INVESTIGATING AN AXONAL FORM OF CHARCOT-MARIE-TOOTH NEUROPATHY USING COMBINED TRANSCRIPTOMIC AND GENOMIC ANALYSIS

by
Dora Yasar

A thesis submitted for the degree of
Master of Philosophy

Faculty of Medicine and Health
The University of Syndey

Primary Supervisor
Professor Marina L. Kennerson

Auxiliary supervisors
Dr. Anthony Nicholas Cutrupi
Dr. Gonzalo Perez-Siles

2024

Statement of Originality

To the best of my knowledge, the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been duly acknowledged.

Dora Yasar

Date: 24/11/2024

TABLE OF CONTENTS

Statement of originality ii
Acknowledgements vii
Abstract ix
Conference presentations x
List of figures xi
List of tables xiv
Abbreviations xvii
Chapter 1: Literature Review 1
1.1| Inherited peripheral neuropathies 1
1.2| Charcot-Marie-Tooth neuropathy 1
1.3| Classification of CMT 3
1.4| Genetics of CMT Type 2 (CMT2) 3
Chapter 1: Part 2 7
1.5| Positional cloning 7
1.6| Next-generation sequencing and variant identification 11
1.6.1| Whole exome sequencing 12
1.6.2| Whole genome sequencing 12
1.7| Variant Identification and prioritisation 14
1.7.1| Variant identification using reference genomes 14
1.7.2| Linkage analysis and filtering for variant prioritisation 15
1.7.3| Multi-omics strategies for prioritising noncoding variants 16
1.8| Family CMT720 17
1.8.1| Initial genetic evaluation of CMT720 17
1.9| More recent molecular investigations of CMT720 20
1.9.1| Linkage analysis 20
1.9.2| Whole-genome sequencing of CMT720 21
1.10| The strategy for identifying the pathogenic variant in CMT720 23
1.11| Project Rationale 26
1.12| Hypothesis 26
1.13| Overall aim 26
1.13| Specific aims 26
Chapter 2: General Materials and Methods 28
2.1| Ethics statement 28
2.2| Standard cleaning and handling procedures 28

2.3| Common reagents 28
2.4| Common equipment 29
2.5| Primer design and manufacture 30
2.6| Standard PCR procedures 30
2.7| Standard gel electrophoresis 31
2.8| Fibroblast cell culture 32
2.9| Whole genome amplification 33
2.10| RNA extraction 34
2.11| Complementary DNA synthesis 35
2.12| Supplementary material 36
Chapter 3: Fine mapping the suggestive linkage loci in CMT720 37
3.1| Introduction 37
3.2| The strategy for eliminating the false positive linkage loci in CMT720 39
3.3| Materials and methods 40
3.3.1| Selection of microsatellite markers 40
3.3.2| Primer design and optimisation 40
3.3.3| Genotyping microsatellite markers 41
3.3.4| Multipoint linkage analysis 44
3.4| Results 47
3.4.1| Fine mapping excludes/deprioritises 3 candidate linkage peaks on chromosome 15, 16 and 18 47
3.4.2| Two suggestive linkage regions on chromosomes 8 and 16 remain as candidate loci for CMT720 52
3.4.3| Haplotype analysis of chromosome 16p12.3-q13 and chromosome 8 q13.2-q21.3 suggestive linkage regions 57
3.4.3.1| Haplotype analysis of chromosome 8q13.2-q21.3 confirms the previously established flanking markers 57
3.4.3.2| Haplotype analysis of chromosome 16p12.3-q13 excludes a 1.44 Mb region 57
3.5| Discussion 62
3.6| Supplementary material 65
Chapter 4: Transcriptome analysis of CMT720 and validation of candidate genes 73
4.1| Introduction 73
4.2| Utilising transcriptomic analysis for prioritising noncoding variants 74
4.3| The strategy for performing transcriptomic analysis on CMT720 75
4.4| Hypothesis 79
4.5| Aims 79
4.6| Methods 80
4.6.1| Selection and culturing of patient and control fibroblasts 80
4.6.2| RNA sequencing and data analysis 80

4.6.3| Identification of novel splicing isoforms and fusion transcripts 81
4.6.4| Differential gene expression analysis 81
4.6.5| Validation of differentially expressed candidate genes with qRT-PCR 82
4.7| Results 83
4.7.1| Novel splicing variants were not identified within the prioritised suggestive linkage regions on chromosome 8 and 16 83
4.7.2| Intrachromosomal gene fusions have been excluded for CMT720 84
4.7.3| Differential gene expression analysis of CMT720 patient derived fibroblasts 86
4.7.4| qRT-PCR validates dysregulation of 4 candidate genes from chromosome 8 and 16 suggestive linkage regions 92
4.8| Discussion 94
4.9| Supplementary material 98
4.9.1| Quality analysis for RNA-seq alignments 98
Chapter 5: Identification and multi-omics analysis of the genomic variants in CMT720 110
5.1| Using lrWGS for identifying variants with higher accuracy and increasing the coverage of the genome 111
5.2| The gapless T2T reference for improved variant detection and analysing previously inaccessible regions in CMT720 112
5.3| Using the draft human pangenome reference and 1KGP resources for variant filtering 114
5.4| Using epigenomics data for identifying potential pathogenic noncoding variants in functionally relevant CREs 117
5.4.1| Using promoter capture Hi-C to select potential regulatory variants impacting dysregulated positional candidates 118
5.5| Hypothesis 121
5.6| Aims 121
5.7| Methods 122
5.7.1| Long read whole genome sequencing 122
5.7.2| Variant identification using srWGS data 122
5.7.3| Strategy developed for filtering SNV and indel calls using genomic resources and bioinformatic tools 123
5.7.3.1| Genotyping patient and control srWGS data using PanGenie 123
5.7.3.2| Preparing benign variant callsets from PanGenie and 1KGP 124
5.7.3.3| Filtering strategy to identify the variants localising to suggestive linkage regions on chromosome 8 and 16 125
5.7.4| Manual SV filtering 128
5.7.5| Filtering RE calls 128
5.7.6| Variant selection and prioritisation 129
5.7.6.1| Using PCHi-C data and transcriptomic findings to prioritise all noncoding

variants 130
5.7.6.2| Additional criteria for prioritising SVs and REs....130
5.7.7| Bioinformatic variant analysis....131
5.7.8| Sanger sequencing 132
5.8| Results 132
5.8.1 | Querying PCHi-C data identifies PIRs for the promoters of the dysregulated positional candidate genes localising to the suggestive linkage regions 132
5.8.2 | No candidate pathogenic RE is identified in CMT720 WGS alignments....135
5.8.3 | No candidate pathogenic SVs is identified in CMT720 WGS alignments 140
5.8.4 | SNVs and indels 145
5.8.4.1| Variant filtering eliminates >99.7% of SNV and indel calls identified in CMT720 patient srWGS and lrWGS 145
5.8.4.2| SNVs and indels localising to *GDAP1* and the genes unique to T2T were deprioritised by bioinformatic analysis....148
5.8.4.3| PCHi-C data prioritises 2 very rare SNVs as candidate pathogenic variants . . 153
5.8.5 | Bioinformatic analysis of the prioritised candidate SNVs g.31287661G>C and g.31037657G>A 155
5.8.5.1| The g.31287661G>C does not localise to CREs or TFBS, and unlikely to be a functionally impactful noncoding variant....155
5.8.5.2| Hg38g.30650690G>A is a TFBS variant with high potential for pathogenicity 156
5.8.6 | Segregation of the hg38g.30650690G>A variant suggests reduced disease penetrance in family CMT720....160
5.9 | Discussion 163
5.10 | Supplementary Material 168
5.10.1. | Quality analysis for srWGS and lrWGS alignments....173
Chapter 6: Final discussion....185
References....199

ACKNOWLEDGEMENTS

I am immeasurably thankful to my primary supervisor Marina Kennerson and co-supervisors Anthony Cutrupi and Gonzalo Perez-Siles. Thanks to their undying support, I did not only become more knowledgeable but also grew as a person with a more grounded sense of purpose. I cannot imagine a better introduction to the world of research.

I would like to express my endless gratitude to Professor Marina Kennerson for allowing me into her team and letting me participate in this fascinating project. Under her guidance, I was able to observe how innovation is driven by rigorous application of scientific theory and daring to go the lengths for exploring the uncharted territories. I felt very lucky and privileged to be able to work with the breakthrough achievements in the field of genomics that she introduced me to. Thanks to her understanding and admirable patience, I was able to become more confident to undertake things I did not know I could achieve. I will always be grateful for the many opportunities she provided for me.

Dr. Anthony Cutrupi has been a mentor that shares very similar interests in molecular biology. I am thankful for all the enthusiastic discussions we had and the brilliant introduction to transcriptomics he provided. The methods and research I heard about thanks to him will guide my next steps.

I would like to express my gratitude for Dr. Gonzalo Perez-Siles for his help with cell culture and always providing great advice that was grounded in practical logic and critical thinking. His calm and rational attitude was always reassuring at times of challenge throughout this project.

Many thanks to Melina Ellis for always helping me find my way around in the lab without getting sick of it, being a very caring person and her great taste in music.

To Dr. Ramesh Narayanan, I would like to express my deep appreciation for understanding what it means to be an international student and his peerless practical advice.

I would like to thank Dr. Bianca Rose Grosz for the savvy tips on all things genetics, especially the experiments I always managed to get concerned about.

To mom, dad and my grandfather Zeki Gürel. Few would make the sacrifices you did to get me where I am right now. I hope I could make you proud.

To the fellow students of Northcott Neuroscience Lab, who were also supportive friends. Ally, Harrison, Janelle and Maddy, you were nothing but nice and always offered help out of care and kindness. I hope I managed to be the same to you.

Finally, I would like to extend my thanks to all CMT720 members who have participated in this research.

ABSTRACT

Charcot-Marie-Tooth disease (CMT) is an inherited peripheral neuropathy (IPN) that leads to the degeneration of the sensory and motor nerves of the peripheral nervous system. We previously reported a Polish family (CMT720) with an autosomal dominant form of axonal CMT. Five suggestive linkage loci were established in this family and all coding mutations in the suggestive linkage regions were excluded using whole genome sequencing (WGS). In this study, we hypothesised that CMT720 is caused by a noncoding mutation. To address the challenge of selecting and analysing noncoding variants, a multi-omics strategy was devised to determine variants worthwhile for future functional analysis. In this project, fine mapping linkage analysis supported two of the five suggestive linkage loci initially reported. By utilising both long and short read WGS and the telomere-to-telomere reference, a full spectrum of variants ranging from single base changes to structural variants were identified in CMT720 patients. Using the diverse set of benign variants from the draft human pangenome and targeting the suggestive linkage regions on chromosome 8 and 16 provided variant filtering power to identify a manageable number of noncoding variants for analysis. Transcriptome profiling of CMT720 patient fibroblasts identified dysregulated gene expression for four positional candidate genes (*IRX6, ZNF704, BCL7C, PRRT2*). Using publicly available epigenomics data, one of the selected noncoding single nucleotide variants was found to localise within a transcription factor binding site that may potentially cause downregulation of the positional candidate gene *PRRT2*.

Overall, this study has demonstrated the use of multi-omics analysis in combination with improved sequencing technologies and genomics resources as a powerful strategy to effectively interrogate the noncoding genome in unsolved IPNs.

CONFERENCE PRESENTATIONS

Oral presentations

Yasar D, Grosz BR, Ellis M, Cutrupi AN, Perez-Siles G, Record C, Samaha G, Mori G, Chew T, Folland C, Ravenscroft G, Deveson I, Chintaliphani S, Stevanovski I, Kochanski A, Reilly MM, Vucic S, Kennerson ML (2024) Integrated transcriptomic and genomic analysis to investigate an axonal form of Charcot-Marie-Tooth disease. *GeneMappers*, Christchurch, New Zeland, August 19-21 2024

Poster presentations

Yasar D, Grosz BR, Ellis M, Record C, Samaha G, Chew T, Kochanski A, Reilly MM, Vucic S, Kennerson ML (2023) Mapping a new gene for axonal Charcot-Marie-Tooth neuropathy. *The XXIII International Congress of Genetics*, Melbourne, Australia, July 16-21 2023

Yasar D, Grosz BR, Ellis M, Cutrupi AN, Perez-Siles G, Record C, Samaha G, Mori G, Chew T, Folland C, Ravenscroft G, Deveson I, Chintaliphani S, Stevanovski I, Kochanski A, Reilly MM, Vucic S, Kennerson ML (2024) Integrated transcriptomic and genomic analysis to investigate an axonal form of Charcot-Marie-Tooth disease. *GeneMappers*, Christchurch, New Zeland, August 19-21 2024

LIST OF FIGURES

Figure 1.1: Diagram illustrating the degeneration of the axon of a motor neuron and atrophy of the denervated muscle in CMT 2

Figure 1.2: Genes known to cause axonal sensorimotor neuropathy in the PNS and the biological processes associated with their pathogenic mechanisms 6

Figure 1.3: Calculation of LOD scores in linkage analysis 8

Figure 1.4: Sample linkage likelihood curves showing the power to exclude genomic regions and also identify linkage loci 10

Figure 1.5: Genomic variation that can be identified by WGS 14

Figure 1.6: The pedigree of family CMT720 19

Figure 1.7: The pedigree of family CMT720 indicating phenotypic status of individuals for "affected only" analysis 22

Figure 1.8: Workflow of the strategy developed to identify the pathogenic variant and the disrupted gene in CMT720 25

Supplementary Figure 2.1: The pedigree of CMT720 showing the DNA samples from individuals that underwent whole genome amplification 36

Figure 3.1: The pedigree used in the microsatellite-based linkage analysis performed on CMT720 43

Figure 3.2: The sliding loci method to perform 3-point linkage analysis using LINKMAP sub-program from the LINKAGE software package 46

Figure 3.3: The suggestive linkage regions excluded or deprioritised based on the results of the fine mapping linkage analysis 48

Figure 3.4: The suggestive linkage regions validated based on the results of the fine mapping linkage analysis 54

Figure 3.5: Haplotype analysis of microsatellite markers from the candidate linkage loci on chromosome 8q13.2-q21.3 and chromosome 16p12.3-q13 in family CMT720 59

Figure 4.1: The summary of RNA-seq workflow for the transcriptomic analysis of CMT720 77

Figure 4.2: PCA plot of RNA-seq data 86

Figure 4.3: The Venn diagram representing the overlap between all DEGs identified by DESeq2, NOISeq and EdgeR 89

Figure 4.4: The Venn diagram representing the overlap between the DEGs identified by DESeq2, NOISeq and EdgeR across the prioritised suggestive linkage regions 91

Figure 4.5: Validating expression profiles of dysregulated positional candidate genes localising to the suggestive chromosome 8 and 16 linkage regions 93

Figure 5.1: The annotated ideograms of chromosomes 8 and 16 in the T2T reference showing comparisons of coverage and gene density with respect to hg38 across the suggestive linkage regions on chromosome 8 and 16 113

Figure 5.2: Representation of variants in pangenome references and workflow of genome inference 116

Figure 5.3: Regulatory interactions mediated by chromatin organisation and 3D chromatin contacts 118

Figure 5.4: Library preparation and experimental workflow of PCHi-C 120

Figure 5.5: The workflow of variant filtering performed on the VCF files obtained from CMT720 patient srWGS and lrWGS T2T alignments 127

Figure 5.6: The summary of the multi-omics analysis that will be used in the current investigation of CMT720 127

Figure 5.7: The Genome Browser screenshots showing the promoter regions of dysregulated positional candidate genes identified with the PCHi-C data 134

Figure 5.8: IGV display of the RE call in IRX6 from patient V:6 visualised in the T2T alignments 139

Figure 5.9: Flowchart diagram showing the results of variant filtering performed on the SVs identified in CMT720 patient IV:4 141

Figure 5.10: Manual SV filtering by visualizing patient, PanGenie and 1KGP calls in IGV 142

Figure 5.11: Variant filtering strategy to prioritise SNVs and indels identified from patient srWGS and lrWGS aligned to the T2T reference 147

Figure 5.12: IGV panel showing the intronic variants identified in the DUSP22 paralog unique to T2T on chromosome 16 in CMT720 patient alignments 151

Figure 5.13: Summary of the bioinformatic analysis performed to predict the functional impact of hg38g.30900198G>C 156

Figure 5.14: Summary of the bioinformatic analysis performed to predict the functional impact of hg38g.30650690G>A 158

Figure 5.15: The pedigree of CMT720 showing the segregation of the candidate TFBS variant hg38g.30650690G>A 161

Figure 5.16: Sanger sequencing chromatograms of the CMT720 family members genotyped for segregation analysis of the noncoding variant hg38g.30650690G>A 162

Supplementary Figure 5.1: The regions showing poor coverage in the PacBio lrWGS T2T alignment of patient IV:4 for the prioritised linkage regions on chromosome 8 and 16 175

LIST OF TABLES

Table 1.1: The genes currently known to cause autosomal dominant CMT2 4

Table 1.2: Suggestive linkage regions identified for CMT720 and the associated LOD scores 21

Table 2.1: Criteria for primers designs 30

Table 2.2: Standard setup for PCR reactions 31

Table 2.3: Standard thermal cycling program used for PCR amplifications 31

Table 2.4: Thermal cycling program used for cDNA synthesis 35

Table 3.1: Thermal cycling program used for optimisation of the primer pairs targeting each microsatellite marker 41

Supplementary Table 3.1: Properties of selected microsatellite markers 65

Supplementary Table 3.2: Sequences of the primers used to amplify the microsatellite markers 66

Supplementary Table 3.3: Optimised PCR conditions for selected microsatellite markers 66

Supplementary Table 3.4: Pedigree information files for suggestive linkage loci 67

Supplementary Table 3.5: Sizes of microsatellite marker alleles 69

Supplementary Table 3.6: Encoded microsatellite marker genotypes 71

Table 4.1: The thermal cycling protocol used for qRT-CPR 83

Table 4.2: List of gene fusions predicted by Defuse in the prioritised suggestive linkage region on chromosomes 8 84

Table 4.3: List of intrachromosomal gene fusions predicted by Arriba in the prioritised suggestive linkage region on chromosome 16 84

Table 4.4: List of gene fusions predicted by FusionCatcher in the prioritised suggestive linkage region on chromosome 16 85

Table 4.5: List of DEG identified by EdgeR localising to chromosome 8 and 16 candidate suggestive linkage regions 87

Table 4.6: List of DEG identified by DESeq2 in chromosome 8 and 16 suggestive linkage regions 87

Table 4.7: List of DEG identified by NOISeq in chromosome 8 and 16 candidate suggestive linkage regions 88

Table 4.8: Gene expression levels of differentially expressed candidate genes in control MNs 92

Supplementary Table 4.1: TaqMan gene expression assay probes 98

Supplementary Table 4.2: Quality control statistics on raw RNA-seq data 98

Supplementary Table 4.3: Quality control statistics on aligned RNA-seq data 99

Supplementary Table 4.4: The script used to perform DGE analysis with DESeq2 99

Supplementary Table 4.5: The script used to perform DGE analysis with edgeR 100

Supplementary Table 4.6: The script used to perform DGE analysis with NOISeq 100

Supplementary Table 4.7: Predictions of abnormal splice isoforms made by StringTie from the prioritised suggestive linkage region on chromosome 8 102

Supplementary Table 4.8: Predictions of abnormal splice isoforms made by StringTie from the prioritised suggestive linkage region on chromosome 16 102

Supplementary Table 4.9: C_T and standard error of C_T values obtained from all batches of qRT-PCR 108

Table 5.1: The number of 3D chromatin contacts and PIRs interacting with the promoters of dysregulated positional candidates identified by accessing neuronal PCHi-C data 133

Table 5.2: The RE calls made by EHDN localising to the suggestive linkage region on chromosome 8 136

Table 5.3: The RE calls made by EHDN localising to the suggestive linkage region on chromosome 16 137

Table 5.4: SV calls prioritised based on proximity to dysregulated candidate genes and subsequently excluded for pathogenic role in CMT720 144

Table 5.5: SNVs and indels identified in GDAP1 and the genes unique to the T2T reference located in the suggestive linkage regions 149

Table 5.6: Prioritised SNVs identified in CMT720 patients 154

Supplementary Table 5.1: Preprocessing of the 1KGP T2T reference VCF file containing SNV and indel calls 168

Supplementary Table 5.2: Preprocessing of the PanGenie VCF files 168

Supplementary Table 5.3: Removing low quality calls 168

Supplementary Table 5.4: Merging the control vcf files 169

Supplementary Table 5.5: Variant filtering pipeline for the SNVs and indels identified in the CMT720 srWGS data aligned against the T2T reference 169

Supplementary Table 5.6: Variant filtering pipeline for the SNVs and indels identified in the CMT720 lrWGS data aligned against the T2T reference 172

Supplementary Table 5.7: Quality control statistics on aligned patient lrWGS data 174

Supplementary Table 5.8: Quality control statistics on aligned patient and control srWGS data 174

Supplementary Table 5.9: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of ZNF704 in DLPFC 177

Supplementary Table 5.10: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of ZNF704 in hippocampus 177

Supplementary Table 5.11: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of PRRT2 in DLPFC 179

Supplementary Table 5.12: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of PRRT2 in hippocampus 180

Supplementary Table 5.13: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of BCL7C in DLPFC 181

Supplementary Table 5.14: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of BCL7C in hippocampus 181

Supplementary Table 5.15: T2T coordinates of the PIRs contacting the promoter region of IRX6 in DLPFC 182

Supplementary Table 5.16: The TF CHiP-seq data accessed during variant analysis 182

Supplementary Table 5.17: The distance between hg38g.30650690G>A and the summits of the CHiP-seq peaks intersected by this candidate variant 183

Supplementary Table 5.18: The canonical TF binding motifs hg38g.30650690G>A is predicted to overlap by Factorbook 183

ABBREVIATIONS

1KGP	1000 Genomes Project
3C	Chromatin conformation capture
3D	3 dimensional
AC	Allele count
AD-CMT2	Autosomal dominant CMT2
AF	Allele frequency
cCRE	Candidate cis-regulatory element
cDNA	Complementary DNA
CHiP	Chromatin immunoprecipitation
CHiP-PCR	Chromatin immunoprecipitation-PCR
CHiP-seq	Chromatin immunoprecipitation sequencing
cM	CentiMorgan
CMAPs	Compound muscle action potentials
CMT	Charcot Marie Tooth neuropathy
CMT1	CMT type 1
CMT2	CMT type 2
CNS	Central nervous system
CRE	Cis-regulatory element
CRISPR	Clustered regularly interspaced short palindromic repeats
Ct	Cycle threshold
CTCF	CCCTC binding factor
DEG	Differentially expressed gene
DGE	Differential gene expression
DGV	Database of Genomic Variants
DHMN	Distal hereditary motor neuropathy
DHMN1	Distal Hereditary Motor Neuropathy Type 1
DHPG	Draft human pangenome
DLPFC	Dorsolateral prefrontal cortex
DMEM	Dulbecco's modified Eagle's medium
DNA	Deoxyribonucleic acid
EDTA	Ethylenediaminetetraacetic acid
FC	Fold change
FDMEM	DMEM supplied with fetal bovine serum
FDR	False discovery rate
Gb	Gigabase
GRCh37/hg19	Genome Reference Consortium Human Build 37/human genome 19
GRCh38/hg38	Genome Reference Consortium Human Build 38/human genome 38
Hi-C	High-resolution chromosome conformation capture
HMN	Hereditary motor neuropathy
HMSN	Hereditary motor and sensory neuropathy

HSAN	Hereditary sensory and autonomic neuropathy
HSN	Hereditary sensory neuropathy
IGV	Integrative Genomics Viewer
Indels	Insertions/deletions
IPN	Inherited peripheral neuropathy
kb	Kilobase
KLF3	Krüppel-like factor 3
LOD	Logarithm of odds
lrWGS	Long read WGS
MAF	Minor allele frequency
Mb	Megabase
miRNA	Micro RNA
MN	Motor neuron
mNCV	Motor nerve conduction velocities
MRFF	Medical Research Future Fund
mRNA	Messenger RNA
NFYA	Specificity protein 1
NFYC	Nuclear transcription factor Y subunit gamma
NGS	Next-generation sequencing
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PC1	Principal component 1
PC2	Principal component 2
PCA	Principal component analysis
PCHi-C	Promoter capture Hi-C
PCR	Polymerase chain reaction
PIR	Promoter interacting region
PKD	Paroxysmal kinesigenic dyskinesia
PNS	Peripheral nervous system
qRT-PCR	Quantitative reverse transcription PCR
RE	Repeat expansion
RNA	Ribonucleic acid
RNA-seq	RNA sequencing
RQ	Relative quantity
SNAPs	Sensory nerve action potentials
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SOX10	SRY-box transcription factor 10
SP1	Specificity protein 1
srWGS	Short read WGS
STR	Short tandem repeat
SV	Structural variant
T2T	Telomere-to-telomere
TAD	Topologically associated domain
TAE	Tris acetate EDTA

TF	Transcription factor
TFBS	Transcription factor binding site
TPM	Transcripts per kilobase million
UTR	Untranslated region
VCF	Variant call format
WES	Whole exome sequencing
WGS	Whole genome sequencing

Chapter 1
Literature review

1.1. Inherited peripheral neuropathies

Inherited peripheral neuropathies (IPN) are a genetically heterogeneous group of diseases causing damage to the nerves of the peripheral nervous system (PNS). IPNs are divided into three groups based on the type of peripheral nerve involvement. Hereditary motor neuropathies (HMN), also known as distal hereditary motor neuropathies (dHMN), affect the motor neurons (MNs) with little to no impact on sensory nerves [1, 2]. Hereditary sensory and autonomic neuropathies (HSAN), also known as hereditary sensory neuropathies (HSN), feature sensory abnormalities with mild autonomic dysfunction and minor motor impairment that may worsen with disease progression [1, 3]. Hereditary motor and sensory neuropathies (HMSN) involve both the motor and sensory nerves, and are also known as Charcot-Marie- Tooth (CMT) neuropathy [4, 5].

1.2. Charcot-Marie-Tooth neuropathy

CMT is named after the neurologists Jean-Martin Charcot, Pierre Marie and Howard Henry Tooth who clinically characterised the disorder in 1886 [6, 7]. The global prevalence of CMT is estimated to reach 1/1200, making this disease the most common IPN [8, 9]. Today more than 1000 causative mutations in over 100 genes are associated with CMT and related disorders, demonstrating the high genetic heterogeneity [10].

The histopathological hallmark of CMT is axonal degeneration or the ‘dying back’ of peripheral nerves from their distal ends (Figure 1) [5, 11]. For CMT, axonal pathology usually progresses slowly over the course of decades in a length-dependent manner, with the longest

peripheral nerves supplying the hands, feet and forelimbs being most vulnerable [5, 12]. Loss of axons in sensory neurons and spinal MNs leads to sensory disturbances [5, 12] and atrophy in the denervated muscles respectively [13, 14]. These symptoms initially affect the feet and forelegs, then emerge in the hands and forearms with disease progression [5, 12]. Classical motor findings such as muscle weakness, gait disturbances, severe muscular atrophy, and pes cavus (high arch) foot deformities result from the chronic denervation of muscles [13, 14], and are often accompanied with the loss of deep tendon reflexes [5, 12]. Patients have substantial sensory abnormalities including insensitivity to pain and temperature and paraesthesia [5, 12]. Due to the impaired regenerative capacity of the PNS in CMT [15, 16] and the absence of curative treatments [17], the disease results in permanent disability with a spectrum of severity ranging from gait disturbances [13, 14] to loss of walking [18] and ambulation [19] in the most severe cases.

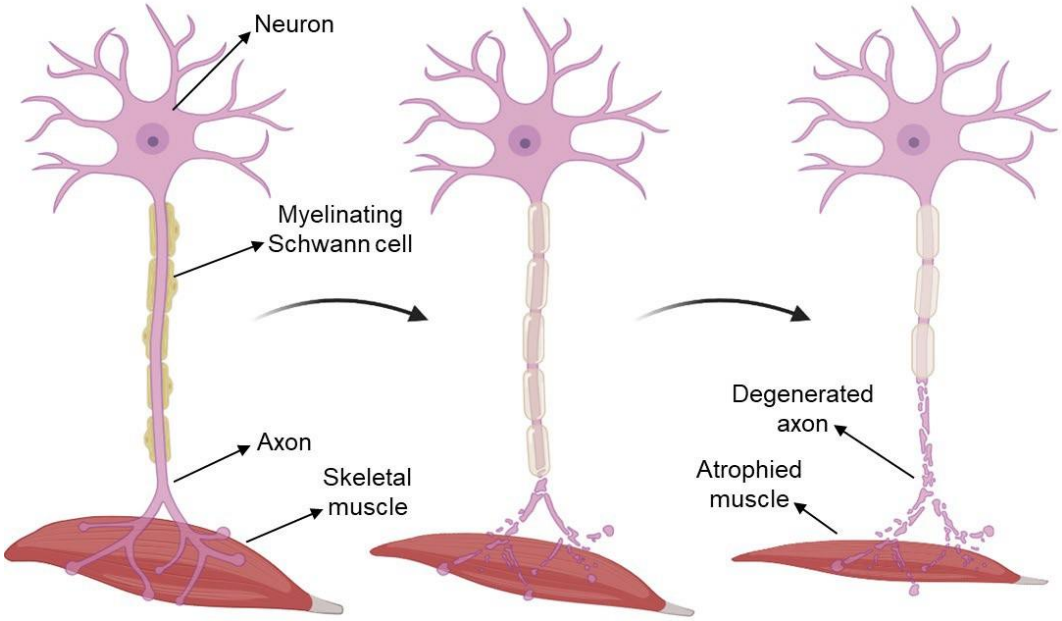


Figure 1.1: Diagram illustrating the degeneration of the axon of a motor neuron and atrophy of the denervated muscle in CMT. Figure was prepared using BioRender.

1.3. Classification of CMT

CMT is traditionally divided into two major groups based on electrophysiological criteria [5, 20]. CMT type 1 (CMT1), also known as the demyelinating form, is caused by the degeneration of the Schwann cells that comprise the myelin sheath in the PNS [21]. This results in reduced median motor nerve conduction velocities (mNCV) below 38 m/s [5, 22]. CMT type 2 (CMT2) also known as axonal CMT is caused by the degeneration of axonal projections [5, 11]. Due to the absence of demyelination in CMT2, mNCVs remain above the 38 m/s threshold, however, sensory nerve action potentials (SNAPs) and compound muscle action potentials (CMAPs) are significantly reduced, indicating axonal loss [5, 21, 23]. An intermediate form of CMT has also been described in which families present electrophysiological and pathological findings for both CMT1 and CMT2 [24, 25]. Intermediate CMT features varying degrees of axonal degeneration and demyelination [24], which results in mNCVs that usually range between 25-45 m/s [20, 25].

1.4. Genetics of CMT type 2 (CMT2)

CMT2 accounts for 12 to 36% of all CMT cases with an approximate prevalence of 1/10000 [8, 26, 27]. Inheritance of CMT2 can be autosomal dominant [28], autosomal recessive [29] or have *de novo* occurrence of gene mutations [9]. Additionally, maternally inherited CMT2 can be caused by pathogenic mutations in the mitochondrial genes *MT-ATP6* [30] and *MT-TV* [31]. The clinical and genetic variability observed for CMT2 often poses major challenges for clinical and genetic investigations. CMT2 is genetically heterogeneous, giving rise to highly similar clinical phenotypes [32]. Interestingly, a significant proportion of the genes associated with CMT2 can also cause other forms of CMT and even other IPNs [33, 34].

The current study will focus on identifying the genetic cause of an unsolved family with autosomal dominant CMT2 (AD-CMT2). The genes known to cause autosomal dominant forms of CMT2 are shown (Table 1.1) (for reviews see [32, 33, 35]).

Table 1.1: The genes currently known to cause autosomal dominant CMT2. Identifier numbers are obtained from Online Mendelian Inheritance in Man (OMIM) database [36]. References indicate the articles that first reported each gene in CMT2.

Gene SYMBOL	Gene name	OMIM ID	CMT2 subtype	References
AARS1	Alanyl-tRNA synthetase 1	601065	CMT2N	[37]
ATP1A1	ATPase Na+/K+ transporting subunit alpha 1	182310	CMT2DD	[38]
CADM3	Cell adhesion molecule 3	609743	CMT2FF	[39]
DHTKD1	Dehydrogenase E1 and transketolase domain containing 1	614984	CMT2Q	[40]
DNM2	Dynamin 2	602378	CMT2M	[41]
DYNC1H1	Dynein cytoplasmic 1 heavy chain 1	600112	CMT2O	[42]
GARS1	Glycyl-tRNA synthetase	600287	CMT2D	[43]
GBF1	Golgi-specific brefeldin-A resistance factor 1	603698	CMT2GG	[44]
GDAP1	Ganglioside-induced differentiation-associated protein 1	606598	CMT2K	[45]
HARS1	Histidyl-tRNA synthetase 1	142810	CMT2W	[46]
HSPB1	Heat shock protein family B (small) member 1	602195	CMT2F	[47]
HSPB8	Heat shock protein family B (small) member 8	608014	CMT2L	[48]
LRSAM1	Leucine rich repeat sterile alpha motif containing 1	610933	CMT2P	[49]
JAG1	Jagged canonical Notch ligand 1	601920	CMT2HH	[50]
MARS1	Methionyl-tRNA synthetase 1	156560	CMT2U	[51]
MFN2	Mitofusin 2	608507	CMT2A	[52]
MME	Membrane metalloendopeptidase	120520	CMT2T	[53]
MORC2	MORC family CW-type zinc finger 2	616661	CMT2Z	[54, 55]
NAGLU	N-acetyl-alpha-glucosaminidase	609701	CMT2V	[56]
NARS1	Asparaginyl-tRNA synthetase 1	108410	Undefined	[57]
NEFH	Neurofilament heavy chain	162230	CMT2CC	[58]
NEFL	Neurofilament light chain	162280	CMT2E	[59]
RAB7A	RAB7A, member RAS oncogene family	602298	CMT2B	[60]
SLC12A6	Solute carrier family 12 member 6	604878	CMT2II	[61]
TRPV4	Transient receptor potential cation channel subfamily V member 4	605427	CMT2C	[62]

Although AD-CMT2 is a primary axonal neuropathy, only a small number of causative genes associated with this disease have highly specific roles in the PNS [63]. Most AD-CMT2 genes are ubiquitously expressed with critical biological roles in the function and structural organisation of axons [63, 64]. The high physiological demands of axons, due to their extreme distance from the neuron soma (up to 1 meter) make them vulnerable to pathogenic insults [63, 65], however the diverse pathomechanisms causing peripheral nerve axonal degeneration remains poorly understood [66, 67]. Figure 1.3 provides an overview on the endogenous roles of AD-CMT2 genes. This figure additionally includes all the genes currently known to cause axonal sensorimotor neuropathy to demonstrate the extensive overlaps and characteristic differences in biological processes that are associated with AD-CMT2 and other disorders with similar neurological symptoms.

