ end of the other primer (SDM_Val-Met_R) is designed to anneal at the nucleotide adjacent to the 5’ end of the mutagenic primer on the reverse strand (and complementary to the ORF sequence) (Table 2–5).

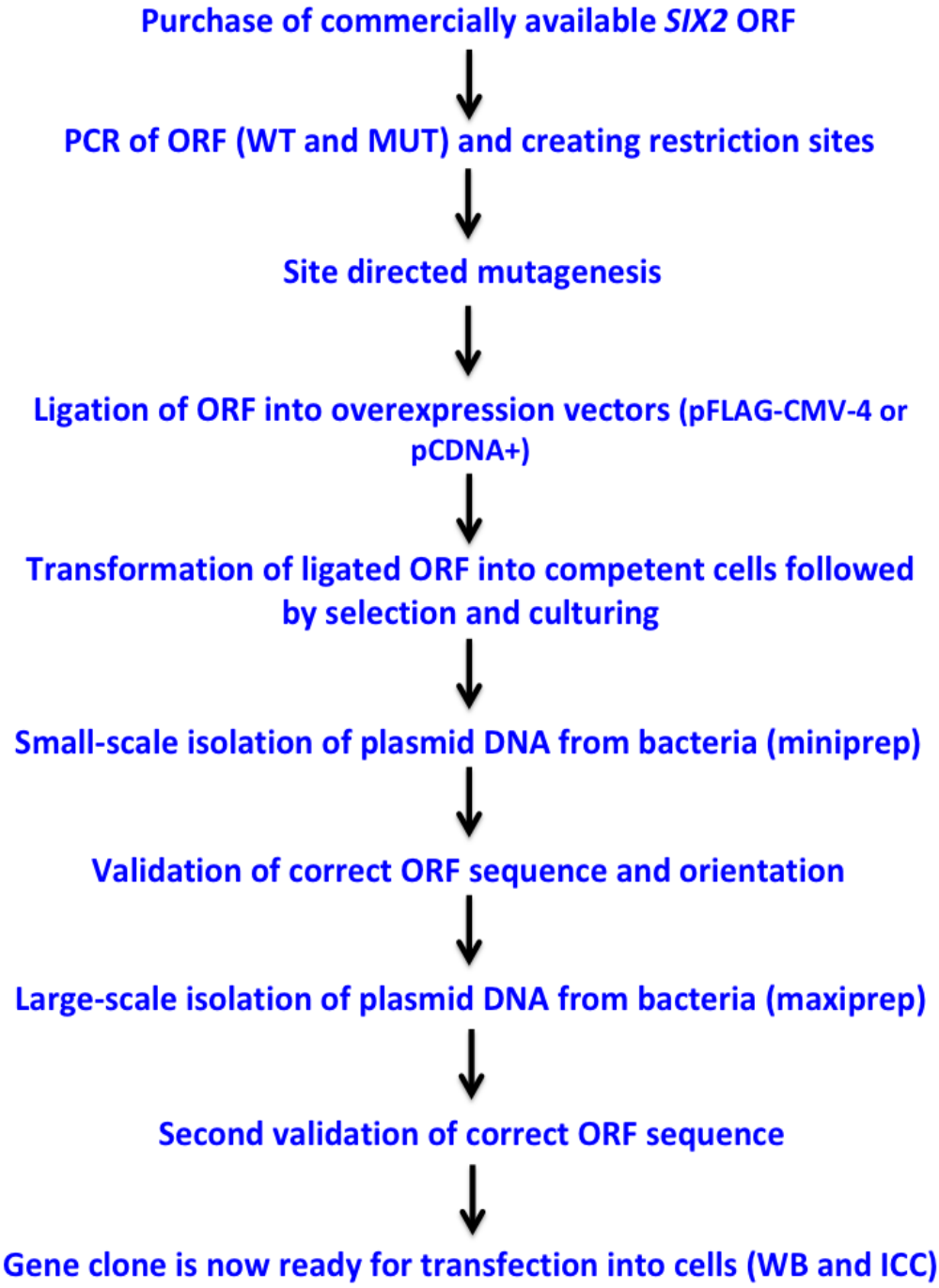


Figure 2–2: Molecular gene cloning procedure. The ORF sequence was commercially purchased. Site-directed mutagenesis was used to introduce desired variation (e.g., *SIX2* variant) to the wild ORF, followed by cutting and ligation of the altered ORF to the plasmid vector. The new cloned vector was then transformed to competent bacteria to allow multiplication, followed by the selection of successfully transformed clones. Systematic checks were performed to confirm correct sequence and orientation of the insert. ORFs either containing the **WT**: Wild-type sequence; or, the homozygous *SIX2* candidate variant, **Mut**: Mutant. **WB**: Western blotting. **ICC**: immunocytochemistry.

Name (F/R)	Oligo	L	%GC	Ta	Tr
SDM_Val-Met_F	5’-AGCCAACCTCaTGGACCTGGG-3’	21	62	66 °C	64.5°C
SDM_Val-Met_R	5’-GACATGGGGTTGAGGATGG-3’	19	58	63 °C	

Table 2–5: Mutagenic primers used to introduce *SIX2* candidate variant during SDM. The SDM primers are designed with 5´ends annealing back-to-back. The lower-case letter in the SDM_Val-Met_F mutagenic primer denotes to the nucleotide to be changed. **L:** Length, **Ta:** primer melting temperature, **Tr:** recommended primers annealing temperature (average of the Ta of the two primers), **%GC:** percentage of G-C base nucleotides. F: forward, R: reverse.

2.7.2.2 SDM exponential PCR amplification

The SDM PCR reaction was assembled from the following components: 12.5µl Q5 Hot Start High-Fidelity 2X Master Mix (1X final concentration), 10µM of forward and reverse primers, 25 ng (1–25 ng/µl) of purified *SIX2* plasmid DNA and finally nuclease-free water to make a 25µl final volume reaction. The reaction components were thoroughly mixed and transferred to a thermal cycling machine (conditions in Table 2–6).

Step	Temp (°C)	Time (seconds)	
Initial denaturation	98	30	
Denaturation	98	10	30 cycles
Primer annealing	66	30	
Extension	72	30	
Final extension	72	120	
Hold	10	∞	

Table 2–6: SDM PCR amplification programme.

2.7.2.3 Kinase, Ligase & DpnI treatment and transformation

Following the SDM-PCR reaction, the amplified linear plasmid was ligated and circularised in a clean PCR tube by assembling the following reagents: 1µl PCR product, 5µl 2X Kinase, Ligase & DpnI (KLD) reaction buffer (1X final concentration), 1µl KLD enzyme mix (1X), and nuclease-free water to make a 10µl final volume reaction. The reagents were mixed by pipetting and incubated for 5 minutes at room temperature.

The circularised plasmids were transformed to NEB 5-alpha Competent *E. coli* cells. A tube of competent cells was thawed on ice, followed by adding 5µl of the KLD ligated mixture prepared earlier. The tube was gently mixed by flicking and incubated on ice for 30 minutes. Afterwards, the cells were heat shocked at 42°C for 30 seconds, followed by incubation on ice for 5 minutes. Super Optimal broth with Catabolite repression (SOC) medium was pre-warmed to room temperature. 950µl SOC was added into the mix and incubated for one hour at 37°C with shaking (250 RPM). Afterwards, cells were mixed and 10–100µl of the mixture was aseptically added to the multiple CR selection plates. The plates were incubated overnight at 37°C.

2.7.2.4 Verification of SDM reaction

To check if the *SIX2* candidate variant was successfully incorporated into the ORF, ten colonies were selected and individually grown in 5ml: LB media with cm antibiotics. Cultured colonies were incubated overnight at 37°C. Plasmid DNA was extracted following the small-scale DNA extraction method (section 2.7.4). The ORFs were Sanger sequenced to confirm the successful incorporation of the *SIX2* homozygous variant.

2.7.3 PCR amplification of *SIX2* Wild-type and *SIX2*-Mutant ORFs using primers with restriction sites and high-fidelity polymerase

Specific gene cloning primers were designed to amplify the desired DNA regions (*SIX2*-WT and *SIX2*-Mut ORFs) and add restriction recognition sequences (Table 2–7). The recognition sequences were added to the ORF to serve as restriction enzyme recognition sites during the digestion and ligation of the ORF and overexpression vectors (more in Section 2.7.3.2). The restriction sites were added to the 5'-end of both forward and reverse primers. Moreover, a polylinker sequence was included upstream of restriction sites to increase the efficiency of binding and cutting of the restriction enzyme.

The Q5 High-Fidelity DNA Polymerase (NEB, M0491) was used to amplify the *SIX*-WT and *SIX*-Mut ORFs. The amplification fidelity of Q5 polymerase is approximately 280 times higher than *Taq* polymerase which, therefore, reduces the chance of introducing error variants by during PCR amplification. The PCR reaction was prepared as following: 5µl 5X Q5 Reaction Buffer, 0.5µl 10 mM dNTPs, 10µM forward and reverse primer, 30ng plasmid DNA template, 0.25µl and nuclease-free water to make a 25µl final volume. The thermocycling condition used is similar to a standard PCR programme, except for primer annealing temperature of 72°C.

Two ORFs were amplified using specific: (1) *SIX2* wild-type ORF (*SIX2*-WT), and (2) *SIX2* mutant containing the *SIX2* variant (*SIX2*-Mut) from SDM. Each ORF was ligated into two different overexpression vectors which were pFLAG-CMV-4 (tagged) cm (Figure 2–3) (Sigma-Aldrich, E7158), and pcDNA3.1+ (CMV-4 promoter; untagged) (Figure 2–4) (Invitrogen, V79020). A total of four vectors were prepared for the functional experiments (Chapter 4): (1) pFLAG-CMV-4/*SIX2*-WT, (2) pFLAG-CMV-4/*SIX2*-Mut, (3) pCDNA3.1+/*SIX2*-WT; and (4) pCDNA3.1+/*SIX2*-WT.

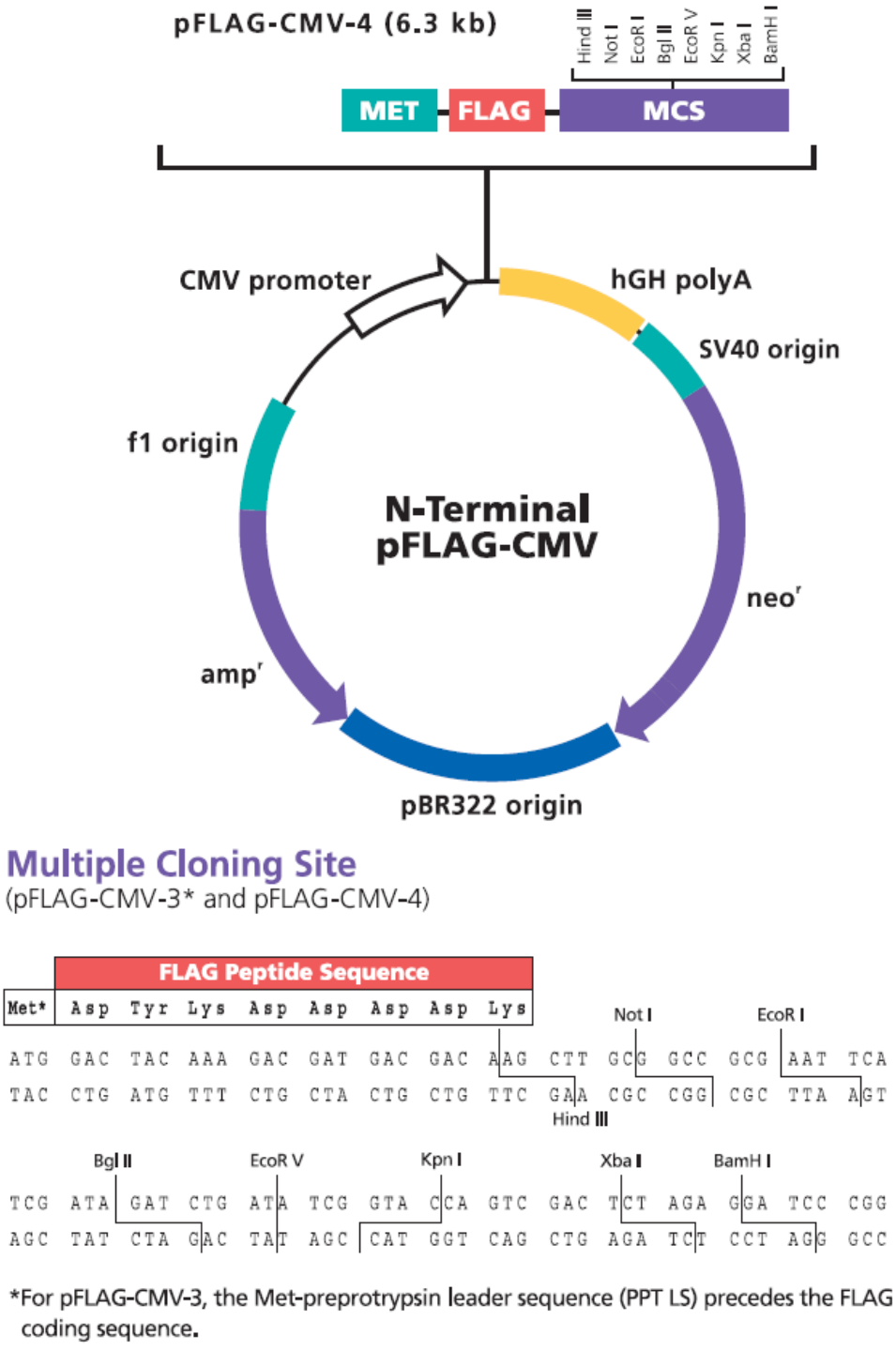
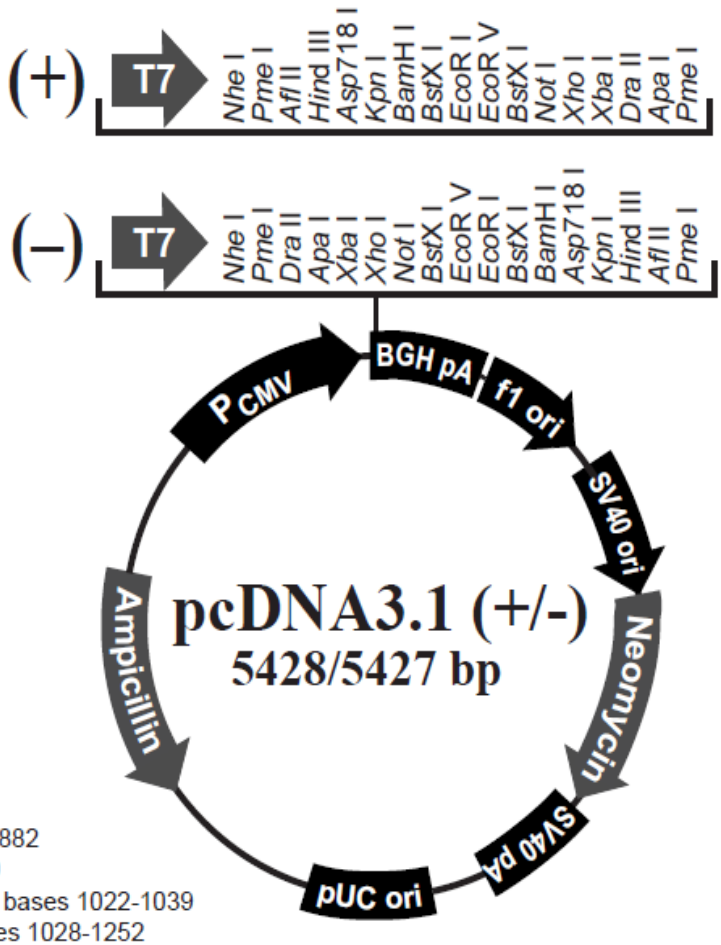


Figure 2–3: Flag-tagged CMV-4 (pFLAG-CMV-4) vector map including the multiple cloning sites. Image reproduced with minor modifications from **(**https://www.sigmaaldrich.com/content/dam/sigma-aldrich/docs/Sigma-Aldrich/Vector/1/pflag-cmv-3-4_expression_vector.pdf).



Comments for pcDNA3.1 (+)
5428 nucleotides

CMV promoter: bases 232-819
T7 promoter/priming site: bases 863-882
Multiple cloning site: bases 895-1010
pcDNA3.1/BGH reverse priming site: bases 1022-1039
BGH polyadenylation sequence: bases 1028-1252
f1 origin: bases 1298-1726
SV40 early promoter and origin: bases 1731-2074
Neomycin resistance gene (ORF): bases 2136-2930
SV40 early polyadenylation signal: bases 3104-3234
pUC origin: bases 3617-4287 (complementary strand)
Ampicillin resistance gene (*bla*): bases 4432-5428 (complementary strand)
ORF: bases 4432-5292 (complementary strand)
Ribosome binding site: bases 5300-5304 (complementary strand)
bla promoter (P3): bases 5327-5333 (complementary strand)

Figure 2–4: Untagged pCDNA3.1 (with CMV4 promoter) overexpression vector map showing detailed vector components and sequences. Image adapted from (https://www.thermofisher.com/order/catalog/product/V79020).

Primer sequence (5' to 3')	Primer Ta °C	Restriction site Enzyme	Vector to be inserted in
CCGCGAATTC**C**ATGTCCATGCTGCCCAC	75	*Eco*RI	pFLAG-CMV-4
CGACGGATCC**G**CTAGGAGCCCAGGTC	72	*Bam*H1	
GCTCGGATCC**T**ATGTCCATGCTGCCCAC	74	*Bam*H1	pCDNA3.1+
CCGTTGGAATTC**G**CTAGGAGCCCAGGTC	72	*Eco*RI	

Table 2–7: Primers with added restriction sites to amplify *SIX2*-WT and *SIX2*-Mut ORFs. The sequence in red refers to the translation start or stop sites. Green sequence refers to a polylinker. Underlined sequences are restriction enzyme recognition site. The nucleotide in **bold** was added to ensure ligation of the ORF and maintain the open reading frame. The primer Ta was calculated based on the following formula: the Ta °C = 2 x (A + T) + 4 x (C + G).

2.7.3.1 Purification of amplified PCR products

GenElute PCR Clean-Up Kit was used to purify PCR products from excess primers, dNTPs and DNA polymerase (Sigma-Aldrich, NA1020). Purified amplification products are required for successful digestion and ligation of ORFs to overexpression vectors. All centrifugations were performed at 14,000 x g and room temperature. First, the GenElute plasmid mini spin column was equilibrated with 500µl of the Column Preparation Solution, followed by centrifuging and discarding of the supernatant. In a clean tube, 5 volumes of the Binding Solution were added to 1 volume of the PCR reaction and mixed properly. The mixed solution was transferred to the equilibrated spin column, spun for 1 minute, followed by discarding flow-through. 500µl diluted Wash Solution was added to the column and centrifuged for one minute. The spin

columns were centrifuged again for two minutes to remove excess washing solution. The columns were transferred to clean collection tubes and 50µl of Elution Solution were added to elute DNA. Finally, the spin columns were centrifuged for two minutes and eluted DNA was stored at -4 °C.

2.7.3.2 Restriction digestion and ligation of ORFs to plasmid vectors

A restriction enzyme digestion/ligation method was used to digest and ligate the amplified ORF to the overexpression vector. 40µl of purified PCR product and 5 µl of either pFLAG or pCDNA3.1+ vectors were double digested with *Eco*RI (Promega, R6011) and *Bam*HI (Promega, R6021). The double digest reaction was prepared as follows: PCR product, overexpression vector, 1.5µl of each restriction enzyme, 6µl 10X multicore restriction buffer E, 0.6µl Bovine Serum Albumin (BSA) and distilled water to make 50µl reaction final volume. The double digestion reaction was subsequently incubated at 37°C for two hours. Appropriate single digestion, double digestion and negative (no enzyme) controls were included in the experiment.

The digested ORF and vector were ligated using T4 DNA ligase kit, according to the manufacturer's instructions (Promega, A1360). The ligation reaction consisted of 1µl T4 DNA ligase (1 unit), 3.5µl insert DNA, 0.5µl plasmid vector and 5µl 2X ligation buffer. The reaction mixture was gently mixed and then incubated overnight at 4°C.

2.7.3.3 Transformation of bacterial cells with plasmid vectors ligated with ORF

Plasmid vectors ligated with ORF were transformed into Alpha-select silver efficiency competent cells (Bioline, BIO-85026). Competent cells were slowly defrosted on ice. In a clean microfuge tube, 50µl of the competent cells were gently mixed with 5µl of

the ORF/plasmid ligated reaction (10% of cells' volume). The cells mixture was incubated on ice for 30 minutes, followed by heat shocking of the cells for 50 seconds at 42°C. The heat-shocked cells were incubated on ice for two minutes. Afterwards, 500µl SOC medium (pre-warmed to room temperature) was added to cells (Invitrogen, 15544034). The cells mixture was incubated for one hour at 37°C on a shaker (250 RPM).

After incubation, 50–100µl of the cells were grown on agar plates (with 100µg/ml ampicillin antibiotics) and incubated overnight at 37°C. Following incubation, ten colonies for each ORF/plasmid were aseptically selected and grown in ~5ml selective LB-broth. Selected colonies were incubated overnight at 37°C with shaking. Clonal cultures were then harvested to extract DNA plasmid (following section).

2.7.4 Small-scale plasmid DNA extraction and purification from bacterial cells using Miniprep kit

Small-scale (<100 µg) plasmid DNA was extracted from bacterial genomic DNA using GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific, K0502), following the manufacturer's instructions. Bacterial cells cultured in 5 ml LB broth were harvested by centrifugation at 6800 x g for 5 minutes at room temperature, followed by completely decanting of the medium. In the subsequent steps, all centrifugations were carried out at 12,000 x g and room temperature. 250µl of the Lysis Solution was added to the bacterial cell pellet and mixed thoroughly by inverting tubes, followed by incubating for a maximum of 5 minutes. Lysis buffer solubilises cell membrane and walls, denatures proteins and converts double-stranded DNA (ds-DNA) to single-stranded DNA (ss-DNA). Afterwards, 350µl of the Neutralization Solution was added to the cell lysates and was mixed immediately, to avoid precipitation of bacterial cell debris. The

neutralization solution allows for the transformation of ss-DNA to ds-DNA. The tubes were then spun for 5 minutes to precipitate cell debris and chromosomal DNA. The clear supernatant was carefully transferred (without cell debris) to supplied GeneJET spin column. The spin columns were centrifuged for two minutes and flow-through was discarded. Subsequently, the spin columns were washed twice with 500µl of the ethanol-diluted Wash Solution, decanting the flow-through each time. An additional two-minute spin was performed to remove all residual washing solution. To elute DNA, the spin columns were placed in clean microcentrifuge tubes, 50 µl of the Elution Buffer was carefully pipetted to the spin column centre and incubated for 3 minutes at room temperature. Lastly, the tubes were centrifuged for 3 minutes and plasmid DNA was stored at -20°C until needed for subsequent DNA sequencing or transfection experiments.

2.7.4.1 Verification of DNA plasmids after digestion and ligation

Similar to SDM verification (Section 2.7.2.4), ten colonies from each ligated ORF/plasmid were checked to ensure correct insert sequence and orientation. On the back of the agar plate, colonies were clearly labelled (#1–10) in case additional culturing from the same colony(s) was needed (e.g., culturing for a large-scale DNA prep). Each plasmid preparation was digested with *Eco*RI and *Bam*HI enzymes and gel electrophoresed (Section 2.5.1.2) to verify correct ligation. Plasmid preps showing correct ligation/digestion were subsequently Sanger sequenced to verify the ORF sequence.

2.7.5 Large-scale plasmid DNA extraction and purification using the Plasmid Maxi Kit

The QIAGEN Plasmid Maxi Kit (Qiagen, 12162) was used to extract larger quantities of plasmid DNA from bacterial cells. After confirmation of sequence and the insert of *SIX2* ORF, one sequence/insert verified colony (labelled previously) was aseptically selected with a sterile toothpick and cultured in 100 ml LB-broth medium. LB-broth was incubated overnight at 37°C (on 250 RPM shaker). After incubation, the tubes were centrifuged at 6000 x g for 15 minutes at 4°C, followed by completely decanting the supernatant. The bacterial cell pellets were resuspended in 10ml Resuspension Buffer P1, ensuring adequate mixing. To lyse the cells, 10 ml of Lysis Buffer P2 was added and mixed properly by inverting of tubes, then incubated at room temperature for up to five minutes. To precipitate cell materials (including plasmid DNA), 10 ml of pre-chilled Neutralizing Buffer P3 was added to the cell lysates and incubated on ice for 20 minutes. The cell lysates were carefully poured into a funnel containing a Whatman filter paper (3mm) to remove membrane-associated proteins and cell debris and were allowed to empty by gravity flow. Afterwards, the supernatant was centrifuged at 20,000 x g for 15 min at 4°C to ensure complete exclusion of all cell debris. Meanwhile, QIAGEN-tip 500 was equilibrated by applying 10 ml Buffer QBT, allowing the column to empty by gravity flow. After centrifugation, the supernatant was carefully transferred to the equilibrated QIAGEN-tip 500 and was allowed to pass through the resin by gravity flow. The QIAGEN-tip 500 was subsequently washed two times with 30ml buffer QC, allowing the washing buffer to pass by gravity flow. To elute DNA, The QIAGEN tip was placed in a clean polypropylene tube, followed by adding 15ml Buffer QF. Afterwards, 0.7 volume of absolute isopropanol was added and mixed to DNA-eluted buffer QF to precipitate DNA. The mixture was centrifuged at 15,000 x g for 30 minutes

at 4°C, followed by decanting of the supernatant. The DNA pellet was then washed with 70% ethanol, centrifuged at 15,000 x g for 30 minutes at 4°C, followed by removing the supernatant. DNA was left to air dry for 5–10 minutes and was subsequently re-dissolved in 350µl TE buffer. DNA was finally stored at -4°C.

2.8 Mammalian cell culture and manipulation

2.8.1 General tissue culture

Using sterile instruments and equipment, all cell handling was performed in class II tissue culture hoods, wiped with 70% ethanol. In a 37°C incubator supplied with 5% CO_2, HEK293 and COS7 cell lines were grown in 35-75 cm^3 vented flask. Both types of cell lines were available from the Molecular and Medical Genetics Laboratory, Institute of Cancer and Genomics, University of Birmingham.

2.8.2 Culture media preparation, cell maintenance and propagation

2.8.2.1 Culture media preparation

HEK293 and COS7 cell lines were cultured in pre-warmed Dulbecco's Modified Eagles Media (DMEM) (Sigma-Aldrich, D6546). To make a complete medium, DMEM was supplemented with 10% (v/v) Fetal Bovine Serum (FBS) (Sigma-Aldrich, F7524) 2mM L-glutamine (Invitrogen, 25030-024) and 1% penicillin/streptomycin (5,000 U/ml) (Invitrogen, 15070-063).

Ensuring enough coverage of cells with medium, sufficient growth medium was added to cell flasks depending on the flasks' volume (8–9 ml for 25 cm^3 flask; 13–14 ml for

75 cm^3). Once the cells were grown to 80-90% confluency, they were split in 1:3-1:5 depending on their growth rate.

2.8.2.2 Passage of confluent cells

To passage HEK293 and COS7 cell lines, the medium was completely removed from flask and the cells were gently washed with pre-warmed phosphate buffered saline (PBS). To detach adherent cells from the flask, 1ml of 1X Trypsin-EDTA (Sigma-Aldrich, T3924) was added to the flask and incubated at 37°C up to five minutes. The cell suspension was spun at 250 x g for 5 minutes at room temperature, followed by careful removal of the supernatant. The cells were lastly rehydrated in a fresh complete culture medium, diluted to chosen volumes and incubated at previously mentioned conditions.

2.8.2.3 Preparation of LB agar plates and broth medium

LB agar plates were prepared by dissolving 7 g of LB broth with agar powder (Sigma-Aldrich, L2897) in 200 ml of distilled water. The mixture was autoclaved and, subsequently, cooled to ~40–50°C before aseptically adding the appropriate antibiotics (100µg/ml). Antibiotics added to the cell culture acts as a selection factor, allowing for the growth of the plasmid of interest (containing the specified antibiotic resistance gene). The agar mixture was carefully poured into sterilized 10 cm diameter Petri dishes (avoiding air bubble formation) and allowed to cool for approximately 45 minutes. The plates were then inverted and stored at 4°C until used.

The LB liquid broth was made by dissolving 4 g of LB broth powder (Sigma-Aldrich, 3022) in 200 ml of distilled water, followed by autoclaving. The broth was cooled down before the addition of desired antibiotics and was stored at 4°C until needed.

2.8.2.4 Freezing cells for prolonged storage

The cell lines were kept long-term in liquid nitrogen. In a sterile cryovial, approximately 12 x 10^6 cells were resuspended in 1 ml of culture FBS with 10% (v/v) dimethyl sulphoxide (DMSO) cryoprotectant. The cryovials were then frozen using a cryopreservation pot to allow for gradual freezing (e.g., -1°C per minute) to reach -80°C. Afterwards, the cryovials were kept in liquid nitrogen for long-term storage.

2.8.2.5 Thawing and reviving of frozen cells

Frozen cells were defrosted in a 37°C water bath until just thawed (30–60 seconds) to reduce the chances of cell toxicity caused by warmed DMSO. The cell suspension was quickly transferred to a polypropylene tube containing 4 ml pre-warmed complete medium, followed by centrifugation of cells at 250 x g for 5 minutes. The supernatant was then discarded carefully. Afterwards, the cell pellet was resuspended in 9ml complete medium, transferred to a fresh flask and incubated accordingly.

2.8.2.6 Cell counting using haemocytometer

The cultured cells were counted using a Countess II Automated Cell Counter instrument (ThermoFisher Scientific, AMQAX1000). The cells were first coloured using trypan dye (ThermoFisher Scientific, SV30084.01), which assesses cell viability by

staining dead cells. The cells were detached, pelleted and adequately resuspended in a known volume of culture medium. 10µl of cell suspension was mixed with equal volume of trypan dye, followed by loading the mixture into the chamber of a Countess Cell Counting disposable slide (Thermo Fisher Scientific, C10228). The slide was then inserted into the instrument and cells were counted automatically (cells/ml).

2.9 Transfection of Cells

All cell transfections were achieved using the FuGENE HD transfection reagent (Promega, E2311), according to manufacturer's recommendations. Cells to be transfected were seeded into 6-well plates with a tissue culture surface (Corning, BC010) or 8-well chamber 0.7 cm^2 polystyrene tissue culture treated glass slide (Falcon, 354108). Approximately 3.0 x 10^5 of HEK293 or 1.0 x 10^4 of COS7 cells were seeded in 6-well plates or 8 well chamber glass slide for western blot (WB) or immunocytochemistry, respectively. Suitable volumes of complete medium were added to the seeded cells and, subsequently, incubated overnight to achieve optimal transfection cell density (~60-80% confluency).

Next day, the transfection mixtures were prepared in a polypropylene tube as following: 200µl Opti-MEM reduced serum medium (Invitrogen, 31985062), 6µl of FuGENE for 2µg DNA (3µl FuGENE: 1µg DNA). The FuGENE solution was directly pipetted and mixed in Opti-MEM medium, followed by adding plasmid DNA (at 1µg/µl concentration). The complete transfection mixture was properly mixed and left to incubate at room temperature for 20 minutes. Subsequently, 200µl of the transfection mixture was directly added to each seeded well with gentle mixing. All transfection experiments were incubated for 48 hours at 37°C (5% CO_2).

2.9.1 Harvesting of transfected cells for protein extraction

Following transfection incubation, the seeded cells were directly harvested using lysis buffer to extract proteins. The Radioimmunoprecipitation assay buffer (RIPA) lysis buffer was prepared from the following components: 50mM Tris pH 8, 150mM NaCl, 0.1% sodium dodecyl sulphate (SDS), 0.5% IGEPAL, 1mM EDTA, 0.5% Na-deoxycholate, and a protease inhibitor cocktail tablet (per 100 ml solution) (cOmplete by Roche, 04693116001). Growth medium was aspirated and cells were carefully washed twice with 0.1M PBS (137mM NaCl, 10mM Na_2HPO_4, 1.8mM KH_2PO_4 and 2.7mM KCl in H_2O at pH 7.4). 200µL of RIPA lysis buffer was directly added to the culture dish and incubated for 5 minutes on ice. Afterwards, the lysed cells were scraped from the culture dish and placed into microfuge tubes. To release the membrane-bound proteins, the harvested lysates were sonicated for ten-second bursts for 3 times. The cell lysates were centrifuged at 14,000 RPM for 30 minutes at 4°C. Subsequently, the clear supernatant containing proteins was carefully transferred into a 1.5 clean microcentrifuge tube and stored at -20°C.

2.10 Protein analysis using Western blot

2.10.1 Protein quantification

A Bio-Rad DC Protein Assay (Bio-Rad, 5000116) was used to quantify the proteins for the WB experiment. In this assay, a colour change is induced upon mixing alkaline copper tartrate solution and Folin reagent with protein. These colour changes are measured to provide theoretical quantifications of the proteins.

To prepare protein standards, a series of dilutions ranging from 0.2–1.6mg/ml were prepared using 2mg/ml BSA standard stock and diluted in RIPA lysis buffer. Subsequently, working solution A was prepared by mixing 20µl reagent S to every 1ml of reagent A.

5µl of samples and protein standards were loaded into a 96-well microtitre plate ensuring careful recording of wells' positions. 25µl of reagent A was added to each well and then 200µl of reagent B. The reagents were by gently mixed and subsequently incubated for 15 minutes at room temperate.

Using a Wallac Victor3 fluorometer (Perkin Elmer) with Microplate Manager software, the absorbance of each sample was read at 690nm wavelength. A standard curve was created, from the calculation of the BSA standards, which therefore generated a regression equation that was utilized to infer the relative concentration of the samples.

2.10.2 Sodium dodecyl sulfate polyacrylamide gel electrophoresis for protein separation in WB

Proteins were separated, based on their sizes, by the sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) method. Prior to electrophoresis, proteins were solubilized in SDS detergent for the purpose of inducing a negative charge and denaturing the secondary and tertiary structure of proteins. Polyacrylamide gels (15%) were prepared as following: acrylamide Protogel (stock National Diagnostics, ED-201), 1% SDS, 375mM Tris at pH 8.8, 10% ammonium persulfate (APS), double distilled water and lastly 0.05% Tetramethylethylenediamine (TEMED) (Sigma, T9281). The gel mixture was dispensed in a clean gel plate (Biorad) to approximately three-quarters gel capacity and was left for 10 minutes to polymerize (solidify).

A stacking gel solution was prepared as follows: 4% acrylamide Protogel, (1.5ml 0.5M) Tris (50mM) pH6.8 (1.5ml 0.5M), 0.08% SDS, 10% APS, 0.06% TEMED and water. Once the gel polymerized, water was carefully decanted, and a comb was cautiously inserted, followed by pouring of stacking gel mixture to create sample wells and solidifying of the mixture. The finally prepared gel was subsequently inserted into an electrophoresis tank filled with running buffer (25mM Tris, 190mM glycine, 0.1 % SDS at pH 8.3).

2.10.2.1 Loading of protein samples and running gel

To load samples onto a polyacrylamide gel, 20µg of purified cell lysates (containing protein, Section 2.9.1) were mixed with 4µl of 5x loading buffer (10 % SDS, 0.313 M Tris HCl at pH 6.8, 0.05 % Bromophenol blue and 50% glycerol) in 20µl total volumes. The stained samples were denatured for 5 minutes at 95°C. The samples were then loaded into the designated wells of the polyacrylamide gel, including a protein marker ladder (GeneFlow, S6-0024). The loaded gel was electrophoresed at 120 volts for 1–1:30 hours until the loading dyes had reached the bottom of the gel, or the protein markers were sufficiently separated.

2.10.3 Protein transfer to nitrocellulose membrane

Following the separation of protein samples, polyacrylamide gels were transferred to a holder transfer cassette with a sheet of Hybond-P nitrocellulose membrane (GE Life Sciences, 10600022). The gel and nitrocellulose membrane were sandwiched between six layers of Whatman filter paper (3mm) and two sponges, all previously soaked in transfer buffer (25mM Tris, 190mM glycine, and 20% methanol). Specifically,

a sponge pad and three layers of filter papers were first placed into a cassette holder. Afterwards, a membrane (activated by soaking in absolute methanol) was carefully placed on top of the gel without introducing air bubbles. The remaining sponge pad and filter papers were lastly placed on top of the stacked cassette. The cassette was closed securely and immersed into the electrophoresis tank; ensuring the membrane was situated toward the (+) anode. The transfer apparatus with an ice cooler was run at 80-90 volts for 60-80 minutes.

2.10.4 Immunodetection of protein blotted to membrane

After transferring proteins, the membrane was incubated on a shaker for 60 minutes in a blocking solution, which was prepared from 5% dried fat-free milk in PBS-Tween buffer 2mM Na_2HPO_4, 0.5mM KH_2PO_4, 1.3mM KCl, 135mM NaCl, 0.05% Tween 20, pH 7.4). After decanting the blocking solution, the membrane was carefully placed into a 50ml polypropylene tube containing 5 ml blocking solution (in PBS-Tween buffer) and the primary antibody diluted to the desired concentration. The tube was then incubated on a tube roller at 4°C overnight.

After overnight incubation, the membrane was washed three times (15 minutes each) in PBS-Tween buffer on a shaker to remove unbound antibodies. Specific secondary antibodies were diluted accordingly in 5ml of 5% dried milk with PBS-Tween. The secondary antibodies solution was added to the tube containing the membrane and incubated for one hour at room temperature on a tube roller. The membrane was subsequently washed in PBS-Tween for three times on an orbital shaker to remove unbound secondary antibodies.

2.10.5 Developing western blots using the enhanced chemiluminescence method

As a part of a chemiluminescent detection system, the Amersham enhanced chemiluminescence (ECL) detection reagent (GE Health Care Life Science, RPN2106) was used to visualize bound antibodies. This system produces a chemiluminescent emission signal that can be detected on a radiographic film at short exposures. The chemiluminescent chemical reaction occurs when the secondary antibodies, conjugated to horseradish peroxidase (HRP), catalyses the oxidation of luminol in the presence of hydrogen peroxide and a catalyst.

Washed membranes were placed and smoothly spread on top of a Saran wrap. ECL reagents A and B were mixed in a 1:1 ratio and were added directly to the membrane (protein-side up). The membrane was incubated for 5 minutes at room temperate, followed by careful removal of excess developing fluids. The membrane was delicately wrapped with Saran wrap and then securely placed in a developing cassette.

In a dark room, autoradiography films (GE Health Care, 10534205) were exposed to the membrane for time periods ranging from 5 seconds to 5 minutes. This exposure time was adjusted based on the antibody signal intensity. The autoradiography films were developed in an automated SRX-101A X-ray film processor (Konica Minolta).

2.10.6 Quantitative analysis of protein expression levels

The protein expression values of protein blots were quantitatively measured using a lightbox and high-resolution camera connected with the Genetools software (GeneSnap, Syngene). Images of developed western blots were captured, followed by measuring the absorbance (arbitrary units) of the individual protein bands. To account for differences in sample loading, protein expression measurements need to be

normalized to a background signal from the developed film, followed by another normalization against a control housekeeping protein (e.g., β-actin). Measurements were also normalized to a transfection control vector (pCMV-4-*JMJD7,* courtesy of Charlotte Eaton) to account for the potential discrepancy in transfection.

2.11 Immunocytochemistry to assess protein cellular localisation

2.11.1 Cell fixation

As explained earlier, COS7 cells were transfected and seeded into an 8-well chamber slide (Section 2.9). On glass slide, COS7 cells were transfected with the following overexpression vectors (one vector for each well): p-FLAG-CMV-4 with *SIX2*-WT ORF, p-FLAG-CMV-4 with *SIX2*-Mut and an empty p-FLAG-CMV-4 (control). After the 48-hour transfection period, glass slides were removed from the 37°C incubator. Growth medium was carefully removed from the wells, the cells were washed three times, followed by decanting the PBS. The cells were subsequently permeabilised and fixed with 500µl of 4% paraformaldehyde in PBS solution for 15 minutes at room temperature. The paraformaldehyde-PBS solution is a cross-linking reagent that preserves and maintains cell structure. Afterwards, the cells were washed three times in PBS each for 10 minutes.

2.11.2 Blocking and staining of cells to be visualized under fluorescent microscopy

Following the PBS washes, the cells were incubated in a blocking solution containing 3% BSA and 0.1% Triton-X in PBS. The blocking solution was used to prevent nonspecific antibody binding.

Antibody diluting buffer was prepared with 3% BSA, 0.05% Triton X-100 in PBS. AlexaFluor-488 conjugated FLAG-tag monoclonal antibody (ThermoFisher scientific, MA1-142-A488) was diluted in 1/100 antibody dilution buffer. Buffer containing antibodies was added to the slide's wells and incubated for 60 minutes on a plate shaker at room temperate. An Alex flour dye produces brighter dye signals comparing with other standard dyes of similar emission. Following incubation periods, the cells were washed three times with PBS (5 minutes each).

To mount cells, the walls of the 8-well chamber slide were removed using the provided removing tool. A few small drops of Antifade Mounting Medium with DAPI counterstain (Vector Labs, H-1200) were placed onto a glass coverslip. DAPI counterstain binds strongly to DNA and, therefore, is an effective cell nucleus marker. DAPI produces blue-cyan dye when excited under ultraviolet light. A coverslip was slowly placed on top of the slide to avoid air bubbles, followed by sealing the edges of the coverslip with nail polish. Afterwards, the slides were wrapped in aluminium foil and kept at 4°C, prior to viewing on fluorescent microscopy. The fluorescent microscopy imaging system used to view the slides and capture images was widefield fluorescence microscope (Zeiss AxioScope).

Chapter 3: Genetic studies into congenital hypothyroidism families using whole exome sequencing

3.1 Introduction

3.1.1 CHT Definition, overview and neonatal screening

CHT is described as the insufficiency or deficiency of thyroid hormone at birth. The thyroid gland produces thyroxine (T4) and triiodothyronine (T3) hormones, both of which are essential for many developmental functions in newborn infants such as growth, brain development and normal metabolism (Shanholtz, 2013). Hence, if left untreated, CHT can lead to serious developmental delays, growth abnormalities and sensorimotor and behaviour problems (Park and Chatterjee, 2005). Initial CHT symptoms are usually mild and can be difficult to detect, including prolonged jaundice, feeding difficulties, hoarse cry and constipation (Park and Chatterjee, 2005).

In the early 1980s, many countries (including the UK) initiated neonatal screening programmes which led to the institution of appropriate and early treatment in neonates born with hypothyroidism (Park and Chatterjee, 2005; Shanholtz, 2013). Newborn screening programmes normally measure levels of T4 and TSH, which is produced by the pituitary gland. While CHT is treatable, 15% of affected children can still show some degree of learning impairment despite early treatment with thyroxine (Dimitropoulos et al., 2009). Although many CHT cases are identified in newborn screening programmes, 75% of infants are born in countries that have not yet initiated newborn screening programmes (Shanholtz, 2013).

3.1.2 Epidemiology

Inadequate uptake of iodine is the major cause of CHT cases worldwide (Bhavani, 2011). In North America and Europe (iodine sufficient countries) the incidence of CHT is one in 3500–4000 newborns, making it the most common inherited endocrine disorder (Park and Chatterjee, 2005). CHT incidence can vary between ethnic groups and geographic location; moreover, it can be twice as common in girls (i.e., female predominance) (Castanet et al., 2001; Leger et al., 2002). CHT incidence is higher in Asian populations, whereas Caucasians and African American ethnicities have the lowest prevalence (Castanet et al., 2001). Notably, the incidence of CHT in consanguineous families is approximately 1 in 700, which is higher than in non-consanguineous families, highlighting the contribution of genetic factors (Hall et al., 1999).

3.1.3 Classification and subtypes: transient versus primary and secondary hypothyroidism

Maternal or neonatal factors can cause transient hypothyroidism, a transitory insufficiency of thyroid hormone production identified at birth, which later reverts to normal thyroid hormone production (Rastogi and LaFranchi, 2010; Bhavani, 2011). The following maternal factors can cause transient CHT: insufficient or excess maternal iodine intake, anti-thyroid medications and maternal thyroid-stimulating hormone receptor (TSHR) antibodies. Neonatal factors that can cause transient CHT include neonatal insufficient iodine, congenital liver haemangiomas and mutations in *DUOX* and *DUOXA2* genes (Rastogi and LaFranchi, 2010).

Primary or secondary (central) causes can be associated with permanent CHT (present at birth), which requires lifetime thyroxine replacement treatment. The most common primary causes are (1) aberrations of thyroid gland development, known as thyroid gland dysgenesis (TGD); and (2) defects in the production of thyroid hormones—thyroid dyshormonogenesis (TDH) (Park and Chatterjee, 2005). Infrequently, other pituitary or hypothalamic diseases can cause insufficient production or binding of TSH, resulting in secondary hypothyroidism along with other pituitary hormone deficiencies. In addition, defects in the TSH β-subunit gene result in congenital TSH deficiency. Impaired thyrotropin releasing hormone can also cause central hypothyroidism (Park and Chatterjee, 2005; Rastogi and LaFranchi, 2010).

3.1.4 Complexity of the aetiology of CHT cases

The aetiology of the majority of CHT cases is not comprehensively understood due to the sporadic nature and wide geographical spread of the cases. Epigenetic and environmental factors are thought to be involved in sporadic cases (Brust et al., 2012). The majority of TGD cases are sporadic; by contrast, biallelic mutations in CHT-related genes have been identified in approximately 15% of TDH cases (more in Sections3.1.5.1 & 3.1.6.1) (Park and Chatterjee, 2005).

A study by Perry et al. (2002) reported five pairs of monozygotic twins who were all discordant for TGD, supporting the involvement of other factors (e.g., environmental) in the pathogenesis of CHT. On the other hand, in 2% of TGD cases, a positive familial history was reported, which suggests the role of genetic factors (Park and Chatterjee, 2005). Some studies have shown that 7.9% of first-degree relatives of congenital TGD-affected individuals have a higher proportion of asymptomatic thyroid dysgenesis anomalies, which also highlights the involvement of genetic factors (Leger et al., 2002).

3.1.5 Thyroid gland dysgenesis

TGD is considered the most common cause of hypothyroidism, accounting for 85% of CHT cases (Park and Chatterjee, 2005). The majority of TGD cases occur sporadically (95–98% without known genetic causes), where the remainder of the cases is familial (Rastogi and LaFranchi, 2010). Thyroid gland anomalies can occur in three forms: (1) an ectopic (abnormal location, yet with normal size or hypoplastic) thyroid, accounting for 35–42% of TGD cases; (2) athyreosis (absence of thyroid tissue) 35–42% of cases; and (3) thyroid hypoplasia (small-sized gland), 24–36% of cases (Pohlenz et al., 2002; Kumar et al., 2009a).

3.1.5.1 Genetics of thyroid gland dysgenesis

Some genes harbouring homozygous or heterozygous changes have been implicated in a minority of TGD cases (approximately 2%), *PAX8*, *TTF-2/FOXE1*, *NKX2-1* and *NXK2-5* (Al Taji et al., 2007; Cherella and Wassner, 2017; Rastogi and LaFranchi, 2010). These genes encode transcription factors that are expressed during thyroid development and are responsible, along with other proteins, for normal embryonic thyroid development and migration (Park and Chatterjee, 2005). Knock-out mice for *Foxe1*, *Nkx2.1* and *Pax8* show anomalies in thyroid gland development, highlighting the important role of these genes in normal thyroid development (Al Taji et al., 2007). In humans, mutations in *FOXE1*, *NKX2-1* and *NKX2-5* can lead to distinct phenotypic features, in addition to TGD, due to the expression of these genes in other tissues (Table 3–1) (Rastogi and LaFranchi, 2010). Additional genes, including *JAG1*, *GLIS3* (Wassner, 2018) and *CDCA8* (Cherella and Wassner, 2017) have been recently

implicated in a few TGD families, accompanied with other congenital abnormities (Table 3–1).

3.1.6 Thyroid dyshormonogenesis

Thyroxine hormone is synthesised in several stages, involving multiple genes (Figure 3–1). Defects in any of these genes can affect the synthesis of thyroid hormones and subsequently cause TDH. When thyroid hormone synthesis is defective, secretion of TSH is increased due to the lack of the negative feedback loop T3/T4 on the pituitary gland (hypothalamic–pituitary–thyroid axis) (Fisher et al., 2000). This increase of TSH secretion can lead to thyroid gland hyperplasia in patients without TGD, resulting in a goitre development, a clinical characteristic of TDH that is rarely seen in infants detected by newborn screening and is usually absent in TGD cases (Park and Chatterjee, 2005; Rastogi and LaFranchi, 2010).

3.1.6.1 Genetics of thyroid dyshormonogenesis

TDH accounts for 15–20% of CHT cases and is generally transmitted in an AR inheritance pattern (Rastogi and LaFranchi, 2010). Mutations in the following genes have been associated with TDH: *SLC5A5*, *SLC26A4*, *TPO*, *TG*, *IYD*, *DUOX2* and *DUOXA2*.

The availability of iodide for thyroid hormone synthesis is controlled by two specific systems in the thyroid gland: (1) SLC5A5 sodium-iodide symporter facilitating iodide uptake/transport (rate-limiting step of T3/T4 synthesis) (Nicola et al., 2011); (2) the recycling of iodide through the deiodination of monoiodotyrosine and diiodotyrosine by nitroreductase enzyme (encoded by *IYD*) (Moreno et al., 2008). Hence, genetic defects

in *SLC5A5* and *IYD* can impair the transportation or the recycling of iodide, respectively, leading to TDH. Heterogeneous clinical manifestations of *SLC5A5* mutations include a variable degree of CHT and goitre, as well as absent or reduced uptake of iodide (Nicola et al., 2011).

Gene	**Tissues expressed in**	**Additional congenital anomalies**
FOXE1 (TTF-2)	Thyroid, pharyngeal endoderm derivatives	Cleft palate, bilateral choanal atresia, spiky hair
NKX2-1 (*TTF-1*)	Thyroid, lung and fetal heart	CHT, predominant dyskinesia, neonatal respiratory distress, and mild hyperthyrotropinemia
NKX2-5	Heart and fetal heart	Thyroid hypoplasia and an increased cardiac malformation
PAX8	Thyroid and kidney	TGD and sometimes kidney-related anomalies
TSHR	Thyroid	Thyroid stimulating hormone resistance that can be accompanied thyroid defects
JAG1	Artery – tibial	Alagille syndrome (variable liver, heart, eye, skeletal, facial anomalies), congenital heart disease
GLIS3	Thyroid	Neonatal diabetes mellitus, congenital glaucoma, developmental delay, hepatic fibrosis, polycystic kidneys
CDCA8	Thyroid	Broad thyroid development anomalies, including agenesis, ectopic to euthyroid with asymmetric thyroid lobes or nodules

Table 3–1: Genes that have been implicated in TGD. In addition to the TGD phenotype, mutations in these genes can be associated with additional congenital defects. Table compiled from information in (Cherella and Wassner, 2017; Park and Chatterjee, 2005; Rastogi and LaFranchi, 2010; Wassner, 2018).

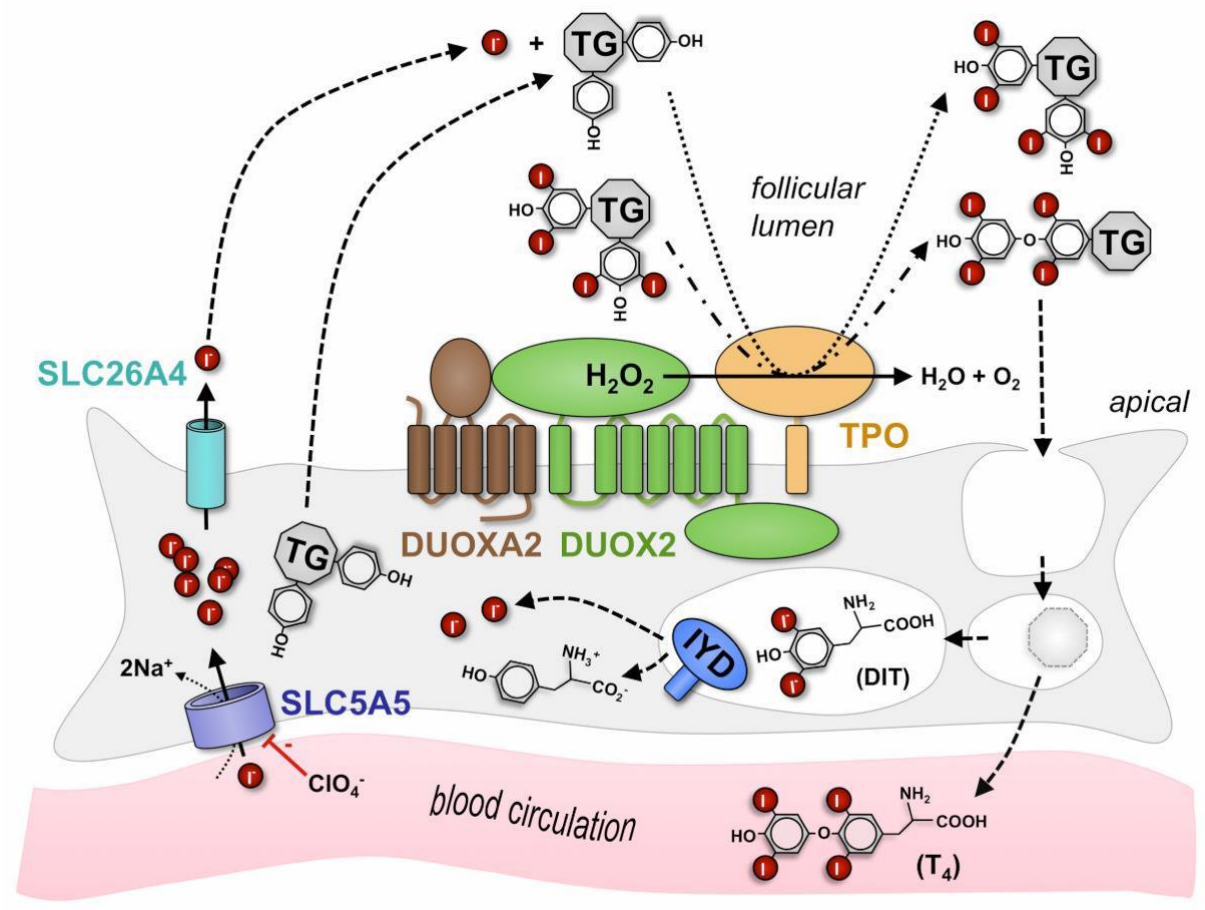


Figure 3–1: Schematic representation of key elements involved in the synthesis of thyroid hormone of a follicular thyroid cell. The steps of thyroid hormone synthesis involve: (1) the uptake of iodide from the circulating blood by the sodium-idodide transporter (SLC5A5); (2) Apical anion channel (SLC26A4) facilitates the flow of iodide into the colloid; (3) Thyroid peroxidase (TPO) catalyses the iodination of the tyrosine groups of thyroglobulin (TG;) (4) Iodinated tyrosines are subsequently coupled within TG via ether-bond formation; Hydrogen peroxide (H_2O_2) is required for steps 3 and 4 as a co-substrate. H_2O_2 generates NADPH-oxidase constituted by dual oxidase 2 (DUOX2) along with its maturation factor, DUOXA2; (5) After the endocytosis, the lysosomal degradation of TG matrix complex liberates the iodothyronines (T4>T3); (6) Iodotyrosine deiodinase (IYD) dehalogenates the concurrently released iodotyrosines which therefore allows the 'reusing' of iodide for further thyroid hormone synthesis. Figure adapted from (Grasberger and Refetoff, 2011).

Pendred syndrome is caused by biallelic inactivating mutations in the *SLC26A4* gene. Pendred syndrome is responsible for 10% of inherited deafness cases, making it the most common cause of inherited deafness (Kopp, 2014). Patients with Pendred syndrome usually have congenital moderate to severe sensorineural hearing loss (Bizhanova and Kopp, 2010). The passive movement of iodide across the apical membrane through the follicular lumen is facilitated by *SLC26A4*, encoding a sulfate transporter (pendrin) that belongs to the SLC26 transporter family (Grasberger and Refetoff; Rastogi and LaFranchi, 2010). Due an impaired function of SLC6A4 in iodide efflux in thyrocytes, patients with Pendred syndrome can consequently develop a thyroid goitre between late childhood and early adolescence (Grasberger and Refetoff, 2011). However, approximately half of patients with impaired SLC26A4 can be asymptomatic for thyroid abnormalities (Grasberger and Refetoff, 2011). TSH levels are often elevated in these patients, possibly leading to development of hypothyroidism (Park and Chatterjee, 2005).

Defects in the *TPO* gene are the most common cause of inherited TDH cases with permanent hypothyroidism (Rastogi and LaFranchi, 2010). Located on the apical membrane surface of thyroid follicular cells, TPO is important for thyroid hormone biosynthesis. Defects in TPO can lead to impaired total iodide organification and consequently result in goitre development (Rastogi and LaFranchi, 2010; Grasberger and Refetoff, 2011).

TG gene encodes a homodimeric glycoprotein that is abundantly and exclusively expressed in the thyroid gland and represents a specialized matrix for thyroid hormone biosynthesis (Targovnik et al., 2011; Grasberger and Refetoff, 2011). The presence of a goitre, low/absent serum TG levels and normal perchloride discharge levels in patients with CHT can be indicative of a TG genetic defect (Targovnik et al., 2011).

Patients had biallelic *TG* mutations show high levels of serum TSH and altered levels of T3/T4 circulating hormones (Grasberger and Refetoff, 2011).

3.2 Aims of the study and the clinical information of CHT families

Although multiple genetic loci are known to be associated with CHT, a comprehensive knowledge of the genetic causes of CHT has not yet been achieved. The number of genes associated with CHT continues to increase, but the known CHT disease-causing genes explain only a fraction of CHT cases. Universal newborn screening regimes have resulted in increased detection of CHT cases; however, newborn screening is subject to false-negative results and, therefore, it cannot be the perfect diagnostic test. The primary aim of this study was to identify know or novel causative gene(s) in four consanguineous CHT families using WES. Once a potential disease-causing variant was identified, other family members or CHT patients will be screened to confirm the segregation of the variant with the disease. The effect of the candidate variant will subsequently be characterised by *in vitro or in vivo* functional studies to understand the pathogenic mechanism in relation to CHT development (more in Chapter 4).

The clinical data of all four families were reviewed by Prof Timothy Barrett (Diabetes Unit, Birmingham Children's Hospital) and are summarised in Table 3–2. The CHT phenotype (i.e. TGD/TDH) was confirmed in all families prior to WES. The pedigrees of the all families are shown in Figure 3–2.

WES was carried out on DNA samples (n=12) (see Section 2.1.1) from four consanguineous CHT families (Family 1–4), involving unaffected parents-affected index (trio approach) (n=2), or unaffected parent and sibling-affected member approach. Three of these families (Family-1, -3, -4) were affected with TGD, whereas

one had TDH (Family-2). WES and initial bioinfromatic analyses were performed by Oxford Gene Technology Company (Oxfordshire, UK) (see Sections 2.3 and 2.3.5.1).

3.3 WES data analysis and results

3.3.1 Quality check metrics and investigation into variants in CHT-associated genes

Prior to the start of WES data analysis, the foremost step was to QC the data to ensure its reliability. Details of QC process and metrics for all families are in Section 8.1 and Appendix Figure 8–1, respectively. All exome-sequenced individuals showed satisfactory QC metric results.

The first step in WES data analysis was to identify whether any affected individuals harboured genetic variants in known CHT-associated genes. Because the coverage of WES is not uniform across the targeted genomic regions, especially at intronic/exonic boundaries, the coverage of all exons of CHT-associated genes was visualised and checked manually using the IGV (more about IGV in Section 2.3.3.2). A 10X read depth threshold was set as the minimum required for genomic regions to be classified as adequately covered (Prof Jean-Baptise Cazier [2015], Director of the Centre for Computational Biology, University of Birmingham, personal communication). Following this read depth threshold, 15 exons in eight CHT genes were not covered adequately across the four families (Table 3–3). These exons were manually amplified and Sanger sequenced using DNA from the affected individuals to ensure potential disease-causing variants were not missed in these regions (primer is Appendix Table 8–13). Combining the manually sequenced exons with WES data, in affected family members, variants detected in CHT-causing genes were investigated further (Table 3–4).

To classify variants as potentially ‘disease-related, in the context of consanguineous families, a variant is expected to be homozygous (in a TDH-associated gene) in a TDH-affected member, heterozygous or homozygous (in TGD-associated gene) in TGD patient and not common in the general population (MAF<1%) using ExAc and dbSNP population databases (more about these databases in Section 2.6.6). Some evidence has supported the association of rare heterozygous mutations in thyroid-related transcription factors in TGD families (Park and Chatterjee, 2005); thus, heterozygous variants in CHT-related genes were evaluated. The MAF represents the relative frequency of an altered allele in a population. Generally, variants with MAF ≥5%, 1–5% or <1% are considered common, low frequency or rare, respectively (Wessel et al., 2015).

After carefully investigating the variants detected in CHT-associated genes, no variants were classified as potentially CHT disease-related as they were either detected in other unaffected family members, common in the general population, or were exome artefacts (Table 3–4). The subsequent step was to filter and prioritise WES data to identify potential novel CHT candidate genes in the families.

CHT Family	TGD / TDH	Index gender	Heel prick TSH (nmol/L)	Venous TSH (mU/L)	Venous FT4 (pmol/L)	Imaging	Additional clinical notes
1	TGD	F	>248	>100	<5.2	USS did not detect thyroid tissue; technetium scan was unsuccessful	Normal growth, development and hearing
2	TDH	F	N/A	16.3	14.7	Ultrasound scan shows normal thyroid gland	N/A
3	TGD	F	>258	>464	4.4	Not performed	Normal growth, development and hearing. No goitre on examination
4	TGD	F	N/A	N/A	N/A	N/A	Has one sister with acquired hypothyroidism at 4.5 years and other sister with TGD

Table 3–2: Clinical information of the four CHT-affected families. TGD: thyroid gland dysgenesis; TDH: thyroid dyshormonogenesis; TSH: thyroid stimulating hormone; FT4: free thyroxine.

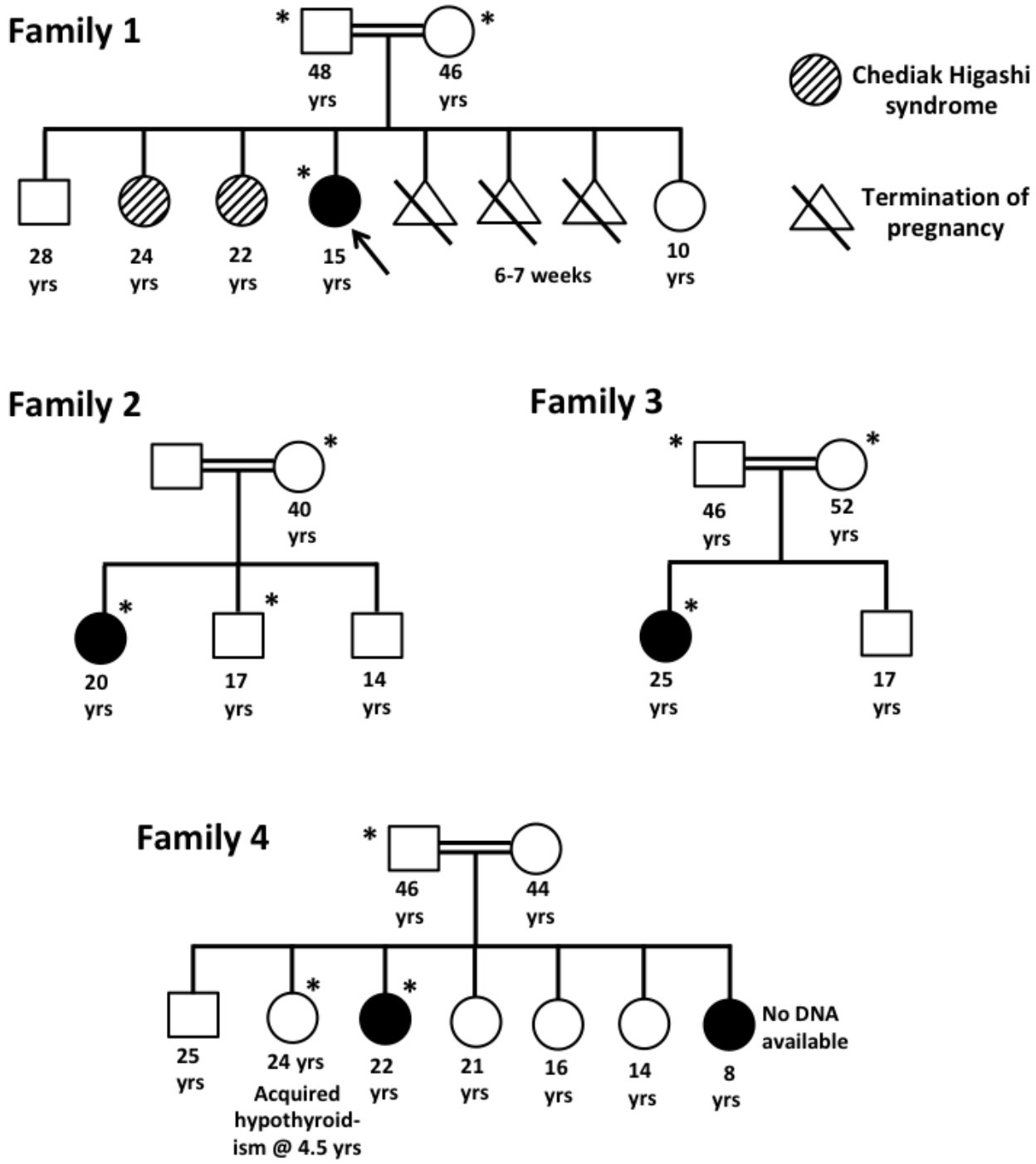


Figure 3–2: The pedigree of CHT families. All families were associated with TGD, except for Family-2 which was a TDH. All families are consanguineous, which is denoted by a double line between the parents. Squares represent males; circles represent females. Black circles represent CHT affected members. Individuals that were exome sequenced are denoted by an asterisk (*)

Gene	Exon not covered in WES	Family
DUOX1	6	Family-2
	7	Family-2
	8	Family-2
	15	Family-2
	30*	Family-2
DUOX2	6	Family-2
	7*	Family-2
	8*	Family-2
SLC16A10	1	Family-2
TPO	8	Family-2
TG	27*	Family-2
SLC5A5	1*	Family-2
	12	Family-2
FOXE1	1*	Family-4
NKX2-1	3*	Family-4

Table 3–3: Exons of CHT-associated gene that were not adequately covered by WES. A minimum of 10 independent exome reads was set as the minimum threshold for an exon to be classified as 'adequately covered'. *Exons that were PCR amplified with the help of a summer student, Shahswar Zearmal (MBChB, University of Birmingham). All the sequencing data were analysed by Naser Ali.

Gene	Variant	Genotype	Detected in	ExAc SA freq	rs entry	Final remarks
TPO	c.483-1G>A; Splice site	Hetro	Family-2-A	Not reported	Not reported	WES sequencing artefact, not detected by Sanger sequencing
SLC26A4	c.1003T>C; p.Phe335Leu	Hetro	Family-2-A	0.00086	rs111033212	Unlikely disease-causing because heterozygous
TG	c.7640T>A; p.Leu2547Gln	Hetro	Family-2-A	0.0374	rs2979042	Unlikely disease-causing because heterozygous and in general population (3.7%)
DUOX2*	c.413C>T; p.Pro138Leu	Homo	Family-2-A	0.9279	rs2001616	Unlikely disease-causing, very common in general population
NKX2-1	c.446A>C; p.Asp149Ala	Homo	Family-2-A,-F,-M	Not reported	Not reported	Unlikely disease-causing because present in two unaffected members
FOXE1	c.505_538 delGCCG…; p.Ala178_Ala179 del	Homo	Family-3-A,-F,-M; Family-4-A	0.0918	rs3021524	Unlikely disease-causing because present in two unaffected members and common in general population
FOXE1*	c.-156T>C; UTR	Homo	Family-4-A	0.82	rs1867279	Unlikely disease-causing because very common in population
FOXE1*	c.-131G>C	Homo	Family-4-A	0.64	rs1867280	Unlikely disease-causing because very common in population

Table 3–4: Variants detected in CHT-associated genes in affected members. Variants were detected from WES data, or by manual Sanger sequencing of inadequately covered genomic regions (denoted by *). Family members denoted by A: affected; F: father; M: mother. UTR: untranslated region; Hetro: heterozygous; Homo: homozygous; ExAc: Exome Aggregation Consortium population database in South Asians (SA); rs: reference SNP ID.

3.3.2 Filtering of WES data to identify potential disease-causing candidate variants

An average of 68,075 SNVs and INDELs were detected in the four families. The large number of identified variants by WES is one of the primary challenges in identifying disease-causing genes (Robinson et al., 2014). Hence, it was necessary to follow a step-wise filtering scheme to reduce this large number of variants and identify disease-causing candidate variants (Figure 3–3). First, synonymous changes do not alter the translated protein and, therefore, were excluded. By contrast, nonsynonymous, stop gain/loss, splice site and INDEL alterations were prioritised (Gilissen et al., 2012). Second, alterations that were present in ExAc and dbSNP population databases with >1% MAF were excluded as these variants are unlikely to be disease-causing in a rare disorder like CHT (Boycott et al., 2013; Gilissen et al., 2012).

Because the affected individuals participating in this study were born to consanguineous parents, the disease-causing variant is expected to follow an AR inheritance pattern. That is, the disease-causing variant should be homozygous in the affected individuals and heterozygous in unaffected individuals. In Family-1 and Family-3, the affected individual and unaffected mother and father were exome-sequenced (trios); hence, variants that are homozygous in the affected member and heterozygous in both parents were prioritised. In Family-2 and Family-4, WES was conducted on an affected member and an unaffected parent and sibling; hence, AR variants were prioritised if they were homozygous in the affected individual and heterozygous in both the unaffected parent and unaffected sibling. AR candidate variants that were present in a homozygous state in any unaffected members of the other CHT families (with the same phenotype) were excluded.

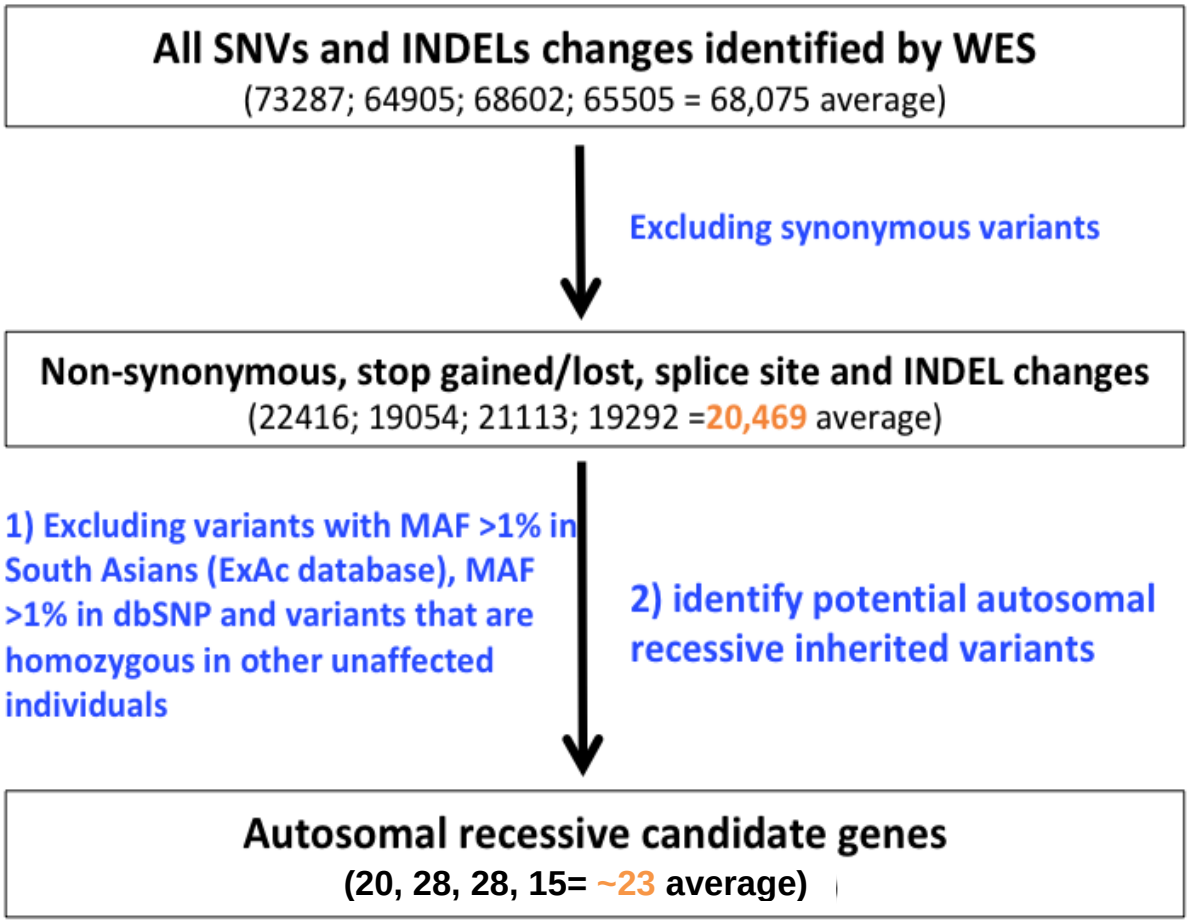


Figure 3–3: Filtering and prioritisation scheme of WES data in CHT families to prioritise potential CHT candidate variants. A step-wise process was followed to reduce the total number of identified variants (Family 1; 2; 3; 4 and average). Synonymous variants: single nucleotide substitutions that do not alter protein products; whereas, nonsynonymous changes are the ones that change the amino acid sequence of the protein. Stop gained/lost: Single nucleotide changes that lead to a premature truncation of the protein (stop gained), or the loss of translation termination site which leads to a longer protein product (stop lost). Splice-site variants: variants that are present at the intron–exon boundary (1–2 basepairs into the intron) which can change the splicing of an exon to its surrounding exons. SNV: single nucleotide variants; INDEL: small insertion/deletion; MAF: minor allele frequency; ExAc: Exome Aggregation Consortium population database.

Following this prioritisation scheme, an average of 23 (range 15–28) AR candidate genes was identified in the CHT families (complete list of candidate genes in Appendix Table 8–6). No candidate gene was commonly mutated among the three TGD families. This number of identified candidate genes required further prioritisation and investigation to pinpoint the most likely disease-related gene(s).

3.3.2.1 Ranking candidate genes based on expression data, gene biological function and animal model databases

The incorporation of computational disease-gene prioritisation analyses is essential to highlight the most promising candidate gene among a list of candidate genes identified by high-throughput technologies. In addition to filtering for rare variants that fit a surmised inheritance mode (AR variants here), integration of expression data, functional annotation of genes and biological information in animal model databases provide researchers with clues to identify the most likely pathogenic candidate gene (Robinson et al., 2014).

A ranking scheme was applied to further prioritise the potential AR candidate genes and pinpoint the most likely disease-related gene (1–5; 1, the least promising candidate; 5, the most promising) (Table 3–5). The first step in this ranking scheme aimed to prioritise genes which their biological function and annotation can be related to CHT. For example, transcription factors are normally associated with TGD families, whereas TDH is usually caused by defects in genes encoding solute carriers (Grasberger and Refetoff, 2011; Park and Chatterjee, 2005; Rastogi and LaFranchi, 2010). Therefore, genes encoding transcription factors or solute carriers are potential candidates for TGD and TDH, respectively. Second, mRNA and/or protein expression of candidate genes in thyroid tissues or in neural crest (thyroid gland develops neural crest cells during early embryological stages, more in discussion) may be of biological

relevance (Gilbert, 2000). Lastly, animal model databases can further support the candidacy of genes by translating what is learned from animal models into useful knowledge about human disease (Robinson et al., 2014). For example, a promising CHT candidate gene is expected to show CHT-related phenotype when knocked down/out in animals, therefore supporting its pathogenic role. Following this scheme, the ranking of the candidate genes identified in the four CHT families is shown in Table 3–6. All candidate genes ranked between 1–3, expect for one candidate gene in Family-1, *SIX2*, which was ranked 4. Because *SIX2* was the top ranked gene, further investigations were conducted to assess its potential candidacy in CHT.

3.3.3 The ranked 4 candidate variant: a homozygous missense change identified in *SIX2* in a TDH Family-1

A homozygous missense variant, c.859G>A; p.Val287Met, was identified in a TGD-affected patient of Family-1. This variant was heterozygous in the unaffected mother and father when visualised using the IGV (Figure 3–4). Using a pair of primers flanking the variant site followed by Sanger sequencing, the candidate variant in *SIX2* was heterozygous in unaffected parents and homozygous in the affected member, confirming the WES finding (Figure 3–5).

In silico prediction tools can be helpful to further classify variants (benign or deleterious); however, applying these predictions exclusively or at the beginning of the variant filtering scheme may lead to inaccurate biological conclusions. For instance, a study by Ng et al. (2010) demonstrated that two pathogenic variants associated with Miller syndrome were accidently ignored at first because they were computationally predicted to be non-deleterious. Therefore, these tools were only used after the potential AR candidate list was generated to avoid accidental removal of the variants of biological significance.

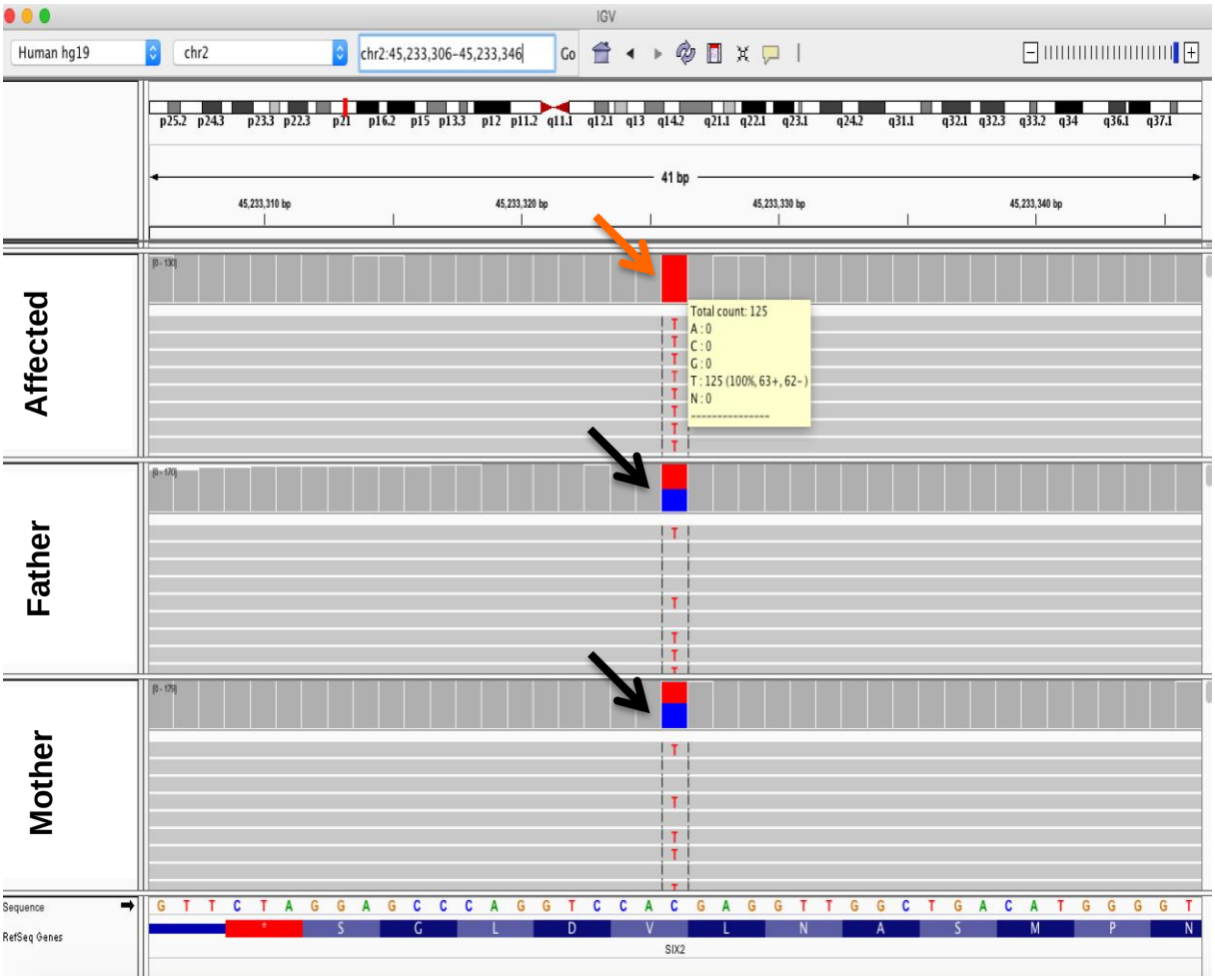


Figure 3–4: *SIX2* candidate variant visualised using the Integrative Genomic Viewer (IGV). The c.859G>A; p.Val287Met variant identified in *SIX2* is present in a homozygous state (orange arrow) in a total of 125 reads (reads are independent sequencing DNA fragments that are covering this nucleotide change). The variant is heterozygous in unaffected parents (black arrow), which is denoted by half of the reads with T allele (red square) and the other half with C allele (blue square). Image reproduced with modifications from IGV viewer).

Ranking	Criteria followed to rank AR candidate genes
1	Gene function/annotation is not relevant to CHT (see text) and there is no mRNA and protein expression in thyroid gland and/or related tissues (e.g., neural crescent) in GeneCards[1] and Protein atlas[2] database
2	Either: (1) the gene function is related to CHT; **or** (2) There is mRNA or protein expression of the gene of interest in thyroid or related tissues in one of the databases only. In protein atlas database, the protein expression level needs to be medium or high
3	Gene function is related to CHT; **and,** there is mRNA or protein expression of the gene of interest in one database only, as specified in Category 2
4	Gene function is related to CHT; **and,** both databases exhibit thyroid-related mRNA and protein expression.
5	Same as Category 4; in addition, animal model data supporting relation to CHT phenotype (Mouse genome informatics [MGI][3], International Mouse Phenotyping Consortium [IMPC][4])

Table 3–5: The ranking scheme applied to prioritise AR candidate genes identified in CHT families. 1: http://www.genecards.org 2: http://www.proteinatlas.org; 3: http://www.informatics.jax.org; 4: http://www.mousephenotype.org

Ranking	Family 1	Family 2	Family 3	Family 4
5 (most relevant)	0	0	0	0
4	1	0	0	0
3	3	4	8	4
2	11	11	13	7
1 (least relevant)	5	13	7	4

Table 3–6: The number of potential AR candidates in the four families in each ranked category. The ranking scheme is explained in **Table 3–5** and in the main text.

SIFT and PolyPhen-2 *in silico* prediction tools (Section 2.3.3.1) were used to predict the potential effect of the *SIX2* homozygous variant on the structure and function of the protein. Both tools predicted a deleterious consequence of the c.859G>A; p.Val287Met on the SIX2 protein. Moreover, the missense substitution affects a highly conserved valine residue (Figure 3–6). Mutations in highly conserved residues infer the potential importance of such mutational events (Watson et al., 2013). The SIX2 protein product is 291 amino acids in length. The p.Val287Met substitution occurs at the 287 residue, which is located in close proximity of the C-terminus.

All *SIX* gene family members (1–6) contain two highly conserved domains: (1) SIX-like homeodomain responsible for DNA binding, and (2) N-terminal SIX domain which can interact with other cofactor proteins that are unable to bind to DNA directly (Neilson et al., 2010). Although the missense candidate variant identified in this study does not reside in any of two previously mentioned conserved domains, the homozygous variant is located in close proximity of the C-terminus of the protein end and in highly conserved amino acid.

SIX2 is a member of the SIX family of homeobox genes, which encode homeodomain-containing transcription factors (Self et al., 2009). Homeobox genes regulate embryonic development by participating in cell growth and differentiation processes in multiple organisms (Self et al., 2009; Zhu et al., 2016). Functional characterisations of SIX2 highlighted the role of SIX2 in the differentiation and organogenesis of the eyes, stomach, branchial arches and kidneys in Drosophila, Xenopus and mice (Boucher et al., 2000; Gong et al., 2007; Self et al., 2006) (more about *SIX2* in Discussion).

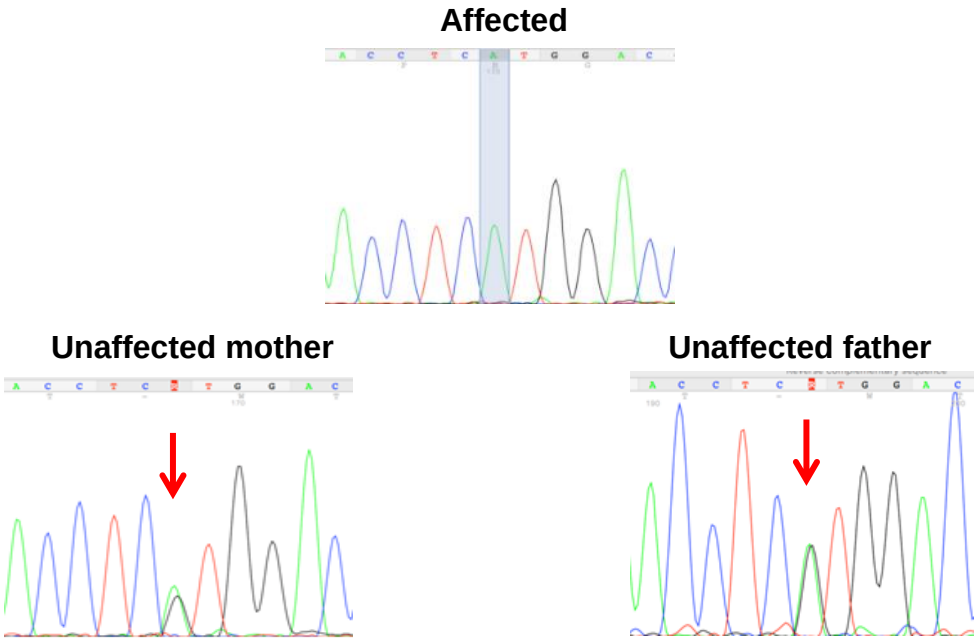


Figure 3–5: Sanger sequencing confirmation of *SIX2* variant (c.859G>A; p.Val287Met) in Family-1. The *SIX2* variant is homozygous in affected member (grey highlight) and heterozygous in unaffected parents (red arrows), confirming the WES finding. Nucleotides A, C, G and T are denoted by green, blue, black and red peaks, respectively.

Species	Amino acids sequence
***Tilapia nilotica* (Nile *tilapi*)**	IPVP----SGPDSMH----HHHSLHHDTILNPMSSNL**V**DLG
Danio rerio (Zebrafish)	IPVP----S-VDSVH----HHHSL-HDTILNPMSSNL**V**DLG
***Takifugu rubripes* (Japanese pufferfish)**	IPVP----GGAESVH----LHHSLHHDTILNPMSSNL**V**DLG
***Salmo salar* (Atlantic salmon)**	YPMTGL--GGAQPLHGMHG-HPHQLQDSLLGPLTSSL**V**DLG
***Mus musculus* (Mouse)**	YSLPGL--TASQPSHGLQA-HQHQLQDSLLGPLTSSL**V**DLG
***Xenopus tropicalis* (Western clawed frog)**	YSLSAL--SASQGGHGLQG-HQHQLQDSLLGPLTSSL**V**DLG
***Equus caballus* (Horse)**	YSLPGL--TASQPGHGLQA-HQHQLQDSLLGPLTSSL**V**DLG
***Cavia porcellus* (Guinea pig)**	YSLPGL--TASQPSHGLQA-HQHQLQDSLLGPLTSSL**V**DLG
Lagothrix lagotricha (Brown woolly monkey)	YSLPGL--TASQPSHGLQA-HQHQLQDSLLGPLTSSL**V**DLG

Figure 3–6: High conservation of the substituted valine amino acid in the *SIX2* candidate variant (c.859G>A; p.Val287Met). A valine is substituted with methionine at amino acid residue 287 which is at close proximity of the C-terminus end. The valine amino acid residue (in red) and the neighbouring amino acids are highly conserved across different species. Table compiled from information provided by PolyPhen2 (http://genetics.bwh.harvard.edu).

3.4 Discussion

NGS technologies are useful tools to study rare diseases. These technologies are increasingly being utilised in clinics to diagnose rare diseases that are difficult for physicians to diagnose so that better counselling options can be provided for families with rare diseases (Alfares et al., 2018). Both WGS and WES are used in research for identifying disease-causing genes and understanding the pathogenic mechanisms of disorders. As stated in the Introduction, WGS sequencing covers approximately 98% of the entire genome (Alfares et al., 2018), whereas WES focuses on the protein-coding subset of the genome in which alterations are considered to have severe phenotypic effects (Bertier et al., 2016). WES is less costly, generates less data output than WGS and, therefore, is considered to be a cost-effective method for identifying disease-causing variants. Herein, a homozygous missense variant in *SIX2* (c.859G>A; p.Val287Met) fitting an AR inheritance was identified in an affected individual with TGD. A detailed literature search was conducted to investigate the candidacy of *SIX2* in TGD.

3.4.1 The identification of the p.Val287Met *SIX2* candidate variant in a male patient with posterior urethral valve

Hwang et al. (2014) reported a male patient with posterior urethral valve harbouring the same p.Val287Met *SIX2* variant identified in our TGD patients, except in a heterozygous state (homozygous in our proband). Posterior urethral valve is a congenital obstructive developmental anomaly of the urethra in males (Agbugui and Omokhudu, 2015). The male patient in the Hwang et al. (2014) study was of South Asian ethnicity, matching the ethnicity of the TGD proband in this study. The authors characterised the *SIX2* variant as likely disease causing because it is rare in the general population (MAF<1%) and is predicted deleterious using two *in silico* tools and

segregating with the disease in the family. No *in vitro* or *in vivo* functional studies were performed to assess the pathogenicity of the variant. However, the classification of the *SIX2* candidate variant as disease causing in a patient with urological anomaly increases the chances of this variant being pathogenic. Although the affected member in Family-1 is a female, an increased risk of renal and urological anomalies is evident in patients with CHT.

3.4.1.1 the increased risk of congenital renal anomalies in children with CHT.

An increased risk of congenital renal and urological anomalies has been reported in children with CHT. A study by Kumar et al. (2009a) reported 54 cases with different renal congenital abnormalities in 1538 congenital hypothyroidism cases (3.5%), including kidney dysplasia, renal agenesis and ectopic kidney. Out of the 54 CHT cases with renal anomalies, one case with posterior urethral valve was reported in a male.

3.4.2 Overview of development, differentiation and morphogenesis of the thyroid gland in vertebrates

The development and differentiation of organs that arise from the pharyngeal region is complex (Xu et al., 2002). The thyroid gland originates from two key structures: (1) neural crest cells, from which the rudimentary lateral thyroid develops; and (2) primitive pharynx, in which the middle portion of the thyroid gland arises (Policeni et al., 2012; Zapanta and Shokri, 2010). In higher vertebrates, thyroid gland develops from the anterior foregut endoderm, in which four important transcription factors (*NKX2-1*/Titf1, *PAX8/Pax8*, *FOXE1*/*Foxe1* AND *HHEX/Hhex*) are expressed in thyroid progenitor cells (Fagman and Nilsson, 2011).

During morphogenesis of the thyroid, embryological tissues surrounding the developing thyroid undergo complex processes, including morphogenetic signals, that lead to the formation of the pharyngeal arches and pouches (Fagman and Nilsson, 2011). The medial anlage, which arises from the midline of the anterior pharyngeal floor amid the first and second branchial arches, consequently leads to the formation of thyroid follicular cells (Park and Chatterjee, 2005). These follicular cells develop into a thyroid bud that delaminates from the pharyngeal endoderm and migrates to its final anatomical location (with the incorporation of thyroid C-cell precursors) into the thyroid (Fagman and Nilsson, 2011). Two lateral anlagen, known as ultimobranchial bodies, develop from the fourth or fifth pharyngeal pouches that eventually become parafollicular C-cells, and these cells contribute to a substantial number of thyroid follicular cells (Park and Chatterjee, 2005).

3.4.3 *SIX2*: tissue expression, role in human disease, and biological function

3.4.3.1 Expression of *SIX2* in human and vertebrates

During human foetus development, *SIX2* RNA is widely expressed at first and second trimester gestation; however, in adult tissues, its expression is limited (Boucher et al., 2000). *SIX2* expression is observed in human foetal tissues, including the skeletal muscle (moderate), limb (moderate), adrenal/kidneys (very strong), heart (weak), liver (weak) and whole eye (strong) (Boucher et al., 2000). In the adult human, strong *SIX2* expression is observed in skeletal muscles but more weakly in the pancreas, ovaries and sclera (Boucher et al., 2000).

During mice embryonic development, the mRNA expression of *Six2* is detected after E.8.5, in which its expression is limited to the cells of the head mesenchyme and the foregut (Oliver et al., 1995). At E9.5, *Six2* expression in the head is widely spread,

possibly including non-neural derivatives of the neural crest origin such as the connective tissue of the head (Oliver et al., 1995). Moreover, Six2 expression is present in a region of the pharyngeal-oesophageal mesenchyme as well as a restricted area of the gut mesenchyme that is likely to resemble the stomach anlage. In mice, the timing of the key morphogenic events during thyroid gland development are as the follows: (1) specification at E8.5, (2) budding E10, (3) migration E10.5-13.5, and (4) follicle formation at E15.5 (Deladoey, 2012). Hence, the timing of *Six2* expression at E8.5 coincides with the early morphogenic events of the thyroid gland development.

3.4.3.2 The molecular function of *SIX2* in thyroid-related tissues

As stated earlier, the *SIX2* homeobox gene controls organ developmental and tissue differentiation processes in vertebrates and insects, including the stomach, branchial arches and kidneys. In mice, *Six2* was identified to have a role in the development of the cranial base, which has a dual embryonic origin, one of which is the cranial neural crest (He et al., 2010). *Six2*-null mice show abnormal growth and elongation of the cranial base, as well as a defective premature fusion of the bones in the cranial base (He et al., 2010). The cranial base has been shown to have a critical role in the development of mammalian thymus, parathyroid and thyroid (Xu et al., 2002). The removal of neural crest cells in chicks led to abnormalities in pharyngeal pouch derivatives, including dysplasia or aplasia of the thymus, cranial facial features and the heart (Xu et al., 2002). As mentioned earlier, the pharyngeal pouches are involved in mammalian thyroid gland development, and, consequently, interfering with these pouches may lead to defects in thyroid development.

A study by He et al. (2010) demonstrated that *Six2*-null newborn mice have a smaller thyroid cartilage and thyroid inferior horns that failed to extend as observed in wild-type mice. Because the authors aimed to detect skeletal abnormalities, the connective

tissues were digested (including the thyroid gland), and, thus, assessment of *Six2*-null mice on thyroid gland structure was not possible (Nicoletta Bobola [2016], corresponding author, personal communication).

3.4.3.3 Association of *SIX2* mutations with renal defects and assessment of renal function in Family-1

An association of *SIX2* mutations has been observed in patients with renal anomalies. Weber et al. (2008) reported the first heterozygous mutations in *SIX2* in five unrelated patients with early defects of renal hypodysplasia, including kidney dysplasia and hypoplasia. Moreover, the authors have identified, through functional experiments, a conserved role of *SIX2* in the development of the renal system. Several other studies have determined a regulatory role of *SIX2* via a complex of protein networks in mammalian nephrogenesis (Park et al., 2012; Self et al., 2009; Xu et al., 2014). For instance, *Six2*-nullzygous mice die soon after birth due to the absence of functional kidneys (Self et al., 2009).

As mentioned earlier, there is an increased risk of kidney anomalies in CHT patients. To examine the renal function in Family-1, blood testing of urea and creatine was performed in the CHT-affected proband, unaffected parents and four unaffected siblings. These tests showed normal range levels of urea (3.2-9.1 mmol/L; normal range: 2.5-8.7) and creatine (31-80 mmol/L; normal range 22-80) for all tested individuals. A kidney ultrasound test was requested for the TGD proband; however, the patient missed the appointment (Prof Timothy Barrett, Diabetes Unit, Birmingham Children's Hospital).

3.4.3.4 Possible interactions of *Six2* with *Eya1*, *Pax8* and *Nkx2.5* in mice

There is literature evidence of interactions between the *Six2* and *Eya1*, as well as *Pax8* and *Nkx2.5*—TGD-related genes in mice. Eya proteins act as transcriptional co-activators and are highly expressed in cranial placodes and branchial arches during organogenesis (Ohto et al., 1999), suggesting a major role of these proteins in the development of vertebrate organs (Xu et al., 2002). Eya1 and Six2 are normally co-localised in the nucleus of metanephric mesenchyme progenitors. However, in *Six2*$^{-/-}$ mice, Eya1 is mostly accumulated in the cytoplasm of *Six2*$^{-/-}$ nephron progenitors (Xu et al., 2014). A study by Xu et al. (2002) on mice determined a critical role of Eya1 during the morphogenesis of organs derived from the pharyngeal region, including the thyroid, parathyroid and thymus. *Eya1*$^{-/-}$ newborn mice showed a hypoplastic thyroid, which was associated with a severe reduction in thyroid parafollicular cells and a 40–60% reduction in the size of thyroid lobes. Altogether, the role of Eya1 protein in thyroid development as well as its abnormal localisation in *Six2*$^{-/-}$ mice can be associated with TGD.

In the mice metanephric mesenchyme, *Eya* genes are co-expressed with *Pax* and *Six* genes, both of which are known for their role in the development of several organs (Gong et al., 2007). In addition to highlighting a critical role of Eya genes in mammalian organogenesis, the co-expression of these genes suggests possible interactions between their gene products (Xu et al., 2002). In humans, *PAX8* is one of the critical transcription factors responsible for early thyroid development (Fagman and Nilsson, 2011). As mentioned previously, *PAX8* mutations cause TGD without other congenital anomalies (Rastogi and LaFranchi). Nevertheless, Meeus et al. (2004) presented a family (two siblings and father) that all harboured a heterozygous *PAX8* mutation and showed thyroid agenesis. In addition, the father had agenesis in one of his kidneys.

In newborn mice, the expression of *Pax8* in the developing thyroid begins at E9.0 (Fagman and Nilsson, 2011), which, as stated earlier, roughly coincides with the beginning of *Six2* expression in mice (at E8.5) and occurs before the beginning of morphogenesis of the thyroid gland (at E10.5) (Xu et al., 2002). A study by Xu et al. (2014) demonstrated a reduction and ectopic localisation in $Pax8^{+}$ cells in $Six2^{-/-}$ mutant newborn mice (personal communication, Pin-Xian Xu corresponding author of Xu , 2014) (Figure 3–7A & B). However, in $Six2^{-/-};Eya1^{+/-}$ double mutant mice, a severe reduction and ectopic localisation of *Pax8* cells was observed, compared with $Six2^{+/+}$ wild-type mice (Figure 3–7C). Although the severe *Pax8* phenotype is mainly observed in the double mutant mice, there is also an abnormal localisation and expression reduction of *Pax8* in $Six2^{-/-}$, which can possibly affect thyroid organogenesis.

Nkx2-5 and *Six2* genes are expressed in the stomach and its surrounding regions (e.g., digestive tract) in developing mice embryos (Self et al., 2009). Mouse embryos deficient in *Nkx2.5* display thyroid bud hypoplasia and cardiac defects (Fagman and Nilsson, 2011). In humans, *NKX2-5* is a transcription factor associated with TGD cases (Fagman and Nilsson, 2011) and 4% of these cases are accompanied with heart defects (Park and Chatterjee, 2005). A study by Self et al. (2009) aimed to assess the effect of inactivated *Six2* in the development of the murine digestive tract. At E18.5, many abnormalities in the stomach-related regions and pyloric sphincter were reported, highlighting the importance of *Six2* during stomach organogenesis. The authors further studied the effect of *Six2-null* on markers that may be conserved in pyloric sphincter formations in mice, one of which is *Nkx2.5*. At the E12.5 stage, compared with the wild-type mice (Figure 3–8A), a weaker and narrower expression domain of Nkx2.5 was found in the pyloric sphincter of *Six2-null* mice (Figure 3–8B). However, at E14.5, no expression difference was observed in *Nkx2.5* in *Six2-null* mice (Figure 3–8C & D). As mentioned previously, in humans, *NKX2-5* is a TGD-related

gene. Therefore, the reduced expression of *NKX2-5* in the *Six2*$^{-/-}$ murine model may affect thyroid-related development.

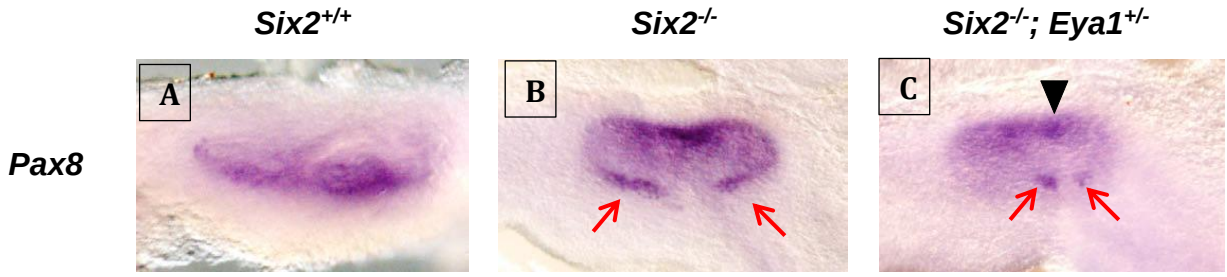


Figure 3–7: *Pax8* expression in the metanephric mesenchyme of mice embryos at E10.5-11.5. Images show whole-mount in situ hybridization for *Pax8*. *Pax8* expression in (A) *Six2* wild-type; (B) *Six*$^{-/-}$. (C) Red arrows point to the reduced *Pax8* on the medullary side of the ureteric bud branching and arrowhead (black) points to ectopic *Pax8*+ cells on the cortical side in *Six2*$^{-/-}$;*Eya1*$^{+/-}$ double mutant mice embryos. Figure reproduced with modifications from (Xu et al., 2014).

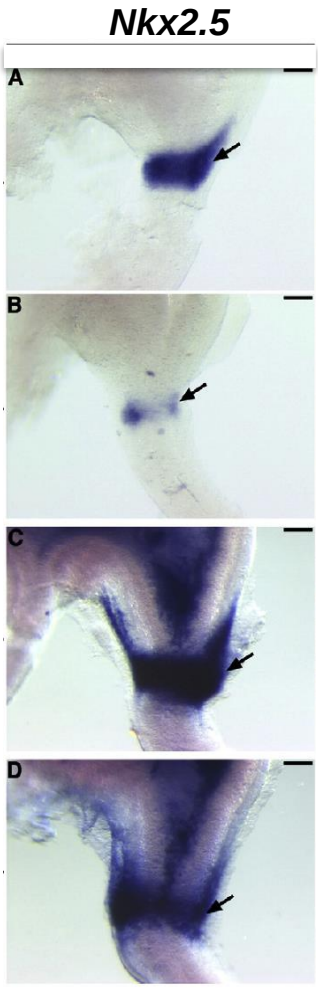


Figure 3–8: Expression of *Nkx2.5* in the murine *Six*$^{-/-}$ stomach. (A) *Nkx2.5* expression in wild type pyloric sphincter territory. (B) reduced expression and narrower expression domain of *Nkx2.5* at E12.5 (black arrow), and (C) At 14.5, the expression of *Nkx2.5* remains in the probable wild type pyloric sphincter territory. (D) Unlike B, the expression of *Nkx2.5* appears normal/unaffected. Figure reproduced with modifications from (Self et al., 2009).

3.5 Summary and conclusion

In this study, WES was conducted in four CHT-affected families, aiming to identify known or novel disease-causing genes associated with CHT. No variants in CHT disease-related genes were identified in any family.

From the WES data, a homozygous missense variant in *SIX2* (c.859G>A; p.Val287Met) was identified in a TGD-affected proband. This variant was predicted deleterious by two *in silico* tools. Using Sanger sequencing, the *SIX2* variant was homozygous in the affected member and heterozygous in the unaffected parents, thus confirming the WES data. The same *SIX2* candidate variant identified in this study was reported in the literature as disease causing in a patient with a urinary tract anomaly (but in a heterozygote state). An increased risk of congenital kidney anomalies is observed in children with CHT; therefore, this observation increases the likelihood of *SIX2* variant being of biological relevance.

A thorough literature search was conducted to investigate the candidacy of *SIX2* in TGD. TGD phenotype is usually caused by defects in transcription factors. *SIX2* is a transcription factor involved in the development and differentiation of multiple organs. *SIX2* is expressed in neural crest derivative tissues, which have a role in thyroid gland development. During mice embryonic development, the beginning of *Six2* expression coincides with the beginning of thyroid gland differentiation, highlighting the potential involvement of *Six2* in thyroid development.

Functional characterisation studies on mice and rodents have identified interactions between *Six2* and *Eya1*, *Nkx2.5* and *Pax8* genes, all of which have been associated with TGD phenotype in mice or humans. Therefore, interactions of *Six2* and the aforementioned TGD-associated genes may indirectly associate with TGD.

Taken together, existing literature evidence supports the potential candidacy of *SIX2* in TGD. Further biological studies were performed to assess the pathogenicity of *SIX2* candidate variant and its involvement in TGD (Chapter 4).

Chapter 4: Investigations into *SIX2* candidacy in thyroid gland dysgenesis using functional experiments and family segregation analysis

4.1 Introduction and aims

NGS technologies are useful tools to study disease-causing genes in families with rare genetic diseases. The application of NGS allows for a comprehensive analysis of genes in the entire genome; thus, sequencing capacity is no longer a limitation in disease-causing identification studies. However, due to the large data output of these technologies, interpreting genotype-phenotype correlations can be challenging (Goodwin et al., 2016). Even after prioritising a single candidate gene, biological assessments and family segregation analysis are required to establish a genotype-phenotype correlation and , in turn, make an accurate genetic diagnosis and provide the best counselling options for the patient's family (Bertier et al., 2016).

In Chapter 3, using WES, a homozygous missense variant (c.859G>A; p.Val287Met) in *SIX2* gene was identified in a patient with TGD, born to consanguineous parents, in Family-1 (Family-1-A). The candidate variant was heterozygous in the unaffected parents, fitting an autosomal recessive inheritance pattern. A detailed literature search provided evidence for the potential candidacy of *SIX2* in TGD. *In silico* tools predicted a damaging effect of the variant on the SIX2 protein structure and function. Although *in silico* bioinformatic tools are useful in predicting the potential pathogenic consequence of a variant, functional characterisation experiments are needed to precisely determine the pathogenic mechanisms of a candidate variant and avoid false biological conclusions (Leung et al., 2018).

Functional characterisation experiments assess the functional consequences of genetic variants at the molecular level, including gene expression, protein products

and protein interactions (Yang, 2012; Zhang et al., 2012). Coding SNVs can substitute an amino acid with another (i.e., missense changes) or introduce a premature stop codon, producing a truncated and likely non-functional protein product (Yang, 2012). Missense changes are considered the most common type of disease-related genetic variations (Bamshad et al., 2011). The *SIX2* candidate variant is a missense change that substitutes a valine for a methionine at codon 287. Although missense variants do not alter the length of the protein product, these changes have been extensively studied to determine their effect on protein structure, stability, folding, interactions and subcellular localisation (Zhang et al., 2012).

To characterise the *SIX2* variant *in vitro*, three initial experiments were planned: (1) an overexpression model to assess protein expression by western blotting using two overexpression vectors containing either an ORF of the *SIX2* wild-type (*SIX2*-WT) or the homozygous candidate variant (*SIX2*-Mut); (2) protein localisation assessment of *SIX2*-WT versus *SIX2*-Mut; and (3) transient silencing of *SIX2* in thyroid progenitor cells using small interfering RNA (siRNA) to investigate the potential effect of *SIX2*-Mut versus *SIX2*-WT on mRNA expression of thyroid-related genes.

DNA from additional unaffected family members of Family-1 was requested to perform family segregation analysis of the *SIX2* candidate variant. With variant segregation analysis, a genotype-phenotype correlation can be determined and, in turn, a causal disease-gene relationship can be established. A decision was made to initially start the first two functional experiments while waiting for the DNA of the additional family members to become available.

4.2 Cloning and site-directed mutagenesis of *SIX2*

To characterise the *SIX2* candidate variant *in vitro*, two mammalian overexpression vectors, a tagged pFlag-CMV-4 and an untagged pcDNA3.1+-CMV-4, containing the

SIX2-WT or *SIX2*-Mut sequences were used. Because the Flag-tagged vector contains a polypeptide protein tag that may alter the protein folding, an untagged pcDNA3.1+-CMV-4 was used as an additional control vector. The *SIX2*-Mut ORF containing the homozygous candidate variant was generated using SDM (see Sections 2.7.1 & 2.7.2). The *SIX2*-WT and *SIX2*-Mut ORFs were amplified, purified and verified by Sanger sequencing and sub-cloned by ligation into pFlag-CMV-4 and pCDNA3.1+-CMV-4 vectors using specifically designed primers and restriction enzymes (see Sections 2.7.3, 2.7.42.7.5).

4.3 Results

4.3.1 Experiment 1: protein expression assessment by western blot

The first experiment aimed to assess the protein expression of the *SIX2*-generated plasmids using WB. HEK293 cells grown to ~80% confluency were transiently transfected with purified pFlag-CMV-4 or pcDNA3.1+-CMV-4 containing either *SIX2*-WT or *SIX2*-Mut ORFs (see Section 2.9). Empty plasmid vectors containing no ORF were used as a control. HEK293 cells were selected because they can be easily grown, manipulated and transfected. To account for potential variations in transfection efficiency, as a control, a HA-tagged-*JMJD4* plasmid (courtesy of Charlotte Eaton, Institute of Cancer and Genomic Sciences, University of Birmingham) was co-transfected in 1:10 ratio with the above-mentioned plasmids. Following the 48-hour transfection period, transfected cells were harvested and purified in RIPA lysis buffer (Section 2.9.1) and quantified using Wallac Victor3 fluorometer (Section 2.10.1). The samples were then electrophoresed on SDS-PAGE gels and western blotted (Sections 2.10.2 & 2.10.5). Protein blots were quantitatively measured using GeneSnap/GeneTools software (Section 2.10.6). Three independent transfection

experiments were performed to account for the potential influence of experimental/technical variability.

4.3.1.1 Results of Experiment 1: western blotting

The SIX2 proteins in pFLAG-CMV-4 and pcDNA3.1-CMV4 was detected using anti-FLAG (Sigma) and anti-SIX2 (Proteintech) antibodies, respectively. The measurement of β-actin, a housekeeping gene, was performed using an α-β-actin antibody (Sigma) to correct for variance in gel protein loading. The protein blots of pFLAG-CMV4/Empty vector, pFLAG-CMV4/*SIX2*-WT and pFLAG-CMV4/*SIX2*-Mut are illustrated in Figure 4–1. Both pFLAG-CMV4 *SIX2*-WT and *SIX2*-Mut plasmids showed protein expression immunoblots; however, no apparent difference in expression was observed between the two plasmids. Similar protein expression results of pcDNA3.1-CMV4/*SIX2*-WT or *SIX2*-Mut plasmids were also observed (data not shown). Quantification of the immunoblots of Flag-tagged *SIX2*-WT and *SIX2*-Mut plasmids shows very similar expression values (Table 4–1). Although no difference in protein expression was observed between the two plasmids, this finding does not rule out the possibility of a difference in the folding or stability of produced proteins.

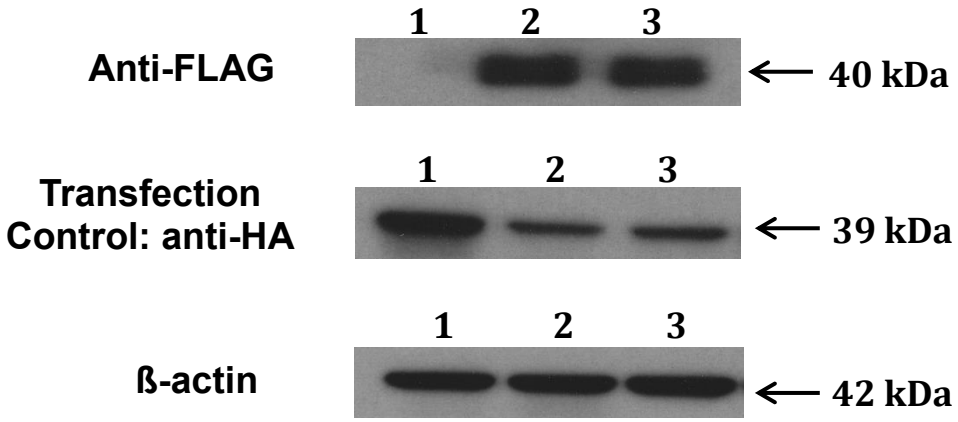


Figure 4–1: Protein expression of *SIX2*-wild type and *SIX2*-Mut plasmids. 1: pFLAG-CMV-4/empty vector; 2: pFLAG/*SIX2*-wildtype; 3: pFLAG/*SIX2*-mutant (containing the homozygous candidate variant). SDS-PAGE immunoblot expression analysis of samples probed with anti-FLAG, anti-HA and anti- β-actin. Expected sizes of the blots are on the right. The immunoblots of HA-Transfection control vectors show very comparable transfection levels.

Plasmid	Mean relative expression, arbitrary units (n=3)
pFLAG-CMV-4/Empty	0.714
pFLAG-CMV-4/SIX2-WT	5.422
pFLAG-CMV-4/SIX2-MuT	5.645

Table 4–1: Mean of quantified protein expression of *SIX2*-Wildtype and *SIX2*-Mut plasmids from three independent western blot experiments. Protein expression from the blots was quantified using GeneSnap tool. *SIX2*-WT plasmid consisted of the *SIX2* wild type ORF, whereas *SIX2*-Mutant plasmid contained the *SIX2* homozygous candidate variant.

4.3.2 Experiment 2: subcellular localisation of the *SIX2* candidate variant versus *SIX2* wild type using immunocytochemistry

The second experiment aimed to assess the subcellular localisation of the pFLAG-CMV4 containing *SIX2*-WT or *SIX2*-Mut plasmids using immunocytochemistry (see Section 2.11). Correct subcellular localisation of a protein is an important factor to maintain protein cellular function and interactions with biological partners required in signalling pathways (Zhang et al., 2012). Thus, mislocalisation of a protein can have a detrimental effect on the protein's function and biological interactions. The full-length SIX2 protein is primarily found in the nucleus, which is expected since *SIX2* is a transcription factor involved in regulating the transcription of other genes (Brodbeck et al., 2004). If the *SIX2*-Mut fully or partially mislocalises to a cellular compartment other than the nucleus, then it can be hypothesised that the *SIX2* candidate variant affects SIX2 subcellular localisation.

A series of preliminary studies were performed to optimise appropriate experimental conditions, including the appropriate antibody concentration, number of seeded cells and transfection conditions. COS7 cell line was selected as the cell line of choice because they are larger and rounder than HEK293 cells and can be readily transfected and manipulated. In an 8-well chamber glass slide, COS7 cells were seeded and then transfected with pFLAG-CMV-4 vectors containing *SIX2*-WT or *SIX2*-Mut ORFs (see Section 2.9). AlexaFluor-488-conjugated FLAG-tag monoclonal antibody was used to detect FLAG-tagged SIX2 protein (Figure 4–2). Proteins produced by both *SIX2*-WT and *SIX2*-Mut vectors were localised to the nucleus.

In conclusion, no mislocalisation of FLAG-tagged SIX2 protein produced by *SIX2*-Mut plasmid was evident. Although WB and immunocytochemical experiments did not show any potential deleterious effect of the *SIX2* variant *in vitro*, additional experiments may highlight its potential biological relevance.

4.3.3 Availability of other unaffected family members of CHT Family-1 under study

The DNA of four unaffected members of Family-1 became available to investigate the segregation of the *SIX2* homozygous candidate variant. Previously, using WES and Sanger sequencing, the *SIX2* candidate variant was confirmed to be homozygous in Family-1-A and heterozygous in both unaffected parents. To confirm the segregation of *SIX2* variant with the TGD phenotype, it was hypothesised that the candidate variant should not be present in a homozygous state in unaffected family members.

Using Sanger sequencing, the candidate variant was assessed in the unaffected family members and was found in a homozygous state in one sister (Family-1-S2), heterozygous in two sisters (Family-1-S1 & S3) and homozygous wild type (no variant) in one brother (Family-1-B) (Figure 4–3). According to previous clinical notes, none of the four unaffected family members showed any CHT phenotype. However, to rule out CHT diagnosis, thyroid function blood tests were requested for these family members. In the four unaffected members, blood tests showed normal levels of TSH ranging from 1.52–2.47 mU/L (normal range, 0.5–5.0) and FT4 ranging from 13.5–14.5 pmol/L (normal range, 9–20.9). Therefore, the clinical notes and thyroid function tests suggests the absence of clinical and biochemical signs of CHT in the four unaffected family members. In conclusion, the homozygous *SIX2* candidate was identified in an unaffected family member and therefore is very unlikely to be TGD disease causing

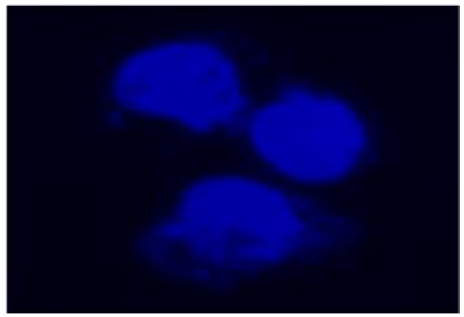
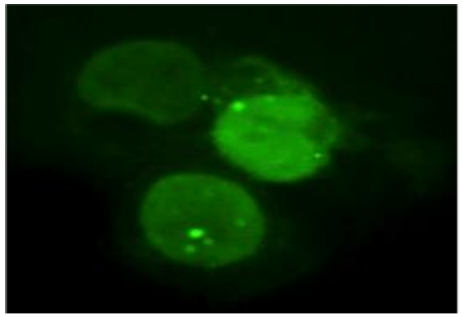
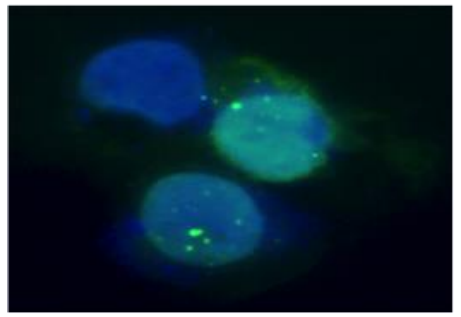
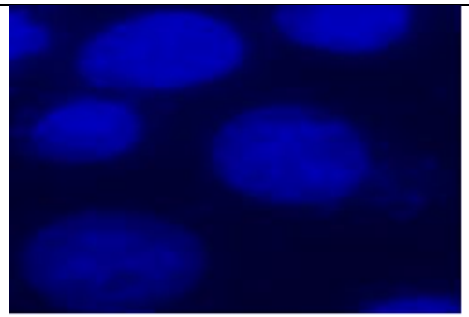
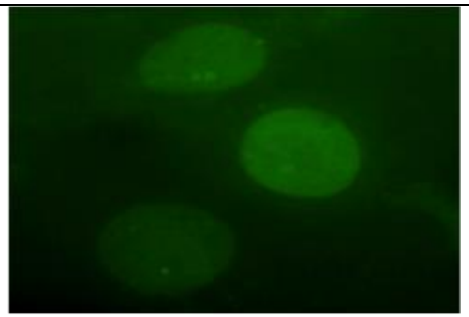
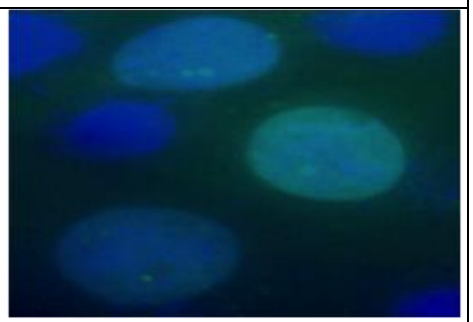
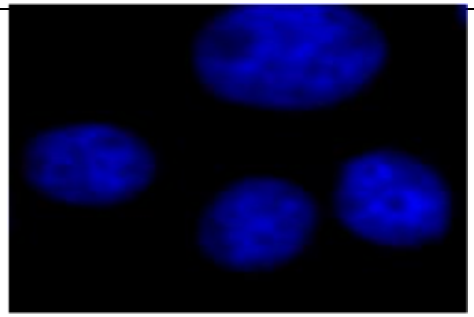
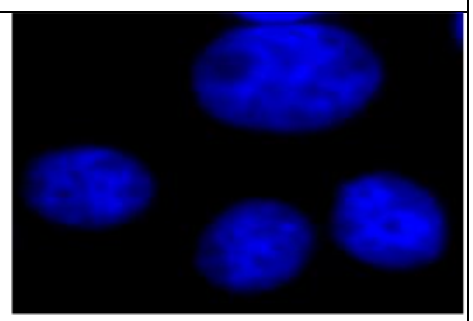
Plasmid	Blue/DAPI	Green/anti-FLAG	Merged
pFLAG/*SIX2*-WT			
pFLAG/*SIX2*-Mut			
Untransfected (control)			

Figure 4–2: COS7 cells fixed and stained with anti-FLAG conjugated antibody (x40 mag), contrast adjusted. In COS7 transfected cells), FLAG-tagged SIX2 proteins produced by *SIX2*-Wildtype and *SIX2*-Mutant (containing *SIX2* candidate variant) plasmids were primarily detected in the nucleus in both the. The extremely bright green spots in anti-FLAG *SIX2*-WT Green and Merged images are staining artefacts (confirmed by Dr Daniel Fulton, Neuroscience and Ophthalmology department, University of Birmingham).

4.3.3.1 Assessing the segregation of rank-3 candidate genes in CHT Family-1

In Chapter 3, using WES, three (*AIRE*, *THADA*, *ZC3H14*) AR candidate genes that were ranked-3 were identified in Family-1-A (heterozygous in unaffected parents and homozygous in affected index) (see Sections 3.3.2.1 & 3.3.3). Because the *SIX2* candidate variant was unlikely be TGD disease causing, the segregation of these three candidate genes was analysed in the four family members using Sanger sequencing mutational analysis. All ranked-3 candidate variants were identified (in homozygous state) in the unaffected Family-1-S2 member, suggesting that these variants are also very unlikely to cause TGD (Table 4–2). Altogether, Family-1-A and unaffected Family-1-S2 members show the same genotype of rank-3 and -4 candidate variants.

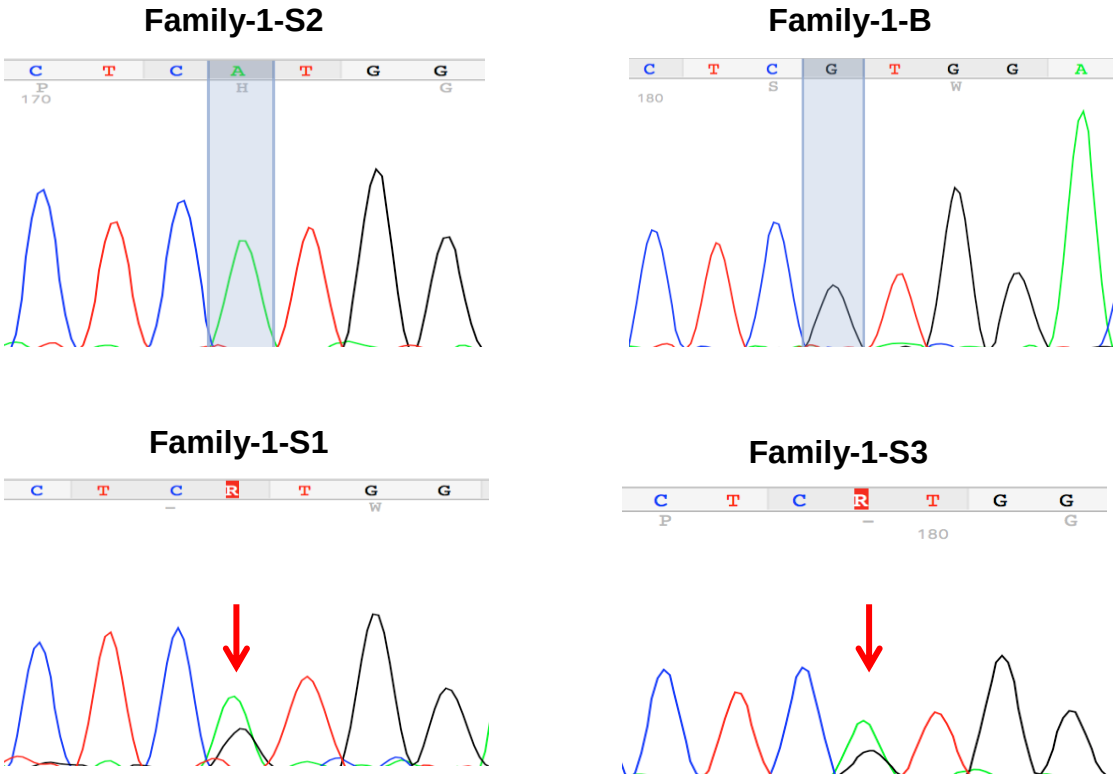


Figure 4–3: Sanger sequencing analysis of *SIX2* candidate variant in four unaffected members of Family-1. The *SIX2* candidate variant (c.859G>A) was identified as homozygous in an unaffected sister (Family-1-S2), homozygous wild type (no change) in an unaffected brother (Family-1-B) and heterozygous in two unaffected sisters (Family-1-S1 & S3) (shaded base pairs and red arrows).

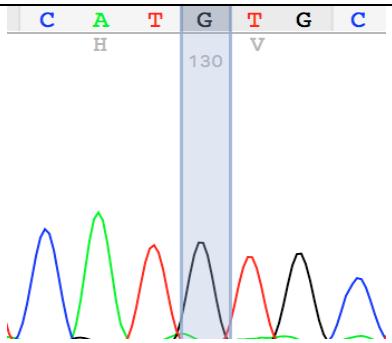
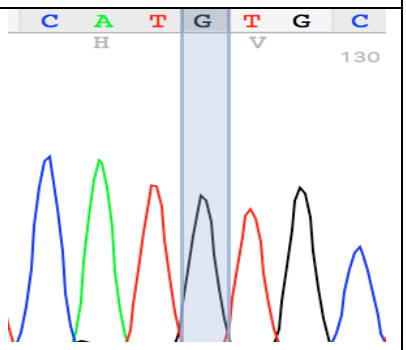
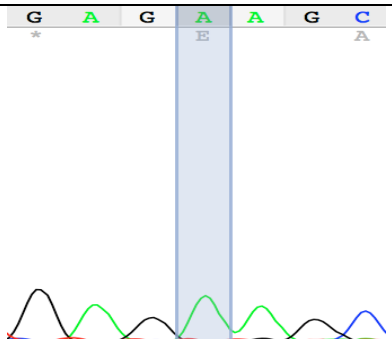
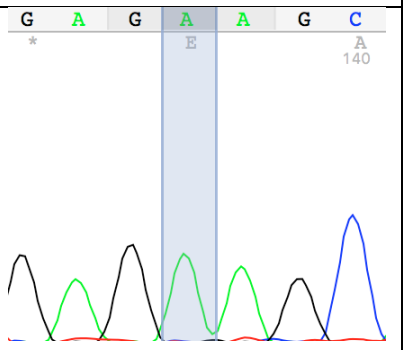
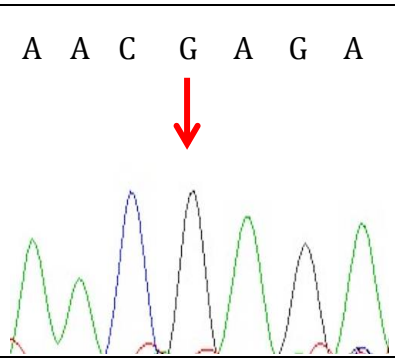
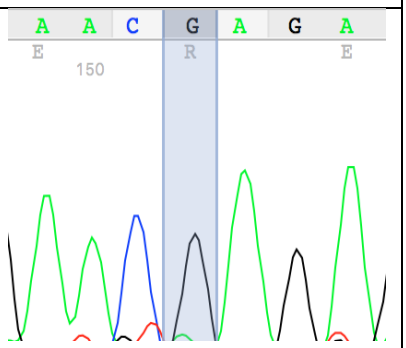
Gene	Change	Family-1-A	Family-1-S2
AIRE	c.5221G>A; p.Glu1741Lys		
THADA	c.788C>G; p.Pro263Arg		
ZC3H14	c.927C>G; p.Ile30 9Met		

Table 4–2: Sanger sequencing analysis of rank-3 candidate genes in TGD-affected (Family-1-A) and an unaffected sister (Family-1-S2). In the *AIRE*, *THADA* and *ZC3H14* genes, the homozygous candidate variants detected in the affected individual were also identified in an unaffected sister. Note: the sequencing image of *ZC3H14* for Family-1-A was taken using the Chromas trace viewer tool, whereas all other images were visualised using the 4Peaks tool.

4.4 Discussion

NGS technologies are powerful tools for identifying genes associated with RGDs. Since the introduction of these technologies in 2009, they have caused an exponential increase in the number of disease-causing genes. As mentioned earlier, WES technology analyses the protein-coding regions, representing ~1.5% of the entire human genome. Although the data output of WES is considerably lower than in WGS, finding the disease-causing gene among thousands of identified variants requires extensive data filtering and prioritisation steps (Gilissen et al., 2012). Moreover, functional characterisation experiments (*in vitro* and *in vivo*) are required to establish an association of the candidate gene with the disease phenotype.

In the current study, the *SIX2* gene identified in Family-1 was classified as a potential candidate gene for TGD. Two biological characterisation experiments and variant segregation analysis in the family were performed to establish a causal disease-gene relationship. The *SIX2* homozygous missense variant did not co-segregate with the disease in the family and, thus, was unlikely to be TGD disease causing.

As mentioned in Chapter 3, TGD is primarily sporadic in occurrence, and only a minority of patients (~2%) show genetic defects in a small subset of genes that generally encode for transcription factors. Monogenic, polygenic and multifactorial factors (e.g., environmental/epigenetics) have been proposed to contribute to TGD pathogenicity (Sun et al., 2018). With such complex disease aetiology, it is not surprising that the rank-3 and -4 variants were classified as unlikely disease causing. The number of families studied here is relatively small (three TGD; one TDH). If more families had been studied, the chances of identifying known or novel CHT-related genes would have been greater.

A study by de Filippis et al. (2017) aimed to assess the potential involvement of oligogenic factors in 177 CHT patients (94 gland-in-situ; 83 TGD). The study showed an oligogenic involvement in 22% of total cases, confirmed by the co-segregation of a sum of rare variants (MAF<0.01) with CHT phenotype in several families. The frequent oligogenic origin of CHT observed by the authors provides a reasonable justification for the complex aetiology of the disease. Assuming an oligogenic involvement in Family-1 of this study, rank-3 and rank-4 variants also failed to co-segregate with the disease. In the Families 1–3, it may be worth assessing the involvement of oligogenic variations in CHT, which would likely require additional DNA from close or extended family members to perform variant segregation analyses.

Monogenic defects harbouring biallelic loss-of-function mutations have been primarily described in families with CHT (Park and Chatterjee, 2005; de Filippis et al., 2017). Heterozygous mutations in thyroid-related transcription factors have also been identified in a few TGD cases, yet with variable penetrance and expressivity (Al Taji et al., 2007; Park and Chatterjee, 2005). Although no heterozygous variants in CHT-known genes co-segregated with the disease in the three TGD families in this study (see Section 3.3.1), rare heterozygous candidate variants in novel genes may be investigated. However, the number of heterozygous changes is expected to be greater than homozygous variants; hence, heterozygous variants would require thorough filtering approaches.

To date, a clear molecular mechanism explaining the majority of TGD has yet to be elucidated. A recent review by Abu-Khudir et al. (2017) proposed a two-hit hypothesis to explain the aetiology of TGD, consisting of first inheriting a germline variant predisposing to TGD, followed by a second acquired somatic change during early post-zygotic divisions or in tissue-specific genes involved in thyroid morphogenesis.

However, there is still no solid literature evidence confirming these mechanisms (Abu-Khudir et al., 2017). A study by Magne et al. (2016) compared WES data of thyroid tissue to matched leukocytes in two patients with TGD and did not observe somatic mutations in their ectopic thyroids. An earlier study by Abu-Khudir et al. (2010) assessed the methylation status of ectopic thyroids (n=2) versus normal thyroid (n=4) and found no epimutation in ectopic thyroids. Hence, further studies are needed to confirm the two-hit hypothesis.