Despite the large number of genes known to cause CMT2, the average diagnostic rate remains at 35% [68], indicating that other causative genes remain to be discovered [69]. In unsolved cases, knowledge of the biological pathways that underly CMT2 can allow identification of candidate genes with similar or complementary roles, and thus, aid gene discovery [70].

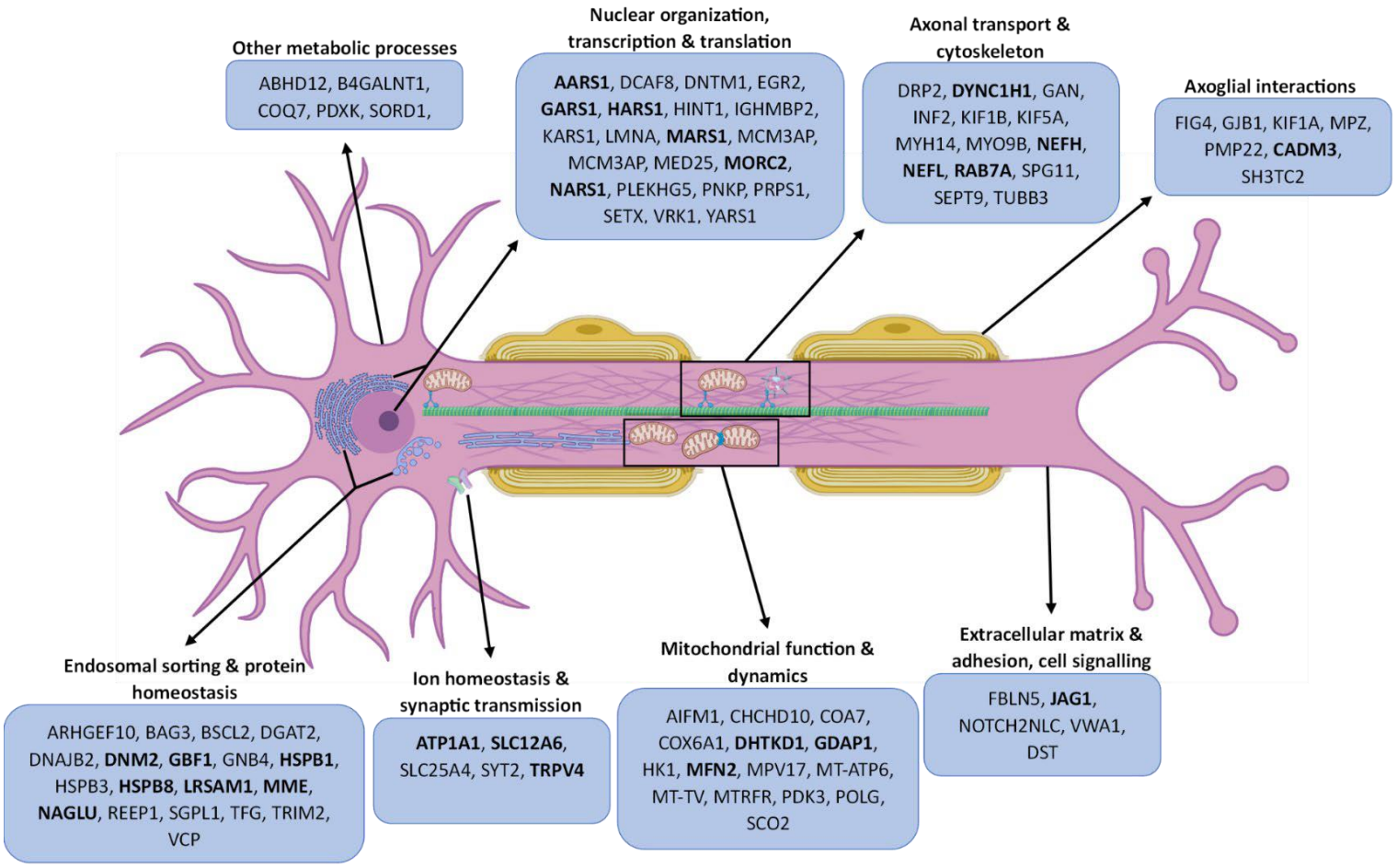


Figure 1.2: Genes known to cause axonal sensorimotor neuropathy in the PNS and the biological processes associated with their pathogenic mechanisms. Genes are grouped based on their biological roles. Genes associated with AD-CMT2 are highlighted in bold. The information presented here is summarised from [10, 71-73] as well as the articles originally reported AD-CMT2 genes as summarised in Table 1.1. Figure was generated using BioRender.

Chapter 1
Part 2

Determining the genetic cause of CMT2 can be challenging and for many families may involve diagnostic odysseys that span decades. For CMT2 families that remain unsolved after the exclusion of known genes, having a strategy to identify candidate disease genes is the cornerstone for delivering precision medicine [74]. The combination of positional cloning and next-generation sequencing (NGS) is now a common approach in Mendelian disorders to identify pathogenic genes in the modern genomic era [75, 76]. The combined use of these tools has been particularly powerful for solving CMT cases by dramatically reducing the number of genes to be analysed and facilitating cost-efficient identification of all potentially pathogenic variants [46, 77].

1.5. Positional cloning

The first step in positional cloning studies is to perform family linkage analysis to identify the disease locus [78]. Genetic linkage is the phenomenon where two loci are more likely to co-segregate during meiosis if they are located closer to each other on the same chromosome, due to the reduced chance of recombination occurring between them [79]. This principal concept is used to determine the linkage between a genomic locus and the disease phenotype segregating in a family [80]. The experimental process in linkage analysis is to genotype individuals in a family with genetic markers with known positions throughout the genome [81]. The aim is to identify a region of DNA with a haplotype of marker alleles that segregate with affected individuals [80]. By tracking the segregation of the marker alleles defining the DNA region (haplotype), a disease locus can be mapped to a region of the genome that is uniquely shared among a group of patients

with a monogenic disorder [80, 82]. Whether such marker haplotypes indicate true linkage or occur by chance is determined using statistical tests [83].

Single nucleotide polymorphisms (SNPs) and microsatellites are commonly used markers in linkage analyses, with each having distinct advantages and limitations [84]. Microsatellites are short tandem repeats (STRs) that comprise highly variable numbers of repeat units [85], therefore, they have more alleles than the biallelic SNPs and are more informative as markers for linkage studies as this increases the chance of genotypes being heterozygous in family members [86, 87]. In contrast, SNP markers can be genotyped using microarray technologies in a high-throughput manner and cover the genome more densely [88, 89]. Although SNPs are biallelic, the high density of SNPs used in microarrays increase the chances of detecting informative recombination events to delimit linkage regions to search for mutations [84].

The logarithm of the odds (LOD) score method is a statistical test in linkage analysis to assess the likelihood that the DNA markers genotyped in a family exhibit linkage to the disease locus [83, 90]. For this method, the number of recombinant and non-recombinant individuals in a pedigree segregating a disease are determined based on the marker genotypes observed in a family [91]. This information is used to calculate the probability of recombination occurring between markers and is expressed as the recombination fraction (θ) [91]. The LOD score is calculated as the logarithmic ratio of the likelihood that disease locus and genotyped marker are linked over the likelihood that they are unlinked (θ=0.5), as shown in Figure 1.3 below:

$$LOD = log_{10} \frac{(1-\theta)^{NR} x \theta^{R}}{0.5^{(NR+R)}}$$

Figure 1.3: Calculation of LOD scores in linkage analysis. R = number of recombinants in a family; NR = number of non-recombinants in a family. The standard LOD score threshold of 3 indicates 1000:1 odds in favor of linkage, whereas a LOD score of -2 indicates 100:1 odd against linkage and is significant evidence for exclusion of the DNA markers tested at specific θ values [83, 92]. LOD scores between -2 and

3 do not provide evidence of significance. For small families that do not have the power to demonstrate significant linkage (LOD>3), LOD scores >1 are considered to indicate “suggestive linkage” and constitute possible target regions for positional cloning studies [93, 94]. In model-based linkage analyses, large families segregating a disease can often provide sufficient numbers of informative meiosis events to establish significant linkage [91, 95]. In contrast, less powerful small families usually yield multiple loci with suggestive LOD scores [75, 96].

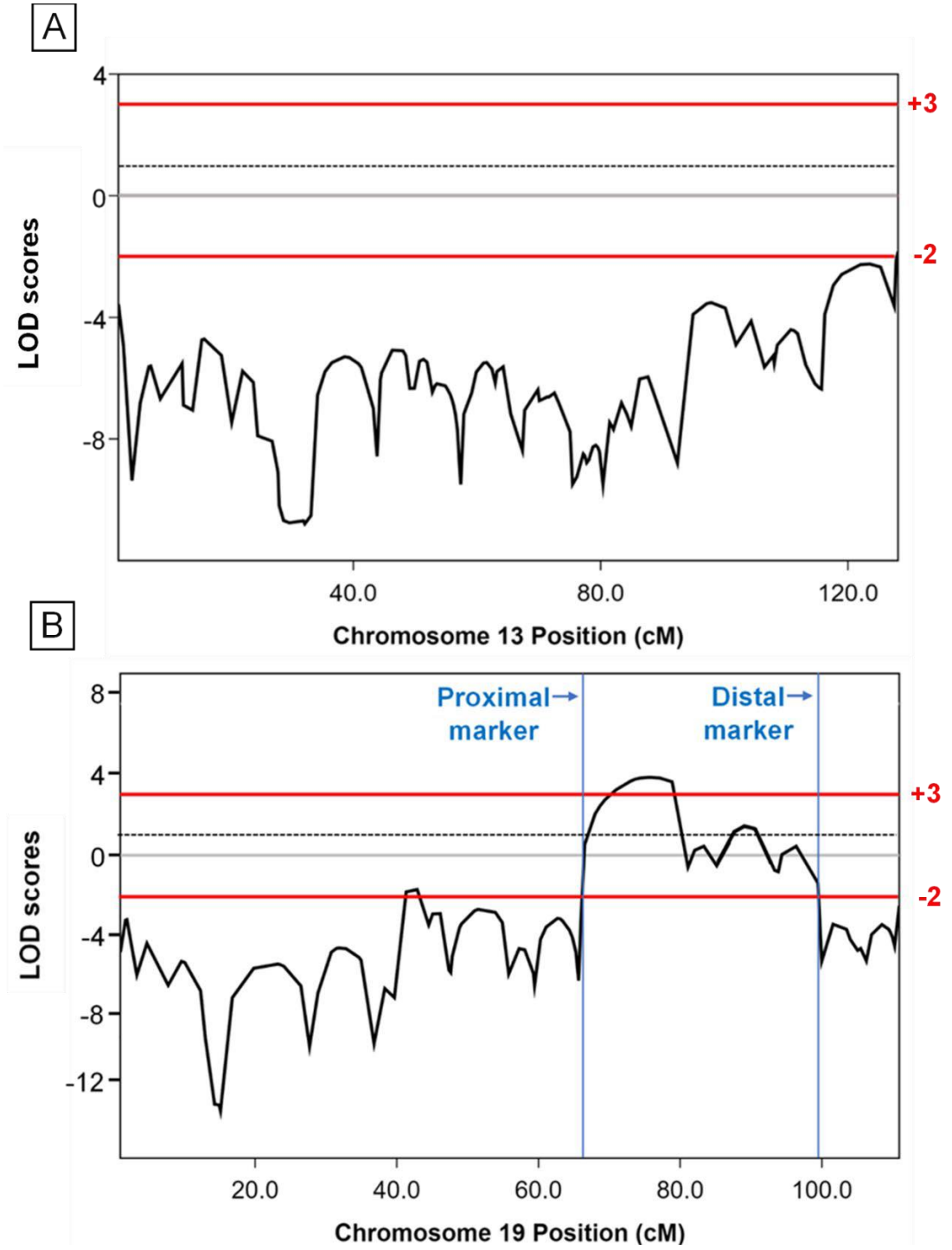


Figure 1.4: Sample linkage likelihood curves showing the power to exclude genomic regions and identify linkage loci. The multipoint LOD scores are plotted against the chromosomal positions to generate a likelihood curve. The upper and lower red lines respectively represent LOD score thresholds of +3 and -2, while the dashed black line representing the LOD score threshold +1 for suggestive linkage. (**A**) Representative markers genotyped on chromosome 13 in a family produced multipoint LOD scores below -2, thereby demonstrating the power of exclusion using linkage analysis. (**B**) Theoretical markers are ordered

from the telomere of the p arm of chromosome 19 to the telomere of the q arm of chromosome 19. Regions of the likelihood curve with maximum LOD scores represent candidate linkage regions mapping in a family. The vertical blue lines represent markers that demarcate the linkage region. This is based on the first marker that excludes (LOD < -2) on either side of the linkage likelihood curve.

For Mendelian gene discovery, mapped linkage loci limit the analysis to a smaller set of positional candidates for identifying the pathogenic mutation [80]. Relevant positional candidate genes that reside within significant or suggestive linkage loci can be identified using genomic repositories including RefSeq [97] and GENCODE [98]. For suggestive linkage regions the yield of gene variants can be in the hundreds [99], therefore, having large-scale NGS sequencing available is key for assessing variants in all positional candidate genes and identifying the pathogenic mutation in small families [96, 100].

1.6. Next-generation sequencing and variant identification

NGS is a group of sequencing technologies that revolutionised gene discovery for Mendelian disorders [101, 102], and caused a surge of pathogenic gene discovery in CMT [103, 104]. In NGS, thousands to millions of fragments of the sample DNA are sequenced simultaneously in a massively-parallel fashion, which allows determining nucleotide identity across large genomic regions at a fraction of the cost and time compared to the classical Sanger sequencing method [105, 106]. Short-read sequencing is the most commonly used NGS method which produces paired-end reads that are 35-300 bp long in length [107, 108]. Recently, long-read NGS technologies that can produce reads ranging from 10 kilobases (kb) to 1 megabase (Mb) have been increasingly used in the field of genomics due to the major advantages over the short-read technologies [109-111], which will be discussed in detail in Chapter 5. High-throughput sequencing and variant calling with NGS has been used to identify all mutations in suggestive linkage loci that are too large to cover with Sanger sequencing [100, 112] and this has led to discovery of pathogenic genes in small families with Mendelian disorders [96], including CMT [46,

104]. NGS can be used for sequencing whole genomes and transcriptomes, but also to specifically target coding sequences [113].

1.6.1. Whole exome sequencing

Using NGS to selectively sequence all coding exons is known as whole-exome sequencing (WES) [114]. The main advantage of WES for pathogenic variant discovery is the highly targeted survey of the coding regions and the reduced costs in which only ~1% of the genome is sequenced [115]. Importantly, in a diagnostic setting, WES is an effective approach as 85% of pathogenic mutations for Mendelian disorders are found in the coding exons of genes [102, 116]. While WES has been used to discover a significant portion of the currently known CMT genes [33, 104], 40% of CMT patients remain without genetic diagnosis following WES [117]. For CMT2, 58% [117] to 76% [118] of families remain unsolved. This has motivated researches to consider alternative methods for targeting the variants that are challenging to interpret or not detected with WES [119-121].

1.6.2 Whole genome sequencing

Using NGS to sequence all genomic DNA is known as whole-genome sequencing (WGS) and this method is the most comprehensive and efficient tool for achieving genetic diagnosis in Mendelian disorders [122, 123]. WGS resolves all coding and noncoding regions of the genome and increases the diagnostic rates in WES-negative cases of CMT by providing better detection of the full spectrum of pathogenic variants [121, 124, 125]. These include the single nucleotide variants (SNVs) and small insertion/deletion events (indels) that form the most diverse group of pathogenic CMT variants [126], but also the relatively rare structural variants (SVs) and repeat expansions (REs) [119, 124] (Figure 1.5). Arguably the strongest advantage of WGS over WES is enabling the investigation of noncoding variants, which remain severely understudied in

Mendelian disorders despite their well-established pathogenic potential [127, 128]. Utilising WGS to study the noncoding variants has substantially furthered our understanding of the molecular disease mechanisms and the elements of genome that underly CMT and IPNs [129-131]. For example, deep intronic SNVs identified through WGS in *MME* and *IGHMBP2* in recessive CMT2 cases are shown to disrupt normal splicing and give rise to abnormal transcripts that are degraded, leading to loss of function in each respective gene [130, 132].

Using WGS, our group discovered a 1.35 Mb translocation at 7q34-q36.2 locus that causes dHMN1 (OMIM:182960) [120] by generating a gene-intergenic fusion transcript formed between a partial copy of *UBE3C* and noncoding intergenic sequence [131]. Duplications that increase the copy number of an upstream distal enhancer have been reported to drive overexpression of *PMP22* and cause CMT1, demonstrating altered regulatory interactions as a potential pathogenic mechanism in CMT [129]. REs are also emerging as pathogenic mutations that highlight the need to use WGS in CMT diagnosis [124]. Recently, REs that disrupt the second intron of *RFC1* were identified in both CMT [133] and cerebellar ataxia with bilateral vestibular areflexia syndrome [134]. While noncoding REs or SVs, or variants localizing with gene regulatory elements have not yet been identified in AD-CMT2, it is important that these types of mutations are considered in families excluded for all coding mutations. These findings also highlight the necessity for interrogating the noncoding genome with WGS to close the diagnostic gap that remains after exhausting the exome [33, 71, 119, 124].

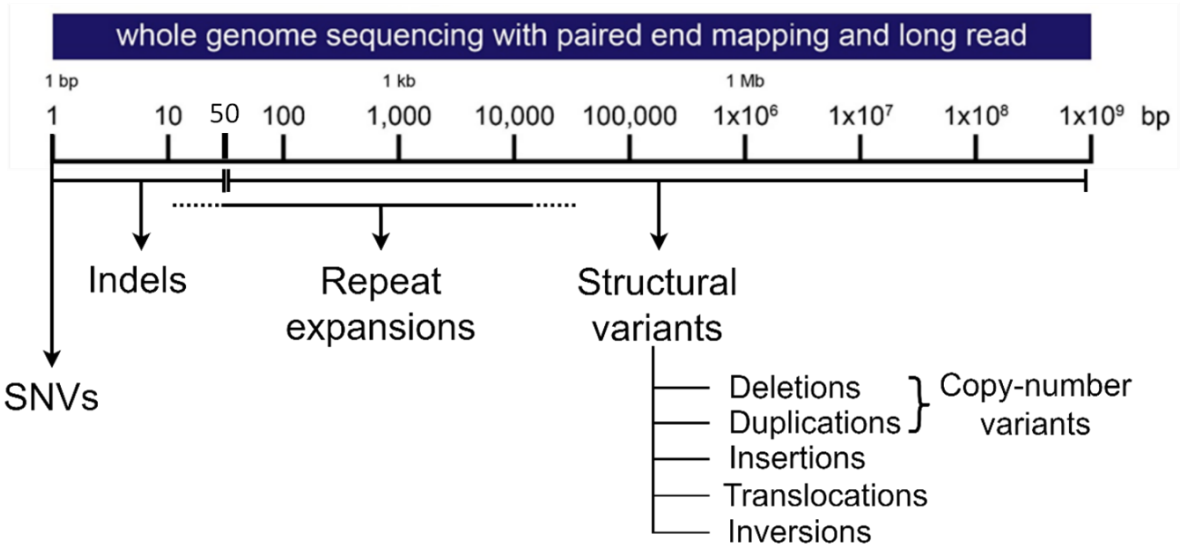


Figure 1.5: Genomic variation that can be identified by WGS. WGS can detect DNA variation ranging in size from single base pairs to millions of base pairs. Indels are <50 bp sequence alterations >50bp are defined as SVs, and are further classified into insertions, inversions, and translocations, as well as copy number variation (CNVs) which encompass duplications and deletions [135, 136]. REs are increases in the number of repeat units that comprise STRs [137] that may span few base pairs [138] up to tens of kilobases [139]. This figure was modified from a conference presentation made by Prof. Marina Kennerson.

1.7. Variant identification and prioritisation

1.7.1 Variant identification using reference genomes

The completeness of a reference genome can have a substantial impact on effectively detecting pathogenic variants. The gaps in the reference genome that has been used for the last decade (GRCh38) are known to obscure or lead to inaccurate representations of variants in genes associated with Mendelian diseases [140, 141]. In a recent review discussing the challenges of solving the remaining IPN families, it was proposed that addressing the gaps in the human reference genome and including genetic diversity would increase the success of identifying the causative variants for unsolved cases of IPNs by revealing previously inaccessible variation [71].

The Telomere-to-Telomere (T2T) consortium released the first gapless human genome reference in 2022, comprising the complete sequence of all chromosomes, including the coverage of challenging telomere, centromere and segmental duplications that were poorly represented or absent in the GRCh38 reference [142] [143]. Compared to GRCh38, aligning WGS to the T2T reference substantially reduced the number of false-positive variants found in clinically relevant genes and improved the detection and representation of pathogenic variants [141].

Inclusion of genetic diversity in a reference genome is also important for interpreting potentially pathogenic mutations [144]. The draft human pangenome reference (DHPG) was released in 2023 and combines the sequence of fully-phased genome assemblies for 47 individuals from diverse genetic backgrounds [144]. By representing the large structural polymorphisms and hypervariable loci in the global population, the DHPG provides a well-resolved profile of benign polymorphisms which represents the “normal” genomic landscape of the human species [144, 145]. A duplication event resolved in the DHPG included an additional copy of the *NOTCH2-NLC* gene associated with CMT2 [144, 146]. This demonstrated the power of pangenome reference to exclude benign variation that may otherwise have appeared potentially pathogenic. Combining the T2T and human pangenome references genome resources represents a promising strategy for accurately identifying variants throughout the whole genome and effectively excluding benign variation in genetic diagnosis studies.

1.7.2. Linkage analysis and filtering for variant prioritisation

While WGS offers the highest potential for diagnosis and gene discovery in CMT [124], it can detect 4 to 5 million variants in a single individual [147], which creates a large burden for variant analysis and requires implementation of robust variant filtering and prioritisation strategies [127, 148]. Many studies will initially employ filtering workflows to select for the variants that show similar properties to the expected pathogenic allele [149] such as variant genotype,

segregation, and population frequency [148, 150]. Targeting linkage loci can also serve as a highly efficient filtering step that may eliminate more than 99% of the extraneous variants identified by NGS [75], and has been used to expedite variant analysis following WGS [100, 112]. Filtering is effective for obtaining a small number of coding variants [148] which can be effectively prioritised using *in silico* tools to predict variant pathogenicity [151, 152].

1.7.3. Multi-omics strategies for prioritising noncoding variants

Even after extensive filtering and restricting the area of search, thousands of noncoding variants identified by WGS often remain [153]. These variants are rarely included in genetic investigations of Mendelian disorders [154, 155] since predicting their impact is difficult due to insufficient knowledge on the function of noncoding sequences [127, 156, 157]. Recently, multi-omics strategies have been successfully implemented to identify the noncoding pathogenic variants across a variety of Mendelian disorders that remain unsolved after WES or use of WGS in isolation [158-160]. These approaches derive experimental evidence of pathogenicity from multiple functional genomics methods, such as transcriptomic and epigenomic analyses, to prioritise the noncoding mutations that are most likely to underly the disease phenotype [159-161].

A variety of functional genomics methods can be incorporated into multi-omics analyses depending on the molecular pathology of the disease being investigated [160]. The aberrant transcripts identified through transcriptomics studies can guide the variant analysis towards intronic variants that disrupt splicing [128, 158] or SVs that give rise to fusion transcripts [162]. Transcriptomics can also be used to detect dysregulated genes and prioritise the noncoding variants disrupting the regulatory 5' and 3' untranslated regions (UTRs) [128]. Gene dysregulation identified by transcriptomic analysis can also be used for identifying pathogenic variants in cis-regulatory elements (CREs)[128], which are noncoding sequences (such as promoters and enhancers) that control the expression of nearby genes [163]. Variants in CREs are implicated in

CMT as demonstrated by the deletions in the promoter of *GJB1* causing CMT1X and the SNVs that disrupt a downstream enhancer of *SH3TC2* and increase the disease severity in patients with demyelinating CMT [164]. Epigenomics can be used in combination with transcriptomics to identify the CREs associated with the dysregulated genes to direct the analysis towards noncoding variants disrupting gene expression [165-167]. Therefore, multi-omics variant prioritisation has high potential for accelerating the discovery of pathogenic variants in CMT by directing the analysis towards the pathogenic noncoding variants in WES-negative cases [129-131].

1.8. Family CMT720

1.8.1 Initial genetic evaluation of CMT720

The current study investigates a multigenerational Polish family (CMT720) that clinically presents with CMT2 and was first reported by our group in 2005 [168]. The family showed the hallmark clinical symptoms of CMT with no CNS involvement and mNCVs that remained above 38 m/s, supporting a classical CMT2 phenotype [168].

The pedigree of CMT720 (Figure 1.6) suggests autosomal dominant inheritance pattern [168]. The phenotypes of individuals V:1, V:3 and V:5 are coded as unknown since they were too young at the time of the clinical visit, and were not evaluated [168]. While no male-to-male inheritance was observed in this family, if there was X-linked inheritance, males would usually present with more severe symptoms which occurs in most cases of X-linked CMT [169]. The lack of differences between phenotypic severity of males and females in CMT720 further supported autosomal dominant inheritance in the family [168]. During the initial genetic evaluation of CMT720, our group excluded the 9 loci (CMT2A, CMT2B, CMT2C, CMT2D, CMT2E, CMT2F, CMT2G, CMT2K and CMT2L) known to cause CMT2 in 2005 by performing haplotype analysis [168]. Furthermore, the *GJB1* and *MPZ* loci were respectively excluded through Sanger sequencing and haplotype analysis to rule out the possibility of an intermediate form presenting

with predominantly axonal symptoms in CMT720 [168]. Despite these early efforts to genetically diagnose CMT720, this family currently remains unsolved.

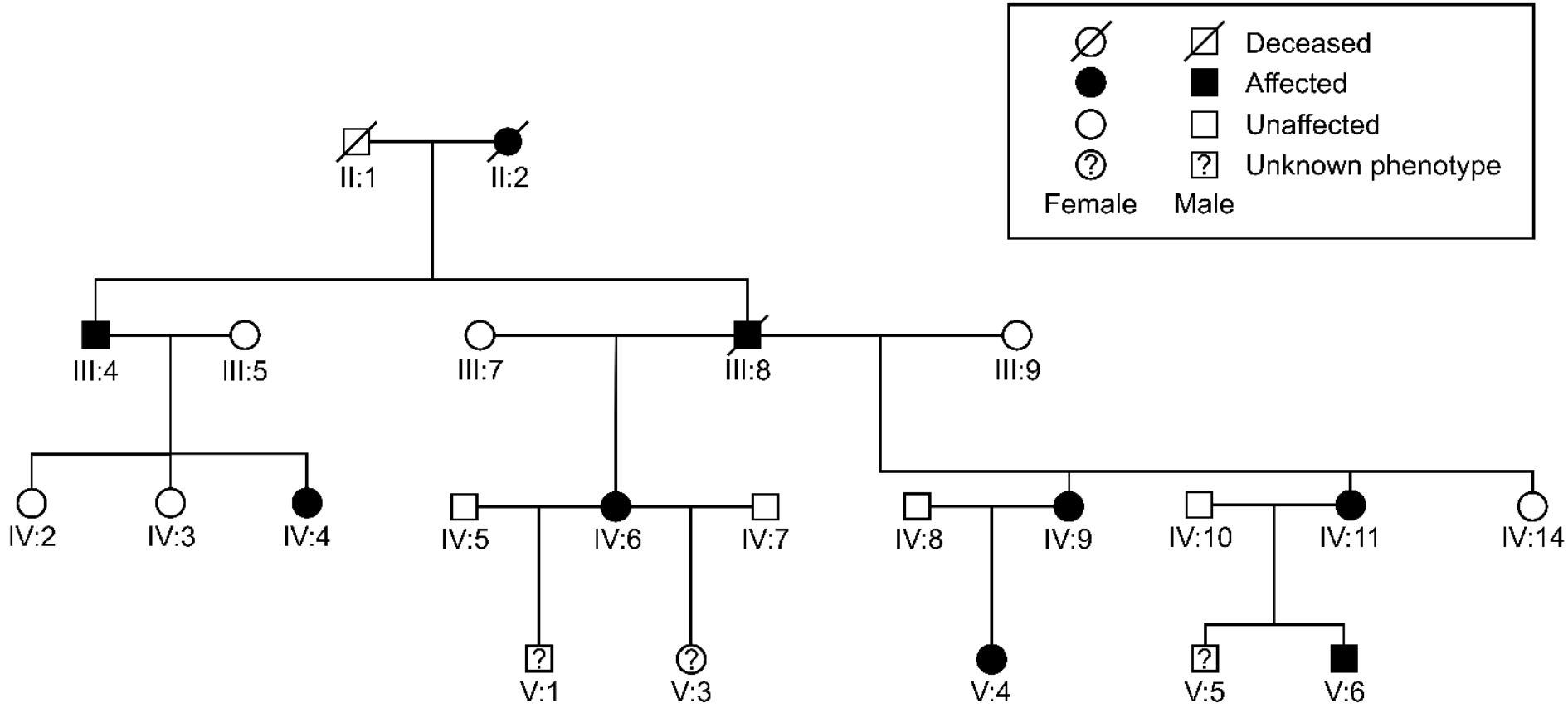


Figure 1.6: The pedigree of family CMT720. The pedigree of family CMT720. Individuals are numbered consecutively in each generation, from left to right. The legend for symbols is provided in the box above.

1.9. More recent molecular investigations of CMT720

Prior to the commencement of this MPhil candidature, WGS and linkage analysis was performed on family CMT720 for analysing all coding regions for pathogenic mutations. The following provides the findings of this investigation.

1.9.1. Linkage analysis

Linkage analysis was carried out on 17 individuals in family CMT720 (Figure 1.7) based on SNP markers that were genotyped using the Linkage V Panel (Illumina). An affected- only method was implemented by coding “unknown” phenotypes to all at risk individuals presenting as unaffected or for individuals that were too young for clinical examination and were not available for further clinical follow up (IV:2, IV:3, IV:14, V:1, V:3 and V:5). Individual V:4 was also coded as “unknown” phenotype because the patient showed mild symptoms in the initial neurological examination [168] and was not available for clinical follow up in 2022. This approach was taken as a conservative measure against reduced penetrance in this family since the primary clinician Andrzej Kochanski informed that patient V:4 shows a much milder clinical phenotype. The incorrect specification of affection status can significantly reduce the power to detect linkage [170], therefore, only the individuals that showed a clear disease phenotype were coded as affected and married-in individuals were coded as unaffected. This analysis resulted in identification of 5 suggestive linkage loci on 4 different chromosomes as summarised in Table 1.2 below. These suggestive regions spanned a total of 83.32 Mb.

Table 1.2: Suggestive linkage regions identified for CMT720 and the associated LOD scores.

Chromosome	Chromosomal position of suggestive linkage region	Flanking SNP markers	Size (Mb)	Maximum LOD score
8	69118917 - 88164384	rs695167-rs1519938	19.1	1.4597
15	61046278 - 71168354	rs782944-rs1348318	10.1	1.4764
16	190281 - 6949202	rs1211375-rs7195006	6.76	1.3505
16	22851014 - 57077090	rs208965-rs41383	34.23	1.502
18	65415362 - 78542601	rs565973-rs1866338	13.13	1.4795

1.9.2. Whole-genome sequencing of CMT720

Using short read WGS (srWGS), the candidate linkage regions were assessed for all coding mutations. The affected individuals IV:4 and V:6 underwent srWGS (Figure 1.7) and the resulting data was uploaded to an online bioinformatic analysis platform Seqr (hosted at the Garvan Institute under the Centre of Population Genomics). This platform provides extensive options for variant filtering, annotation, and *in silico* prediction for pathogenicity [171]. All SNVs and indels were identified in both patient and underwent variant filtering. The coding sequences of all IPN and CMT genes were screened, however, no pathogenic mutations were identified. The suggestive loci that were mapped by linkage analysis in the family, were also screened for coding variants. No variants were identified that segregated with affected family members. Accordingly, the combined srWGS and linkage analysis excluded all coding SNVs and indels in known CMT genes and linkage positional candidate genes in CMT720.

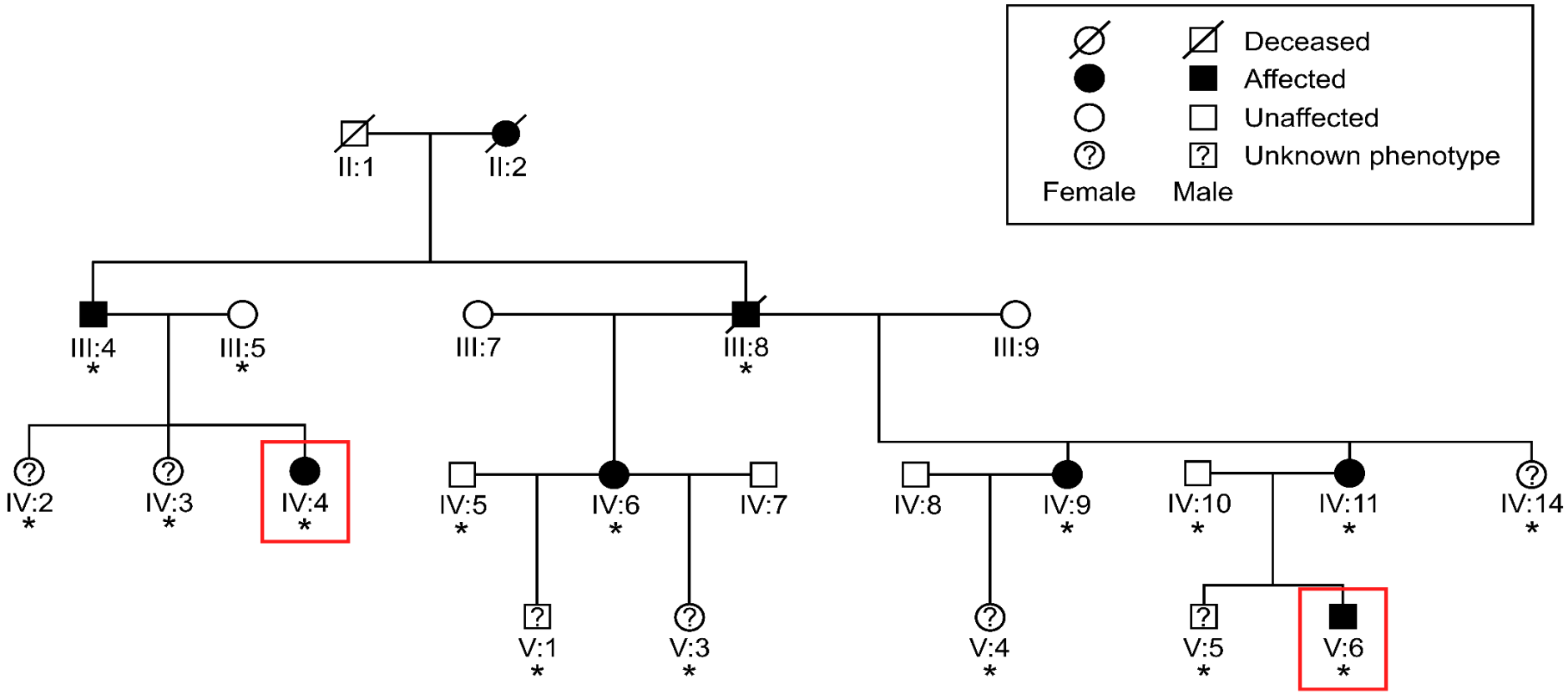


Figure 1.7: The pedigree of family CMT720 indicating phenotypic status of individuals for "affected only" analysis. Individuals are numbered consecutively in each generation, from left to right. Individuals genotyped for the SNP based linkage analysis are indicated by asterisks. The individuals that underwent WGS are indicated with a red box. The legend for symbols is provided in the box above. This pedigree was adapted from the original publication by Kochanski *et al.* [168] and the individual identifiers were retained from the original pedigree.