Although genes are usually biallelically expressed, autosomal monoallelic expression (AME) (one copy) of some genes is evident (Magne et al., 2016). AME with random choice between paternal and maternal alleles represents an uncommon class of disease-causing genes. Mechanisms causing monoallelic expression include random AME or gene imprinting, in which genes are preferentially expressed from either the paternal or the maternal alleles (Hermanns et al., 2011). The sporadic nature of TGD argues for the possible involvement of altered autosomal somatic gene expression in thyroid gland diseases (including TGD). Hermanns et al. (2011) hypothesised that monoallelic expression of *PAX8* in thyroid tissue may be associated with a familial case of athyreosis (complete absence of thyroid tissue). Another study by Magne et al. (2016) investigated genes that were subjected to AME in ectopic (n=2) or orthotopic (n=4) thyroids. The authors found an average of 22 genes subjected to AME, which may account for the discordance of CHT in monozygotic twins, as well as the sporadic occurrence of CHT cases. The study was performed on a small number of ectopic thyroids, and therefore, a larger TGD cohort is needed to clearly ascertain this association of AME. However, the difficulties of obtaining ectopic thyroid tissues from asymptomatic patients need to be considered.

Additional subsequent experiments were also planned (if PhD time permitted): (1) assessment of protein stability by measuring the relative differences in protein half-lives of *SIX2*-WT versus *SIX2*-Mut; and (2) measurement of mRNA expression levels using real-time (quantitative) PCR of cells transfected with *SIX2*-WT versus *SIX2*-Mut. However, these experiments were not performed due to the failure of *SIX2* candidate variant to segregate with the disease in Family-1 and limited PhD time.

4.5 Chapter summary

In Chapter 3, a homozygous missense variant in *SIX2* was identified in a TGD patient (Family-1-A), fitting an autosomal recessive inheritance. *In silico* tools and an in-depth literature review supported the potential candidacy of the *SIX2* variant in TGD pathogenicity. *In vitro* functional characterisation experiments were performed to assess the pathogenicity of the candidate variant. The first experiment assessed the protein expression of *SIX2*-WT and *SIX2*-Mut plasmids transfected to HEK293 cells; no protein expression differences were observed between the two plasmids. In the second experiment, the subcellular localisation of SIX2 proteins produced by *SIX2*-WT and *SIX2*-Mut constructs was investigated in COS7-transfected cells using immunocytochemistry analysis. As a transcription factor, SIX2 proteins produced by both plasmids were localised to the nucleus, and hence, no mislocalisation in *SIX2*-Mut was evident.

The DNA of four additional unaffected family members became available, allowing for assessment of the segregation of the *SIX2* variant in the entire family. The homozygous *SIX2* variant was identified in an unaffected Family-1-S2 member, suggesting that the variant is very unlikely disease causing. Afterward, the segregation of three rank-3 genes was assessed in the unaffected members and were also

identified as homozygous in Famil-1-S2 unaffected member. Altogether, the homozygous variants in *SIX2*, *AIRE*, *THADA*, and *ZC3H14* genes are very unlikely to be TGD disease-causing.

4.6 Future work

In Chapter 3, we primarily aimed to detect small insertions/deletions and point mutation changes using WES data. Larger structural genomic CNV are another form of genetic alterations that can be disease causing. CNV are preferentially studied using aCGH technology; however, WES data can be utilised to identify CNV alterations (Sathirapongsasuti et al., 2011). Due to PhD limited time, this analysis was not performed; however, it could be conducted in the future.

Because WES focuses on protein-translated regions of the genome, variants in gene promoter and intronic regions will be missed as these regions are not targeted by WES. A request has been made to involve the four CHT families of this study in the UK's 100,000 Genomes Project, which performs WGS on families with rare disorders and cancers. WGS provides a uniform sequencing coverage of the entire genome and therefore will detect variants in promoter and intronic regions, as well as exonic regions that are not uniformly covered by WES. A larger CHT cohort involving sequenced families across the UK such is the UK's 100,000 Genomes Project can be very helpful in establishing causal disease-gene relationships.

Chapter 5: Genetic and transcriptomic profiling of undifferentiated pleomorphic sarcoma of bone using NGS technologies

5.1 Introduction

5.1.1 UPSb tumours

UPSb is a rare type of bone sarcoma that represents 2–5% of all primary bone tumours (Niini et al., 2011; ESMO/European Sarcoma Network Working Group, 2014). The tumour normally develops in long bones of the lower extremities, commonly the femur followed by the tibia and pelvis. UPSb is classified as high-grade with a metastatic rate of at least 50%, especially to lungs (Doyle, 2014; Niini et al., 2011). UPSb usually occurs in older patients and usually affects males more than females (ESMO/European Sarcoma Network Working Group, 2014). During clinical examination, these patients generally report pain and an increased occurrence of fracture. The recommended treatment of UPSb usually involves neoadjuvant therapy followed by wide excision (Gerrand et al., 2016).

The morphological characteristics of UPSb tumours consist of atypical spindled and pleomorphic cells that can be arranged in a storiform, fascicular or haphazard form. Notably, chondroid and osteoid matrix mineralisations are absent in these tumours (Chen et al., 2017a). UPSb shares similar clinical presentation with dedifferentiated chondrosarcoma, such as arising in similar anatomical sites. Although both UPSb and dedifferentiated chondrosarcoma are aggressive tumours, the latter is extremely aggressive with a poorer prognosis (Christopher et al., 2013). The majority of patients with dedifferentiated chondrosarcoma die within two years of initial diagnosis, whereas patients with UPSb have a median survival time of approximately 5 years (Chen et al.,

2017a). For patients with advanced dedifferentiated chondrosarcoma, conventional chemotherapies have very limited efficacy, so complete surgical excision is the preferred treatment method. By contrast, neoadjuvant or adjuvant chemotherapy is more beneficial in patients with UPSb with an increased 5-year survival rate of 59% (Chen et al., 2017a).

UPSb diagnosis is one of exclusion; hence, thorough sampling is required for accurate pathological diagnosis (Doyle, 2014). Due to the morphological similarities between dedifferentiated chondrosarcoma, osteosarcoma and UPSb, pathological diagnosis of UPSb can be challenging for pathologists and clinicians (Chen et al., 2017a; Niini et al., 2011). Sometimes a UPSb tumour is diagnosed as an osteosarcoma or dedifferentiated chondrosarcoma after resection (Gerrand et al., 2016). UPSb can be distinguished from an osteosarcoma by the absence of malignant osteoid deposition (Romeo et al., 2012). However, in a limited biopsy sample, differentiating between these bone lesions is challenging. Therefore, patients would benefit from a molecular diagnostic method that can differentiate between UPSb tumours and other bone sarcoma subtypes. Unfortunately, definitive diagnostic testing to differentiate between these lesions has yet to be established.

The genetics of UPSb is not fully understood. Complex karyotypes (Christopher et al., 2013) and low frequency *TP53* mutations have been reported in UPSb (Kawaguchi et al., 2002). A study by Niini et al. (2011) identified complex copy number karyotypes in most UPSb samples with frequent losses in 9p22-pter and 13q21-q22 regions and frequent gains in 18q12-q22; 1q21-q23, 6p21.1, 7p12-pter, 7q22-q31, 8q21.3-qter and 9q32-qter regions. Of the 20 patients with UPSb, 8 patients exhibited homozygous deletions of *CDKN2A*; 7 patients, *RB1*; and 3 patients, *TP53*. Furthermore, in study by

Sarhadi et al. (2014) that assessed CNVs in chromosome 9, heterozygous and homozygous deletions in *CDKN2A* were found in UPSb tumours.

5.1.2 NGS technologies in the field of cancer

NGS technologies have been extensively used to understand the molecular mechanisms of tumourigenesis and identify potential drug-gene targets that can be incorporated into clinical practice (Bertier et al., 2016). As mentioned earlier, WES and RNA-Seq techniques provide a comprehensive high-throughput analysis of coding exons and transcriptomic alterations, respectively, that can identify genetic changes contributing to cancer malignancy (Hardwick et al., 2017). The increased practicality and decreased cost of WES in recent years have enabled researchers to study genomic alterations, including INDELs and SNVs, in protein coding regions (Bertier et al., 2016). RNA-Seq can be used to identify gene fusion transcripts that arise from genomic structural alterations in tumours. Gene fusions have attracted many researchers in the cancer field due to their potential pathogenic role in cancer development and applications as molecular and diagnostic markers when recurrently identified in a specific tumour subtype (Byron et al., 2016).

5.2 Aims of the study

To our knowledge, no exomic or transcriptomic profiling of UPSb tumours has been conducted using NGS technologies. Moreover, no molecular diagnosis criteria have been established to improve the diagnosis of UPSb. The primary aim of this project was to investigate the genetic and transcriptomic landscape of UPSb using WES and RNA-Seq. First, ten normal-paired and four normal-unpaired (tumour only) UPSb

samples were exome-sequenced to find exonic genetic changes, including INDELs and SNVs. Second, eight UPSb tumours were RNA-sequenced to detect gene fusions. Comprehensive assessment of the genetic alterations landscape in UPSb tumours will expand the understanding of their pathogenicity, leading to development of accurate molecular diagnosis tests and personalised treatments.

5.3 WES results and data analysis

5.3.1 WES samples information and somatic variant calling tools

A total of 24 samples (14 UPSb tumours and 10 corresponding normal samples) were exome-sequenced. The clinicopathological information of all tumours were reviewed and confirmed by Dr Sumathi Vaiyapuri (The Royal Orthopaedic Hospital NHS Foundation Trust, University of Birmingham) (see Section 2.1.2).

Nine matched normal-tumour pairs (UPSb-T1 to UPSb-T9) and three tumour-only (normal-unpaired) samples (UPSb-T10 to UPSb-T12) were exome-sequenced by OGT Company. In addition, one matched normal-tumour pair (UPSb-T13) and one tumour-only (UPSb-T14) sample were exome-sequenced, courtesy of Prof Eamonn Maher's laboratory (University of Cambridge, UK). All tumour DNA was extracted from fresh frozen cancer tissues, except for one FFPE tumour (UPSb-T12). By contrast, DNA from corresponding normal samples was obtained from tumour-free muscle tissue adjacent to the tumour site (see Section 2.2). All DNA samples were quality checked (see Section 2.2.4).

WES data from aligned reads (hg19 reference assembly) were received in BAM file format (see Section 2.3.4). In all normal-matched tumours and tumour-only samples, SNVs and INDELs variants were identified using VarScan2 and MuTect somatic

variant callers (Section 2.3.5.2). As mentioned earlier, VarScan2 analysis of UPSb-T1 to UPSb-T12 samples were performed by OGT Company. For UPSb-T13 and UPSb-T14, VarScan2 tool was run by Source BioScience Company (Nottingham, UK). For all UPSb tumours, MuTect tool was run by our collaborator, Dr David Huen (School of Biology, Chemistry and Forensic Science, University of Wolverhampton).

5.3.2 Quality check metrics of WES data using FastQC tool

QC metrics of raw WES data is one of the first essential steps in the data analysis pipeline. The QC step provides an overview of the sequencing quality and identifies any bias with the data that can affect subsequent data analysis or lead to false biological conclusions. Using either BAM or FastQ file inputs, QC metrics of the WES data were generated using the FastQC tool to spot issues in either the starting library material or the sequencing itself (details in Appendix Section 8.1). An average of 95.4% targeted bases with 10X coverage was achieved in UPSb samples (for QC metric details see Appendix Table 8–2).

Overall, all UPSb samples showed satisfactory QC scores with minimal bias across all samples, except for the FFPE tumour that showed slightly suboptimal QC results. This finding was expected since DNA extracted from FFPE tissue is generally degraded and in smaller sized fragments (<300 bp). In addition, formalin fixation of tissues introduces DNA-protein crosslinks, which cannot be completely reversed during the DNA extraction process (Dietrich et al., 2013).

5.3.3 *IDH1/2* hotspot mutations

IDH1 and *IDH2* genes encode isocitrate dehydrogenases 1 and 2 enzymes, respectively, both of which act as reaction catalysts in many cellular processes. R132 and R172 heterozygous hotspot mutations in *IDH1* and *IDH2*, respectively, have been identified in many cancers, such as gliomas/secondary glioblastomas (Yan et al., 2009) and acute myeloid leukaemia (Abbas et al., 2010). Moreover, *IDH1/2* mutations were detected in 87% (20/23) of dedifferentiated chondrosarcoma (Chen et al., 2017a) but were not identified in 222 osteosarcomas (Amary et al., 2011).

To investigate *IDH1/2* mutations in our UPSb samples, BAM files were checked for R132 and R172 hotspot mutations. Adequate coverage of at least 20 reads for both hotspot changes was achieved in all UPSb tumours. No mutations in R132 nor R172 hotspots were identified in any of the 14 UPSb samples. The absence of *IDH1/2* mutations in these UPSb tumours ruled out the diagnosis of dedifferentiated chondrosarcoma.

5.3.4 Filtering and identification of somatic candidate variants in WES data

Using VarScan2 and MuTect, a total of 15,311 unfiltered SNVs and 187,373 unfiltered INDELs were identified in the samples (see Appendix Table 8–7). With such a large number of called variants, a series of filtering and prioritisation steps was followed (details in Section 2.3.6) to identify true positive candidate variants that can be of biological relevance (Figure 5–1). In short, somatic changes were prioritised if they altered the sequence of the produced protein (missense, nonsense, splice sites, INDELs) and were not common in the general population (MAF <0.1%). Missense variants were prioritised if predicted ‘deleterious’ by at least one *in silico* tool (SIFT or

PolyPhen2). For tumours that were not paired with a corresponding normal sample, the following additional filtering steps were applied: (1) discard variants reported in population databases regardless of MAF, (2) rescue variants with MAF of less than 0.01% in ExAc if they are reported somatic in COSMIC, and (3) eliminate variants with a VAF of >50% (to avoid possible germline variants).

Following the abovementioned filtering scheme, Figure 5–2 shows the total number of filtered somatic variants identified in each UPSb tumour (n=14). A total of 138 INDELs and 794 SNVs were detected in all UPSb tumours. Of the SNVs identified, 746 nonsynonymous variants were identified, accounting for the majority of SNVs (94%). In addition, of SNVs, 33 nonsense variants (4.2%) and 14 splice sites variants (1.8%) were discovered. SNVs were more prevalent than INDELs, which was expected because SNVs are the most common type of disease-related mutations (Bamshad et al., 2011).

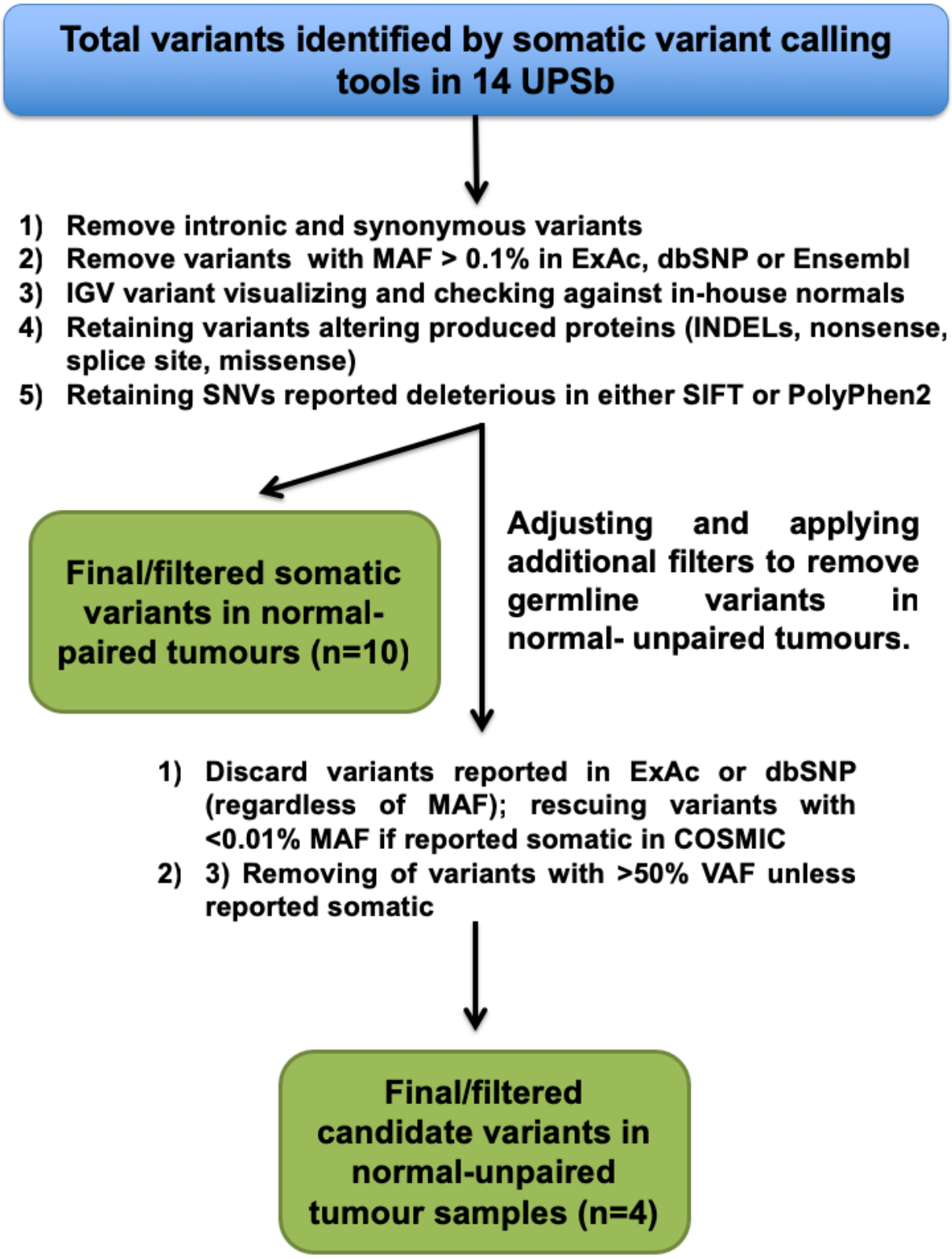


Figure 5–1: The prioritization scheme used to filter and prioritise WES variants identified in UPSb tumours. VAF: variants allele frequency; MAF: minor allele frequency in population databases; SNVs: single nucleotide variants; INDELs: small insertions/deletions; ExAc: The Exome Aggregation Consortium database

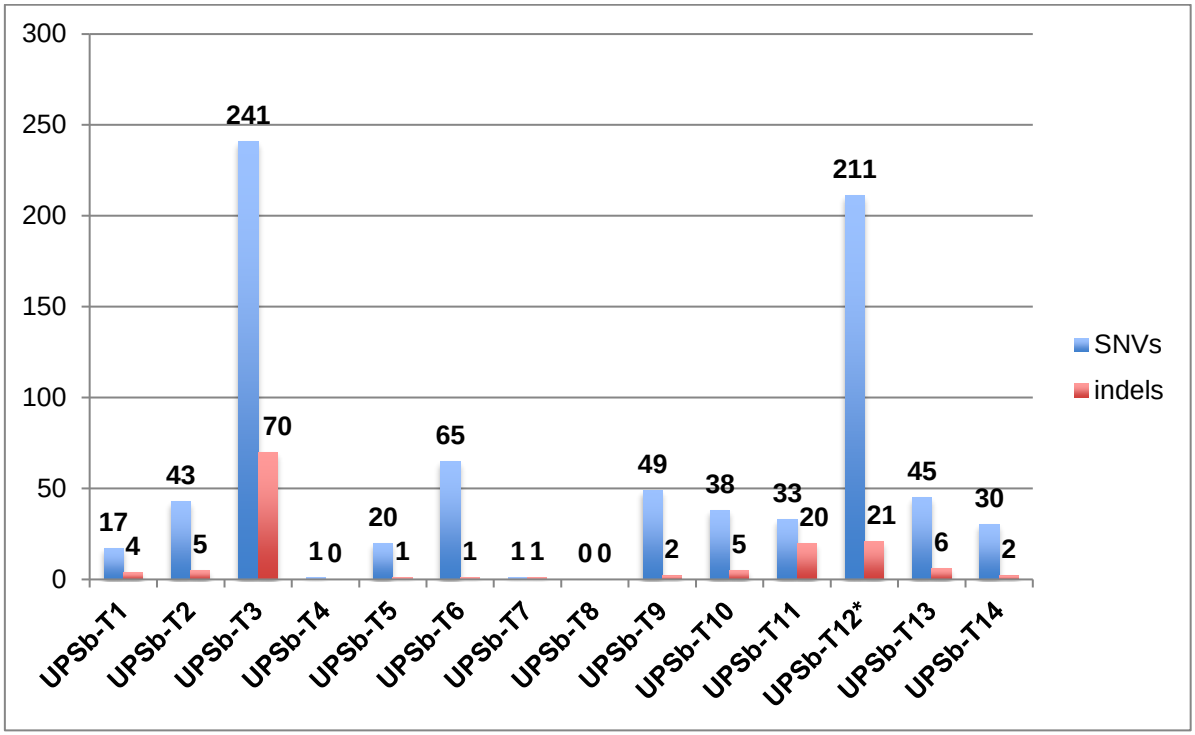
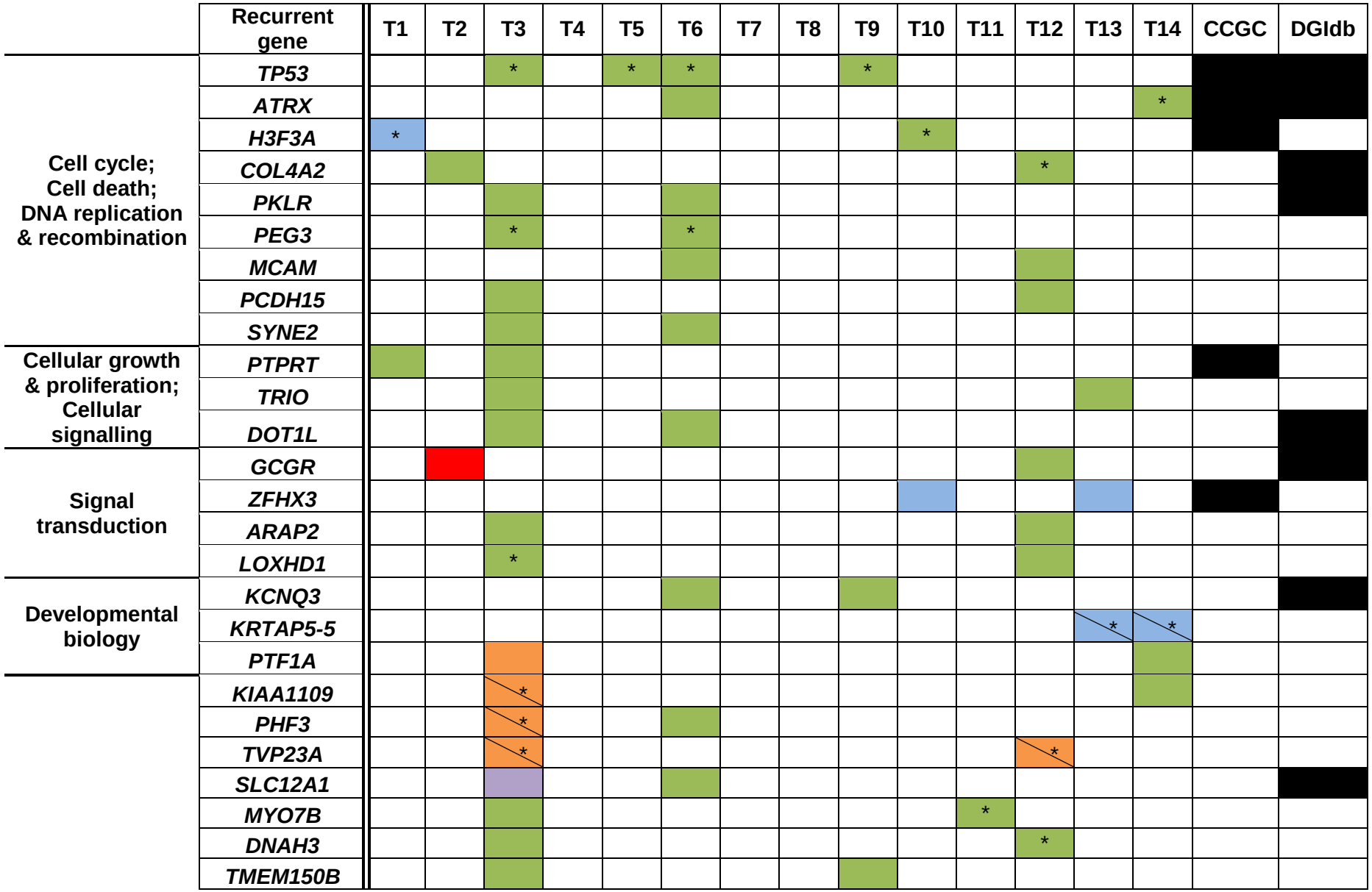


Figure 5–2: The total number of filtered WES alterations identified in 14 UPSb samples. UPSb-T12* is an FFPE tumour. INDELs: small insertions/deletions; SNVs: single nucleotide variants.

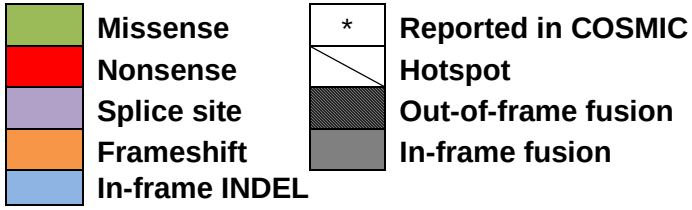
5.3.5 Identification of recurrent candidate genes in UPSb samples

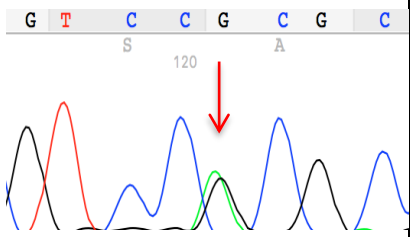
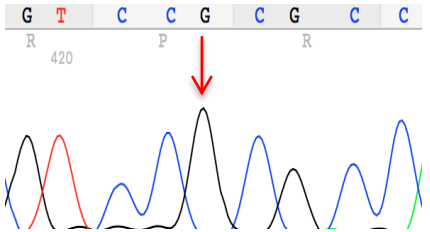
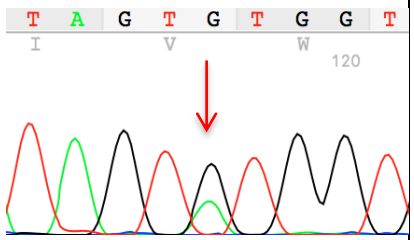
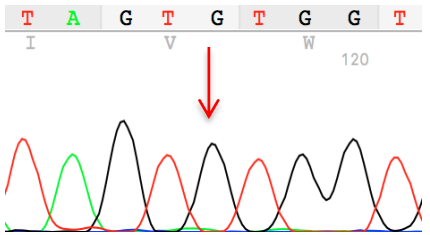
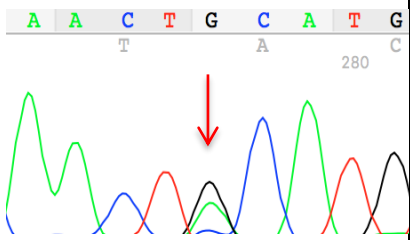
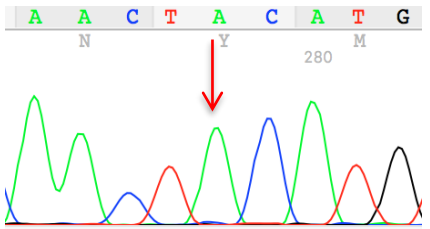
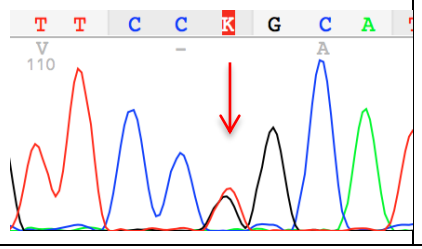
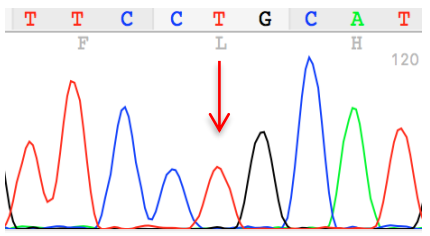
Recurrent mutated genes are genes that are mutated in more than one tumour sample. These genes can be cancer driver mutations and of potential biological and therapeutic relevance. A total of 31 recurrent genes were identified in 2 out of 14 UPSb samples (14.3%), except for *TP53* gene which was altered in four tumours (28.6%) (Table 5–1). Sequencing confirmation of few recurrent gene is in Table 5–2 (more sequencing results in Appendix Table 8–8). Recurrent genes were investigated further to identify the ones of potential tumourigenesis relevance.


Recurrent gene
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
T12
T13
T14
CCGC
DGIdb
Cell cycle; Cell death; DNA replication & recombination
Cellular growth & proliferation; Cellular signalling
Signal transduction
Developmental biology
TP53
ATRX
H3F3A
COL4A2
PKLR
PEG3
MCAM
PCDH15
SYNE2
PTPRT
TRIO
DOT1L
GCGR
ZFHX3
ARAP2
LOXHD1
KCNQ3
KRTAP5-5
PTF1A
KIAA1109
PHF3
TVP23A
SLC12A1
MYO7B
DNAH3
TMEM150B

Recurrent gene	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	CCGC	DGIdb
MAMDC4																
ZBTB34																
TTBK1																
CSMD3																

Table 5–1: Recurrent candidate genes identified in 14 UPSb samples using WES. CCGC: COSMIC Cancer Census Gene. **DGIdb**: The Drug Gene Interaction Database. Somatic recurrent genes are classified according to their biological function based on Ingenuity Pathway Analysis (IPA) (https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/) and Reactome Database (https://reactome.org/).



Gene	ID	WES variant details	Depth; %VAF	COSMIC ID / Transcript ID	SIFT; PolyPhen2	Tumour	Matched normal
TP53	T3	Missense; c.473G>A; p.Arg158His;	51; 57%	COSM1640853 / ENST00000413465	Tolerated; Damaging		
	T5	Missense; c.646G>A; p.Val216Met;	37; 21%	COSM1644280 / ENST00000413465	Deleterious; Damaging		
	T6	Missense; c.707A>G; p.Tyr236Cys;	54; 68%	COSM116674 / ENST00000413465	Deleterious; Damaging		
	T9	Missense; c.724T>G; p.Cys242Gly	38 / 16%	COSM3717645 / ENST00000413465	Deleterious; Damaging		

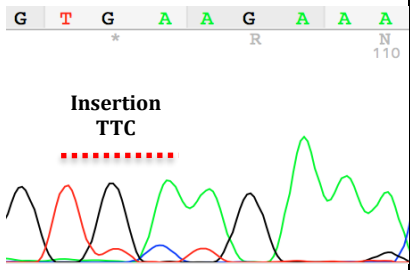
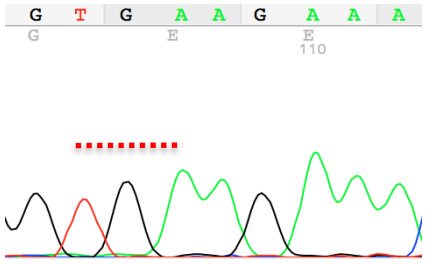
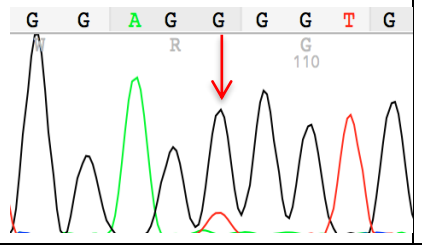
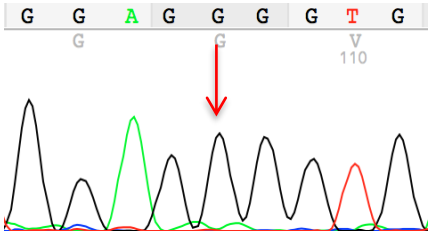
H3F3A	T1	Frameshift; c.106_107insTTC;p.Val36_Lys37insLeu	20; 21%	COSM5574356 / ENST00000366814	N/A		
	T10	Missense; c.104G>T; p.Gly35Val	70 / 25%	COSM502595 / ENST00000366814	Deleterious ; Damaging		

Table 5–2: Sanger sequencing confirmation of WES variants in UPSb tumours. All variants here are heterozygous and somatic. WES variants details column consists: (1) the type of WES variant (e.g., missense), (2) coding variants position (c.), for example, c.38G>A represents a G to A nucleotide substitution at nucleotide 38 of the coding region, where c.1 is nucleotide A of ATG start codon, and (3) affected translated protein (p.), for example, p.Gly13Asp represents a substitution of glycine to aspartic acid amino acid at codon 13, where methionine (ATG) start codon is #1. %VAF: altered variant allele frequency. COSMIC# represent the accession number of somatic variants previously reported in the COSMIC database, if present. All variants were confirmed by sequencing both reverse and forward direction; however, for illustrative purposes, sequencing electropherograms of one direction is presented.

5.3.5.1 Investigating recurrent genes in COSMIC cancer gene census, inTOgen and the drug interaction databases

To identify potential cancer driver genes, the 31 recurrent candidate genes were investigated in the CCGC list (v83) (http://cancer.sanger.ac.uk/cosmic/curation) and inTOgen database (https://www.intogen.org/search). CCGC-curated genes are genes that have shown strong evidence for being causally implicated in cancer development and pathogenicity. Of all recurrent genes, six genes (*TP53*, *ATRX*, *ZFHX3*, *H3F3A*, *CSMD3*, *PTPRT* and *TRIO*) were listed in the CCGC and four were classified as a 'cancer driver' in inTOgen (*TP53*, *ATRX*, *ZFHX3* and *TRIO*) (Table 5–1).

Recurrent candidate genes were investigated in DGIdb to identify potential 'druggable' genes. DGIdb classifies a gene as 'druggable' if a known interaction (e.g., inhibition) between a drug compound and a target gene has been reported in the literature. From the UPSb recurrent gene list, eight genes (*TP53*, *ATRX*, *DOT1L*, *GCGR*, *COL4A2*, *KCNQ3*, *SLC12A1* and *PKLR*) were classified as potential drug targets (Table 5–1).

5.4 RNA sequencing data analysis/processing and results

5.4.1 Overview of RNA-Seq experiment and processing of RNA-Seq raw reads

In the second part of the study, eight fresh-frozen UPSb tumours were transcriptome sequenced (RNA-Seq) to identify gene fusions. Gene fusions are produced as a result of intra-chromosomal (within chromosomes) or inter-chromosomal (between chromosomes) rearrangements. RNA from corresponding normal samples for all tumours was not RNA-sequenced but was available for any required laboratory work.

Prior to library preparation, extracted RNA (see Section 2.2.3) were quality checked by measuring RIN values to ensure RNA samples are mainly mRNA, non-degraded and of high-quality (examples in Material and Methods Figure 2–1).

The QC checks of the RNA sequenced data is useful to ensure the reliably of RNA-Seq data and provide quality metrics about the different steps of RNA-Seq technique such as library preparation and read alignment. High-quality mRNA samples yield a uniform coverage of RNA reads mapped to known exons, which is important for subsequent sequencing and data analysis (Conesa et al., 2016). RNA-Seq was performed at the Genomics Birmingham facility to achieve a minimum of 50 million read coverage. Sequenced data were uploaded into the BaseSpace hub (See Section 2.4.3.1).

The TopHat Alignment v1.0 and Cufflinks Assembly bundle tools v1.1 were used to perform the initial data processing of raw RNA-Seq (details in Section 2.4.3.1.1 & Appendix Section 8.4). To ensure reliability of the RNA-Seq data, the quality metrics of all samples were checked and showed satisfactory results (QC details in Appendix Table 8–4).

5.4.2 RNA-Seq data and identification of gene fusions

After preprocessing of RNA-Seq data, using TopHat2-Fusion and STAR-Fusion tools (see Section 2.4.4), a total of 111 candidate gene fusions were identified in eight UPSb tumours. Of the 111 gene fusions, two recurrent candidate gene fusions were identified: (1) *BACH1-GRIK1* in four tumour samples, and (2) *CTSC-RAB38* in two tumour samples. These recurrent fusions were investigated further in the literature and the following gene fusion databases: (1) ConjoinG, a database of conjoined genes (https://metasystems.riken.jp/conjoing/); (2) Tumor Fusion Gene Data Portal, which

uses gene fusion data from TCGA (http://www.tumorfusions.org); and (3) COSMIC fusion database (http://cancer.sanger.ac.uk/cosmic/fusion). None of the recurrent gene fusions (with same fusion gene partners) was reported previously in any database; however, a gene fusion involving *BACH1* and a different gene partner, *MAP3K7CL*, was reported in urothelial bladder carcinoma (BLCA) (TCGA.BT.A3PK.01A, http://www.tumorfusions.org). The *CTSC-RAB38* fusion has been previously reported as germline (present in non-neoplastic samples) (Babiceanu et al., 2016) and a transcription read-through event (Grosso et al., 2015). Read-through RNA chimeras usually result from an RNA processing event in which the transcription machinery reads through the intragenic regions of two neighbouring and similarly oriented genes in the absence of a DNA rearrangement, resulting in a spliced together chimeric transcript (Grosso et al., 2015; Jia et al., 2016).

Two sets of primers were designed to validate the recurrent gene fusions on the cDNA level by RT-PCR. *BACH1-GRIK1* was confirmed in both normal and tumour cDNA of four samples and, therefore, was classified as germline. As expected, *CTSC-RAB38* was also confirmed germline in two samples. Although read-through transcripts have been identified in many neoplastic samples (Babiceanu et al., 2016; Grosso et al., 2015) their presence in non-neoplastic tissues makes their biological importance unclear, and therefore, read-throughs are normally discarded from RNA sequencing analysis (Delespaul et al., 2017; Giacomini et al., 2013; Jia et al., 2016).

To filter the remaining 109 non-recurrent gene fusions, a series of filtering steps were followed (Figure 5–3, details in 2.4.4). In short, gene fusions were prioritised if they were supported by >3 junction/supporting fusion reads and the fusion breakpoints were within the exons or exonic boundaries. By contrast, gene fusions in intronic regions were discarded. Subsequently, nine candidate gene fusions were prioritised and

assessed in tumour and corresponding normal samples using RT-PCR. Eight gene fusions were confirmed somatic (present in the tumour and absent from the corresponding normal sample) whereas one fusion was classified as germline (RT-PCR gels in Appendix Figure 8–9).

The eight somatic fusions were further prioritised to identify fusions that can potentially be involved in UPSb tumourigenesis or therapeutically targeted (Table 5–3). Fusion genes were prioritised if (1) they were reported previously in the literature or in gene fusions databases (ConjoinG, TCGA fusion portal or COSMIC gene fusion), or (2) fusion gene partners are classified as a 'cancer driver' in the inTOgen cancer database (https://www.intogen.org/search) or CCGC list (http://cancer.sanger.ac.uk/cosmic/curation). Gene fusions were also considered of biological relevance if one or both gene fusion partners exhibit fundamental roles in cellular processes or homeostasis (e.g., oncogenes, tumour suppressor or transcription factors) (Abate et al., 2014; Atak et al., 2013). A gene fusion was considered potentially 'druggable' if one or both fusion partners were classified as a potential drug target by DGIdb (http://www.dgidb.org). Lastly, a chimeric transcript including an in-frame kinase gene partner was also considered potentially druggable (Tamura et al., 2015) (more in Discussion).

Three of eight somatic fusions were considered potentially druggable, *PKNOX2-MMP20*, *ASAP2-ADAM17* and *FARP1-STK24*. In addition, two of the eight genes identified in this study (*CLTC-VMP1* and *FARP1-STK24)* have been previously reported in the literature in other cancers, highlighting their potential involvement in tumourigenesis. The RT-PCR and LR-PCR analyses of these two fusions are discussed in detail below.

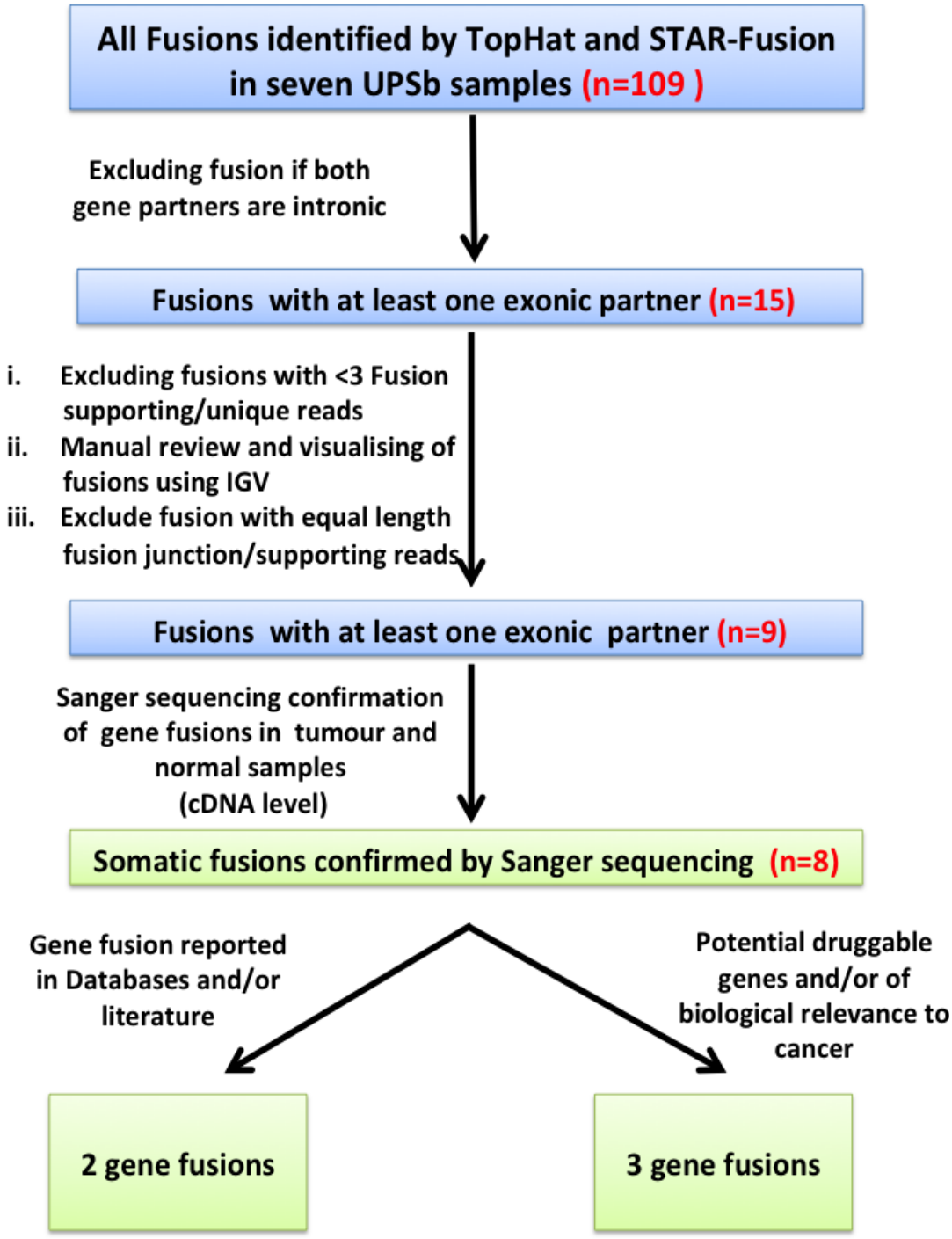


Figure 5–3: Filtering and prioritization scheme of gene fusions identified by RNA-Seq in eight UPSb tumours. The scheme shows a series of filtering steps to identify genuine candidate gene fusions. The fusions were further prioritised based being reported in the literate or of biological relevance.

Sample	Gene fusion	Novel? (literature search)	inTOgen cancer driver	COSMIC gene fusions/COSMIC curated genes	TCGA Fusion Gene Data Portal	Kinase gene partner, or related to cancer	DGIdb drug-gene interaction
UPSb-T2	*PKNOX2-MMP20*	Novel	No reported	No reported	Not reported	***PKNOX2*, T.F.**	**Yes, *MMP20***
	CMAS-PYROXD1	Novel	No reported	No reported	Not reported	None	No drugs identified
UPSb-T6	*ASAP2-ADAM17*	Novel	No reported	No reported	Not reported	None	**Yes, *ADAM17***
	FARP1-STK24	**Reported previously**	No reported	No reported	Not reported	***STK24*, kinase**	**Yes, *STK24***
UPSb-T9	*OSBPL2-CABLES2*	Novel	No reported	No reported	Not reported	None	No drugs identified
	MICAL3-UFD1L	Novel	No reported	No reported	Not reported	None	No drugs identified
UPSb-T13	*CLTC-VMP1*	**Reported previously**	**Reported, *CLTC***	***CLTC* in curated gene list**	**Reported**	None	No drugs identified
	APOL1-MYH9	Novel	**Reported, *MYH9***	***MYH9* in curated gene list**	Not reported	None	No drugs identified

Table 5–3: Literature and database search to prioritise validated UPSb somatic gene fusions. COSMIC curated gene list involves genes that have been experimentally implicated in cancer. Genes were classified as related to cancer if they were tumour suppressors, oncogenes or transcription factors involved in fundamental cellular processes or homeostasis. Details about gene fusions breakpoints and junction/supporting and spanning reads in Appendix Table 8–9. inTOgen: Integrative Onco Genomics Database; COSMIC: Catalogue of Somatic Mutations in Cancer; TCGA: The Cancer Genome Atlas; DGIdb: The Drug Gene Interaction Database; T.F.: transcription factor.

5.4.2.1 *CLTC-VMP1* gene fusion in UPSb-T13

The *CLTC-VMP1* gene fusion is supported by 156 fusion junction/supporting RNA-Seq reads that are uniformly distributed around the fusion junction/breakpoint (Figure 5–4A). The *CLTC-VMP1* is out-of-frame (more in Discussion) and formed by fusing exons 1–14 (NCBI Reference Sequence NM_001288653.1) of *CLTC* (5′-gene partner) to the last two exons of *VMP1* (NCBI Reference Sequence: NM_030938) (3′-gene partner) (Figure 5–5**)**. *CLTC-VMP1* has been previously reported in multiple cancers (more in Discussion) (Giacomini et al., 2013; Liu et al., 2012; Robinson et al., 2011a). Using RT-PCR, a set of primers annealing to *CLTC* and *VMP1* gene partners was used to validate and amplify the fusion junction/breakpoint region. *CLTC-VMP1* was confirmed somatic in UPSb-T13 (Figure 5–5C). The GAPDH housekeeping gene was used as a positive control.

5.4.2.2 *FARP1-STK24* gene fusion in UPSb-T6

The *FARP1-STK24* fusion is supported by 15 evenly distributed fusion junction/supporting reads and was confirmed somatic in UPSb-T6 by RT-PCR **(**Figure 5–4B). This gene fusion forms by joining the first two exons and part of exon 3 of *FARP1* (NCBI Reference Sequence: NM_001001715.3) to the first five exons of *STK24* (NCBI Reference Sequence: NM_003576) (Figure 5–6). The 5′-ends of both genes are fused together (5′-5′ gene fusion representation) due to an interstitial (intra-chromosomal) deletion (more later). A study by Veeraraghavan et al. (2014) reported *FARP1-STK24* fusion in a breast cancer tumour in a somatic state. The DGIdb database identifies a potential kinase inhibitor drug named Bosutinib for *STK24* gene (http://www.dgidb.org/drugs/BOSUTINIB#_summary).

5.4.3 Identification of fusion breakpoints at the genomic level using long-range PCR

Unlike the breakpoints at the transcriptomic/cDNA level, the genomic points of the fusions were unknown and, therefore, required designing multiple primer spanning exonic and intronic regions of both gene partners—a process known as genomic walking. During LR-PCR genome walking, one primer (usually forward) is anchored (unchanged), while multiple reverse primers are used to map genomic breakpoints.

Focusing on gene fusions with potential relevance in tumourigenesis or druggability, genomic breakpoint analysis of five fusions was performed to confirm that the chimeric fusion transcript occurred as a result of chromosomal rearrangements at the genomic/DNA level rather than a defective non-genomic transcription machinery event (e.g., read-through). The five fusions were *CLTC-VMP1*, *FARP1-STK24, PKNOX2-MMP20, APOL1-MYH9* and *ASAP2-ADAM17.* LR-PCR was conducted to identify the genomic breakpoint of each gene fusion partner (using tumour and normal DNA), followed by Sanger sequencing confirmation (see Section 2.5.3.2). Using LR-PCR, all five gene fusion were confirmed somatic (present in tumours but not in corresponding normal samples) and occurred as a result of genomic rearrangements (LR-PCR gels in Appendix Figure 8–10). The two previously reported gene fusions (*CLTC-VMP1 and FARP1-STK24*) are caused by interstitial deletions (details in the following sections).

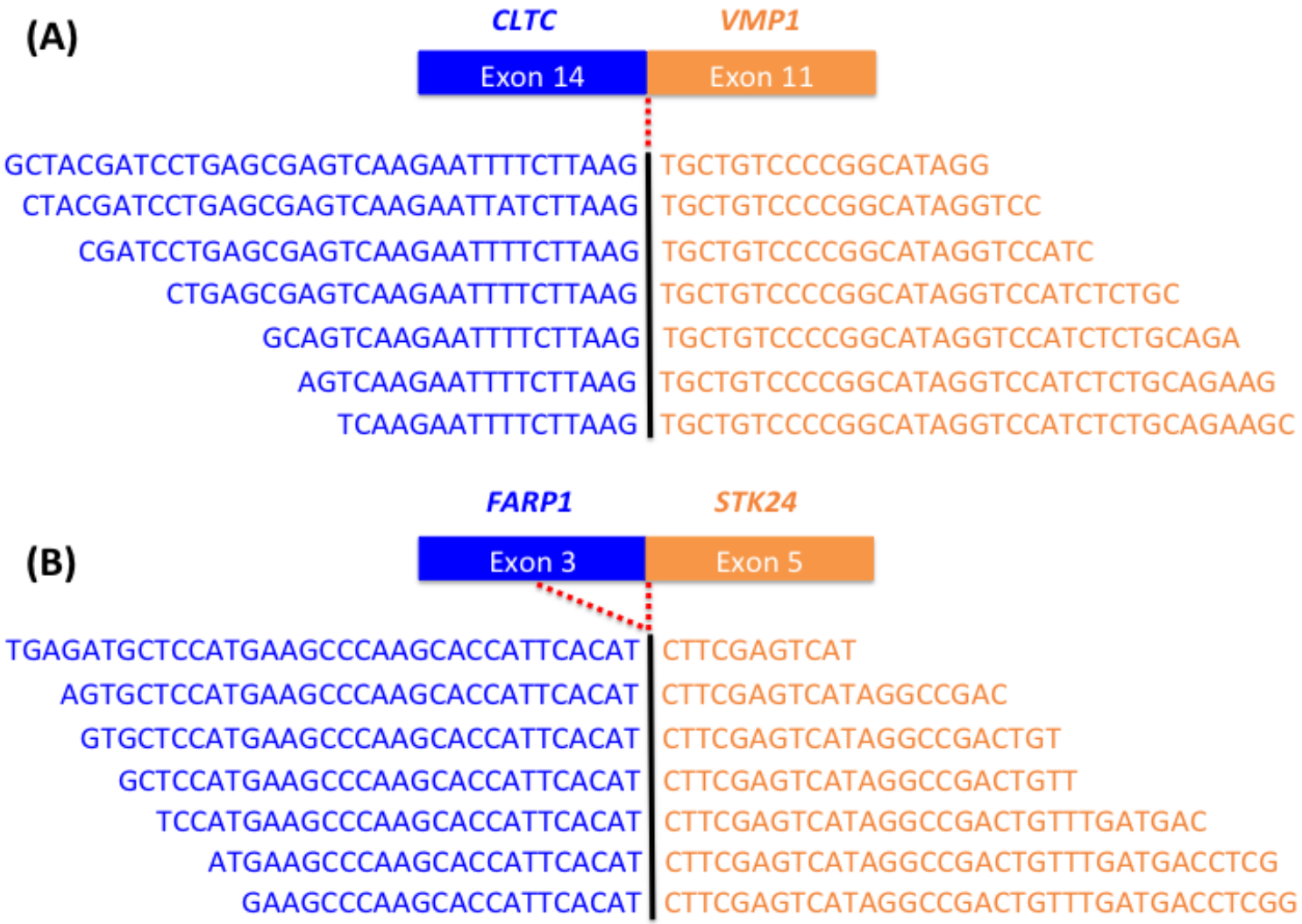


Figure 5–4: The distribution of junction/supporting reads for (A) *CLTC-VMP1* and (B) *FARP1-STK24* gene fusions. The vertical black line represents the fusion break point. Reads are coloured, in blue or orange, according to each fusion gene partner. For illustrative purposes, only seven fusion junction reads are shown in here.

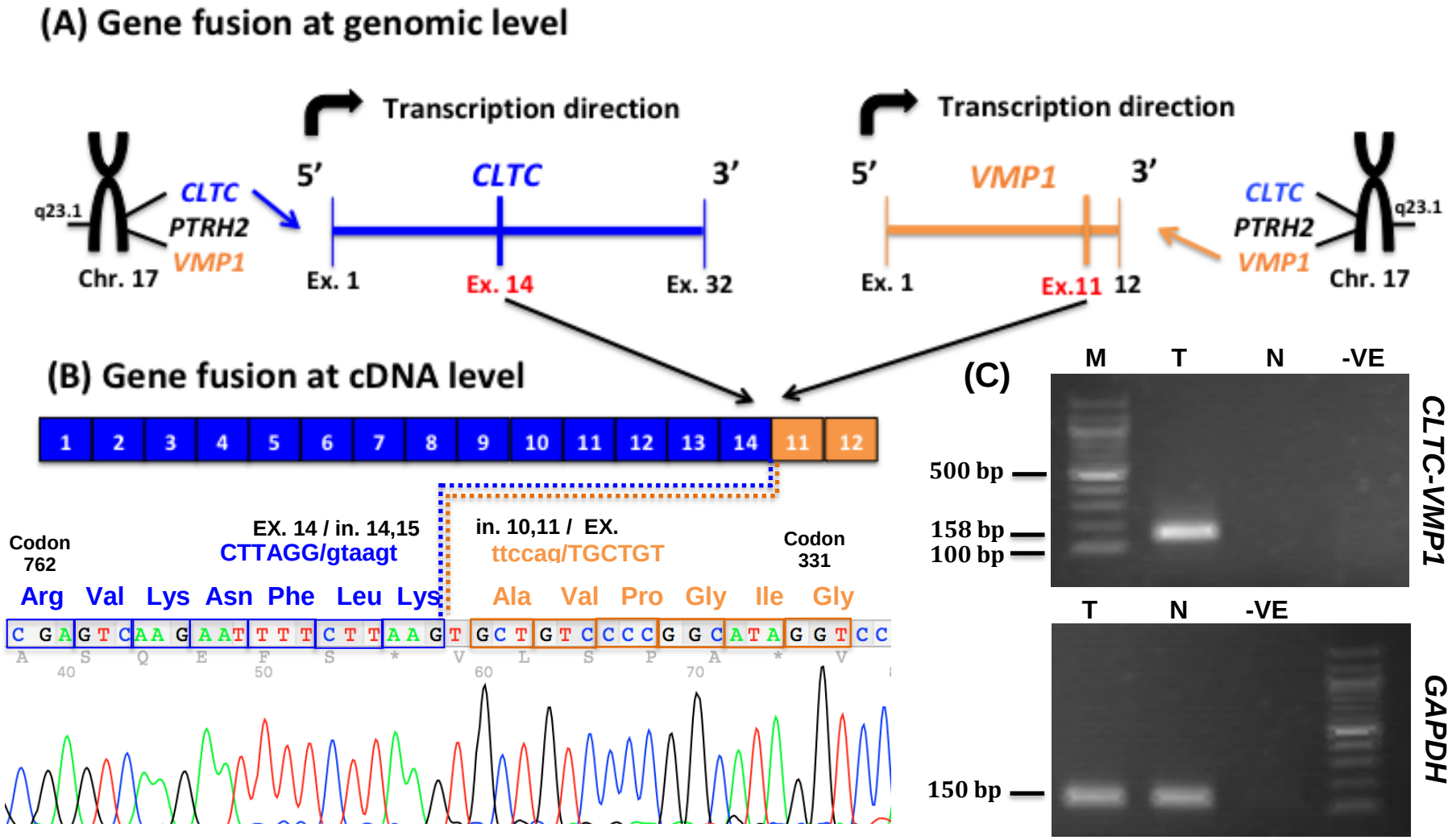


Figure 5–5: Schematic representation and RT-PCR confirmation of *CLTC-VMP1* fusion in UPSb-T13. (A) Representation of gene fusions partners, in blue (*CLTC*) and orange (*VMP1*), at the genomic level. Exons (Ex.) are represented by vertical lines. Both genes are transcribed from the forward (sense) strand. (B) Transcriptomic representation of the gene fusion. Translated amino acid codons are in blue and orange boxes for *CLTC* and *VMP1*, respectively. The unboxed 'T' nucleotide belongs to the amino acid proceeding Ala (not shown here) resulting in an out-of-frame fusion. For each gene partner, exonic (EX.) nucleotides are in CAPS whereas intronic sequences (in.) are in lowercase. (C) The agarose gel (2.5%) shows the RT-PCR products of *CLTC-VMP1* fusion and GAPDH (housekeeping gene control) in T: tumour; N: corresponding normal; -VE: negative control (water), M: DNA marker/ladder.

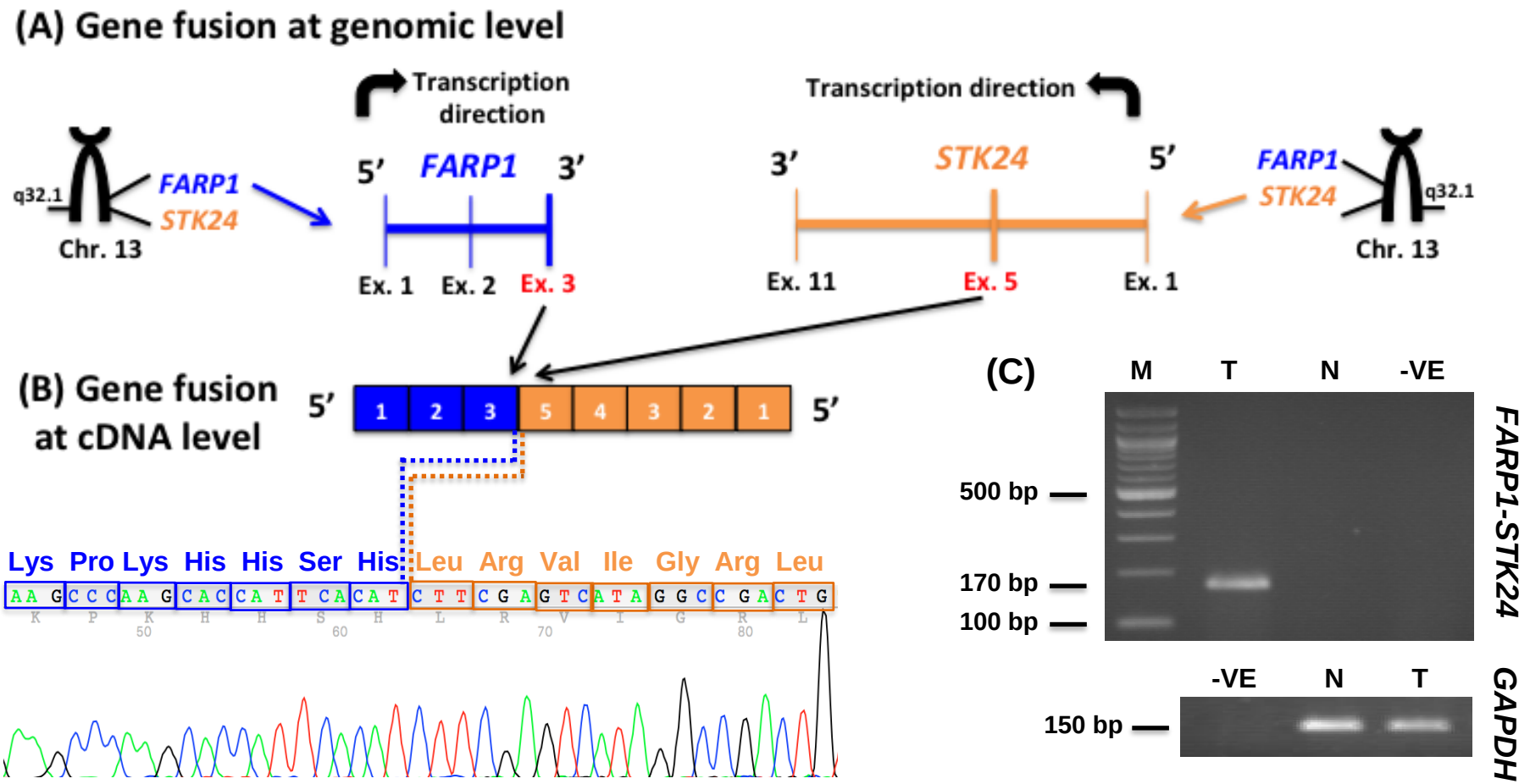


Figure 5–6: Schematic representation and RT-PCR confirmation of *FARP1-STK24* fusion in UPSb-T6. (A) Genomic representation of the gene fusion. *FARP1* and *STK24* are coloured in blue and orange, respectively. *FARP1* and *STK24* are transcribed from the forward (sense) and reverse (antisense) strands, respectively. (B) Gene fusion representation at the transcriptomic level. The 5' segments of both genes are fused together. In-frame translated amino acids of *FARP1* and *STK24* are in blue and orange boxes, respectively. (C) The agarose gel (2.5%) images show RT-PCR confirmation of *FARP1-STK24* gene fusion, including GAPDH positive control. T: tumour sample; N: corresponding normal; -VE: negative control (water) and M: DNA marker/ladder.

5.4.3.1 Genomic mapping of *CLTC-VMP1* gene fusion breakpoints in UPSb-T13

The breakpoints at the cDNA level of the *CLTC* and *VMP1* genes were at the end of exon 14 and beginning of exon 11, respectively (Figure 5–5). We speculated that the genomic breakpoints are likely to be within intron 14/15 for *CLTC* and intron 10/11 for *VMP1* (Figure 5–7). The majority of gene fusion genomic breakpoints occur within intronic/intergenic regions due to the longer length of non-coding regions than those of exons. To map genomic breakpoints, sets of forward and reverse primers (spaced approximately 750–1500 bp apart) were designed to genome walk through the intronic regions mentioned above with fewer primers annealing to exons 14 of *CLTC* and exons 11 of *VMP1*. When visualised using agarose gels, ladder-shaped LR-PCR amplicons infer that LR-PCR products are becoming smaller in size, and, therefore, moving toward the fusion genomic breakpoints (Figure 5–7C).

Ten primer pairs consisting of one anchored/universal forward primer (F1) and ten reverse primers (R1-10) were used in the genome walking analysis. All primer sets amplified products successfully, except for F1-R10, suggesting that the fusion breakpoint is upstream from the annealing site of R10 primer. As expected, the fusion genomic breakpoints were mapped to introns 14/15 of the *CLTC* and introns 10/11 of *VMP1* fusion gene partners (Figure 5–7A). Sanger sequencing of the smallest LR-PCR product (F1/R9) mapped the *CLTC* genomic breakpoint 4879 bp downstream from the end of exon 14 and the *VMP1* breakpoint 10,888 bp upstream from the beginning of exon 11 (Figure 5–7B). The genomic mapping analysis shows that intron 14/15 of *CLTC* is joined to intron 10/11 of *VMP1* as a result of a ~158kb interstitial (within chromosome) deletion at 17q23.1 between the *CLTC* and *VMP1* genes. The entire *PTRH2* gene that is located between *CLTC* and *VMP1* is also deleted (more in

Discussion). *CLTC-VMP1* was confirmed as somatic for second time by LR-PCR using DNA of the tumour and the corresponding normal tissue (Figure 5–7C).

5.4.3.2 Genomic mapping of *FARP1-STK24* fusion breakpoints in UPSb-T6

From the RNA-Seq data and RT-PCR confirmation, the transcriptomic gene fusion breakpoints were in the middle of exon 3 (NCBI Reference Sequence: NM_001001715.3) for *FARP1* gene partner and at end of exon 5 for *STK24* gene (NCBI Reference Sequence: NM_003576.4) (Figure 5–6). Since the *FARP1* cDNA/transcriptomic breakpoint occurred in the middle of exon 3, the theoretical *FARP1* genomic/DNA breakpoint is likely to be the same or immediately downstream from the cDNA breakpoint. The theoretical genomic breakpoint of *STK24* is likely to be within intron 5/6. Similar to the genomic mapping assessment of *CLTC-VMP1* fusion, genome walking analysis was performed.

Unexpectedly, the LR-PCR genomic breakpoint assessment of *FARP1* did not match the theoretical breakpoint. The *FARP1* DNA breakpoint was 4871 bp downstream from the *FARP1* cDNA fusion in an intergenic region between *FARP1* and *STK24* genes (Figure 5–8A). On the other hand, as expected, the *STK24* genomic breakpoint was in intron 5/6, 6019 bp downstream of the end of exon 5 (Figure 5–8B). The *FARP1-STK24* fusion occurs as result of a ~219kb interstitial deletion between the two genes. The *FARP1-STK24* fusion was confirmed somatic by LR-PCR (Figure 5–8C).