1.10. The strategy for identifying the pathogenic variant in CMT720

Since the previous investigations have excluded all coding SNVs and indels in CMT720, this study will consider the remaining variants within the suggestive linkage loci that have not been assessed. This includes two broad categories: 1) all forms of noncoding variants and 2) SVs and REs that may be impacting positional candidate genes in the suggestive linkage regions. WGS is the optimal choice to detect all these variant types, however, using this method is anticipated to yield tens of thousands of variants across our suggestive linkage loci. Therefore, in this study, we will develop a filtering and prioritisation strategy for reducing the number of potentially pathogenic candidate variants that will be detected via WGS to facilitate the identification of the pathogenic mutation in CMT720.

Since SNVs and indels are the most common types of pathogenic coding variants in CMT [18, 126], their exclusion in CMT720 suggests that the pathogenic variant in this family is likely to be non-coding. The main focus of this study, will be to effectively prioritise and analyse the non-coding variants identified in CMT720, which will constitute the vast majority of the variants found across our suggestive linkage loci. To address the possibility that a previously unexplored type of variant could be impacting the coding region of positional candidate genes in our suggestive linkage loci, the region will also be analysed for SVs and REs which were not queried in previous investigations of CMT720 WGS data.

To achieve these goals, we will initially perform a fine mapping linkage analysis within the previously identified 5 candidate linkage regions using microsatellite markers to eliminate the false positive suggestive linkage peaks and prioritise the remaining suggestive linkage loci for variant identification. By targeting the region of search to prioritised linkage loci, this will reduce the number of variants requiring analysis [172]. A transcriptomic profile of affected individuals from CMT720 will be constructed using patient derived fibroblast cell lines to identify the aberrant transcripts and dysregulated genes observed within our prioritised linkage regions. The functional

evidence for pathogenicity provided by these transcriptomic abnormalities can be used to guide the selection and analysis of noncoding variants in CMT720 that are likely to be impactful [128, 158]. Based on the findings from the transcriptomic analysis, all variants in the non-coding regions of genes (introns or regulatory elements) that give rise to abnormal transcripts or show dysregulation will be prioritised. Additionally, epigenomics datasets will be integrated in the multi-omics prioritisation strategy to identify the noncoding variants that may disrupt the CREs associated with dysregulated candidate genes. By following this variant identification and prioritisation strategy in our current investigation of CMT720, we expect to conduct a comprehensive and effective analysis for all types of variants that may cause CMT2 in this family.

Linkage analysis is a powerful method that provides substantial filtering power for studies of genetic disease utilising NGS [96], and multiomic variant prioritisation is currently the most effective strategy to facilitate identification of impactful noncoding variants [127]. By combining these two methods in a single workflow to analyse all forms of genomic variation, our study of CMT720 will provide a comprehensive and novel model for exploring the variation beyond the exome for genetic diagnosis of CMT.

Below is a flow chart (Figure 1.8) summarising the integrated use of linkage analysis, WGS and multi-omics (transcriptome and epigenomics) we will employ for variant prioritisation in this current study of CMT720.

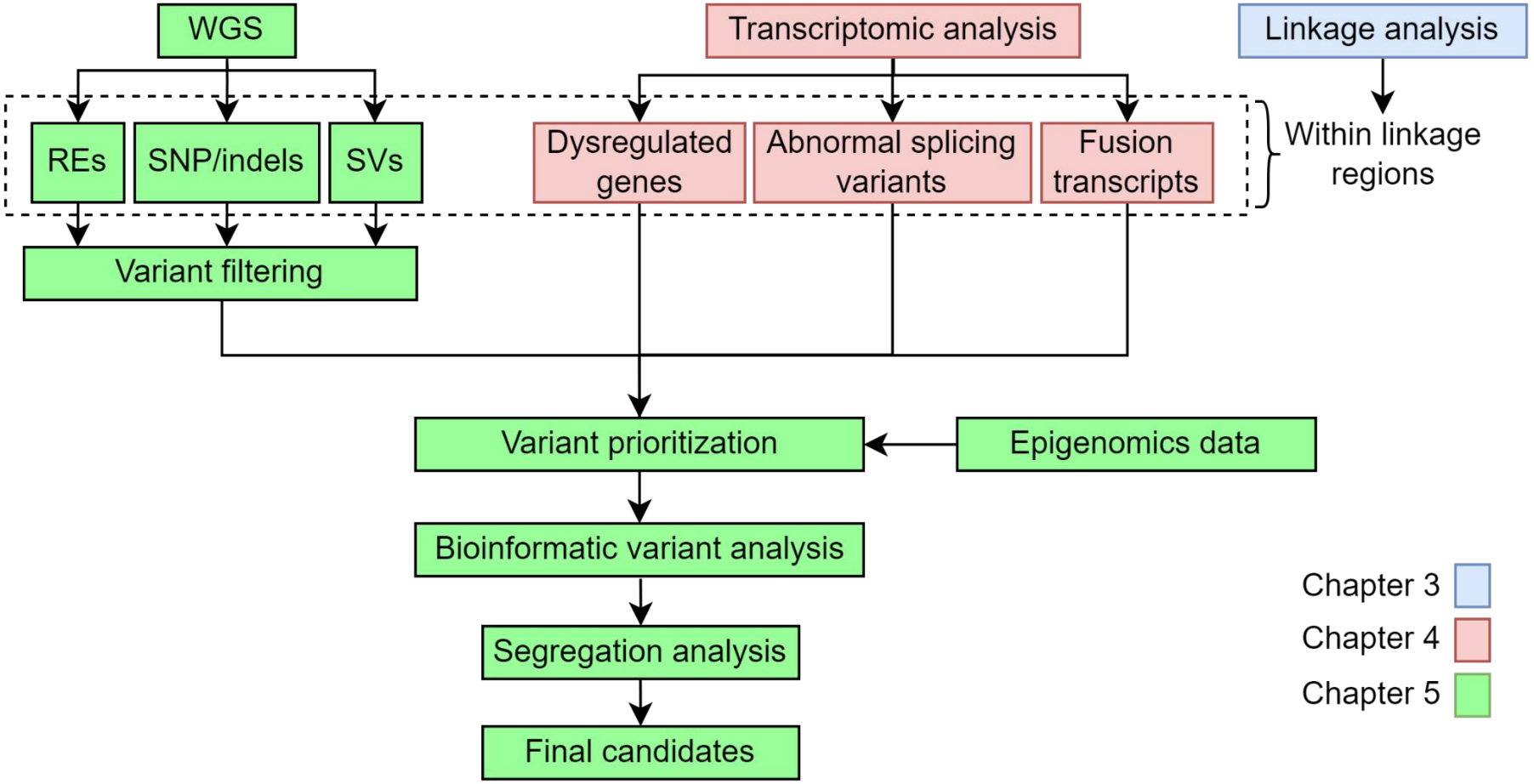


Figure 1.8: Workflow of the strategy developed to identify the pathogenic variant and the disrupted gene in CMT720. The methods used in each chapter have been colour coded as indicated in the legend at the bottom right.

1.11. Project Rationale

The current project will target the full spectrum of noncoding genomic variants for a pathogenic role in CMT720. Refining the candidate suggestive linkage regions and constructing the transcriptomic profile of CMT720 using patient fibroblast cell lines will provide a useful guide for variant filtering selection and prioritisation. Utilising epigenomics data will guide the selection of noncoding variants that may impact regulatory elements associated with the dysregulated positional candidates. Combining linkage analysis, WGS and multi-omics variant prioritisation will maximise the chance of detecting the pathogenic variant causing CMT720.

1.12. Hypothesis

That the combination of next generation sequencing, modern genomic resources, linkage analysis and multiomics will facilitate variant prioritisation and identify the pathogenic non-coding variant in CMT720.

1.13. Overall aim:

Identify the non-coding pathogenic mutation and gene disrupted in CMT720.

1.14. Specific aims:

- Refine the suggestive linkage loci identified in CMT720 by fine mapping using microsatellite markers.
- Generate a transcriptomic profile for patient fibroblasts from family CMT720 and examine the refined linkage loci for abnormal transcripts and gene dysregulation.

- Identify candidate SNVs, indels, SVs or REs from WGS data, test for segregation in the remaining family members and provide supporting bioinformatic evidence that the variants selected, may underly the transcriptome abnormalities of positional candidate genes localising to the refined suggestive linkage loci.

Chapter 2
General Materials and Methods

2.1. Ethics statement

All protocols used in this study were approved by Sydney Local Health District Human Ethics Review Committee (2019/ETH07839) and informed consent was obtained from all participants.

2.2. Standard cleaning and handling procedures

All glassware, 1.5mL microcentrifuge tubes, pipette tips without aerosol barriers and glassware were sterilised using autoclave. Before and after each use, benchtops in the lab and inside the laminar flow hoods were wiped with 70%(v/v) ethanol (Lomb Scientific). The laminar flow cabinets were sterilised under ultraviolet light for 30 min after each use. Procedures with fluorophore-labelled reagents were carried out in the dark.

2.3. Common reagents

UltraPure™ nuclease-free distilled water (Invitrogen) was used in PCR reactions, and for the storage of DNA samples. Millipore Milli-Q® UltraPure (Merck), after sterilisation was used to prepare TAE buffer and media used to culture fibroblasts. Gels were prepared using agarose powder (Bioline) and 2%(v/v) TAE buffer prepared from 50x TAE buffer stock solution (Astral Scientific) and MilliQ water (Merck).

2.4. Common equipment

Pipetman (Gilson) P2 (0.2-2µL), P20 (2-20µL), P200 (20-200µL) and P1000 (100-1000µL), as well as Acura (Socorex) 1-10µL multichannel micropipettes were used for pipetting. 10µL, 20µL, 200µL and 1000µL pathogen and nuclease free aerosol barrier tips (Interpath Services) were used in all procedures except when loading PCR products on agarose gels, for which autoclaved 20µL**,** Sapphire pipette tips (Greiner) were used. All microcentrifuge tubes used in experiments were 1.5mL flat capped conical tubes (Interpath Services). All PCR reactions and samples submitted for capillary electrophoresis were prepared in UltraFlux 200µL 8-strip PCR tubes with dome caps (SSbio). Samples submitted for Sanger sequencing were prepared in flat-top non-skirted 96-well PCR plates (Scientific Specialties). All qRT-PCR reactions were prepared in MicroAmp Fast Optical 96-Well Reaction Plate with Barcode (Applied Biosystems). All 96-well plates were centrifuged in Allegra X12R (Beckman-Coulter), covered with ABsolute QPCR Seal (Applied Biosystems) and heat-sealed using Combi Thermo-sealer (Applied Biosystems).

The contents of 1.5ml microcentrifuge tubes and 8-strip tubes were mixed using REAX top (Heidolph) vortex machine. E-centrifuge (WEALTEC) was used for brief centrifugation of 1.5ml microcentrifuge tubes, whereas, Centrifuge 5417C (Eppendorf) was used when a specific g force and duration was needed. 8-strip tubes were centrifuged using Capsulefuge PMC-860 (TOMY). Thermomixer Compact (Eppendorf) heating block was used to incubate all reactions carried out in 1.5 mL microcentrifuge tubes. Samples submitted for capillary electrophoresis or Sanger sequencing were dried at 95°C for 10 min using GeneAmp PCR System 9700 (Applied Biosystems).

2.5. Primer design and manufacture

All primers for Sanger sequencing were designed using the NCBI Primer Design Tool [173]. Briefly, the sequences of the 300 bp regions immediately upstream and downstream of the variant of interest imported into the Primer Design Tool to obtain complementary forward and reverse primers using the default parameters. Primer pairs were selected based on the criteria summarised as per common guidelines [174, 175] (Table 2.1). All primers were manufactured by Sigma-Aldrich (USA).

Table 2.1 Criteria for primers designs.

Selection criteria	Parameter range
Primer length	18-22bp
Distance of primer to the variant	>150bp
Difference of Tm* between forward and reverse primers	<5°C
3' self-complementarity	<5 bp
GC content	40-60%

***Tm: melting temperature**

2.6. Standard PCR procedures

Preparation of PCR experiments were performed in a BSC 180 (Gelman Sciences) biological safety cabinet. All oligonucleotide primers were diluted to 10 µM working stocks. DNA samples were prepared at a concentration of 5 ng/µL in nuclease-free distilled water (Invitrogen). All reaction mixtures were prepared using final volumes provided in Table 2.2. Two different polymerase mastermixes (MyTaq™ HS Red Mix 2x (Bioline) or ImmoMix™ Red 2x (Bioline)) were used to optimise and run PCR reactions depending on the performance of individual primer pairs. The mastermix used with each primer pair will be indicated in relevant chapters. Reactions were performed in a final reaction volume of 10 µL.

Table 2.2. Standard setup for PCR reactions.

Reagent	Volume (µL)
Mastermix (2X)	5
Forward primer (10 µM)	0.4
Reverse primer (10 µM)	0.4
Nuclease-free water	2.2
DNA sample (5 ng/ µL)	2
Total	10

Thermal cycling was performed using a Mastercycler Pro (Eppendorf). The standard thermal cycling program used for PCR reaction conditions as indicated below (Table 2.3). Annealing temperatures were specifically adjusted by performing optimisation runs for the primer pairs that did not produce robust amplification using the standard PCR conditions indicated below. PCR conditions optimised for each primer pair will be indicated in relevant chapters.

Table 2.3. Standard thermal cycling program used for PCR amplifications.

Step	Number of cycles	Duration (minutes:seconds)	Temperature (°C)
Heat activation of polymerase	1	1:00	95.0
Denaturation		1:00	95.0
Annealing	35	0:40	60.0
Extension		0:30	72.0
Final extension	1	2:00	72.0
End	-	Indefinite	20.0

2.7. Standard gel electrophoresis

Agarose gels (1.5% w/v) were prepared in 1X TAE buffer. 1X SYBR Green Safe DNA Gel Stain (Invitrogen) was added into molten agar. Size fractionation of PCR products was performed at 5V/cm for 50 min in 1X TAE buffer using either a Hoefer HE 33 mini electrophoresis tanks (Amersham Biosciences) or a multiSUB Choice Wide Midi horizontal electrophoresis tank (Cleaver Scientific) connected to a 3000Xi power supply (BioRad). HyperLadder 100 bp (Bioline) was used as the size standard. Gels were viewed at 470nm using

the Safe Imager 2.0 (Invitrogen) transilluminator and imaged using PowerShot S5 IS (Canon) with a Hoya O(G) filter.

2.8. Fibroblast cell culture

All fibroblasts were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with foetal bovine serum (FDMEM), prepared by adding 1% (v/v) penicillin/streptomycin (Gibco), 1% L-glutamine (Gibco) and 10% (v/v) fetal bovine serum (SAFC Biosciences) to DMEM (Gibco). All cells were incubated in Heracell 150 (Thermo Fisher Scientific) incubator at 37°C in humidified air with 5% CO_2. All media change, plating and preparation for cryopreservation was performed in 20 BH-EN 4DRT (Gelaire) laminar flow cabinet. Media was transferred using Costar Stripette 5mL, 10 mL, 20 mL, and 100 mL serological pipettes (Cornings) with an electronic pipette filler (Edwards). 1.5 mL microcentrifuge tubes (Interpath Services) and 15mL conical polypropylene centrifuge tubes (Corning) were used for collecting and pelleting cells. 2 mL internal thread cryogenic vials (Corning) were used to freeze cells for long term storage. 15ml and 50ml polypropylene centrifuge tubes (Corning) were used for making aliquots of the media and reagents. Centrifugation of conical tubes was performed using a Megafuge 8 (Thermo Fisher Scientific).

An overview of all primary fibroblast lines used in this study is provided in Table 2.4. Primary fibroblast lines (passage 3) from affected individuals IV:4 and IV:9 were obtained from our collaborator (Professor Mary Reilly) at University College London. These two lines were initially expanded for biobanking using the standard fibroblast cell culture protocol. Briefly, the frozen cell pellet was thawed at 37°C in a water bath and resuspended in 9 mL of FDMEM in 15 mL conical tubes, (Corning) and pelleted by centrifugation at 300 g for 5 min. The supernatant was discarded and the pellet was resuspended in 10 mL of FDMEM and seeded

in T-75 flasks (Corning). The fibroblasts were cultured with media changes every 3 days until 100% confluent. The confluent cells were washed with 5ml of 1X Dulbecco's phosphate-buffered saline (Gibco) and were dissociated by incubation in 1 mL of 0.5% v/v trypsin-EDTA (Gibco) for 1 min. FDMEM (8 mL) was added to the flask and cells were collected and pelleted by centrifugation at 300 g for 4 min. The cell pellet was resuspended in 5 mL of 10% dimethyl sulfoxide (v/v) dissolved in FDMEM. Aliquots (1mL) of cell resuspension were distributed into 5 cryogenic vials (Corning) and stored at -80°C.

Table 2.4: Primary fibroblast lines used in this study

Cell ID	Disease status	Sex
Patient IV:4	Affected	Female
Patient IV:9	Affected	Female
Control 1	Control	Female
Control 2	Control	Male
Control 3	Control	Male
Control 4	Control	Female

2.9. Whole genome amplification

Whole genome amplification was carried out using the REPLI-g® UltraFast Mini kit (Qiagen) as per the manufacturer’s guidelines to increase the quantity of the scarce DNA samples available on 8 affected and 9 unaffected family members as shown in Supplementary Figure 1. Briefly, 12.5% (v/v) denaturation buffer D1 and 10% (v/v) neutralizing buffer N1 were prepared by diluting Buffer DLB (supplied in the kit) and Stop Solution (supplied in the kit) in nuclease-free water (Invitrogen) respectively. 1µl of each patient DNA sample and 1µl of Buffer D1 was placed into an individual microcentrifuge tube to denature DNA, and mixed by vortexing. The tubes were incubated at room temperature (15–25°C) for 3 min. 2 µl of Buffer N1 was added to each tube to stop denaturation and mixed by vortexing. 15 µl UltraFast Reaction Buffer (supplied in the kit) and 1 µl DNA polymerase (supplied in the kit) was added

into each reaction tube containing 4 µl of denatured DNA and the mixture was incubated at 30°C for 1.5 hours. Following incubation, REPLI-g the DNA polymerase was inactivated by heating the sample for 3 min at 65°C. Whole-genome amplified DNA was diluted 1:25 (v/v) in nuclease-free water (Invitrogen) to prepare a working solution for all downstream applications involving PCR amplification, while the rest of the stock solution was stored at -20°C.

2.10. RNA extraction

Total RNA was extracted from the fibroblast cells (number of cells) of both patients and controls using the RNeasy Mini kit (QIAGEN) as per manufacturer's guidelines. Prior to experimental work, the bench surface, gloved hands, pipettes, tubes, and containers were treated with RNAse Zap (Sigma-Aldrich). Briefly, frozen fibroblast pellets were thawed on ice and Buffer RLT was added to each tube to lyse the thawed cells. The suspension was mixed by vortex until cell clumps disappeared. The cells were passed through a 25G x 25mm hypodermic needle (Terumo) using a 1cc RNAse-free syringe (Terumo) to shear samples and facilitate effective lysis. Ethanol (70% v/v) (Lomb Scientific) was added to the lysate and mixed by pipetting up and down 10 times to allow RNA to bind to the column membrane in the following step. Contents of each tube was loaded on the RNeasy spin column and centrifuged for 15 s at 8000 g. The RNA captured by the column membrane was washed using the protocols for buffers supplied with the kit. The RNA was eluted from the column in 30µl of RNAse-free water. The purity and concentration of the extracted RNA was determined using the NanoDrop ND-1000 Spectrophotometer (Thermo-Fisher Scientific). RNA was stored at -80°C until further use.

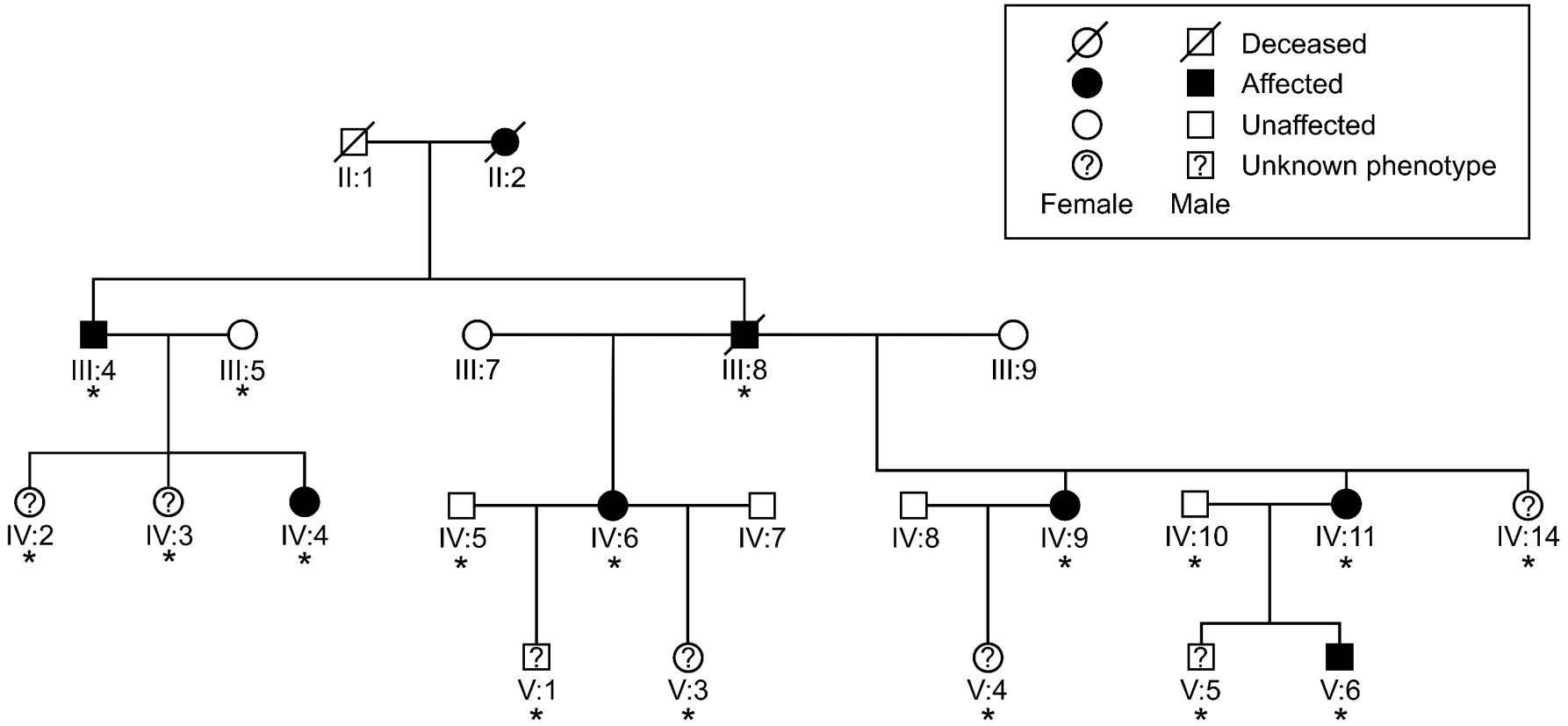
2.11. Complementary DNA synthesis

Total RNA was used to prepare reverse transcribed complementary DNA (cDNA) using the iScript cDNA Synthesis Kit (BioRad) according to the manufacturer's protocol. Briefly, a reaction mixture was prepared comprising of RNA template (1 µg), 5x iScript Reaction Mix (4 µL), 1 µl iScript Reverse Transcriptase and nuclease-free water (supplied in the kit) to a reaction volume to 20 µL. The cDNA synthesis was performed using a Mastercycler Pro (Eppendorf) and the thermal cycling program is shown (Table 2.5).

Table 2.5: Thermal cycling program used for cDNA synthesis.

Step	Number of cycles	Duration (min:s)	Temperature (°C)
Priming	1	5:00	25.0
Reverse transcription	1	20:00	46.0
Reverse transcriptase inactivation	1	1:00	95.0
End	1	Indefinite	4.0

2.12. Supplementary Material



Supplementary Figure 2.1: The pedigree of CMT720 showing the DNA samples from individuals that underwent whole genome amplification. The individuals marked with asterisks are genotyped and included in the analysis. Legend of symbols is provided in the box at the top right. This pedigree was adapted from the original publication by Kochanski et al. [168] and the individual identifiers were retained from the original pedigree.

Chapter 3

Fine mapping the suggestive linkage loci in CMT720

3.1. Introduction

A major challenge in modern positional cloning studies in CMT is that large families providing sufficient power for establishing significant linkage loci are often unavailable [104] while linkage analyses conducted on small families produce multiple false positive linkage loci with suggestive LOD scores [46, 172, 176]. Suggestive linkage loci however, can be highly useful for targeting DNA regions to search for mutations and reducing the number of candidate pathogenic coding variants identified by NGS [75, 172]. When considering the inclusion of all types of noncoding variants in our current investigation of CMT720, this still represents a large number of candidates to investigate across the 5 suggestive linkage regions. Since CMT720 is a monogenic disease, only one of these regions is expected to be the real disease locus, and pursuing the variants in the remaining loci adds unnecessary burden to analysis [96]. Therefore, eliminating the false positive linkage loci in family CMT720 remains a priority for maximizing the exclusion of variants with no pathogenic impact and facilitate a more efficient analysis.

Fine mapping linkage analysis has been frequently used for verifying or excluding linkage to the loci with suggestive LOD scores that result from the genome wide scan analyses [172, 177, 178]. In this method an informative set of markers are used to perform a second linkage analysis within the peaks of suggestive LOD scores to detect the recombination events

that may have been missed during the initial genome wide linkage analysis [177, 179]. Following fine mapping, the loci that do not maintain suggestive LOD scores are deprioritised or excluded, while the regions that produce similar or higher levels of LOD scores remain as candidate linkage loci [172, 177]. The biallelic SNP markers used in the original genome wide scan linkage analysis may have limited informativeness [86, 87], and it is possible recombination events that could exclude some of the false positive linkage loci in family CMT720 may have been missed.

Microsatellite markers have more than 10 alleles on average [180], and are often highly heterozygous in the population, which makes them more informative than SNPs as genotyping markers [86, 87]. Heterozygosity is a measure of marker informativeness that indicates the fraction of individuals in the population with a heterozygous genotype for a given marker [80, 181]. Heterozygous markers increase the chance of determining which paternal or maternal chromosome an allele originates from (i.e. determining phase of the segregation of alleles), and permits identification of recombination events [80, 181]. Microsatellite-based fine mapping linkage analysis has allowed eliminating false positive linkage regions identified by genome wide scans using SNP arrays in investigations of IPNs [182] and inherited neurological disorders [183]. Our lab has previously used this approach to eliminate 5 false positive linkage loci identified through a SNP-based genome wide scan linkage analysis in a small family with autosomal recessive CMT [172]. Prioritizing the variants in the two remaining candidate loci facilitated identification of the likely pathogenic SNV c.A118C in *AHNAK2* [172]. Additionally, smaller families can provide sufficient power for obtaining LOD scores below -2 [92], which can be used for the exclusion and refinement of the linkage loci obtained in CMT families of similar size and power to CMT720 [172]. Genotype information obtained from microsatellite markers during fine mapping is useful for constructing haplotypes and identifying key recombinant individuals that can define the flanking markers delimiting a linkage region [77,

184].

3.2. The strategy for eliminating the false positive linkage loci in CMT720

The aim of this chapter is to exclude the false positive linkage loci identified in CMT720 by performing fine mapping linkage analysis using microsatellite markers. Fine mapping linkage analysis will use microsatellite markers within the five candidate linkage regions identified in family CMT720. Parametric (model based) multipoint linkage analysis will be the choice of analysis for this study to provide the highest power for further the defining the linkage loci in CMT720 [185, 186]. Parametric linkage analysis models the disease trait including penetrance, mode of inheritance and the frequency of the disease allele [185, 186]. For multipoint linkage analysis, linkage between a modelled disease locus in a family is tested against a haplotype of markers, which incorporates patterns of marker segregation and recombination events between markers to increase the power to detect linkage [185, 186]. By performing fine mapping using this strategy, the exclusion of false positive linkage loci identified in CMT720 is expected. This will assist in eliminating a significant proportion of the variants and positional candidate genes and facilitate variant prioritisation. Additionally, the number of positional candidate genes within the remaining suggestive linkage regions will be feasible for manual analysis using a data visualisation software such as Integrative Genomics Viewer (IGV) [187]. This will also minimise the number of SVs and REs to consider that may impact coding sequences within the linkage region which were not addressed by previous investigations in family CMT720.

3.3. Materials and methods

3.3.1. Selection of microsatellite markers

The identity of microsatellite markers with a heterozygosity of 0.7 or higher were determined across the candidate linkage regions by referring to the Marshfield human genetic maps [188]. The heterozygosity of most markers selected for this study was above 0.75 to ensure that the marker informativeness was sufficient for detecting linkage in an autosomal dominant disease [84, 86, 89]. Additionally, we aimed to maintain intermarker distances between 5 cM (~5 Mb) to 10 cM (~10 Mb) across each suggestive linkage locus to obtain sufficient density for detecting linkage [189, 190] and a suitable number of markers that is amenable to genotyping within the timeframe of this project [191]. The identifiers and properties of all microsatellite markers selected for the current linkage analysis are summarised in Supplementary Table 3.1. The amplicon sizes and coordinates for each marker in the GRCh38 human genome reference [143] were determined by referring to the STS Markers database on Genome Browser [192].

3.3.2. Primer design and optimisation

Sequences of the forward and reverse oligonucleotide primer pairs targeting each marker (Supplementary Table 3.2), were provided to Sigma-Aldrich (USA) for primer manufacturing. The forward primers of each pair were conjugated to the fluorescein dye to tag DNA fragments and allow for determining the amplicon length using capillary electrophoresis. PCR reactions were optimised to determine the annealing temperature and mastermix combinations that yielded robust amplification with each primer pair using the thermal cycling program shown in Table 3.1 below. The optimal annealing temperature and mastermix (see

Section 2.6) combinations determined for each marker are summarised in Supplementary Table 3.3.

Step	Number of cycles	Duration (minutes:seconds)	Temperature (°C)
Heat activation of polymerase	1	1:00	95.0
Denaturation	35	1:00	95.0
Annealing		0:40	Variable
Extension		0:30	72.0
Final extension	1	2:00	72.0
End	-	-	20.0

Temperature (°C)
51.1
52.5
54.3
56.2
58.3
60.2
62.0
63.5

Table 3.1: Thermal cycling program used for optimisation of the primer pairs targeting each microsatellite marker. The gradient of annealing temperatures used during optimisation is provided in the adjacent detail table.

3.3.3. Genotyping microsatellite markers

17 family members were genotyped as indicated in Figure 3.1 following amplification of DNA samples through whole genome amplification as described in Chapter 2. Each microsatellite marker was amplified using the PCR protocols optimised for the primers targeting selected markers (Supplementary Table 3.1). PCR products were dried and sent to Garvan Institute of Medical Research (Australia) for size fractionation through capillary electrophoresis. The size marker LIZ600 was used. Marker alleles were identified by scoring the electrofluorogram peaks in the traces produced by capillary electrophoresis. The traces were visualised using Gene Marker software, version 1.60 (SoftGenetics) and fluorescent signals were scored to determine the sizes of each allele (Supplementary table 3.5). Cyrillic, version 2.1 (Cherwell Scientific Publishing) was used to construct a pedigree file in which marker genotypes for each individual and genealogical relationships between individuals were indicated. The microsatellite marker genotype for each individual, was re-coded as an integer corresponding to the number of different allele sizes genotyped for each marker (Supplementary figure 3.6). An affected-only method was implemented by coding the

phenotypes of all at risk individuals (IV:2, IV:3, IV:14, V:1, V:3 and V:5) as unknown. In addition, V:4 was coded as an unknown phenotype as clinicians were not able to re-assess this individual who initially had a very mild phenotype (personal communication of clinician Andrzej Kochanski). Tools within Cyrillic were used to export the pedigree and genotype data in the Linkage package file formats. The subprogram LINKMAP of the Linkage Package was used to perform the multipoint linkage analysis [90].