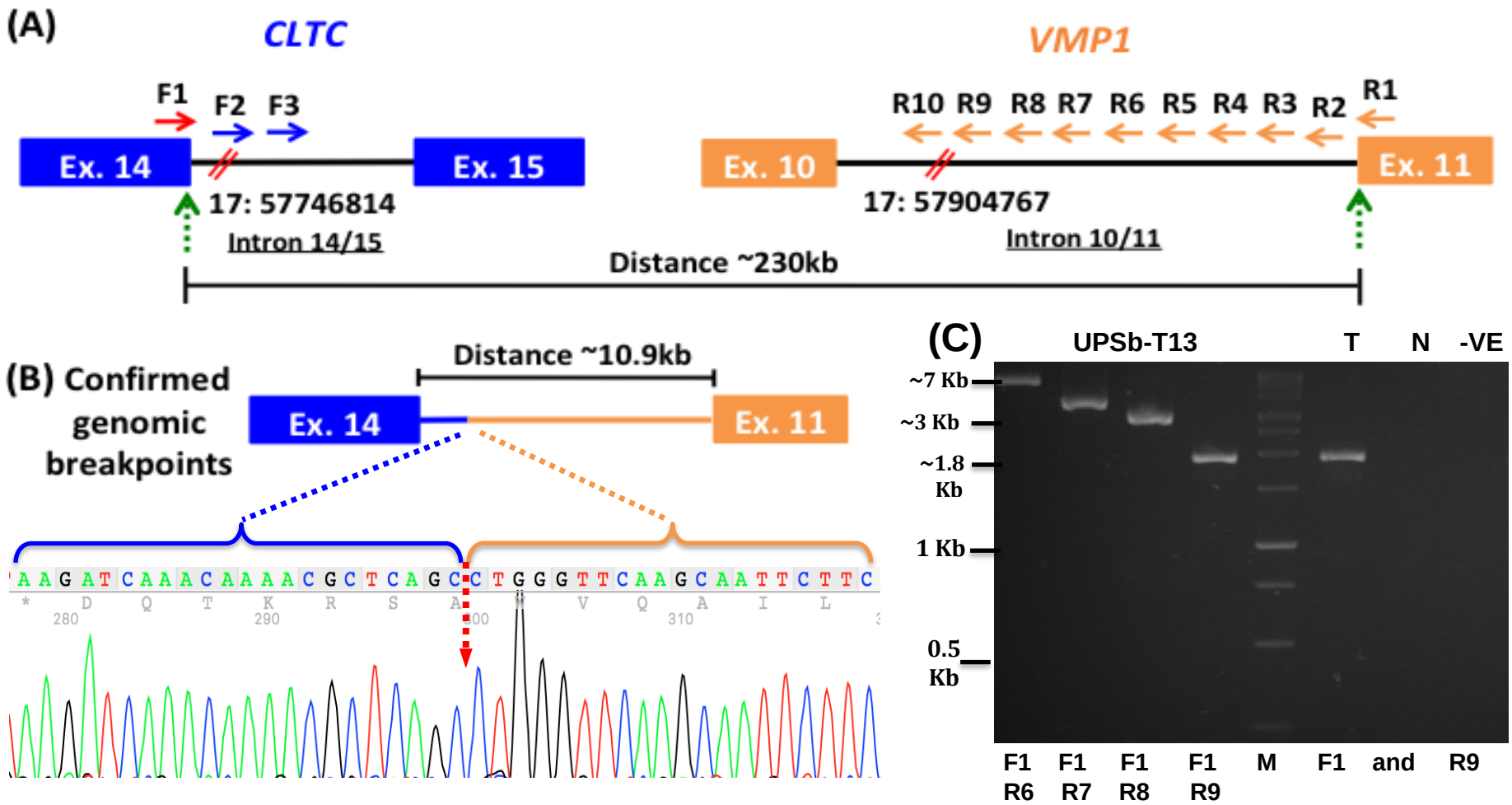


Figure 5–7: Genomic representation of *CLTC-VMP1* gene fusion and LR-PCR confirmation of the genomic breakpoints. (A) The *CLTC* and *VMP1* gene fusion partners are coloured in blue and orange, respectively. Green dotted arrow represents cDNA breakpoint and the genomic distance between the cDNA breakpoints of fusion gene partners. Black solid horizontal refers to intronic regions. Red double slash (//) represents the mapped genomic DNA breakpoints in *CLTC* (Chr17:57746814) and *VMP1* (Chr17:57904767).The small horizontal arrows represent F: forward and R: reverse primers; the red arrow represents the anchored primer (F1). (B) Sanger sequencing confirming the breakpoint (vertical red-dotted arrow) in *CLTC* (blue) and *VMP1* (orange) and the newly formed genomic distance between fused exons. (C) The agarose gel (1%) shows a ladder-shaped of LR-PCR amplicons using different primers. The somatic status of the fusion was assessed using F1&R9 primers in T: tumour; N: corresponding normal; -VE: negative control (water). M: 1 Kb DNA marker/ladder.

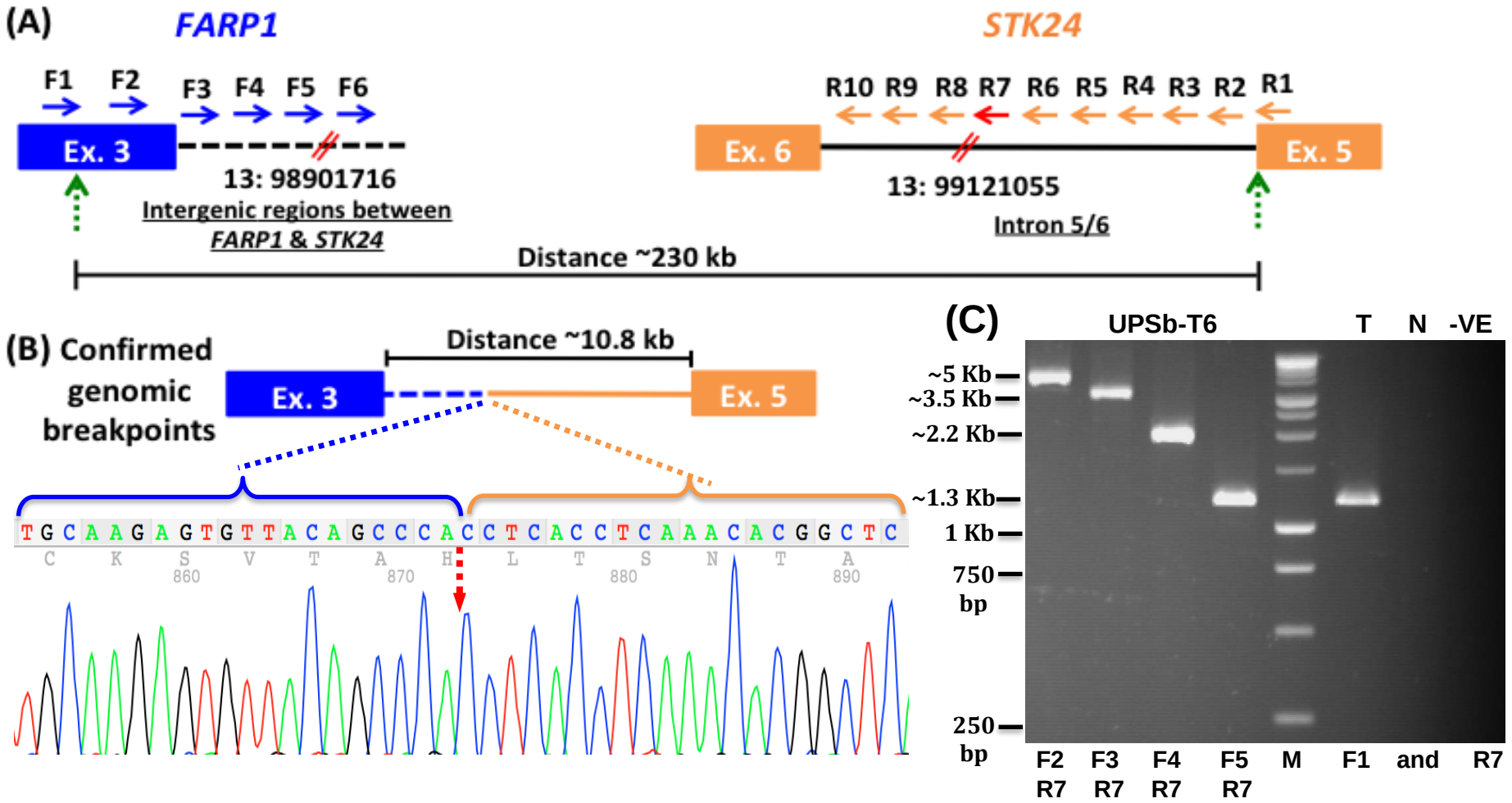


Figure 5–8: Genomic representation of *FARP1-STK24* gene fusion and LR-PCR confirmation of genomic/DNA breakpoints. (A) Details of *FARP1* and *STK24* gene fusion partners are coloured in blue and orange, respectively. Green dotted arrow represents the cDNA breakpoint, black dashed horizontal line represents the intergenic region between *FARP1* and *STK24,* and black solid horizontal refers to *STK24* intronic region. Red double slash (//) represents the mapped genomic DNA breakpoints of *FARP1* (Chr13:98901708) and *STK24* (Chr13:99121055). The small horizontal arrows denote F: forward and R: reverse primers, where the red arrow represents the anchored primer (R7). (B) Sanger sequencing confirms the breakpoint (vertical red-dotted arrow) in *FARP1* (blue) and *STK24* (orange) and the altered genomic distance, due to the chromosomal deletion. The agarose gel (1%) shows a ladder-shaped LR-PCR bands using different primers. The fusion somatic status was assessed using F1 & R7 primer set in T: tumour; N: corresponding normal; -VE: negative control (water). M: 10 Kb DNA marker/ladder.

5.5 Summary of findings

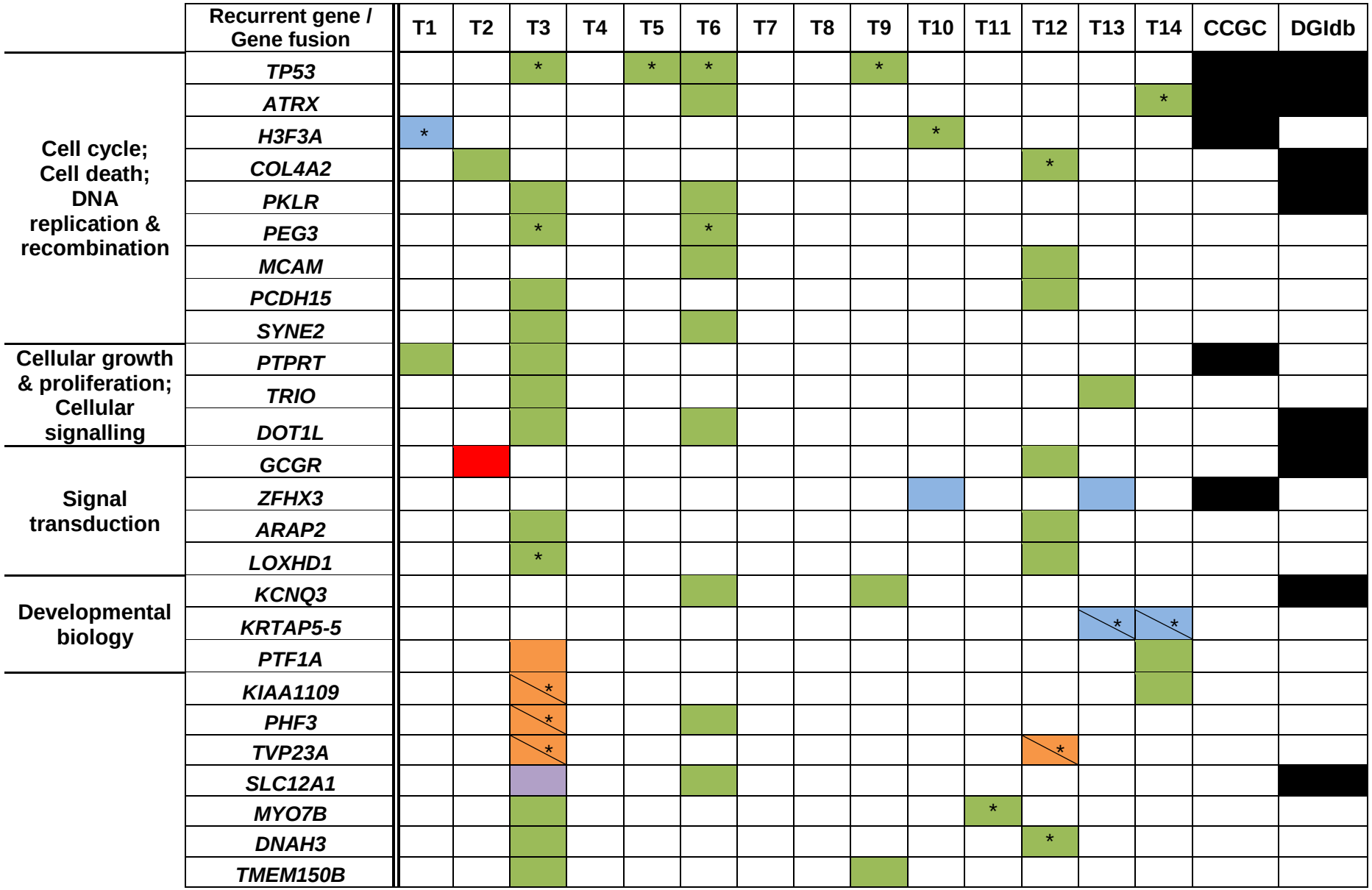
The first part of this project aimed to identify somatic SNVs and small INDELs in 14 UPSb samples using WES. Ten UPSb samples were paired with their corresponding normal samples, while four tumour samples were exome-sequenced alone (without normal samples). A series of filtering and prioritisation steps were followed to identify potentially deleterious somatic alterations in all samples. A total of 794 SNVs and 138 INDELs were identified in 14 UPSb tumours. A total of 31 somatic recurrent genes were altered in 2 of 14 UPSb tumours (14.3%), except for TP53 which was mutated in four samples (28.6%) (Table 5–4). Recurrent candidate genes were prioritised further to identify genes that may be involved in UPSb tumourigenesis (i.e., cancer driver genes) or targeted by drugs. Of the 31 recurrent genes, eight were classified as 'cancer driver' genes by COSMIC and inTOgen databases. Eight genes were classified as 'potentially drug-targeted' by DGIdb based on possible drug-gene interactions. The landscape of WES alterations in UPSb samples is shown in Table 5–4.

In the second part of the study, RNA-Seq was performed on eight UPSb tumours to identify gene fusions using two splice-junction identifier tools. Following a series of filtering steps, eight somatic gene fusions were detected in four UPSb tumours (Table 5–4). All gene fusions were tumour-specific (private); that is, none of the gene fusions were recurrently identified in more than one sample, except for two germline fusions that were not of biological interest. Somatic gene fusions were further prioritised to infer their potential biological importance and identify fusions that can be potentially drug targeted. Of the eight somatic gene fusions, two were previously reported in other cancers, and three have shown potential drug-gene interactions according to the DGId (Table 5–4).

5.6 Discussion

5.6.1 Discussion part one: WES findings

Due to their importance in tumourigenesis, identifying somatic genomic alterations is one of the primary aims of cancer genetic studies. This knowledge can help provide accurate diagnoses and prognosis options and develop targeted therapeutics for patients (Dees et al., 2012; Hofree et al., 2016). NGS technologies allow researchers to investigate genetic alterations in a high-throughput and in-depth genome-wide perspective, revealing the complex landscapes of human cancer biology (Wang et al., 2013). Since the first proof of concept WES study in 2009 (Ng et al., 2009), numerous studies have focused on examining the cancer exome, aiming to identify and discriminate between genuine cancer driver genes and passenger (unlikely pathogenic) genes (Hofree et al., 2016; Lawrence et al., 2013). As stated previously, although it only focuses on protein coding regions of the genome, WES is considered to be an efficient, cost-effective method to identify exonic genomic alterations, which can have severe effects on the translated proteins (Bertier et al., 2016). A comprehensive and accurate identification of actual driver genes of UPSb will significantly help in translating these findings into clinical application for patient diagnosis. While potential drug-gene interactions were identified in eight recurrent genes, a more comprehensive understanding from additional biological studies are required before translating these findings into clinical practice.


Recurrent gene / Gene fusion
T1
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
T12
T13
T14
CCGC
DGIdb
Cell cycle; Cell death; DNA replication & recombination
Cellular growth & proliferation; Cellular signalling
Signal transduction
Developmental biology
TP53
ATRX
H3F3A
COL4A2
PKLR
PEG3
MCAM
PCDH15
SYNE2
PTPRT
TRIO
DOT1L
GCGR
ZFHX3
ARAP2
LOXHD1
KCNQ3
KRTAP5-5
PTF1A
KIAA1109
PHF3
TVP23A
SLC12A1
MYO7B
DNAH3
TMEM150B

Recurrent gene	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	CCGC	DGIdb
MAMDC4																
ZBTB34																
TTBK1																
CSMD3																
CLTC-VMP1															***CLTC***	
FARP1-STK24																
APOL1-MYH9															***MYH9***	
ADAM17-ASAP2																
PKNOX2-MMP20																
OSBPL2-CABLES2																
MICAL3-UFD1L																
CMAS-PYROXD1																

Table 5–4: WES and RNA-Seq genetic landscape in 14 UPSb samples. CCGC: COSMIC Cancer Census Gene. **DGIdb**: The Drug Gene Interaction Database. Somatic recurrent genes are classified according to their biological function based on Ingenuity Pathway Analysis (IPA) (https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/) and Reactome database (https://reactome.org/). Gene fusions reported previously in the literature are underlined.

Missense
Nonsense
Splice site
Frameshift
In-frame INDEL

* **Reported in COSMIC**
Hotspot
Out-of-frame fusion
In-frame fusion

5.6.1.1 *TP53* cancer-associated gene

5.6.1.1.1 Overview of *TP53* mutation frequency in cancer

The *TP53* gene is a well-defined cancer gene that has been associated with more than 50% of human cancers (Baugh et al., 2018). The p53 protein, a transcription factor involved in signalling and metabolic pathways, is known for its function as a tumour suppressor (Baugh et al., 2018; Berkers et al., 2013; Leroy et al., 2014). The p53 protein is activated during various cellular stress events, arising during malignant transformation to maintain cellular homeostasis. Depending on the context and stress characteristics of the cell, p53-mediated stress response can induce DNA damage repairing or cell death (apoptosis) responses in mild or severe cellular stress events, respectively (Berkers et al., 2013).

In cancer, the frequency of *TP53* alteration varies considerably depending on the type of cancer, ranging from 5–90% in occurrence (Leroy et al., 2014) (Figure 5–9). Approximately 90% of *TP53* mutations are missense mutations that produce a faulty protein, whereas the remaining are loss-of-function mutations (e.g., frameshift, nonsense, larger deletions) that produce a truncated protein (Baugh et al., 2018). For missense mutations, guanine to adanine is most common substituted nucleotide (COSMIC, http://cancer.sanger.ac.uk/). In this study, all recurrent *TP53* alterations were missense substitutions, two of which were guanine to adanine substitutions.

Osteosarcoma is one of the few cancers that display a high frequency (~50%) of *TP53* gene deletion and are usually associated with poor prognosis (Joseph et al., 2014; Leroy et al., 2014). *TP53* nonsynonymous SNVs were also reported in 7% of Ewing sarcomas (Tirode et al., 2014) and synovial sarcomas (Joseph et al., 2014).

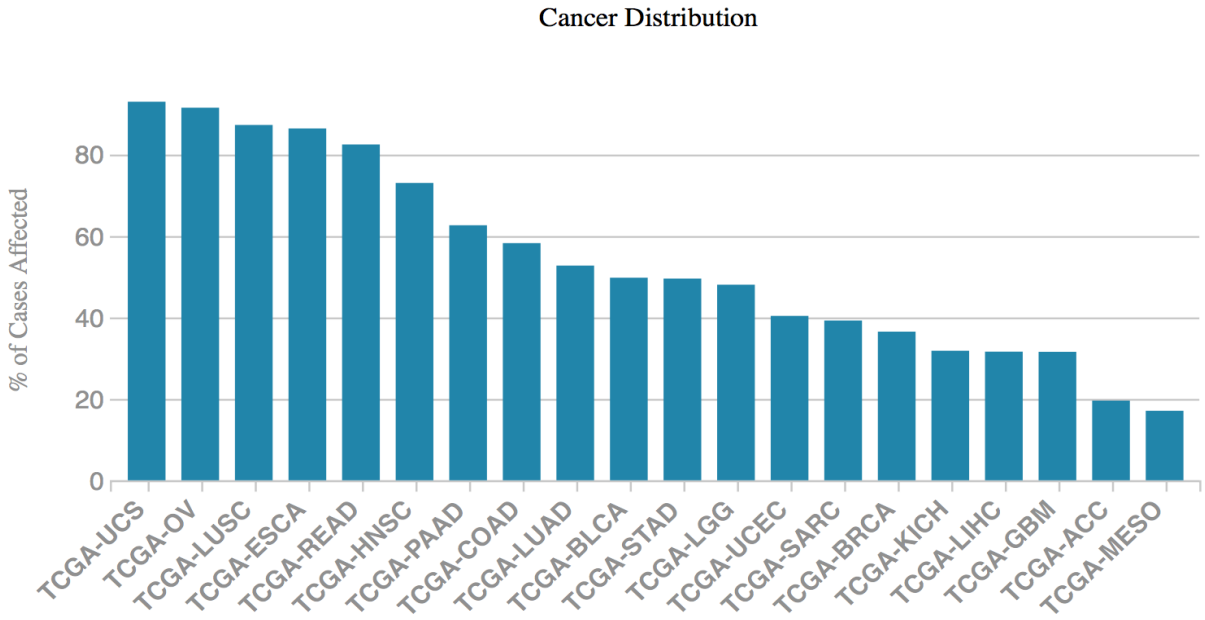


Figure 5–9: Distribution of *TP53* alterations across 20 top-mutated cancer projects. A total of 4,008 cases are affected by 1,291 *TP53* mutations across 32 of projects of The Cancer Genome Atlas (TCGA). The highest percentage of *TP53* mutations (92.98%) is found in uterine carcinosarcoma (TCGA-UCS) whereas the lowest percentage (0.56%) is reported in pheochromocytoma and paraganglioma (not shown here). Image obtained from National Cancer Institute, GDC Data Portal https://portal.gdc.cancer.gov).

5.6.1.1.2 Four somatic *TP53* variants identified in four UPSb samples by WES

The four somatic heterozygous missense variants identified in four different UPSb tumours (Table 5–4) are described as somatic in COSMIC and reported in various other cancer types, suggesting a pathological role of these variants in tumourigenesis (Table 5–5). Moreover, p.Arg158His, p.Val216Met and p.Tyr236Cys variants are reported among the 50 most common *TP53* mutations from the International Agency for Research on Cancer (IARC) (Baugh et al., 2018).

The c.646G>A; p.Val216Met variant identified in UPSb-T5 was also reported in one UPSb tumour and one osteosarcoma . A study by Kawaguchi et al. (2002) found four missense alternations, including the c.646G>A; p.Val216Met variant, in 4 of 18 UPSb samples (22.2%). A WES study by Joseph et al. (2014) reported the c.646G>A;

p.Val216Met variant in a patient aged 16 years and diagnosed with primary osteosarcoma (small cell) of the femur.

UPSb ID	Variant details	COSMIC ID	Reported #	Top five cancer tissues associated with *TP53* alterations	Reported in any bone cancer
T3	c.473G>A p.Arg158 His	COSM1640853	27	(1) Central nervous system; (2) Large intestine; (3) Upper aerodigestive tract; (4) Liver; (5) Stomach	None
T5	c.646G>A p.Val216 Met	COSM10667	91	(1) Haematopoietic and lymphoid; (2) Breast; (3) Oesophagus; (4) Upper digestive tract; (5) Central nervous system	(1) UPSb; OS
T6	c.707A>G p.Tyr236 Cys	COSM10731	71	(1) Oesophagus; (2) Ovary; (3) Breast (4) Stomach; (5) Lung	None
T9	c.724T>G p.Cys242 Gly	COSM3717645	9	(1) Liver; (2) Lung; (3) Skin; (4) Breast; (5) Large intestine	None

Table 5–5: COSMIC findings of the recurrent *TP53* somatic variants identified in four UPSb samples. Three *TP53* variants (underlined) are reported among the 50 most common somatic mutations in *TP53* from the International Agency for Research on Cancer database (http://p53.iarc.fr) (Table compiled from information reported in the COSMIC database http://cancer.sanger.ac.uk/cosmic). Os: osteosarcoma.
#This column represents the total number of times the variant was reported as somatic in various cancer tissues in COSMIC.

5.6.1.1.3 Localisation of recurrent *TP53* variants in p53 DNA-binding domain and their biological relevance in literature

Most *TP53* mutations reported in cancer occur within the p53 DNA-binding domain between amino acids residues 95–288 (Leroy et al., 2014). All *TP53* mutations identified in this study also occurred within the p53 DNA-binding domain (Figure 5–10). These amino acids provide specificity to DNA binding—a gene expression regulator of diverse cellular processes (Leroy et al., 2014). *TP53* missense mutations located within the DNA-binding domain can therefore alter the amino acid residues that are involved in DNA binding, resulting in a conformational change of the DNA-binding domain and consequently contributing to tumourigenesis (Baugh et al., 2018).

A study by Zerdoumi et al. (2017) assessed the p53 transcriptional level response to DNA damage in p.Arg158His (same variant as in UPSb-T3) and found a reduction in p53 transcriptional activity, similar to that observed in null mutations. In addition, the authors found a drastic alteration of p53 transcriptional activity in p.Val216Met (same variant as in UPSb-T5), similar to that found in well-studied dominant-negative missense changes. These findings illustrate the bona fide detrimental impact of p.Arg158His and p.Val216Met on p53 transcriptional response to DNA damage.

A study by Neskey et al. (2015) functionally characterised (*in vitro*) *TP53* mutations, one of which is the p.Tyr236Cys identified in the current study. The p.Tyr236Cys showed similar characteristics to wild-type p53 in invasion and cell proliferation assays suggesting that this variant may maintain some tumour suppressive functions.

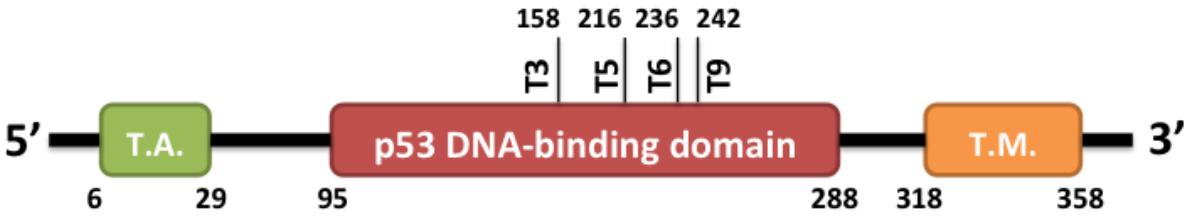


Figure 5–10: The location of the *TP53* somatic variants in relation to p53 functional domains. The three major functional domains of the p53 protein (transcript ID: ENST00000413465) are shown: p53 transactivation domain (T.A.), p53 DNA-binding domain and p53 tertramerisation domain (T.M.). All *TP53* variants identified in UPSb (vertical black lines) occur in the DNA-binding domain. Original figure, compiled from information in Ensembl database).

5.6.1.2 *ATRX* cancer-associated gene

5.6.1.2.1 Overview

ATRX is classified as a tumour suppressor gene based on the nature of its mutations, gene length and gene features (Kadoch and Crabtree, 2015). ATRX protein is as a member of the SWI/SNF2 (SWItch/Sucrose Non-Fermentable) ATP-dependent chromatic remodelling protein complex (Nandakumar et al., 2017). Along with histone modifying enzymes, SWI/SNF2 remodelling proteins regulate gene expression activation and repression of thousands of genes in a coordinated process through remodelling of the chromatic structure (Prasad et al., 2015; Tang et al., 2010). Moreover, ATRX protein is involved in cell cycle regulation and maintenance of genome stability (Nandakumar et al., 2017).

ATRX mutations are frequently observed in astrocytic (brain) tumours (Ikemura et al., 2016). *ATRX* mutations are potential cancer-driving alternations in gliomas (Nandakumar et al., 2017), small cell lung cancers (Kadoch and Crabtree, 2015) and six adult soft tissue sarcomas (including UPS, leiomyosarcoma, dedifferentiated liposarcoma) (The Cancer Genome Atlas Research Network, 2017).

5.6.1.2.2 Potential biological relevance of recurrent *ATRX* variants identified in UPSb tumours

Two somatic *ATRX* variants (predicted deleterious by SIFT and PolyPhen2 tools) were identified in two different UPSb tumours (14.3% of UPSb tumours) (Table 5–4). The first variant, c.6326G>T; p.Ser2109Ile, is identified in UPSb-T6. The second variant, c.1052A>T; p.Glu351Val, is detected in UPSb-T14 and is reported somatic in COSMIC database (COSMIC ID: COSM6608613) in a large intestine carcinoma sample.

The first variant (p.Ser2109Ile) is located in the helicase/ATPase domain of ATRX (Figure 5–11). The helicase/ATPase subunit, part of the SWI/SNF2 complex, is the enzymatic core and molecular motor of the ATRX protein (Gibbons et al., 2008). The ATPase subunit is made of conserved domains that are involved in ATP hydrolysis, which enables SNF2 protein members including ATRX to translocate along nucleic acids and catalyse chromatic structural transformations (Narlikar et al., 2013; Mitson et al., 2011). It has been shown that ATRX, along with other SWI/SNF2 protein members, is recruited to sites of DNA damage, suggesting a role in genetic stability of tumours (Koschmann et al., 2016). Hence, the p.Ser2109Ile variant may alter the conformation or structure of the ATRX ATPase region and subsequently interfere with the catalytic activity of the SWI/SNF2 complex and DNA damage response.

The second variant (p.Glu351Val) is located at the enhancer zeste homologue 2 (EZH2) interacting regions of ATRX (Figure 5–11). EZH2 is the catalytic subunit protein of the PRC2 complex, which mediates the histone methyltransferase activity of Polycomb repressive complex 2 (PRC2) (Kaneko et al., 2010). A potential oncogenic activity of EZH2 has been highlighted in the development and progression of various cancer types (Comet et al., 2016). Interaction between ATRX and EZH2 proteins has been identified through the EZH2 interaction region (Ratnakumar and Bernstein,

2013). The p.Glu351Val variant may therefore interfere with ATRX and EZH2 protein interactions and subsequently disrupt the stability of the PRC2 complex. PRC2/EZH2 inhibitors are currently in clinical trials but its efficacy is still not very clear due to the complex oncogenic and tumour suppressive roles of PRC2 protein (Comet et al., 2016). Although further pathogenic understanding of the p.Glu351Val *ATRX* variant effect is required, PRC2/EZH2 inhibitors can be potentially used as a targeted therapeutic agent.

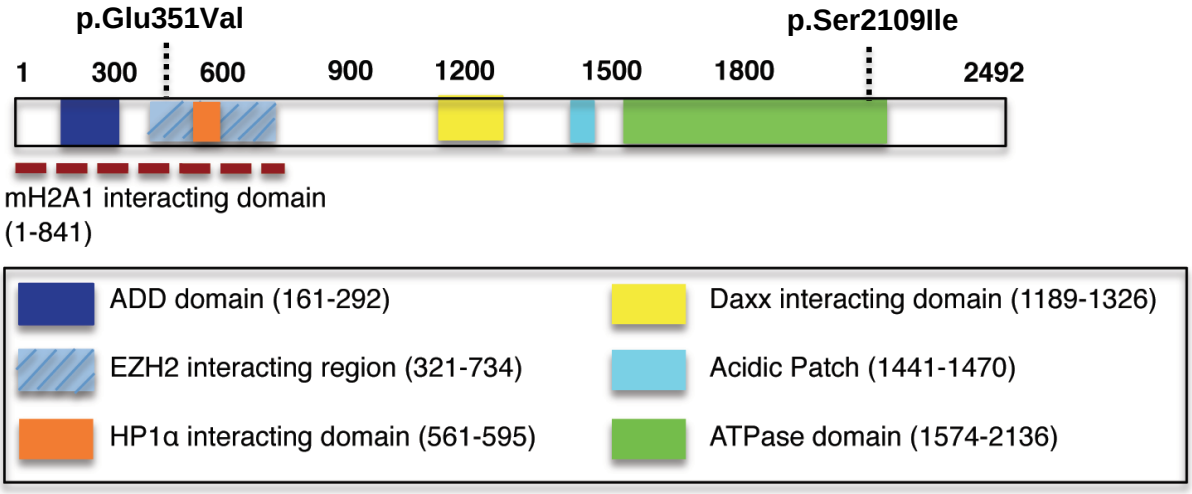


Figure 5–11: Multi domains of the ATRX protein. The p.Glu351Val variant resides in the enhancer zeste homologue 2 (EZH2) protein interaction regions. The p.Ser2109Ile variant lays in the ATPase/Helicase domain. Figure adapted from (Ratnakumar and Bernstein, 2013).

5.6.1.3 *H3F3A* gene

An in-frame insertion (c.106_107insTTC; p.Val36_Lys37insLeu) and a missense change (c.104G>T; p.Gly35Val) in *H3F3A* were identified in UPSb-T1 and UPSb-T10 tumours, respectively (Table 5–4). Both of these changes are reported in the COSMIC

database in giant cell tumours of bone but with unknown somatic status. However, both variants were confirmed somatic in the current study (Table 5–2).

A study by Behjati et al. (2013) showed highly specific tumour-type mutations in *H3F3B* and *H3F3A* in 73/77 (95%) cases of chondroblastoma and 49/53 (92%) of giant cell tumours of bone, respectively. Mutations in chondroblastoma predominantly affected the K36M amino acid, while 48 cases of giant cell bone tumours harboured a G34W change. A low frequency of *H3F3A/B* mutations were observed in osteosarcoma (2%; 2/103), conventional chondrosarcoma (1%; 1/75) and clear cell chondrosarcoma (7%, 1/15); however, no H3.3 changes were identified in chondromyxoid fibroma (n=43), chordoma (n=25) or soft tissue/synovial chondroma (n=7). A WES study by Schwartzentruber et al. (2012) identified recurrent mutations in 31% of glioblastoma multiforme tumours involving K27M or G34R/G34V missense changes in *H3F3A*. The absence of truncating mutations and the non-random identification of the same mutations in various cancers suggest that these *H3F3A* variations are likely of a gain-of-function phenotype (Schwartzentruber et al., 2012).

The Schwartzentruber et al. (2012) study identified a co-occurrence of *ATRX* mutations in 100% of glioblastoma multiforme tumours harbouring G34R/V changes in H3.3. In addition, the authors detected somatic *TP53* mutations in 86% of tumours with *H3F3A* or *ATRX* mutations. In this study, neither *ATRX* nor *TP53* variants were identified in UPSb tumours harbouring *H3F3A* variations.

The correlation of H3.3 mutations in multiple cancers has sparked researchers to assess the biological functional of H3.3 mutations in, for example, post-translation regulations (Kervarrec et al., 2017). Lysine amino acid residue (K) is a critical site and target for post-translational methylation (Behjati et al., 2013). H3.3 lysine to methionine

(K36M), G34R and G34V mutations result in a reduced methylation of K36 residue through the inhibition of SET domain-containing enzymes (Behjati et al., 2013).

Here, the p.Val36_Lys37insLeu and p.Gly35Val variants are particularly of biological interest because they are near or at positions in the highly conserved amino-terminal tail of the H3.3 protein that is associated with essential post-translational modifications, specifically activation at the K36 residue (Figure 5–12) (Behjati et al., 2013). The p.Val36_Lys37insLeu H3.3 variant inserts a leucine between valine and lysine amino acids, which can therefore interfere with leucine post-translational methylation sites.

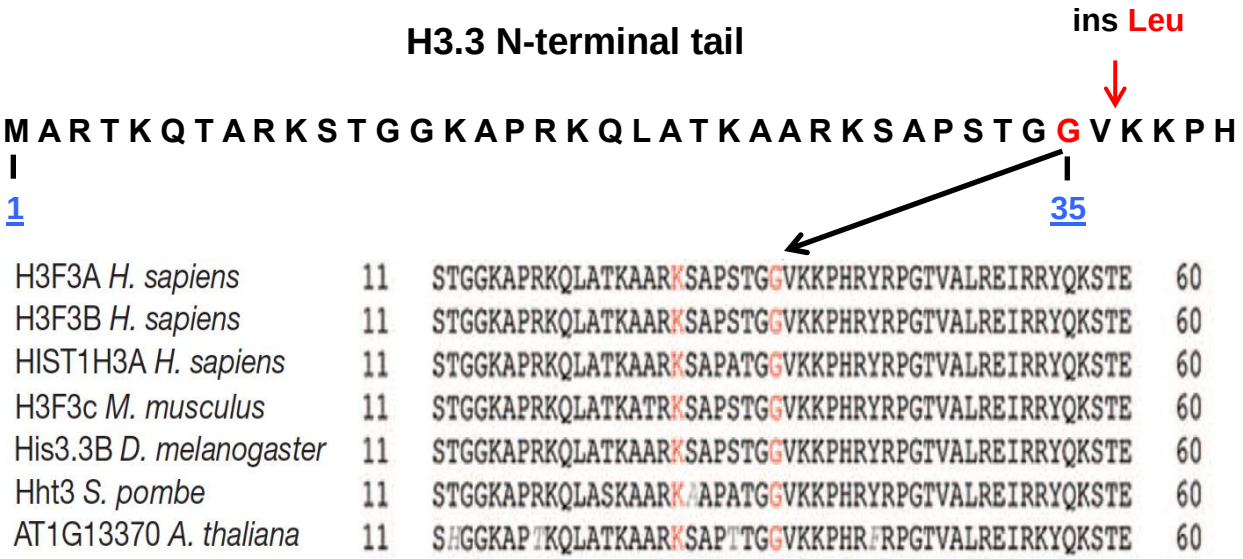


Figure 5–12: The location of the *H3F3A* variants in the N-terminal tail of H3.3. Top raw letters represent the translated amino acid of *H3F3A* where methionine (M) is residue 1. The p.Gly35Val p.Val36_Lys37insLeu variants identified in UPSb are in red within the common post-translational modification of H3.3 N-terminal tail. Amino acids 11 to 60 are highly conserved across mammals and plant species including the G, V and K amino acid residues. Image adapted from (Schwartzentruber et al., 2012).

5.6.1.4 Limitations of part 1 WES study and WES technology

The processing and prioritisation of the large volume of data produced by WES requires extensive data processing, which can be challenging especially in drawing biological significance and conclusions (Bertier et al., 2016). The use of several

bioinformatics tools and algorithms has facilitated the analysis of WES data and the identifications of potential candidate driver genes (Hofree et al., 2016). However, it can be challenging to precisely pinpoint the actual driver gene(s), especially when the list of recurrent genes is quite large as in UPSb tumours (n=31 recurrent genes).

The identification of somatic variants in exome-sequenced tumours without their matched normal can be challenging. The DNA from matched normal samples was used in laboratory validation, allowing us to establish stringent criteria to differentiate variants that are highly likely somatic from those that are germline. However, validating the somatic status of all called variants in matched normal DNA is very laborious and time consuming. Although the majority of somatic variants have <50% VAF, some somatic variants can be present at higher than 50% VAF especially homozygous variants. Therefore, the 50% VAF cut-off applied to tumours without exome-sequenced corresponding normal tissue could be too stringent and may have discarded true somatic calls (Shain et al., 2015).

A limitation of the UPSb WES study was the rarity of UPSb samples, which led to difficulties in finding a large number of samples. Additional UPSb tumours can serve as a validation set to further validate the recurrent genes identified in this WES study. By focusing on high frequency recurrent genes, a validation set can help discard low frequency recurrent genes and identify genuine tumourigenesis-related genes.

One of the biggest limitations of WES technology is that it only focuses on ~1.5% coding regions of the genome and therefore will likely miss non-coding variants. As mentioned previously, WGS analyses the entire genome and may be beneficial especially in UPSb-T8, where WES failed to identify protein-coding somatic changes. Nevertheless, WGS remains relatively costly and requires extensive data processing and bioinformatics analysis with complex algorithmic pipelines (Payne et al., 2018).

Moreover, interpretations and drawing meaningful biological conclusions of non-coding variants remain difficult (Belkadi et al., 2015).

5.6.2 Discussion part two: RNA-Seq findings

RNA-Seq studies provide a comprehensive analysis of the transcriptome, revealing insights into transcriptomic alterations related to cancer pathogenesis. Unlike DNA, which is highly identical across all cells, RNA provides a transcriptome representation that reflects the unique biology and regulatory mechanisms of cells. Transcriptomic profiling not only reveals gene expression profiling but also provides insights of the structure and variations in individual transcripts (Cieslik and Chinnaiyan, 2018). Gene fusions, occurring as a result of chromosomal rearrangements, represent an important tumour-related subtype of genetic alterations. As stated earlier, because of the overlapping clinical presentations, distinguishing between UPSb and dedifferentiated chondrosarcoma or osteosarcoma can be challenging for pathologists and clinicians (Chen et al., 2017a). Hence, the identification of cancer-specific gene fusions can provide insights into tumourigenesis mechanisms and enable better diagnosis and patient management, especially for heterogeneous tumours such as UPSb.

Gene fusions have been associated with almost one third of soft tissue neoplasms and approximately half of these gene fusions are recurrent (Mertens et al., 2016). Gene fusions involving *PRDM10* and *MED12* or *CITED2* gene partners have been identified at low frequency (5%) in UPS of soft tissue (Hofvander et al., 2015). A more recent study by Delespaul et al. (2017) identified *TRIO* with *LINC01504* or *ZNF558* as a recurrent gene fusion in four different non-translocation-related soft tissue sarcomas, including UPS of soft tissue. In the current study, no recurrent somatic gene fusion has been identified in the UPSb. However, two gene fusions that were previously reported

in other cancers were identified in two UPSb tumours, *CLTC-VMP1* and *FARP1-STK24*.

5.6.2.1 *CLTC-VMP1* gene fusion in sample UPSb-T9

A gene fusion involving the clathrin heavy chain (*CLTC*) and vacuole membrane protein 1 (*VMP1*), *CLTC-VMP1,* was confirmed somatic in UPSb-T9 at the cDNA and DNA levels using RT-PCR (Section 5.4.2.1) and LR-PCR (Section 5.4.3.1), respectively. This gene fusion transcript formed as a result of an interstitial chromosomal deletion (Section 5.4.3.1) joining the first 14 exons of *CLTC* (NCBI Reference Sequence NM_001288653.1) to the last 2 exons of *VMP1* (NCBI Reference Sequence: NM_030938.4) **(**Figure 5–5**).**

A *CLTC-VMP1* fusion has been previously reported in BT-549 breast cancer cell lines (Robinson et al., 2011a; Giacomini et al., 2013), HCC1954 breast cancer cell lines (Giacomini et al., 2013), hypopharynx tumour (Nair et al., 2015), and large-cell lung carcinoma (Liu et al., 2012). In studies by Giacomini et al. (2013); Nair et al. (2015), the reported cDNA breakpoint for *CLTC* was at the end of exon 27 in HCC1954/BT-549 cell lines and at exon 15 in the hypopharynx tumour; by contrast, the *VMP1* cDNA breakpoint was at the beginning of exon 11 in both the BT-549 cell line and hypopharynx tumour. In the UPSb tumour, the *CLTC* breakpoint is different from the previously mentioned breakpoint; however, the *VMP1* breakpoint is the same as in the BT-549 (Giacomini et al., 2013), HCC1954 cell lines and hypopharynx tumour.

In RNA splicing, intronic sequences are spliced/removed from the nascent mRNA, and exons are joined together to form mature mRNA that can be subsequently translated into amino acid sequences. Most gene fusions use the canonical GT-AG splicing rule for donor and acceptor sites to produce a chimeric transcript by a splicing reaction

(Babiceanu et al., 2016; Sibley et al., 2016). The *CLTC-VMP1* chimeric transcript identified in this study is produced following canonical GT-AG splicing rule (Figure 5–13A).

The *CLTC-VMP1* fusion was described as out-of-frame in the BT-549 and HCC1954 cell lines (Giacomini et al., 2013), and hypopharynx tumour (Nair et al., 2015). The *CLTC-VMP1* fusion identified in UPSb-T9 is also out-of-frame. Because the fusion is out-of-frame, a premature stop codon (TGA) is introduced after 75 amino acids of the VMP1 protein, downstream from the *CLTC* cDNA breakpoint (Figure 5–13B). That is, the first 764 amino acids of the *CLTC* partner is joined to the 75 amino acids of *VMP1,* forming a total chimeric protein length of 749 amino acids.

5.6.2.1.1 Potential tumour suppressor role of *CLTC-VMP1*

The *CLTC-VMP1* fusion results in a loss of six repeats of the clathrin heavy chain/VPS 7-fold repeats domain of *CLTC* (Figure 5–13B). Clathrin is made of 7 heavy chain repeats, which form a polyhedral protein lattice that covers the intercellular membrane involved in endocytosis, lysosomal degradation and intercellular trafficking (Wakeham et al., 2003). Moreover, the clathrin protein is involved in chromosome segregation and Golgi reassembly during mitosis, protein-protein interactions, and protein sorting (Young, 2007). The loss of six clathrin heavy chain repeats in *CLTC-VMP1* chimeric transcript can therefore have an adverse effect on cellular integrity.

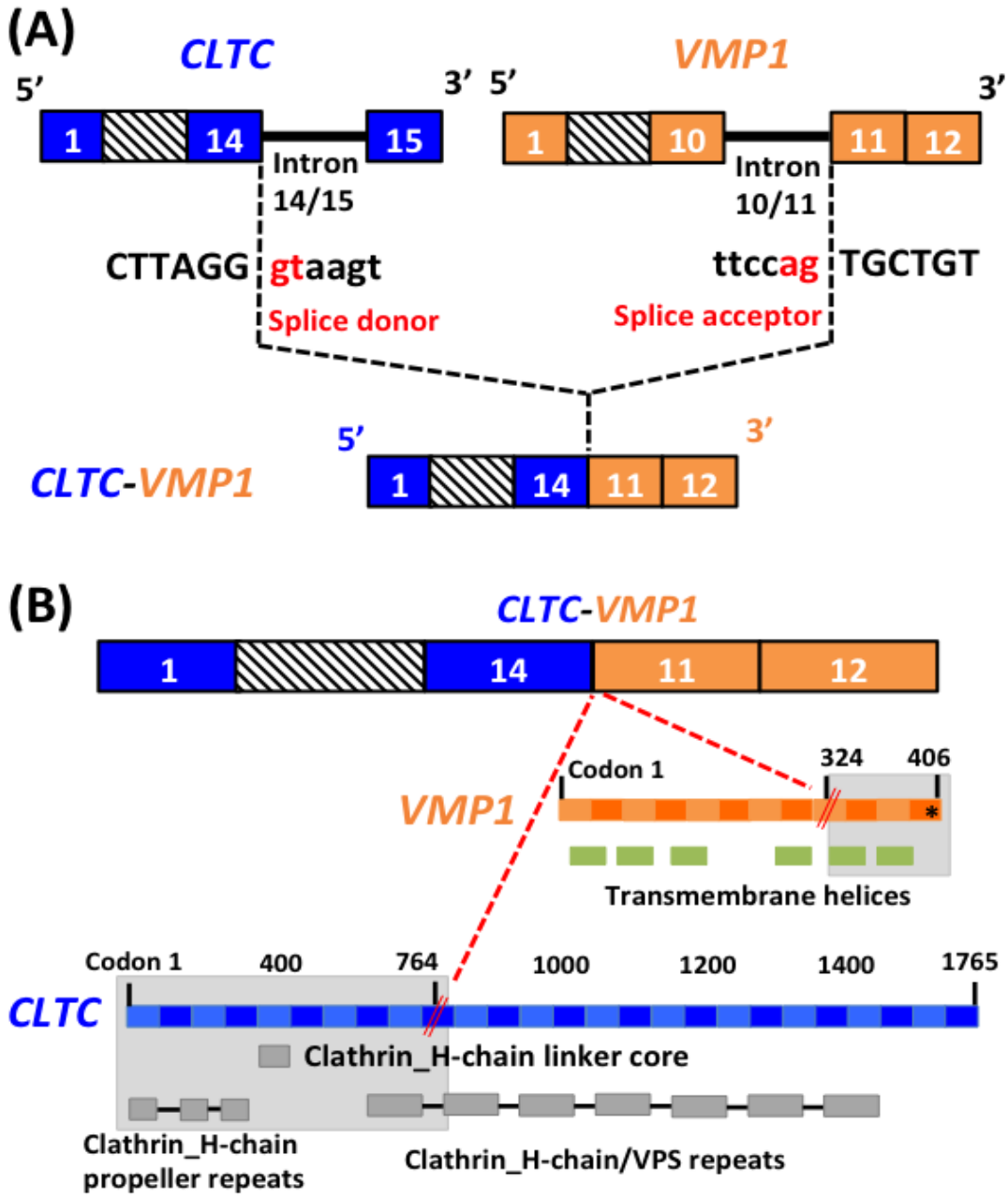


Figure 5–13: Canonical splicing of *CLTC-VMP1* gene fusion and the affected functional biological domains of the fusion gene partners. (A) Exons of *CLTC* and *VMP1* genes are in blue or orange boxes, respectively. As shown, *CLTC-VMP1* gene fusion follows the standard GT-AG canonical splicing rule, joining exons 14 of *CLTC* to exon 11 of *VMP1*. (B) Representation of the functional domains present in CLTC and VMP1 protein products. The cDNA breakpoint for each gene partner is denoted by red double slash (//). The grey shaded area represents the retained domains of each fusion partner. The black asterisk (*) represent the premature codon introduced in the chimeric transcript as a result of the fusion being out of frame.

As stated earlier, the genomic breakpoint analysis of *CLTC-VMP1* fusion revealed an interstitial deletion leading to a complete deletion of the *PTRH2* gene. Chromosomal rearrangements involving *CLTC-PRTH2-VMP1* genes have been described previously in multiple cancer entities, including glioblastoma, lung cancer, breast cancer and leukaemias (Giacomini et al., 2013). PTRH2 is a mitochondrial protein and a part of an integrin-signalling pathway that regulates cell survival and death (apoptosis) (Giacomini et al., 2013; Jan et al., 2004). Overall, the genomic rearrangement of the *CLTC-PRTH2-VMP1* locus and the out-of-frame characteristic of *CLTC-VMP1* suggest that one or more of these three genes may function as a tumour suppressor (Giacomini et al., 2013).

5.6.2.2 *FARP1-STK24* fusion identified in UPSb-T6

A gene fusion involving FERM, ARH/RhoGEF and pleckstrin domain protein 1 (*FARP1*) and serine/threonine kinase 24 (*STK24*) was confirmed somatic in UPSb-T6 by RT-PCR (Section 5.4.2.2) and LR-PCR (Section 5.4.3.2). The *FARP1-STK24* gene fusion was previously identified in an invasive breast cancer tumour; however, the fusion breakpoints were not reported (Veeraraghavan et al., 2014). The *FARP1-STK24* gene fusion is in-frame, joining the first 88 amino acids of *FARP1* (NM_001001715.3) and 211 amino acids of *STK24* (NM_003576.4). The fusion results in a chimeric protein of 299 amino acids in length **(**Figure 5–14**)**. The chimeric transcript is spliced following the CT-AC donor-acceptor splicing rule, the reverse complementary of standard GT-AG canonical splicing.

The *FARP1* and *STK24* genes are transcribed from the positive (sense) or negative (antisense) strands, respectively. In the *FAPR1-STK24* fusion, the two gene partners are joined together in opposing orientation, classifying the fusion as 5′-to-5′ (sense-to-

antisense) fusion orientation. Approximately half of the cancer-associated genomic rearrangements link genes in opposite orientations (5′-to-5′ or 3′-to-3′) (Shlien et al., 2016). In 5′-to-5′ (sense-to-antisense) fusions, the 5′ transcription regulatory apparatus is retained in both genes, and therefore, both sense and antisense genes can start transcription that would extend into the other gene partner (Shlien et al., 2016). Theoretically, in *FARP1-STK24*, transcription can start from *FARP1* (sense) and extend to *STK24* (antisense), or vice versa.

The potential biological function of sense-to-antisense fusions is not very clear; however, some evidence in the literature supports their potential pathological role. Shlien et al. (2016) found 50 5′-to-5' gene fusions in 23 breast cancers (44%), linking the sense gene fusion partner with an intronic, exonic or a partially exonic antisense partner. The authors also reported a gene fusion linking the oestrogen receptor gene *ESR1* with multiple spliced antisense transcripts involving *SYNE1*. Recurrent gene fusions involving *ESR1* gene partner are described in invasive breast tumours, highlighting their pathogenic role (Li et al., 2013; Veeraraghavan et al., 2014).

5.6.2.2.1 Potential apoptotic and oncogenic roles of *FARP1* and *STK24* genes

Aberrant gene expression of *FARP1* has been observed in multiple cancers. A significant decrease in *FARP1* expression has been found in pheochromocytomas (Croise et al., 2016). By contrast, a substantial increase in *FARP1* expression is detected in breast cancers (Croise et al., 2016). *FARP2* is a close homologue of *FARP1* that shares the same FARP1 domain structure and has a role in bone homeostasis by activating RAC1 or CDC42 (He et al., 2013). The knock down of FARP1 in the rat adrenal gland cell line (PC12) leads to a significant inhibition of Cdc42 activity, a gene that has been linked to cellular processes associated with cancer

biology (Croise et al., 2016). Although this finding suggests that FARP1 may also regulate bone homeostasis or alter cancer biology, further studies are needed to confirm these roles.

The ~219kb genomic interstitial deletion within *FARP1* and *STK24* leads to the deletion of almost all three main functional domains of *FARP1*, 4.1, ezrin, radixin, and moesin (FERM), RhoGEF, pleckstrin homology (PH) (He et al., 2013) (Figure 5–14). Although the majority of *STK24* kinase domain is intact, the regulatory domain is missing (Tsatsanis et al., 2007) (Figure 5–14).

FARP1 is a member of the Rho guanine nucleotide exchange factor (Rho GTPase) family, which belongs to the Ras superfamily (Lawson et al., 2016). Rho GTPases have been strongly implicated in many fundamental cell processes, including cell cycle integrity, cell migration and apoptosis (Croise et al., 2016; Lawson et al., 2016). Altered Rho GTPase activity is associated with the hallmarks of cancer mechanisms, including oncogenic transformation and tumour suppression (Orgaz et al., 2014). Complex signalling cascade signals are responsible for regulating the activation and inactivation of Rho GTPases (Croise et al., 2016). The active GTP-bound form of Rho GTPases is presumed to drive the oncogenic phenotype when overexpressed or aberrantly activated. By contrast, the GDP-bound inactive form is believed to inhibit tumourigenesis (Lawson et al., 2016).

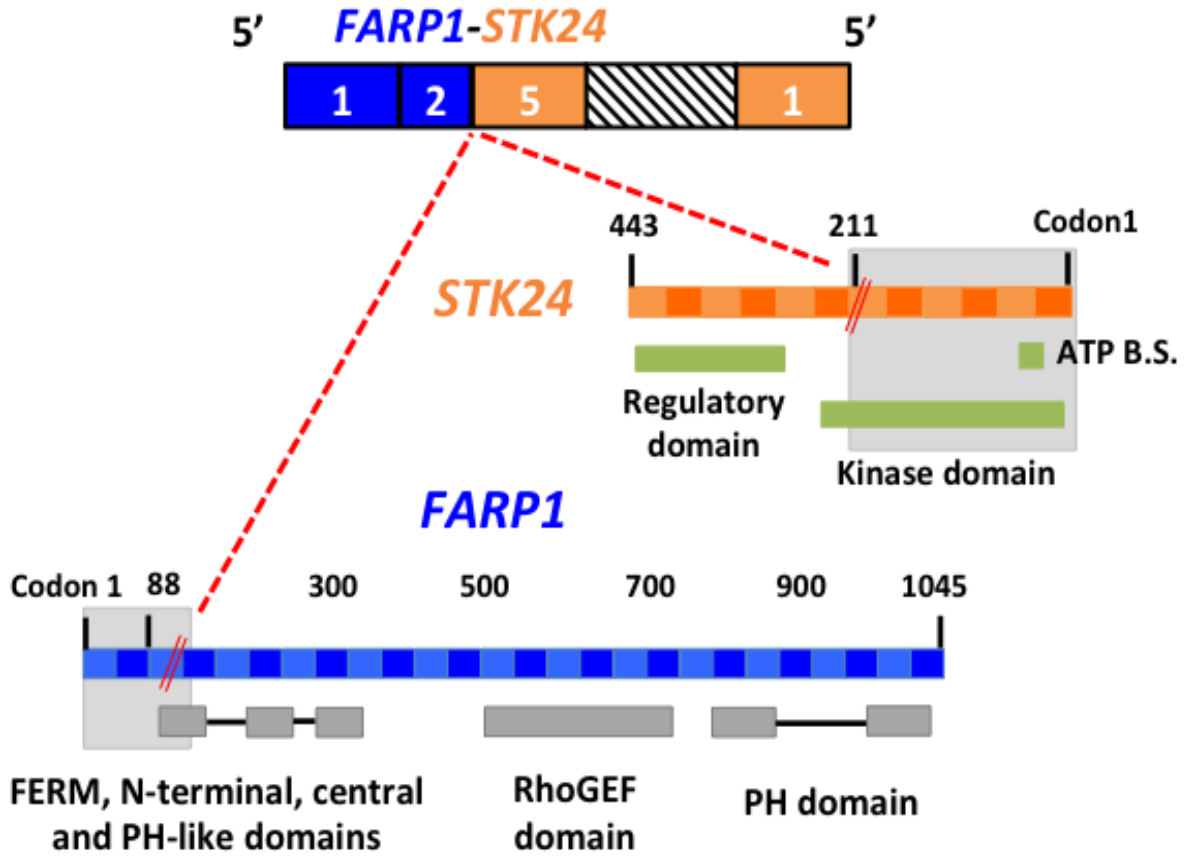


Figure 5–14: The impact of *FARP1-STK24* fusion on the functional domains of FARP1 and STK24 proteins. The grey shaded area represents the retained portions of functional domains of *FARP1* and *STK24* genes. The red double slash sign (//) represents the cDNA breakpoint. ATP B.S.: ATP binding site.

The role of Rho GTPases in cancer is complex and involves various regulators and signalling pathways. Although several studies demonstrated the association of aberrant Rho GTPase expression in cancer tumours and metastases, the precise role of Rho GTPase in tumourigenesis is not fully understood (Croise et al., 2016). The complete loss of RhoGEF and PH functional domains in *FARP1-STK24* may interfere with their biological activity and interactions with other substrates. Deciphering the

precise pathological role of the *FARP1* in *FARP1-STK24* would require further functional characterisation experiments.

STK24, also known as *MST3*, is a serine/threonine protein kinase that belongs to the mammalian Sterile20-like (MST) kinase family, a key group of signalling molecules that regulate cancer-associated cellular processes, including cell division cycle, cell morphogenesis, apoptosis and oncogenic transformation (Cho et al., 2016; Thompson and Sahai, 2015; Tsatsanis et al., 2007). Oncogenic proteins contribute to cellular functions by inducing signals from the extracellular compartment into the cytoplasm and finally toward the nucleus, leading to transcription initiation of proteins associated with oncogenic functionality (Tsatsanis et al., 2007; Cho et al., 2016). Elevated expression level of *STK24* has been linked to aggressive breast cancer subtypes, suggesting a potential oncogenic role of *STK24* (Thompson and Sahai, 2015).

On the other hand, a caspase-dependent apoptotic role of STK24 protein has been recognised. Caspase-3 protein cleaves STK24 at the boundary region between the protein kinase and the regulatory domains of STK24 protein (Fuller et al., 2012). Once cleaved and activated, STK24 is translocated into the nucleus (Cho et al., 2016). Consequently, the accumulation of the active STK24 kinase domain in the nucleus can promote apoptosis responses (Olesen et al., 2016).

The dual apoptotic and oncogenic roles of STK24 may complicate its precise pathological function in *FARP1-STK24*. However, based on the altered domains and biological functionality of *STK24*, two theories can explain the potential pathological role of *STK24*. First, the loss of the STK24 boundary sequence and regulatory domain may hinder cleavage/activation and nuclear accumulation of STK24 protein and, subsequently, interfere with its apoptotic role. Second, the loss of the regulatory domain and retention of the catalytic kinase domain of STK24 may lead to the

abnormal accumulation of altered ST24 protein in the extracellular environment, leading to an undesired initiation of an oncogenic mechanism.

5.6.2.3 Potential targeted therapeutics for fusion genes composed of protein kinase genes

Gene fusions constituted of kinase proteins have received significant attention as potential therapeutic targets. The use of specific kinase inhibitors in tumours harbouring kinase-related gene fusions can improve tumour prognosis and patient outcomes (Tamura et al., 2015). For example, the use of ALK inhibitors in patients with non-small cell lung cancer harbouring *EML4-ALK* fusion has resulted in significant progression-free survival and a reduction of cancer symptoms in clinical trials (Solomon et al., 2014). Thus, gene fusions that are composed of in-frame kinase transcripts retaining kinase catalytic activity, such as *FARP1-STK24*, can potentially lead to the identification or development of novel targeted therapeutics. A study by Olesen et al. (2016) identified 14 previously unknown inhibitors of *STK24*, some of which are currently used in cancer therapeutics that resulted in enhanced cancer cell apoptosis. STK24 inhibitors can be a potential targeted therapy for tumours harbouring *STK24* fusions; nonetheless, additional studies are required to assess the precise pathological role and efficacy of STK24 inhibitors.

5.6.2.4 Limitations of RNA-Seq technology

RNA-Seq utilises the capability of high-throughput sequencing to provide a quantitative and exploratory analysis of the whole transcriptome, including the identification of gene fusions (Byron et al., 2016; Wang et al., 2013). In contrast to hybridisation-based

technologies (e.g., FISH and aCGH), RNA-Seq provides a quantitative analysis of the whole transcriptome at a single base-pair resolution (Han et al., 2015).

Although RNA-Seq itself is considered unbiased, RNA sample preparation and library construction may introduce undesired bias (Conesa et al., 2016). RNA library construction targets the total mRNA using polyadenylation (Poly-A) enrichment of mRNA. The poly-A capturing of mRNA material does not allow for adequate capturing of partially degraded or low-quality mRNA samples (Byron et al., 2016). Although none of the RNA-sequenced UPSb samples were FFPE stored, library construction of FFPE samples can be challenging, and therefore, optimisation and evaluation of library construction protocols in these samples are needed. Fragmentation of RNA during library preparation can be another source of bias. During RNA fragmentation, the library construction favours internal regions of transcripts while producing shorter fragments of transcript ends, which can get lost during size selection steps, leading to non-uniform coverage across the entire transcript length (Jiang et al., 2015).

Even after applying heuristic bioinformatic filtering steps, artefacts resulting from misalignment of reads, such as in highly homologous genes, can still be called (Conesa et al., 2016). This issue can be partially resolved by manual inspecting and visualising the uniformity of mapped reads across gene fusion junctions and gene fusion partner genes. However, this manual inspection is time consuming when the called gene fusion list is long.

As with any high-throughput technologies, the large-scale processing and analysis of data such as read alignments and fusion calling require sophisticated algorithmic pipelines. Although numerous NGS computational frameworks have been developing, the large number of generated data can make data analysis and identification of gene fusion challenging and prone to errors. Therefore, further refinement of data analysis

applications and establishment of benchmark gold standards are needed, especially when translating RNA-Seq into clinical and diagnostic applications (Han et al., 2015; Wang et al., 2013).

5.7 Future work

As mentioned previously, finding additional UPSb samples can serve as a validation set to confirm and expand the WES and RNA-Seq findings of the current study. WES or a more cost-effective targeted exome approach can be utilised to validate WES findings. WGS may be considered in UPSb-T8 sample that lacks any WES alterations. On the other hand, CNV analysis can also be performed to identify somatic chromosomal gains or losses among UPSb samples. As of August 2018, Prof Farida Latif has established a collaboration to perform CNV analysis in UPSb tumours.

No recurrent gene fusions were found among UPSb samples. However, it is worth screening the somatic-confirmed gene fusions identified in this study, especially *CLTC-VMP1* and *FARP1-STK24*, in a larger UPSb cohort. Moreover, functional studies can be conducted on these two fusions to characterise their biological role and assess their possible drug target ability.

A collaboration to perform the analysis of differentially expressed genes using RNA-Seq data in UPSb has been established (more in Chapter 7). Determining the expression profiles of the UPSb can identify a molecular biomarker that can be of diagnostic importance. Although corresponding normal samples were not RNA-sequenced, RNA samples from normal samples are available, and differential expression assessment of selective significant genes can be performed using quantitate RT-PCR.

5.8 Peer reviewed publications

The work presented in this chapter has been published in:

> **Ali, N. M.**, Niada S., Brini A.T., Morris, M. R., Kurusamy, K., Alholle, A., Huen, D., Antonescu, C., Tirode, F., Sumathi, V., et al. 2018. Genomic and transcriptomic characterisation of undifferentiated pleomorphic sarcoma of bone. *J Pathol*. 247:166–176.

Part of the work presented in this chapter has been published in:

> Chen, S., Fritchie, K., Wei, S., **Ali, N.,** Curless, K., Shen, T., Brini, A. T., Latif, F., Sumathi, V., Siegal, G. P., et al. 2017. Diagnostic utility of IDH1/2 mutations to distinguish dedifferentiated chondrosarcoma from undifferentiated pleomorphic sarcoma of bone. *Hum Pathol,* 65**,** 239-246.

Chapter 6: Genetic and transcriptomic analyses of adamantinoma and osteofibrous dysplasia-like adamantinoma bone tumours using NGS technologies

6.1 Introduction

Classic adamantinoma (henceforth referred to as adamantinoma) is a low-grade primary malignant bone tumour that arises in long bones (97% of cases), predominately the tibia (~85%) and followed by the fibula (~10%) (Jain et al., 2008; Taylor et al., 2012). The tumour is rare, accounting for 0.1–0.5% of all primary bone tumours, and commonly occurs in the second to fifth decade of life (Jain et al., 2008). Histopathologically, adamantinoma tumours are characterised by a combination of both epithelial and osteofibrous dysplasia components that can be intermixed with each other in various proportions and differentiation patterns (Hatori et al., 2006). Adamantinoma is usually indolent; however, it can be aggressive (Camp et al., 2008). Treatment usually involves a wide resection with a limited use of radiation therapy and chemotherapy (Camp et al., 2008). Long-term follow-up of patients is highly recommended to observe late local recurrence or metastasis (Scholfield et al., 2017). In a clinicopathologic study by Keeney et al. (1989), in 85 adamantinomas of long bones, 26 (31%) had recurrent local disease, 13 (15%) showed lung metastasis, and 6 (7%) developed lymph node metastasis.

Adamantinoma of the long bones can share clinical, radiological and histological similarities with other bone tumours named OFD and OFD-like adamantinoma (also known as differentiated adamantinoma) which encompasses features of adamantinoma and OFD (Camp et al., 2008; Taylor et al., 2012). OFD-like adamantinoma is characterised by a predominance of osteofibrous dysplasia tissue

with small clusters of epithelial cells that are identified upon careful microscopy examination or immunohistochemistry studies (Christopher et al., 2013; Hatori et al., 2006). Similar to adamantinoma, OFD-like adamantinoma commonly affects the tibia and the fibula (Taylor et al., 2012). However, OFD-like adamantinoma frequently occurs at a younger age than adamantinomas (20 years or younger) (Kahn, 2003). The prognosis for OFD-like adamantinoma is usually more favourable than that for adamantinoma tumours (Kahn, 2003; Puchner et al., 2016). Because of the usual benign nature of OFD-like adamantinoma and its occurrence in young children, the treatment opinions are contentious but are usually divided between observation or radical surgery for symptomatic lesions (Scholfield et al., 2017).

There have been unresolved issues about the relationship of adamantinoma, OFD-like adamantinoma, and OFD, suggesting a histogenetic relationship between these lesions. Histologically, OFD-like adamantinoma is proposed to sit between adamantinoma (malignant) and OFD (usually benign) (Scholfield et al., 2017). The principle theory explaining the relationship between these three entities states that OFD may be a precursor to OFD-like adamantinoma, which can be a subsequent precursor to adamantinoma (Scholfield et al., 2017). Some well-documented OFD-like adamantinoma cases reported subsequent progression to the adamantinoma phenotype, suggesting that OFD-like adamantinoma may be the precursor lesion of adamantinoma (Hatori et al., 2006). However, there is no well-established knowledge supporting this theory (Scholfield et al., 2017).

Common cytogenetic alterations involving trisomies 7, 8 12 and/or 21 have been reported in adamantinoma and OFD-like adamantinoma cases, possibly linking the two entities genetically. However, clear cytogenetic profiles that discriminate between the two entities are not described (Camp et al., 2008). Moreover, trisomies 7, 8 and/or 12

can also be found in other bone lesions such as chondrosarcomas and Ewing sarcomas (Gleason et al., 2008). In addition, no molecular diagnostic tests are currently available to differentiate between these lesions. To our knowledge, no genetic profiling studies using NGS technologies have been conducted on adamantinoma or OFD-like adamantinoma tumours. Therefore, expanding the genetic knowledge to differentiate between the entities is nontrivial.

6.2 Aim of the study

As mentioned previously, NGS technologies are widely used in cancer genetic studies. Similar to the aims of the UPSb study, WES and RNA-Seq technologies were used to investigate the genetic (exomic) and transcriptomic alterations, respectively, in eight adamantinoma and four OFD-like adamantinoma tumours. In this study, when referring to sample numberings, adamantinoma and OFD-like adamantinoma samples are abbreviated to ADA and OFD-like-ADA, respectively. Of the eight adamantinoma tumours, six tumours were paired with corresponding normal tissues (ADA-T1 to ADA-T6), whereas two were normal-unpaired tumour samples (tumour only) (ADA-T7 and ADA-T8). Three of the four OFD-like adamantinoma tumours were normal-paired (OFD-like-ADA-T1 to OFD-like-ADA-T3), whereas one tumour (OFD-like-ADA-T4) was unpaired with a corresponding normal sample.

Understanding the landscape of genetic and transcriptomic alterations in these two lesions can provide a broader understanding of tumourigenesis and shed a light on potential personalised therapeutics. Because adamantinoma can be differentially diagnosed as OFD-like adamantinoma, this study can potentially establish molecular diagnosis criteria to differentiate between the two tumour entities and consequently provide better prognosis options.

6.3 Part 1: WES data results and analysis

6.3.1 Details of WES samples, somatic variant callers and quality check metrics

WES was carried out on a total of 14 samples involving adamantinoma tumours and their corresponding normal samples (six normal-paired tumours and two normal-unpaired tumours), as well as seven samples including three normal-paired and one normal-unpaired OFD-like adamantinoma tumours. The clinicopathological information of all cases was reviewed by Dr Sumathi Vaiyapuri (The Royal Orthopaedic Hospital NHS Foundation Trust, University of Birmingham) (see Section 2.1.3).

All samples were exome sequenced by Oxford Gene Technology (Oxfordshire, UK), except for one normal-paired OFD-like adamantinoma tumour sequenced at Prof. Eamonn Maher's laboratory (University of Cambridge, UK). All DNA from tumour and normal tissues was extracted from fresh frozen tissues (see Section 2.2). VarScan2 and MuTect somatic calling tools (Section 2.3.5.2) were used to identify SNVs and insertions and deletions INDELs. VarScan2 was run by OGT (Oxfordshire, UK) for all adamantinoma and OFD-like adamantinoma samples, except for OFD-like-ADA-T3 run by Source BioScience company (Nottingham, UK). For all tumours, the MuTect variant calling tool was run by our collaborator, Dr David Huen (School of Biology, Chemistry and Forensic Science, University of Wolverhampton). The data was delivered as specified earlier in Section 2.3.4.

As mentioned previously, QC metrics are an essential preliminary step to ensure the reliability of WES data (details in Appendix 8.1). The QC metrics for adamantinoma and OFD-like samples are in Appendix Table 8–3. Overall, all samples showed satisfactory QC metrics.

6.3.2 Filtering of WES variants to identify somatic candidate variants

Using VarScan2 and MuTect, a total of 7580 and 78663 unfiltered INDELs and SNVs, respectively, were identified in eight adamantinoma tumours. In four OFD-like adamantinoma tumours, 2510 INDELs and 26149 SNVs were detected (details in Appendix Table 8–10). A stepwise filtering scheme was followed to reduce this large number of variants and identify genuine somatic calls (details in Section 2.3.6). In short, intronic (apart from splice sites) and synonymous variants that are common in the general population (MAF >0.1% in databases) were excluded. By contrast, missense (predicted damaging by at least one *in silico* tool), nonsense, splice site and small INDEL changes were prioritised. An additional filtering step was applied to normal-unpaired tumours (n=3): excluding any reported variants in population frequency databases (regardless of MAF), except for variants with MAF < 0.01% and described somatic in the COSMIC database.

After applying the abovementioned prioritising scheme, the total number of filtered somatic candidate variants for each sample was generated, as presented in Figure 6–1. A total of 21 INDELs and 182 SNVs were identified in eight adamantinoma tumours (Figure 6–1A). SNVs consisted of 156 (85.7%) missense, 19 nonsense (10.4%) and 7 splice site alterations (3.8%).

In four OFD-like adamantinoma tumours, a total of 7 INDELs and 59 SNVs were detected (Figure 6–1B). In these tumours, SNVs were composed of 54 missense, 3 nonsense and 2 splice site changes. Two of the four tumours lacked any WES variants, and one tumour harboured only one INDEL alteration.

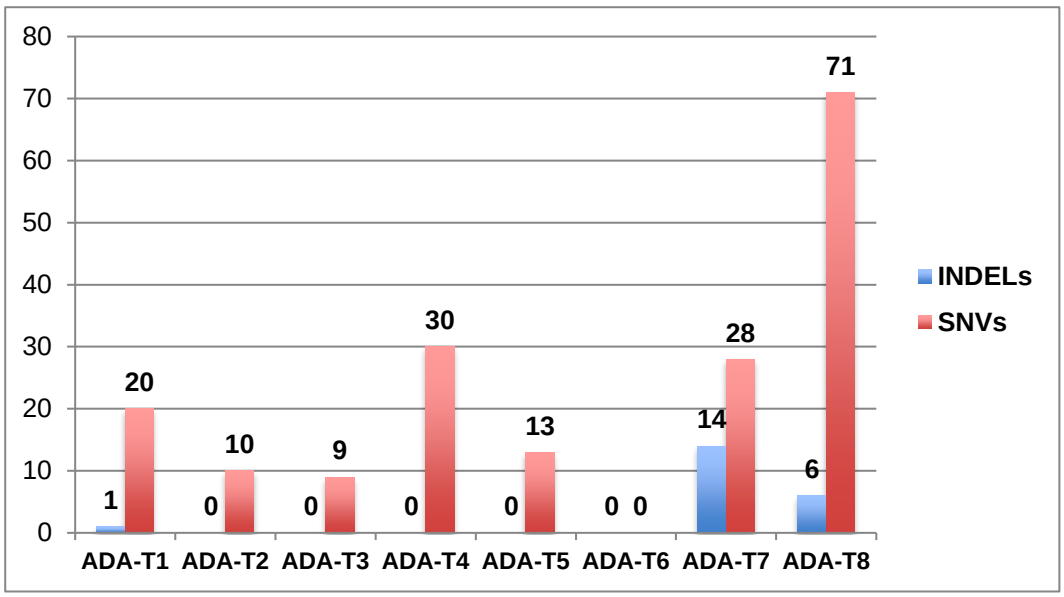
The calculated average coding mutation rate (overall coding somatic mutation burden) for eight adamantinoma is 0.84/megabase (range, 0–2.57) (details in Appendix Table

8–11). For four OFD-like adamantinoma tumours, this average rate was 2.20/megabase; however, three samples showed a mutation rate close to zero, whereas one sample had a mutation rate of 2.20/megabase (see Appendix Table 8–11). Thus, the average coding mutation rate in OFD-like adamantinoma may be overrepresented because of the sample with 2.20/megabase coding mutation rate. A low coding mutation rate average is observed in Ewing's sarcoma (0.42/megabase) (Lawrence et al., 2013) and in 206 soft tissue sarcomas (average 1.06/megabase) (The Cancer Genome Atlas Research Network, 2017), which are consistent with the findings in this study.

6.3.3 Recurrent candidate gene analysis in adamantinoma and OFD-like adamantinoma tumours

KMT2D (MLL2/MLL4) was recurrently mutated in 2/8 adamantinoma tumours (25%) (Table 6–1). The first variant detected in ADA-T2, c.16154C>A; p.Ser5385Ter, is a nonsense substitution described as somatic in the COSMIC database in a squamous cell lung carcinoma tumour (COSM3954819). The second variant identified in ADA-T8, c.10930C>T; p.Pro3644Ser, is a missense substitution that is predicted deleterious by SIFT and PolyPhen2 *in silico* tools.

A



B

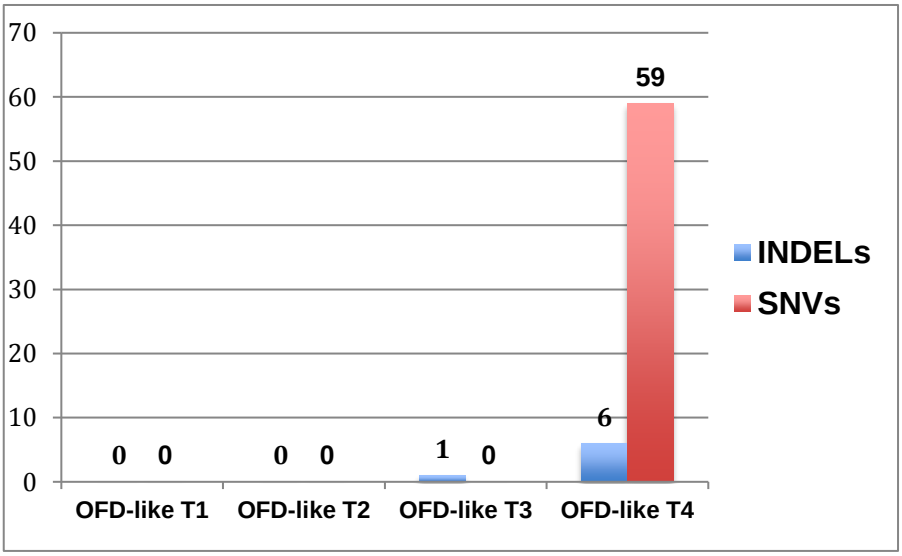


Figure 6–1: The total number of filtered WES variants identified in adamantinoma and OFD-like adamantinoma tumours. The graphs represent the filtered somatic variants identified in (A) eight adamantinoma (ADA) and (B) four OFD-like adamantinoma tumours (OFD-like). SNVs: single nucleotide variants; INDELs: insertions or deletions.

In four OFD-like adamantinoma tumours, no recurrent gene mutations were identified. However, a missense substitution (c.6437C>T; p.Pro2146Leu) in *KMT2D* was identified in OFD-like-ADA-T4 (Table 6–1). This variant is described as somatic in the COSMIC database in acute myeloid leukaemia and malignant melanoma samples (COSM3461578).

KMT2D was investigated in cancer genomic studies databases to infer its potential biological relevance in carcinogenesis. KMT2D is classified as a 'causally implicated gene in cancer' in the CCGC list (https://cancer.sanger.ac.uk/cosmic/census) and as a mutational cancer driver in the inTOgen database (http://intogen.org) in the following cancer types: prostate adenocarcinoma, bladder carcinoma, breast carcinoma, medulloblastoma, lung squamous cell carcinoma, and head and neck squamous cell carcinoma. A potential drug-gene interaction was identified for KMT2D by DGIdb (http://www.dgidb.org/), classifying the gene as potentially drug-targeted.

6.3.4 Ingenuity Pathway Analysis for somatic candidate genes identified in adamantinoma and OFD-like samples

Due to the lack of recurrent mutation in OFD-like adamantinoma and the identification of only one recurrent gene in adamantinoma, a pathway analysis approach was conducted to identify putative enriched canonical pathways, as well as disease and biological functions (bio function), among each tumour subtype. IPA tool transforms the filtered list of genes with potential biological role of each tumour subtype into meaningful molecular pathways/networks using curated literature findings and the IPA knowledge base (Qiagen; http://www.qiagen.com/ingenuity). IPA was run with the experimentally observed confidence filter to obtain a significant analysis output (p-value ≤ 0.05; right-tailed Fisher's exact test) that is based on literature-confirmed

findings. The p-value reflects the likelihood that mutated genes are associated with significant processes/pathways rather than being random chance events. Relevant canonical and bio function pathways identified in adamantinoma and OFD-like adamantinoma tumours are presented in Table 6–2.

6.4 Part 2: RNA sequencing data analyses and results

6.4.1 Overview of the RNA-Seq experiment and data analysis

The aim of the RNA-Seq study was to identify gene fusions in adamantinoma and OFD-like adamantinoma tumours. RNA specimens from five fresh frozen adamantinoma and three OFD-like adamantinoma tumours were transcriptome-sequenced using RNA-Seq. RNA-Seq was carried out on ADA-T1, ADA-T3, ADA-T4, ADA-T5 and ADA-T8 adamantinoma tumours and OFD-like-ADA-T2 through OFD-like-ADA-T4 OFD-like adamantinoma tumours. RNA from the corresponding normal tissues were not RNA-sequenced but were available for laboratory confirmation tests. RNA-Seq was performed by the Genomics Birmingham facility to achieve a minimum of 50 million read coverage (75 bp read length) per tumour sample (Institute of Cancer & Genomic Sciences, University of Birmingham) (see Section 2.4). Subsequent data analyses were conducted using the Illumina BaseSpace hub (https://basespace.illumina.com) (see Section 2.4.3).

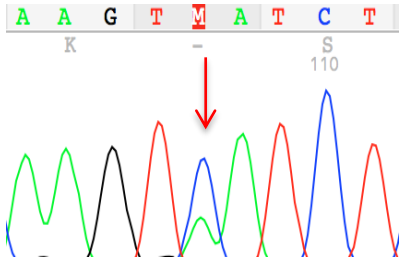
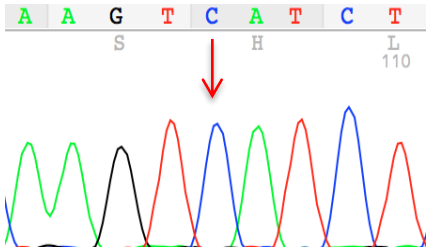
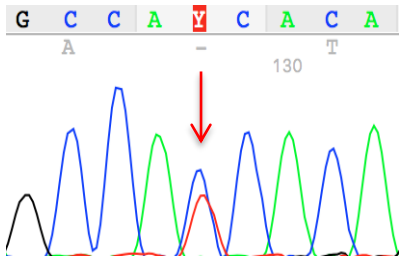
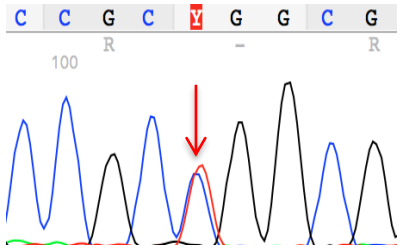
Gene	ID	WES variant details	Depth; %VAF	COSMIC ID / Transcript ID	SIFT; PolyPhen2	Tumour	Matched normal
KMT2D	**ADA-T2**	Nonsense; c.16154C>A; p.Ser5385Ter	118; 30%	COSM3954819 / ENST00000526209	N/A		
	ADA-T8	Missense c.10930C>T; p.Pro3644Ser	71; 42%	Not reported / ENST00000301067	Deleterious ; Damaging		Sample not available
	OFD-like ADA-T4	Missense; c.6437C>T; p.Pro2146Leu	56; 47%	COSM3461578 / ENST00000301067	Deleterious ; Damaging		Sample not available

Table 6–1: Details and Sanger sequencing confirmation of WES variants detected in *KMT2D*. *KMT2D* is recurrently mutated in two adamantinoma samples. *KMT2D* is also mutated (non-recurrent) in one OFD-like adamantinoma tumour. WES variant details column consists of the following: (1) the type of WES variant (e.g., missense); (2) coding variants position (c.), for example, 16154C>A represents a C to A nucleotide substitution at nucleotide 16154 of the gene coding region, where c.1 is nucleotide A of ATG start codon; and (3) affected translated protein (p.), for example, p.Ser5385Ter represents a nonsense substitution at codon 5385 causing a premature stop, where methionine is 1. All variants are heterozygous (red arrow). %VAF: altered variant allele frequency. The variant in sample ADA-T2 was confirmed as somatic; however, DNA from corresponding normal samples for the other two tumours was not available to confirm the somatic status of their variants.

Adamantinoma (n=8)		OFD-like adamantinoma (n=4)	
Affected canonical/bio function pathway	**p-value**	**Affected canonical/bio function pathway**	**p-value**
Calcium signalling	1.34E-02	Calcium signalling	3.55E-03
FAK signalling	1.57E-02	JAK/Stat signalling	2.84E-02
Regulation of cellular mechanics by Calpain protease	1.67E-02	Calcium transport I	3.12E-02
Mitochondrial dysfunction	2.33E-02	Cellular compromise	2.38E-03
Role of BRCA1 in DNA damage response	4.38E-02	Cell cycle	3.16E-03
IL-8 signalling	4.38E-02	Cell morphology	3.16E-03
Cellular function and maintenance	4.7E-04	Cell-to-Cell signalling and interaction	3.16E-03
Cell-to-cell signalling and interaction	5.25E-04	Cellular assembly and organisation	3.16E-03
Cell morphology	3.08E-03		
Cell signalling	3.37E-03		
Cellular development	4.38E-03		

Table 6–2: Canonical and bio function pathways affected in adamantinoma and OFD-like adamantinoma tumours. Pathway analysis was performed using Ingenuity Pathway Analysis (IPA) software (Qiagen; http://www.qiagen.com/ingenuity). For each tumour subtype, the list of filtered genes with potential pathological role was used. Canonical pathways (underlined) are well-established pathways that are associated with important cellular and molecular processes. Bio function analysis predicts the altered biological pathway based on the mutated genes. The top five bio function pathways (based on p-value) were selected.

Using the TopHat Alignment v1.0 and Cufflinks Assembly bundle tools v1.1, a multistep data analysis approach was followed consisting of performing QC of raw data details in Appendix Section 8.4 and Table 8–5) and identification of gene fusions using TopHat2-Fusion algorithm, available at BaseSpace Illumina hub (https://basespace.illumina.com). All tumours showed satisfactory QC results. A second splice junction calling tool, named STAR-Fusion (Haas et al., 2017), was used to detect gene fusions (Courtesy of Yun Shao Sung, MS and Cristina R. Antonescu, MD, Memorial Sloan-Kettering Cancer Center, New York) (details about TopHat2-Fusion and STAR-Fusion tools in Section 2.4.3.1 Appendix Section 8.4.2).

6.4.2 Identification of a recurrent fusion, *KANSL1-ARL17A,* in adamantinoma and OFD-like adamantinoma tumours

One recurrent fusion, *KANSL1-ARL17A*, was identified in 3/5 (60%) adamantinoma and in 2/3 (66.7%) OFD-like adamantinoma tumours by RT-PCR using a pair of primers flanking the gene fusion breakpoint (Figure 6–2, fusion details and breakpoints in Appendix Table 8–12). The *KANSL1-ARL17A* (named: *KANSARL*) has been previously reported as a germline fusion transcript (Kinsella et al., 2011). *KANSARL* was also classified germline in four adamantinoma corresponding normal samples. *KANSARL* was recently described as the first germline cancer-predisposing fusion gene, which was detected in normal tissue samples, breast cancer, lung cancer and prostate cancer cell lines, as well as glioblastoma and prostate tumours (Zhou et al., 2017).

KANSL1-ARL17A by RT-PCR was screened in two additional tumours, one adamantinoma (ADA-T1) and one OFD-like adamantinoma (OFD-like-ADA-1). These tumours were not RNA-sequenced but their RNA samples were available for laboratory

screening. The *KANSL1-ARL17A* was identified in ADA-T1 and OFD-like-ADA-1 tumours. Altogether, *KANSL1-ARL17A* was detected in 4/6 (66.7%) adamantinoma tumours and in 3/4 (75%) OFD-like adamantinoma tumours (Figure 6–2).

The *KANSARL* fusion transcript is formed by an inversion genomic rearrangement, leading to *KANSL1→ARL17A* orientation (Figure 6–3A) (Zhou et al., 2017). Six different *KANSARL* fusion transcript isoforms have been identified so far, involving different exons and using different splice junctions (Zhou et al., 2017). The *KANSARL* fusion transcript identified in this study occurred by splicing together the first three exons of both *KANSL1* and *ARL17A* fusion gene partners (Figure 6–3B). Sanger sequencing of the RT-PCR amplicons confirmed the fusion transcript breakpoint at the cDNA level (Figure 6–3C).

The *KANSARL* fusion is rarely detected in the tumours from patients of Asian or African ethnicity; however, it is familially inherited and preferentially associated with tumours from patients of European ancestry origin (Zhou et al., 2017). Notably, the adamantinoma and OFD-like adamantinoma tumours used in this study were from patients of European ethnicity.

6.4.3 Prioritising non-recurrent gene fusion candidates in adamantinoma and OFD-like adamantinoma samples

A total of 42 non-recurrent candidate gene fusions were identified in adamantinoma and OFD-like adamantinoma samples. A series of filtering steps were followed to identify genuine somatic gene fusion (Figure 6–4 & Section 2.4.4.1). In short, gene fusions were prioritised if at least one gene fusion partner was exonic and the gene fusion breakpoint is supported by ≥3 fusion junction reads. Following this filtering scheme, a total of three gene fusions were identified in adamantinoma tumours,

whereas no gene fusions were detected in OFD-like adamantinoma samples (fusions details in Appendix Table 8–12). The three gene fusions identified in adamantinoma samples were assessed by RT-PCR using cDNA from tumour and corresponding normal samples. Of the three gene fusions identified in adamantinoma, two fusions were confirmed germline (present in tumour and normal tissues) in ADA-T3, *XXYLT1-ACAP2* and *C15orf57-CBX3*, whereas one gene fusion was confirmed somatic in ADA-T4, *EPHB4-MARCH10,* which is discussed in details in the following section.

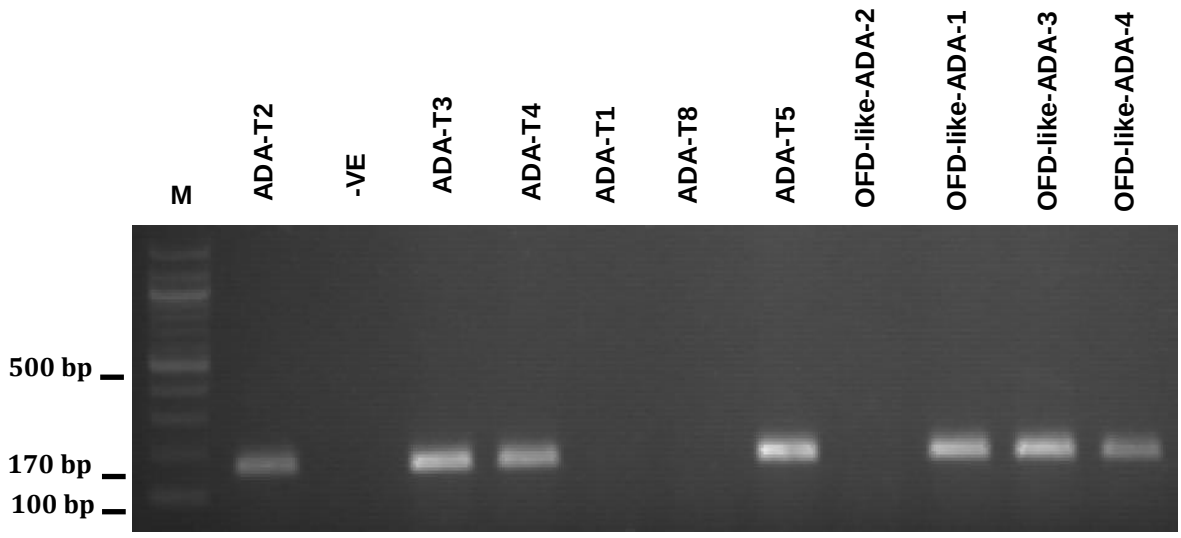


Figure 6–2: RT-PCR confirmation of *KANSL1-ARL17A* fusion transcript in six adamantinoma and four OFD-like adamantinoma tumours. RT-PCR amplicons represent samples for which the fusion transcript is detected. M: 100 base pair DNA marker; –VE: negative control; ADA: adamantinoma; OFD-like-ADA: osteofibrous dysplasia-like adamantinoma.

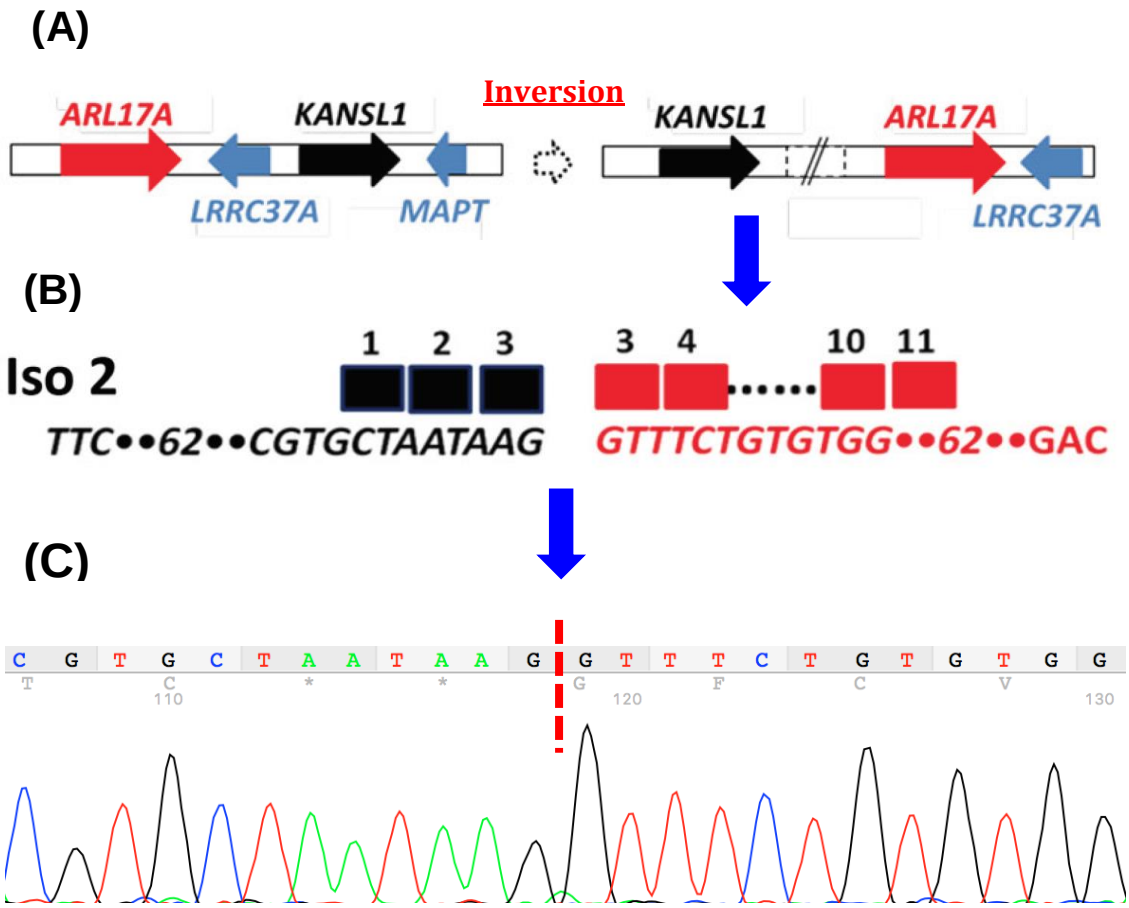


Figure 6–3: Schematic representation of the *KANSL1-ARL17A* fusion transcript. (A) An inversion chromosomal rearrangement at 17q21.31 chromosomal band results in *KANSL1→ARL17A* formation. Genomic breakpoint is denoted by a double slash (//). (B) The isoform 2 fusion transcript is formed by joining the first three exons of each gene fusion partner, *KANSL1* and *ARL17A*. (C) Sanger sequencing of the RT-PCR amplicons confirms the fusion breakpoint at the cDNA level. Figure panel A & B are reproduced with modifications from (Zhou et al., 2017)**.**

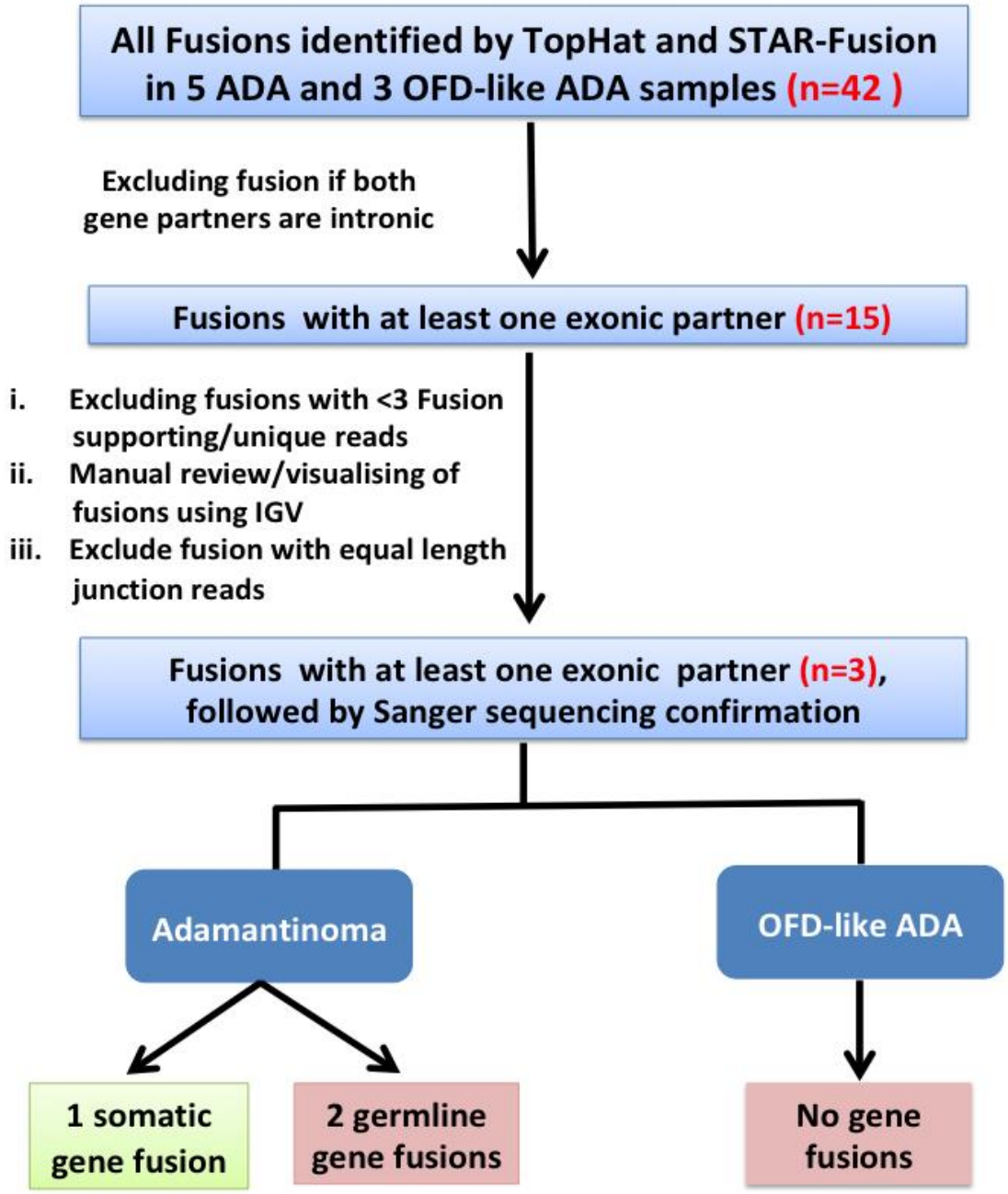


Figure 6–4: Filtering and prioritisation steps used to identify genuine somatic gene fusions in adamantinoma and OFD-like adamantinoma tumours. A multistep scheme was followed to filter candidate gene fusions identified using RNA-Seq. ADA: adamantinoma. OFD-like ADA: OFD-like adamantinoma.

6.4.3.1 *EPHB4-MARCH10* gene fusion in ADA-T4 sample: RT-PCR and LR-PCR analyses

The *EPHB4-MARCH10* gene fusion identified in the ADA-T4 tumour was supported by 109 fusion junction/supporting reads that are distributed uniformly across the fusion breakpoint (see Appendix Table 8–12 for fusion details). The fusion transcript is novel and has not been described previously in the literature. *EPHB4-MARCH10* was screened in ADA-T2 and OFD-like-ADA-T1 tumours (not RNA-sequenced) but was not detected in either sample. *EPHB4* fusion gene partner is classified as a potentially druggable gene by DGIdb (http://www.dgidb.org/).

EPHB4-MARCH10 is formed by joining the first eight exons of the 5′-gene partner *EPHB4* (Transcript ID: ENST00000358173.3) to the last exon (exon 2) of the 3′-gene partner *MARCH10* (Transcript ID: ENST00000582568.1) (Figure 6–5 A & B). The fusion breakpoint was confirmed by Sanger sequencing at the cDNA level (Figure 6–5B). *EPHB4-MARCH10* was confirmed somatic in the ADA-T4 sample using RT-PCR and a set of primers flanking the breakpoints (Figure 6–5C). The *EPHB4-MARCH10* fusion transcript is in-frame and predicted to produce a chimeric protein product of 640 amino acid in length, which consists of 529 amino acids from *EPHB4* and 111 amino acids belonging to *MARCH10*.

The cDNA/transcriptomic breakpoints for *EPHB4* and *MARCH10* were at the end of exon 8 and the beginning of exon 2, respectively. Therefore, we speculated the theoretical genomic/DNA translocation to be in intron 8/9 of *EPHB4* and intron 1/2 of *MARCH10*. Using LR-PCR, a series of six forward (F1–F6) and 13 reverse (R1–R13) primers were used in the genome walking analysis (Figure 6–6A) (more about genome walking in Section 5.4.3). All primer sets amplified products except for the F6/R13 primer set, suggesting that the breakpoint is upstream of the annealing sites of the F6

and R13 primers. As expected, the genome walking analysis mapped genomic/DNA breakpoints to introns 8/9 and 1/2 of *EPHB4* and *MARCH10*, respectively. Sanger sequencing of the smallest PCR product (F5/R12) determined the DNA breakpoint of *EPHB4* located 2,469 bp downstream of the end of exon 8 and the *MARCH10* breakpoint mapped 7749 bp upstream of the beginning of exon 2 (Figure 6–6 B & C). Moreover, LR-PCR analysis revealed a somatic chromosomal translocation rearrangement between chromosome 7 and 17 that was present in the tumour but not in the corresponding normal tissue, linking the fusion gene partners together (Figure 6–6C).

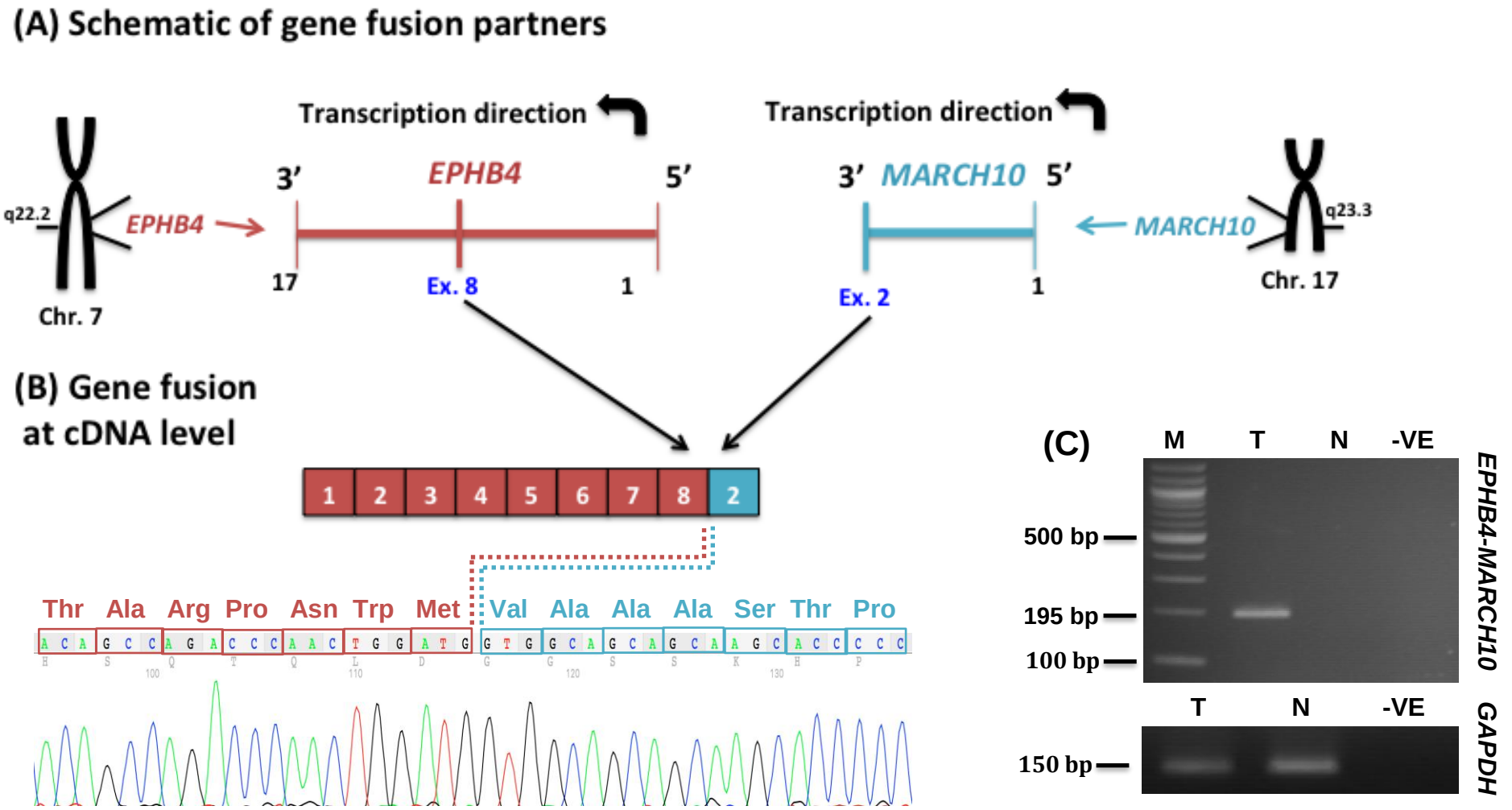


Figure 6–5: Schematic representation and RT-PCR confirmation of *EPHB4-MARCH10* fusion in ADA-T4 tumour. (A) Genomic representation of fusion gene partners. Exons are numbered and coloured red and cyan for *EPHB4* (ENST00000358173.3) and *MARCH10* (ENST00000582568.1), respectively. (B) The representation of *EPHB4-MARCH10* chimeric transcript at the cDNA level. (C) Amplified RT-PCR products are visualised on agarose gel (2.5%), including GAPDH positive control. M: 100 bp DNA marker; T: tumour; N: corresponding normal; -VE: negative control (water).

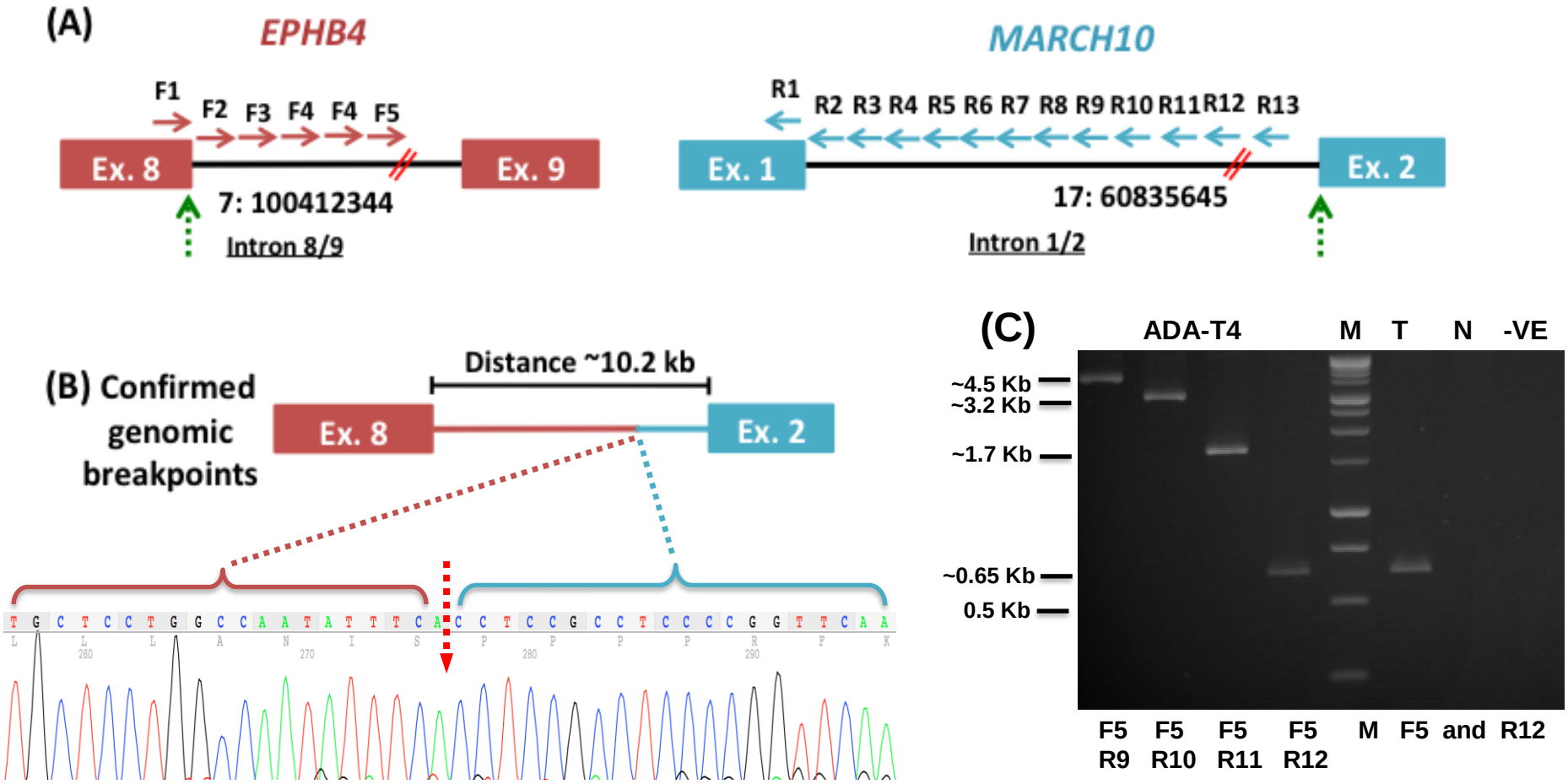


Figure 6–6: Genomic characterisation and LR-PCR validation of *EPHB4-MARCH10*. (A) Details of the *EPHB4* (red) and *MARCH10* (cyan) gene partners. Forward (F) and reverse (R) primers (coloured arrows) used in the genome walking analysis. The green arrow represents the cDNA breakpoints, red double slash represents the mapped genomic/DNA breakpoints. (B) Sanger sequencing confirming the genomic breakpoint (red dotted arrow), joining the gene fusion partners together and showing the newly formed distance between the exons of the gene fusion partners (~10.2 Kb). (C) Agarose gel (1%) showing LR-PCR amplicons decreasing in size during the genome walking analysis and confirming the somatic status of the gene fusion (present in tumour [T] and absent in matched-normal [N]). –VE: negative control (water). M: 1 Kb DNA marker.

6.5 Summary of findings: WES and RNA-Seq

In the first part of the study, eight adamantinoma and four OFD-like adamantinoma tumours were exome-sequenced, aiming to identify coding somatic alterations. Six adamantinoma and three OFD-like adamantinoma tumours were paired with corresponding normal tissues; however, no corresponding normal tissues were available for the remaining three tumours. Following a series of filtering and prioritisation steps, a total of 21 INDELs and 182 SNVs were detected in eight adamantinoma samples, and 7 INDELs and 59 SNVs were identified in four OFD-like adamantinoma tumours. One gene, *KMT2D*, was recurrently altered in two adamantinoma tumours (25%) (Table 6–3). Interestingly, *KMT2D* was also mutated in one OFD-like adamantinoma tumour (Table 6–3). *KMT2D* is classified as a 'cancer driver' gene by the COSMIC and IntOGen databases, as well as potentially druggable by DGIdb database. Using IPA, multiple cellular and biological pathways were significantly altered in adamantinoma and OFD-like adamantinoma tumours (Table 6–2). In the second part, RNA-Seq was conducted on five adamantinoma and three OFD-like adamantinoma tumours to identify somatic gene fusions resulting from chromosomal rearrangements. Two different splice-junction calling tools were used to detect gene fusions. One gene fusion previously reported as a cancer-predisposing germline fusion (*KANSL1-ARL17A*) was confirmed by RT-PCR in 4/6 adamantinoma (66.7%) and in 3/4 OFD-like adamantinoma (75%) tumours (Table 6–3). *EPHB4-MARCH10* was the only somatic fusion identified in this study, which was characterised by RT-PCR and LR-PCR in ADA-T4 (Table 6–3). Genomic characterisation analysis of *EPHB4-MARCH10* revealed a somatic chr7:17 translocation event linking the two gene fusion partners together at the DNA level. DGIdb database classifies *EPHB4* as

a potentially drug-targeted gene. Apart from *KANSL1-ARL17A* cancer-predisposing germline fusion, no somatic gene fusions were identified in OFD-like adamantinoma tumours.

Mutated gene	**ADA-T1**	**ADA-T2**	**ADA-T3**	**ADA-T4**	**ADA-T5**	**ADA-T6**	**ADA-T7**	**ADA-T8**	**OFD-like-ADA-T1**	**OFD-like-ADA-T2**	**OFD-like-ADA-T3**	**OFD-like-ADA-T4**
KMT2D		*										*

Gene fusion												
KANSL1-ARL17A												
EPHB4-MARCH10												

Table 6–3: WES and RNA-Seq alterations landscape in adamantinoma and OFD-like adamantinoma tumours. WES was conducted on 12 tumours in total. *KMT2D* was a recurrently mutated gene in two adamantinoma samples. *KMT2D* was mutated in one OFD-like adamantinoma tumour. Orange and green boxes represent nonsense and missense changes, respectively. *Reported in COSMIC database. RNA-sequenced samples are underlined. RNA samples for ADA-T2 and OFD-T1 were available for laboratory screening of the two gene fusions. RNA material was not available for ADA-T6 and T7 to perform RNA-Seq or laboratory screening. ADA: adamantinoma; OFD-like-ADA: osteofibrous dysplasia-like adamantinoma.

6.6 Discussion

The application of NGS technologies allows researchers to investigate the landscape of genetic alterations in cancer. Specifically, WES enables sequencing analysis of both exonic and splice site alterations (Petersen et al., 2017), while RNA-Seq allows for comprehensive transcriptome profiling, identification of gene fusions and determination of gene expression profiles in cancer samples (Byron et al., 2016). Elucidating the genetic profiling of tumours samples facilitates the understanding of tumourigenesis, improves prognosis and patient outcomes, and identifies potential targets for therapeutics.

6.6.1 *KMT2D* recurrent gene mutations identified in adamantinoma tumours by WES

6.6.1.1 Overview and association with cancer

Chromatin is the primary source of genetic information, containing DNA-packed material and chromosomal proteins (Ford and Dingwall, 2015). Chromatin remodelling through modifications of the chromatin structure is a fundamental key regulator of gene expression in cells. These epigenetic modifications are complex and involve various protein complexes. Interfering with such pathways can disrupt cellular differentiation and regulatory controls, leading to imbalances in cellular homeostasis and development (Ford and Dingwall, 2015). Methylation of histone H3 at the lysine 4 residue (H3K4) controlled by histone lysine methyltransferases is one of the epigenetic mechanisms influencing the transcriptional activity at enhancer or promoter DNA regions (Froimchuk et al., 2017; Liu et al., 2015).

Histone lysine N-methyltransferase 2D (*KMT2D*), also known as *MLL4* or *MLL2*, is a histone-modifier tumour suppressor gene belonging to the suppressor of variegation, enhancer of zeste, trithorax (SET1) class of lysine methyltransferases (KMTs) (Ford and Dingwall, 2015; Hillman et al., 2018; Liu et al., 2015). This SET1 family of KMTs is a part of a multimeric protein complex that acts as coactivators of transcription factors (Ford and Dingwall, 2015).

Somatic mutations in *KMT2D* have been identified in various types of cancers, including those of the brain, lymph nodes, blood and lungs (Figure 6–7) (Froimchuk et al., 2017). *KMT2D* mutations have also been classified as cancer drivers in mantle cell lymphoma and squamous cell carcinomas of the head and neck (Ford and Dingwall, 2015). Moreover, *KMT2D* is one of the most frequently mutated genes in multiple paediatric malignancies (Huether et al., 2014).

6.6.1.2 *KMT2D* variants identified in adamantinoma and OFD-like adamantinoma tumours

Two *KMT2D* variants were identified in two adamantinoma tumours (ADA-T2, ADA-T8). Interestingly, one *KMT2D* variant was detected in one OFD-like adamantinoma tumour (OFD-like-ADA-T4) (Table 6–3). The clinical features of the OFD-like-ADA-T4 sample was reassessed to rule out the possibility of the tumour being actually of adamantinoma classification. However, due to limited sample material and no additional clinical features, histopathological assessment was not possible.

The c.16154C>A; p.Ser5385Ter identified in ADA-T2 is a nonsense change, occurring just upstream of the SET domain (Figure 6–8). This nonsense change leads to a truncated protein product, missing the entire SET domain at the C-terminus. Frameshift and nonsense alterations occurring in the SET domain represent 37% of the total

number of *KMT2D* mutations described in tumours (Rao and Dou, 2015). The frequent occurrence of these heterozygous-inactivating mutations in multiple cancers is indicative of a haploinsufficient tumour suppressor phenotype (Ford and Dingwall, 2015).

The enzymatically active SET domain is responsible for H3K4 methyltransferase activity and the protein stability of KMT2D in cells (Dorighi et al., 2017). Set1-containing KMTs, including KMT2D, are part of large highly conserved multicomponent complexes known as SET/COMPASS (COMplex of Proteins ASsociated with Set) (Ford and Dingwall, 2015). The SET/COMPASS complexes are responsible for the majority of global H3K4 dimethylation and trimethylation, which are marked with active promoters or enhancers (Froimchuk et al., 2017).

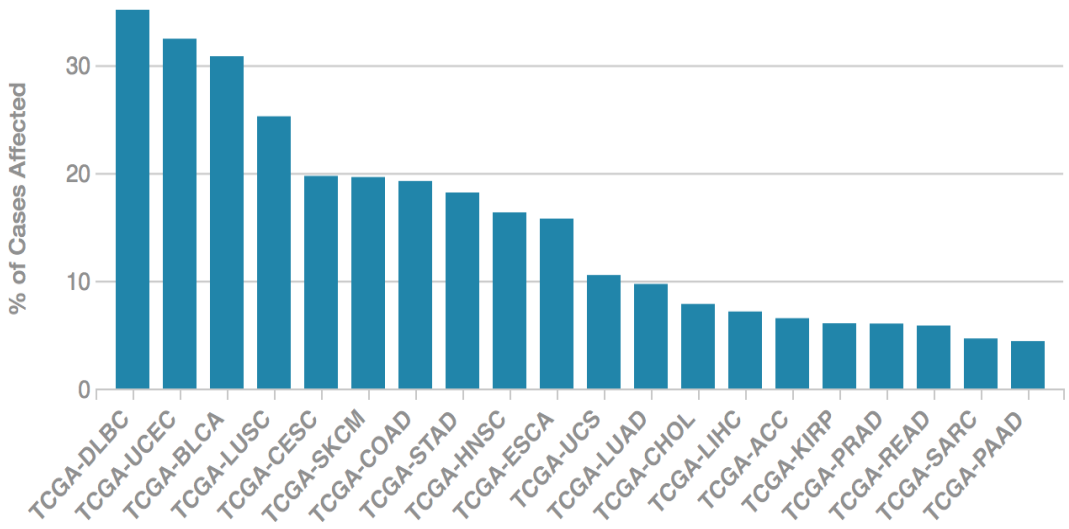


Figure 6–7: The distribution *KMT2D* mutation across 20 cancer projects. A total of 1,140 cases were affected with 1,498 *KMT2D* mutations across 31 cancer projects of the Cancer Genome Atlas (TCGA). The highest percentage of *KMT2D* mutations was detected in lymphoid neoplasm diffuse large B-cell lymphoma (TCGA-DLBC) (35.14%), whereas, the thyroid carcinoma harboured the lowest percentage of mutations (0.61%) (not shown). Image obtained from National Cancer Institute, GDC Data Portal https://portal.gdc.cancer.gov).

To understand the outcome of targeted inactivation of the SET domain in mice, a study by Lee et al. (2009) inactivated the methyltransferase activity of Kmt2C, a closely related paralogue to *Kmt2D* with extended SET domain similarity, by replacing the wild-type Kmt2C with a mutant Kmt2C that contains an in-frame deletion of a 61-aa catalytic core in the SET domain. The *Kmt2C*$^{\Delta SET/\Delta SET}$ mice showed incomplete embryonic lethality and exhibited cellular hyperproliferation and kidney ureter urothelium tumours, accompanied by increased levels of DNA damage. Altogether, the *KMT2D* truncating mutation identified in this study, which leads to complete loss of the SET domain, is likely tumourigenesis-related, characterised by loss of tumour suppressing activity.

A missense variant, c.10930C>T; p.Pro3644Ser, was detected in ADA-T8 and occurs in close proximity to the high mobility group (HMG-I) that consists of short motifs that interact with the minor groove of DNA sequences (Figure 6–8). These binding motifs support promoter recognition and stabilisation of DNA binding and have a regulatory role in DNA replication and transcription (Bochynska et al., 2018). The last missense variant identified in OFD-like-ADA-T4, c.6437C>T; p.Pro2146Leu, does not reside in a recognised functional domain of KMT2D (Figure 6–8). Both missense variants require further investigations to elucidate their potential pathogenic roles.

6.6.2 *EPHB4-MARCH10* gene fusion identified by RNA-Seq

A novel gene fusion involving ephrin receptor (*EPHB4)* and membrane-associated ring-CH-type finger 10 (*MARCH10*) genes was confirmed somatic in ADA-T4 tumour by RT-PCR and LR-PCR analyses (Section 0). *EPHB4-MARCH10* fusion results from a chr7:17 chromosomal translocation, joining the first eight exons of *EPHB4* to the last exon of *MARCH10*. The splicing of the chimeric transcript follows the CT-AC donor-

acceptor splicing rule (the reverse complementary of standard GT-AG canonical splicing).

EPHB4 encodes a receptor tyrosine kinase (RTK) that belongs to the erythropoietin-producing hepatocellular carcinoma (Eph) family, the largest class of RTKs (Salgia et al., 2018). RTKs are transmembrane proteins that regulate key biological processes, including cell differentiation and proliferation. Upon the binding of signalling molecules, RTKs undergo conformational changes producing active forms of the enzymes that subsequently initiate cellular responses (Salgia et al., 2018). Impaired Eph/ephrin cellular interactions have been implicated in carcinogenesis, contributing to tumour growth, metastasis, chemoresistance and tumour angiogenesis (Merchant et al., 2017; Salgia et al., 2018). Mutations in *EPHB4* have been identified in several lung cancers including lung adenocarcinoma and small cell lung cancer (Ferguson et al., 2015).

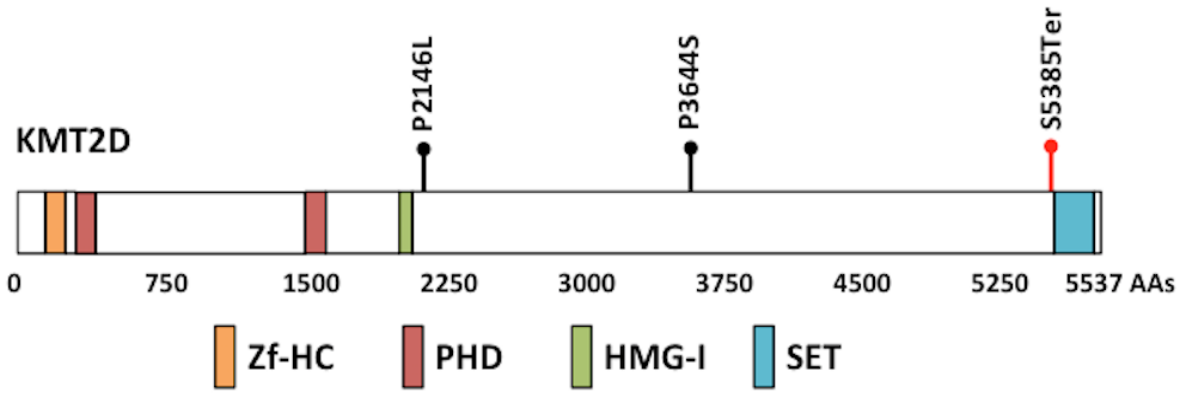


Figure 6–8: Schematic of *KMT2D* alterations and their relative protein positions in adamantinoma and OFD-like adamantinoma tumours. Two recurrent in *KMT2D* were identified in two adamantinoma tumours; one *KMT2D* variant was detected in an OFD-like adamantinoma tumour. The black and red balls represent missense and nonsense changes, respectively, and their positions in the KMT2D protein. AAs: amino acids numbers. Zf-HC: PHD-like zinc-binding domain; PHD: PHD zinc finger; HMG-I: high-mobility-group (HMG)-box domain; SET: SET domain. Original figure, compiled from information in (Froimchuk et al., 2017; Tan et al., 2015).

MARCH10 encodes a microtubule-associated E3 ubiquitin ligase belonging to the membrane-associated RING-CH (MARCH) family, which consists of 11 gene members in mammals (Iyengar et al., 2011). A comprehensive physiological role of MARCH proteins has not yet been achieved, requiring further investigations (Samji et al., 2014). However, it has been speculated that these proteins are involved in the ubiquitination of cell-surface immune regulatory molecules, endoplasmic reticulum-related degradation and endosomal protein trafficking (Iyengar et al., 2011).

6.6.2.1 *EPHB4-MARCH10* effect on protein domains organisation and the dual functionality of EPHB4 in cancer

The *EPHB4-MARCH10* chimeric transcript is in-frame, resulting from joining the first 529 amino acids of *EPHB4* (ENST00000358173.3) to the last 111 amino residues of *MARCH10* (ENST00000582568.1) (Figure 6–9). In the *EPHB4-MARCH10* chimeric transcript, the organisation of the protein domains of EPHB4 is as follows: the ephrin receptor ligand binding, putative ephrin-receptor like and fibronectin type III (repeats) functional domains are retained, while ephrin type-A receptor 2 transmembrane, catalytic protein tyrosine kinase and sterile alpha motif domains are completely lost (Figure 6–9). The N-terminus ephrin receptor ligand binding and putative ephrin-receptor like domains mediate cellular signalling between Eph receptors and the ephrins in adjacent cells, in addition to participating in the clustering of ephrin complexes (Chen et al., 2017b). The kinase domain is the catalytic domain activated upon the binding of ephrin ligands to the globular domain of the receptors (Chen et al., 2017b). The sterile alpha motif domain, a protein-protein interaction domain, is involved in the homodimerisation and oligomerisation of receptors (Salgia et al., 2018).

.

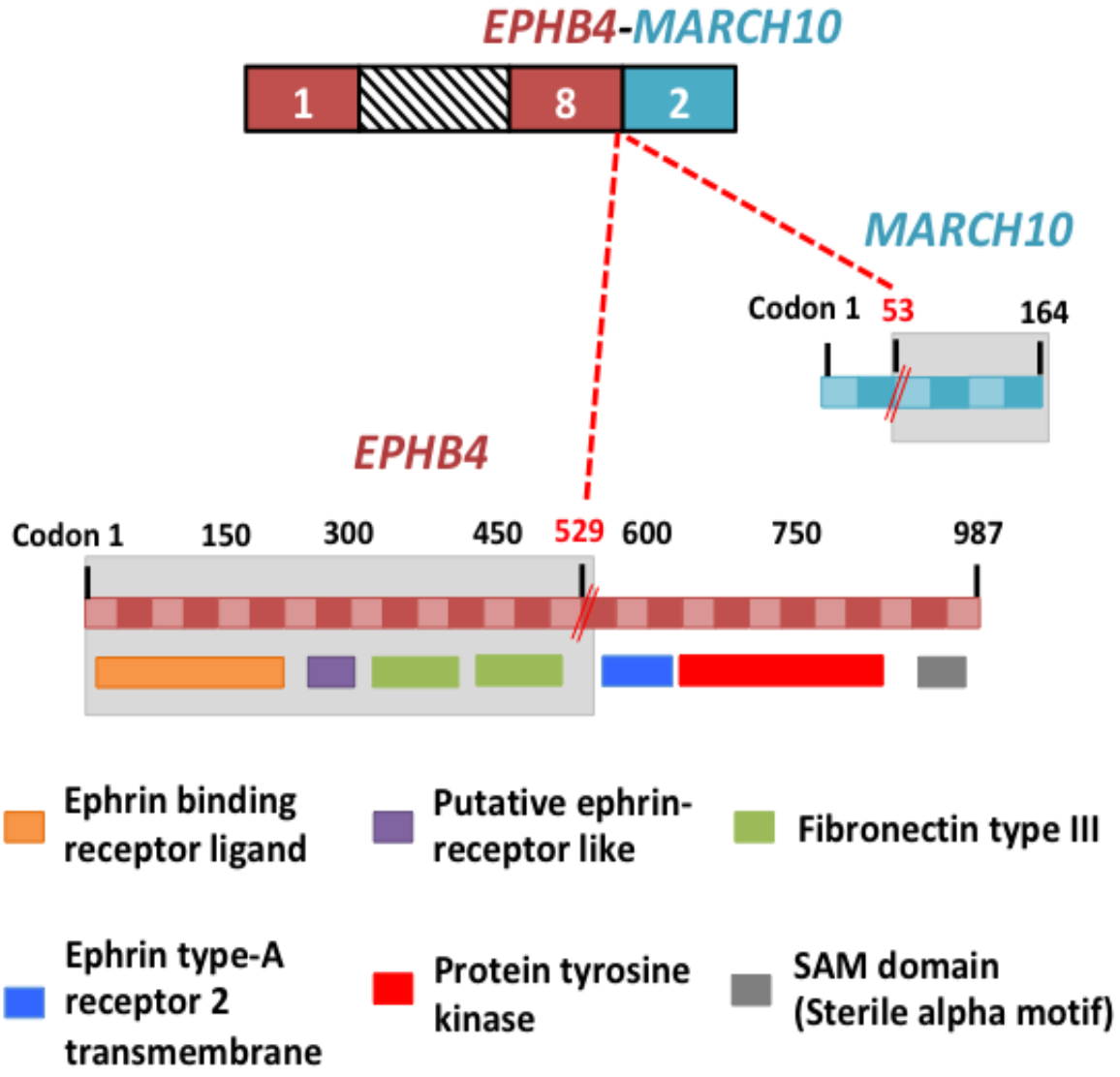


Figure 6–9: *EPHB4-MARCH10* gene fusion effect on the protein domain organization of the fusion gene partners. The schematic shows the functional protein domains of *EPHB4* (red), ENST00000358173.3. No functional protein domains are present in *MARCH10* (cyan) for the ENST00000582568.1 transcript. The cDNA breakpoint of each gene fusion partner is denoted by a red double slash (//), with the corresponding amino acid number (in red). The Grey shaded area represents the retained functional domains in the *EPHB4-MARCH10*.

The precise biological role of Eph-receptor/ephrin-ligand in cancer development is complex due to the various expression levels observed in cancer (Chen et al., 2017b). In cancer cells, both tumour-suppressing and tumour-promoting (oncogenic) activities have been described for EPHB4 (Rutkowski et al., 2012). A downregulation of EPHB4 in the majority of breast carcinoma tumour cells versus strong expression at the periphery of the tumour specimen supports the complex and dual roles of EPHB4 in cancer development (Kaenel et al., 2012). Overexpression of EPHB4 has been reported in multiple cancers, including those of the colon, prostate, oesophageal and breast (Chen et al., 2017b). Overexpression of EPHB4 correlates with poor prognosis in multiple malignancies (Merchant et al., 2017). However, low levels of Eph or ephrin has also been reported in cancer cell lines and tumour samples (Pasquale, 2010).

A bidirectional intracellular signalling consisting of forward and reverse signals between the EPHB4 receptor and ephrinB2 ligand regulates cell morphology, adhesion and migration (Chen et al., 2017b; Merchant et al., 2017) (Figure 6–10). The forward signalling mechanism occurs in receptor-expressing cells and is initiated by autophosphorylation events that subsequently activate the tyrosine kinase catalytic domain (Salvucci and Tosato, 2012). The reverse signalling depends on Src family kinases for propagation in the ligand (ephrin)–expressing cells (Salgia et al., 2018).

Both of these EPHB4-receptor/ephrin-ligand forward and reverse signalling has been associated with tumour suppressive and oncogenic phenotypes (Chen et al., 2017b). For example, after activation of EPHB4 receptor with its ligand ephrin-B2, an EPHB4 forward signalling mechanism actives the anti-oncogenic Abl-Crk pathway, exerting a tumour-suppressing effect (Kaenel et al., 2012). On the other hand, there is a mounting evidence of the ability of the EPHB4-receptor/ephrin-ligand forward and reverse signalling in enhancing malignant transformation (Chen et al., 2017b).