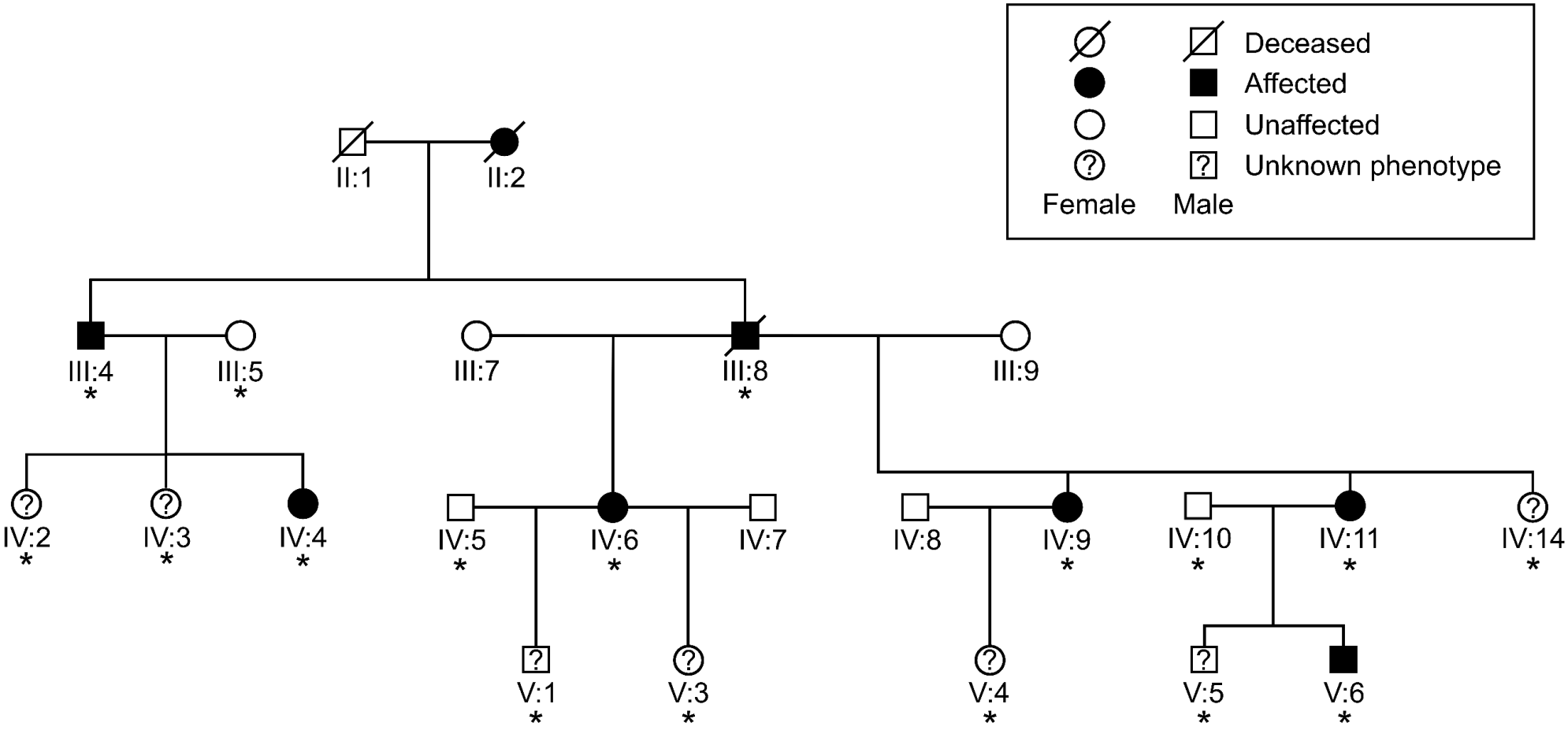


Figure 3.1: The pedigree used in the microsatellite-based linkage analysis performed on CMT720. The individuals marked with asterisks are genotyped and included in the analysis. Legend of symbols is provided in the box at the top right. This pedigree was adapted from the original publication by Kochanski *et al.* [168] and the individual identifiers were retained from the original pedigree.

3.3.4. Multipoint linkage analysis

Parametric multipoint linkage analysis was carried out using the MAKEPED, Linkage Control Program, LINKMAP and LINKAGE REPORT PROGRAM components of the LINKAGE software package [90] following the workflows described [82]. Initially, a .pre file containing the pedigree information of family CMT720 was exported from the Cyrillic program which contained genealogical relationships affection status and marker genotype data. The resulting pedigree file (.pre) was processed into a .ped file by the MAKEPED to assign genealogical pointers necessary for analysis by the subsequent programs. A data file (.dat file) containing the disease model and the marker loci being tested in the pedigree was constructed by the Cyrillic program as part of exporting file formats for the Linkage package. For the disease locus autosomal dominant inheritance was assumed with 90% penetrance and disease allele frequency of 0.0001. Equal allele frequencies (1/n where n = the number of alleles genotyped for each marker) were used for each microsatellite marker. Using the .dat (marker information) and .ped files (pedigree and genotype information) the Linkage Control Program prepared a batch file to run the LINKMAP program using the sliding multipoint analysis method. Recombination fractions (θ) between consecutive markers were calculated by converting the intermarker distances in centimorgans (cM) obtained from the Marshfield Genetic Map [188] using Haldane's mapping function shown below, where the variable "x" is substituted with intermarker distance:

$$\theta = \frac{1 - e^{-2x}}{2}$$

The recombination fractions were provided to LINKMAP. No sex difference in recombination fractions was assumed. 3-point linkage analysis was performed by testing for linkage between a haplotype of 2 markers and a hypothetical disease locus at 5 equal intervals between each marker to generate a likelihood curve plot (Figure 3.2). Each marker was covered by a "sliding haplotype" analysis to help reduce computational power needed to process high haplotype

numbers from multiallelic markers [90]. An example of the sliding method is shown for a multipoint analysis involving the disease and a map of 5 markers (Figure 3.2). The location scores used to plot a likelihood curve generated by the LINKMAP program (multipoint LOD score against marker location) were summarised using the LINKAGE REPORT PROGRAM. For the extended haplotype analysis, disease haplotypes were constructed based on the minimal number of marker recombinations, with the disease.

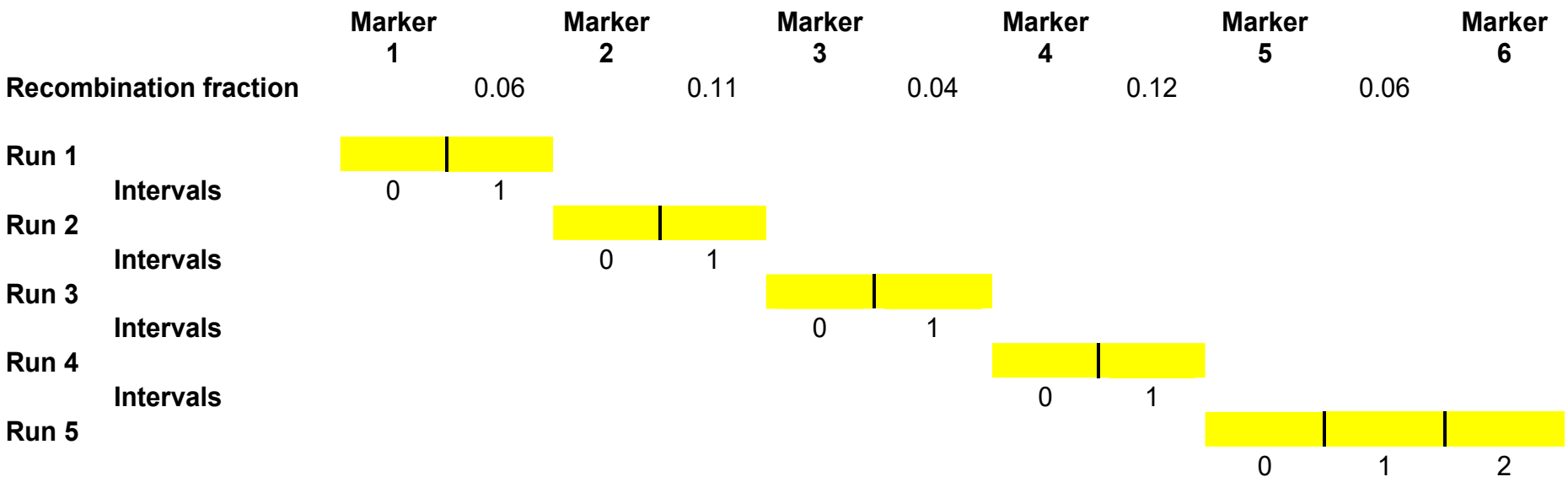


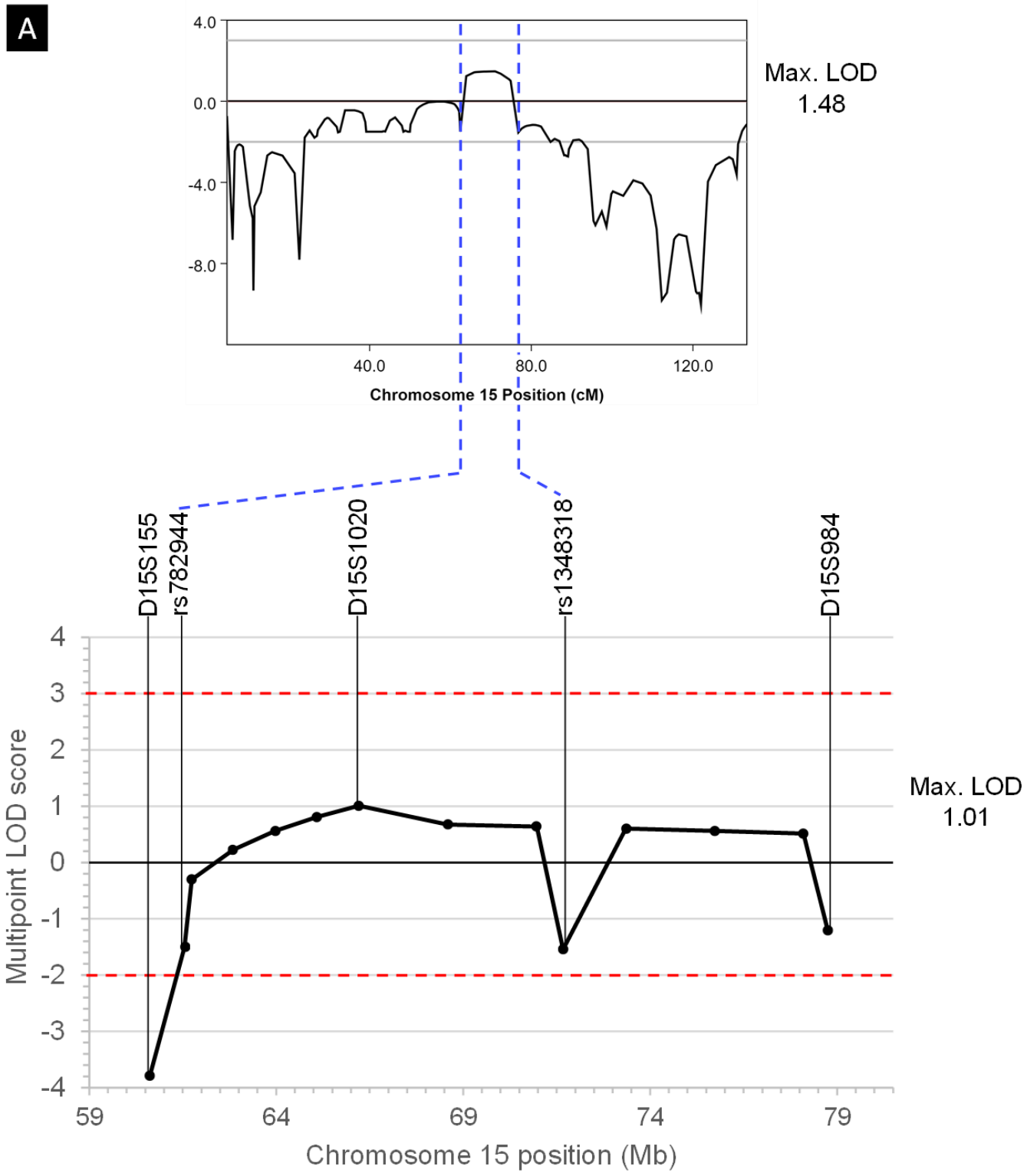
Figure 3.2: The sliding loci method to perform 3-point linkage analysis using LINKMAP sub-program from the LINKAGE software package. The recombination fraction between each marker is calculated from the map distances and specified for each run for a pair of markers. Interval 0 is initially calculated in each run to obtain the log likelihood at θ = 0.5 left of the leftmost marker. This information is used to test for linkage between the hypothetical disease locus and each pair of markers at 5 equal intervals across interval 1 in the subsequent step. This process is reiterated for each pair of markers until the entire locus of interest is covered. In the final run, linkage is lastly tested across interval 2 which is located right to the rightmost marker. The information provided here was summarised from the works of Lathrop *et al.*, and Ott and Terwilliger [82, 90].

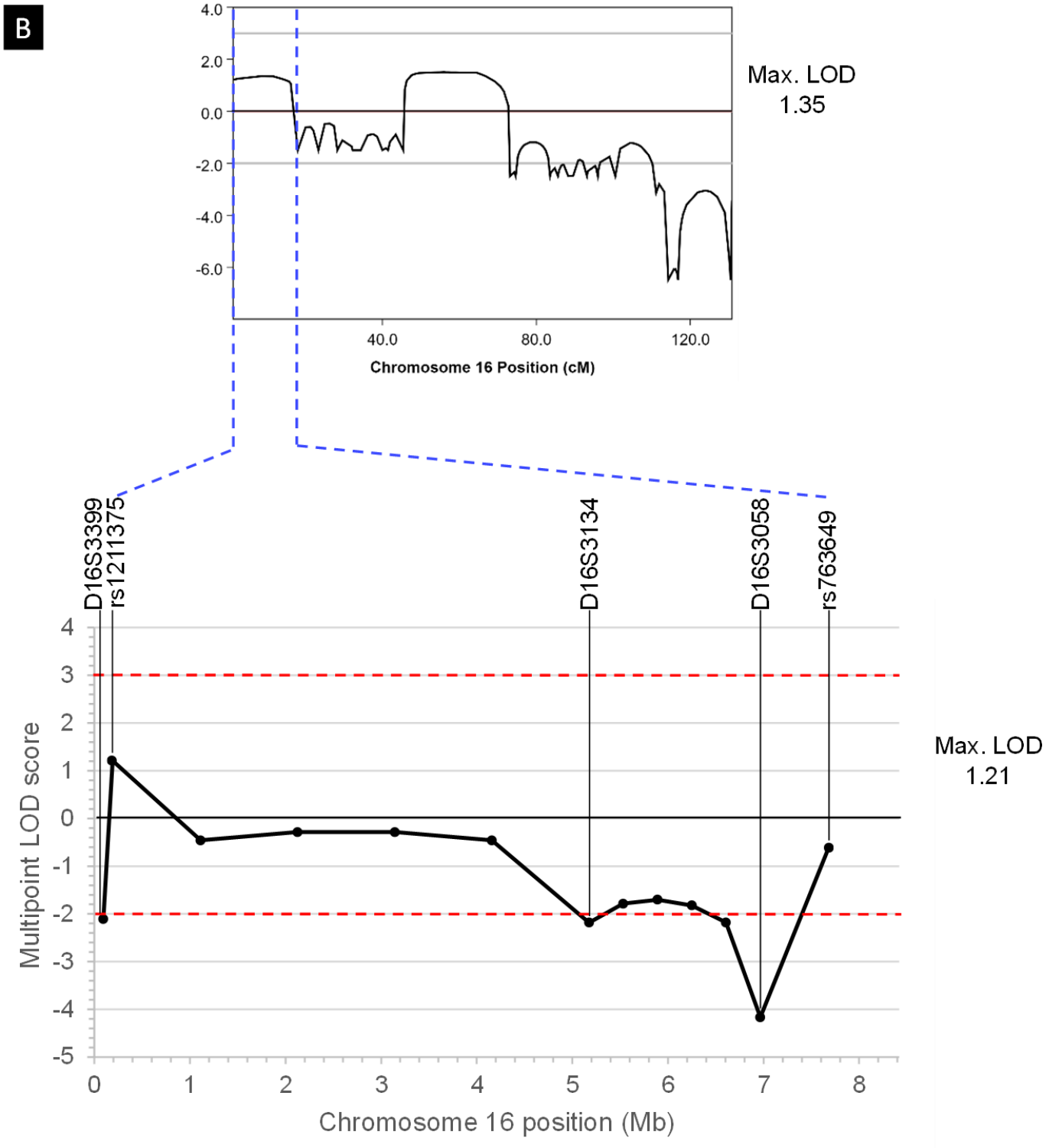
3.4. Results

3.4.1. Fine mapping excludes/deprioritises 3 candidate linkage peaks on chromosome 15, 16 and 18

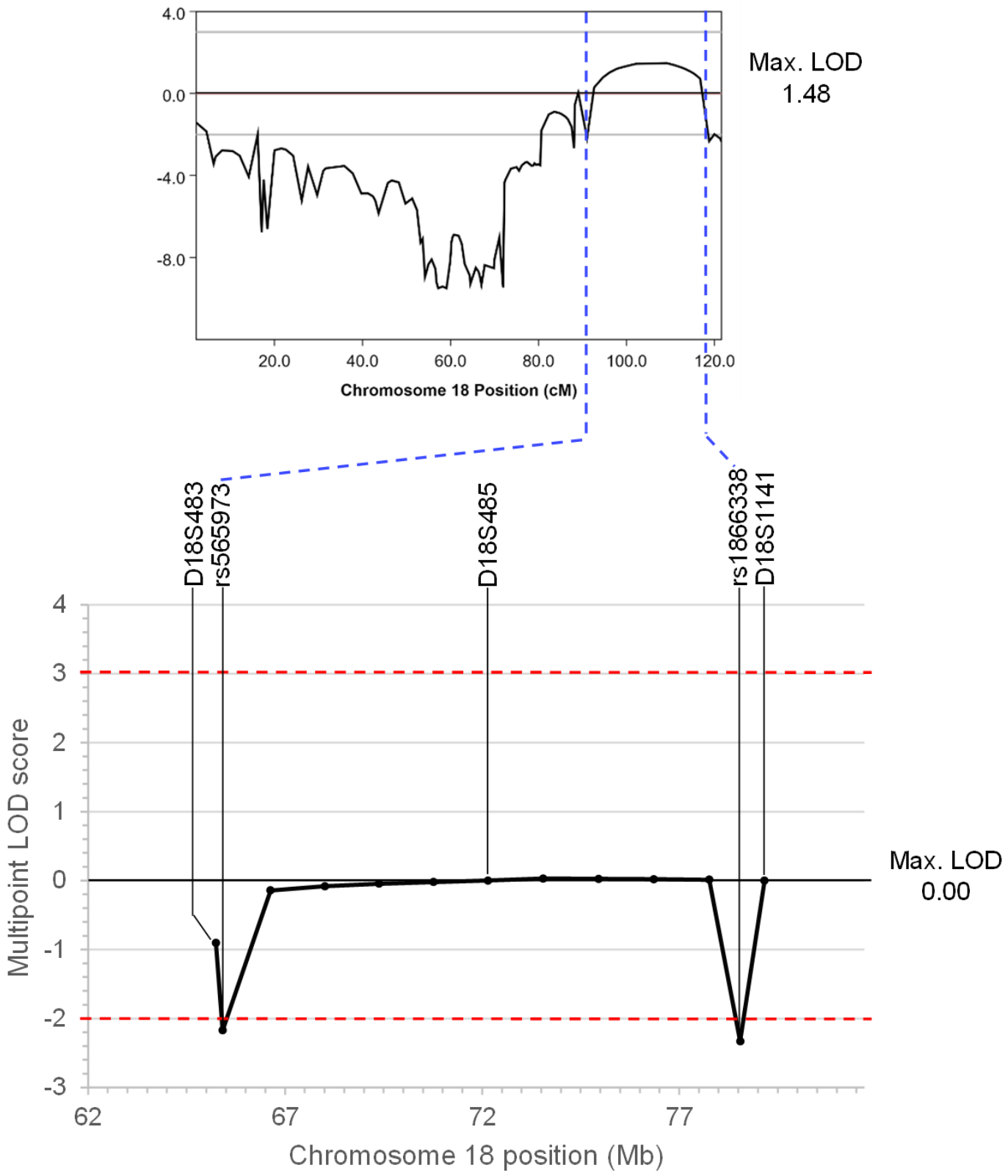
The suggestive linkage locus identified at the telomeric end of the p arm of chromosome 16 (peak 1) was excluded since multipoint linkage analysis produced LOD scores below -2 for all microsatellite markers (Figure 3.3B). The suggestive linkage locus on chromosomes 18 was deprioritised for further analysis since the multipoint LOD scores did not reach the LOD>1 threshold of suggestive linkage (Figure 3.3C). For chromosome 15, only D15S1020 reached the threshold of suggestive linkage (LOD 1.01), whereas all remaining markers gave negative LOD scores (Figure 3.3A). This region was also deprioritised for analysis. Overall, fine mapping with microsatellite markers definitively excluded one suggestive linkage region and de-prioritised two suggestive linkage loci for this study. This corresponded to removing a total of 29.99 Mb for analysis for this study.

Figure 3.3: The suggestive linkage regions excluded or deprioritised based on the results of the fine mapping linkage analysis. Multipoint LOD score likelihood curves for the microsatellite-based fine mapping linkage analysis (bottom panels) shown in relation to the LOD score likelihood curves of the SNP- based genome wide scan linkage analysis (top panels). The LOD score likelihood curves for the candidate linkage region on (**A**) chromosome 15, (**B**) chromosome 16 peak 1 and (**C**) chromosome 18 do not support the suggestive linkage identified by the SNP-based genome wide scan linkage analysis. Blue dashed lines represent the position of the SNP markers demarcating the boundaries of the candidate linkage regions determined by the genome wide scan linkage analysis.


A
4.0
0.0
-4.0
-8.0
40.0
80.0
120.0
Chromosome 15 Position (cM)
Max. LOD
1.48
D15S155
rs782944
D15S1020
rs1348318
D15S984
Multipoint LOD score
4
3
2
1
0
-1
-2
-3
-4
59
64
69
74
79
Chromosome 15 position (Mb)
Max. LOD
1.01


B
4.0
2.0
0.0
-2.0
-4.0
-6.0
40.0
80.0
120.0
Chromosome 16 Position (cM)
Max. LOD
1.35
D16S3399
rs1211375
D16S3134
D16S3058
rs763649
Multipoint LOD score
4
3
2
1
0
-1
-2
-3
-4
-5
0
1
2
3
4
5
6
7
8
Chromosome 16 position (Mb)
Max. LOD
1.21

C


4.0
0.0
-4.0
-8.0
20.0
40.0
60.0
80.0
100.0
120.0
Chromosome 18 Position (cM)
Max. LOD
1.48
D18S483
rs565973
D18S485
rs1866338
D18S1141
Multipoint LOD score
4
3
2
1
0
-1
-2
-3
62
67
72
77
Chromosome 18 position (Mb)
Max. LOD
0.00

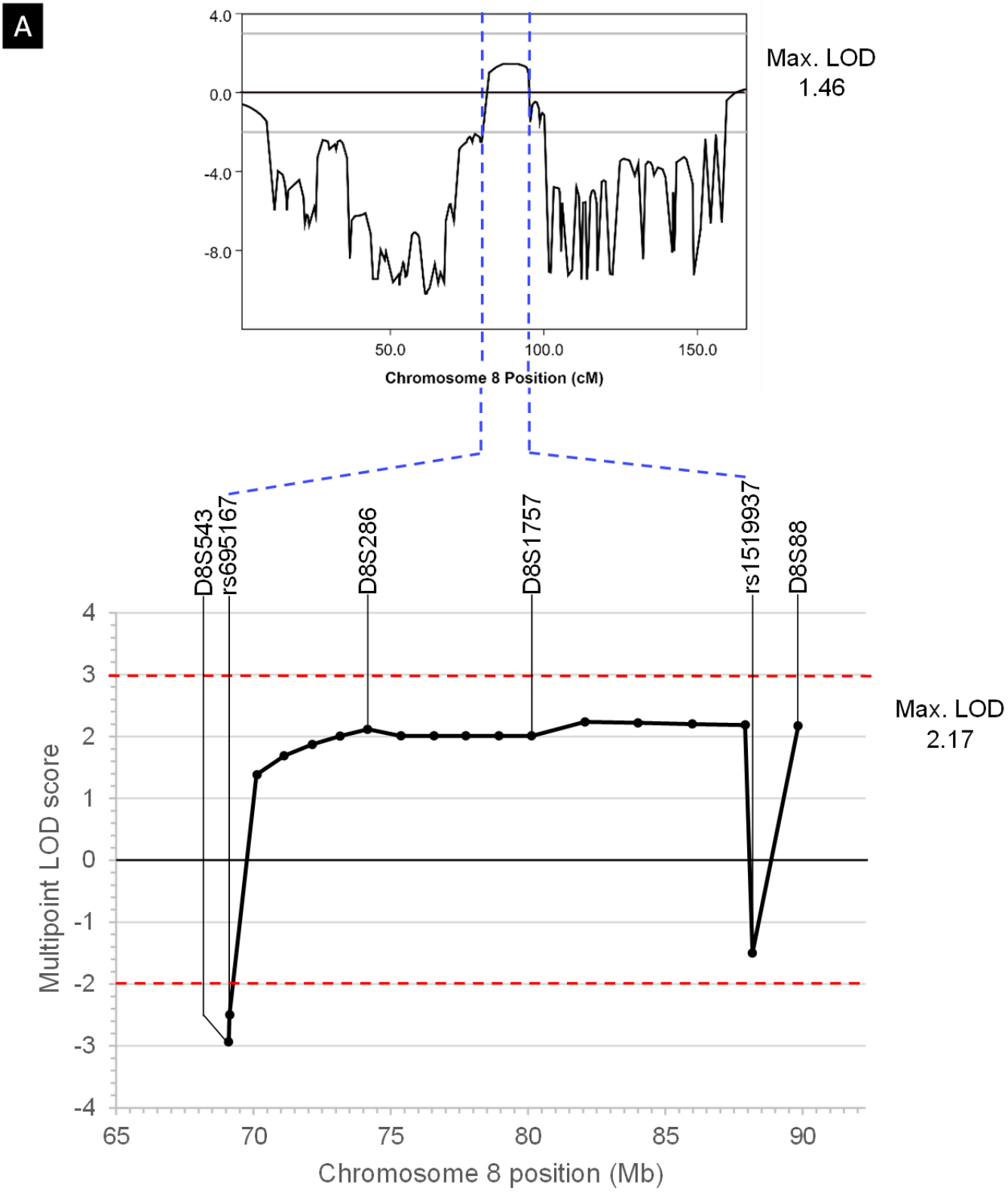
3.4.2. Two suggestive linkage regions on chromosomes 8 and 16 remain as candidate loci for CMT720

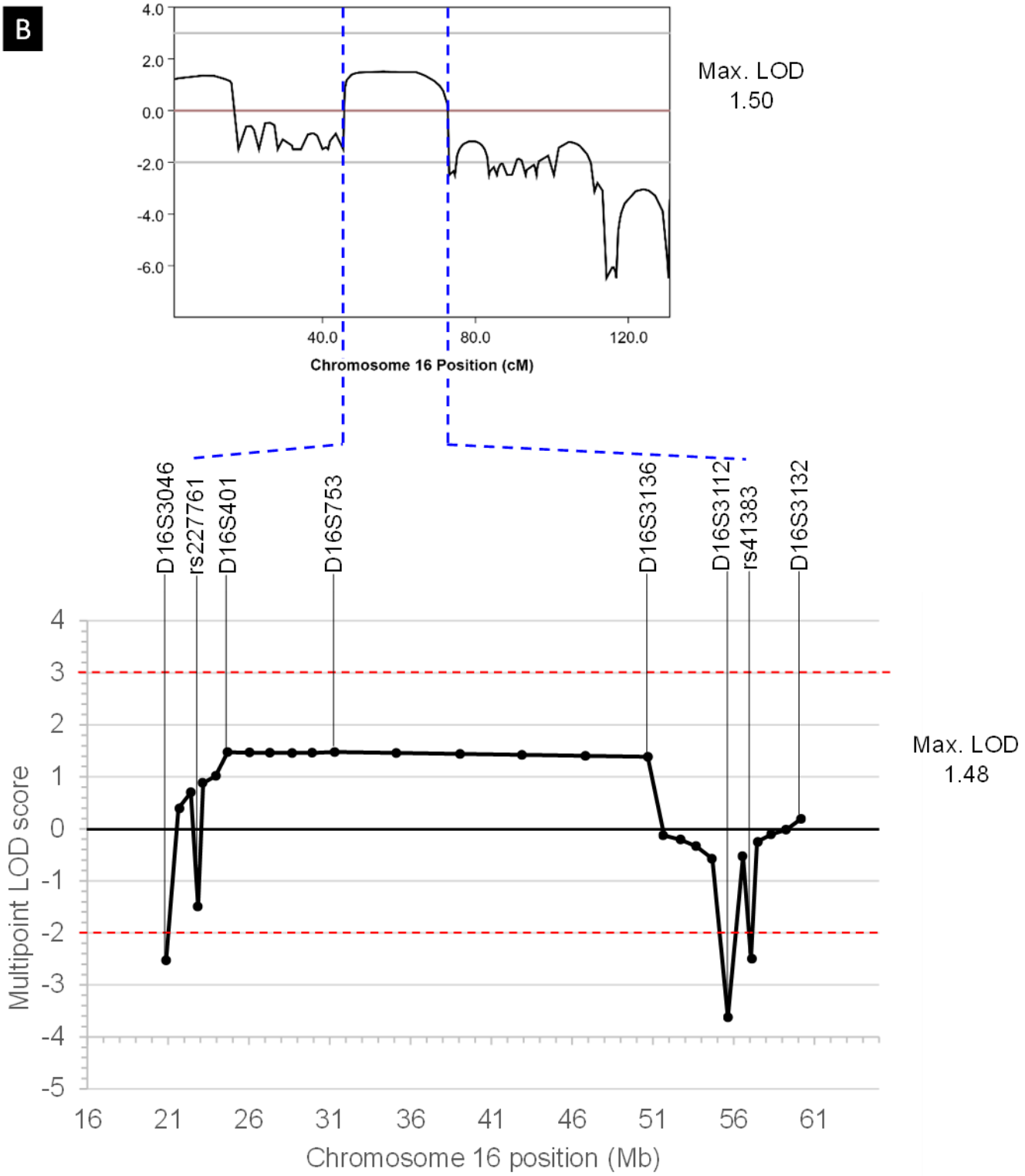
Multipoint linkage analysis with microsatellite markers produced a maximum LOD score of 2.23 in the candidate linkage peak on chromosome 8q13.2-q21.3, with the LOD scores for the markers D8S286 (LOD 2.12) and D8S1757 (LOD 2.01) remaining above the threshold of suggestive linkage (Figure 3.4A). These findings indicate that the fine mapping linkage analysis with microsatellite markers continues to support the suggestive linkage on chromosome 8 identified by the genome wide scan linkage analysis using SNP markers. Multipoint linkage analysis produced a negative LOD score of -2.93 at the marker D8S543 (Figure 3.4A, bottom panel), indicating the marker has recombined with the candidate disease locus. D8S543 is located centromeric to the proximal flanking marker rs695167 previously identified through genome wide scan linkage analysis, therefore, the proximal marker remains as rs695167 (Figure 3.4A, bottom panel). D8S88 produced a LOD score of 2.17 and is located telomeric to the previously established distal boundary demarcated by rs1519937. Since rs1519937 has recombined with the candidate disease locus as indicated by a LOD score of -1.50 it remains as the distal marker (Figure 3.4A, bottom panel).

Multipoint linkage analysis with microsatellite markers spanning chromosome 16 peak 2, produced a maximum LOD score of 1.48 at the markers D16S401 and D16S753 in candidate linkage peak on chromosome 16p12.3-q13 (Figure 3.4B, bottom panel). These findings indicate that the fine mapping linkage analysis with microsatellite markers continues to support the suggestive linkage on chromosome 16 identified by the genome wide scan linkage analysis using SNP markers. Marker D16S3046 shows a LOD score of -2.53 and is located telomeric to the distal flanking marker rs208965 on the p arm of chromosome 16 previously identified through genome wide scan linkage analysis (Figure 3.4B, bottom panel). Accordingly, rs208965 remains as the distal flanking marker defining the linkage interval at chromosome 16p12.3-q13. The microsatellite marker D16S3112 on the q arm of chromosome

16 shows a LOD score of -3.40 and is located centromeric to the distal flanking marker rs41383 previously identified through genome wide scan linkage analysis (Figure 3.4B, bottom panel). This indicates D16S3112 is the new distal marker on the q arm portion of this suggestive linkage region and excludes a 1.44 Mb region between the markers D16S3112 and rs41383.

Figure 3.4: The suggestive linkage regions validated based on the results of the fine mapping linkage analysis. Detailed view of multipoint LOD score likelihood curves for the microsatellite-based fine mapping linkage analysis (bottom panels) shown in relation to the LOD score likelihood curves of the SNP- based genome wide scan linkage analysis (top panels). The LOD score likelihood curves for the candidate linkage regions on (**A**) chromosome 8 and (**B**) chromosome 16 peak 2 support the suggestive linkage identified by the SNP-based genome wide scan linkage analysis. Blue dashed lines represent the position of the SNP markers demarcating the boundaries of the candidate linkage regions determined by the genome wide scan affected only linkage analysis.