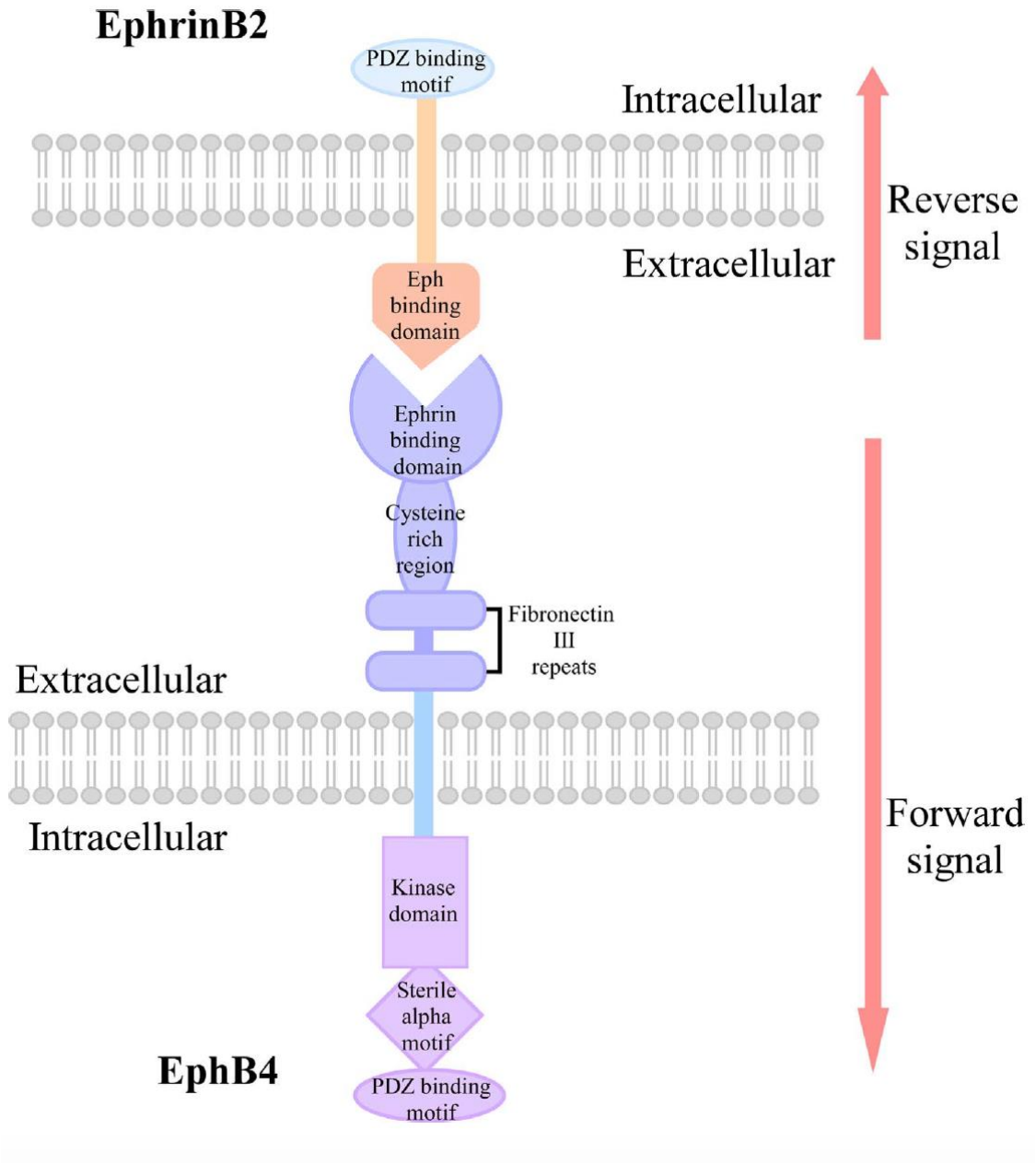


Figure 6–10: A schematic representation of the domain structure of EPHB4 and ephrin ligand. Eph receptors, including EPHB4, contain extracellular and intracellular compartments. The extracellular compartment consists of: an ephrin-binding domain, a cysteine rich region and two fibronectin type III repeats, while the tyrosine kinase domain, sterile alpha motif and PSD95, DLG, ZO1-binding motif (PDZ)-binding motif are intracellular. A bidirectional forward or reverse signalling from Eph receptor or ephrin ligand, respectively, participates in multiple cellular processes. Image adapted from (Chen et al., 2017b).

As mentioned earlier, *EPHB4-MARCH10* fusion leads to a complete loss of the three domains located in the intracellular compartment of EPHB4. Therefore, the forward EPHB4-receptor/ephrin-ligand signalling is likely disrupted by the loss of the three domains including the kinase catalytic activity domain (Figure 6–10). A study by Hu et al. (2014) assessed the overexpression of wild-type EPHB4 and mutant EPHB4 (kinase-inactive) in oesophageal squamous cell carcinoma cells and found that both wild-type and mutant EPHB4 promoted cell growth and migration. This finding suggests that EPHB4 promoted growth of the tumour cells independent of its kinase activity. With the complex and dual role of EPHB4 in tumourigenesis, the precise tumour-suppressing or -promoting role of *EPHB4-MARCH10* requires further investigations.

6.7 Remarks and future work

WES and RNA-Seq technologies are useful tools to study cancer genetics. The limitations of these two high-throughput technologies are described in Section 0 & 5.6.2.4. In WES study, variants in the tumour-associated *KMT2D* gene were detected in 25% of adamantinoma tumours, which have not been previously described in this tumour subtype. The assessment of *KMT2D* mutations in additional adamantinoma tumours (a validation set), for example by a targeted sequencing approach, can further validate the tumourigenesis association of *KMT2D* in adamantinoma. Regarding the RNA-Seq findings, screening the somatic *EPHB4-MARCH10* gene fusion in additional tumours can also be performed.

No recurrently mutated genes were identified in OFD-like adamantinoma tumours (n=4) by WES. Two of the four tumours lacked any coding changes. Moreover, RNA-Seq failed to identify genuine somatic gene fusions in OFD-like adamantinoma

tumours. Because WES covers a small fraction of the genome, WGS can be considered to further investigate these tumours. Additional tumour samples are required to increase the cohort size.

Similar to recommendations for UPSb tumours, CNVs analysis from WES data and differentially expressed genes from RNA-Seq data can be conducted for adamantinoma and OFD-like adamantinoma tumours. CNVs analysis may be necessary especially for OFD-like adamantinoma because no recurrent small INDELs or SNVs were detected. As of September, 2018, Prof Farida Latif initiated collaborations to perform WES CNV and RNA-Seq differential expression analyses in adamantinoma and OFD-like adamantinoma tumours.

Chapter 7: General discussion

Establishing causal relationships between genes and the pathophysiological mechanisms they control is one of the central goals of medical genetics studies. Before the introduction of NGS, genetic studies primarily utilised conventional approaches to decipher genetic variants, including genetic linkage-based studies in inherited disease or the candidate gene approach in cancer (Doostparast Torshizi and Wang, 2018). Although these approaches led to many breakthroughs, they primarily relied on traditional Sanger sequencing, which has a limited throughput and can be laborious (Goodwin et al., 2016).

By contrast, the rapid advancement of NGS technologies have revolutionised the medical genetics field, fulfilling medical science/research demands and ushering in a new era of personalised genomic medicine. Moreover, NGS is increasingly being utilised for the clinical diagnosis of inherited diseases and cancer (Hardwick et al., 2017). NGS techniques provide unbiased, cost-effective and rapid analysis of the genome to identify genetic alterations associated with human disease, including rare genetic variants and transcriptomic alterations (e.g., gene fusions) (Byron et al., 2016; Rabbani et al., 2014). The wealth of data generated by NGS can be used to investigate and understand increasingly complex mechanisms of human disease. WES analyses the protein-coding regions of the genome, an important and functional subset in which genetic variants are likely to have a severe impact on produced proteins (Bamshad et al., 2011).

In the last decade, WES has accelerated the rate of discovering disease-causing genes in RGDs (Boycott et al., 2017) and deciphered complex genetic alteration landscapes of many sporadic cancer subtypes (Levy and Myers, 2016). RNA-Seq enables high-throughput sequencing and quantifying of the entire transcriptome,

detecting genetic events (e.g., gene fusions) that can be of biological relevance to cancer development and progression (Cieslik and Chinnaiyan, 2018). Although NGS technologies have led to many discoveries in rare diseases and sporadic cancer, the aetiology of some of these diseases remains poorly or incomprehensively understood. In the current thesis, NGS technologies were employed to investigate the genetic alteration spectrum and broaden the molecular understanding of a rare disease (CHT) and three rare sporadic bone tumours (UPSb, adamantinoma and OFD-like adamantinoma).

7.1 WES in studying RGDs, an example being CHT

In Chapter 3, WES sequencing was conducted on four consanguineous families with CHT to identify variants fitting AR inheritance. Because consanguineous families exhibit a higher prevalence of CHT than non-consanguineous families, there is a higher likelihood of identifying disease-causing variants in consanguineous families. Moreover, in consanguineous families, disease-causing variants are expected to be homozygous and inherited from a common ancestor, allowing prioritisation of homozygous variants and elimination of heterozygous ones.

In CHT families, two WES analytical approaches were applied, consisting of exome-sequencing: (1) unaffected parents-affected member trios (n=2), and (2) unaffected parent-unaffected sibling-affected members (n=2). In the four families, an average of 68,075 initial unfiltered changes (SNVs and INDELs) were identified by WES. To reduce this large number of variants and identify homozygous variants fitting AR inheritance pattern, we followed a series of filtering and prioritization steps. Consequently, the average number of filtered homozygous candidate variants for both approaches was comparable—27 in the first approach and 22 in the second one—

suggesting that both approaches were successful to similar extents in drastically reducing the number of initially identified variants. Following an additional filtering step to pinpoint variants of likely biological relevance, a homozygous candidate gene in *SIX2* fitting an AR inheritance pattern was detected in a patient affected with TGD (Family-1). Functional characterisation experiments and family segregation analyses were performed to infer the pathological significance of *SIX2* in TGD (Chapter 4).

Family segregation analyses using additional unaffected members classified the *SIX2* homozygous candidate variant as very unlikely to be disease causing. The availability of DNA from four additional members enabled a family segregation analysis on a larger family, allowing for the confident exclusion of non-segregating variants. Although a causal disease-gene relationship was not established in this study, the availability of other family members was very beneficial for classifying the *SIX2* homozygous variant as non-causal in CHT.

Because TGD is primarily sporadic in occurrence and has complex disease aetiology, the difficulty in establishing disease-gene causal relationships in the three TGD families of this study was not unexpected. However, additional investigations (as stated in Section 4.6) can be performed to further investigate other potential disease-gene relationships in these families.

A major limitation of this study was its small cohort size. However, with the emergence of large sequencing projects and international collaborative work, new insights into CHT are likely to emerge. For example, the UK's 100,000 Genomes Project is currently the largest nationwide sequencing project in the world that aims to implement a new genomic medicine service for patients with rare diseases and cancers. By combining WGS sequencing data with medical records, this project will offer new approaches for diagnosis and treatment of diseases. The new Genome Medicine Service will

standardise diagnostic and care pathways for many if not all rare diseases, allowing all NHS patients access to genomic testing. Furthermore, the UK's 100,000 Genomes is currently recruiting the four patients with CHT who participated in the current study.

7.1.1 Rare diseases and accelerating disease-gene discovery

Although RGDs are individually uncommon, they collectively affect the wellbeing of a significant proportion of people (primarily children) around the world. In the European-derived general population, RGDs collectively affect at least 1 in 50 individuals (Boycott et al., 2017). Although the pace of discovery for RGD-causing genes has substantially increased over the past decade, the current diagnostic rate is approximately 50% in general clinical genetics clinics (Boycott et al., 2017). Moreover, thousands of rare diseases are without approved treatment (Austin et al., 2018). This problem has led to realisation that global collaboration is required to maximise the outcome of rare disease research. The International Rare Diseases Research Consortium (IRDiRC) established in 2011 aims to unite rare disease researchers and organisations to identify current and future barriers in the RGD research field and implement strategies to overcome these bottlenecks (Austin et al., 2018; Boycott et al., 2017). With wide-scale international collaboration, data can be shared between clinical and research communities to collectively accelerate disease-causing gene discovery in undiagnosed patients.

7.2 WES for studying rare bone tumours

7.2.1 UPSb

In addition to investigating a RGD (i.e., CHT), NGS technologies were utilised to study rare sporadic bone tumours. In Chapter 5, WES and RNA-Seq were conducted on high-grade UPSb bone tumours to characterise their genetic (SNVs and INDELs) and transcriptomic (gene fusions) somatic alteration landscape. Although conventional approaches such as aCGH and Sanger sequencing of candidate genes have been conducted on UPSb (Niini et al., 2011; Kawaguchi et al., 2002), these tumours remain poorly understood genetically and lack standard molecular diagnostic testing (Christopher et al., 2013). To our knowledge, this study represents the first comprehensive genetic and transcriptomic profiling of these tumours using NGS technologies with the goals of characterising tumourigenesis mechanisms and identifying cancer driver genes that can be of potential therapeutic relevance.

Using WES, 31 recurrently mutated genes were identified in UPSb tumours, including *TP53*, *ATRX*, *H3F3A* and *DOT1L.* These recurrent genes were altered in 2/14 samples (14.3%), except for *TP53* mutated in 4/14 samples (28.6%). *ATRX*, *H3F3A* and *DOT1L* chromatin remodelling genes were collectively altered in 5/14 samples (35.7%), suggesting the potential involvement of altered chromatin remodelling pathways in UPSb pathogenesis. Chromatin remodelling genes are cancer-related genes and play an important role in regulating chromatin structure and the expression of thousands of genes (Prasad et al., 2015). To our knowledge, chromatin remodelling genes have not been previously described in UPSb.

TP53 was the most frequently mutated gene in UPSb and was classified as a potential gene-drug target by DGIdb. A study by Leijen et al. (2016) showed that tumours (e.g.,

melanoma, lung cancers, colorectal cancer) harbouring *TP53* mutations or deregulated p53 pathway benefited from combining AZD1775 with cytotoxic chemotherapy. Patients with these *TP53*-mutated solid tumours showed a 21% response rate, compared with a 12% response rate observed in patients with *TP53* wild-type tumours. Although additional investigations are required, UPSb tumours harbouring *TP53* mutations may benefit from AZD1775 therapy.

ATRX mutations have been associated with alternative elongation of telomeres (ALT), a phenomenon that prevents shortening of telomeres and, therefore, induces immortalisation of tumour cells (Koelsche et al., 2016). Because *ATRX* was recurrently mutated in two UPSb tumours, the potential association of impaired ATRX and ALT in these tumours may be of biological significance. Although complexity and heterogeneity of the ALT phenotype is evident (Koelsche et al., 2016), some targeted drugs have shown promising anti-proliferative activity outcomes in tumours harbouring *ATRX* alterations. For example, Trabectedin chemotherapy drug initiates a series of biological events that interfere with DNA-binding proteins and DNA repair pathways, causing cell cycle arrest and apoptosis (D'Incalci and Galmarini; Pompili et al., 2017). Trabectedin is a therapeutic option in patients with soft tissue sarcomas associated with ALT-induced poor prognosis (Pompili et al., 2017). Thus, UPSb with *ATRX* alterations can be potentially of therapeutic relevance; however, further investigations are needed.

The variants identified in the *H3F3A* gene affect the G34 and V35 hotspot amino acids, which have been mutated in multiple cancers such as 92% of giant cell tumours of bone (Behjati et al., 2013), 31% of paediatric glioblastomas (Schwartzentruber et al., 2012) and 3.37% of primary malignant bone tumours (e.g., osteosarcoma) (Amary et al., 2017). Although giant cell tumour of bone is considered benign, in the current study

and in Amary et al. (2017), *H3F3A* mutations are reported in aggressive bone tumour subtypes. Therefore, a malignant phenotype should be evaluated and considered in tumours harbouring *H3F3A* mutations.

In the second part of the UPSb study, RNA-Seq was conducted on eight UPSb tumours and identified eight somatic genes fusions. Two of the eight fusions were reported previously in other cancers, *CLTC-VMP1* and *FARP1-STK24*. Genomic analyses using LR-PCR revealed that both genes fusions are somatic, caused by interstitial deletions and have potential oncogenic and tumour-suppressing roles.

The *CLTC-VMP1* gene fusion is out of frame and was identified in the UPSb-T13 sample that lacked WES alterations in *TP53*, *H3F3A*, *ATRX,* or *DOT1L.* The interstitial deletion mechanism of *CLTC-VMP1* leads to a complete deletion of *PTRH2*, a gene that has a role in inducing apoptosis (Griffiths et al., 2011; Jan et al., 2004). The *CLTC-PTRH2-VMP1* rearrangement and the out-of-frame nature of *CLTC-VMP1* fusion suggest that the *CLTC-VMP1* fusion is possibly interfering with tumour-suppressing activity (Giacomini et al., 2013).

The *FARP1-STK24* gene fusion forms by joining the 5′ transcription regulatory apparatus of both fusion gene partners. Theoretically, in this fusion, both *FARP1* and *STK24* can initiate transcription that can continue to the other gene partner. *STK24* is a serine/threonine protein kinase that has a role in cell cycle and apoptosis as well as an oncogenic phenotype (Cho et al., 2016; Thompson and Sahai, 2015). Because kinase-constituted gene fusions are classified as potential drug targets (Tamura et al., 2015), *FARP1-STK24* can be potentially targeted with kinase inhibitors.

Due to limited PhD time, differential gene expression profiling using RNA-Seq was not performed. As mentioned earlier, no established molecular testing is available for UPSb. Differential expression analysis can identify genes unregulated in UPSb, which

can serve as potential molecular markers in diagnostic testing. Moreover, hierarchical clustering analyses can be conducted to further explore the gene expression profile of UPSb. Hierarchical clustering analyses of samples are beneficial in biologically classifying distinct tumour subgroups (Dey et al., 2017). Because UPSb shares clinical presentations with two other malignant bone tumours, osteosarcoma and dedifferentiated chondrosarcoma, hierarchical clustering analyses of these three tumours can be performed to determine whether UPSb can be molecularly distinguished from the other two bone tumours.

Prof Farida Latif has established a collaboration to perform gene expression and hierarchical clustering analyses on UPSb tumours using RNA-Seq data. Preliminary results identified a potential molecular biomarker for UPSb. In addition, hierarchical clustering analysis clustered UPSb tumours from other bone and soft tissue sarcomas, thus molecularly distinguishing UPSb. Although further confirmatory experiments have been initiated (as of August 2018), these preliminary findings would be helpful for differentiating between these bone tumours molecularly, optimising treatment and management options and confirming UPSb diagnosis by excluding other differential diagnoses.

7.2.2 Adamantinoma and OFD-like adamantinoma

Similar to the procedures of the UPSb study, WES and RNA-Seq were conducted on adamantinoma and OFD-like adamantinoma bone tumours (Chapter 6). Although these tumours are extremely rare with an aetiology that is poorly understood, an overlap in the histopathological features has been noted between these two tumours (Christopher et al., 2013). Moreover, a controversial theory of progression from OFD-like adamantinoma to adamantinoma has been proposed; however, no founded

guidance of this theory currently exists in the literature. To the best of our knowledge, no previous WES or RNA-Seq profiling studies have been conducted on these tumours. Therefore, the current study represents the first high-throughput sequencing study to molecularly characterise these bone tumours.

Using WES, *KMT2D* was the only recurrently mutated gene in 2/8 adamantinoma samples (25%). *KMT2D* has not been reported in adamantinoma previously. No recurrently mutated genes were identified in OFD-like adamantinoma tumours (n=4). However, one OFD-like adamantinoma tumour (OFD-like-ADA-T4) harboured a missense variant in *KMT2D.* It was speculated that this OFD-like adamantinoma may in fact be an adamantinoma. Careful examination of the clinical notes did not reveal potential adamantinoma characteristics in OFD-like-ADA-T4. Moreover, OFD-like-ADA-T4 had very limited sample material, and therefore, performing histopathological analyses were not possible.

KMT2D, which encodes a histone methyltransferase enzyme, is a tumour suppressor gene that is frequently altered in multiple cancers (Hillman et al., 2018). Drug compounds targeting KMT2D have not been reported in the literature (Koren and Bentires-Alj, 2017). However, multiple molecular drugs targeting methyltransferases are currently in early clinical trials (Koren and Bentires-Alj, 2017). Therefore, therapeutics targeting *KMT2D* can be developed in the future.

The average coding mutation rate for adamantinoma tumours (n=8) was 0.84/megabase (range, 0–2.57). By contrast, in three OFD-like adamantinoma tumours, this mutation rate was close to zero, except for OFD-like-ADA-T4 which showed a coding mutation rate of 2.20/megabase. Because OFD-like-ADA-T4 is the same sample that harboured the *KMT2D* variant, overall, this sample shows a mutational profile similar to adamantinoma. However, conclusive adamantinoma

histology in OFD-like-ADA-T4 cannot be established due to the reasons mentioned earlier. The higher mutation rate of adamantinoma than that of OFD-like adamantinoma may explain the more malignant phenotype of adamantinoma tumours. However, the number of analysed samples in this study was limited, and a larger cohort is recommended to confirm these findings.

In the second part of the study, RNA-Seq was conducted on five adamantinoma and three OFD-like adamantinoma tumours. Following a series of filtering and prioritisation steps, one recurrent germline gene fusion (*KANSL1-ARL17A*) was detected in 4/6 (66.7%) adamantinoma tumours and in 3/4 (75%) OFD-like adamantinoma tumours. The *KANSL1-ARL17A* has been recently reported as the first cancer predisposition (germline) gene fusion in many types of cancer identified in patients of European ancestry origin (Zhou et al., 2017).

One somatic gene fusion was identified in an adamantinoma tumour, *EPHB4-MARCH10*; by contrast, no somatic gene fusions were detected in OFD-like adamantinoma tumours. Because *EPHB4* encodes a receptor tyrosine kinase involved in cell differentiation and proliferation (Salgia et al., 2018), *EPHB4-MARCH10* shows potential oncogenic or tumour suppressor activity. However, further investigations are needed to confirm this activity.

Similar to differential expression and hierarchical clustering analyses conducted on UPSb, these analyses can be performed on adamantinoma and OFD-like adamantinoma tumours to investigate potential molecular diagnosis criteria. As of September 2018, Prof Farida Latif has established a collaboration to conduct these analyses in the near future. Hierarchical clustering analysis of adamantinoma and OFD-like adamantinoma tumours would be beneficial to molecularly distinguish

between these two entities, especially in the OFD-like-ADA-T4 tumour that exhibited a mutational profile similar to that in adamantinomas.

The extreme rarity nature of these tumours makes large cohort studies challenging. International collaboration between cancer centres is therefore required to confirm the findings of this study and, in turn, decipher the molecular landscape of adamantinoma and OFD-like adamantinoma tumours.

7.3 Final conclusions

Understanding the genetic landscape of diseases provides insights into translating these discoveries into prevention strategies, diagnostic applications and therapeutic and management opportunities. In the current thesis, NGS technologies were successful in expanding the understanding of genetic alterations and molecular pathogenic mechanisms of rare diseases and cancers. Rare diseases and cancers are often considered as two different groups of conditions; however, this is far from reality and the current thesis is a practical example that demonstrates how these two conditions can be addressed by similar molecular technologies. The identification of recurrently mutated genes and somatic gene fusions in UPSb and adamantinoma is of high interest due to their potential diagnostic and treatment applications.

With their unprecedented abilities to simultaneously analyse the entire genome in a single test, NGS technologies has the potential to revolutionise the diagnosis and treatment of rare diseases and cancers. Moreover, with the advent of national and collaborative sequencing projects and the continuous decrease in WGS cost, NGS will continue to provide novel insights into medical research, human disease and treatment options. Currently, the UK's 100,000 Genome Project is applying NGS technologies to

the diagnosis of both rare diseases and cancers, revolutionising personalised medicine approaches in these two groups of conditions.

Chapter 8: Appendix and supplementary information

8.1 QC metrics for WES data of the CHT and bone tumours projects

8.1.1 WES FastQC QC scores

FastQC tool assess the overall quality of NGS data, including sequenced alignments and bases. A QC HTML report for each sample is generated that consists of two sections: read alignment and sequence quality scores. The read alignment quality section provides QC metrics such as the total number of reads examined, percentage of aligned reads, high quality (≥ Q20 Phred score) aligned bases and percentage of targeted bases with zero, 10X and 20X coverage. A Q20 Phred base quality indicates that a base has a 1 in 100 probability of being incorrect (99% accuracy of base calls). The WES QC alignment scores for CHT samples (Chapter 3) are in Table 8–1; UPSb tumours in Table 8–2; adamantinoma and OFD-like tumours in Table 8–3.

The second section, FastQC tool generates QC metrics for the sequenced bases of the aligned reads. First, the base pair sequence quality graph provides a summary of the quality scores of bases at different positions across the aligned reads (Figure 8–1A). Second, the overall sequences quality score, per sample, provides an overall QC score for all sequences as well as identify if a subset of sequences is of poor quality (Figure 8–1B). In addition to base and sequence quality QC scores, FastQC also provides additional information about the % G-C content across the read, sequence duplication level and insert size from paired end data which are illustrated in Figure 8–2**.** As mentioned earlier, the FastQC sequence QC metrics for all WES projects presented in this thesis showed satisfactory results. For illustrative purposes only the FastQC scores of the UPSb project is shown.

	Lowest value across affected	Highest value across affected	Mean in affected	Lowest value across unaffected	Highest value across unaffected	Mean in unaffected
Total number of reads examined	41,009,650	56,761,530	47,994,014	44,041,090	88,674,564	55,698,076
% reads aligned	99.80%	99.91%	99.84%	99.41%	99.92%	99.79%
High quality (≥ Q20 Phred score) aligned bases	3,916,675,170	6,661,772,045	4,977,879,642	4,282,538,490	10,407,613,596	5,925,312,358
% Duplication	2.31%	3.63%	2.72%	2.24%	3.42%	2.85%
% of targeted bases with zero coverage	0.93%	1.01%	0.98%	0.80%	0.98%	0.90%
% of targeted bases with 10X coverage	96.87%	98.35%	97.61%	97.29%	98.88%	97.92%

Table 8–1: FastQC read alignment QC metrics of the WES data of the four CHT families. The table shows the QC metrics for affected members (n=4) and unaffected family members (n=9) (Original table, compiled from QC analysis provided by the FastQC tool, courtesy of Oxford Gene technology (OGT) company).

	Lowest value across N's	**Highest value across N's**	**Mean in N's**	**Lowest value across T's**	**Highest value across T's**	**Mean in T's**
Total number of reads examined	**40,477,648**	**62,270,318**	51,373,983	**39,565,626**	**67,192,116**	53,378,871
% reads aligned	**99.31%**	**99.74%**	99.33%	**99.35%**	**99.73%**	99.54%
High quality (≥ Q20 Phred score) aligned bases	**3,823,835,106**	**5,874,601,779**	4,849,218,443	**3,952,472,742**	**6,342,421,897**	5,266,210,984
% Duplication	**3.10%**	**4.66%**	3.88%	**3.03%**	**4.21%**	9.52%
% of targeted bases with zero coverage	**0.81%**	**1.04%**	0.93%	**0.86%**	**1.00%**	0.98%
% of targeted bases with 10X coverage	**96.01%**	**97.95%**	96.98%	94.07%	**97.90%**	96.19%
% of targeted based with 20X coverage	**85.93%**	**93.97%**	89.95%	**86.14%**	**94.18%**	92.97%

Table 8–2: FastQC read alignment QC metrics of the WES data for UPSb samples. The values shown here are from nine normal/tumour paired (n=20) and four unpaired tumour samples N: normal; T: tumour. Original table, compiled from information provided by the FastQC tool, ourtesy of OGT company).

	Lowest value across N's	Highest value across N's	Mean in N's	Lowest value across T's	Highest value across T's	Mean in T's
Total number of reads examined	42,219,176	52,561,056	47,004,256	44,181,694	59,705,220	50,968,247
% reads aligned	99.28%	99.36%	99.33%	99.28%	99.37%	99.51%
high quality (≥ Q20 Phred score) aligned bases	3,955,546,098	4,943,335,878	4,412,857,121	4,158,677,835	5,591,713,117	5,448,206,952
% duplication	1.92%	2.90%	2.26%	1.73%	2.99%	4.78%
% of targeted bases with zero coverage	0.85%	0.99%	0.91%	0.84%	0.99%	0.96%
% of targeted bases with 10X coverage	96.90%	97.71%	97.40%	96.48%	98.01%	97.77%
% of targeted bases with 20X coverage	88.40%	91.92%	90.94%	87.90%	94.39%	93.31%

Table 8–3: FastQC read alignment QC metrics of the WES data for adamantinoma and OFD-like adamantinoma samples. The values here are from eight normal/tumour paired (n=20) and three unpaired tumour samples N: normal; T: tumour. Original table, compiled from information provided by FastQC tool, Courtesy of OGT company).

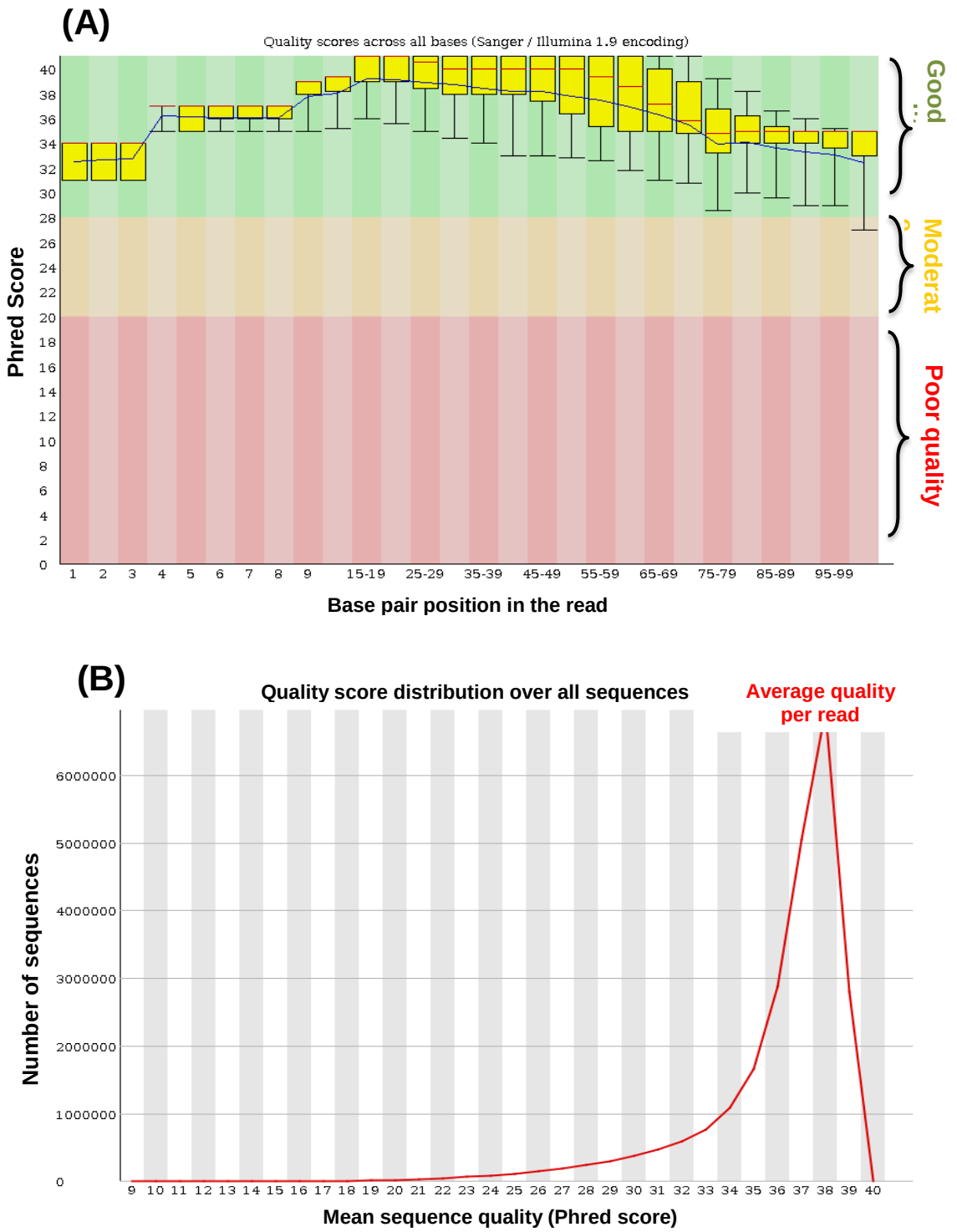


Figure 8–1: A subset of the QC metrics of the sequence quality part in USPb-T1 as an example. (A) An overview of the Phred quality scores for #1-100 bases of the100-bp-length read. A higher Phred (30-40) score increases the probability of the base call being correct. Phred scores of 30 and 40 represents a 99.9% and 99.99% correct base probability, respectively. (B) This graph shows the overall quality score distribution of all the sequences. It allows to detect if subset of sequences is of low quality (as an additional low peak) which is not observed in here since the majority of sequences are of 38 score (one high peak). Generally, a quality score peak of ≥20 is considered acceptable. Images adapted from FastQC tool with added information from (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

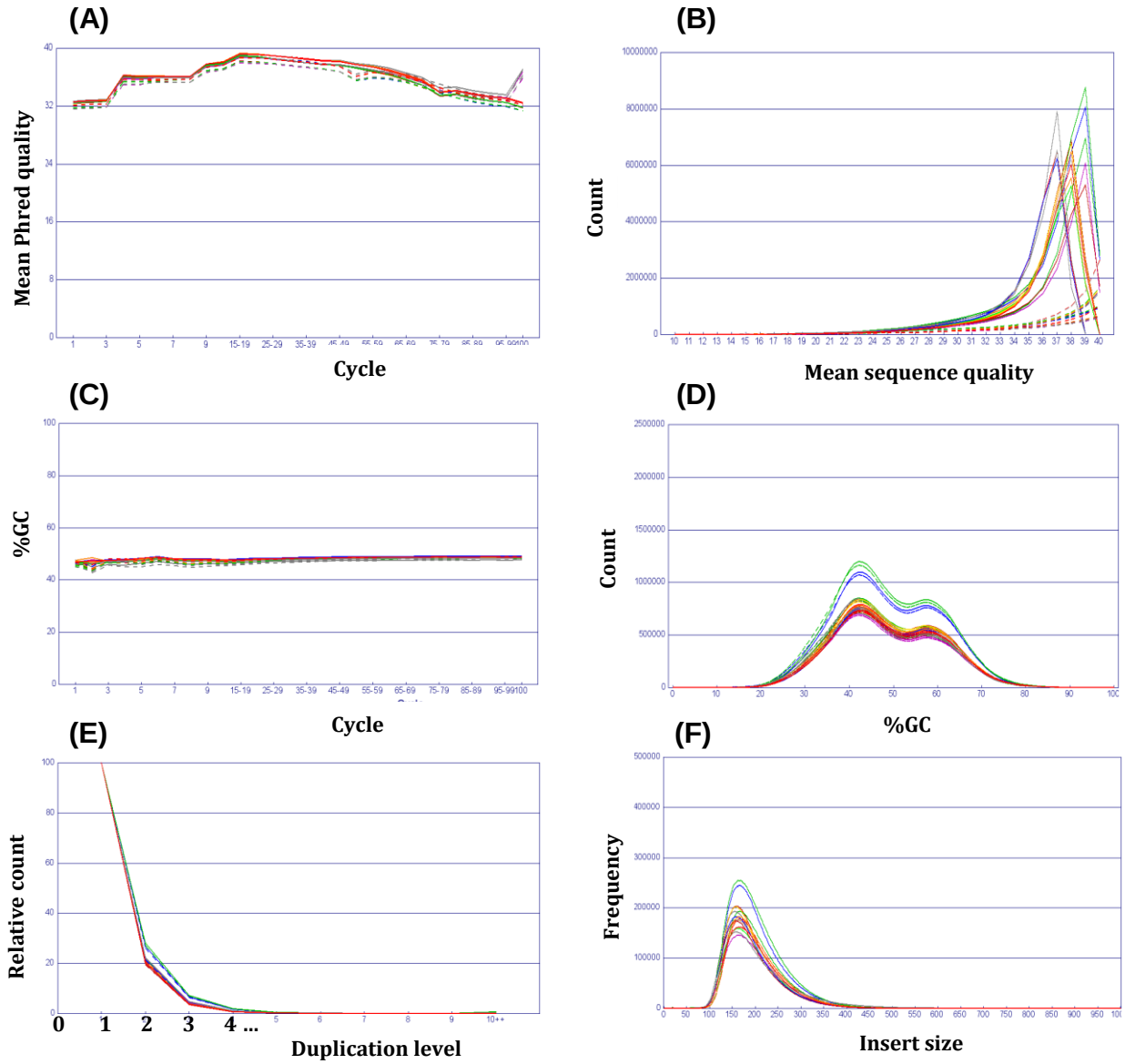


Figure 8–2: Combined FastQC sequence quality graphs for all WES of eight normal-paired UPSb samples. Each coloured graph curve (straight or dotted) represents one of the eight normal-paired paired tumours (n=16). (A) Graph represents the mean Phred base quality score at each position of 100-bp reads. Phred mean quality is expected to decline towards the end of a read. (B) Plot measures the mean of sequence quality for all reads for each sample. The majority of reads are of ≥36 score which are indicative of good sequencing quality. (C) Graph represents the %GC content of a read (at different base positions). A straight graph line indicates a uniform GC content and any deviations from this line (not present here) may reflect sample preparation issues. (D) Plot represents the mean %GC content across the entire read in the whole cohort. The graph curves reflect the GC content and should be of a similar distribution. Although, two paired samples show a slightly higher %GC count, the curves are of similar distribution (i.e. at the same 40% and 60% GC content) comparing to the other samples. (E) The plot represents the degree of duplication for every sequence in a library. A low level of duplication (<2-3), as above, is indicative of minimal enrichment bias (e.g., PCR amplification bias). (F) The graph plots represent the frequency of the observed insert sizes of paired-end reads. The insert size should be similar across the samples from the same sequencing run (Graphs are courtesy of OGT company, with additional information from FastQC, Babraham Bioinformatics http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

8.2 VarScan2 for the detection of somatic alterations in bone tumour projects

Using WES data, VarScan2 was used to identify somatic alterations in UPSb (Chapter 5), adamantinoma and OFD-like adamantinoma (Chapter 6) tumours. By applying a series of heuristic and statistical algorithms, VarScan2 performs a direct comparison of read alignments (from .BAM format), base calls and normalized sequence depth between normal/tumour paired samples to identify germline (inherited), somatic, and loss of heterozygosity (LOH) alterations (Koboldt et al., 2012) (Figure 8–3). Each of these categorized variants is further classified as high confidence or low confidence. A somatic variant is classified as high confidence if the VAF in tumour is ≥10% and <5% in normal (Koboldt et al., 2012). To further refine identified variants, VarScan2 applies a false-positive filter to discard false positive calls that occur as result of sequencing or alignments artefacts. This filter assesses every variant for nine empirically derived criteria (e.g., read position, VAF, map/read quality difference) to differentiate true calls from likely false ones (Koboldt et al., 2012).

The default VarScan2 settings for variant detection are: ≥ 8X read depth, ≥20 Phred base quality, ≥10% VAF and a significant *P*-value of <0.05 (Koboldt et al., 2012). Variants identified in both tumour and normal samples are classified germline. If a variant is identified in one sample (e.g., tumour and not in corresponding normal), the statistical significance is computed by Fisher's exact test in which both the variant and reference allele are compared to expected distribution in relation to sequencing error alone (Koboldt et al., 2012). This variant is classified somatic if the calculated *P*-value is statistically significant. A variant is classified LOH if it is heterozygous in normal and homozygous in tumour; homozygous if it has a VAF of ≥75% (Koboldt et al., 2012).

Because of the various algorithmic calculations, Varscan2 recommends that normal-tumour pairs have been processed under same experimental settings (e.g., exon

capturing and sequencing platform) (Koboldt et al., 2012). This recommendation was achieved in all exome-sequenced tumour samples.

In addition to variant detection, VarScan2 can also identify somatic CNVs based on a calculated $\log_2$ ratio of normal and tumour sequence depth for each contiguous region. After applying minimum read depth and Phred requirements, significant deviations from the log-ratio are classified as chromosomal losses or gains (Koboldt et al., 2012).

8.3 Identification of somatic point variations using MuTect

MuTect was the second somatic calling tool used in the bone tumour projects (Chapter 5 and 6). By following a series of algorithmic steps, MuTect analyses the aligned sequence data (.BAM file) from normal and tumour samples independently for SNVs detection (Figure 8–4). The MuTect variant detection methodology consists of four main steps: (1) filtering/removing of lower quality sequencing data, (2) variant identification in tumour data by applying a Bayesian classifier, (3) a second filtering to further remove false positive calls missed in step one, and (4) classifying variants as somatic or germline (Cibulskis et al., 2013).

To remove lower quality sequencing data, MuTect preforms standard preprocessing steps that consist of: labelling duplicate reads, recalibration of base quality scores and local realignment of reads. Although both VarScan2 and MuTect analyse tumour and normal alignments, MuTect pre-processes normal and tumour reads differently and according to how they will be used. That is, only high-quality variants are called in tumour to eliminate false positive calls; whereas, less stringent filters are applied to the normal sample in order to detect systemic artefacts more easily or for classifying variant somatic status (see Figure 8–4) (Cibulskis et al., 2013).

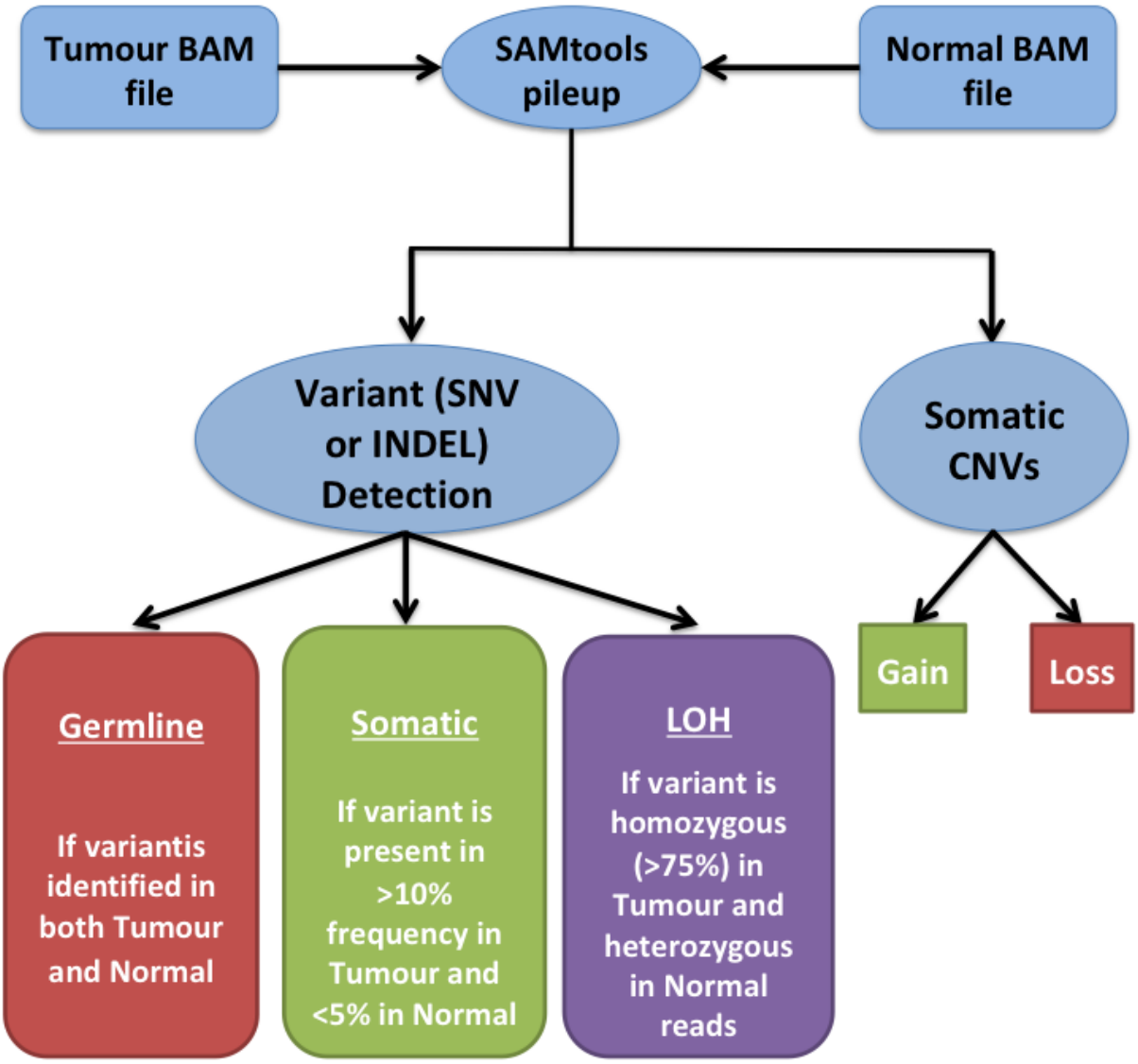


Figure 8–3: VarScan2 pipeline to identify somatic alterations in tumour/normal paired samples. First, tumour and normal BAM alignments are simultaneously analysed by performing pairwise comparison of base calls and normalized read depth to create a SAMtools pileup. Alteration are then classified into two categories 1) Detected Variants: single nucleotide variants (SNVs) or small insertions/deletions (INDELs), and 2) chromosomal copy number alterations. For each identified SNV or INDEL, VarScan2 performs a comparison between tumour and normal genotypes. A variant is classified 1) germline: if present in ≥10% variant allele frequency (VAF) in both tumour and normal samples, 2) somatic: if present in ≥10% VAF in tumour and <5% in normal, and 3) Loss-of-heterozygosity (LOH): if heterozygous in the normal and homozygous (≥75% VAF) in tumour. Copy number variations are identified by comparing the reads' depth (with at least Phred Q20 score) between normal and tumour samples, followed by normalization of read depths. Figure redrawn with modifications from (Koboldt et al., 2012).

After removing low quality reads, MuTect performs statistical analyses to identify sites that harbour somatic variations in preprocessed tumour alignments. A Bayesian probability classifier is applied to detect tumour sites that deviate from the genome reference (i.e., SNVs) (Cibulskis et al., 2013). A standard SNVs callset is produced and subsequently passes through six filters to further discard non-independent sequencing artefacts and generate a high confidence callset (Figure 8–4). Finally, a second Bayesian probability classifier is used to ensure variants identified in the first Bayesian classifier are not present in the normal sample (Cibulskis et al., 2013). In theory, germline classification of variants is determined based on calculating a log odds (LOD) score and comparing it to a cutoff that is calculated based on of the log-ratios of: (1) probability of tumour site deviating from the reference; (2) probability of the absence of non-reference site in normal sample (Cibulskis et al., 2013).

8.4 RNA sequencing

8.4.1 RNA-Seq data processing: overview of data QC of reads for bone tumour projects

RNA-Seq data analysis requires a series of processing steps which were performed using the TopHat Alignment v1.0 and Cufflinks Assembly bundle tools v1.1, available in the Tuxedo package (Figure 8–5). Prior to analysis, for each sample, the raw reads are assembled into FastQ files by TopHat Alignment bundle tool. The foremost data processing step is to quality check the RNA-Seq data. TopHat Alignment bundle tools provide a detailed RNA-Seq QC metrics report, and visual representation for each sample using the Picard tools (more in the following section) (http://broadinstitute.github.io/picard/).

Total aligned reads percentage is an essential mapping quality parameter of aligned reads and can be indicative of the overall RNA-sequencing accuracy and to detect DNA contamination of RNA sample (Conesa et al., 2016). The percentage of total aligned reads, RNA-sequenced reads mapping to human genome, is expected to be 70-90% (Dobin et al., 2013) which was achieved in all UPSb, adamantinoma and OFD-like adamantinoma tumours. Details of the number QC metrics of RNA-sequenced reads, percentage of total aligned and unaligned reads for UPSb are in Table 8–4; adamantinoma and OFD-like adamantinoma in Table 8–5.

8.4.2 RNA-Seq read alignment and identification of gene fusions using TopHat-Fusion and STAR-Fusion

Following QC checks, the TopHat Alignment tool uses Bowtie (v0.12.9) and TopHat2 tools (v2.0.7) to align reads and identify candidate gene fusion (Illumina App Guide, 2016) (Figure 8–5). Bowtie is a high-throughput bioinfomatic tool that simultaneously applies multiple processors to align large datasets (e.g., paired-end reads) to a reference genome (Langmead et al., 2009). Bowtie alignment data processing also forms the basis for TopHat-2 tool that includes TopHat-Fusion splice junction mapper (Langmead et al., 2009). TopHat2 is a high-throughput splice junction mapper that used aligned RNA-Seq reads to identify splice junctions between exons (Kim et al., 2013). TopHat2 also provided analysis metrics such as the number of reads generated (read depth), reads spanning both ends/partners of a fusion and reads spanning fusion junctions

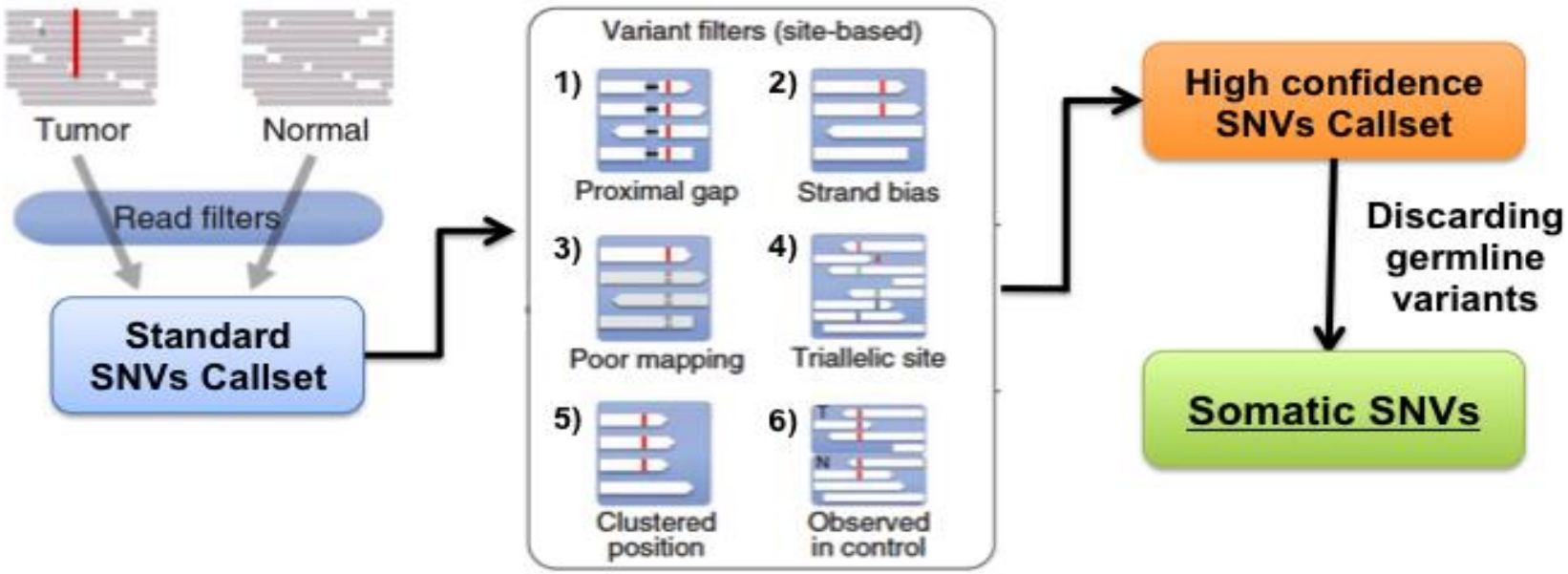


Figure 8–4: Overview of the MuTect analyses steps for somatic SNVs detection. Alignment reads are preprocessed (read filters) as following 1) in tumour, reads are only retained if they pass six parameters, some of which are: alignment mapping quality of ≥1, base quality score of ≥5, 2) In matched normal, two parameters are applied: 1) ≥5 base quality score; 2) reads that disagree with the reference (i.e., show a variant) are retained. A red vertical line denotes to a variant observed in tumour reads but not in the normal sample. To generate high confidence callset, six filters are further applied 1) Proximal gap: discard false positive calls produced by nearby misaligned deletions/insertions, 2) Poor mapping: discard false positive calls produced by sequence similarity in genomic regions, 3) Triallelic site: remove false positive calls if a site has three alleles (A, B and C), 4) Standard bias: discards false positive calls that are identified in reads with a single direction (rather both direction reads), 5) Clustered position: if alternate alleles are clustered at a consistent positions at the reads (e.g., start of read), these false calls are rejected., and 6) Observed in control/normal: discard variants that are present in ≥3% MAF in normal sample. Lastly, these high confidence calls are classified as somatic or germline (Figure reproduced with modifications from Cibulskis et al., 2013)

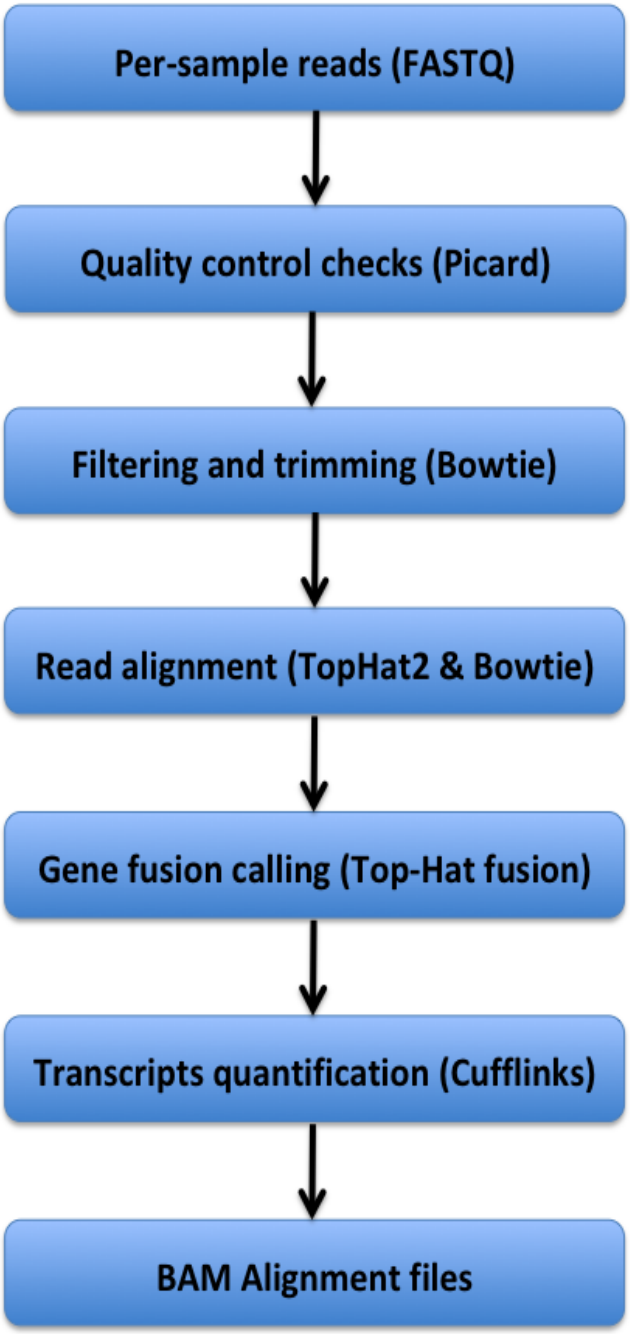


Figure 8–5: TopHat Alignment bundle tool workflow used to analyse RNA-Seq data and identify gene fusions. Original figure, compiled from information in (Illumina App Guide, 2016).

Reads	Number of reads[a]	% Total Aligned[b]	% Unaligned[c]
UPSb-T1	49,812,773	86.95%	8.49%
UPSb-T2	52,868,488	94.13%	6.85%
UPSb-T5	54,190,792	92.84%	5.30%
UPSb-T6	64,083,181	92.70%	4.62%
UPSb-T9	55,395,338	91.97%	4.61%
UPSb-T10	68,527,027	93.27%	3.36%
UPSb-T13	75,770,827	93.09%	3.38%
UPSb-T14	57,132,505	75.58%	24.42%
Mean in all samples	**59,722,616**	**90.07%**	**7.63%**

Table 8–4: Details of the generated RNA-Seq reads in eight UPSb tumours. [a] reads passing QC filter. [b] the percentage of passed filter reads that are successfully aligned to the reference genome (GRCh37). [c] the percentage of passed reads that failed to align to the reference genome (Table reproduced from QC metrics output provided by TopHat2 alignment tool in the Illumina BaseSpace hub).

Reads	Number of reads[a]	% Total Aligned[b]	% Unaligned[c]
ADA-T1	69,044,024	94.61%	5.39%
ADA-T3	72,742,406	94.23%	5.77%
ADA-T4	56,220,144	93.05%	6.95%
ADA-T5	82,604,711	94.33%	5.67%
ADA-T8	82,085,066	93.91%	6.09%
OFD-like-ADA-T2	57,891,707	93.47%	6.53%
OFD-like-ADA-T3	76,757,173	93.70%	6.30%
OFD-like-ADA-T4	60,940,983	93.70%	6.30%
Mean in all samples	69,785,777	93.88%	6.13%

Table 8–5 Details of the generated RNA-Seq reads in five adamantinoma and three OFD-like adamantinoma tumours. [a] reads passing QC filter. [b] the percentage of passed filter reads that are successfully aligned to the reference genome (GRCh37). [c] the percentage of passed reads that failed to align to the reference genome (Table reproduced from QC metrics output provided by TopHat2 alignment tool in the Illumina BaseSpace hub).

First, input reads were filtered to remove abundant sequences (listed in an internal library of the tool) such as mitochondrial and ribosomal RNA. Paired reads are discarded if at least one read aligns to an abundant sequence (Illumina App Guide, 2016). Bowtie trims two bases from the 5'-end of the read due to the high mismatch rate in these bases. Second, filtered paired-end sequence reads are aligned to the reference genome (GRCh37/hg19 assembly) by Bowtie tool (Illumina App Guide, 2016).

After mapping the reads that align completely to exons, Bowtie identifies initially unmapped reads and splits each read into 25-bp segments (for 75bp read lengths). These 25bp segment reads are mapped to the reference genome, enabling for

detection of segments that map to the fusion gene partners (discordant reads). TopHat2-Fusion is designed used to discover transcripts resulting from fusion gene products (Kim and Salzberg, 2011) (Figure 8–6). To reduce the likelihood of false positive gene fusion calls, TopHat2-Fusion requires that fusion junction/supporting reads have at least 13-bp aligning to both left and right partners of the fusion with ≤ 2 mismatches. TopHat2-Fusion identifies fusion candidate genes if: (1) the two sides/genes of the fusion reside on different chromosome; or (2) both gene fusion partners reside on the same chromosome and are separated by at least 100 Kb to exclude majority of read-through events. A gene fusion candidate must have at least one fusion junction/supporting read.

After identifying fusion candidates, the 25-bp segments are combined/stitched together and re-aligned again to produce full read alignments. Subsequently, a heuristic scoring system is applied to identify best-read alignments and discard multi-mapped reads that map to multiple genomic locations. Lastly, for confident fusion candidates, TopHat-2 provide statistics about the genomic locations of both gene partners, number of fusion junction/supporting reads and fusion spanning reads—reads that map to either side/gene of the fusion breakpoint.

STAR-Fusion was the second calling tool used for gene fusion identification. Similar to TopHat-Fusion, STAR-Fusion rapidly maps RNA-Seq reads to reference genome to identify discordant mapping reads (details in Figure 8–7).

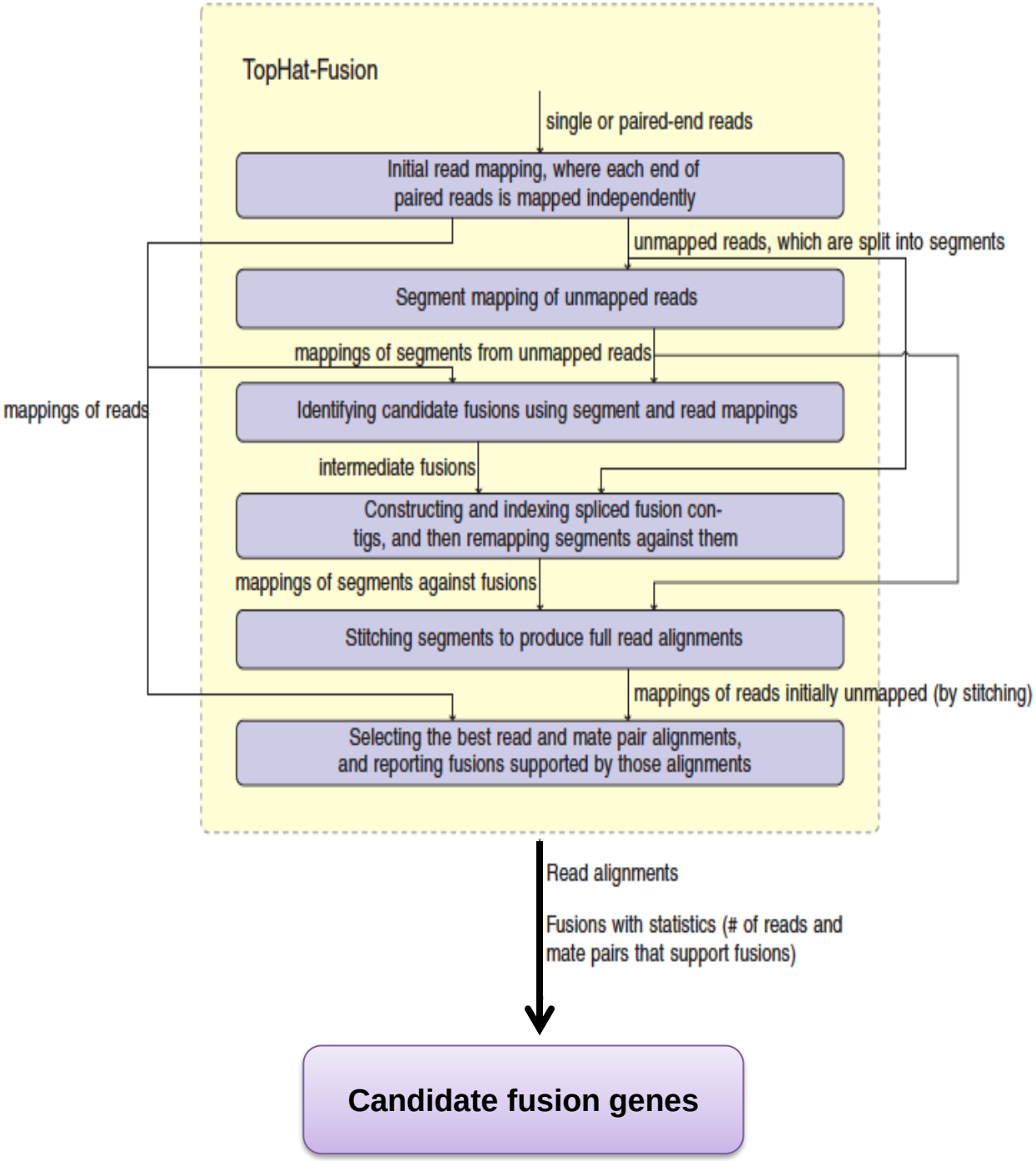


Figure 8–6: TopHat-Fusion algorithm pipeline for the identification of gene fusions. The figure summarizes the series of steps followed by TopHat-Fusion for gene fusion calling. Figure reproduced with minor modifications from (Kim and Salzberg, 2011).

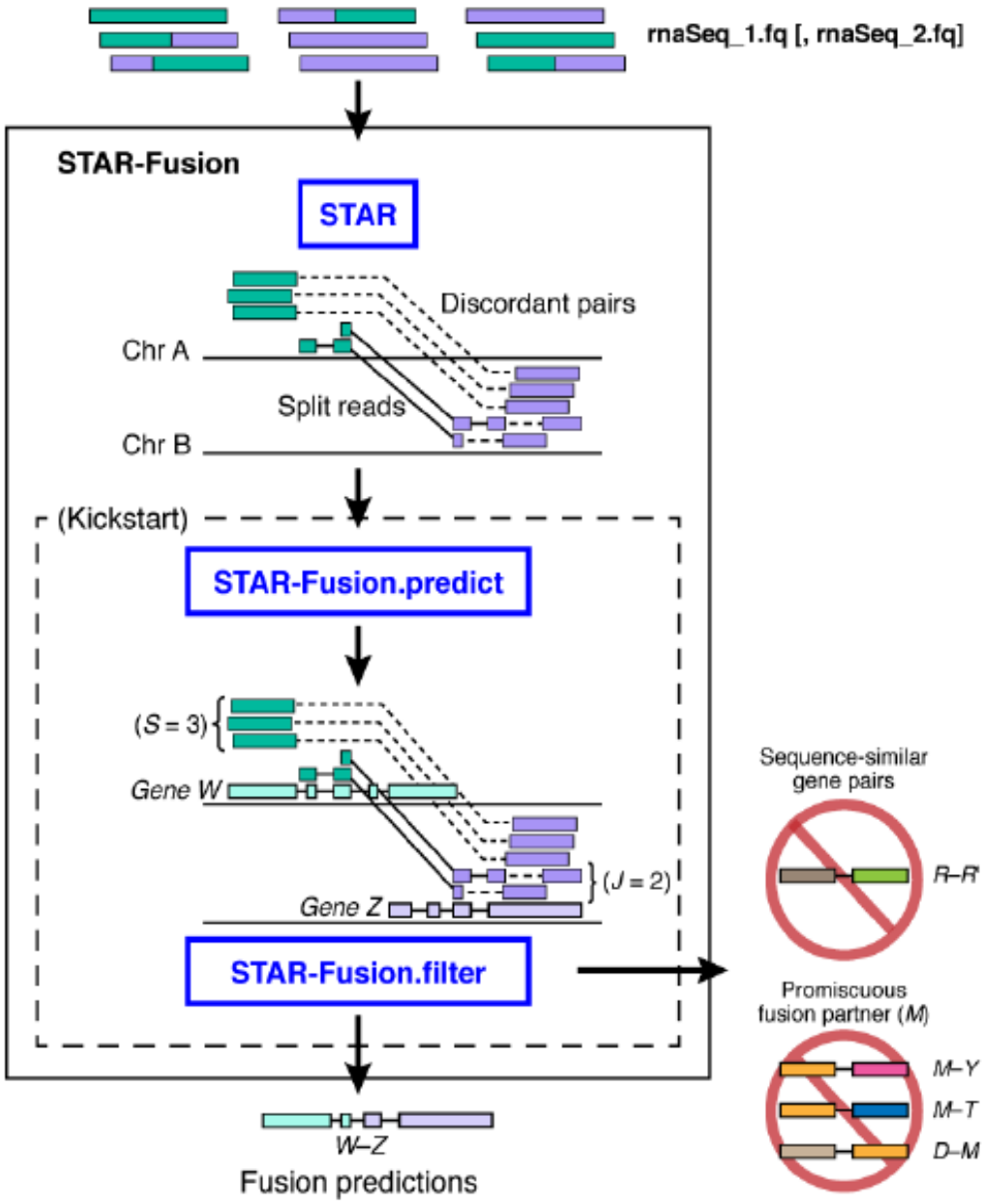


Figure 8–7: Overview of STAR-Fusion workflow for gene fusion detection. STAR algorithm is used to align RNA-Seq reads to the reference assembly (GRCh37). In STAR-Fusion predict algorithm, discordant/split reads, indicative of candidate gene fusions, are identified and mapped to reference transcript annotations. These identified discordant reads are then mapped to the exons of reference transcript annotations, followed by identifying splits reads—reads that span and support the fusion breakpoints (fusion gene partners). To discard likely artefacts, STAR-Fusion filter algorithm discards similar fusions pairs (partners sharing high sequence similarity). Promiscuous fusion genes, known as gene partners found to fuse with multiple fusion partners in one sample, are also discarded. Lastly, gene fusion passing the abovementioned filers are labelled as potential candidate fusions. Figure reproduced with minor modifications from (Haas et al., 2017)

8.5 Supplementary figures and tables

8.5.1 Chapter 3: CHT families study using WES

Family-1	Family-2	Family-3	Family-4
C1orf222	*LAMC1*	*TRIM65*	*RP11-181C3.1*
NBEAL2	*DZANK1*	*HLA-A*	*MEX3C*
FAM71D	*SLC39A14*	*PALLD*	*BCL2L12*
MUC4	*NCAM1*	*KTI12*	*TIAM1*
PCDHB8	*PPL*	*HLA-DQB1*	*NOM1*
PLCH2	*ZNF66*	*HLA-DQB2*	*SYNPO*
MASP2	*HSPBP1*	*KRTAP10-10*	*ERP29*
THADA	*SIRPA*	*PODXL*	*MBOAT7*
SIX2	*DISC1*	*EXOG*	*EIF2B4*
ZC3H14	*PAK1IP1*	*FYCO1*	*IGHV3-30*
MAP1A	*AHNAK2*	*SETD2*	*AKT1S1*
AIRE	*MYO9A*	*PARP3*	*URB1*
PREPL	*SYNM*	*ATF6B*	*CRIPAK*
C21orf2	*RHBDD2*	*NDUFS7*	*PHGR1*
SIK1	*OR51I1*	*LRRC15*	*ZNF714*
AC022532.1	*SH3TC1*	*HIST1H3C*	
ADAM21	*RPP25*	*SLC25A47*	
SEC24C	*ZSCAN2*	*CYP4B1*	
SEH1L	*TLE3*	*ZDBF2*	
COL18A1	*ANKS3*	*OR8B8*	
	SUSD4	*ZNF479*	
	HLA-DRB1	*SORL1*	

Family-1	Family-2	Family-3	Family-4
	TNFRSF10A	*GJB6*	
	CDHR5	*SLC1A7*	
	CYP11A1	*TRAK1*	
	MMP25	*FANCE*	
	ZNF714	*C6orf165*	
	VSTM2B	*COL14A1*	

Table 8–6: List of the autosomal recessive candidate genes identified by WES in the four CHT families. All families are associated with thyroid gland dysgenesis except for Family-2 which is associated with thyroid dyshormonogenesis.