A
4.0
0.0
-4.0
-8.0
50.0
100.0
150.0
Chromosome 8 Position (cM)
Max. LOD
1.46
D8S543
rs695167
D8S286
D8S1757
rs1519937
D8S88
4
3
2
1
0
-1
-2
-3
-4
Multipoint LOD score
65
70
75
80
85
90
Chromosome 8 position (Mb)
Max. LOD
2.17


B
4.0
2.0
0.0
-2.0
-4.0
-6.0
40.0
80.0
120.0
Chromosome 16 Position (cM)
Max. LOD
1.50
D16S3046
rs227761
D16S401
D16S753
D16S3136
D16S3112
rs41383
D16S3132
Multipoint LOD score
4
3
2
1
0
-1
-2
-3
-4
-5
16
21
26
31
36
41
46
51
56
61
Chromosome 16 position (Mb)
Max. LOD
1.48

3.4.3. Haplotype analysis of chromosome 16p12.3-q13 and chromosome 8 q13.2-q21.3 suggestive linkage regions

3.4.3.1. Haplotype analysis of chromosome 8q13.2-q21.3 confirms the previously established flaking markers

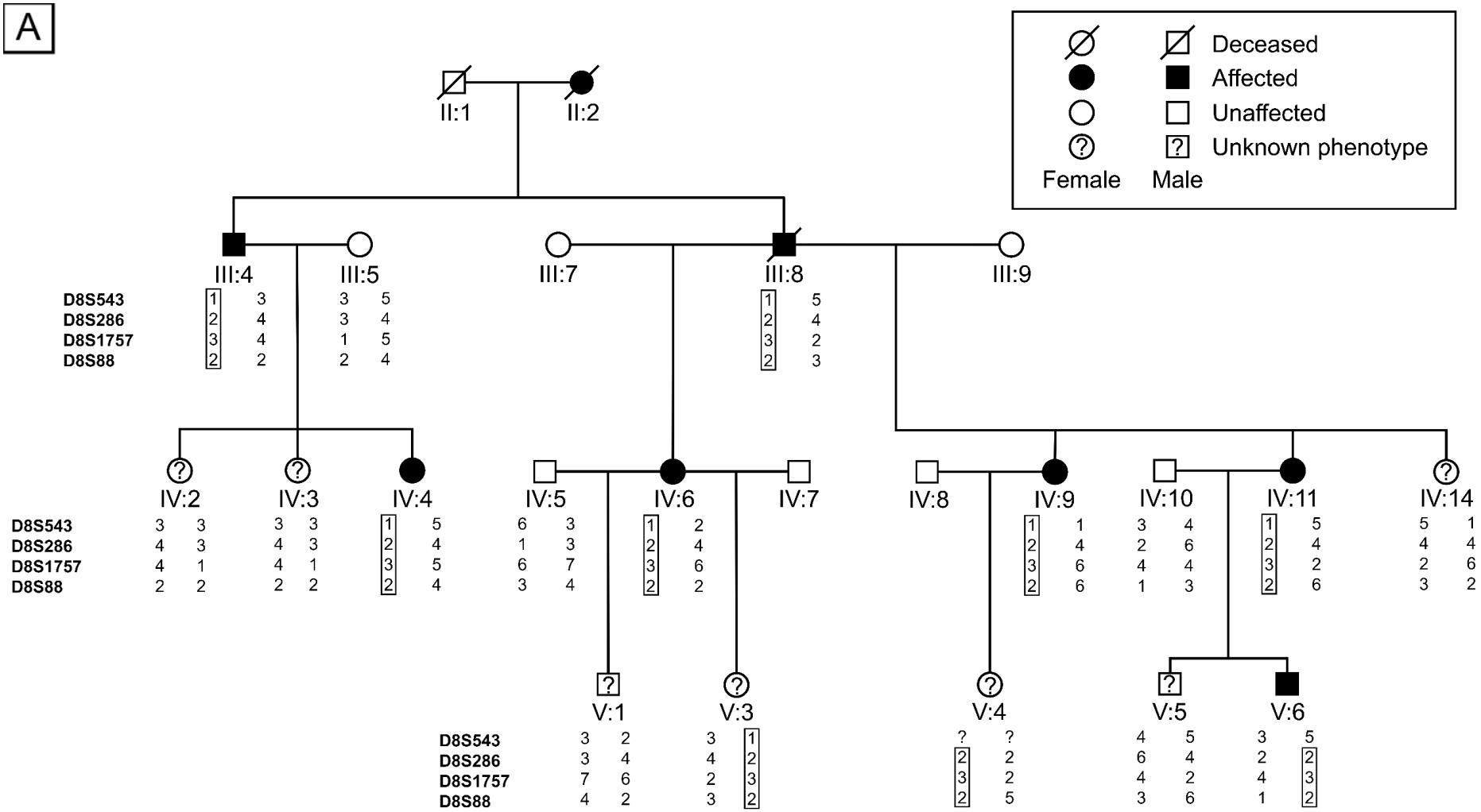
Extended haplotypes of individuals were constructed according to the Marshfield genetic map [188] based on minimizing intermarker recombination. Haplotype analysis detected no recombination between the suggestive disease locus on chromosome 8 q13.2-q21.3 and the markers D8S286, D8S1757 and D8S88. This haplotype was also observed in an individual coded as “unknown” phenotype due to the age of the individual (V:3). This analysis shows recombination between markers D8S543 and D8S286 in a single affected family member (V:6) (Figure 3.5A). Haplotype analysis did not detect a distal recombination site, however, revealed that D8S88 is homozygous in two affected individuals (III:4 and IV:6), making it uninformative for observing the recombination events in 4 unaffected (IV:2, IV:3, V:1 and V:3) and one affected family member (IV:4) (Figure 3.5A). This finding suggests that D8S88 is not highly informative, which can potentially explain the LOD score of 2.17 obtained from this marker despite the LOD score of -1.50 obtained from the distal flanking marker rs1519937.

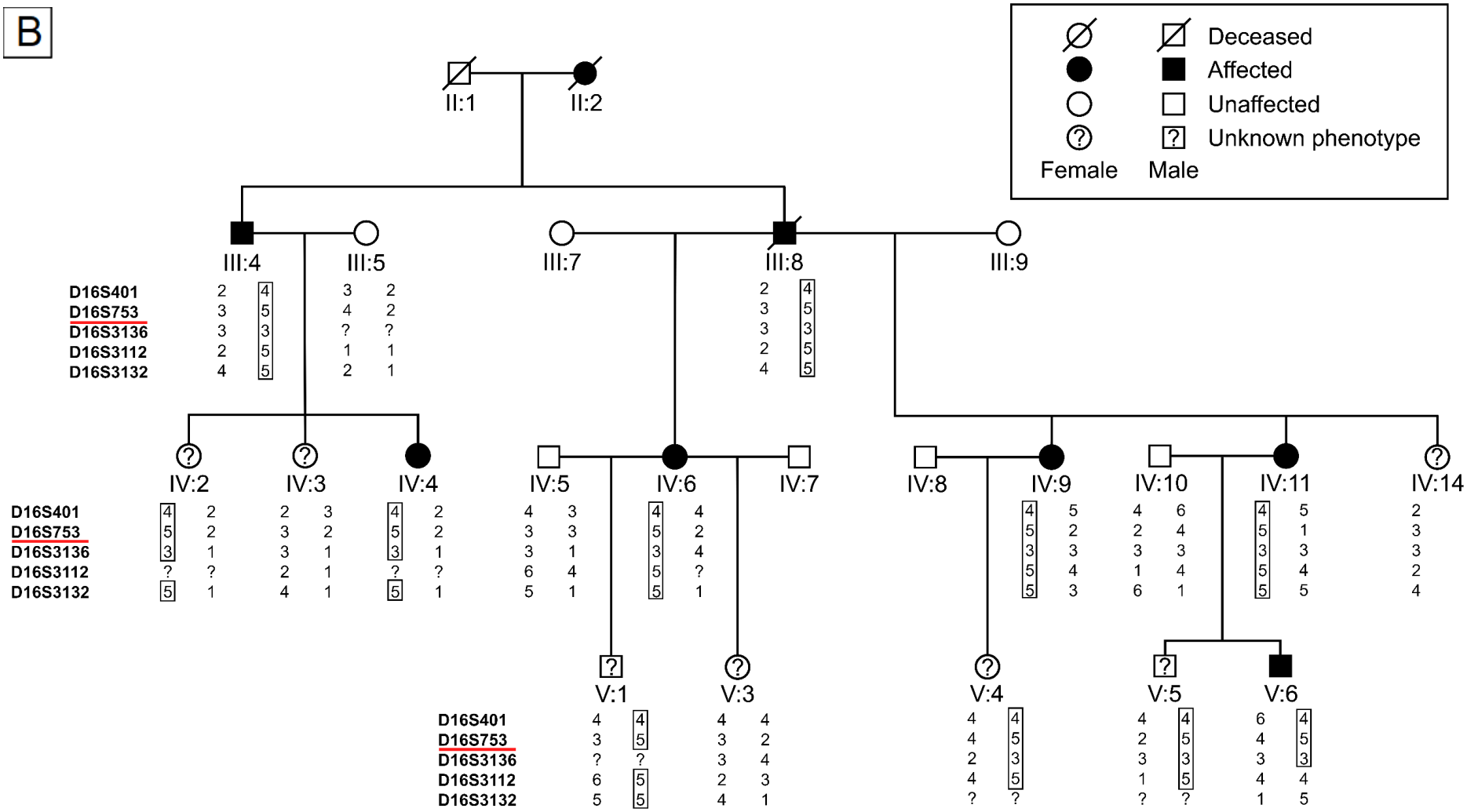
3.4.3.2. Haplotype analysis of chromosome 16 16p12.3-q13 excludes a 1.44 Mb region

Extended haplotype analysis showed no recombination between the modelled disease locus and the markers D16S753, D16S3136 and D16S3112 localising to chromosome 16p12.3-q13. This haplotype was also observed in 2 unaffected individuals (IV:2 and V:1) (Figure 3.5B). Since rs208965 remains as the distal marker on the p arm of chromosome 16, D16S3046 was not included in the haplotype analysis. Based on the analysis, the distal recombination site at the q arm of chromosome 16 is between D16S3136 and D16S3112.

These findings support D16S3112 as the distal demarcating marker on the p arm of chromosome 16 and allow us to exclude a 1.44 Mb region centromeric to the previously established distal marker rs41383.

Figure 3.5: Haplotype analysis of microsatellite markers from the candidate linkage loci on chromosome 8q13.2-q21.3 and chromosome 16p12.3- q13 in family CMT720. The haplotype segregating with the disease is boxed. Marker alleles are provided below each individual. Individuals are numbered consecutively in each generation, from left to right. The legend of pedigree symbols is provided in the box at the top right. (**A**) The haplotypes of microsatellite markers from the candidate linkage peak on chromosome 8q13.2-q21.3. The markers are presented in order of centromere (top) to telomere (bottom). (**B**) The haplotypes of microsatellite markers from the candidate linkage peak on chromosome 16p12.3-q13. The markers are presented in order of telomere (top) to centromere (bottom) on the p arm of chromosome 16, and centromere (top) to telomere (bottom) on the q arm. The position of the centromere relative to the markers is indicated by the horizontal red line. These pedigrees were adapted from the original publication by Kochanski *et al.* [168] and the individual identifiers were retained from the original pedigree.


A
Deceased
Affected
Unaffected
Unknown phenotype
Female
Male
II:1
II:2
III:4
III:5
III:7
III:8
III:9
D8S543
D8S286
D8S1757
D8S88
IV:2
IV:3
IV:4
IV:5
IV:6
IV:7
IV:8
IV:9
IV:10
IV:11
IV:14
V:1
V:3
V:4
V:5
V:6


B
Deceased
Affected
Unaffected
Unknown phenotype
Female
Male
II:1
II:2
III:4
III:5
III:7
III:8
III:9
IV:2
IV:3
IV:4
IV:5
IV:6
IV:7
IV:8
IV:9
IV:10
IV:11
IV:14
V:1
V:3
V:4
V:5
V:6
D16S401
D16S753
D16S3136
D16S3112
D16S3132

3.5. Discussion

In this chapter, we performed fine mapping linkage analysis using microsatellite markers across the 5 suggestive linkage peaks identified for CMT720 by an initial genome-wide scan linkage analysis with SNP markers. The goal of this experiment was to exclude false positive linkage loci.

Our current linkage analysis has allowed us to deprioritise 2 candidate linkage loci on chromosomes 15 and 18 as likely false positives since LOD scores in these regions dropped below the threshold of suggestive linkage. In contrast, for linkage peak 1 on chromosome 16, the multipoint LOD scores were -2 or less which confirmed exclusion of this region of DNA as candidate for the disease locus in CMT720.

Fine mapping supported the 2 candidate linkage peaks on chromosomes 8 and 16 that were identified in the initial genome wide scan. The linkage region on chromosome 8 increases the maximum LOD score to 2.19, indicating this region of DNA should still be considered for causative mutations [177]. For chromosome 16 peak 2, the maximum LOD score was 1.48, indicating that it still supports suggestive linkage and should also be considered as a candidate disease locus for causative mutations in CMT720. By doing the “affecteds only” analysis, the extended haplotype analysis has shown selected at-risk individuals carrying the disease haplotypes for both the chromosome 8 and 16 suggestive linkage regions. With the continuing effort to map genes for unsolved IPN families it is becoming apparent that for small families the presence of reduced penetrance can confound a linkage analysis. A recent study investigating a recurrent *ITPR3* mutation (p.Thr1424Met) in families demonstrated a high level of clinical variability of disease severity [193]. Some individuals in families had very mild phenotypes and were considered normal on clinical examination [193]. It is therefore important to undertake an “affected only” analysis for these families to avoid non-penetrant individuals appearing to represent a recombination between the disease and a putative linkage region.

Since our fine mapping LOD scores and haplotype analysis did not indicate new proximal or distal recombination sites for the prioritised linkage region on chromosome 8, the flanking markers remain as rs695167 and rs1519937 respectively, corresponding to an interval that spans 19.04 Mb. On chromosome 16, a crossover in individual V:6 at marker D16S3112 has allowed us to refine the prioritised suggestive linkage region from 34.23 Mb to a 32.8 Mb interval between the flanking markers rs41383 and D16S3112. The combined prioritised candidate linkage regions cover a total of 51.83 Mb, indicating that only 1.7% of the genome will be targeted in the subsequent experiments for investigating CMT720 in this study. By querying the gene annotations on MANE project [194] using Table Browser [192], we have identified 66 and 196 genes within the prioritised linkage regions on chromosome 8q13.2-q21.3 and 16p12.3-q13, which implicates a total of 262 positional candidate genes for the current investigation.

The main weakness of the current linkage study in CMT720 in is the small sample size of family members available that does not provide sufficient power for identifying a locus with significant linkage [104]. CMT720 did not have the power to obtain a multipoint LOD>3 across any of the suggestive loci investigated in our current microsatellite-based analysis, instead yielding loci with suggestive LOD scores. Further prioritisation and reduced the analytical burden for this study would have been helpful if a single linkage region was prioritised. Instead, there will be two regions to consider.

Linkage studies performed by Zimoń *et al.* identified pathogenic dominant *GDAP1* mutations that segregate in both affected and unaffected individuals across families with different genetic backgrounds [195]. Interestingly, our suggestive linkage peak in chromosome 8 harbours *GDAP1*, however all coding mutations were excluded. It will be important to screen this known CMT2 gene for SVs and REs, and intronic variants that could impact GDAP1 as this was not done in previous investigations of CMT720.

Nonetheless, small families that are not permissive for producing significant evidence of linkage can provide LOD scores below -2 for excluding linkage [92, 172]. By using more informative microsatellite markers, we were able to detect recombination events that were missed by using SNP markers, allowing us exclude the 1.44 Mb region in the candidate linkage region on chromosome 16. While they are more informative than SNPs, it is possible to observe microsatellite markers with reduced informativeness, such as D8S88 from the prioritised candidate linkage peak on chromosome 8, which did not allow us to confidently determine a distal recombination site in our haplotype analysis. To address this, we may run additional microsatellite markers flanking D8S88 in the future to detect the recombination events at the distal end of this suggestive linkage region missed in the current fine mapping analysis.

In the Chapter 4, the functional data required to prioritise the noncoding variants will be generated by performing transcriptomic analysis using CMT720 patient-derived tissue to identify dysregulated positional candidate genes or aberrant transcripts (splicing or gene fusions) that can guide the selection and prioritisation of candidate DNA variants within the linkage regions on chromosome 8 and 16.

3.6. Supplementary Material

Supplementary Table 3.1: Properties of selected microsatellite markers.

Marker ID	Heterozygosity	Size (bp)	Chromosome	Coordinates (start-end)	Position (cM)	Interval (cM)
D8S543	0.74	116-140	8	69100616-69100998	87.54	7.07
D8S286	0.81	220-238	8	74169349-74169690	94.61	2.67
D8S1757	0.80	266-292	8	80127509-80127923	97.28	5.34
D8S88	0.82	76-86	8	89836549-89836796	102.62	N/A
D15S155	0.73	237-269	15	60120846-60121173	52.33	8.95
D15S1020	0.86	211-231	15	65706243-65706563	61.28	12.24
D15S984	0.92	204-256	15	77604525-77604847	73.52	N/A
D16S3399	0.77	180	16	95247-95429	0.00	7.61
D16S3134	0.75	161-174	16	5174460-5174689	7.61	5.51
D16S3058	0.74	112-140	16	6965081-6965415	13.12	N/A
D16S3046	0.74	84-108	16	20875076-20875388	40.65	6.29
D16S401	0.77	166-180	16	24674636-24674873	46.94	10.85
D16S753	0.79	252-276	16	31262128-31262471	57.79	4.32
D16S3136	0.70	175-211	16	50672322-50672571	62.11	11.80
D16S3112	0.84	267-281	16	55636639-55636997	73.91	6.09
D16S3132	0.72	223-241	16	60162486-60162849	80.00	N/A
D18S483	0.79	197-225	18	65247639-65247916	99.04	7.77
D18S485	0.79	176-190	18	72154033-72154441	106.81	17.30
D18S1141	0.76	263-293	18	79163659-79163991	124.11	N/A

Supplementary Table 3.2: Sequences of the primers used to amplify the microsatellite markers.

Marker ID	Sequence of forward primer (5'→3')	Sequence of reverse primer (3'→5')
D8S543	TGGTGTCATTGCTTTCTAGTCT	TGCACAGGTGAGTAAATTTGTAA
D8S286	GCTGTTTATTTGCCCATGT	GCATGAAACTGTCACTGAGA
D8S1757	ATGGAGCACTGCCAAGAA	TTTGAGCCCTATTTTTGAGAGA
D8S88	TCCAGCAGAGAAAGGGTTAT	GGCAAAGAGAACTCATCAGA
D15S155	TTTTCTAGGCAGGTAGTCCCA	GATTTCCATAGCACACATTTGAGT
D15S1020	TGCACAATGGATACTAAACAGC	CGATAGAGCAAGACTGTCTCAA
D15S984	GCAGACACGCTCGCAT	GAGGCTCCGAGGGCAG
D16S3399	ACCTAGATCCCTCCAGGTTT	GGGCCATTATTCAGCCAATC
D16S3134	CTGGGAAATTCTGGGA	GGCCAAGGTGTTTGTT
D16S3058	CACTACAGCCTGGGAAACA	CAGGACTAGAATGACCAAACATAA
D16S3046	CCCAGAATAAACTGCGTG	TTCATGGACCCCCTATTG
D16S401	TTCTCTTACAACACTGCCCC	ATTTGGATGGCTTGACAGAG
D16S753	CAGGCTGAATGACAGAACAA	ATTGAAAACAACTCCGTCCA
D16S3136	ATTGCCCTCAAGAACAGC	GTGCTATGCCATCCCAG
D16S3112	TACTTTGGAGCCCGAGG	AGCCCCCAGTGGTGTTTAT
D16S3132	ATGCTTTGTGGGCTGT	CCTTGGTTAATGTATTTGGA
D18S483	TTCTGCACAATTTCAATAGATTC	GAACTGAGCAAACGAGTATGA
D18S485	CCACATGAGGATATGGTGAG	GCCCCTATTATGAAGTATTAAAAG
D18S1141	TCTTTTGACAAATAACCCC	GGACAGTGCGAGACCT

Supplementary Table 3.3: Optimised PCR conditions for selected microsatellite markers.

Marker ID	Annealing temperature (°C)	Mastermix
D8S543	60.0	ImmoMix™ Red 2x
D8S286	60.0	ImmoMix™ Red 2x
D8S1757	60.0	ImmoMix™ Red 2x
D8S88	60.0	MyTaq™ HS Red Mix 2x
D15S155	60.0	ImmoMix™ Red 2x
D15S1020	56.2	MyTaq™ HS Red Mix 2x
D15S984	60.0	ImmoMix™ Red 2x
D16S3399	60.0	MyTaq™ HS Red Mix 2x
D16S3134	52.5	MyTaq™ HS Red Mix 2x
D16S3058	60.0	ImmoMix™ Red 2x
D16S3046	56.2	MyTaq™ HS Red Mix 2x
D16S401	60.0	ImmoMix™ Red 2x
D16S753	60.0	ImmoMix™ Red 2x
D16S3136	60.0	MyTaq™ HS Red Mix 2x
D16S3112	60.0	ImmoMix™ Red 2x

D16S3132	56.2	MyTaq™ HS Red Mix 2x
D18S483	60.0	MyTaq™ HS Red Mix 2x
D18S485	60.0	MyTaq™ HS Red Mix 2x
D18S1141	60.0	ImmoMix™ Red 2x

Supplementary Table 3.4: Pedigree information files for suggestive linkage loci.

The pedigree information file (.pre files) of CMT720 containing 6 columns of information is shown below. The file was created by tabulating the output of Cyrillic using Microsoft Excel. The “Individual” and “Genotyped” columns in the adjunct table indicate the identity of each individual as shown in the pedigree of family CMT720 and whether and individual was genotyped, respectively, and were not included in the original file.

Legend

1) FID: Family identifier

2) IID: Individual identifier

3) PID: IID of the father of individual (0 if individual is a founder)

4) MID: IID of the mother of individual (0 if individual is a founder)

5) Sex: Sex of individual (male=1, female=2)

6) Aff: Affection status, (0=unknown, 1=unaffected, 2=affected)

FID	IID	PID	MID	Sex	Aff.	Individual	Genotyped
CMT720	1	0	0	1	1	II:1	No
CMT720	2	0	0	2	2	II:2	No
CMT720	3	1	2	1	2	III:4	Yes
CMT720	4	0	0	2	1	III:5	Yes
CMT720	5	1	2	1	2	III:8	Yes
CMT720	6	0	0	2	1	III:7	No
CMT720	7	0	0	2	1	III:9	No

CMT720	8	3	4	2	0	IV:2	Yes
CMT720	9	3	4	2	0	IV:3	Yes
CMT720	10	3	4	2	2	IV:4	Yes
CMT720	11	0	0	1	1	IV:5	Yes
CMT720	12	5	6	2	2	IV:6	Yes
CMT720	13	0	0	1	1	IV:7	No
CMT720	14	0	0	1	1	IV:8	No
CMT720	15	5	7	2	2	IV:9	Yes
CMT720	16	0	0	1	1	IV:10	Yes
CMT720	17	5	7	2	2	IV:11	Yes
CMT720	18	5	7	2	0	IV:14	Yes
CMT720	19	11	12	1	0	V:1	Yes
CMT720	20	13	12	2	0	V:3	Yes
CMT720	21	14	15	2	0	V:4	Yes
CMT720	22	16	17	1	0	V:5	Yes
CMT720	23	16	17	1	2	V:6	Yes

Supplementary Table 3.5: Sizes of microsatellite marker alleles.

Individual	D8S543	D8S286	D8S1757	D8S88	D15S155	D15S1020	D15S984	D16S3399	D16S3134	D16S3058
II:1	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
II:2	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
III:4	110 126	222 230	270 276	80 80	257 259	213 219	212 214	178 180	161 162	106 128
III:5	126 132	228 230	260 278	80 84	257 265	217 225	206 212	170 176	161 164	124 126
III:8	110 132	222 230	268 270	80 82	257 263	213 221	212 214	172 180	161 161	106 106
III:7	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
III:9	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:2	126 126	228 230	260 276	80 80	259 265	217 219	206 212	0 0	161 164	124 128
IV:3	126 126	228 230	260 276	80 80	257 257	219 225	206 212	0 0	161 161	106 126
IV:4	110 132	222 230	270 278	80 84	259 265	213 217	212 214	176 178	162 164	124 128
IV:5	126 134	218 228	280 284	82 84	259 261	211 215	0 0	174 178	160 161	106 114
IV:6	110 120	222 230	270 280	80 80	259 263	213 221	212 214	172 172	159 161	106 106
IV:7	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:8	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:9	110 110	222 230	270 280	80 88	263 265	213 215	212 214	172 172	159 161	106 106
IV:10	126 130	222 236	276 276	76 82	257 265	209 217	200 230	176 178	159 161	122 126
IV:11	110 132	222 230	268 270	80 88	257 263	213 219	212 214	172 172	159 161	106 106
IV:14	110 132	230 230	268 280	80 82	257 263	213 215	212 214	178 180	159 161	106 106
V:1	120 126	228 230	280 284	80 84	259 263	211 213	212 220	172 174	159 160	106 106
V:3	110 126	222 230	268 270	80 82	255 263	213 221	212 216	172 178	159 161	106 106
V:4	0 0	222 232	268 270	80 86	257 263	213 223	200 212	172 172	161 161	106 106
V:5	130 132	230 236	268 276	82 88	257 257	217 219	200 214	172 178	159 161	106 122
V:6	126 132	222 222	270 276	76 80	263 265	209 213	214 230	172 176	159 161	106 122

Supplementary Table 3.5 (Continued): Sizes of microsatellite marker alleles.

Individual	D16S3046	D16S401	D16S753	D16S3136	D16S3112	D16S3132	D18S843	D18S485	D18S1141
II:1	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
II:2	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
III:4	98 98	166 170	301 305	256 268	201 201	263 269	193 213	176 176	273 273
III:5	96 96	166 168	282 301	252 264	0 0	261 261	215 215	176 176	273 287
III:8	98 100	170 166	301 305	256 268	201 201	263 269	193 193	176 176	273 281
III:7	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
III:9	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:2	96 98	166 170	301 301	252 268	199 201	0 0	213 215	176 176	273 287
IV:3	96 98	166 168	301 305	252 256	199 201	261 263	213 215	176 176	273 287
IV:4	96 98	166 170	301 301	252 268	199 201	0 0	213 215	176 176	273 287
IV:5	96 98	168 170	290 336	256 256	199 201	267 271	193 213	176 180	273 277
IV:6	96 100	170 170	301 305	252 268	201 207	0 0	193 213	176 176	273 281
IV:7	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:8	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:9	96 100	170 172	301 305	252 268	201 201	267 269	193 211	164 176	0 0
IV:10	94 96	170 176	305 317	252 264	201 201	261 267	193 211	176 180	265 273
IV:11	96 100	170 172	301 344	248 268	201 201	267 269	193 211	174 176	273 291
IV:14	98 98	160 166	305 344	248 256	201 201	263 267	193 211	174 176	273 291
V:1	96 96	170 170	301 336	256 268	0 0	269 271	0 0	176 176	273 277
V:3	94 100	170 170	305 340	252 256	201 207	263 265	207 213	176 176	273 285
V:4	98 100	170 170	301 328	264 268	201 203	267 269	0 0	164 176	273 291
V:5	96 100	170 170	301 305	252 268	201 201	261 269	193 193	176 180	265 291
V:6	94 100	170 176	301 317	264 268	201 201	267 267	211 211	176 176	273 273

Supplementary Table 3.6: Encoded microsatellite marker genotypes.

Individual	D8S543	D8S286	D8S1757	D8S88	D15S155	D15S1020	D15S984	D16S3399	D16S3134	D16S3058
II:1	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
II:2	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
III:4	1 3	2 4	3 4	2 2	3 2	3 6	4 3	5 6	3 4	1 6
III:5	3 5	3 4	1 5	2 4	2 6	9 5	2 3	1 4	3 5	4 5
III:8	1 5	2 4	2 3	2 3	2 5	3 7	3 4	2 6	3 3	1 1
III:7	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
III:9	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:2	3 3	4 3	4 1	2 2	3 6	5 6	2 3	0 0	3 5	6 4
IV:3	3 3	4 3	4 1	2 2	2 2	6 9	2 3	0 0	3 3	1 5
IV:4	1 5	2 4	3 5	2 4	3 6	3 5	4 3	5 4	4 5	6 4
IV:5	3 6	1 3	6 7	3 4	3 4	2 4	0 0	3 5	2 3	1 2
IV:6	1 2	2 4	3 5	2 2	3 5	3 7	3 4	2 2	3 1	1 1
IV:7	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:8	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:9	1 1	2 4	3 6	2 6	5 6	3 4	3 4	2 2	3 1	1 1
IV:10	3 4	2 6	4 4	1 3	2 6	1 5	1 7	4 5	1 3	3 5
IV:11	1 5	2 4	2 3	2 6	2 5	3 6	3 4	2 2	3 1	1 1
IV:14	1 5	4 4	2 6	2 3	2 5	3 4	3 4	6 5	3 1	1 1
V:1	3 2	3 4	7 6	4 2	3 5	2 3	3 6	3 2	2 1	1 1
V:3	3 1	2 4	2 3	3 2	1 5	3 7	3 5	5 2	1 3	1 1
V:4	0 0	5 2	2 3	2 5	2 5	8 3	1 3	2 2	3 3	1 1
V:5	4 5	6 4	4 2	3 6	2 2	5 6	1 4	5 2	1 3	3 1
V:6	3 5	2 2	4 3	1 2	5 6	1 3	4 7	4 2	1 3	3 1

Supplementary Table 3.6 (Continued): Encoded microsatellite marker genotypes.

Individual	D16S3046	D16S401	D16S753	D16S3136	D16S3112	D16S3132	D18S843	D18S485	D18S1141
II:1	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
II:2	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
III:4	3 3	2 4	3 5	2 2	2 5	4 5	1 4	3 3	2 2
III:5	2 2	2 3	2 4	0 0	1 1	1 2	5 5	3 3	2 6
III:8	3 4	4 2	3 5	2 2	2 5	4 5	1 1	3 3	2 4
III:7	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
III:9	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:2	3 2	4 2	5 2	2 1	0 0	5 1	4 5	3 3	2 6
IV:3	3 2	2 3	3 2	2 1	2 1	4 1	4 5	3 3	2 6
IV:4	3 2	4 2	5 2	2 1	0 0	5 1	1 4	3 4	2 3
IV:5	2 3	3 4	3 3	1 2	4 6	1 5	1 4	3 4	2 3
IV:6	4 2	4 4	5 2	2 4	0 0	5 1	1 4	3 3	2 4
IV:7	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:8	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0	0 0
IV:9	4 2	4 5	5 2	2 2	5 4	5 3	1 3	3 1	0 0
IV:10	1 2	4 6	2 4	2 2	1 4	1 6	1 3	3 2	2 7
IV:11	4 2	4 5	5 1	2 2	5 4	5 5	1 3	3 2	2 7
IV:14	4 2	4 5	3 1	2 2	2 4	4 5	1 3	3 2	2 7
V:1	2 2	4 4	3 5	0 0	6 5	5 5	0 0	3 3	3 2
V:3	1 4	4 4	3 2	2 4	2 3	4 1	2 4	3 3	5 2
V:4	3 4	4 4	4 5	3 2	4 5	0 0	0 0	1 3	2 7
V:5	2 4	4 4	2 5	2 2	1 5	0 0	1 1	4 3	1 7
V:6	1 4	6 4	4 5	2 2	4 4	1 5	3 3	3 3	2 7

Chapter 4

Transcriptome analysis of CMT720 and validation of candidate genes

4.1. Introduction

In the previous chapter, two suggestive linkage regions on chromosomes 8 and 16 have been supported using fine mapping linkage analysis. This refinement will allow focusing specifically on the variants localising within 1.7% of the genome in the current investigation of CMT720. Despite a substantial reduction in analytical burden, variant identification using WGS will likely produce tens of thousands of noncoding variants in these regions [153]. To effectively interrogate all classes of non-coding variants for a pathogenic role in CMT720, a prioritisation strategy using different “omics” tools is needed to address the major challenges that make noncoding mutations refractory to analysis.

Noncoding variants can cause inherited neuropathy through a wide range of molecular mechanisms [130, 131, 146] as summarised in Chapter 1. However, the knowledge gaps regarding the function of noncoding sequences greatly limits the accuracy of predictive software and prevents high-throughput prioritisation of impactful noncoding mutations [157, 196]. To overcome these limitations, transcriptome directed variant analysis is a strategy to identify non-coding mutations that may be impacting gene regulation, formation of aberrant transcripts or causing gene fusions [158, 162]. In this chapter, we will use transcriptome analysis of CMT720 using patient derived fibroblasts to identify any candidate genes within the suggestive linkage regions that show differential expression or aberrant transcripts (splicing or fusion). This approach will identify candidate genes impacted by differential transcriptome changes and facilitate identifying and analysing a feasible number of noncoding variants for a pathogenic role

in CMT720.

4.2. Utilising transcriptomic analysis for prioritizing noncoding variants

Transcriptomics combined with genomic analysis is a key identification and prioritisation strategy for rare diseases [197] and has a genetic diagnostic success rate of 9% [158] to 35% [198] in cases previously unsolved by WES or WGS alone. Transcriptomics can capture the functional impact of a wide range of causative noncoding variants implicated in IPNs such as the mutations that disrupt splicing [130, 132], cause formation of pathogenic fusion transcripts [131] or differential gene expression between patients and controls [129, 164]. For example, in a cohort of 50 patients with rare neuromuscular diseases undiagnosed after WES or WGS, transcriptomic analysis identified pathogenic intronic variants that led to abnormal splicing in 20% of the cases [198]. Since it is often not possible to predict whether the gene fusions are expressed based on the genomics sequence alone, integration of transcriptomics into structural variation (SV) analysis is considered as the best practice for detecting clinically relevant gene fusions [162, 199]. Our group detected a pathogenic gene-intergenic *UBE3C* fusion transcript formed by the intrachromosomal DHMN1 translocation through combined analysis of WGS data and transcriptomic analysis [131]. Differential gene expression analysis is a core component of most transcriptomic studies [200, 201], and is highly effective for identifying the variants that disrupt regulatory sequences but remain elusive in genomic analyses [128, 202]. In a cohort of unsolved rare diseases, detection of differential gene expression (DGE) identified variants disrupting UTR or promoter regions which accounted for 21% of the successful diagnoses made by transcriptome directed analysis [128]. Since transcriptomics can be highly informative for functional prioritisation of a wide range of noncoding variants, our strategy will be to analyse the transcriptome of family CMT720, with the goal of identifying aberrant splicing, fusion transcripts

and gene dysregulation.

4.3. The strategy for performing transcriptomic analysis on CMT720

For transcriptomic analysis to accurately represent the underlying disease biology, appropriate tissue and methods of quantification are needed [203]. Ideally, transcriptomic analysis of CMT720 would require access to the patient spinal MNs and sensory neurons of dorsal root ganglia. These cell types cannot be obtained non-invasively currently the only viable way to study axonal IPNs in disease relevant tissue is reprogramming patient-derived induced pluripotent stem cells (iPSC) into MNs [131, 204, 205]. For this exploratory study, using patient fibroblasts can provide an adequate alternative cell type for investigating the transcriptome of CMT720 in a timely and cost-efficient manner [206]. Compared to other easily accessible patient tissue such as lymphoblasts and whole blood, it has been shown that fibroblasts express genes associated with IPNs at similar levels to neurons [207] and share a higher number of alternatively spliced transcripts with neurons [208]. Furthermore, previous investigations of CMT2 have detected abnormal splicing variants and gene dysregulation caused by the pathogenic variants using patient derived fibroblasts [209, 210]. Therefore, in the current study, fibroblasts obtained from CMT720 patients will be used to perform the transcriptomic analysis.

In this investigation, RNA-seq will be performed on patient fibroblasts by following the common experimental and data analysis practices summarised in Figure 4.1. The RNA-seq data will be queried for abnormally spliced transcripts and fusion transcripts for prioritising potential intronic variants that disrupt splicing or SVs that give rise to expressed gene fusions, respectively. Additionally, DGE analysis will be undertaken to detect gene dysregulation that may be caused by candidate pathogenic variants disrupting UTRs and cis-regulatory elements (CREs). Differentially expressed genes (DEGs) will be validated by quantitative real-time PCR (qRT-PCR), which is accepted as the gold-standard method for quantifying transcript abundance and is commonly used to validate the results of DGE analyses as an orthogonal method [211].

Accordingly, we expect our transcriptomic analysis to be highly informative for guiding the selection and prioritisation of noncoding variants by providing candidate genes for variant context in family CMT720.

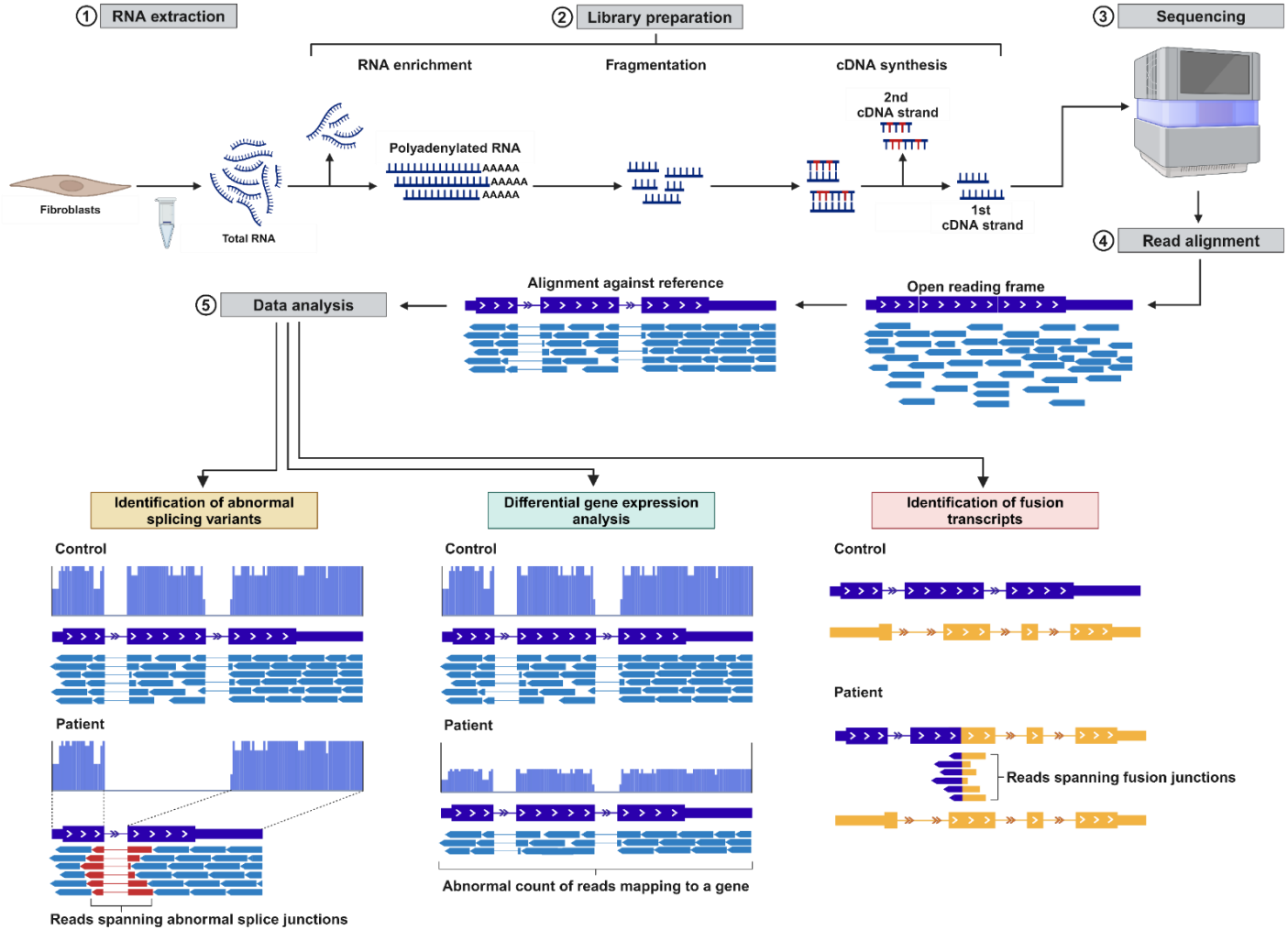


Figure 4.1: The summary of RNA-seq workflow for the transcriptomic analysis of CMT720. RNA from patient and control fibroblasts will be extracted and enriched for polyadenylated transcripts. A stranded RNA-seq library will be constructed and sequenced using short-read NGS. The transcriptome of CMT720

will be constructed by alignment and the resulting data will be analysed by a selection of algorithms to identify fusion transcripts, abnormal splice isoforms and differential gene expression. This figure was created with BioRender.

4.4. Hypothesis

We hypothesise that performing multi-omics variant prioritisation by generating the transcriptomic profile of CMT720 using RNA-seq will allow specific selection of high priority candidate noncoding pathogenic variants for CMT720. The specific aims of this chapter are:

4.5. Aims

1) Performing RNA-seq using CMT720 and control fibroblasts to identify differential gene expression, fusion transcripts, and aberrantly spliced transcripts located in the prioritised candidate linkage loci.

2) Validating the expression of the DEGs identified by RNA-seq analysis using qRT-PCR.

4.6. Methods

4.6.1. Selection and culturing of patient and control fibroblasts

Biopsies to generate patient fibroblast lines from two affected members in the family was kindly organised by Professor Mary Reilly (University College London). Cryopreserved fibroblasts from the two affected individuals (IV:4 and IV:9), and 4 unrelated healthy controls (C1, C2, C3 and C4) were thawed and cultured using our standard protocol as described (Chapter 2, Section 8) to obtain the patient and control RNA for the RNAseq and qRT-PCR experiments.

4.6.2. RNA sequencing and data analysis

Total RNA was isolated from patient (n=2) and control (n=4) fibroblasts using the RNEasy Mini kit (QIAGEN) following the manufacturer's protocol as described in Chapter 2, Section 10. RNA sequencing, alignment of sequencing data and transcript quantification were outsourced to Macrogen (Seoul, Korea). TruSeq Stranded mRNA kit (Illumina) was used to prepare the RNA sequencing library where the transcripts with 5' poly-A tails were captured. Paired-end sequencing with a read length of 151 bp was performed by using the NovaSeq6000 platform (Illumina). Quality control on the unaligned sequencing data was performed by Macrogen (Seoul, Korea) using FastQC [212]. Adapter sequences and low- quality bases were trimmed using Trimmomatic [213] and the trimmed reads were aligned against the NCBI_109.20200522 genome annotation based on the GRChg38 human genome assembly [143] using the splice-aware aligner HISAT2 [214]. Quantification of transcripts at the level of genes and individual isoforms were performed using the StringTie program [215]. Transcript abundance is presented as raw counts and normalised counts indicated as transcripts per kilobase million (TPM).

4.6.3. Identification of novel splicing isoforms and fusion transcripts

Bioinformatic analysis to identify novel splicing isoforms and fusion transcripts was outsourced to Macrogen (Seoul, Korea). Briefly, novel alternatively spliced transcripts in the RNA-seq alignments of patients and controls were identified and quantified by StringTie [215]. StringTie combines alignment against known transcripts and local *de novo* assembly of mapped reads to detect novel splicing isoforms [215]. Arriba [216], Defuse [217] and FusionCatcher [218] algorithms were used to detect gene fusions based on chimeric and multimapping reads in RNA-seq alignments of patients and controls.

4.6.4. Differential gene expression analysis

Due to the variability in gene expression patterns in RNA-seq experiments, algorithms to predict differential gene expression may generate high rates of false positives or fail to detect dysregulated genes [219, 220]. Comparative benchmarking studies of DGE algorithms have shown that the highest rates of true positives are obtained by combining the outputs of multiple algorithms [219]. Therefore, in this study we used three commonly used algorithms: DESeq2 [221], edgeR [222] and NOISeq [223] to identify DGE in CMT720. In each DGE analysis pipeline the count data was normalised against the library size prior to calculations to account for the variability across samples [224, 225] and low count genes that produce inaccurate estimates of differential expression are removed [221, 226]. DGE was determined using algorithm-specific thresholds adjusted to the false discovery rate (FDR)(DESeq2: FDR = 0.1; edgeR: FDR = 0.05; NOISeq: probability of differential expression > 95%). Specific parameters for low count filtering and determining DGE for each algorithm can be found in the Supplementary materials.

DEGs identified by multiple algorithms were determined by using DiVenn2 to visualise the intersections between datasets produced by each tool [227]. DEGs within our prioritised suggestive linkage regions that were identified by at least two different algorithms were determined as the dysregulated positional candidate genes. This approach was taken to obtain the most accurate predictions of differential expression since the algorithms used for performing DGE analysis on RNA-seq data often produce high numbers of false-positive DEGs [219].

4.6.5. Validation of differentially expressed candidate genes with qRT-PCR

qRT-PCR was performed to validate the differential expression of the dysregulated positional candidate genes predicted by the DGE analysis. The iScript cDNA Synthesis Kit (BioRad) was used to reverse transcribe the total RNA extracted from patient and control fibroblasts as described in Chapter 2, Section 11. qRT-PCR was performed in triplicate with 20 µL reactions using the same RNA extracts that underwent RNA-seq. The reaction mixture contained final concentrations of (1X) TaqMan Gene Expression Assay (1 µL), (1X) TaqMan Gene Expression Mastermix (10 µL), 50 ng cDNA template (1 µL) and 8 µL of nuclease-free water (Invitrogen). TaqMan gene expression assays that targeted the largest number of transcripts for each selected gene was used as summarised in Supplementary Table 4.1. *GAPDH* and *RPLP0* were used as endogenous control genes. DNase/RNAse-free water was used as the negative control for each gene expression assay. Thermal cycling was carried out using Step One Plus Real-Time PCR machine (Applied Biosystems) with the program shown in Table 4.1. Due to the number of assays used, qRT-PCR was performed in 3 separate batches.

Table 4.1. The thermal cycling protocol used for qRT-PCR.

Step	Number of cycles	Duration (min:s)	Temperature (°C)
Initial denaturation	1	10:00	95.0
Denaturation Annealing	40	00:15 1:00	95.0 60.0

Relative quantitation of gene expression was carried out using the Step One Plus Analysis software (Applied Biosystems). The expression of each candidate gene was quantified with respect to each endogenous control gene using the $2^{\Delta\Delta Ct}$ method [228] and was presented as relative fold change (FC). One tailed, independent samples t-test was used to determine whether the significant up or downregulation predicted by the RNA-seq DGE analysis is validated. The results were visualised using the GraphPad Prism software (10.1.2).

4.7. Results

Transcriptome analysis was performed on 2 patients and 4 control fibroblast cell lines. The quality metrics of the RNA-seq experiment are provided in Supplementary Section 1.

4.7.1. Novel splicing variants were not identified within the prioritised suggestive linkage regions on chromosome 8 and 16

StringTie identified 9 novel splicing transcripts originating from 7 genes in the prioritised suggestive linkage region on chromosome 8 (Supplementary Table 4.7), and 106 novel splicing transcripts originating from 57 genes in the prioritised suggestive linkage region on chromosome 16 (Supplementary Table 4.8). The predicted novel splicing transcripts however, were not called in both affected individuals. Accordingly, all novel splicing isoforms detected by StringTie were excluded from further analysis.

4.7.2. Intrachromosomal gene fusions have been excluded for CMT720

The Defuse tool predicted a total 758 fusion transcripts across all samples. Ony one fusion transcript detected in patient IV:4 by Defuse localised to the prioritised suggestive linkage region on chromosome 8, while Defuse identified no fusion transcripts in the suggestive linkage region on chromosome 16 (Table 4.2). The fusion transcript identified in patient IV:4 was absent in patient V:9 (Table 4.2), therefore, it was excluded as a potential pathogenic transcript.

Table 4.2: List of gene fusions predicted by Defuse in the prioritised suggestive linkage region on chromosomes 8.

Sample	Gene1	Gene2	Gene1 breakpoint (strand)	Gene2 breakpoint (strand)	Probability
IV:4	*TRPA1*	*AC022905.1*	8:72053001 (-)	8:72227389 (+)	0.5695783

The Arriba tool identified 479 fusion transcripts. One transcript detected in patient IV:4 localised to the prioritised suggestive linkage region on chromosome 16 with both gene partners (*CRNDE* and *LOC1053712758*) derived from within the suggestive linkage region (Table 4.3). This call was given a low confidence score by Arriba, indicating that it is unlikely to reflect an expressed gene fusion and is most likely a false positive [216]. The fusion transcript predicted in patient IV:4 was not present in the other affected individual (IV:9), therefore, it was excluded as a candidate pathogenic transcript. No fusion transcripts were predicted by Arriba in the prioritised suggestive linkage region on chromosome 8.

Table 4.3: List of intrachromosomal gene fusions predicted by Arriba in the prioritised suggestive linkage region on chromosome 16.

Sample	Gene1	Gene2	Breakpoint1 (strand)	Breakpoint2 (strand)	Confidence
IV:4	*CRNDE*	*LOC105371275*	16:54923585 (-)	16:54893483 (-)	low

FusionCatcher predicted no intrachromosomal gene fusions within the chromosome 8 suggestive linkage region. Only one intrachromosomal gene fusion was predicted on chromosome 16 with both gene partners derived from within the suggestive linkage region (Table 4.4). The intrachromosomal gene fusion predicted by FusionCatcher was not exclusively observed in both affected individuals, therefore, this call was excluded as a candidate pathogenic fusion product. FusionCatcher has also predicted 192 interchromosomal gene fusions with at least one gene partner in the prioritised suggestive linkage region on chromosome 8 and 758 interchromosomal gene fusions with at least one partner in the prioritised suggestive linkage region on chromosome 16. Due to the prohibitively large number of interchromosomal fusion predictions made by FusionCatcher, these could not be analysed to identify the calls that were unique to the affected individuals.

Table 4.4: List of gene fusions predicted by FusionCatcher in the prioritised suggestive linkage region on chromosome 16.

Sample	5' fusion partner	3' fusion partner	5' partner: strand	3' partner: strand	Predicted effect
IV:9	*MMP2*	*LPCAT2*	16:55505579:+	16:55525508:+	UTR/CDS(truncated)

4.7.3. Differential gene expression analysis of CMT720 patient derived fibroblasts

Principal component analysis (PCA) was carried out using DESeq2 (Supplementary Table 4.4) based on the number of the filtered counts (>3 counts per sample or >10 counts across all samples) to determine the variability in gene expression between samples. PCA indicated clear grouping of samples according to affection status along principal component 1 (PC1) (Figure 4.2), suggesting that affection status accounts for the largest proportion of variance amongst samples.

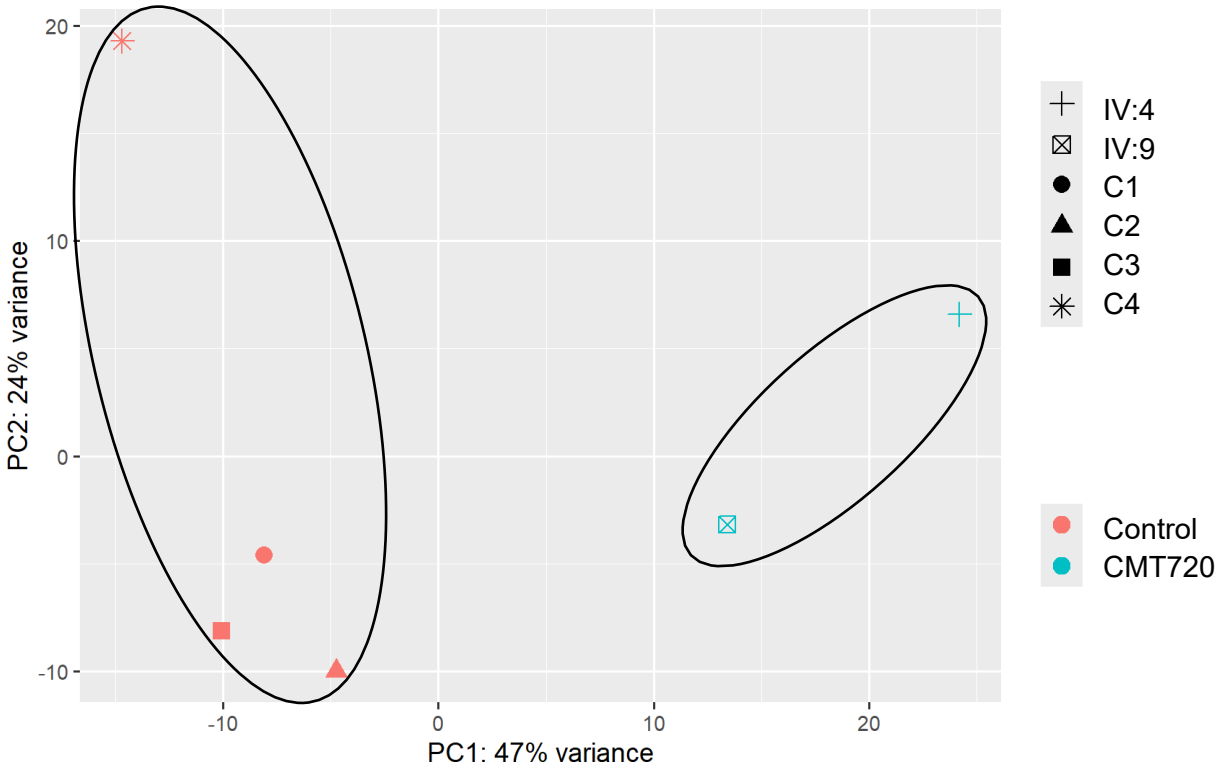


Figure 4.2: PCA plot of RNA-seq data. The PCA plot indicates samples clustering by affection status according to PC1 as indicated by the ovals. The spread along X-axis represents the variability between samples due to the principal component 1 (PC1) while the spread along Y-axis represents the variability due to principal component 2 (PC2). The percentage of variability among all samples explained by PC1 and PC2 are given after colons along the labels for the x- and y-axes respectively.

EdgeR identified a total of 188 DEGs (FDR<0.05) (Figure 4.3). Of these, *ZNF704*

localised to the prioritised suggestive linkage region on chromosome 8 and *IRX3* localised the suggestive linkage region on chromosome 16 (Table 4.5). DESeq2 detected 925 DEGs (FDR<0.1)(Figure 4.3). Of these, 7 were located in the prioritised suggestive linkage region on chromosome 8 while 13 were located to the suggestive linkage region on chromosome 16 (Table 4.6). A total of 1862 DEGs were detected by NOISeq (Probability of DGE >95%)(Figure 4.3) with 7 and 13 genes respectively localising to the chromosome 8 and 16 suggestive linkage regions (Table 4.7). Out of all DEGs identified by EdgeR, DESeq2 and NOISeq, 629 DEGs were supported by at least 2 algorithms (Figure 4.3).

Table 4.5: List of DEGs identified by EdgeR localising to chromosome 8 and 16 candidate suggestive linkage regions.

Gene SYMBOL	Log 2 fold change	FDR-adjusted p-value	Chromosome
ZNF704	3.048710926	0.003505292	8
IRX3	1.063863289	0.005625577	16

Table 4.6: List of DEGs identified by DESeq2 in chromosome 8 and 16 suggestive linkage regions.

Gene SYMBOL	Log 2 fold change	FDR-adjusted p-value	Chromosome
C8ORF38	1.639605097	0.000014151	8
FABP5	-1.571240653	0.098261038	8
LRRCC1	-1.067249182	0.061918491	8
MSC	0.863228801	0.043123490	8
RDH10	2.028518642	0.092121201	8
ZFHX1	0.670519189	0.028811233	8
ZNF704	3.226673337	0.000000006	8
BCL7C	-0.677055925	0.014074401	16
CORO1A	-1.853906707	0.024814285	16
IRX3	1.073974228	0.003282927	16
IRX6	3.201526566	0.048714166	16
MT1E	-0.983375188	0.046454651	16
N4BP1	0.423035211	0.099249343	16
PRRT2	-1.405957157	0.008673855	16

PYCARD	-0.79858	0.042955453	16
SHCBP1	-1.710480012	0.089486001	16
STX1B	-1.171699479	0.019520049	16
TBX6	-1.960525604	0.054229503	16
YPEL3	0.759464394	0.085170692	16
ZNF423	2.55360794	0.003826371	16

Table 4.7: List of DEGs identified by NOISeq in chromosome 8 and 16 candidate suggestive linkage regions.

Gene SYMBOL	Probability of differential expression	Log 2 fold change	Chromosome
LOC105375630	0.964563127	-3.240710547	8
RDH10	0.98233194	2.02370291	8
SBSPON	0.959590937	4.644389678	8
STMN2	0.994502121	2.572496502	8
SULF1	0.951174088	1.256045868	8
TRAM1	0.973290607	0.650072973	8
ZFHX4	0.958418	0.657656	8
ZNF704	0.989711884	3.210528725	8
ADCY7	0.957673007	0.63571963	16
ANKRD26P1	0.984920854	4.129269394	16
BCL7C	0.963787358	-0.69076525	16
CD19	0.963674711	-3.163321565	16
DOC2A	0.979898901	-3.655795772	16
FUS	0.965709919	-0.421040207	16
HERPUD1	0.978554	0.677841	16
HMGN2P41	0.960673676	-0.8082824	16
IRX3	0.976750041	1.059789002	16
IRX6	0.960388346	3.169294499	16
KIF22	0.964892012	-1.278453996	16
LOC112694756	0.958622607	-0.447550919	16
METTL9	0.964261736	-0.400324577	16
MMP2	0.998380216	0.452026449	16
MT1E	0.962672395	-0.995615198	16
MT2A	0.967833294	-0.746639995	16
NUPR1	0.960029653	0.532706653	16
PLK1	0.972608	-1.80999	16
PRRT2	0.964441	-1.42114	16
SHCBP1	0.971000196	-1.726830474	16
YPEL3	0.976316903	0.746356974	16

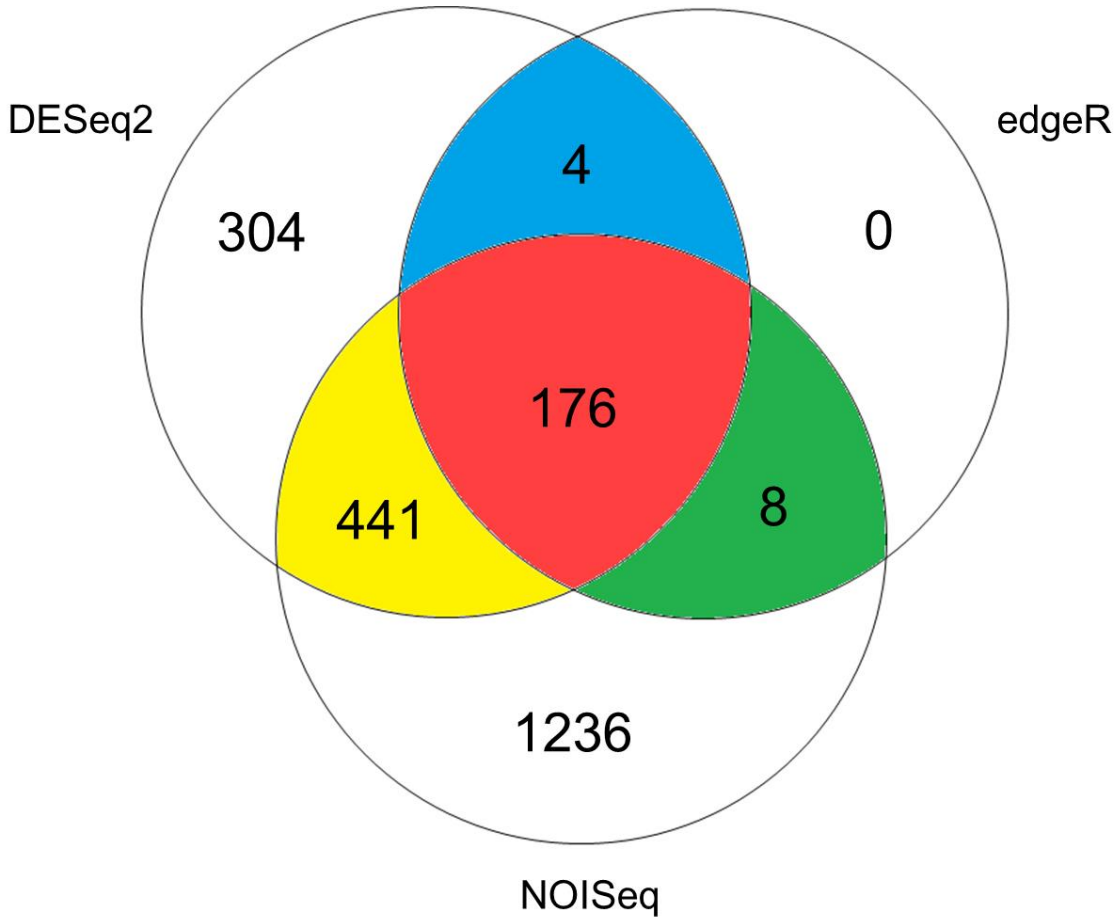


Figure 4.3: The Venn diagram representing the overlap between all DEGs identified by DESeq2, NOISeq and EdgeR. 441 DEGs were supported by both DESeq2 and NOISeq, while 8 DEGs were identified by both edgeR and NOISeq. 4 genes were detected by DESeq2 and edgeR. 1236 DEGs were detected only by NOISeq, whereas 304 DEGs were detected only by DESeq2. No DEGs were identified only by edgeR.

2 DEGs (*ZFHX4* and *RHD10*) in the prioritised suggestive linkage region on chromosome 8 were supported by both DESeq2 and NOISeq and *ZNF704* was supported by all three algorithms (Figure 4.4A). Six candidate genes (*BCL7C*, *IRX6, MT1E*, *PRRT2*, *SHCBP1* and *YPEL3*) localising to the chromosome 16 were supported by NOISeq and DESeq2, while *IRX3* was supported by all three algorithms used to detect differential gene expression (Figure 4.4B). Based on 2 or more algorithms supporting the differential expression of these genes, they were selected as the candidate dysregulated genes to potentially use for

transcriptome guided analysis in CMT720.

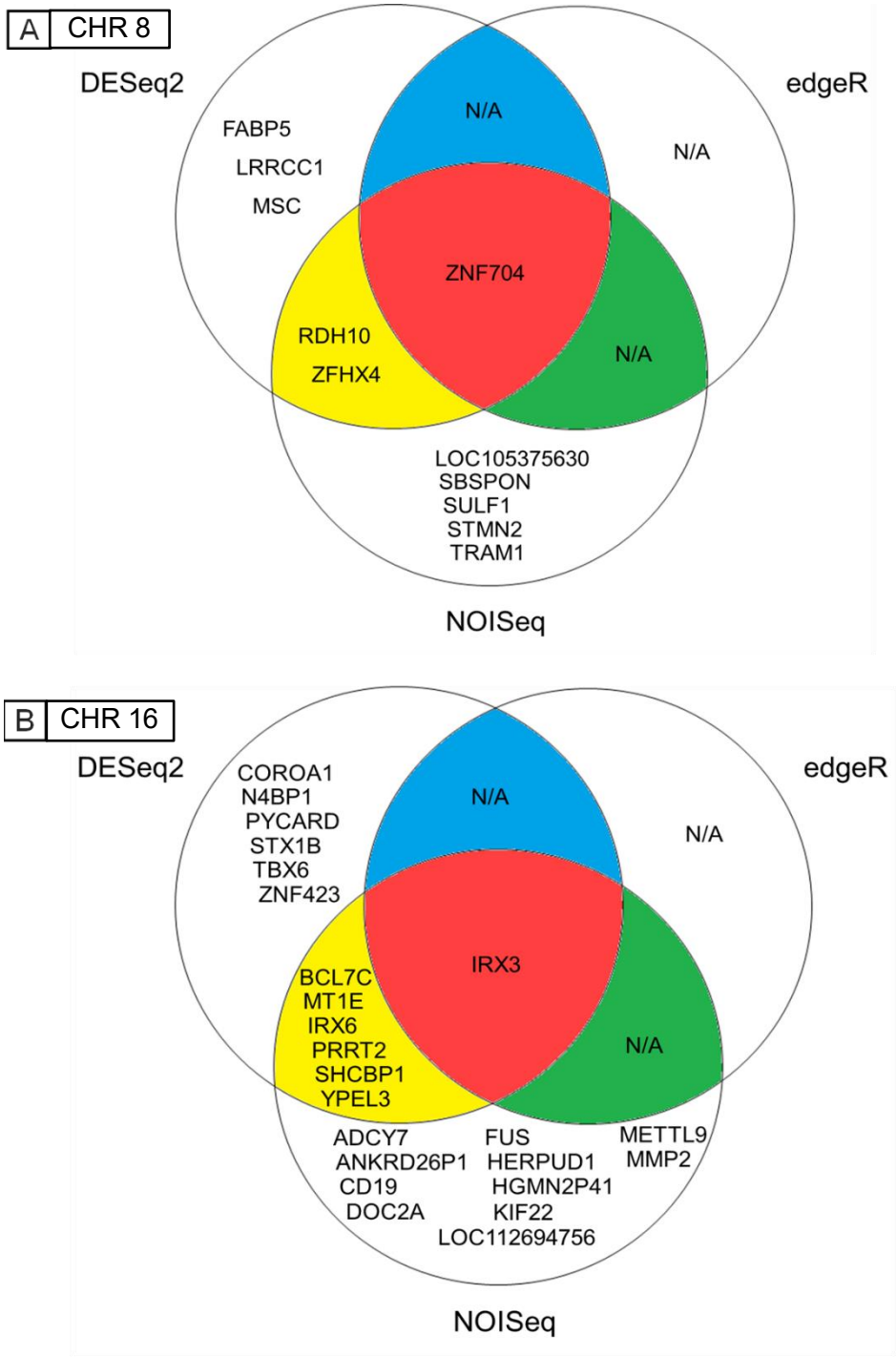


Figure 4.4: The Venn diagram representing the overlap between the DEGs identified by DESeq2, NOISeq and EdgeR across the prioritised suggestive linkage regions. (A) The DEGs identified by DESeq2, NOIseq and edgeR in our prioritised suggestive linkage region on chromosome 8 **(B)** The DEGs identified by DESeq2, NOIseq and edgeR in our prioritised suggestive linkage region on chromosome 16.

To assess the biological relevance of the candidate dysregulated genes identified in the patient fibroblasts, neuronal expression was determined using in-house RNA-seq data from healthy MNs derived from iPSC control lines (MN C1, MN C2, MN C3)(Table 4.8)[131]. *MT1E* was not expressed in the control MNs (Table 4.8), therefore, this gene was deprioritised and not assessed by qRT-PCR analysis in this study.

Table 4.8: Gene expression levels of differentially expressed candidate genes in control MNs.

Gene SYMBOL	Unnormalised counts			TPM			Chromosome
	MN C1	MN C2	MN C3	MN C1	MN C2	MN C3	
RDH10	841	702	735	6.70	5.29	6.28	8
ZFHX4	939	1009	620	2.12	2.15	1.50	8
ZNF704	13112	8301	12455	27.78	16.53	28.14	8
BCL7C	3299	2844	3028	76.15	59.77	78.86	16
IRX3	1072	1173	661	12.70	13.26	8.43	16
IRX6	656	1099	670	7.41	11.59	7.99	16
MT1E	0	0	0	0.00	0.00	0.00	16
PRRT2	24575	24683	21076	352.07	326.43	325.40	16
SHCBP1	58	64	34	0.41	0.36	0.22	16
YPEL3	3310	3661	3681	107.72	109.35	119.56	16

TPM: Transcripts per kilobase million

4.7.4. qRT-PCR validates dysregulation of 4 candidate genes from chromosome 8 and 16 suggestive linkage regions

Quantitative RT-PCR (qRT-PCR) was used to validate the DGE identified by RNA-seq analysis for the 9 candidate DEGs confirmed to be expressed in neurons. *RPLP0* was selected as the endogenous control for relative quantification of gene expression since it displayed more stable expression across all batches and samples when compared to *GAPDH* (Supplementary Table 4.9). The direction and magnitude of relative expression observed by

qRT-PCR is shown for the candidate dysregulated genes (Figure 4.5). Among the candidate DEGs in the prioritised suggestive linkage region on chromosome 16, *BCL7C* (FC= 0.66, p=0.0383) and *PRRT2* (FC= 0.26 , p=0.0427) showed significant downregulation, while *IRX6* (FC=2.17, p=0.0137) showed significant upregulation (Figure 4.5), recapitulating the DGE identified in the RNA-seq experiment. For candidate genes in the suggestive linkage region on chromosome 8, only the upregulation of *ZNF704* (FC=15.08, p=0.0004) was supported by qRT-PCR (Figure 4.5). Since dysregulation of gene expression identified by RNA-Seq for *BCL7C*, *IRX6*, *PRRT2* and *ZNF704* was supported by qRT-PCR, these genes were selected as the high priority dysregulated positional gene candidates in our suggestive linkage loci.

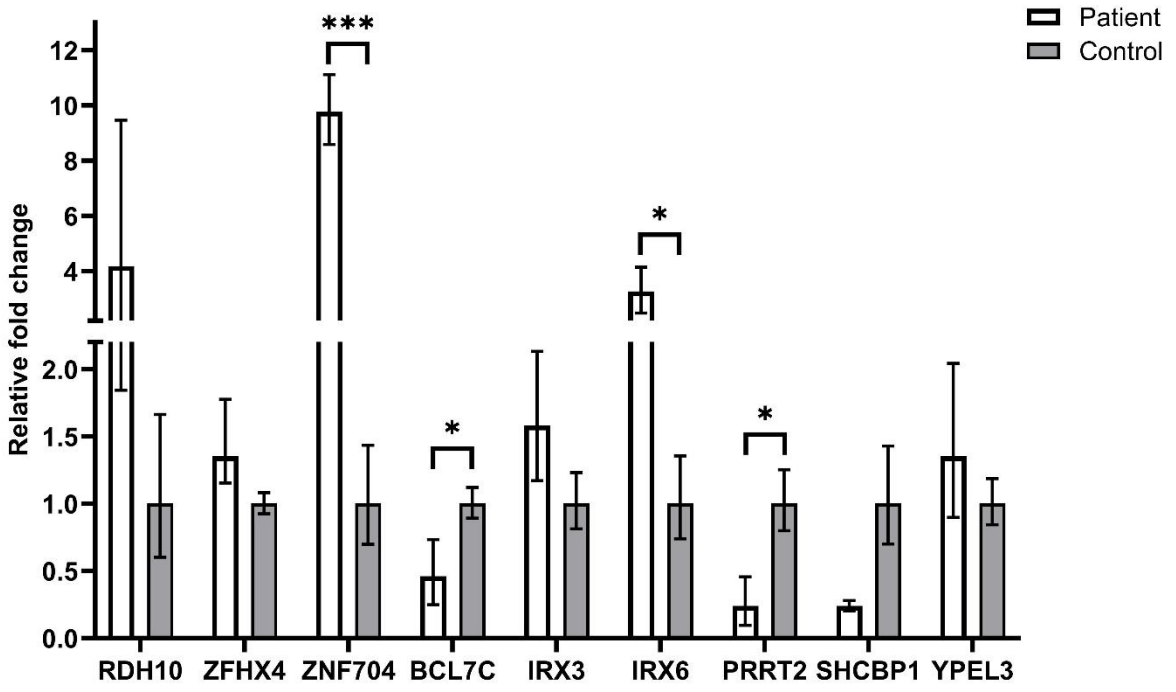


Figure 4.5: Validating the expression profiles of dysregulated positional candidate genes localising to the suggestive chromosome 8 and 16 linkage regions. The histogram showing relative expression of the 9 candidate dysregulated genes in CMT720 patients with respect to controls determined by qRT-PCR analysis after normalisation against *RPLP0*. The error bars represent the standard error of the mean relative quantity (RQ) in each group. *ZNF704* and *IRX3* show significant upregulation, whereas, *BCL7C* and *PRRT2* show significant downregulation (*p<0.05, ***p<0.0005).