8.5.2 Chapter 5: UPSb study using WES and RNA-Seq

Sample	INDELs	SNVs
UPSb-T1	89	417
UPSb-T2	99	940
UPSb-T3	1610	1761
UPSb-T4	88	374
UPSb-T5	81	498
UPSb-T6	100	627
UPSb-T7	356	75
UPSb-T8	64	399
UPSb-T9	211	3058
UPSb-T10*	2,549	37802
UPSb-T11*	3009	41585
UPSb-T12*	3880	63164
UPSb-T13*	2934	35724
UPSb-T14	241	949

Table 8–7: The total number of unfiltered variants identified in UPSb. *Tumours that are unpaired with own corresponding normal. Underlined sample is FFPE. The coding mutational rate (overall somatic tumour burden) was calculated by the dividing the total number of identified somatic variant by the length of the entire exome (30 megabase).

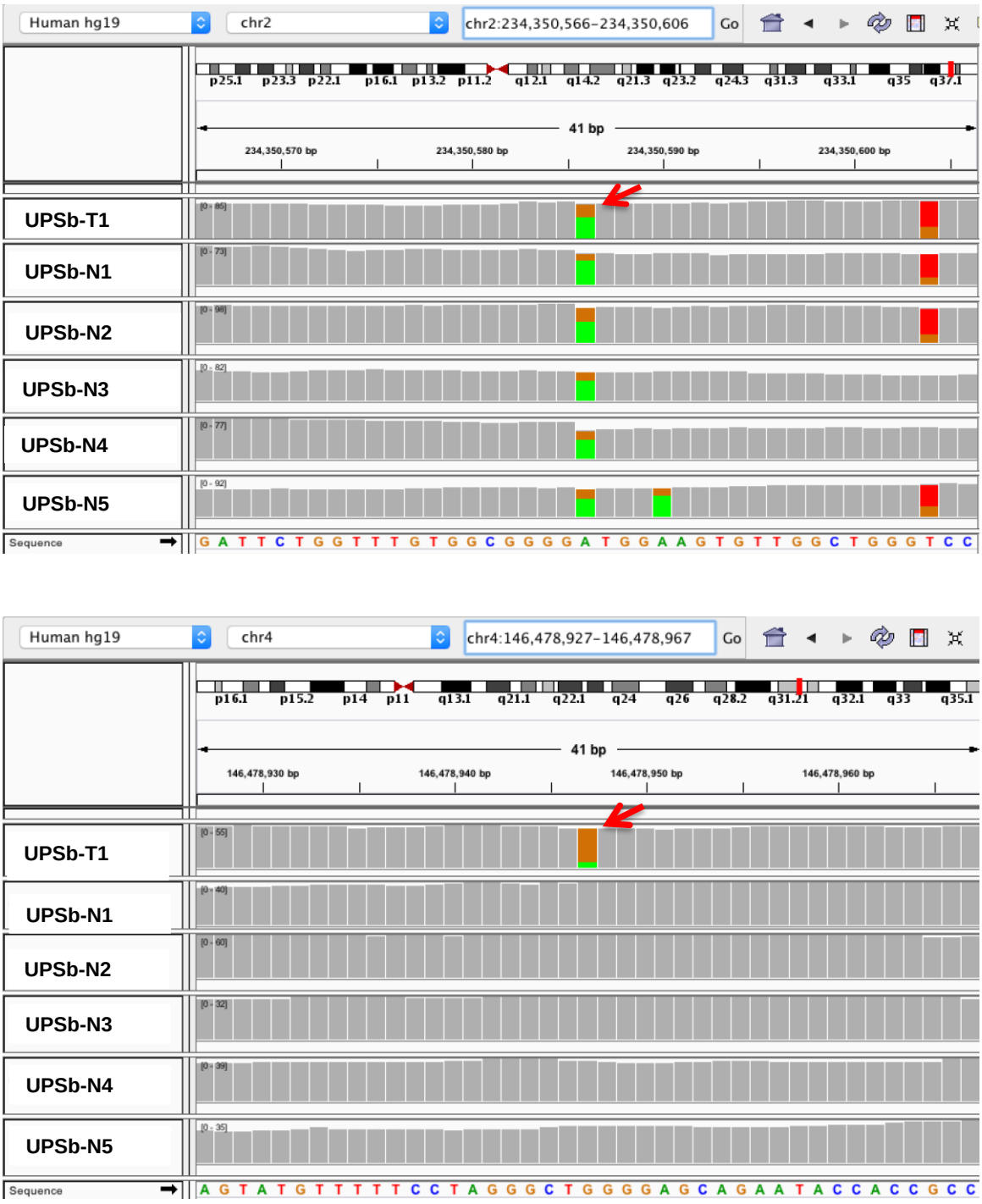
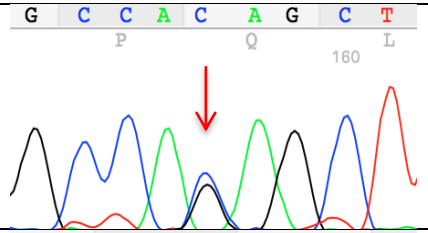
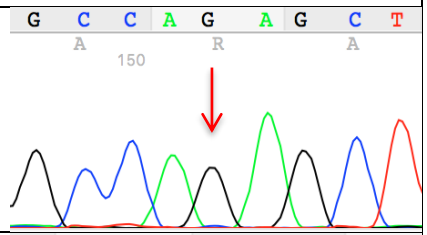
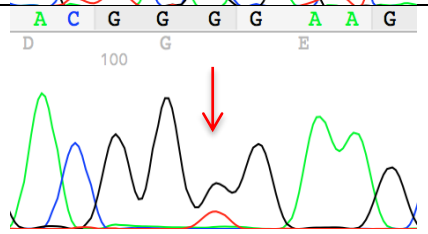
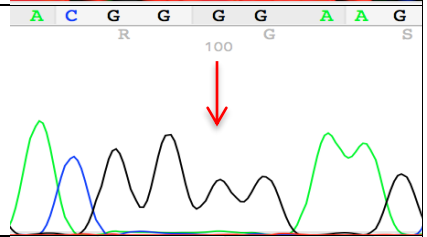
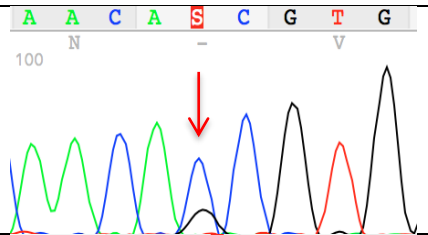
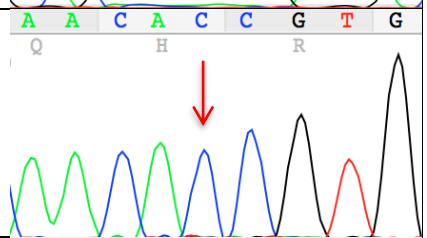
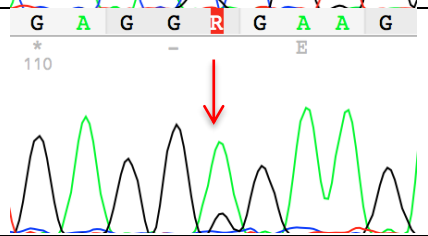
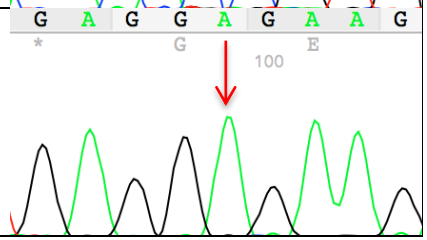
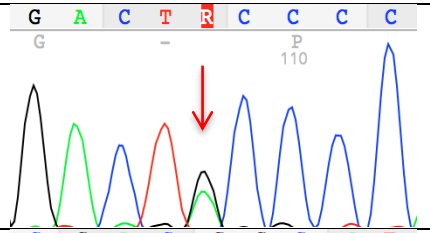
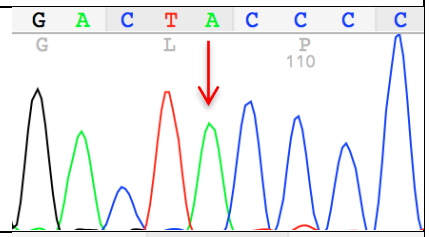
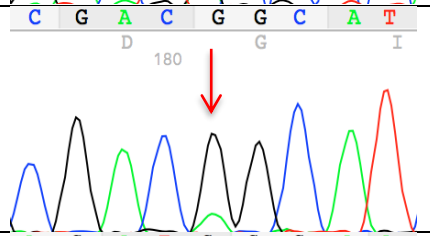
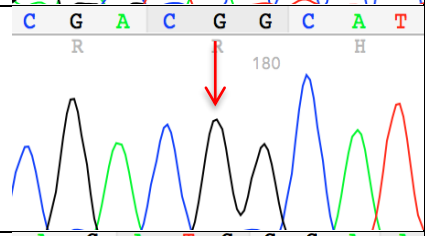
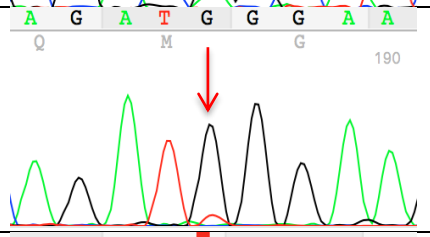
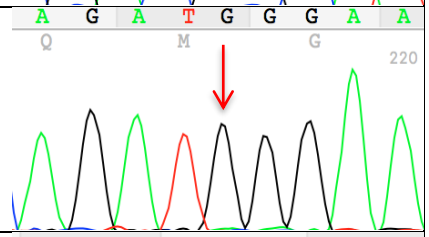
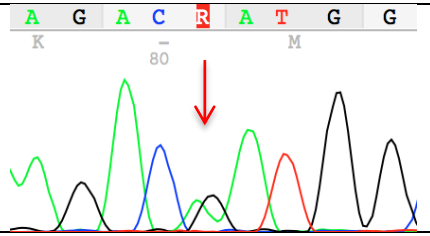
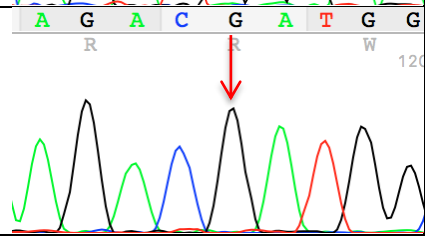
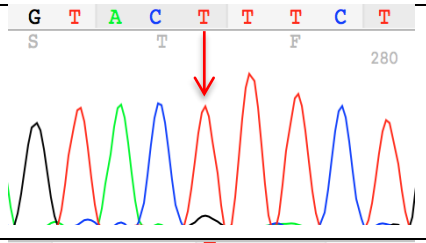
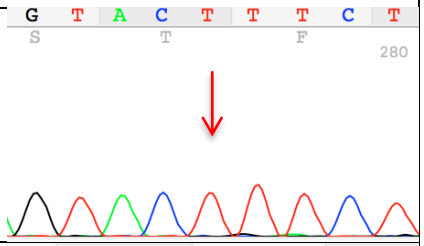
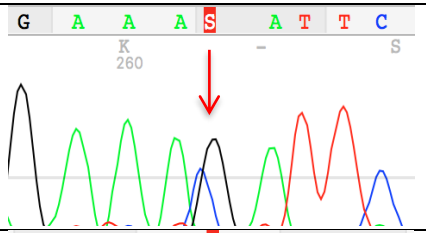
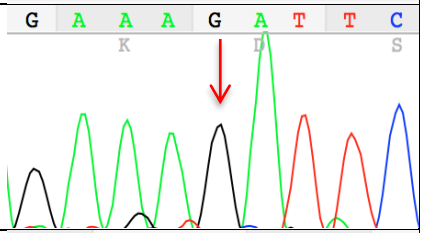
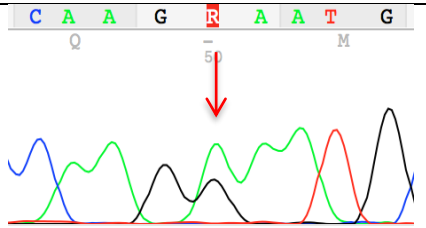
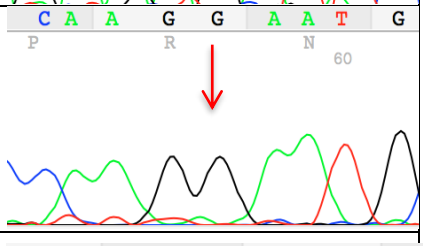
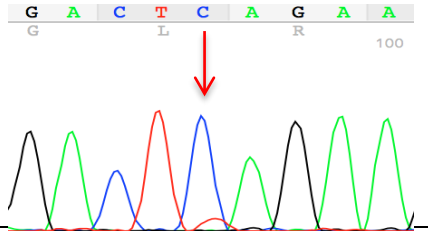
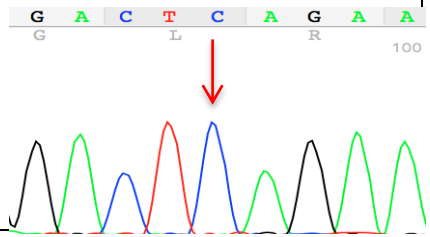
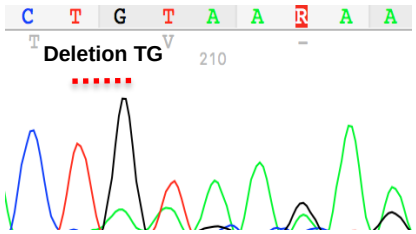
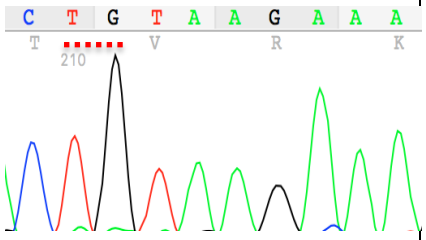
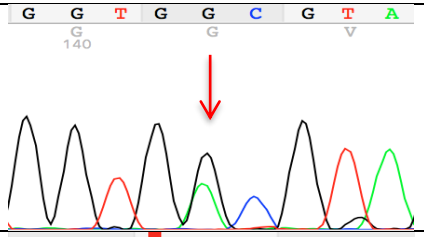
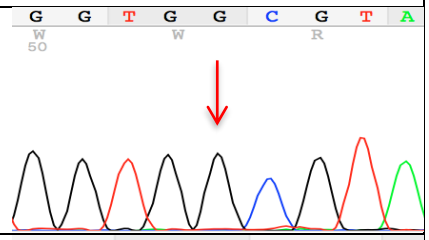
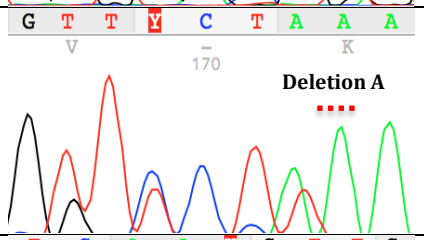
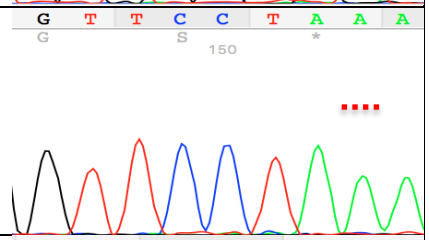
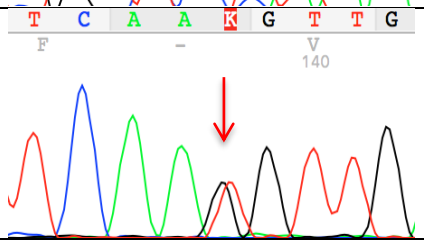
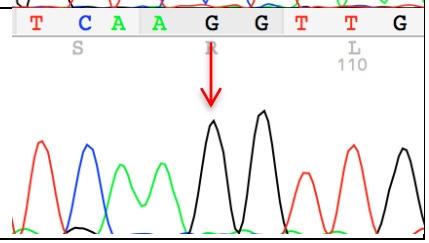


Figure 8–8: Visualising somatic variants using the Integrative Genomics Viewer (IGV) to identify genuine somatic calls from false positive and/or germline ones. Top IGV image shows an A>G called variant (red arrow) in UPSb-T1 tumour. This variant is also present in five other normal samples and therefore is very likely germline or sequencing artefact. Bottom IGV image shows a G>A variant that present in UPSb-T1 and absent from the corresponding normal and four additional normal samples. Therefore, this variant is very likely a genuine somatic call. Note: for illustrative purposes, only five sequenced normal samples are shown in here rather than the total normal samples (n=10) used in BAM visualising of variants.

Gene	ID	WES variant details	Depth; %VAF	COSMIC ID; rs ID	SIFT; PolyPhen2	Tumour	Matched normal
DOT1L	T3	Missense; c.1785G>C; p.Gln595His; (Somatic)	37; 49%	Not reported	Tolerated; Damaging		
	T6	Missense; c.920G>T; p.Gly307Val; (Somatic)	122; 36%	Not reported	Deleterious; Damaging		
PTPRT	T1	Missense; c.3794C>G; p.Thr1265Ser ; (Somatic)	30; 17%	Not reported	Deleterious; Damaging		
	T3	Missense; c.1499A>G; p.Glu500Gly; (Somatic)	25; 30%	Not reported	Deleterious; Damaging		

Gene	ID	WES variant details	Depth; %VAF	COSMIC ID; rs ID	SIFT; PolyPhen2	Tumour	Matched normal
PKLR	T3	Missense; c.611A>G; p.Tyr204Cys; (Somatic)	62; 46%	Not reported	Deleterious; Damaging		
	T6	Missense; c.994G>A p.Gly332Ser; (Somatic)	123; 10%	rs773626254	Deleterious; Damaging		
PEG3	T3	Missense; c.1198G>T; p.Gly400Trp; (Somatic)	81; 19%	COSM6485653	Deleterious; Damaging		
	T6	Missense; c.790G>A; p.Asp264Asn ; (Somatic)	32; 48%	COSM1712957	Deleterious; Damaging		

Gene	ID	WES variant details	Depth; %VAF	COSMIC ID; rs ID	SIFT; PolyPhen2	Tumour	Matched normal
SYNE2	T3	Missense; c.221T>G; p.Leu74Arg; (Somatic)	126; 11%	Not reported	Deleterious; Damaging		
	T6	Missense; c.7714G>C; p.Asp2572His; (Somatic)	137; 37%	Not reported	Deleterious; Damaging		
COL4A2	T12	Missense; c.3035G>A; p.Gly1012Glu (Somatic)	37 / 35%	COSM4541607	Deleterious; Damaging		
EGF	T5	Nonsense; c.3055C>T; p.Gln1019Ter (Somatic)	14; 21%	Not reported	N/A		

Gene	ID	WES variant details	Depth; %VAF	COSMIC ID; rs ID	SIFT; PolyPhen2	Tumour	Matched normal
COL1A2	T1	Frameshift; c.87_88delTG; p.Val30LysfsTer8; (Somatic)	22; 32%	Not reported	N/A		
KRAS	T10	Missense; c.38G>A; p.Gly13Asp (Somatic)	37; 35%	COSM1140132	Deleterious; Damaging		
FAT2	T3	Frameshift; c.8055delA; p.Val2686TyrfsTer41; (Somatic)	55: 20%	Not reported	N/A		
RB1	T2	Splice site; c.940-1G>T; (Somatic)	32; 37.5%	Not reported	N/A		

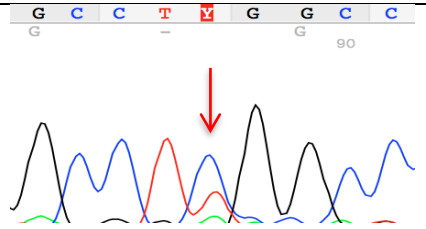
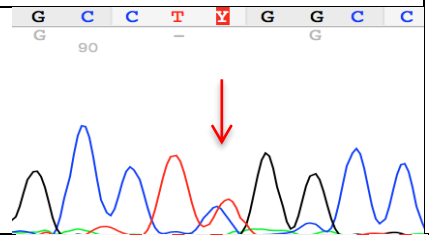
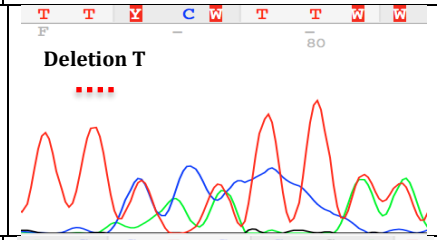
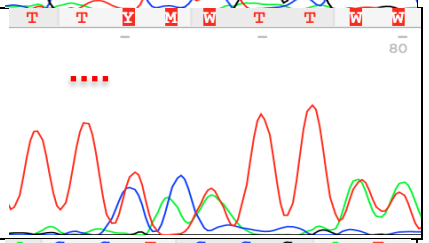
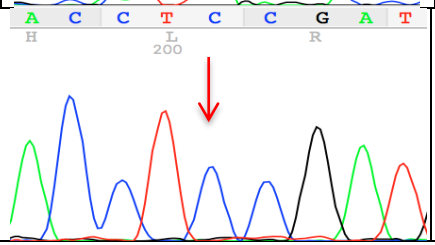
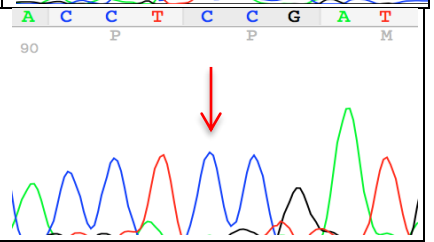
Gene	ID	WES variant details	Depth; %VAF	COSMIC ID; rs ID	SIFT; PolyPhen2	Tumour	Matched normal
ROBO2	T11	Missense; c.41C>T; p.Arg15Trp; (Germline)	169; 51%	rs62250276	Deleterious; Damaging		
TXNRD1	T11 [	Frameshift; c.1432delT; p.Ser480Hisf sTer38; (Germline)	19; 62%	Not reported	N/A		
ERCC3	T12	Missense; c.989C>T; p.Ser330Phe (False positive)	44; 15%	Not reported	Deleterious; Damaging		

Table 8–8: Sanger sequencing confirmation of WES variants identified in UPSb tumours. All variants identified are heterozygous. WES variants details column consists: **1)** the type of WES variant (e.g., missense); **2)** coding variants position (c.), **3)** Affected translated protein (p.) **4)** Somatic, germline or false positive variant status. %VAF: altered variant allele frequency. COSMIC# represent the accession number of somatic variants identified in the COSMIC database, if present. rs ID represents the accession number of a known polymorphism, if present.

Sample	Gene fusion	Right gene genomic coordinates	Left gene genomic coordinates	No. of fusion junction/supporting reads	No. of fusion spanning reads
UPSb-T2	*PKNOX2-MMP20*	chr11:125237881	chr11:102449873	96	11
	CMAS-PYROXD1	chr12:22218219	chr12:21620413	22	9
UPSb-T6	*OSBPL2-CABLES2*	chr20:60831277	chr20:60971648	8	8
	MICAL3-UFD1L	chr22:18507047	chr22:19459331	34	7
UPSb-T9	*CLTC-VMP1*	chr17:57746301	chr17:57915656	156	101
	APOL1-MYH9	chr22:36661642	chr22:36681999	126	61
UPSb-T13	*ADAM17-ASAP2**	chr2:9637243	chr2:9525554	68	87
	*FARP1-STK24**	chr13:98896837	chr13:99127075	15	38

Table 8–9: Somatic gene fusions identified by TopHat-Fusion and STAR-Fusion and validated by RT-PCR in UPSb tumours. Gene fusions with an asterisk (*) represent gene fusions formed in a 5' to 5' gene orientation; that is, the 5' region of the right gene is fused with the 5' region of the left gene.

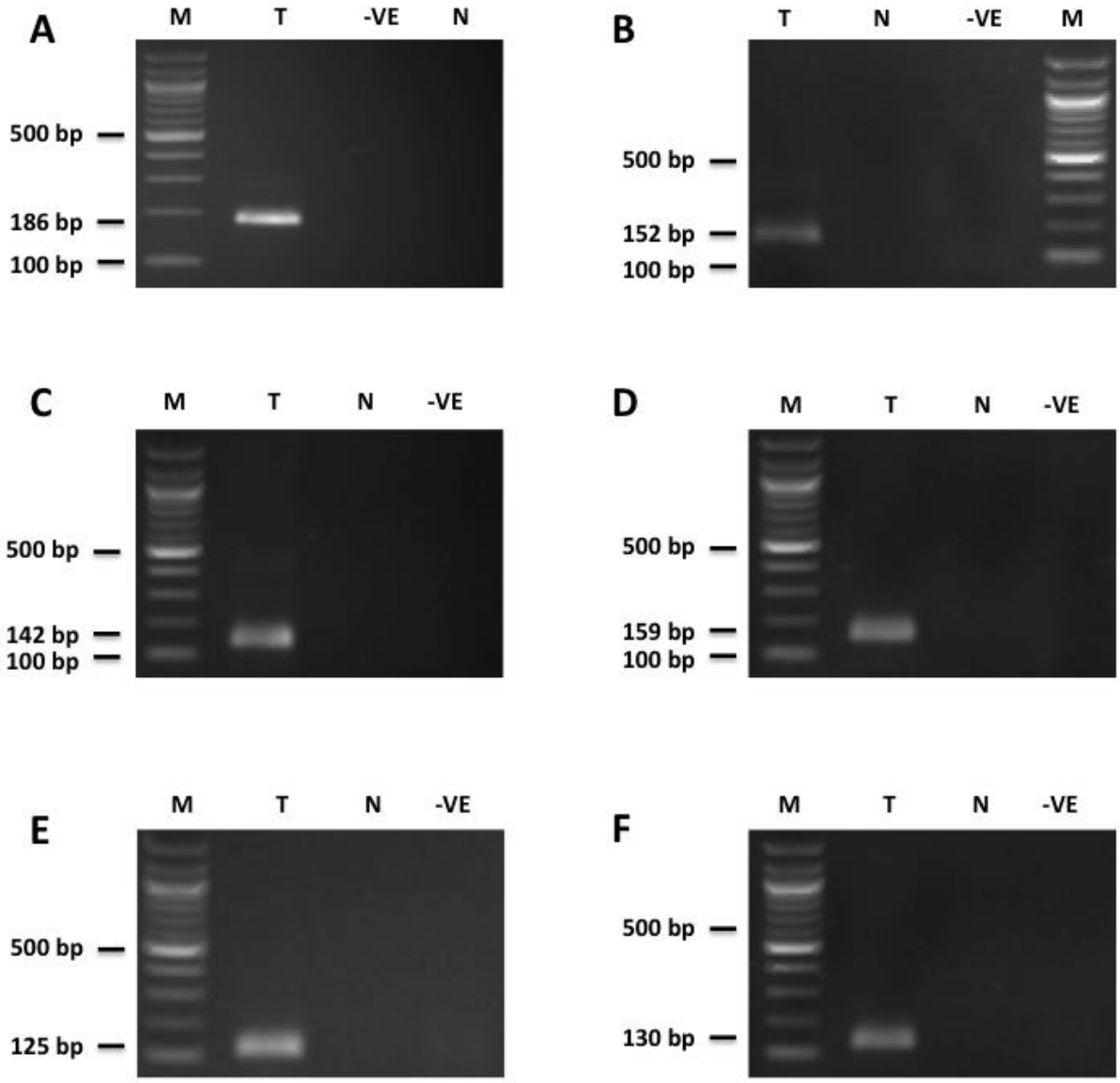


Figure 8–9: RT-PCR somatic confirmation of six gene fusions identified in UPSb tumours at the cDNA level. Primers flanking gene fusion junction confirmed the following gene fusions (A) *PKNOX2-MMP20*, (B) *ASAP2-ADAM17*, (C) *APOL1-MYH9*, (D) *CMAS-PYROXD1*, (E) *MICAL3-UFD1L*, and (F) *OSBPL2-CABLES2*. M: 100 basepair DNA marker; T: tumour; N: corresponding normal; –VE: negative control (water).

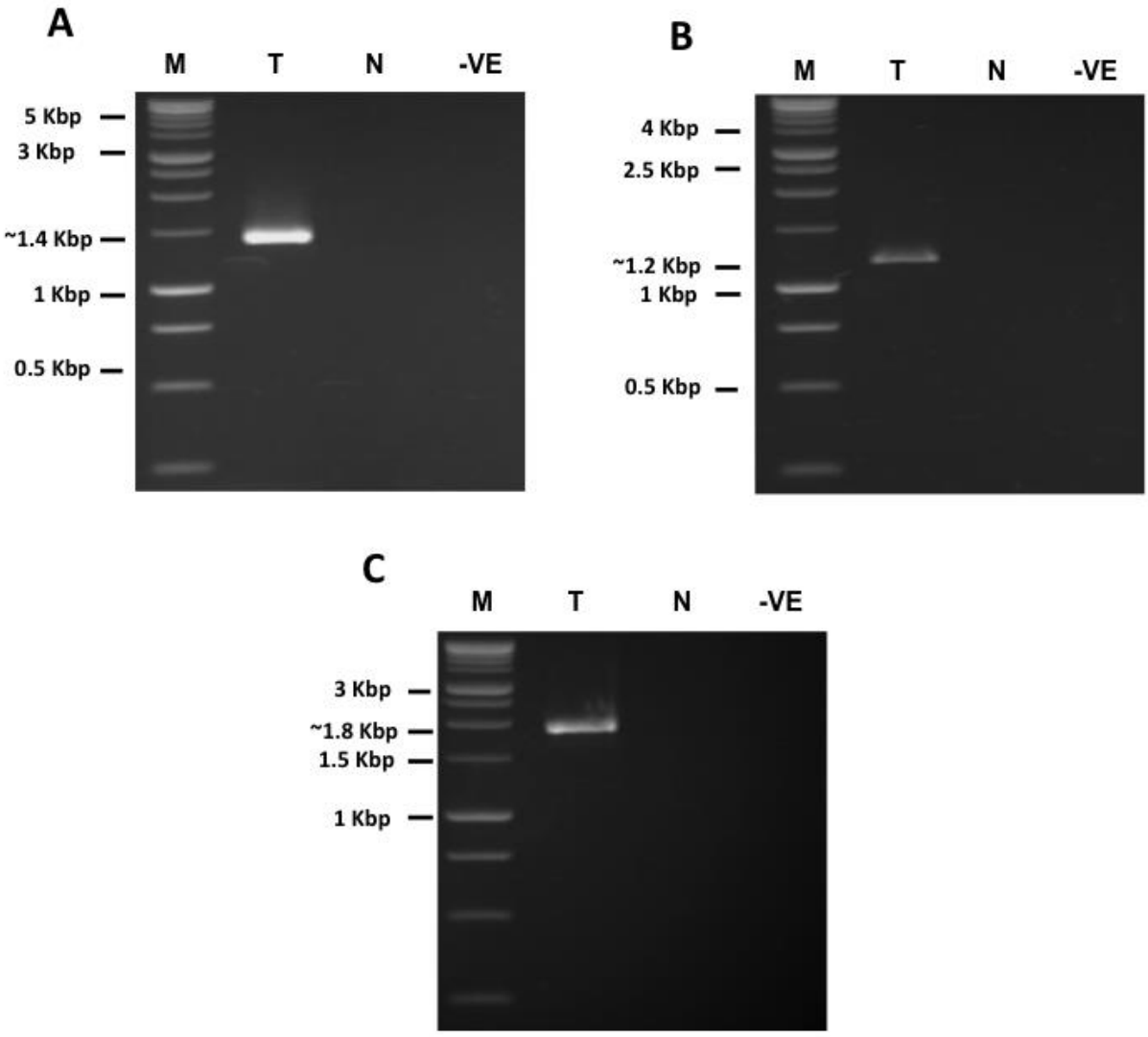


Figure 8–10: LR-PCR confirmation of three gene fusions identified in UPSb tumours. The following gene fusions were confirmed somatic by LR-PCR at the DNA level: (A) *PKNOX2-MMP20*, (B) *ASAP2-ADAM17*, and (C) *APOL1-MYH9*. M: 1 Kbp DNA marker; T: tumour; N: corresponding normal; –VE: negative control (water).

8.5.3 Chapter 6: Adamantinoma and OFD-like adamantinoma study using WES and RNA-Seq

Sample	INDELs	SNVs
ADA-T1	93	574
ADA-T2	76	537
ADA-T3	69	501
ADA-T4	65	497
ADA-T5	56	418
ADA-T6	52	343
ADA-T7*	2362	23624
ADA-T8*	2297	26020
OFD-like T1	65	323
OFD-like T2	48	359
OFD-like T3	52	292
OFD-like T4*	2345	25175

Table 8–10: The total number of unfiltered WES variants in adamantinoma and OFD-like adamantinoma. *Tumours unpaired with own corresponding normal (tumour only).

Sample	No. of coding mutations/variations	Coding mutation rate	Approximated mutations per Mb
ADA-T1	21	7.00E-07	0.70
ADA-T2	10	3.33E-07	0.33
ADA-T3	9	3.00E-07	0.30
ADA-T4	13	4.33E-07	0.43
ADA-T5	30	1.00E-06	1.00
ADA-T6	0	0.00E+00	0.00
ADA-T7	42	1.40E-06	1.40
ADA-T8	77	2.57E-06	2.57
Mean in ADA's	25.25	8.42E-07	0.84
OFD-like T1	0	0.00E+00	0.00
OFD-like T2	0	0.00E+00	0.00
OFD-like T3	1	3.33E-08	0.03
OFD-like T4	65	2.17E-06	2.17
Mean in OFD's	66	2.20E-06	2.20

Table 8–11: The coding mutation rate of adamantinoma and OFD-like adamantinoma tumours. The coding mutation rate in each sample is calculated by dividing the total number of identified variants by the total length of the exome (30 megabase).

Sample	Gene fusion	Status	Right gene genomic coordinates	Left gene genomic coordinates	No. of fusion junction supporting reads	No. of fusion spanning reads
Multiple ADA & OFD-like tumours	*KANSL1- ARL17A*	Germline	chr7:44171923	chr7:44430293	12	10
ADA-T3	*XXYLT1-ACAP2*	Germline	chr3:194947584	chr3:195047613	4	3
	C15orf57-CBX3	Germline	chr15:40854971	chr7:26241389	3	1
ADA-T4	*EPHB4-MARCH10*	Somatic	chr7:100414813	chr17:60827816	109	16

Table 8–12: Gene fusions identified in adamantinoma and OFD-like adamantinoma tumours by RNA-Seq. The details of the gene fusions identified in ADA: adamantinoma; OFD-like: OFD-like adamantinoma.

8.6 List of primers used in confirmational experiments of NGS data

Gene	Primer	Primer sequence	Expected product size
THADA	F	CACCAGGGATTTGAAGCAGT	250
	R	GCAGCGATAAGACAATGCAA	
AIRE	F	GCTGTTTTGGGAAGGAGGTG	299
	R	TCTTGGATGGGAGAGGCTTG	
ZC3H14	F	ATGGTTCTTCTCGCCCTTCT	250
	R	ACCTCCATACTCATTTTCTCCGA	
SIX2	F	AGACCACTCATCATCCAGCC	325
	R	CTGCGGGTCTTTCAGTACCT	
***DUOX1* (Ex. 6)**	F	AGACCCCATGTTCGACCCC	600
	R	AAGGCTGCATCCGACGTG	
***DUOX1* (Ex. 7)**	F	GAGCTTCTCCAGGGGACAG	498
	R	TGGCGGTTGTCCACAGAT	
***DUOX1* (Ex.8)**	F	CTGGTCCTCATCTCCACACG	295
	R	AGAAATGGCCGGTATCCTGG	
***DUOX1* (Ex. 15)**	F	TTCCCCTCAGCTACCCAGA	300
	R	TAGCTTGCATCTCTCACCCC	
***DUOX1* (Ex. 30)**	F	AGGGTGAGCTTCTGATGGG	250
	R	CAGTGTGAAGGGGTGGTACT	
***DUOX2* (Ex. 6)**	F	TTCCTCAACATCCGCATCCC	651
	R	TTCCCGGTTCCCTCTCTCT	
***DUOX2* (Ex. 7)**	F	TCGGGGCAGAGAGGGAA	270
	R	GTTTCTGGGTGGCGGTTGTC	
***DUOX2* (Ex. 8)**	F	GCCCCACACCATCCAACT	233
	R	CAGGAAGGAGACGGTGATGA	
FOXE1	F	ACTCCCGAGCCTCTGTCT	466
	R	ATGAACTTGTAGATGCCGCC	
NKX2-1	F	TCCAGAACCACCGCTACAAA	474
	R	ATAGCAAGGTGGAGCAGGAC	
SLC16A10	F	AGAGGCTTCAGTGTCGATCC	372
	R	CTCCACCTCCACCTTCTCG	
TPO	F	ACCGTGTATGGCAGCTCC	454
	R	GTACCTGGGAGAGAGAAGCC	
TG	F	CAAATGCTCTCAGGGGACAG	209
	R	TGGCTCTCCTGGTCATATGT	
***SLC5A5* (Ex. 1)**	F	GGACATCGACAGCCCATAGA	280
	R	TGGACACCAGGAGCATGAG	
***SLC5A5* (Ex. 12)**	F	GAGCCTTCATCTTGGGAATGT	381
	R	CCCTTCAAGATCACACCATCC	

Table 8–13: List of the Sanger primers used in the WES study of four CHT families.

Gene	Primer	Primer sequence	Expected product size
TP53	F	CCTGCTTGCCACAGGTCT	296
	R	TGTGATGAGAGGTGGATGGG	
	F	ATGGTTGCATGAAAGGAGG	200
	R	TGTATTTTAGTAGAGACGGGG	
	F	TGTCTCCTTCCTCTTCCTACAG	293
	R	GCCAGACCTAAGAGCAATCAG	
	F	GGCCTCTGATTCCTCACTGA	248
	R	GGAGGTCAAATAAGCAGCAGG	
H3F3A	F	TGCTGGTAGGTAAGTAAGGAGG	390
	R	CCTCCAGGTAAGATTATGGCTTC	
DOT1L	F	GGGGCTTACTTCACCTTGGA	430
	R	TGAGAAACCATCCAGCCGAG	
	F	GTGGGTGAGGTCTGCATGG	396
	R	AGCGTAGACTAGGGGAGACA	
ATRX	F	CCTGGGCATATGTATTATCATCC	464
	R	TCCAACTTTGTTTCCCTCTCT	
MAMDC4	F	CCACATAGCCACCGACTTTG	127
	R	ACTGTTCCGGCTGTGGTC	
COL2A1	F	TGGGTAGCAAGGCATCTACT	250
	R	ACTTGGGTCCTTTGGGTTTG	
SYNE2	F	CGCAGAAAGGGTGTTGTGTT	428
	R	CCTTTATCCCGAGGCTGTGA	
	F	GAGCATGTCTTCTCAGTGATTT	208
	R	GGCATCTCAAGGTCTCCAAT	
ROBO2	F	AAGCCTTTACCTCCTCTCAA	265
	R	TGGCTAATCTACCATGACTACA	
	F	ACAGTAGTCTCGTTACCAGAAAC	395
	R	TCATCATCCACCTTGTCAGC	
	F	TCAGGTCCTTTAGTAGACTGC	225
	R	GAAGGGTTCCAGGGGTAC	
MYO7B	F	TGTTCGACCGGGAGCTGT	296
	R	GAGAGGGAGGAAGGACAGGG	
	F	CGTCCATGCCAACAACAAGG	224
	R	TACTCATCCCTCTGCCCCAT	
	F	TCTCCTCTCTGCTCCCATCA	227
	R	CATTCCCTCAACTCCCGACC	
TET2	F	CCTAATGGTGCTACAGTTTCT	231
	R	CAGCTCAGAGTTAGAGGTCT	
	F	GGCCTGTGATGCTGATGATG	155
	R	GCTTTGTGGTTCCCTGGATG	
MACF1	F	TGGGGTGCTGGTAATGTTCT	335
	R	AAGCTTCCACCTCCTCTTCC	
	F	AGCACAAACTAGAAGGGGCT	164
	R	ACTTACATGGTGCTTTGCGA	

Gene	Primer	Primer sequence	Expected product size
MGA	F	CCAGGGTCTGTGATGGGAAT	235
	R	AGCTTTGTGGTTCAGGAGGA	
	F	GGCGGGCTTTTATTAGTAAGGT	249
	R	TCTTGTTGGAAAGGAGGGAATC	
KDM4B	F	GGGCTGCTTGGATTTGAGAT	249
	R	ACTCTGGGCATGAAGTGGAT	
	F	TTTCCCTCTGCAGATCACGC	250
	R	GAAACGGCTGCTGGACCA	
AMOTL2	F	CAGTATGCCTAGGAGTGGGG	249
	R	AGGTGGACTGAGGGAGCT	
	F	ACGTGGTCACTGTCACTGAT	269
	R	GCTTCTCACAGGCTGCCT	
FAT2	F	AGAAGACCACAGGCAGCAT	205
	R	ACACCACATCCACAGAAAACC	
PLEC	F	GCGAGTGGGAGGAGATCAC	214
	R	GAACGGGAGCGGAAACCA	
C1orf86	F	AGGTCCTTCAGTCATGCCTG	205
	R	AGGCTGGTCTCAAACTCCTG	
COL4A2	F	AATTGAAAACTGGAGGGCGG	154
	R	TCTCCCTTGACTCCTTTGATGT	
MCAM	F	AGCTGCTCACTTGGCTCT	154
	R	TCCACCGCAGACCCCTAG	
ABCC12	F	TAGCATTGTCCCTGTCCACA	194
	R	TTTCCCCTGAGCGTCTTCTT	
	F	CGCTTTTCTTTCCAGGTGCT	206
	R	CAAAAGCCTGCCGTTACCTG	
PROM1	F	AGTTTCCTCTTTTGCTGGAT	150
	R	GTTCAAGTCCTTTCATAATGGG	
DST	F	AAGTGAAGCCAAGCAGTACC	149
	R	TGCGAAAATTCAGGAGGTTCTC	
PFKP	F	ACTCCCACGCTTGTCTGA	147
	R	TGCCCCGATGCTGAAATC	
BAIAP2L1	F	TACTGCTACACTCCCTCCAA	148
	R	GCTGAGTAAGGTCTTGTTGGA	
ABI3BP	F	CCTTTCCTCAGTATCCAGAATCA	145
	R	CAGGTTTGGTCTCAGGTGC	
TTBK1	F	TTTCGACAGCAAAGAGTGGG	223
	R	CAAATGCTGCAGGTCCTCC	
TXNRD1	F	TGGTCTGATGTTCTGATGTT	210
	R	AGTGTTCTTGATGGTATGCT	
	F	CTGCCCCTTCTGCTCATTTT	225
	R	GTCCCAGCTACTCAGGAGAC	
PPP1R42	F	AGTTGATGAAGCTGTGGAAA	204
	R	TCACAGATCCAGATGTCAGA	

Gene	Primer	Primer sequence	Expected product size
PECR	F	GTCTAGAGCAATCAATCAGGCG	178
	R	TGTGTTCAGCAGGGGAAAGA	
BEND4	F	CGTCTCCATCAGCAGCAG	146
	R	GAGTCCTCAAGTGGCCCT	
KRAS	F	GCCTGCTGAAAATGACTGAA	172
	R	ATCAAAGAATGGTCCTGCAC	
TYRO3	F	CCCTCATCCTGCTTCGAAAG	261
	R	CCATTTGTCCTTCCCACGAT	
SCG3	F	AGGAAGGAATCAAGCTGGAGT	281
	R	TCCACATAATCACAACCACACAG	
PRMT10	F	ACACACTTTTCAGATGGGCT	393
	R	CAATAAGGGGTGGCACTCAG	
NPHP3	F	GCTATGAAGGAGGAGATTTTGA	192
	R	ACTGCTGCTATTACCTTTGT	
ERCC3	F	ATGCACTGTCAGAAAACGCT	235
	R	GGTCTTGAGCCACTCCATGA	
NEB	F	TCCCAACACCTATCACTCCA	198
	R	GATTTGGGTTGAAAACTGAGGT	
PDZD2	F	CAGAAACAGCATTCCAGGGG	238
	R	TATGCCCTGTCCATTAGCCA	
CYP1A1	F	GCTTGCCTGTCCTCTATCCT	220
	R	CACCGATACACTTCCGCTTG	
CBX4	F	ATGAACCCATAGACTTGCGC	200
	R	GAAGGGCTTGAACTCGCTC	
PCDH12	F	CATGAAAGTGCAAGGGGA	209
	R	GCATCAGTCACCCTAAAGTT	
B2M	F	GGGCATTCCTGAAGCTGACA	199
	R	GGAGAAGGGAAGTCACGGAG	
MGAT3	F	ACCAATGGCACCTTCGAGTA	191
	R	ATCTCGTCCGCATCGTCAAT	
PAOX	F	ATGAGGAGACCAGGAAGCTG	245
	R	CCTCCCATACCCAAGCAAGA	
POSTN	F	TTCCTGTATCTGACTGCCTG	419
	R	GGAATGCATCATCTACTTTGCT	
CCBE1	F	GCCCTTTGTCCTTCTCATGT	347
	R	GGACCACTCATAAGGCTTACC	
ZNF540	F	GCCTTTATGCTTCGTTCAGTC	231
	R	ATAGGGCTTTTCACCAGTGT	
PTPRT	F	GCAACTTTGTCAACCCCTGT	233
	R	TAAGCCAATGGATGCAGCAC	
	F	TCAGTCCAGGCCTCATCATC	224
	R	CTTGGCCTCCTCCTACCTG	
AMFR	F	GAACCTGCTTAGGAGGGAGG	244
	R	CGAGCTTGCTGGAGGAGTT	

Gene	Primer	Primer sequence	Expected product size
FCGBP	F	GGGTCCCTTGAAAGATTGCA	288
	R	TAGTGACTGTTCTGAGGGCA	
	F	GGAATGTGTGTCCAAGCCAT	395
	R	CTCAGAGGAAGGAAAGTGGC	
TRIO	F	GGGCTTCAGGAGAGAGAGA	298
	R	CTTTTCACTTGCTGTTCAGGA	
RB1	F	TGCATGCGAACTCAGTGTAT	257
	R	ACCATGTGCAATACCTGTCT	
ATM	F	ACATGTGGTTTCTTGCCTTTGT	285
	R	GTGCAAAGAACCATGCCCA	
PHF3	F	AGGACTCGCTGAAGAACATG	246
	R	CCCTAAATTTACTACCCTCCTCA	
	F	GGGTCCTCTTCCATTTCTGC	295
	R	CTGTGGAGACCTCGCTGTAT	
KMT2C	F	GGGCCTCAAACAAGTCAGTC	395
	R	TGGGTGACACAGTGAGACTC	
PRDM10	F	CAGTAGCTTCACCCTCCCTTT	246
	R	CACTTCGCTGCTTCCGTTC	

Table 8–14: Sanger sequencing primers used in the validation of WES data of UPSb samples.

Gene	Primer	Primer sequence	Expected product size
ATP7B	F	TGGCAGAGCAGTGTGGAATA	381
	R	TCTCTCAGGATGGGGAAAGC	
	F	CCCTCCCTTTCACTTCACCC	300
	R	GTGGTGCTCTCTGTGGTTTG	
TNK1	F	GACGCTGTGGGAGATGTTCT	239
	R	GGACTGTGTATCTGGTGAGGT	
	F	CCACAGCCTATCAGTTGCCT	242
	R	TCAAGCCACCTCATTCCCAG	
KMT2D	F	AGAGCTGTCTTGTCATGGATT	236
	R	TCTAGGTCCTTGGCTGCATA	
	F	CCTACATCTCCGCATTCCCC	392
	R	CAGGTGGGGTAGTGTGGAAT	
	F	AGAATATCGGAACAAGCAGCAG	248
	R	GCTGTATTAAGGAAGGGGCC	
CPEB2	F	CTTCAAACCGAGTCTGCACC	320
	R	GCGAGAGGTGCTGCTGAT	
MTUS2	F	TCGTGGTCCTGTATAGAACTCAC	250
	R	ATGACACACAGTTACACAGG	
	F	TTTGGGTTCGTTTTAGCAGA	286
	R	ATGACACACAGTTACACAGG	

Gene	Primer	Primer sequence	Expected product size
COL18A1	F	CCACCTGGTTGCGCTCAA	346
	R	GGAGCCTGCCCTGAAGTC	
	F	CTGCATTCCTGGGCGTGT	208
	R	TCTCGTGCCAGCTTTCATCT	
MYO5C	F	AGGAGCTGGATTATTCTGTCAA	220
	R	GCAAGCTTACATTCCAGTGC	

Table 8–15: Sanger sequencing primers used in the validation of WES data of adamantinoma and OFD-like adamantinoma tumours.

Gene fusion	Primer	Sequence	Product size
BATCH-GRIK1	F	AAACTGCCATTCAATGCACAACG	284
	R	ATTTGCCCCTCTGTTGGTCTGA	
CTSC-RAB38	F	ACTGCTTCAAATGTGGCTGG	205
	R	CACCACTGACTGCTGAAATACA	
CLTC-VMP1	F	GCAAGACTGGGCAAATCAAAGA	156
	R	CTTCTGCCGTTGAGCCTCC	
PKNOX2-MMP20	F	CGATTTGGCCATTTACTCCTGA	186
	R	CTGTCCACATCTCTGCCCC	
FARP1-STK24	F	CCTCATCCTCCTTCCTCAAA	170
	R	ACCCAGATCAAAAGGAACACC	
MICAL3-UFD1L	F	GGTGATGCGGCTGTGAAG	125
	R	CTCATCAGCCACAAACTCCAG	
CMAS-PYROXD1	F	GCTGATGCCTGTTCTACTGC	159
	R	CTTTGCTGCATACCATCCCATC	
OSBPL2-CABLES2	F	AGTAGAAGAGCACATGTCAGGG	130
	R	CATCTTCCAGAAACTCCAGGGA	
APOL1-MYH9	F	CTACGGAAAGAAGTGGTGGACA	142
	R	GATCTCGTCAGCCAGCTCATC	
ASAP2-ADAM17	F	CCTGGATGAAAGTGATGACGAC	152
	R	CTTTCTGCGAGAGGGAACAG	
KANSL1-ARL17A*	F	AGACCGGGCAGCTATTGTC	170
	R	AACATCCCAGACAGCGAAGG	
EPHB4-MARCH10*	F	CGTCAGAAAACCGGGCAGAG	195
	R	AGAAAGCAAAGACGGAGCAGTA	

Table 8–16: List of primers used in RT-PCR confirmation of gene fusions (RNA-Seq) detected in UPSb, adamantinoma and OFD-like adamantinoma tumours. * gene fusion identified in adamantinoma and OFD-like tumours. # gene fusion detected in an adamantinoma tumour. All other gene fusions are identified in UPSb tumours.

Gene Fusion	Primer name	Primer sequence
CLTC-VMP1	F1	TGGGCAAATCAAAGAAGTAGAAAGA
	R1	GGAGAAGAAACCGGATTAAAGTCTT
	R2	TGTTACATAGGGCTAAGGAGGTAAA
	R3	GAGGGAGAAAAGAGAGATCAGTGAG
	R4	AAAACTGCAAATTTATGTGGCCG
	R5	TTCACACCTGTAATCCTAGCACTTT
	R6	GAAACTCCGTCTCAAAACAAAACAA
	R7	TGCCCTTAAGCTCTGATTAAATTCA
	R8	CCTACTCCCATTACAAGCTGATCTGTTAT
	R9	TGTTTCTTTACATCCAGGATCTTGG
	R10	TGGTGCTCTAGTCTAAATTTCCAAC
FARP1-STK24	F1	GTCCTCATCCTCCTTCCTCAAA
	F2	CACTTTGTGGGATGTTTATGTGAGAAC
	F3	TTGTTAGAGTCTCAAAATAAGTGGTGTGTG
	F4	GATAAGAAAGTCCAGGCAAACAGAAACAG
	F5	CATTCTTCATGGATTCCGTATTTGTGAA
	F6	GAACTTCTGGTCTGCTTACTGTTTTACTTT
	R1	CTGACAGACACCCAGATCAAAAG
	R2	CTTTTGTGAACTCTGTTAGATGGGA
	R3	CCATTTTCAGTGAGTTACCTGCCT
	R4	CATTCTCTCTTGCATGACTTGTAGA
	R5	GGTTATCCCTGGTTTTCTTCTTGAT
	R6	TACTCATAGTTAGATGTCTGGTGGG
	R7	AATTACTCATAGTTAGATGTCTGGTGGGAG
	R8	CTCTTCCACCTCCATTCTACAGAT
MMP20-PKNOX2	F1	CTTTTCTTCCCACTGTAGCACACGTA
	F2	GGTGGTAATTAGGGCTGTTATTCTCATTCT
	F3	GGAACCTAAGCAGCAGACTTGATAAGAATC
	F4	GTCTAATCCGAGCAGATCCCATGATTG
	F5	GCCCTGTTTATAATGCACTTGTTTCTGTC
	F6	CTCTCCAGGTTTGAGGCTAGACAGATAG
	R1	CTTACCATTTAATTCTACAGCAGCATCGAT
	R2	GCTCCCTGATTTATAATATTGTGCTGTCCT
	R3	GGTTTCTAAGATGGTTTCAGGAAGTTTTCT
	R4	AGACCTAGACAATCACTATCTACTTTCTGT
	R5	TATTCTAGGCCCGTGAGTTTCGAGTTTC
	R6	ATTCAGACTGTGTATTTCTTCTTGACTCGG
	R7	AGTCTGTTTTGCCTGATAATAGTATAGCCATTC
	R8	GTTTCCCTCTAATTCAGTCACATTGCTAGG
	R9	GGCCTGCTAGAAAGAGAATAGGAAGGAATA
APOL1-MYH9	F1	CTTGTACTCTTGGAACCTGGGATGG
	R1	CTTCTTCAGCCGGTCGTTGATCAG

ASAP2-ADAM17	F1	AGATTGTCAGGAATTTGGATGGTACTGTG
	F2	GGCAACAAGAATGAAACTCCATCTCAAAA
	F3	CCATTTTAACCAGAAGACTCCCAAAACTTT
	F4	TCTTAGGGTAGGAGTCAGCAAACTATGG
	F5	CACAATTTTACCTGGTTTCACATCTGACTG
	R1	CCCACTTGTGTTTTGATTTGTTTTCATCTG
	R2	GCAGAAATGCCAGAAAATTAAACCATACTTAGG
	R3	CAAAGTGATGGGATTACAGGTGTGAGTC
	R4	CAGCAAGAAGAAACACCAACAGGATAG
	R5	TCACTTGTATGAGAAACAGCGTATTATGC
	R6	GATCACTTGAGGTCAGGAGTTTGAGAG
	R7	CATTTCTGCCATTCTTATGTATGCCCAC
	R8	AGTGTGACCTCAGAAACTACTCCTTGG
EPHB4-MARCH10*	F1	AGACATCTCTCTTGTTTACCTGCTTTTCAC
	F2	GAACTGTATTTCCCTTTTATGCTTCCTT
	F3	CATTGCATTCCAGCCTGAGAAAGTGAG
	F4	CAGTGGCTTTCTAAACCTTGTAATTTTCTG
	F5	GTTTTGAACACACGATTTTGCTTTGTCAC
	F6	TCAGGAGATCACAGTTCATTGCTCAAAG
	R1	ATTTTGCTATTTGTAGTAGAGACGGGGATT
	R2	GGAAGAAGGACATGAAGATAGGACTTTGAC
	R3	CTAAGGTGGAAGGATTGGTTGAGCTC
	R4	CAATGCACTATGATCAGACCATTACATTCC
	R5	CACCATGCTCGGCTAATTTGTGTATTTT
	R6	GCTACAGGAGGTAAGAAATCAAATGAGTCT
	R7	CAAAGACGGAGCAGTAAGGGGTGTATT
	R8	CCAAGTTCTGCCAGCTTTACTTCTTAAATG
	R9	GCCAAATTCTAACCACTCTTACATCTCAAA
	R10	TTCATGATGTGGGATTTCTGAGCTGAC
	R11	AAATCTGTTTGTAGAGGGCATTGATAGTTC
	R12	TAGACAAACTAAAGCATTGGGTGGATCTAA
	R13	GAGATGAGCCCGAGAGACATTAAAGTTTC

Table 8–17: List of primers used in LR-PCR confirmation of gene fusions (RNA-Seq) identified in UPSb and adamantinoma tumours. * gene fusion detected in adamantinoma tumours. All other genes fusions are identified in UPSb.

Chapter 9: Reference list

Abate, F., Zairis, S., Ficarra, E., Acquaviva, A., Wiggins, C. H., Frattini, V., Lasorella, A., Iavarone, A., Inghirami, G. & Rabadan, R. 2014. Pegasus: a comprehensive annotation and prediction tool for detection of driver gene fusions in cancer. *BMC Syst Biol,* 8**,** 97.

Abbas, S., Lugthart, S., Kavelaars, F. G., Schelen, A., Koenders, J. E., Zeilemaker, A., Van Putten, W. J., Rijneveld, A. W., Lowenberg, B. & Valk, P. J. 2010. Acquired mutations in the genes encoding IDH1 and IDH2 both are recurrent aberrations in acute myeloid leukemia: prevalence and prognostic value. *Blood,* 116**,** 2122-6.

Abu-Khudir, R., Larrivee-Vanier, S., Wasserman, J. D. & Deladoey, J. 2017. Disorders of thyroid morphogenesis. *Best Pract Res Clin Endocrinol Metab,* 31**,** 143-159.

Abu-Khudir, R., Paquette, J., Lefort, A., Libert, F., Chanoine, J. P., Vassart, G. & Deladoey, J. 2010. Transcriptome, methylome and genomic variations analysis of ectopic thyroid glands. *PLoS One,* 5**,** e13420.

Acosta, J., Wang, W. & Feldser, D. M. 2018. Off and back-on again: a tumor suppressor's tale. *Oncogene,* 37**,** 3058-3069.

Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet,* Chapter 7**,** Unit7.20.

Agbugui, J. O. & Omokhudu, O. 2015. Posterior urethral valve: an unusual cause of primary male infertility. *J Reprod Infertil,* 16**,** 113-5.

Agnihotri, S., Jalali, S., Wilson, M. R., Danesh, A., Li, M., Klironomos, G., Krieger, J. R., Mansouri, A., Khan, O., Mamatjan, Y., et al. 2016. The genomic landscape of schwannoma. *Nat Genet,* 48**,** 1339-1348.

Ahluwalia, J. K., Hariharan, M., Bargaje, R., Pillai, B. & Brahmachari, V. 2009. Incomplete penetrance and variable expressivity: is there a microRNA connection? *Bioessays,* 31**,** 981-92.

Al Taji, E., Biebermann, H., Limanova, Z., Hnikova, O., Zikmund, J., Dame, C., Gruters, A., Lebl, J. & Krude, H. 2007. Screening for mutations in transcription factors in a Czech cohort of 170 patients with congenital and early-onset hypothyroidism: identification of a novel PAX8 mutation in dominantly inherited early-onset non-autoimmune hypothyroidism. *Eur J Endocrinol,* 156**,** 521-9.

Alfares, A., Aloraini, T., Subaie, L. A., Alissa, A., Qudsi, A. A., Alahmad, A., Mutairi, F. A., Alswaid, A., Alothaim, A., Eyaid, W., et al. 2018. Whole-genome sequencing offers additional but limited clinical utility compared with reanalysis of whole-exome sequencing. *Genet Med*.

Amary, F., Berisha, F., Ye, H., Gupta, M., Gutteridge, A., Baumhoer, D., Gibbons, R., Tirabosco, R., O'donnell, P. & Flanagan, A. M. 2017. H3F3A (Histone 3.3) G34W Immunohistochemistry: A Reliable Marker Defining Benign and Malignant Giant Cell Tumor of Bone. *Am J Surg Pathol,* 41**,** 1059-1068.

Amary, M. F., Bacsi, K., Maggiani, F., Damato, S., Halai, D., Berisha, F., Pollock, R., O'donnell, P., Grigoriadis, A., Diss, T., et al. 2011. IDH1 and IDH2 mutations are frequent events in central chondrosarcoma and central and periosteal chondromas but not in other mesenchymal tumours. *J Pathol,* 224**,** 334-43.

Anis, E., Hawkins, I. K., Ilha, M. R. S., Woldemeskel, M. W., Saliki, J. T. & Wilkes, R. P. 2018. Evaluation of Targeted Next-Generation Sequencing for Detection of Bovine Pathogens in Clinical Samples. *J Clin Microbiol,* 56.

Apolinario, T. A., Paiva, C. L. & Agostinho, L. A. 2017. REVIEW-ARTICLE Intermediate alleles of Huntington's disease HTT gene in different populations worldwide: a systematic review. *Genet Mol Res,* 16.

Atak, Z. K., Gianfelici, V., Hulselmans, G., De Keersmaecker, K., Devasia, A. G., Geerdens, E., Mentens, N., Chiaretti, S., Durinck, K., Uyttebroeck, A., et al. 2013. Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. *PLoS Genet,* 9**,** e1003997.

Austin, C. P., Cutillo, C. M., Lau, L. P. L., Jonker, A. H., Rath, A., Julkowska, D., Thomson, D., Terry, S. F., De Montleau, B., Ardigo, D., et al. 2018. Future of Rare Diseases Research 2017-2027: An IRDiRC Perspective. *Clin Transl Sci,* 11**,** 21-27.

Babiceanu, M., Qin, F., Xie, Z., Jia, Y., Lopez, K., Janus, N., Facemire, L., Kumar, S., Pang, Y., Qi, Y., et al. 2016. Recurrent chimeric fusion RNAs in non-cancer tissues and cells. *Nucleic Acids Res,* 44**,** 2859-72.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A. & Shendure, J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet,* 12**,** 745-55.

Baugh, E. H., Ke, H., Levine, A. J., Bonneau, R. A. & Chan, C. S. 2018. Why are there hotspot mutations in the TP53 gene in human cancers? *Cell Death Differ,* 25**,** 154-160.

Beaulieu, C. L., Majewski, J., Schwartzentruber, J., Samuels, M. E., Fernandez, B. A., Bernier, F. P., Brudno, M., Knoppers, B., Marcadier, J., Dyment, D., et al. 2014. FORGE Canada Consortium: outcomes of a 2-year national rare-disease gene-discovery project. *Am J Hum Genet,* 94**,** 809-17.

Becker, J., Semler, O., Gilissen, C., Li, Y., Bolz, H. J., Giunta, C., Bergmann, C., Rohrbach, M., Koerber, F., Zimmermann, K., et al. 2011. Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta. *Am J Hum Genet,* 88**,** 362-71.

Beckingsale, T. B. & Shaw, C. 2017. Epidemiology of bone and soft-tissue sarcomas. *Orthopaedics and Trauma,* 31**,** 144-150.

Behjati, S., Tarpey, P. S., Presneau, N., Scheipl, S., Pillay, N., Van Loo, P., Wedge, D. C., Cooke, S. L., Gundem, G., Davies, H., et al. 2013. Distinct H3F3A and H3F3B driver mutations define chondroblastoma and giant cell tumor of bone. *Nat Genet,* 45**,** 1479-82.

Belkadi, A., Bolze, A., Itan, Y., Cobat, A., Vincent, Q. B., Antipenko, A., Shang, L., Boisson, B., Casanova, J. L. & Abel, L. 2015. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A,* 112**,** 5473-8.

Berkers, C. R., Maddocks, O. D., Cheung, E. C., Mor, I. & Vousden, K. H. 2013. Metabolic regulation by p53 family members. *Cell Metab,* 18**,** 617-33.

Bertier, G., Hetu, M. & Joly, Y. 2016. Unsolved challenges of clinical whole-exome sequencing: a systematic literature review of end-users' views. *BMC Med Genomics,* 9**,** 52.

Bhavani, N. 2011. Transient congenital hypothyroidism. *Indian J Endocrinol Metab,* 15**,** S117-20.

Biesecker, L. G. & Green, R. C. 2014. Diagnostic clinical genome and exome sequencing. *N Engl J Med,* 370**,** 2418-25.

Bishop, R. 2010. Applications of fluorescence in situ hybridization (FISH) in detecting genetic aberrations of medical significance. *Bioscience Horizons,* 3**,** 85-95.