4.8. Discussion

In this chapter we have performed transcriptome profiling on patient-derived fibroblast cells from family CMT720 with the goal of identifying and validating any dysregulated genes, abnormally spliced variants, or gene fusion transcripts localising within the suggestive linkage regions on chromosome 8 and 16 refined in Chapter 3.

Linkage analysis is a highly versatile tool that can expedite the transcriptomic analysis of Mendelian disorders [229]. Using the suggestive linkage regions to direct the analysis of patient RNA-seq data, 98.6% abnormally spliced transcripts, 98.4% gene fusions and 99.6% of DEGs were excluded for a pathogenic role in CMT720. The filtering power provided by linkage analysis was instrumental for obtaining a manageable number of DEGs for validation by qRT-PCR and analysing all aberrant transcripts except for the interchromosomal gene fusion calls made by FusionCatcher. These findings demonstrate the effectiveness of linkage analysis for facilitating the transcriptomic analysis of Mendelian disorders.

Analysis for abnormally spliced transcripts unique to the affected individuals was performed by StringTie [215] and no aberrant transcripts were identified in the suggestive linkage regions on chromosome 8 and 16. Using a single algorithm to detect abnormal splicing raises the possibility some abnormally spliced transcripts could remain undetected since these tools can have high rates of false negatives [230]. This can be addressed in future transcriptomic analysis conducted on CMT720 by developing a pipeline that incorporates additional algorithms to increase the sensitivity for detecting abnormally spliced transcripts. Due to time constraints this was not possible in the current investigation.

The gene fusion analysis indicated that the causative variant is unlikely to be an intrachromosomal rearrangement disrupting positional candidate genes. No segregating gene

fusion transcripts with both gene partners originating from chromosome 8 or 16, were detected within the suggestive linkage regions. In contrast, the high number of FusionCatcher interchromosomal gene fusion predictions was prohibitive for systematically selecting calls unique to the CMT720 patients, as further filtering would be needed beyond the linkage localisation. These calls most likely represent the high rates of false positives that may be observed in some fusion prediction algorithms [231]. In this situation, SVs identified by WGS can aid in identifying and validating functionally relevant interchromosomal gene fusion predictions [162]. This strategy will be useful to revisit the predictions of interchromosomal gene fusions in the event interchromosomal translocation SVs are detected by WGS and localise to the suggestive linkage regions on chromosome 8 and 16.

The DGE pipeline predicted 629 DEGs that were supported by at least 2 algorithms. Together with the PCA analysis that shows clear grouping of samples according to affection status, these findings strongly suggest that CMT720 fibroblasts display global gene dysregulation. Of the 629 DEGs, segregation of abnormal gene expression patterns was identified in the suggestive linkage regions with 10 positional candidates supported by 2 or more algorithms used in the DGE analysis. The qRT-PCR analysis of the positional candidate genes recapitulated the dysregulated expression in 4 of the genes (*BCL7C, IRX6*, *PRRT2* and *ZNF704*). These dysregulated positional candidate genes will be used for directing the selection of noncoding variants to be assessed from patient WGS data within the suggestive linkage regions on chromosome 8 (*ZNF704*) and chromosome 16 (*BCL7C, IRX6*, *PRRT2*). Since all aberrant splicing predictions were excluded, the downregulation of *PRRT2* and *BCL7C* is unlikely to result from nonsense mediated decay. Accordingly, the most likely mechanisms that explain the dysregulation of these prioritised DEGs include disruption of regulatory sequences such as cis-regulatory elements (CREs) and microRNA binding sites [164, 232]. Overall, the findings from our DGE analysis implicate the noncoding variants in

the introns and UTRs of prioritised DEGs, and those located in intergenic CREs as the variants with the highest priority for analysis.

The dysregulated positional candidate genes were confirmed to be expressed in control iPSC-derived MNs which provided further support for biological relevance in appropriate tissue affected in CMT2 disease. However, based on the role of *IRX6* and *PRRT2*, these two positional candidates hold the highest functional priority. *IRX6* is involved in the differentiation of motor neurons [233], while *PRRT2* regulates synaptic activity in CNS neurons and is associated with paroxysmal disorders [234], episodic ataxia, and neurodevelopmental disorders [235]. Interestingly, there is also a report in the literature associating *PRRT2* with polyneuropathy, although, purely based on clinical evaluation with no electrophysiological testing [236]. The extensive association of *PRRT2* with movement disorders and its potential involvement in polyneuropathy makes this gene a highly interesting candidate for CMT720. It was interesting to note that no gene dysregulation or prediction of aberrant transcripts was identified for *GDAP1*, suggesting that non-coding mutations impacting the gene regulation of *GDAP1* are unlikely to be the cause of axonal CMT in family CMT720. Nevertheless, as a known CMT2 gene localising to the suggestive linkage locus on chromosome 8, intronic variants in *GDAP1* still need to be analysed as a conservative measure against the transcriptomic abnormalities that may have been missed in the current investigation due to use of non-neuronal tissue.

In chapter 5, the full spectrum of DNA variants (SNVs, indels, SVs and REs) will be identified using WGS and the noncoding variants located in the introns and UTRs of the dysregulated positional candidate genes identified in this chapter will be prioritised for further analysis. In addition, epigenomic data will be used to determine the CREs that control the expression of the dysregulated positional candidates, and the noncoding variants within these

regulatory regions will also be prioritised. Overall, the results in this chapter provide a transcriptome guided framework to select relevant non-coding DNA candidate variants as the pathogenic cause of disease in CMT720.

4.9. Supplementary Material

Supplementary Table 4.1: TaqMan gene expression assay probes.

Gene ID	Probe ID
BCL7C	Hs00191102_m1
GAPDH	Hs99999905_m1
IRX3	Hs00735523_m1
IRX6	Hs01584109_m1
PRRT2	Hs00293604_m1
RDH10	Hs00416907_m1
RPLP0	Hs99999902_m1
SHCBP1	Hs01090784_m1
YPEL3	Hs00368883_m1
ZFHX4	Hs01016103_m1
ZNF704	Hs01367662_m1

Supplementary Section 4.9.1. Quality analysis for RNA-seq alignments

On average, RNA-seq has generated 17.5 gigabases (Gb) of sequence per sample (Supplementary Table 4.2). Across all samples, an average of 92.24% of the base calls had a Phred score of 30 or above and >90% of all base calls in each sample had a Phred score of 30 or above, indicating high sequencing quality (Supplementary Table 4.2). High alignment quality was achieved across each sample with >98% the reads mapping against the reference (Supplementary Table 4.3).

Supplementary Table 4.2: Quality control statistics on raw RNA-seq data.

Sample	Total bases	Total reads	GC (%)	Q20 (%)	Q30 (%)
IV:4	19,999,658,268	132,448,068	51.27	97.39	92.91
IV:9	16,293,111,702	107,901,402	51.64	96.94	92.06
C1	15,746,031,756	104,278,356	51.65	96.89	92
C2	19,992,863,872	132,403,072	50.13	97.26	92.57
C3	18,057,874,372	119,588,572	51.37	96.91	91.98
C4	15,140,521,756	100,268,356	51.55	96.89	91.91

GC: Guanine/cytosine content

Supplementary Table 4.3: Quality control statistics on aligned RNA-seq data.

Sample	Processed reads	Mapped reads (%)	Unmapped reads (%)
IV:4	129,914,200	128,306,673 (98.76)	1,607,527 (1.24)
IV:9	105,359,160	104,185,165 (98.89)	1,173,995 (1.11)
C1	101,660,596	100,305,227 (98.67)	1,355,369 (1.33)
C2	129,578,458	128,065,737 (98.83)	1,512,721 (1.17)
C3	116,517,242	115,256,647 (98.92)	1,260,595 (1.08)
C4	97,769,358	96,609,797 (98.81)	1,159,561 (1.19)

Supplementary Table 4.4: The script used to perform DGE analysis with DESeq2.

```
1  setwd(“D:/Downloads/result_RNAseq/Exploratory stats and differential
   expression analysis/Deseq2”)
2  cts<-read.table("DY_fbRNAseqCounts.csv", sep = ",", header = TRUE,
   row.names = 1)
3  cts <- as.matrix(cts)
4  condition <- factor(c(rep("Control", 4), rep("CMT720", 2)))
5  samples <- factor(c("Control 1", " Control 2", " Control 3", "Control 4", "IV:4",
   "IV:9"))
6  coldata <- data.frame(row.names=colnames(cts), condition, samples)
7  library(DESeq2)
8  dds <- DESeqDataSetFromMatrix(countData = cts, colData = coldata, design = ~
   condition)
9  dds
10 keep <- rowSums(counts(dds) >= 10) >=3
11 dds <- dds[keep,]
12 dds$condition <- relevel(dds$condition, ref = "Control")
13 plotPCA(rld, intgroup = c("condition", "samples"))
14 pcaData <- plotPCA(rld,intgroup = c("condition", "samples"), returnData = TRUE)
15 percentVar <- round(100 * attr(pcaData, "percentVar"))
16 ddsDEG <- DESeq(dds)
17 res <- results(ddsDEG)
18 MaxCooks <- apply(assays(ddsDEG)[["cooks"]], 1, max)
19 LowCounts <- res$baseMean < metadata(res)$filterThreshold
20 res$Outlier = res$baseMean > 0 & is.na(res$pvalue)
21 res$MaxCooks = MaxCooks
22 res$LowCounts = LowCounts
23 res
24 resSig010 <- subset(res, padj < 0.1)
25 write.csv(as.data.frame(resSig010), file = "DESeq2_Results_padj010.csv")
```

Supplementary Table 4.5: The script used to perform DGE analysis with edgeR.

```
1  setwd(“D:/Downloads/result_RNAseq/Exploratory stats and differential
   expression analysis/Deseq2”)
2  library(edgeR)
3  library(dplyr)
4  countdata<-read.table(“add your file.csv", sep = ",", header = TRUE, row.names
   = 1)
5  countdata <- as.matrix(countdata)
6  condition <- factor(c(rep(“group”, numberofreplicates), rep(“group”,
   numberofreplicates)))
7  coldata <- data.frame(row.names=colnames(countdata), condition)
8  targets <- read.table(“a file containing all your variables.txt", header = TRUE)
9  cpms = cpm(countdata)
10 keep <- rowSums(cpms >1) >=1
11 counts = cpms[keep,]
12 d = DGEList(counts = counts, group = targets$Condition)
13 d$samples
14 d = calcNormFactors(d)
15 d$samples
16 d<-estimateDisp(d)
17 d$samples
18 plotBCV(d)
19 et <- exactTest(d)
20 summary(decideTests(et))
21 allsiget<-data.frame(topTags(et,30000))
22 Filter_DEGsET <-allsiget %>% subset(FDR<0.05)
23 View(Filter_DEGsET
```

Supplementary Table 4.6: The script used to perform DGE analysis with NOISeq.

```
1  setwd("D:/Downloads/result_RNAseq/Original/Expression_profile/StringTie")
2  cts<-read.table("DY_fbRNAseqCounts.csv", sep = ",", header = TRUE,
   row.names = 1)
3  cts <- as.matrix(cts)
4  condition <- factor(c(rep("Control", 4), rep("CMT720", 2)))
5  samples <- factor(c("Control 1", "Control 2", "Control 3", "Control 4", "IV:4",
   "IV:9"))
6  coldata <- data.frame(row.names=colnames(cts), condition, samples)
7  head(coldata)
8  head(cts)
9  library(NOISeq)
10 library(dplyr)
```

```
11 myfilt = filtered.data(cts, factor = "condition", norm = FALSE, depth = NULL,
   method = 1, cv.cutoff = 100, cpm = 1, p.adj = "fdr")
12 mydata <- readData(data = myfilt, factors = coldata)
13 mydata
14 mynoiseqbio = noiseqbio(mydata, k = 0.5, norm = "tmm", factor = "condition", lc =
   0, r = 20, adj = 1.5, plot = FALSE, a0per = 0.9, random.seed = 12345, nclust =
   50)
15 results <- as.data.frame(mynoiseqbio@results[[1]])
16 mynoiseq.degU = degenes(mynoiseqbio, q = 0.95, M = 'up')
17 mynoiseq.degD = degenes(mynoiseqbio, q = 0.95, M = 'down')
18 mynoiseq.degALL = degenes(mynoiseqbio, q = 0.95, M = NULL)
19 write.table(mynoiseq.degALL,"Filter_ mynoiseq.degALL.tsv", sep="\t",
   col.names=NA, quote=F)
```

Supplementary Table 4.7: Predictions of abnormal splice isoforms made by StringTie from the prioritised suggestive linkage region on chromosome 8.

Nearest RefSeq transcript	Gene SYMBOL	Gene name	TPM values					
			C1	C2	C3	C4	IV:4	IV:9
NM_015170	*SULF1*	sulfatase 1	0	0	0	17.5443	0	0
NM_001317804	*TRAM1*	translocation associated membrane protein 1	22.2237	60.466	96.7249	58.5216	63.7447	74.2813
NR_033652	*MSC-AS1*	MSC antisense RNA 1	0	0	7.92732	0	8.83398	0
NR_033334	*TMEM70*	transmembrane protein 70	0.00013	0.00041	0.00037	0.00089	0.0039	0.00823
NM_001199214	*STMN2*	stathmin 2	0	0	0	0	7.88546	0
NM_014018	*MRPS28*	mitochondrial ribosomal protein S28	9.05813	6.33387	11.9861	9.49898	4.3734	3.75272
NM_014018	*MRPS28*	mitochondrial ribosomal protein S28	8.34542	5.56903	5.03227	5.80186	4.28164	4.94298
XM_011517528	*RBIS*	ribosomal biogenesis factor	2.51054	1.43512	1.54334	0.84113	1.20354	2.074
XM_011517528	*RBIS*	ribosomal biogenesis factor	3.38229	3.68985	4.37033	4.70588	6.03817	4.29441

Supplementary Table 4.8: Predictions of abnormal splice isoforms made by StringTie from the prioritised suggestive linkage region on chromosome 16.

Nearest RefSeq transcript	Gene SYMBOL	Gene name	TPM values					
			C1	C2	C3	C4	IV:4	IV:9
NM_001039503	*PRSS53*	serine protease 53	0	0	5.95397	0	17.2131	19.5789
NM_001308293	*SH2B1*	SH2B adaptor protein 1	0	3.2062	4.03694	1.1683	7.24856	9.39658
NM_001348078	*LONP2*	lon peptidase 2, peroxisomal	8.26934	5.6209	8.1688	7.5199	4.09614	3.97809
NM_001272096	*STX4*	syntaxin 4	14.1548	11.37	13.1374	12.341	10.2831	8.44451
NM_001199142	*EIF3C*	eukaryotic translation initiation factor 3 subunit C	0.00264	0.0397	0	0.0178	0.05194	0.06446
NR_037609	*SLX1B-SULT1A4*	SLX1B-SULT1A4 readthrough (NMD candidate)	2.92306	2.0887	3.05878	3.635	1.62886	1.76791
XM_011545719	*ATXN2L*	ataxin 2 like	0.57376	11.355	0.57681	3.3286	9.50501	19.0215

NR_037608	*SLX1A-SULT1A3*	SLX1A-SULT1A3 readthrough (NMD candidate)	5.46924	4.6065	5.09873	3.8919	3.61827	3.73507
XM_011545961	*HSD3B7*	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 7	9.83562	1.5468	16.054	10.91	2.80475	0.55726
NM_001039503	*PRSS53*	serine protease 53	0	0	0	0	0	5.85453
gene-RAB43P1	*RAB43P1*	RAB41 member RAS homolog family pseudogene 1	0	0	0	0	11.3318	0
XM_017023376	*LPCAT2*	lysophosphatidylcholine acyltransferase 2	0	0	0	0	119.517	0
NM_001272096	*STX4*	syntaxin 4	0	0	0.00717	0	0	7.45362
NR_134855	*INO80E*	INO80 complex subunit E	3.30911	3.4943	1.63067	4.638	1.24046	2.21856
XM_011522809	*IRX5*	iroquois homeobox 5	1.18189	2.1016	1.41402	5.8748	9.79672	3.42993
NM_001031835	*PHKB*	phosphorylase kinase regulatory subunit beta	4.96964	4.9285	22.9792	2.841	22.8009	17.4374
NM_001348078	*LONP2*	lon peptidase 2, peroxisomal	17.1934	22.296	21.3422	16.899	23.2101	21.927
NM_001365304	*LOC112694756*	uncharaterized LOC112694756	3.17472	8.9338	0.02923	5.0394	0.66369	0
NM_001042432	*CLN3*	CLN3 lysosomal/endosomal transmembrane protein, battenin	0.60914	1.1509	1.01118	1.1808	0.91143	2.68364
XM_011522838	*ADCY7*	adenylate cyclase 7	4.28844	25.984	31.7041	28.449	37.4023	30.7279
XM_011545719	*ATXN2L*	ataxin 2 like	4.88717	0.5664	0.54393	1.1923	4.7687	2.82819
XM_011545719	*ATXN2L*	ataxin 2 like	7.6527	3.5177	2.03609	3.101	3.82614	11.1854
NM_001199142	*EIF3C*	eukaryotic translation initiation factor 3 subunit C	0.51382	2.9096	16.1237	28.388	0.17779	1.10162
NR_134855	*INO80E*	INO80 complex subunit E	0.06777	0.0722	0.92854	0.2373	0.89902	0.52046
NM_001199142	*EIF3C*	eukaryotic translation initiation factor 3 subunit C	0	17.844	13.9019	0.3525	0.02798	0
NM_001039503	*PRSS53*	serine protease 53	0.0729	0.9089	1.22533	0.1942	0.03825	0.19563
XM_017023909	*SETD1A*	SET domain containing 1A, histone lysine methyltransferase	6.01421	5.6101	8.40506	5.0134	6.22395	2.77566

gene-HMGN2P3	*HMGN2P3*	high mobility hroup nucleosomal binding domain 2 pseudogene 3	1.33548	0	0	1.1714	4.90865	0
NM_001039503	*PRSS53*	serine protease 53	20.2963	11.156	26.4328	16.77	5.65796	17.8865
NM_006662	*SRCAP*	Snf2 related CREBBP activator protein	17.8663	17.938	8.34108	18.775	18.908	20.9363
NM_001308293	*SH2B1*	SH2B adaptor protein 1	2.55923	6.2626	6.45751	1.972	2.83802	1.74971
NM_001382779	*FBXL19*	F-box and leucine rich repeat protein 19	0	2.1036	0	0	0	4.51785
NR_037609	*SLX1B-SULT1A4*	SLX1B-SULT1A4 readthrough (NMD candidate)	2.7677	2.1744	0.36638	1.1669	0.85518	0.681
NR_039744	*MIR4519*	microRNA 4519	9.20512	8.1312	7.07997	9.7514	8.28084	6.69119
XR_002957780	*ZNF785*	zinc finger protein 785	2.61248	2.3704	2.79469	2.2624	2.5759	2.83624
XM_011545961	*HSD3B7*	hydroxy-delta-5-steroid dehydrogenase, 3 beta- and steroid delta-isomerase 7	2.25326	2.0611	2.84738	5.1221	1.95307	2.10077
XM_006721047	*MAZ*	MYC associated zinc finger protein	4.4146	5.446	4.00641	10.042	3.22786	4.54321
NR_037608	*SLX1A-SULT1A3*	SLX1A-SULT1A3 readthrough (NMD candidate)	4.30109	2.8933	3.461	2.4006	3.0505	2.15733
NM_017458	*MVP*	major vault protein	26.6751	0.2562	12.0196	23.711	6.27042	7.86071
NM_152288	*ORAI3*	ORAI calcium release-activated calcium modulator 3	0.70967	0.8216	2.29009	1.7024	4.50198	0.70445
NM_001308293	*SH2B1*	SH2B adaptor protein 1	0	1.0753	1.47412	0.8085	0.72648	2.06892
NM_017458	*MVP*	major vault protein	0.92084	1.3402	6.48305	2.6188	1.41335	0.88138
NM_001199142	*EIF3C*	eukaryotic translation initiation factor 3 subunit C	0.0352	0	0	0.0002	0	0.05589
NR_002966	*SNORA30*	small nucleolar RNA, H/ACA box 30	0.05589	0	0.06747	0.0593	0.04478	0
NR_002557	*LOC613038*	SAGA complex associated factor 29 pseudogene	1.68611	0.3646	0.26914	0.2656	0.16908	0.22612
NR_037609	*SLX1B-SULT1A4*	SLX1B-SULT1A4 readthrough (NMD candidate)	6.96285	4.9676	5.04827	3.2745	4.62913	3.47859

NM_001270940	*XPO6*	exportin 6	1.68684	2.9825	2.6336	3.5848	2.53949	1.9219
NR_037608	*SLX1A-SULT1A3*	SLX1A-SULT1A3 readthrough (NMD candidate)	2.96366	1.4135	0.24119	0.1768	0.35276	0.53668
XM_011545719	*ATXN2L*	ataxin 2 like	3.73875	1.6845	0.76199	1.3764	1.13379	1.22456
XM_011522838	*ADCY7*	adenylate cyclase 7	6.67738	11.419	6.67699	10.679	2.84089	25.2927
NM_017458	*MVP*	major vault protein	36.6386	35.244	7.87107	11.337	17.1945	11.5384
XM_011545719	*ATXN2L*	ataxin 2 like	12.6598	2.4598	8.57907	18.869	11.9585	17.1805
gene-RAB43P1	*RAB43P1*	RAB43 pseudogene 1	0	0.2849	8.46462	0	0	0
NM_001308293	*SH2B1*	SH2B adaptor protein 1	0.03104	3.5856	0	0.1221	0.04069	0
XM_017023909	*SETD1A*	SET domain containing 1A, histone lysine methyltransferase	0	5.1162	0.02312	0	0	0
XM_011545827	*IL4R*	interleukin 4 receptor	19.3695	0	0	0	0	0
NR_106922	*MIR6862-1*	microRNA 6862-1	0.02205	0	0	0	0	0
NM_001199142	*EIF3C*	eukaryotic translation initiation factor 3 subunit C	0	0	0	0.0803	0	0
NR_049833	*MIR3680-2*	microRNA 3680-2	0.01731	0	0	0	0	0
NR_049833	*MIR3680-2*	microRNA 3680-2	0.00451	0	0	0	0	0
XM_011545997	*RNF40*	ring finger protein 40	0	0	0	4.896	0	0
NM_052874	*STX1B*	syntaxin 1B	0	5.2877	0	0	0	0
NM_018206	*VPS35*	VPS35 retromer complex component	0	0	30.2402	0	0	0
NM_005953	*MT2A*	metallothionein 2A	153.471	0	0	0	0	0
NR_107058	*MIR6862-2*	microRNA 6862-2	0.01409	0	0	0	0	0.00016
XM_011545719	*ATXN2L*	ataxin 2 like	7.03575	0.1008	0.02863	0.2021	0.01675	0.31277
gene-BCLAF1P2	*BCLAF1P2*	BCL2 associated transcription factor 1 pseudogene 2	5.83301	0	0	0	6.77894	0
NM_004530	*MMP2*	matrix metallopeptidase 2	7.10923	19.668	287.008	20.876	15.6823	25.1023
NM_006662	*SRCAP*	Snf2 related CREBBP activator protein	3.08098	3.4616	12.2125	2.5856	3.72437	2.89609
NM_001270940	*XPO6*	exportin 6	3.03859	3.7021	4.40804	3.0502	3.68053	2.76548

NM_001199142	*EIF3C*	eukaryotic translation initiation factor 3 subunit C	0	0.0404	0	0.1578	0.13868	0.03449
NR_134855	*INO80E*	INO80 complex subunit E	10.4032	9.7829	9.22425	13.793	9.32102	10.5351
XR_933606	*N/A*	N/A	12.1106	9.5841	6.15526	9.2585	6.25996	10.0647
XM_011545719	*ATXN2L*	ataxin 2 like	0	4.5596	0.77809	0.2909	0.6926	0.57532
XM_024450231	*TNRC6A*	trinucleotide repeat containing adaptor 6A	18.5159	23.609	20.7334	18.026	21.9354	20.3506
XR_950809	*ARHGAP17*	Rho GTPase activating protein 17	8.94789	5.8414	6.71374	1.9198	4.95975	4.7429
NM_006319	*CDIPT*	CDP-diacylglycerol--inositol 3-phosphatidyltransferase	8.31944	9.6571	6.05763	12.244	10.1199	9.79039
NM_001039503	*PRSS53*	serine protease 53	31.9501	34.446	8.52179	26.028	19.5562	22.9878
NR_039742	*MIR4517*	microRNA 4517	2.65913	0.9535	1.69512	2.4777	2.84924	1.64807
NR_039742	*MIR4517*	microRNA 4517	2.51625	0.919	1.60964	2.3542	2.66616	1.5788
NM_017458	*MVP*	major vault protein	3.20658	2.2989	13.4364	5.2334	8.06032	7.21485
NM_001173984	*BRD7*	bromodomain containing 7	9.72096	8.9638	0	6.8832	11.1263	4.80211
NM_001365304	*LOC112694756*	uncharaterized LOC112694756	0.80356	1.1219	10.1828	1.2294	8.90108	1.16532
XR_950809	*ARHGAP17*	Rho GTPase activating protein 17	5.76072	6.328	19.9641	15.356	21.2913	7.62444
NR_037573	*DCTN5*	dynactin subunit 5	4.13153	10.472	9.25709	0	2.12305	6.83063
NM_001170634	*FUS*	FUS RNA binding protein	87.9152	36.448	66.7225	61.267	54.8335	59.6427
NR_031576	*MIR762*	microRNA 762	0.00407	0	0.00705	0.0031	0.00527	0
XM_017023626	*NPIPB8*	nuclear pore complex interacting protein family member B8	3.69974	1.6704	2.75203	2.774	2.2321	3.72559
XM_017023376	*LPCAT2*	lysophosphatidylcholine acyltransferase 2	44.4325	49.356	86.8886	65.63	22.6024	117.401
XM_011545719	*ATXN2L*	ataxin 2 like	0.26561	0.7006	1.28896	0.7103	0.94493	0.35181
NM_001308293	*SH2B1*	SH2B adaptor protein 1	4.80789	2.6512	3.16185	4.3477	3.52462	4.35822
rna-MIR4519	*MIR4519*	microRNA 4519	0.00016	0	4	0.0003	0.00019	0
NM_001039503	*PRSS53*	serine protease 53	18.4766	10.892	35.5906	14.666	34.3338	9.48792

NM_001308293	*SH2B1*	SH2B adaptor protein 1	8.51088	4.1698	4.03158	1.2346	5.31462	2.86053
XM_006721047	*MAZ*	MYC associated zinc finger protein	2.7969	2.2485	1.69507	1.684	0.42801	3.45052
XR_001752034	*TAOK2*	TAO kinase 2	19.7843	12.662	10.8467	16.37	17.6863	13.2311
XM_011545971	*KAT8*	lysine acetyltransferase 8	14.1562	16.132	13.8902	12.934	14.2593	14.5231
XM_024450448	*ARMC5*	armadillo repeat containing 5	0	3.5826	0	0	0.33053	1.15908
NM_024816	*RABEP2*	rabaptin, RAB GTPase binding effector protein 2	0.77751	0.8678	1.32484	1.1956	1.08948	1.03651
NM_001270940	*XPO6*	exportin 6	1.24885	4.2007	1.8925	2.0765	2.99161	1.92103
NR_039872	*MIR4721*	microRNA 4721	0.01087	0.005	0.03384	0.1006	0.06102	0.00765
NR_037608	*SLX1A-SULT1A3*	SLX1A-SULT1A3 readthrough (NMD candidate)	1.18823	1.2841	0.81198	2.101	1.18349	1.44363
XM_024450467	*SPNS1*	sphingolipid transporter 1 (putative)	6.78422	8.0387	5.79877	3.8894	8.22055	4.18767
gene-HNRNPA1P48	*HNRNPA1P48*	heterogeneous nuclear ribonucleoprotein A1 like 3	0	0	0	6.4018	3.16085	0
NM_017458	*MVP*	major vault protein	7.12131	6.2377	13.8318	9.0158	7.99416	10.0918
NR_106829	*MIR6771*	microRNA 6771	0	0	0	0	0	0

Supplementary Table 4.9. Cycle threshold (C_T) and standard error of C_T values obtained from all batches of qRT-PCR.

Batch 1			
Sample Name	**Target Name**	**C_T Mean**	**Standard error of C_T**
C2	GAPDH	19.7006	0.053
C1	GAPDH	19.5663	0.053
C3	GAPDH	20.0597	0.053
C4	GAPDH	20.4087	0.053
IV:4	GAPDH	19.8560	0.053
IV:9	GAPDH	20.2057	0.053
C2	RPLP0	20.9511	0.049
C1	RPLP0	20.8018	0.049
C3	RPLP0	20.8518	0.049
C4	RPLP0	21.3702	0.049
IV:4	RPLP0	20.7950	0.049
IV:9	RPLP0	20.4627	0.049
C2	RDH10	26.8728	0.305
C1	RDH10	25.8996	0.305
C3	RDH10	29.3817	0.305
C4	RDH10	28.1008	0.305
IV:4	RDH10	24.1219	0.305
IV:9	RDH10	26.1516	0.305
C2	ZFHX4	28.7498	0.075
C1	ZFHX4	28.6361	0.075
C3	ZFHX4	29.0635	0.075
C4	ZFHX4	29.0375	0.075
IV:4	ZFHX4	27.9794	0.075
IV:9	ZFHX4	28.1671	0.075
C2	ZNF704	33.7830	0.335
C1	ZNF704	34.1531	0.335
C3	ZNF704	32.3446	0.335
C4	ZNF704	32.5638	0.335
IV:4	ZNF704	29.5375	0.335
IV:9	ZNF704	29.5782	0.335
Batch 2			
Sample Name	**Target Name**	**C_T Mean**	**Standard error of C_T**
C2	*GAPDH*	19.6331	0.074
C1	*GAPDH*	19.5805	0.074
C3	*GAPDH*	19.7091	0.074
C4	*GAPDH*	19.9153	0.074
IV:4	*GAPDH*	19.5149	0.074
IV:9	*GAPDH*	20.6989	0.074
C2	*RPLP0*	20.8494	0.044
C1	*RPLP0*	20.9558	0.044
C3	*RPLP0*	20.8915	0.044
C4	*RPLP0*	21.3360	0.044
IV:4	*RPLP0*	20.9443	0.044
IV:9	*RPLP0*	20.5088	0.044
C2	*PRRT2*	33.1277	0.215
C1	*PRRT2*	33.7032	0.215
C3	*PRRT2*	32.3620	0.215
C4	*PRRT2*	32.7423	0.215

IV:4	*PRRT2*	34.0549	0.215
IV:9	*PRRT2*	35.9747	0.215
C2	*BCL7C*	28.9392	0.108
C1	*BCL7C*	28.7060	0.108
C3	*BCL7C*	28.2377	0.108
C4	*BCL7C*	28.8127	0.108
IV:4	*BCL7C*	29.0590	0.108
IV:9	*BCL7C*	30.1819	0.108
C2	*YPEL3*	26.0002	0.104
C1	*YPEL3*	26.8275	0.104
C3	*YPEL3*	27.2429	0.104
C4	*YPEL3*	27.1871	0.104
IV:4	*YPEL3*	25.7205	0.104
IV:9	*YPEL3*	26.4722	0.104
	Batch 3		
Sample Name	***Target Name***	**C_T Mean**	**Standard error of C_T**
C2	*GAPDH*	19.9201	0.030
C1	*GAPDH*	20.0746	0.030
C3	*GAPDH*	20.2241	0.030
C4	*GAPDH*	20.1090	0.030
IV:4	*GAPDH*	20.1381	0.030
IV:9	*GAPDH*	20.4653	0.030
C2	*RPLP0*	21.7926	0.054
C1	*RPLP0*	21.4611	0.054
C3	*RPLP0*	21.5886	0.054
C4	*RPLP0*	22.0547	0.054
IV:4	*RPLP0*	21.4580	0.054
IV:9	*RPLP0*	21.0964	0.054
C2	*IRX6*	34.5815	0.234
C1	*IRX6*	35.1546	0.234
C3	*IRX6*	34.3629	0.234
C4	*IRX6*	36.6914	0.234
IV:4	*IRX6*	32.8071	0.234
IV:9	*IRX6*	33.1852	0.234
C2	*IRX3*	26.8564	0.112
C1	*IRX3*	27.6840	0.112
C3	*IRX3*	27.7246	0.112
C4	*IRX3*	27.2908	0.112
IV:4	*IRX3*	26.0304	0.112
IV:9	*IRX3*	26.5343	0.112
C2	*SHCBP1*	29.9586	0.190
C1	*SHCBP1*	29.2172	0.190
C3	*SHCBP1*	27.9743	0.190
C4	*SHCBP1*	28.1128	0.190
IV:4	*SHCBP1*	30.8467	0.190
IV:9	*SHCBP1*	30.0208	0.190

Chapter 5

Identification and multi-omics analysis of the genomic variants in CMT720

In the previous chapter we performed transcriptomic profiling on patient fibroblasts from CMT720 and identified 4 positional candidate genes from RNA-seq data that was further validated by qRT-PCR analysis. Since RNA-seq analysis on fibroblasts identified no aberrant transcripts, our findings suggest that the molecular mechanism of the putative noncoding variant causing the disease in CMT720 most likely involves disruption of gene regulation. In this chapter we will use WGS to identify the full spectrum of non-coding genomic variants (SNVs, indels, SVs and REs) in CMT720 patients. Following variant filtering, noncoding mutations that are most likely to dysregulate the 4 positional candidate genes will be selected as the variants with the highest potential for pathogenicity. Epigenomics data will be used to complement the transcriptomic findings by identifying the CREs associating with the dysregulated positional candidates. Variants localising to the introns, UTRs and CREs of the dysregulated positional candidates will be prioritised and undergo variant analysis with the goal of identifying the potential noncoding variant causing CMT720.