Bizhanova, A. & Kopp, P. 2010. Genetics and phenomics of Pendred syndrome. *Mol Cell Endocrinol,* 322**,** 83-90.

Bochynska, A., Luscher-Firzlaff, J. & Luscher, B. 2018. Modes of Interaction of KMT2 Histone H3 Lysine 4 Methyltransferase/COMPASS Complexes with Chromatin. *Cells,* 7.

Botstein, D. & Risch, N. 2003. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet,* 33 Suppl**,** 228-37.

Boucher, C. A., Winchester, C. L., Hamilton, G. M., Winter, A. D., Johnson, K. J. & Bailey, M. E. 2000. Structure, mapping and expression of the human gene encoding the homeodomain protein, SIX2. *Gene,* 247**,** 145-51.

Boycott, K. M., Rath, A., Chong, J. X., Hartley, T., Alkuraya, F. S., Baynam, G., Brookes, A. J., Brudno, M., Carracedo, A., Den Dunnen, J. T., et al. 2017. International Cooperation to Enable the Diagnosis of All Rare Genetic Diseases. *Am J Hum Genet,* 100**,** 695-705.

Boycott, K. M., Vanstone, M. R., Bulman, D. E. & Mackenzie, A. E. 2013. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nat Rev Genet,* 14**,** 681-91.

Boyd, K. P., Korf, B. R. & Theos, A. 2009. Neurofibromatosis type 1. *J Am Acad Dermatol,* 61**,** 1-14; quiz 15-6.

Brodbeck, S., Besenbeck, B. & Englert, C. 2004. The transcription factor Six2 activates expression of the Gdnf gene as well as its own promoter. *Mech Dev,* 121**,** 1211-22.

Brust, E. S., Beltrao, C. B., Chammas, M. C., Watanabe, T., Sapienza, M. T. & Marui, S. 2012. Absence of mutations in PAX8, NKX2.5, and TSH receptor genes in patients with thyroid dysgenesis. *Arq Bras Endocrinol Metabol,* 56**,** 173-7.

Burkhart, D. L. & Sage, J. 2008. Cellular mechanisms of tumour suppression by the retinoblastoma gene. *Nat Rev Cancer,* 8**,** 671-82.

Burrell, R. A., Mcclelland, S. E., Endesfelder, D., Groth, P., Weller, M. C., Shaikh, N., Domingo, E., Kanu, N., Dewhurst, S. M., Gronroos, E., et al. 2013. Replication stress links structural and numerical cancer chromosomal instability. *Nature,* 494**,** 492-496.

Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. 2016. Translating RNA sequencing into clinical diagnostics: opportunities and challenges. *Nat Rev Genet,* 17**,** 257-71.

Camp, M. D., Tompkins, R. K., Spanier, S. S., Bridge, J. A. & Bush, C. H. 2008. Best cases from the AFIP: Adamantinoma of the tibia and fibula with cytogenetic analysis. *Radiographics,* 28**,** 1215-20.

Carr, I. M., Bhaskar, S., O'sullivan, J., Aldahmesh, M. A., Shamseldin, H. E., Markham, A. F., Bonthron, D. T., Black, G. & Alkuraya, F. S. 2013. Autozygosity mapping with exome sequence data. *Hum Mutat,* 34**,** 50-6.

Castanet, M., Polak, M., Bonaiti-Pellie, C., Lyonnet, S., Czernichow, P. & Leger, J. 2001. Nineteen years of national screening for congenital hypothyroidism: familial cases with thyroid dysgenesis suggest the involvement of genetic factors. *J Clin Endocrinol Metab,* 86**,** 2009-14.

Cavenee, W. K., Dryja, T. P., Phillips, R. A., Benedict, W. F., Godbout, R., Gallie, B. L., Murphree, A. L., Strong, L. C. & White, R. L. 1983. Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature,* 305**,** 779-84.

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M. & Wilson, J. F. 2018. Runs of homozygosity: windows into population history and trait architecture. *Nat Rev Genet,* 19**,** 220-234.

Chen, S., Fritchie, K., Wei, S., Ali, N., Curless, K., Shen, T., Brini, A. T., Latif, F., Sumathi, V., Siegal, G. P., et al. 2017a. Diagnostic utility of IDH1/2 mutations to distinguish dedifferentiated chondrosarcoma from undifferentiated pleomorphic sarcoma of bone. *Hum Pathol,* 65**,** 239-246.

Chen, Y., Zhang, H. & Zhang, Y. 2017b. Targeting receptor tyrosine kinase EphB4 in cancer therapy. *Semin Cancer Biol*.

Cherella, C. E. & Wassner, A. J. 2017. Congenital hypothyroidism: insights into pathogenesis and treatment. *Int J Pediatr Endocrinol,* 2017**,** 11.

Cho, C. Y., Lee, K. T., Chen, W. C., Wang, C. Y., Chang, Y. S., Huang, H. L., Hsu, H. P., Yen, M. C., Lai, M. Z. & Lai, M. D. 2016. MST3 promotes proliferation and tumorigenicity through the VAV2/Rac1 signal axis in breast cancer. *Oncotarget,* 7**,** 14586-604.

Christopher, D., Fletcher, J. A. & Bridge, P. 2013. WHO classification of tumours of soft tissue and bone. *International agency for research on cancer. 4th edition. Lyon*.

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S. & Getz, G. 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol,* 31**,** 213-9.

Cieslik, M. & Chinnaiyan, A. M. 2018. Cancer transcriptome profiling at the juncture of clinical translation. *Nat Rev Genet,* 19**,** 93-109.

Comet, I., Riising, E. M., Leblanc, B. & Helin, K. 2016. Maintaining cell identity: PRC2-mediated regulation of transcription and cancer. *Nat Rev Cancer,* 16**,** 803-810.

Comoglio, P. M., Trusolino, L. & Boccaccio, C. 2018. Known and novel roles of the MET oncogene in cancer: a coherent approach to targeted therapy. *Nat Rev Cancer,* 18**,** 341-358.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., Mcpherson, A., Szczesniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., et al. 2016. A survey of best practices for RNA-seq data analysis. *Genome Biol,* 17**,** 13.

Cooper, D. N., Krawczak, M., Polychronakos, C., Tyler-Smith, C. & Kehrer-Sawatzki, H. 2013. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. *Hum Genet,* 132**,** 1077-130.

Cotto, K. C., Wagner, A. H., Feng, Y. Y., Kiwala, S., Coffman, A. C., Spies, G., Wollam, A., Spies, N. C., Griffith, O. L. & Griffith, M. 2018. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res,* 46**,** D1068-d1073.

Croce, C. M. 2008. Oncogenes and cancer. *N Engl J Med,* 358**,** 502-11.

Croise, P., Houy, S., Gand, M., Lanoix, J., Calco, V., Toth, P., Brunaud, L., Lomazzi, S., Paramithiotis, E., Chelsky, D., et al. 2016. Cdc42 and Rac1 activity is reduced in human pheochromocytoma and correlates with FARP1 and ARHGEF1 expression. *Endocr Relat Cancer,* 23**,** 281-93.

Cui, C., Shu, W. & Li, P. 2016. Fluorescence In situ Hybridization: Cell-Based Genetic Diagnostic and Research Applications. *Front Cell Dev Biol,* 4**,** 89.

Curtis, C., Lynch, A. G., Dunning, M. J., Spiteri, I., Marioni, J. C., Hadfield, J., Chin, S. F., Brenton, J. D., Tavare, S. & Caldas, C. 2009. The pitfalls of platform comparison: DNA copy number array technologies assessed. *Bmc Genomics,* 10**,** 588.

Cutting, G. R. 2015. Cystic fibrosis genetics: from molecular understanding to clinical application. *Nat Rev Genet,* 16**,** 45-56.

D'incalci, M. & Galmarini, C. M. 2010. A review of trabectedin (ET-743): a unique mechanism of action. *Mol Cancer Ther,* 9**,** 2157-63.

Dai, W., Zheng, H., Cheung, A. K., Tang, C. S., Ko, J. M., Wong, B. W., Leong, M. M., Sham, P. C., Cheung, F., Kwong, D. L., et al. 2016. Whole-exome sequencing identifies MST1R as a genetic susceptibility gene in nasopharyngeal carcinoma. *Proc Natl Acad Sci U S A,* 113**,** 3317-22.

De Filippis, T., Gelmini, G., Paraboschi, E., Vigone, M. C., Di Frenna, M., Marelli, F., Bonomi, M., Cassio, A., Larizza, D., Moro, M., et al. 2017. A frequent oligogenic involvement in congenital hypothyroidism. *Hum Mol Genet,* 26**,** 2507-2514.

Dees, N. D., Zhang, Q., Kandoth, C., Wendl, M. C., Schierding, W., Koboldt, D. C., Mooney, T. B., Callaway, M. B., Dooling, D., Mardis, E. R., et al. 2012. MuSiC: identifying mutational significance in cancer genomes. *Genome Res,* 22**,** 1589-98.

Deladoey, J. 2012. Congenital Hypothyroidism due to Thyroid Dysgenesis: From Epidemiology to Molecular Mechanisms. *A New Look at Hypothyroidism.* InTech.

Deladoey, J., Belanger, N. & Van Vliet, G. 2007. Random variability in congenital hypothyroidism from thyroid dysgenesis over 16 years in Quebec. *J Clin Endocrinol Metab,* 92**,** 3158-61.

Delespaul, L., Lesluyes, T., Perot, G., Brulard, C., Lartigue, L., Baud, J., Lagarde, P., Le Guellec, S., Neuville, A., Terrier, P., et al. 2017. Recurrent TRIO Fusion in Nontranslocation-Related Sarcomas. *Clin Cancer Res,* 23**,** 857-867.

Dey, K. K., Hsiao, C. J. & Stephens, M. 2017. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genet,* 13**,** e1006599.

Dietrich, D., Uhl, B., Sailer, V., Holmes, E. E., Jung, M., Meller, S. & Kristiansen, G. 2013. Improved PCR performance using template DNA from formalin-fixed and paraffin-embedded tissues by overcoming PCR inhibition. *PLoS One,* 8**,** e77771.

Dilliott, A. A., Farhan, S. M. K., Ghani, M., Sato, C., Liang, E., Zhang, M., Mcintyre, A. D., Cao, H., Racacho, L., Robinson, J. F., et al. 2018. Targeted Next-generation Sequencing and Bioinformatics Pipeline to Evaluate Genetic Determinants of Constitutional Disease. *J Vis Exp*.

Dimitropoulos, A., Molinari, L., Etter, K., Torresani, T., Lang-Muritano, M., Jenni, O. G., Largo, R. H. & Latal, B. 2009. Children with congenital hypothyroidism: long-term intellectual outcome after early high-dose treatment. *Pediatr Res,* 65**,** 242-8.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics,* 29**,** 15-21.

Dolled-Filhart, M. P., Lordemann, A., Dahl, W., Haraksingh, R. R., Ou-Yang, C. W. & Lin, J. C. 2012. Personalizing rare disease research: how genomics is revolutionizing the diagnosis and treatment of rare disease. *Per Med,* 9**,** 805-819.

Doostparast Torshizi, A. & Wang, K. 2018. Next-generation sequencing in drug development: target identification and genetically stratified clinical trials. *Drug Discov Today*.

Dorighi, K. M., Swigut, T., Henriques, T., Bhanu, N. V., Scruggs, B. S., Nady, N., Still, C. D., 2nd, Garcia, B. A., Adelman, K. & Wysocka, J. 2017. Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. *Mol Cell,* 66**,** 568-576.e4.

Doyle, L. A. 2014. Sarcoma classification: an update based on the 2013 World Health Organization Classification of Tumors of Soft Tissue and Bone. *Cancer,* 120**,** 1763-74.

Eckstein, O. S., Wang, L., Punia, J. N., Kornblau, S. M., Andreeff, M., Wheeler, D. A., Goodell, M. A. & Rau, R. E. 2016. Mixed-phenotype acute leukemia (MPAL) exhibits frequent mutations in DNMT3A and activated signaling genes. *Exp Hematol,* 44**,** 740-4.

Esmo/European Sarcoma Network Working Group 2014. Bone sarcomas: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Annals of Oncology,* 25**,** iii113-iii123.

Evola, F. R., Costarella, L., Pavone, V., Caff, G., Cannavo, L., Sessa, A., Avondo, S. & Sessa, G. 2017. Biomarkers of Osteosarcoma, Chondrosarcoma, and Ewing Sarcoma. *Front Pharmacol,* 8**,** 150.

Fagman, H. & Nilsson, M. 2011. Morphogenetics of early thyroid development. *J Mol Endocrinol,* 46**,** R33-42.

Ferguson, B. D., Tan, Y. H., Kanteti, R. S., Liu, R., Gayed, M. J., Vokes, E. E., Ferguson, M. K., Iafrate, A. J., Gill, P. S. & Salgia, R. 2015. Novel EPHB4 Receptor Tyrosine Kinase Mutations and Kinomic Pathway Analysis in Lung Cancer. *Sci Rep,* 5**,** 10641.

Ferguson-Smith, M. A. 2015. History and evolution of cytogenetics. *Mol Cytogenet,* 8**,** 19.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D. & Bray, F. 2015. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer,* 136**,** E359-86.

Fisher, D. A., Schoen, E. J., La Franchi, S., Mandel, S. H., Nelson, J. C., Carlton, E. I. & Goshi, J. H. 2000. The hypothalamic-pituitary-thyroid negative feedback control axis in children with treated congenital hypothyroidism. *J Clin Endocrinol Metab,* 85**,** 2722-7.

Forbes, S. A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C. G., Ward, S., Dawson, E., Ponting, L., et al. 2017. COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res,* 45**,** D777-d783.

Ford, D. J. & Dingwall, A. K. 2015. The cancer COMPASS: navigating the functions of MLL complexes in cancer. *Cancer Genet,* 208**,** 178-91.

Froimchuk, E., Jang, Y. & Ge, K. 2017. Histone H3 lysine 4 methyltransferase KMT2D. *Gene,* 627**,** 337-342.

Fuller, S. J., Mcguffin, L. J., Marshall, A. K., Giraldo, A., Pikkarainen, S., Clerk, A. & Sugden, P. H. 2012. A novel non-canonical mechanism of regulation of MST3 (mammalian Sterile20-related kinase 3). *Biochem J,* 442**,** 595-610.

Fung, Y. K., Murphree, A. L., T'ang, A., Qian, J., Hinrichs, S. H. & Benedict, W. F. 1987. Structural evidence for the authenticity of the human retinoblastoma gene. *Science,* 236**,** 1657-61.

Garcia-Sanchez, A. & Marques-Garcia, F. 2016. Review of Methods to Study Gene Expression Regulation Applied to Asthma. *Methods Mol Biol,* 1434**,** 71-89.

Garofalo, A., Sholl, L., Reardon, B., Taylor-Weiner, A., Amin-Mansour, A., Miao, D., Liu, D., Oliver, N., Macconaill, L., Ducar, M., et al. 2016. The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. *Genome Med,* 8**,** 79.

Gazzo, A., Raimondi, D., Daneels, D., Moreau, Y., Smits, G., Van Dooren, S. & Lenaerts, T. 2017. Understanding mutational effects in digenic diseases. *Nucleic Acids Res,* 45**,** e140.

Gerrand, C., Athanasou, N., Brennan, B., Grimer, R., Judson, I., Morland, B., Peake, D., Seddon, B., Whelan, J. & British Sarcoma, G. 2016. UK guidelines for the management of bone sarcomas. *Clin Sarcoma Res,* 6**,** 7.

Giacomini, C. P., Sun, S., Varma, S., Shain, A. H., Giacomini, M. M., Balagtas, J., Sweeney, R. T., Lai, E., Del Vecchio, C. A., Forster, A. D., et al. 2013. Breakpoint analysis of transcriptional and genomic profiles uncovers novel gene fusions spanning multiple human cancer types. *PLoS Genet,* 9**,** e1003464.

Gianferante, D. M., Mirabello, L. & Savage, S. A. 2017. Germline and somatic genetics of osteosarcoma - connecting aetiology, biology and therapy. *Nat Rev Endocrinol,* 13**,** 480-491.

Gibbons, R. J., Wada, T., Fisher, C. A., Malik, N., Mitson, M. J., Steensma, D. P., Fryer, A., Goudie, D. R., Krantz, I. D. & Traeger-Synodinos, J. 2008. Mutations in the chromatin-associated protein ATRX. *Hum Mutat,* 29**,** 796-802.

Gilbert, S. 2000. Developmental biology, 6th edn (Sunderland, Mass., Sinauer Associates).

Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. 2012. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet,* 20**,** 490-7.

Giordano, A. & Lee, W. H. 2006. Preface. *Oncogene,* 25**,** 5189.

Gleason, B. C., Liegl-Atzwanger, B., Kozakewich, H. P., Connolly, S., Gebhardt, M. C., Fletcher, J. A. & Perez-Atayde, A. R. 2008. Osteofibrous dysplasia and adamantinoma in children and adolescents: a clinicopathologic reappraisal. *Am J Surg Pathol,* 32**,** 363-76.

Gong, K. Q., Yallowitz, A. R., Sun, H., Dressler, G. R. & Wellik, D. M. 2007. A Hox-Eya-Pax complex regulates early kidney developmental gene expression. *Mol Cell Biol,* 27**,** 7661-8.

Goodwin, S., Mcpherson, J. D. & Mccombie, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet,* 17**,** 333-51.

Grasberger, H. & Refetoff, S. 2011. Genetic causes of congenital hypothyroidism due to dyshormonogenesis. *Curr Opin Pediatr,* 23**,** 421-8.

Griffiths, G. S., Grundl, M., Leychenko, A., Reiter, S., Young-Robbins, S. S., Sulzmaier, F. J., Caliva, M. J., Ramos, J. W. & Matter, M. L. 2011. Bit-1 mediates integrin-dependent cell survival through activation of the NFkappaB pathway. *J Biol Chem,* 286**,** 14713-23.

Gronwald, J., Byrski, T., Huzarski, T., Oszurek, O., Janicka, A., Szymanska-Pasternak, J., Gorski, B., Menkiszak, J., Rzepka-Gorska, I. & Lubinski, J. 2008. Hereditary breast and ovarian cancer. *Hered Cancer Clin Pract,* 6**,** 88-98.

Grosso, A. R., Leite, A. P., Carvalho, S., Matos, M. R., Martins, F. B., Vitor, A. C., Desterro, J. M., Carmo-Fonseca, M. & De Almeida, S. F. 2015. Pervasive transcription read-through promotes aberrant expression of oncogenes and RNA chimeras in renal carcinoma. *Elife,* 4.

Haas, B., Dobin, A., Stransky, N., Li, B., Yang, X., Tickle, T., Bankapur, A., Ganote, C., Doak, T., Pochet, N., et al. 2017. STAR-Fusion: Fast and Accurate Fusion Transcript Detection from RNA-Seq. *bioRxiv*.

Hakim, D. N., Pelly, T., Kulendran, M. & Caris, J. A. 2015. Benign tumours of the bone: A review. *J Bone Oncol,* 4**,** 37-41.

Hall, S. K., Hutchesson, A. C. & Kirk, J. M. 1999. Congenital hypothyroidism, seasonality and consanguinity in the West Midlands, England. *Acta Paediatr,* 88**,** 212-5.

Hamamy, H. 2012. Consanguineous marriages : Preconception consultation in primary health care settings. *J Community Genet,* 3**,** 185-92.

Han, Y., Gao, S., Muegge, K., Zhang, W. & Zhou, B. 2015. Advanced Applications of RNA Sequencing and Challenges. *Bioinform Biol Insights,* 9**,** 29-46.

Hanahan, D. & Weinberg, R. A. 2011. Hallmarks of cancer: the next generation. *Cell,* 144**,** 646-74.

Hangauer, M. J., Vaughn, I. W. & Mcmanus, M. T. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet,* 9**,** e1003569.
Hardwick, S. A., Deveson, I. W. & Mercer, T. R. 2017. Reference standards for next-generation sequencing. *Nat Rev Genet,* 18**,** 473-484.
Hassanpour, S. H. & Dehghani, M. 2017. Review of cancer from perspective of molecular. *Journal of Cancer Research and Practice*.
Hatori, M., Watanabe, M., Hosaka, M., Sasano, H., Narita, M. & Kokubun, S. 2006. A classic adamantinoma arising from osteofibrous dysplasia-like adamantinoma in the lower leg: a case report and review of the literature. *Tohoku J Exp Med,* 209**,** 53-9.
He, G., Tavella, S., Hanley, K. P., Self, M., Oliver, G., Grifone, R., Hanley, N., Ward, C. & Bobola, N. 2010. Inactivation of Six2 in mouse identifies a novel genetic mechanism controlling development and growth of the cranial base. *Dev Biol,* 344**,** 720-30.
He, X., Kuo, Y. C., Rosche, T. J. & Zhang, X. 2013. Structural basis for autoinhibition of the guanine nucleotide exchange factor FARP2. *Structure,* 21**,** 355-64.
Hejmadi, M. 2009. *Introduction to cancer biology*, Bookboon.
Hermanns, P., Grasberger, H., Refetoff, S. & Pohlenz, J. 2011. Mutations in the NKX2.5 gene and the PAX8 promoter in a girl with thyroid dysgenesis. *J Clin Endocrinol Metab,* 96**,** E977-81.
Hillman, R. T., Celestino, J., Terranova, C., Beird, H. C., Gumbs, C., Little, L., Nguyen, T., Thornton, R., Tippen, S., Zhang, J., et al. 2018. KMT2D/MLL2 inactivation is associated with recurrence in adult-type granulosa cell tumors of the ovary. *Nat Commun,* 9**,** 2496.
Hofree, M., Carter, H., Kreisberg, J. F., Bandyopadhyay, S., Mischel, P. S., Friend, S. & Ideker, T. 2016. Challenges in identifying cancer genes by analysis of exome sequencing data. *Nat Commun,* 7**,** 12096.
Hofvander, J., Tayebwa, J., Nilsson, J., Magnusson, L., Brosjo, O., Larsson, O., Von Steyern, F. V., Domanski, H. A., Mandahl, N. & Mertens, F. 2015. RNA sequencing of sarcomas with simple karyotypes: identification and enrichment of fusion transcripts. *Lab Invest,* 95**,** 603-9.
Hoischen, A., Van Bon, B. W., Rodriguez-Santiago, B., Gilissen, C., Vissers, L. E., De Vries, P., Janssen, I., Van Lier, B., Hastings, R., Smithson, S. F., et al. 2011. De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. *Nat Genet,* 43**,** 729-31.
Hu, F., Tao, Z., Shen, Z., Wang, X. & Hua, F. 2014. Down-regulation of EphB4 phosphorylation is necessary for esophageal squamous cell carcinoma tumorigenecity. *Tumour Biol,* 35**,** 7225-32.
Huether, R., Dong, L., Chen, X., Wu, G., Parker, M., Wei, L., Ma, J., Edmonson, M. N., Hedlund, E. K., Rusch, M. C., et al. 2014. The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. *Nat Commun,* 5**,** 3630.
Hui, J. Y. 2016. Epidemiology and Etiology of Sarcomas. *Surg Clin North Am,* 96**,** 901-14.
Hwang, D. Y., Dworschak, G. C., Kohl, S., Saisawat, P., Vivante, A., Hilger, A. C., Reutter, H. M., Soliman, N. A., Bogdanovic, R., Kehinde, E. O., et al. 2014. Mutations in 12 known dominant disease-causing genes clarify many congenital anomalies of the kidney and urinary tract. *Kidney Int,* 85**,** 1429-33.
Idikio, H. A. 2011. Human cancer classification: a systems biology- based model integrating morphology, cancer stem cells, proteomics, and genomics. *J Cancer,* 2**,** 107-15.
Ikemura, M., Shibahara, J., Mukasa, A., Takayanagi, S., Aihara, K., Saito, N., Aburatani, H. & Fukayama, M. 2016. Utility of ATRX immunohistochemistry in diagnosis of adult diffuse gliomas. *Histopathology,* 69**,** 260-7.
Interiano, R. B., Malkan, A. D., Loh, A. H., Hinkle, N., Wahid, F. N., Bahrami, A., Mao, S., Wu, J., Bishop, M. W., Neel, M. D., et al. 2016. Initial diagnostic management of pediatric bone tumors. *J Pediatr Surg,* 51**,** 981-5.
Iyengar, P. V., Hirota, T., Hirose, S. & Nakamura, N. 2011. Membrane-associated RING-CH 10 (MARCH10 protein) is a microtubule-associated E3 ubiquitin ligase of the spermatid flagella. *J Biol Chem,* 286**,** 39082-90.

Jain, D., Jain, V. K., Vasishta, R. K., Ranjan, P. & Kumar, Y. 2008. Adamantinoma: a clinicopathological review and update. *Diagn Pathol,* 3**,** 8.

Jan, Y., Matter, M., Pai, J. T., Chen, Y. L., Pilch, J., Komatsu, M., Ong, E., Fukuda, M. & Ruoslahti, E. 2004. A mitochondrial protein, Bit1, mediates apoptosis regulated by integrins and Groucho/TLE corepressors. *Cell,* 116**,** 751-62.

Jia, Y., Xie, Z. & Li, H. 2016. Intergenically Spliced Chimeric RNAs in Cancer. *Trends Cancer,* 2**,** 475-484.

Jiang, L., Huang, J., Higgs, B. W., Hu, Z., Xiao, Z., Yao, X., Conley, S., Zhong, H., Liu, Z., Brohawn, P., et al. 2016. Genomic Landscape Survey Identifies SRSF1 as a Key Oncodriver in Small Cell Lung Cancer. *PLoS Genet,* 12**,** e1005895.

Jiang, Z., Zhou, X., Li, R., Michal, J. J., Zhang, S., Dodson, M. V., Zhang, Z. & Harland, R. M. 2015. Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. *Cell Mol Life Sci,* 72**,** 3425-39.

Joseph, C. G., Hwang, H., Jiao, Y., Wood, L. D., Kinde, I., Wu, J., Mandahl, N., Luo, J., Hruban, R. H., Diaz, L. A., Jr., et al. 2014. Exomic analysis of myxoid liposarcomas, synovial sarcomas, and osteosarcomas. *Genes Chromosomes Cancer,* 53**,** 15-24.

Kadoch, C. & Crabtree, G. R. 2015. Mammalian SWI/SNF chromatin remodeling complexes and cancer: Mechanistic insights gained from human genomics. *Sci Adv,* 1**,** e1500447.

Kaenel, P., Mosimann, M. & Andres, A. C. 2012. The multifaceted roles of Eph/ephrin signaling in breast cancer. *Cell Adh Migr,* 6**,** 138-47.

Kahn, L. B. 2003. Adamantinoma, osteofibrous dysplasia and differentiated adamantinoma. *Skeletal Radiol,* 32**,** 245-58.

Kaneko, S., Li, G., Son, J., Xu, C. F., Margueron, R., Neubert, T. A. & Reinberg, D. 2010. Phosphorylation of the PRC2 component Ezh2 is cell cycle-regulated and up-regulates its binding to ncRNA. *Genes Dev,* 24**,** 2615-20.

Kang, H., Tan, M., Bishop, J. A., Jones, S., Sausen, M., Ha, P. K. & Agrawal, N. 2017. Whole-Exome Sequencing of Salivary Gland Mucoepidermoid Carcinoma. *Clin Cancer Res,* 23**,** 283-288.

Kawaguchi, K., Oda, Y., Sakamoto, A., Saito, T., Tamiya, S., Iwamoto, Y. & Tsuneyoshi, M. 2002. Molecular analysis of p53, MDM2, and H-ras genes in osteosarcoma and malignant fibrous histiocytoma of bone in patients older than 40 years. *Mod Pathol,* 15**,** 878-88.

Keeney, G. L., Unni, K. K., Beabout, J. W. & Pritchard, D. J. 1989. Adamantinoma of long bones. A clinicopathologic study of 85 cases. *Cancer,* 64**,** 730-7.

Kervarrec, T., Collin, C., Larousserie, F., Bouvier, C., Aubert, S., Gomez-Brouchet, A., Marie, B., Miquelestorena-Standley, E., Le Nail, L. R., Avril, P., et al. 2017. H3F3 mutation status of giant cell tumors of the bone, chondroblastomas and their mimics: a combined high resolution melting and pyrosequencing approach. *Mod Pathol,* 30**,** 393-406.

Khan, A., Khan, I., Suleman, S., Zahid, K. & Nabi, G. 2015. A Comprehensive Review on Various Aspects of Genetic Disorders. *Journal of Biology and Life Science,* 6**,** 110-118.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. & Salzberg, S. L. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol,* 14**,** R36.

Kim, D. & Salzberg, S. L. 2011. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol,* 12**,** R72.

Kinsella, M., Harismendy, O., Nakano, M., Frazer, K. A. & Bafna, V. 2011. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. *Bioinformatics,* 27**,** 1068-75.

Kircher, M. & Kelso, J. 2010. High-throughput DNA sequencing--concepts and limitations. *Bioessays,* 32**,** 524-36.

Knudson, A. G. 2001. Two genetic hits (more or less) to cancer. *Nat Rev Cancer,* 1**,** 157-62.

Knudson, A. G., Jr. 1971. Mutation and cancer: statistical study of retinoblastoma. *Proc Natl Acad Sci U S A,* 68**,** 820-3.

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., Mclellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L. & Wilson, R. K. 2012. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res,* 22**,** 568-76.

Koelsche, C., Renner, M., Johann, P., Leiss, I., Sahm, F., Schimmack, S., Wardelmann, E., Renker, E. K., Schirmacher, P., Korshunov, A., et al. 2016. Differential nuclear ATRX expression in sarcomas. *Histopathology,* 68**,** 738-45.

Kopp, P. 2014. Mutations in the Pendred Syndrome (PDS/SLC26A) gene: an increasingly complex phenotypic spectrum from goiter to thyroid hypoplasia. *J Clin Endocrinol Metab,* 99**,** 67-9.

Koren, S. & Bentires-Alj, M. 2017. Tackling Resistance to PI3K Inhibition by Targeting the Epigenome. *Cancer Cell,* 31**,** 616-618.

Koschmann, C., Calinescu, A. A., Nunez, F. J., Mackay, A., Fazal-Salom, J., Thomas, D., Mendez, F., Kamran, N., Dzaman, M., Mulpuri, L., et al. 2016. ATRX loss promotes tumor growth and impairs nonhomologous end joining DNA repair in glioma. *Sci Transl Med,* 8**,** 328ra28.

Kreahling, J. M., Foroutan, P., Reed, D., Martinez, G., Razabdouski, T., Bui, M. M., Raghavan, M., Letson, D., Gillies, R. J. & Altiok, S. 2013. Wee1 inhibition by MK-1775 leads to tumor inhibition and enhances efficacy of gemcitabine in human sarcomas. *PLoS One,* 8**,** e57523.

Ku, C. S., Cooper, D. N. & Patrinos, G. P. 2016. The Rise and Rise of Exome Sequencing. *Public Health Genomics,* 19**,** 315-324.

Kuhlenbaumer, G., Hullmann, J. & Appenzeller, S. 2011. Novel genomic techniques open new avenues in the analysis of monogenic disorders. *Hum Mutat,* 32**,** 144-51.

Kumar, A., White, T. A., Mackenzie, A. P., Clegg, N., Lee, C., Dumpit, R. F., Coleman, I., Ng, S. B., Salipante, S. J., Rieder, M. J., et al. 2011. Exome sequencing identifies a spectrum of mutation frequencies in advanced and lethal prostate cancers. *Proc Natl Acad Sci U S A,* 108**,** 17087-92.

Kumar, J., Gordillo, R., Kaskel, F. J., Druschel, C. M. & Woroniecki, R. P. 2009a. Increased prevalence of renal and urinary tract anomalies in children with congenital hypothyroidism. *J Pediatr,* 154**,** 263-6.

Kumar, P., Henikoff, S. & Ng, P. C. 2009b. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc,* 4**,** 1073-81.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol,* 10**,** R25.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature,* 499**,** 214-218.

Lawson, C. D., Fan, C., Mitin, N., Baker, N. M., George, S. D., Graham, D. M., Perou, C. M., Burridge, K., Der, C. J. & Rossman, K. L. 2016. Rho GTPase Transcriptome Analysis Reveals Oncogenic Roles for Rho GTPase-Activating Proteins in Basal-like Breast Cancers. *Cancer Res,* 76**,** 3826-37.

Lee, J., Kim, D. H., Lee, S., Yang, Q. H., Lee, D. K., Lee, S. K., Roeder, R. G. & Lee, J. W. 2009. A tumor suppressive coactivator complex of p53 containing ASC-2 and histone H3-lysine-4 methyltransferase MLL3 or its paralogue MLL4. *Proc Natl Acad Sci U S A,* 106**,** 8513-8.

Lee, J. C., Jeng, Y. M., Su, S. Y., Wu, C. T., Tsai, K. S., Lee, C. H., Lin, C. Y., Carter, J. M., Huang, J. W., Chen, S. H., et al. 2015. Identification of a novel FN1-FGFR1 genetic fusion as a frequent event in phosphaturic mesenchymal tumour. *J Pathol,* 235**,** 539-45.

Leger, J., Marinovic, D., Garel, C., Bonaiti-Pellie, C., Polak, M. & Czernichow, P. 2002. Thyroid developmental anomalies in first degree relatives of children with congenital hypothyroidism. *J Clin Endocrinol Metab,* 87**,** 575-80.

Leijen, S., Van Geel, R. M., Pavlick, A. C., Tibes, R., Rosen, L., Razak, A. R., Lam, R., Demuth, T., Rose, S., Lee, M. A., et al. 2016. Phase I Study Evaluating WEE1 Inhibitor AZD1775

As Monotherapy and in Combination With Gemcitabine, Cisplatin, or Carboplatin in Patients With Advanced Solid Tumors. *J Clin Oncol,* 34**,** 4371-4380.
Leroy, B., Anderson, M. & Soussi, T. 2014. TP53 mutations in human cancer: database reassessment and prospects for the next decade. *Hum Mutat,* 35**,** 672-88.
Leung, G. K. C., Luk, H. M., Tang, V. H. M., Gao, W. W., Mak, C. C. Y., Yu, M. H. C., Wong, W. L., Chu, Y. W. Y., Yang, W. L., Wong, W. H. S., et al. 2018. Integrating Functional Analysis in the Next-Generation Sequencing Diagnostic Pipeline of RASopathies. *Sci Rep,* 8**,** 2421.
Levy, S. E. & Myers, R. M. 2016. Advancements in Next-Generation Sequencing. *Annu Rev Genomics Hum Genet,* 17**,** 95-115.
Li, S., Shen, D., Shao, J., Crowder, R., Liu, W., Prat, A., He, X., Liu, S., Hoog, J., Lu, C., et al. 2013. Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep,* 4**,** 1116-30.
Liu, F., Mischel, P. S. & Cavenee, W. K. 2017. Precision cancer therapy is impacted by oncogene-dependent epigenome remodeling. *NPJ Precis Oncol,* 1**,** 1.
Liu, J., Lee, W., Jiang, Z., Chen, Z., Jhunjhunwala, S., Haverty, P. M., Gnad, F., Guan, Y., Gilbert, H. N., Stinson, J., et al. 2012. Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events. *Genome Res,* 22**,** 2315-27.
Liu, L., Kimball, S., Liu, H., Holowatyj, A. & Yang, Z. Q. 2015. Genetic alterations of histone lysine methyltransferases and their significance in breast cancer. *Oncotarget,* 6**,** 2466-82.
Lopes, P. & Oliveira, J. L. 2013. An innovative portal for rare genetic diseases research: the semantic Diseasecard. *J Biomed Inform,* 46**,** 1108-15.
Macarthur, D. G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J. K., Montgomery, S. B., et al. 2012. A systematic survey of loss-of-function variants in human protein-coding genes. *Science,* 335**,** 823-8.
Magne, F., Ge, B., Larrivee-Vanier, S., Van Vliet, G., Samuels, M. E., Pastinen, T. & Deladoey, J. 2016. Demonstration of Autosomal Monoallelic Expression in Thyroid Tissue Assessed by Whole-Exome and Bulk RNA Sequencing. *Thyroid,* 26**,** 852-9.
Mahdieh, N. & Rabbani, B. 2013. An overview of mutation detection methods in genetic disorders. *Iran J Pediatr,* 23**,** 375-88.
Makinen, N., Aavikko, M., Heikkinen, T., Taipale, M., Taipale, J., Koivisto-Korander, R., Butzow, R. & Vahteristo, P. 2016. Exome Sequencing of Uterine Leiomyosarcomas Identifies Frequent Mutations in TP53, ATRX, and MED12. *PLoS Genet,* 12**,** e1005850.
Martelotto, L. G., De Filippo, M. R., Ng, C. K., Natrajan, R., Fuhrmann, L., Cyrta, J., Piscuoglio, S., Wen, H. C., Lim, R. S., Shen, R., et al. 2015. Genomic landscape of adenoid cystic carcinoma of the breast. *J Pathol,* 237**,** 179-89.
Martincorena, I. & Campbell, P. J. 2015. Somatic mutation in cancer and normal cells. *Science,* 349**,** 1483-9.
Matthew, F., Nicola, D., Jackie, C., Gill, L. & Rob, G. 2013. Bone and soft tissue sarcomas UK incidence and survival: 1996 to 2010. *National Cancer Intelligence Network (NCIN), UK*.
Mavrogenis, A. F. & Ruggieri, P. 2015. Therapeutic approaches for bone sarcomas. *Bone Cancer (Second Edition).* Elsevier.
Mclaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P. & Cunningham, F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics,* 26**,** 2069-70.
Meeus, L., Gilbert, B., Rydlewski, C., Parma, J., Roussie, A. L., Abramowicz, M., Vilain, C., Christophe, D., Costagliola, S. & Vassart, G. 2004. Characterization of a novel loss of function mutation of PAX8 in a familial case of congenital hypothyroidism with in-place, normal-sized thyroid. *J Clin Endocrinol Metab,* 89**,** 4285-91.
Merchant, A. A., Jorapur, A., Mcmanus, A., Liu, R., Krasnoperov, V., Chaudhry, P., Singh, M., Harton, L., Agajanian, M., Kim, M., et al. 2017. EPHB4 is a therapeutic target in AML and promotes leukemia cell survival via AKT. *Blood Adv,* 1**,** 1635-1644.

Mertens, F., Antonescu, C. R. & Mitelman, F. 2016. Gene fusions in soft tissue tumors: Recurrent and overlapping pathogenetic themes. *Genes Chromosomes Cancer,* 55**,** 291-310.

Metzker, M. L. 2010. Sequencing technologies - the next generation. *Nat Rev Genet,* 11**,** 31-46.

Mitson, M., Kelley, L. A., Sternberg, M. J., Higgs, D. R. & Gibbons, R. J. 2011. Functional significance of mutations in the Snf2 domain of ATRX. *Hum Mol Genet,* 20**,** 2603-10.

Moreno, J. C., Klootwijk, W., Van Toor, H., Pinto, G., D'alessandro, M., Leger, A., Goudie, D., Polak, M., Gruters, A. & Visser, T. J. 2008. Mutations in the iodotyrosine deiodinase gene and hypothyroidism. *N Engl J Med,* 358**,** 1811-8.

Nair, J., Jain, P., Chandola, U., Palve, V., Vardhan, N. R., Reddy, R. B., Kekatpure, V. D., Suresh, A., Kuriakose, M. A. & Panda, B. 2015. Gene and miRNA expression changes in squamous cell carcinoma of larynx and hypopharynx. *Genes Cancer,* 6**,** 328-40.

Nakagawa, H., Wardell, C. P., Furuta, M., Taniguchi, H. & Fujimoto, A. 2015. Cancer whole-genome sequencing: present and future. *Oncogene,* 34**,** 5943-50.

Nandakumar, P., Mansouri, A. & Das, S. 2017. The Role of ATRX in Glioma Biology. *Front Oncol,* 7**,** 236.

Narlikar, G. J., Sundaramoorthy, R. & Owen-Hughes, T. 2013. Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. *Cell,* 154**,** 490-503.

Neilson, K. M., Pignoni, F., Yan, B. & Moody, S. A. 2010. Developmental expression patterns of candidate cofactors for vertebrate six family transcription factors. *Dev Dyn,* 239**,** 3446-66.

Neskey, D. M., Osman, A. A., Ow, T. J., Katsonis, P., Mcdonald, T., Hicks, S. C., Hsu, T. K., Pickering, C. R., Ward, A., Patel, A., et al. 2015. Evolutionary Action Score of TP53 Identifies High-Risk Mutations Associated with Decreased Survival and Increased Distant Metastases in Head and Neck Cancer. *Cancer Res,* 75**,** 1527-36.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., et al. 2010. Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet,* 42**,** 30-5.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature,* 461**,** 272-6.

Nicola, J. P., Nazar, M., Serrano-Nascimento, C., Goulart-Silva, F., Sobrero, G., Testa, G., Nunes, M. T., Munoz, L., Miras, M. & Masini-Repiso, A. M. 2011. Iodide transport defect: functional characterization of a novel mutation in the Na+/I- symporter 5'-untranslated region in a patient with congenital hypothyroidism. *J Clin Endocrinol Metab,* 96**,** E1100-7.

Niini, T., Lahti, L., Michelacci, F., Ninomiya, S., Hattinger, C. M., Guled, M., Bohling, T., Picci, P., Serra, M. & Knuutila, S. 2011. Array comparative genomic hybridization reveals frequent alterations of G1/S checkpoint genes in undifferentiated pleomorphic sarcoma of bone. *Genes Chromosomes Cancer,* 50**,** 291-306.

Ohto, H., Kamada, S., Tago, K., Tominaga, S. I., Ozaki, H., Sato, S. & Kawakami, K. 1999. Cooperation of six and eya in activation of their target genes through nuclear translocation of Eya. *Mol Cell Biol,* 19**,** 6815-24.

Olesen, S. H., Zhu, J. Y., Martin, M. P. & Schonbrunn, E. 2016. Discovery of Diverse Small-Molecule Inhibitors of Mammalian Sterile20-like Kinase 3 (MST3). *ChemMedChem,* 11**,** 1137-44.

Oliver, G., Wehr, R., Jenkins, N. A., Copeland, N. G., Cheyette, B. N., Hartenstein, V., Zipursky, S. L. & Gruss, P. 1995. Homeobox genes and connective tissue patterning. *Development,* 121**,** 693-705.

Orgaz, J. L., Herraiz, C. & Sanz-Moreno, V. 2014. Rho GTPases modulate malignant transformation of tumor cells. *Small GTPases,* 5**,** e29019.

Ozsolak, F. & Milos, P. M. 2011. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet,* 12**,** 87-98.

Pabinger, S., Dander, A., Fischer, M., Snajder, R., Sperk, M., Efremova, M., Krabichler, B., Speicher, M. R., Zschocke, J. & Trajanoski, Z. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform,* 15**,** 256-78.

Park, J. S., Ma, W., O'brien, L. L., Chung, E., Guo, J. J., Cheng, J. G., Valerius, M. T., Mcmahon, J. A., Wong, W. H. & Mcmahon, A. P. 2012. Six2 and Wnt regulate self-renewal and commitment of nephron progenitors through shared gene regulatory networks. *Dev Cell,* 23**,** 637-51.

Park, S. M. & Chatterjee, V. K. 2005. Genetics of congenital hypothyroidism. *J Med Genet,* 42**,** 379-89.

Pasquale, E. B. 2010. Eph receptors and ephrins in cancer: bidirectional signalling and beyond. *Nat Rev Cancer,* 10**,** 165-80.

Payne, K., Gavan, S. P., Wright, S. J. & Thompson, A. J. 2018. Cost-effectiveness analyses of genetic and genomic diagnostic tests. *Nat Rev Genet,* 19**,** 235-246.

Perry, R., Heinrichs, C., Bourdoux, P., Khoury, K., Szots, F., Dussault, J. H., Vassart, G. & Van Vliet, G. 2002. Discordance of monozygotic twins for thyroid dysgenesis: implications for screening and for molecular pathophysiology. *J Clin Endocrinol Metab,* 87**,** 4072-7.

Peters, T. L., Kumar, V., Polikepahad, S., Lin, F. Y., Sarabia, S. F., Liang, Y., Wang, W. L., Lazar, A. J., Doddapaneni, H., Chao, H., et al. 2015. BCOR-CCNB3 fusions are frequent in undifferentiated sarcomas of male children. *Mod Pathol,* 28**,** 575-86.

Petersen, B. S., Fredrich, B., Hoeppner, M. P., Ellinghaus, D. & Franke, A. 2017. Opportunities and challenges of whole-genome and -exome sequencing. *BMC Genet,* 18**,** 14.

Pflueger, D., Terry, S., Sboner, A., Habegger, L., Esgueva, R., Lin, P. C., Svensson, M. A., Kitabayashi, N., Moss, B. J., Macdonald, T. Y., et al. 2011. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Res,* 21**,** 56-67.

Pierron, G., Tirode, F., Lucchesi, C., Reynaud, S., Ballet, S., Cohen-Gogo, S., Perrin, V., Coindre, J. M. & Delattre, O. 2012. A new subtype of bone sarcoma defined by BCOR-CCNB3 gene fusion. *Nat Genet,* 44**,** 461-6.

Pineda-Alvarez, D. E., Dubourg, C., David, V., Roessler, E. & Muenke, M. 2010. Current recommendations for the molecular evaluation of newly diagnosed holoprosencephaly patients. *Am J Med Genet C Semin Med Genet,* 154c**,** 93-101.

Plant, J. & Cannon, S. 2016. Diagnostic work up and recognition of primary bone tumours: a review. *EFORT Open Rev,* 1**,** 247-253.

Pogue, R. E., Cavalcanti, D. P., Shanker, S., Andrade, R. V., Aguiar, L. R., De Carvalho, J. L. & Costa, F. F. 2018. Rare genetic diseases: update on diagnosis, treatment and online resources. *Drug Discov Today,* 23**,** 187-195.

Pohlenz, J., Dumitrescu, A., Zundel, D., Martine, U., Schonberger, W., Koo, E., Weiss, R. E., Cohen, R. N., Kimura, S. & Refetoff, S. 2002. Partial deficiency of thyroid transcription factor 1 produces predominantly neurological defects in humans and mice. *J Clin Invest,* 109**,** 469-73.

Policeni, B. A., Smoker, W. R. & Reede, D. L. 2012. Anatomy and embryology of the thyroid and parathyroid glands. *Semin Ultrasound CT MR,* 33**,** 104-14.

Pompili, L., Leonetti, C., Biroccio, A. & Salvati, E. 2017. Diagnosis and treatment of ALT tumors: is Trabectedin a new therapeutic option? *J Exp Clin Cancer Res,* 36**,** 189.

Prasad, P., Lennartsson, A. & Ekwall, K. 2015. The roles of SNF2/SWI2 nucleosome remodeling enzymes in blood cell differentiation and leukemia. *Biomed Res Int,* 2015**,** 347571.

Puchner, S. E., Varga, R., Hobusch, G. M., Kasparek, M., Panotopoulos, J., Lang, S., Windhager, R. & Funovics, P. T. 2016. Long-term outcome following treatment of Adamantinoma and Osteofibrous dysplasia of long bones. *Orthop Traumatol Surg Res,* 102**,** 925-932.

Puliti, A., Caridi, G., Ravazzolo, R. & Ghiggeri, G. M. 2007. Teaching molecular genetics: chapter 4-positional cloning of genetic disorders. *Pediatr Nephrol,* 22**,** 2023-9.

Rabbani, B., Tekin, M. & Mahdieh, N. 2014. The promise of whole-exome sequencing in medical genetics. *J Hum Genet,* 59**,** 5-15.

Rao, R. C. & Dou, Y. 2015. Hijacked in cancer: the KMT2 (MLL) family of methyltransferases. *Nat Rev Cancer,* 15**,** 334-46.
Rastogi, M. V. & Lafranchi, S. H. 2010. Congenital hypothyroidism. *Orphanet J Rare Dis,* 5**,** 17.
Ratnakumar, K. & Bernstein, E. 2013. ATRX: the case of a peculiar chromatin remodeler. *Epigenetics,* 8**,** 3-9.
Richette, P., Bardin, T. & Stheneur, C. 2008. Achondroplasia: from genotype to phenotype. *Joint Bone Spine,* 75**,** 125-30.
Riegel, M. 2014. Human molecular cytogenetics: From cells to nucleotides. *Genet Mol Biol,* 37**,** 194-209.
Robinson, D. R., Kalyana-Sundaram, S., Wu, Y. M., Shankar, S., Cao, X., Ateeq, B., Asangani, I. A., Iyer, M., Maher, C. A., Grasso, C. S., et al. 2011a. Functionally recurrent rearrangements of the MAST kinase and Notch gene families in breast cancer. *Nat Med,* 17**,** 1646-51.
Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. 2011b. Integrative genomics viewer. *Nat Biotechnol,* 29**,** 24-6.
Robinson, P. N., Kohler, S., Oellrich, A., Wang, K., Mungall, C. J., Lewis, S. E., Washington, N., Bauer, S., Seelow, D., Krawitz, P., et al. 2014. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res,* 24**,** 340-8.
Romeo, S., Bovee, J. V., Kroon, H. M., Tirabosco, R., Natali, C., Zanatta, L., Sciot, R., Mertens, F., Athanasou, N., Alberghini, M., et al. 2012. Malignant fibrous histiocytoma and fibrosarcoma of bone: a re-assessment in the light of currently employed morphological, immunohistochemical and molecular approaches. *Virchows Arch,* 461**,** 561-70.
Rutkowski, R., Mertens-Walker, I., Lisle, J. E., Herington, A. C. & Stephenson, S. A. 2012. Evidence for a dual function of EphB4 as tumor promoter and suppressor regulated by the absence or presence of the ephrin-B2 ligand. *Int J Cancer,* 131**,** E614-24.
Ryu, D., Joung, J. G., Kim, N. K., Kim, K. T. & Park, W. Y. 2016. Deciphering intratumor heterogeneity using cancer genome analysis. *Hum Genet,* 135**,** 635-42.
Sakoparnig, T., Fried, P. & Beerenwinkel, N. 2015. Identification of constrained cancer driver genes based on mutation timing. *PLoS Comput Biol,* 11**,** e1004027.
Salgia, R., Kulkarni, P. & Gill, P. S. 2018. EphB4: A promising target for upper aerodigestive malignancies. *Biochim Biophys Acta,* 1869**,** 128-137.
Salvucci, O. & Tosato, G. 2012. Essential roles of EphB receptors and EphrinB ligands in endothelial cell function and angiogenesis. *Adv Cancer Res,* 114**,** 21-57.
Samji, T., Hong, S. & Means, R. E. 2014. The Membrane Associated RING-CH Proteins: A Family of E3 Ligases with Diverse Roles through the Cell. *Int Sch Res Notices,* 2014**,** 637295.
Samuel, G. N. & Farsides, B. 2017. The UK's 100,000 Genomes Project: manifesting policymakers' expectations. *New Genet Soc,* 36**,** 336-353.
Sankaran, V. G., Ghazvinian, R., Do, R., Thiru, P., Vergilio, J. A., Beggs, A. H., Sieff, C. A., Orkin, S. H., Nathan, D. G., Lander, E. S., et al. 2012. Exome sequencing identifies GATA1 mutations resulting in Diamond-Blackfan anemia. *J Clin Invest,* 122**,** 2439-43.
Sarhadi, V. K., Lahti, L., Scheinin, I., Ellonen, P., Kettunen, E., Serra, M., Scotlandi, K., Picci, P. & Knuutila, S. 2014. Copy number alterations and neoplasia-specific mutations in MELK, PDCD1LG2, TLN1, and PAX5 at 9p in different neoplasias. *Genes Chromosomes Cancer,* 53**,** 579-88.
Sathirapongsasuti, J. F., Lee, H., Horst, B. A., Brunner, G., Cochran, A. J., Binder, S., Quackenbush, J. & Nelson, S. F. 2011. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics,* 27**,** 2648-54.
Schaffer, A. A. 2013. Digenic inheritance in medical genetics. *J Med Genet,* 50**,** 641-52.

Scholfield, D. W., Sadozai, Z., Ghali, C., Sumathi, V., Douis, H., Gaston, L., Grimer, R. J. & Jeys, L. 2017. Does osteofibrous dysplasia progress to adamantinoma and how should they be treated? *Bone Joint J,* 99-b**,** 409-416.

Schuurs-Hoeijmakers, J. H., Geraghty, M. T., Kamsteeg, E. J., Ben-Salem, S., De Bot, S. T., Nijhof, B., Van De, V., Ii, Van Der Graaf, M., Nobau, A. C., Otte-Holler, I., et al. 2012. Mutations in DDHD2, encoding an intracellular phospholipase A(1), cause a recessive form of complex hereditary spastic paraplegia. *Am J Hum Genet,* 91**,** 1073-81.

Schwartzentruber, J., Korshunov, A., Liu, X. Y., Jones, D. T., Pfaff, E., Jacob, K., Sturm, D., Fontebasso, A. M., Quang, D. A., Tonjes, M., et al. 2012. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature,* 482**,** 226-31.

Self, M., Geng, X. & Oliver, G. 2009. Six2 activity is required for the formation of the mammalian pyloric sphincter. *Dev Biol,* 334**,** 409-17.

Self, M., Lagutin, O. V., Bowling, B., Hendrix, J., Cai, Y., Dressler, G. R. & Oliver, G. 2006. Six2 is required for suppression of nephrogenesis and progenitor renewal in the developing kidney. *Embo j,* 25**,** 5214-28.

Shain, A. H., Garrido, M., Botton, T., Talevich, E., Yeh, I., Sanborn, J. Z., Chung, J., Wang, N. J., Kakavand, H., Mann, G. J., et al. 2015. Exome sequencing of desmoplastic melanoma identifies recurrent NFKBIE promoter mutations and diverse activating mutations in the MAPK pathway. *Nat Genet,* 47**,** 1194-9.

Shanholtz, H. J. 2013. Congenital hypothyroidism. *J Pediatr Nurs,* 28**,** 200-2.

Sheridan, M. B., Wohler, E., Batista, D. A., Applegate, C. & Hoover-Fong, J. 2015. The Use of High-Density SNP Array to Map Homozygosity in Consanguineous Families to Efficiently Identify Candidate Genes: Application to Woodhouse-Sakati Syndrome. *Case Rep Genet,* 2015**,** 169482.

Shlien, A., Raine, K., Fuligni, F., Arnold, R., Nik-Zainal, S., Dronov, S., Mamanova, L., Rosic, A., Ju, Y. S., Cooke, S. L., et al. 2016. Direct Transcriptional Consequences of Somatic Mutation in Breast Cancer. *Cell Rep,* 16**,** 2032-46.

Sibinga Mulder, B. G., Mieog, J. S. D., Farina Sarasqueta, A., Handgraaf, H. J., Vasen, H. F. A., Swijnenburg, R. J., Luelmo, S. a. C., Feshtali, S., Inderson, A., Vahrmeijer, A. L., et al. 2018. Diagnostic value of targeted next-generation sequencing in patients with suspected pancreatic or periampullary cancer. *J Clin Pathol,* 71**,** 246-252.

Sibley, C. R., Blazquez, L. & Ule, J. 2016. Lessons from non-canonical splicing. *Nat Rev Genet,* 17**,** 407-421.

Siller, C. & Lewis, I. 2010. Update and review of the management of bone tumours. *Paediatrics and Child Health,* 20**,** 103-108.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A. & Ponting, C. P. 2014. Sequencing depth and coverage: key considerations in genomic analyses. *Nat Rev Genet,* 15**,** 121-32.

Singh, R. R., Patel, K. P., Routbort, M. J., Aldape, K., Lu, X., Manekia, J., Abraham, R., Reddy, N. G., Barkoh, B. A., Veliyathu, J., et al. 2014. Clinical massively parallel next-generation sequencing analysis of 409 cancer-related genes for mutations and copy number variations in solid tumours. *Br J Cancer,* 111**,** 2014-23.

Siva, N. 2015. UK gears up to decode 100,000 genomes from NHS patients. *Lancet,* 385**,** 103-4.

Smittenaar, C. R., Petersen, K. A., Stewart, K. & Moitt, N. 2016. Cancer incidence and mortality projections in the UK until 2035. *Br J Cancer,* 115**,** 1147-1155.

Solomon, B. J., Mok, T., Kim, D. W., Wu, Y. L., Nakagawa, K., Mekhail, T., Felip, E., Cappuzzo, F., Paolini, J., Usari, T., et al. 2014. First-line crizotinib versus chemotherapy in ALK-positive lung cancer. *N Engl J Med,* 371**,** 2167-77.

Spencer, D. H., Sehn, J. K., Abel, H. J., Watson, M. A., Pfeifer, J. D. & Duncavage, E. J. 2013. Comparison of clinical targeted next-generation sequence data from formalin-fixed and fresh-frozen tissue specimens. *J Mol Diagn,* 15**,** 623-33.

Stachler, M. D., Taylor-Weiner, A., Peng, S., Mckenna, A., Agoston, A. T., Odze, R. D., Davison, J. M., Nason, K. S., Loda, M., Leshchiner, I., et al. 2015. Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nat Genet,* 47**,** 1047-55.

Stratton, M. R., Campbell, P. J. & Futreal, P. A. 2009. The cancer genome. *Nature,* 458**,** 719-24.

Sun, F., Zhang, J. X., Yang, C. Y., Gao, G. Q., Zhu, W. B., Han, B., Zhang, L. L., Wan, Y. Y., Ye, X. P., Ma, Y. R., et al. 2018. The genetic characteristics of congenital hypothyroidism in China by comprehensive screening of 21 candidate genes. *Eur J Endocrinol,* 178**,** 623-633.

Tamborero, D., Gonzalez-Perez, A., Perez-Llamas, C., Deu-Pons, J., Kandoth, C., Reimand, J., Lawrence, M. S., Getz, G., Bader, G. D., Ding, L., et al. 2013. Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci Rep,* 3**,** 2650.

Tamura, R., Yoshihara, K., Yamawaki, K., Suda, K., Ishiguro, T., Adachi, S., Okuda, S., Inoue, I., Verhaak, R. G. & Enomoto, T. 2015. Novel kinase fusion transcripts found in endometrial cancer. *Sci Rep,* 5**,** 18657.

Tan, J., Ong, C. K., Lim, W. K., Ng, C. C., Thike, A. A., Ng, L. M., Rajasegaran, V., Myint, S. S., Nagarajan, S., Thangaraju, S., et al. 2015. Genomic landscapes of breast fibroepithelial tumors. *Nat Genet,* 47**,** 1341-5.

Tang, L., Nogales, E. & Ciferri, C. 2010. Structure and function of SWI/SNF chromatin remodeling complexes and mechanistic implications for transcription. *Prog Biophys Mol Biol,* 102**,** 122-8.

Tarailo-Graovac, M., Shyr, C., Ross, C. J., Horvath, G. A., Salvarinova, R., Ye, X. C., Zhang, L. H., Bhavsar, A. P., Lee, J. J., Drogemoller, B. I., et al. 2016. Exome Sequencing and the Management of Neurometabolic Disorders. *N Engl J Med,* 374**,** 2246-55.

Targovnik, H. M., Citterio, C. E. & Rivolta, C. M. 2011. Thyroglobulin gene mutations in congenital hypothyroidism. *Horm Res Paediatr,* 75**,** 311-21.

Taylor, R. M., Kashima, T. G., Ferguson, D. J., Szuhai, K., Hogendoorn, P. C. & Athanasou, N. A. 2012. Analysis of stromal cells in osteofibrous dysplasia and adamantinoma of long bones. *Mod Pathol,* 25**,** 56-64.

Teeuw, M. E., Loukili, G., Bartels, E. A., Ten Kate, L. P., Cornel, M. C. & Henneman, L. 2014. Consanguineous marriage and reproductive risk: attitudes and understanding of ethnic groups practising consanguinity in Western society. *Eur J Hum Genet,* 22**,** 452-7.

The Cancer Genome Atlas Research Network 2017. Comprehensive and Integrated Genomic Characterization of Adult Soft Tissue Sarcomas. *Cell,* 171**,** 950-965.e28.

Thompson, B. J. & Sahai, E. 2015. MST kinases in development and disease. *J Cell Biol,* 210**,** 871-82.

Tirode, F., Surdez, D., Ma, X., Parker, M., Le Deley, M. C., Bahrami, A., Zhang, Z., Lapouble, E., Grossetete-Lalami, S., Rusch, M., et al. 2014. Genomic landscape of Ewing sarcoma defines an aggressive subtype with co-association of STAG2 and TP53 mutations. *Cancer Discov,* 4**,** 1342-53.

Torre, L. A., Siegel, R. L., Ward, E. M. & Jemal, A. 2016. Global Cancer Incidence and Mortality Rates and Trends--An Update. *Cancer Epidemiol Biomarkers Prev,* 25**,** 16-27.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L. & Pachter, L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc,* 7**,** 562-78.

Tsatsanis, C., Zafiropoulos, A. & Spandidos, D. A. 2007. Oncogenic Kinases in Cancer. *eLS*.

Van Allen, E. M., Wagle, N., Stojanov, P., Perrin, D. L., Cibulskis, K., Marlow, S., Jane-Valbuena, J., Friedrich, D. C., Kryukov, G., Carter, S. L., et al. 2014. Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nat Med,* 20**,** 682-8.

Veeraraghavan, J., Tan, Y., Cao, X. X., Kim, J. A., Wang, X., Chamness, G. C., Maiti, S. N., Cooper, L. J., Edwards, D. P., Contreras, A., et al. 2014. Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers. *Nat Commun,* 5**,** 4577.

Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., Jr. & Kinzler, K. W. 2013. Cancer genome landscapes. *Science,* 339**,** 1546-58.

Wakeham, D. E., Chen, C. Y., Greene, B., Hwang, P. K. & Brodsky, F. M. 2003. Clathrin self-assembly involves coordinated weak interactions favorable for cellular regulation. *Embo j,* 22**,** 4980-90.
Waldrop, M. M. 2016. The chips are down for Moore's law. *Nature,* 530**,** 144-7.
Wan, T. S. 2014. Cancer cytogenetics: methodology revisited. *Ann Lab Med,* 34**,** 413-25.
Wang, Q., Xia, J., Jia, P., Pao, W. & Zhao, Z. 2013. Application of next generation sequencing to human gene fusion detection: computational tools, features and perspectives. *Brief Bioinform,* 14**,** 506-19.
Wang, Q. Y., Hu, B., Liu, H., Tang, L., Zeng, W., Wu, Y. Y., Cheng, Z. P. & Hu, Y. 2016. A genetic analysis of 23 Chinese patients with hemophilia B. *Sci Rep,* 6**,** 25024.
Wang, Z., Gerstein, M. & Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet,* 10**,** 57-63.
Wassner, A. J. 2018. Congenital Hypothyroidism. *Clin Perinatol,* 45**,** 1-18.
Watson, I. R., Takahashi, K., Futreal, P. A. & Chin, L. 2013. Emerging patterns of somatic mutations in cancer. *Nat Rev Genet,* 14**,** 703-18.
Weber, S., Taylor, J. C., Winyard, P., Baker, K. F., Sullivan-Brown, J., Schild, R., Knuppel, T., Zurowska, A. M., Caldas-Alfonso, A., Litwin, M., et al. 2008. SIX2 and BMP4 mutations associate with anomalous kidney development. *J Am Soc Nephrol,* 19**,** 891-903.
Wessel, J., Chu, A. Y., Willems, S. M., Wang, S., Yaghootkar, H., Brody, J. A., Dauriz, M., Hivert, M. F., Raghavan, S., Lipovich, L., et al. 2015. Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nat Commun,* 6**,** 5897.
Whicher, D., Philbin, S. & Aronson, N. 2018. An overview of the impact of rare disease characteristics on research methodology. *Orphanet J Rare Dis,* 13**,** 14.
Whyte, M. P., Greenberg, C. R., Salman, N. J., Bober, M. B., Mcalister, W. H., Wenkert, D., Van Sickle, B. J., Simmons, J. H., Edgar, T. S., Bauer, M. L., et al. 2012. Enzyme-replacement therapy in life-threatening hypophosphatasia. *N Engl J Med,* 366**,** 904-13.
Xu, J., Wong, E. Y., Cheng, C., Li, J., Sharkar, M. T., Xu, C. Y., Chen, B., Sun, J., Jing, D. & Xu, P. X. 2014. Eya1 interacts with Six2 and Myc to regulate expansion of the nephron progenitor pool during nephrogenesis. *Dev Cell,* 31**,** 434-47.
Xu, P. X., Zheng, W., Laclef, C., Maire, P., Maas, R. L., Peters, H. & Xu, X. 2002. Eya1 is required for the morphogenesis of mammalian thymus, parathyroid and thyroid. *Development,* 129**,** 3033-44.
Yadav, V. K., Degregori, J. & De, S. 2016. The landscape of somatic mutations in protein coding genes in apparently benign human tissues carries signatures of relaxed purifying selection. *Nucleic Acids Res,* 44**,** 2075-84.
Yan, H., Parsons, D. W., Jin, G., Mclendon, R., Rasheed, B. A., Yuan, W., Kos, I., Batinic-Haberle, I., Jones, S., Riggins, G. J., et al. 2009. IDH1 and IDH2 mutations in gliomas. *N Engl J Med,* 360**,** 765-73.
Yang, X. 2012. Use of functional genomics to identify candidate genes underlying human genetic association studies of vascular diseases. *Arterioscler Thromb Vasc Biol,* 32**,** 216-22.
Young, A. 2007. Structural insights into the clathrin coat. *Semin Cell Dev Biol,* 18**,** 448-58.
Zapanta, P. & Shokri, T. 2010. Embryology of the Thyroid and Parathyroids, Thyroid Embryology Clinical Correlations.
Zerdoumi, Y., Lanos, R., Raad, S., Flaman, J. M., Bougeard, G., Frebourg, T. & Tournier, I. 2017. Germline TP53 mutations result into a constitutive defect of p53 DNA binding and transcriptional response to DNA damage. *Hum Mol Genet,* 26**,** 2812.
Zhang, Z., Miteva, M. A., Wang, L. & Alexov, E. 2012. Analyzing effects of naturally occurring missense mutations. *Comput Math Methods Med,* 2012**,** 805827.
Zhao, M., Kim, P., Mitra, R., Zhao, J. & Zhao, Z. 2016. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res,* 44**,** D1023-31.
Zhou, J. X., Yang, X., Ning, S., Wang, L., Wang, K., Zhang, Y., Yuan, F., Li, F., Zhuo, D. D., Tang, L., et al. 2017. Identification of KANSARL as the first cancer predisposition fusion

gene specific to the population of European ancestry origin. *Oncotarget,* 8**,** 50594-50607.

Zhu, S. M., Chen, C. M., Jiang, Z. Y., Yuan, B., Ji, M., Wu, F. H. & Jin, J. 2016. MicroRNA-185 inhibits cell proliferation and epithelial-mesenchymal transition in hepatocellular carcinoma by targeting Six2. *Eur Rev Med Pharmacol Sci,* 20**,** 1712-9.