It is predicted that a considerable proportion of non-coding mutations will localise to repetitive and structurally complex regions that constitute 54% of the human genome [237]. Notably, the suggestive linkage region on chromosome 16 contains large segmental duplications [238] as well as centromeric and pericentromeric repeat arrays [142], which are the most challenging regions to achieve sufficient coverage for accurate variant calls using WGS [106, 109, 239]. Furthermore, most of these regions remain as gaps in the hg38 reference which is the predominant reference used [142]. The major task in this chapter will be to determine the appropriate sequencing platform for performing WGS and the genome

reference for aligning the sequence data to maximise the detection of variants within the prioritised suggestive linkage regions identified in CMT720.

5.1. Using lrWGS for identifying variants with higher accuracy and increasing the coverage of the genome

The insufficient length of the short sequencing reads produced by srWGS platforms, results in misalignment of reads in repetitive regions, leading to both false positive and false negative variant calls, as well as inaccurate estimations of variant size [240-242]. The limitations imposed by read length are most detrimental for the accurate detection of large or repetitive variants such as REs and SVs, but also cause poor detection of SNVs and indels localising to repetitive regions [243, 244]. For example, with more than 72% of SVs locating within or flanked by STRs [245], srWGS produces up to 89% false positive SV calls [241, 246] and misses more than half of the SVs in an average genome [247]. Additionally, it is estimated that 748 genes remain inaccessible to sequencing with srWGS due to poor alignment quality [248].

Long read WGS (lrWGS) platforms have been increasingly implemented in genetic diagnosis of rare diseases due to substantial improvements in variant detection compared to srWGS platforms [161, 249]. The most widely used lrWGS platforms are provided by Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) [250]. The ONT platforms provide the longest read length of all WGS technologies, ranging on average from 10-100 kb but can reach 1 Mb depending on the library preparation methods [106, 111]. This platform however has a high base calling errors reaching 8% [251]. The most recent PacBio platforms produce highly accurate long reads (HiFi reads) that are10-20 kb long and show 99.9% base calling accuracy [109]. While both ONT and PacBio HiFi platforms can detect SVs at

comparable rates [242], PacBio HiFi achieves >99% accuracy at detecting SNVs and indels while ONT misses approximately 3% and 20% of these variant types in benchmark studies, respectively [252]. Compared to Illumina srWGS, PacBio HiFi can detect an additional 33% more REs >150bp [240] and 82% more SVs [242], while the SNV and indel calls show high concordance between these platforms averaging at 92% [243, 253]. Furthermore, PacBio can achieve high coverage across 52% the repetitive regions that cannot be sequenced by srWGS, and improves mapping and variant calling quality across 193 clinically relevant genes [109]. In the current investigation of CMT720, we will use the existing srWGS data from patients IV:4 and V:6, and will also perform PacBio HiFi sequencing on patient IV:4 to identify the noncoding variants with higher accuracy across a larger proportion of the suggestive linkage regions on chromosomes 8 and 16.

5.2. The gapless T2T reference for improved variant detection and analysing previously inaccessible regions in CMT720

Approximately 8% of the hg38 human reference genome that is used by most genetic studies remain as gaps of unresolved sequence [142], which gives rise to poor mapping quality, as well as false positive and false negative variant calls in both srWGS and lrWGS alignments [141]. The novel telomere-to-telomere (T2T) reference adds 182 Mb of sequence that is absent from hg38 [142], which reduces the rate of mismatching bases in both Illumina srWGS and PacBio lrWGS, and facilitates even coverage of the genome [141]. The increased quality of alignment against T2T results in higher accuracy and sensitivity for detecting all types of variants with the largest improvements observed in the detection of SVs [141, 237, 254]. Aligning PacBio HiFi reads against the T2T reference results in an increase of 16% in the number of SVs identified and eliminates the false positive insertion and duplication calls caused by shortened STR representations in hg38 [141, 255]. To improve the detection of

variants and coverage of the genome with WGS, existing srWGS data from patients IV:4 and V:6, and the lrWGS data that will be obtained from patient IV:4 will be aligned against T2T. By using T2T for alignment, we expect to identify a higher number of variants in CMT720 patients with greater accuracy. Additionally, alignment against T2T can potentially allow access to 722 kb and 16.3 Mb of sequence respectively, across the suggestive linkage regions on chromosomes 8 and 16 that are absent in the hg38 alignment (Figure 5.1)[142]. These gaps correspond to segmental duplications and centromeric repeat arrays that are harbouring gene densities unique to the T2T reference (Figure 5.1) which may include functional paralogues of protein coding genes [256] or novel protein coding genes identified in the T2T assembly [142]. If any protein coding genes unique to the T2T reference are identified across the gaps completed in our suggestive linkage loci, these previously inaccessible genes will be screened for mutations.

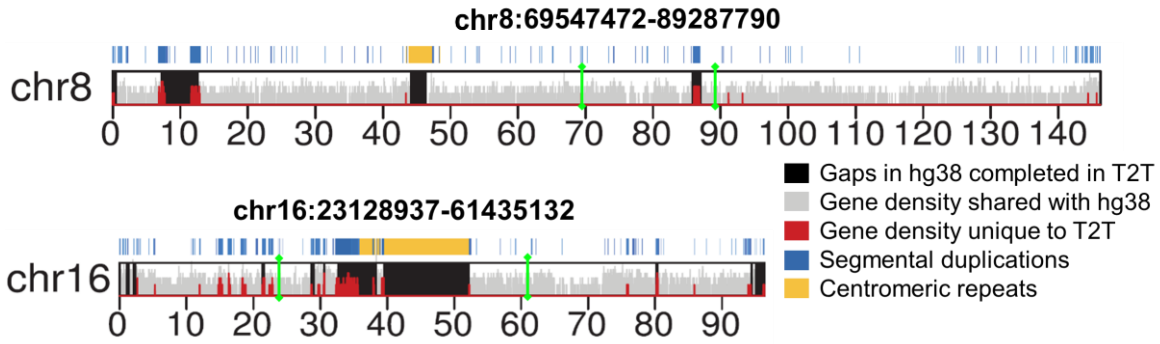


Figure 5.1: The annotated ideograms of chromosomes 8 and 16 in the T2T reference showing comparisons of coverage and gene density with respect to hg38 across the suggestive linkage regions on chromosome 8 and 16. The green bars demarcate the borders of suggestive linkage regions and the coordinates of each linkage region in the T2T reference is provided above each ideogram. The descriptions of the genomic features annotated on ideograms are provided in the legend. Ideograms adapted from Nurk *et al.* [142].

5.3. Using the draft human pangenome reference and 1000 Genomes Project resources for variant filtering

With increased detection of variants anticipated due to the use of T2T and lrWGS, effective variant filtering remains crucial for facilitating the analysis of the candidate pathogenic variants identified in CMT720 patient srWGS and lrWGS alignments. While T2T offers unique advantages for variant detection, T2T annotation of variant databases are currently scarce for this genome reference, limiting the resources available for variant filtering. Currently, only two population variant databases have been developed by aligning WGS data against T2T: the SNVs and indel reference based on the realignment of 1000 Genome Project (1KGP) srWGS data [142] and the SVs identified in the PacBio HiFi alignments of the 1KGP samples [247]. While the variants reported in the population database dbSNP [257] against hg38 have been lifted to T2T [142], lift over is known to cause erroneous variant representations [258, 259], accordingly, this dataset will not be used in the current investigation as a conservative measure. Therefore, the references of normal variants that are suitable to use in this study are further limited to the 1KGP resources based on 3,202 individuals, which may fail to provide sufficient filtering power to reduce the number of noncoding mutations to a manageable number across the suggestive linkage regions on chromosome 8 and 16.

In addition to serving as a template for aligning WGS data, the draft human pangenome (DHPG) reference is a resource of highly accurate variant calls that can be used to identify the normal variation in the human genome [144]. The DHPG was constructed based on 94 fully-phased and *de novo* assembled haplotypes of 47 healthy individuals from a broad ethnic background which substantially increases the diversity of variants and minimises reference biases [144]. Due to these advantages, DHPG comprises approximately 20 million SNVs, 7 million indels and 400 thousand SVs [144] that captures the “normal” variation in the global population much more effectively compared to the population databases constructed against

a single reference genome (see Figure 5.2A for an example of variant representation in the pangenome reference) [144]. Since existing genome assemblies can be integrated into the pangenome, there are publicly available versions of the DHPG incorporating the T2T or hg38 references, which allows representation of all variation in the DHPG against T2T coordinates [144]. For this study, the DHPG will have high utility as a population reference of benign polymorphic variants with respect to the T2T reference.

To harness the power of filtering using the DHPG, the tool PanGenie will be utilised [260]. PanGenie is a genotyping algorithm, that determines whether the variants present in DHPG are also present in sample srWGS data [260]. This is achieved by a process known as genome inference [261] where the sequence of each variant in the DHPG is used to construct unique k-mers, which are sequence fragments with a set length of “k” that are used by PanGenie like *in silico* probes [260] (Figure 5.2B). PanGenie then searches the FASTQ files containing unaligned srWGS reads for matches with the k-mers that uniquely represent an allele, sums the counts of matches, and compares these counts to the expected sequencing depth to make a genotyping call (Figure 5.2B) [260]. In cases where the numbers of unique k-mers are too low for a genotype call, PanGenie uses the nearby polymorphic loci in the reference haplotypes as markers to calculate the linkage between variant alleles and determines the most likely genotype [260, 262]. Since genome inference does not involve read alignment, PanGenie overcomes the limitations in variant detection caused by the read mapping errors in srWGS alignments and increases the detection of benign SVs by 104% and small variants (<50 bp) by 38% compared to the variants identified in the 1KGP alignments against the hg38 reference [144, 260]. PanGenie provides the largest improvements in identification of deletions <300 bp and insertions [144], which are particularly challenging to detect and call with srWGS alignments [263]. In this investigation, we will use PanGenie to genotype srWGS data from both CMT720 patients and our in-house controls with the T2T-inclusive build of the DHPG. This strategy will increase filtering power by permitting extraction

of the latent information from patient srWGS data [144] to obtain a set of benign SNVs, indels and SVs identified against T2T coordinates with minimal reference biases.

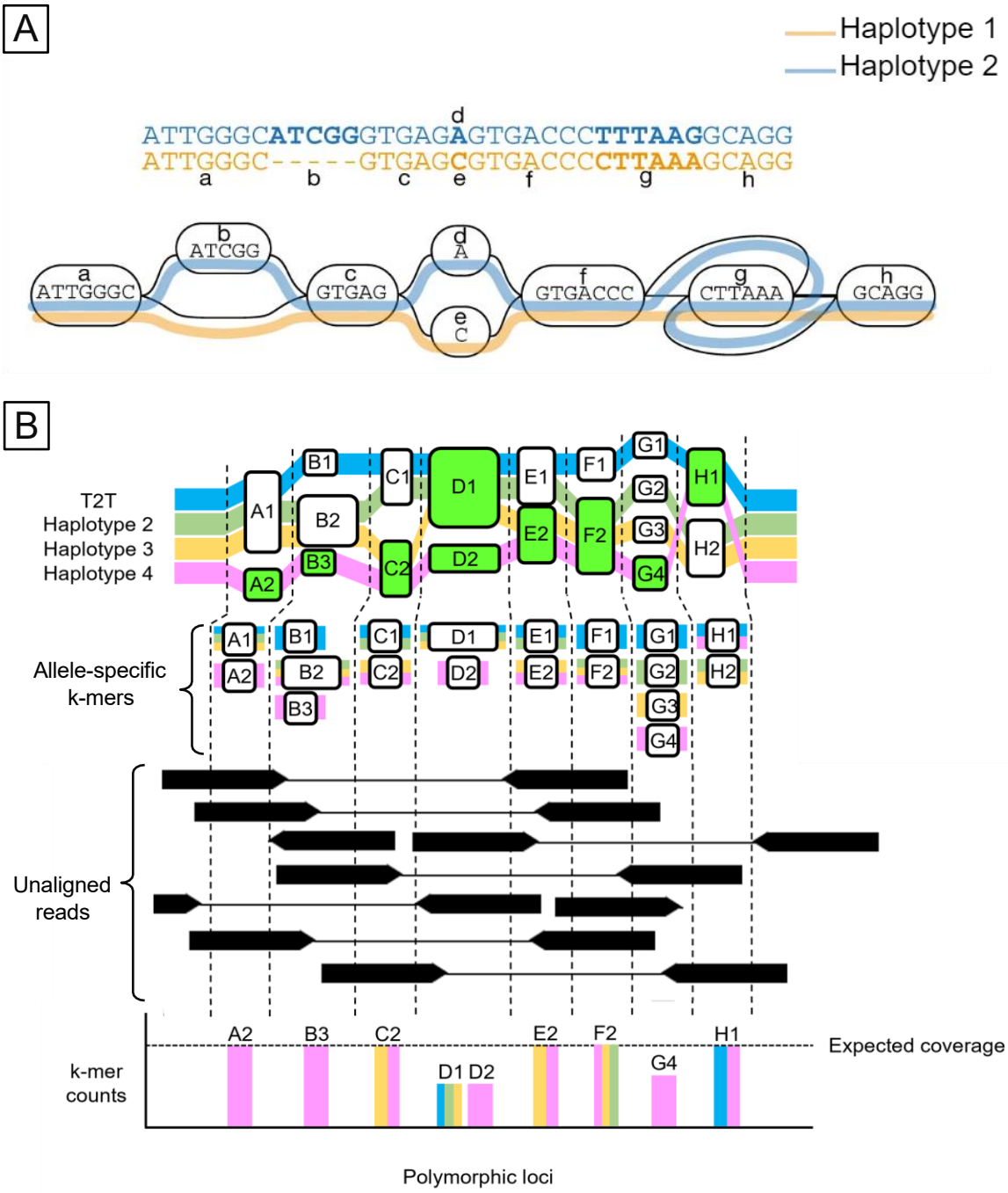


Figure 5.2: Representation of variants in pangenome references and workflow of genome inference. (**A**) Representation of DNA sequence variation represented in an alignment between two haplotypes (top) and in a pangenome graph (bottom). Corresponding sequences between the two representations are indicated with lowercase letters. In a pangenome graph, sequences that are common between the haplotypes are represented as a single node (a, c and f) while the variants are represented as divergent nodes (b, d, e and g). Panel A was adapted from [144]. (**B**) The genome

inference performed by PanGenie that will be utilised in the current investigation of CMT720 as described by Ebler *et al.* [260] and Hantze *et al.* [262]. PanGenie uses the T2T FASTA file and a VCF file that represents the variants in a the DHPG with their coordinates indicated against the T2T reference. These files are used to design k-mers that are unique to the sequence of each allele at a polymorphic locus. K-mers are then used to search for matching sequences in the unaligned srWGS data. The number of matches with the allele-specific k-mers are counted and compared against the expected level of coverage to make homozygous or heterozygous genotype calls.

5.4. Using epigenomics data for identifying potentially pathogenic noncoding variants in functionally relevant CREs

As briefly mentioned in Chapter 1, CREs are regions of noncoding DNA motifs that control the expression of nearby genes by recruiting transcription factors (TF) [163] and interacting with the promoters of their gene targets through 3 dimensional (3D) chromatin contacts (Figure 5.3A)[264, 265]. Promoters and enhancers are the most well-studied types of CREs that predominantly have positive influence on gene expression [266] whereas silencers negatively regulate transcription by recruiting repressor proteins [267, 268]. Another class of CREs known as insulators or boundary elements organise groups of other regulatory elements and their target genes into topologically associated domains (TADs) (Figure 5.3A) in which regulatory chromatin interactions take place frequently while CRE-promoter contacts rarely occur across TADs [264, 269]. Determining the functional association between a CRE and a gene can be challenging since it is estimated that there are approximately one million CREs in the human genome each of which may interact with multiple genes [270]. Furthermore, CREs might be located up to 1 Mb away from their gene targets [271] and most show tissue specific and temporally restricted activity [272, 273].

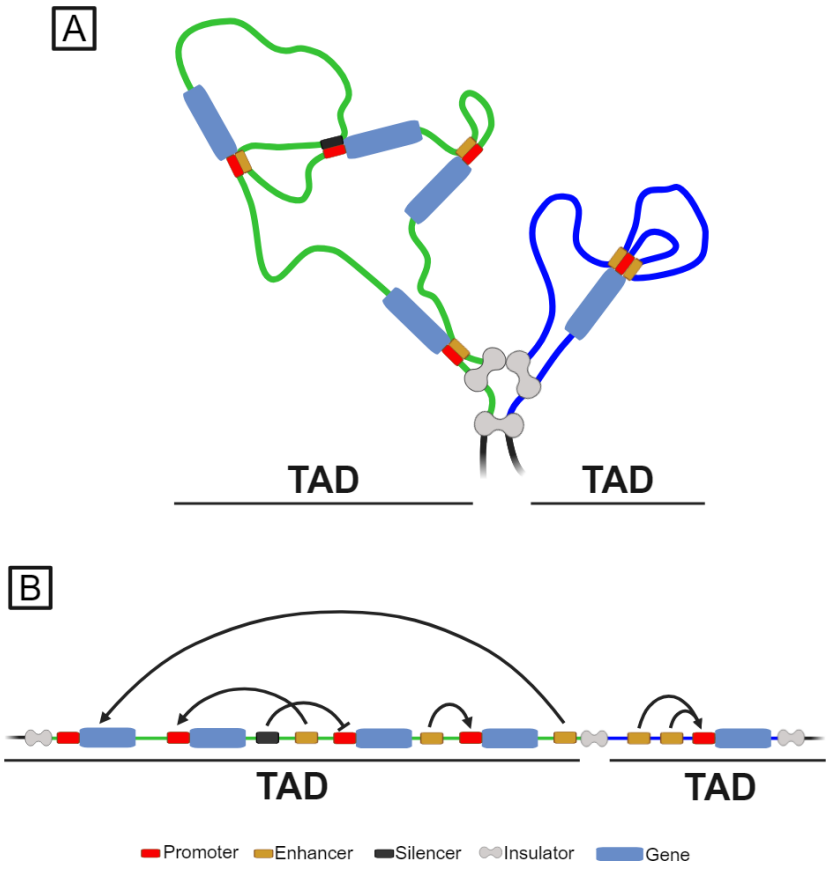


Figure 5.3: Regulatory interactions mediated by chromatin organisation and 3D chromatin contacts. (**A**) A simplified diagram showing regulation of gene expression by 3D contacts between CREs and the promoters of their gene targets. Insulator elements organise regulatory interactions into discrete groups by forming loops that comprise TADs [264] (**B**). The linear representation of the 3D regulatory contacts shown in panel B. This figure was created with BioRender.

5.4.1. Using promoter capture Hi-C to select potential regulatory variants impacting dysregulated positional candidates

The most efficient epigenomic method for determining the regulatory relationship between CREs and genes with high specificity is to use assays for analysing the 3D chromatin contacts in tissues of interest [273, 274]. The 3D chromatin contacts are identified by a method called chromatin conformation capture (3C) [275]. In 3C, two distantly located regions of DNA brought into close proximity by 3D chromatin contacts, are fixed by crosslinking in cells (Figure

5.4A) [275]. This is followed by the digestion of crosslinked DNA with enzymes, filling in fragment ends and biotin labelling (Figure 5.4A) [275]. The DNA fragments are then ligated together [275] (Figure 5.4A). The DNA is then purified and the chimeric fragments with biotin are pulled down for sequencing [275] (Figure 5.4A). By sequencing these chimeric fragments, the fragments can be aligned to the regions participating in chromatin interactions [276] (Figure 5.4A). Combined use of 3C with NGS to capture and sequence all regions across the genome that participate in 3D chromatin interactions is known as Hi-C, and this method produces a map of coordinates delineating each pair of interacting regions [269, 277]. Promoter capture Hi-C (PCHi-C) is a variation of the Hi-C technologies that utilises probes to capture all chimeric fragments containing gene promoters, which achieves 11 to 15 fold enrichment of chromatin interactions with gene promoters (Figure 5.4B) [265, 278]. On average, each significant promoter interacting region (PIRs) identified by PCHi-C (Figure 5.4B) represents a robust chromatin interaction supported by thousands of sequencing reads [274]. PIRs are highly enriched for CREs, making PCHi-C an ideal epigenomic method for identifying tissue-specific regulatory interactions [265, 274, 279].

Due to the time restrictions that prevent obtaining patient derived-MNs, it is not feasible to perform PCHi-C on CMT720 using disease relevant tissue in the current study to identify the CREs associated with the dysregulated positional candidates. However, through large collaborative efforts of the ENCODE [273] project, collections of PCHi-C data obtained from neuronal tissues are publicly available. In this chapter, online PCHi-C data obtained from neuronal tissue will be used for determining the PIRs that contact with the promoters of dysregulated positional candidate genes identified in Chapter 4. The noncoding variants located in these PIRs will be prioritised for analysis. Since PIRs often contain other sequences in addition to CREs [274], accessing additional epigenomics datasets during bioinformatic variant analysis will be necessary to ensure that the prioritised variants are located within CREs. Using additional epigenomics data can also provide further evidence of pathogenicity

by confirming the localisation of a regulatory variant to a TF binding site (TFBS) [127].

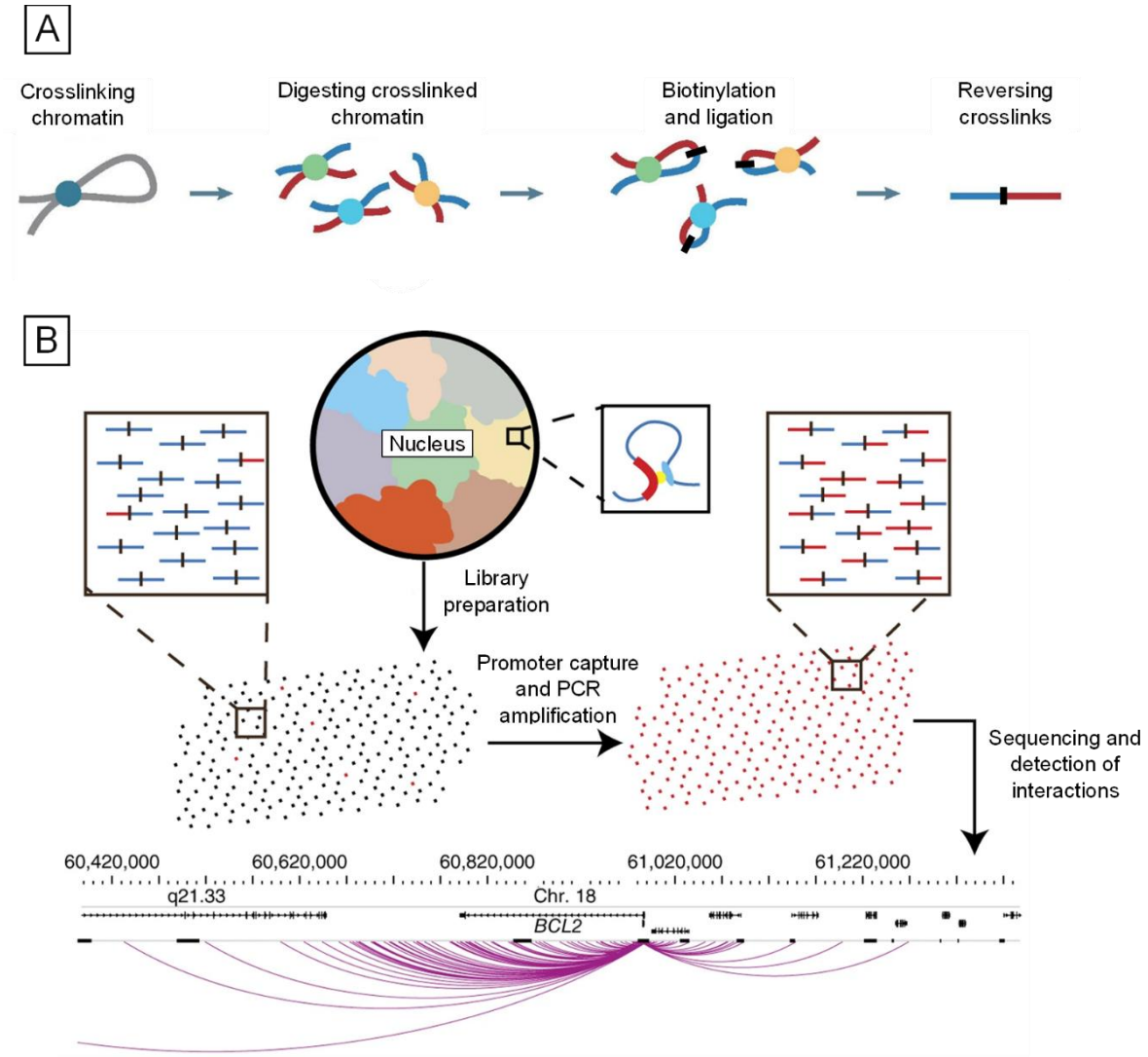


Figure 5.4: Library preparation and experimental workflow of PCHi-C. (**A**) A common library preparation workflow is used by 3C-based methods. The regions of DNA participating in 3D contacts are crosslinked, digested with restriction enzymes, biotinylated and ligated. Crosslinks are reversed to obtain chimeric fragments. Information adapted from Li *et al.* [280]. (**B**) In PCHi-C, the chimeric fragments harbouring gene promoters are selectively captured by oligonucleotide probes, amplified by PCR and sequenced using NGS. Mapping the sequences of chimeric fragments allow identifying the PIRs interacting with the promoters of genes. In this exemplary contact map, the location of PIRs on chromosome 21 interacting with the promoter of *BCL2* are indicated by purple arcs. Black boxes represent the positions of all promoters captured in this PCHi-C experiment. Information was adapted from Mifsud *et al.* [279]. Panel A was adapted from Li *et al.* [280] and panel B was adapter from Mifsud *et al.* [279].

Using transcriptomic analysis combined with epigenomics data to select and prioritise

the candidate pathogenic variants identified by WGS in patients from family CMT720 will likely highlight multi-omics variant prioritisation as an effective approach for IPN research. In addition to contributing to the molecular characterisation of CMT720, the findings of this investigation will provide insights into the utility of multi-omics strategies as well as completeness and representation of genetic diversity in human genome references for investigating the noncoding genome in unsolved IPNs.

5.5. Hypothesis

We hypothesise that lrWGS and new improved genomics resources will facilitate accurate identification and efficient filtering for the full spectrum of noncoding variants in CMT720 patients. Using transcriptome guided analysis and PCHi-C data will assist in the selection of candidate noncoding pathogenic variants associated with the dysregulated positional gene candidates in the chromosome 8 and 16 suggested linkage regions.

5.6. Aims

The specific aims are to:

1) Perform lrWGS on CMT720 patient IV:4 and identify SNVs, indels, REs and SVs using both srWGS and lrWGS alignments of CMT720 patients.

2) Perform variant filtering and use PCHi-C data to select all variants localizing to the introns, UTRs or PIRs of the dysregulated positional candidates for analysis.

3) Perform segregation and bioinformatic analysis on the selected candidate variants to prioritise those that will undergo further functional studies to support a pathogenic role in family CMT720.

5.7. Methods

5.7.1. Long read whole genome sequencing

As part of grant support from Medical Research Future Fund (MRFF) Genomics Health Futures Mission (APP 2007681) in which Professor Kennerson is CIB, PacBio long-read WGS was outsourced to Dr. Ira Deveson at the Garvan Institute of Medical Research (Australia). Dr. Deveson coordinated DNA preparation, library preparation, sequencing of the sample and performed the bioinformatic pipelines for sequence alignment and variant calling. Fibroblasts of patient IV:4 from CMT720 were cultured to confluence as previously described (Chapter 4) and ~3 million cells were sent for DNA preparation. Briefly, Nanobind CBB kit (PacBio) was used to extract high molecular weight DNA (50-300kb) from patient fibroblasts and SMRT Bell 3.0 prep kit (PacBio) was used for ligating adapter sequences to construct a circularised sequencing library. The PacBio Revio platform was used to perform WGS at a sequencing depth of >30x coverage. The patient FASTQ file was aligned to both the GRCh38 (hg38) reference with no alternative haplotypes [143] and the chm13v2.0 (T2T) reference [142] using minimap2 [281]. SNVs and indels in both hg38 and T2T alignments of the lrWGS data were identified using clair3 and SVs were identified using clair3 [282] and Sniffles2 [283]. Quality control on patient lrWGS alignment was performed by using Samtools [284].

5.7.2. Variant identification using srWGS data

Alignment of srWGS to the T2T reference and variant calling and the PanGenie genotyping was outsourced to Dr. Georgie Samaha through the Sydney Informatics Hub (SIH), a Core Research Facility of the University of Sydney and the Australian BioCommons which is enabled by NCRIS via Bioplatforms Australia. For the high performance computing the project utilised the National Computational Infrastructure (NCI) supported by the Australian

Government and the Sydney Informatics Hub HPC Allocation Scheme, which is supported by the Deputy Vice-Chancellor (Research), University of Sydney and the ARC LIEF, 2019: Smith, Muller, Thornber et al., Sustaining and strengthening merit-based access to National Computational Infrastructure (LE190100021).

Briefly, srWGS FASTQ files from CMT720 patients IV:4 and V:6, and 3 healthy controls (Controls 1 to 3) were aligned against the chm13v2.0 (T2T) reference [142] using the fq2bam tool from the Clara Parabricks software suite (NVIDIA). SNVs and indels from the T2T alignments were identified using DeepVariant [285]. This pipeline has been made available to the public. Quality control on patient srWGS alignment was performed by using Samtools [284].

The REs were called against the hg38 alignments of the srWGS data from CMT720 patients IV:4 and V:6 through a collaboration with Dr. Chiara Folland at University of Western Australia. This pipeline utilises the ExpansionHunter Denovo (EHDN) caller which can detect all REs across the genome without a reference of known REs [240]. Cohort-level RE calls were identified in the CMT720 patients and the HiSeq X Diversity Cohort from the Polaris Project (Illumina) comprising 150 healthy samples selected from the 1KGP. A post-processing script that is available from the str-analysis repository of Broad Institute (https://github.com/broadinstitute/str-analysis.git) was deployed to annotate and perform case-control analysis on the resulting REs.

5.7.3. Strategy developed for filtering SNV and indel calls using genomic resources and bioinformatic tools

5.7.3.1. Genotyping patient and control srWGS data using PanGenie

PanGenie genotyping the control and CMT720 patient srWGS data against the DHPG was performed using the FASTQ files from the patients IV:4 and V:6 and 3 controls [260] with

a default k-mer length of 21 bp. The genotypes were computed using the T2T reference genome (chm13v2.0.fa) (https://github.com/marbl/CHM13.git) [142] and a graph VCF comprising 88 phased haplotypes (HPRC-CHM13 pangenome graph) (https://github.com/eblerjana/pangenie.git), which represents the sequence of each variant as a phased genotype co-ordinate resolved against the T2T reference [144, 260]. As the reference VCF used by PanGenie represents a pangenome graph, all overlapping genotype calls made by PanGenie are combined into multiallelic variant records [144, 260], which are not compatible for filtering against the variants in CMT720 patients identified by alignment against the T2T reference. Therefore, the complex multiallelic genotype calls produced by PanGenie from the controls and CMT720 patients were converted to biallelic calls using the 'convert-to-biallelic.py' script made available by the developers of PanGenie (https://github.com/eblerjana/pangenie.git). To make the other files used in the filtering process compatible with the biallelic PanGenie calls, all patient, control and reference callsets will also be converted to biallelic calls.

5.7.3.2. Preparing benign variant callsets from PanGenie and 1KGP

The variant call file (VCF) [286] manipulation program BCFtools [284] was used to prepare the VCF files obtained from population genomics resources (1KGP) and the PanGenie genotyping for filtering out normal variation from patient callsets. The VCF files containing all variants on chromosome 8 and chromosome16 from the 1KGP T2T alignment callset were obtained from the GitHub T2T reference repository (https://github.com/marbl/CHM13.git) [142]. Variants in the 1KGP T2T VCF files with minor allele frequencies less than 1/1000 (MAF<0.001) were removed (Supplementary Table 5.1). This step was performed to obtain callsets of benign variants with MAF>0.001, which will be excluded from patient calls during filtering process. All multiallelic variant calls in the filtered 1KGP VCF files were converted to

biallelic calls (Supplementary Table 5.1). All PanGenie VCF files containing the biallelic genotype calls from patients and controls were merged into a single file (Supplementary Table 5.2).

5.7.3.3. Filtering strategy to identify the variants localising to suggestive linkage regions on chromosome 8 and 16

The filtering strategy to identify candidate variants in the suggestive linkage regions is shown in Figure 5.5. Variant calls from the patient and control VCF files with a Phred-scaled quality score of <20 were removed (Supplementary Table 5.3) (Figure 5.5A). Control VCF files were combined into a single VCF file (Supplementary Table 5.4) and patient VCF files were merged by retaining only the variant calls present in both patients (Figure 5.5B) (Supplementary Table 5.5). This step could not be performed for the lrWGS VCF (patient IV:4) as a second patient with lrWGS was not available (Figure 5.5B). Variants in controls were excluded from the patient V:6 and IV:4 combined callset (Figure 5.5C). The coordinates of prioritised suggestive linkage regions were lifted from hg38 to T2T using UCSC LiftOver tool [192] and variant calls located outside the prioritised suggestive linkage regions were removed (Figure 5.5D). Variants on chromosomes 8 and 16 in the pre-processed 1KGP callsets were excluded from the patient calls to remove the benign variants with MAF>0.001 (Figure 5.5E). Normal variants in the merged PanGenie VCF were removed from the patient callset (Figure 5.5F). Homozygous calls were removed (Figure 5.5G) and the variant calls from each prioritised linkage region were grouped into separated VCF files (Figure 5.5H). All code used for filtering the VCF files of patient srWGS and lrWGS alignments are provided in Supplementary Table 5.5 and Supplementary Table 5.6 respectively.

In all steps, intermediate files were compressed using bgzip and were indexed using tabix when required [284]. Variant counts were obtained by using the BCFtools stats command

on the intermediate and output files.

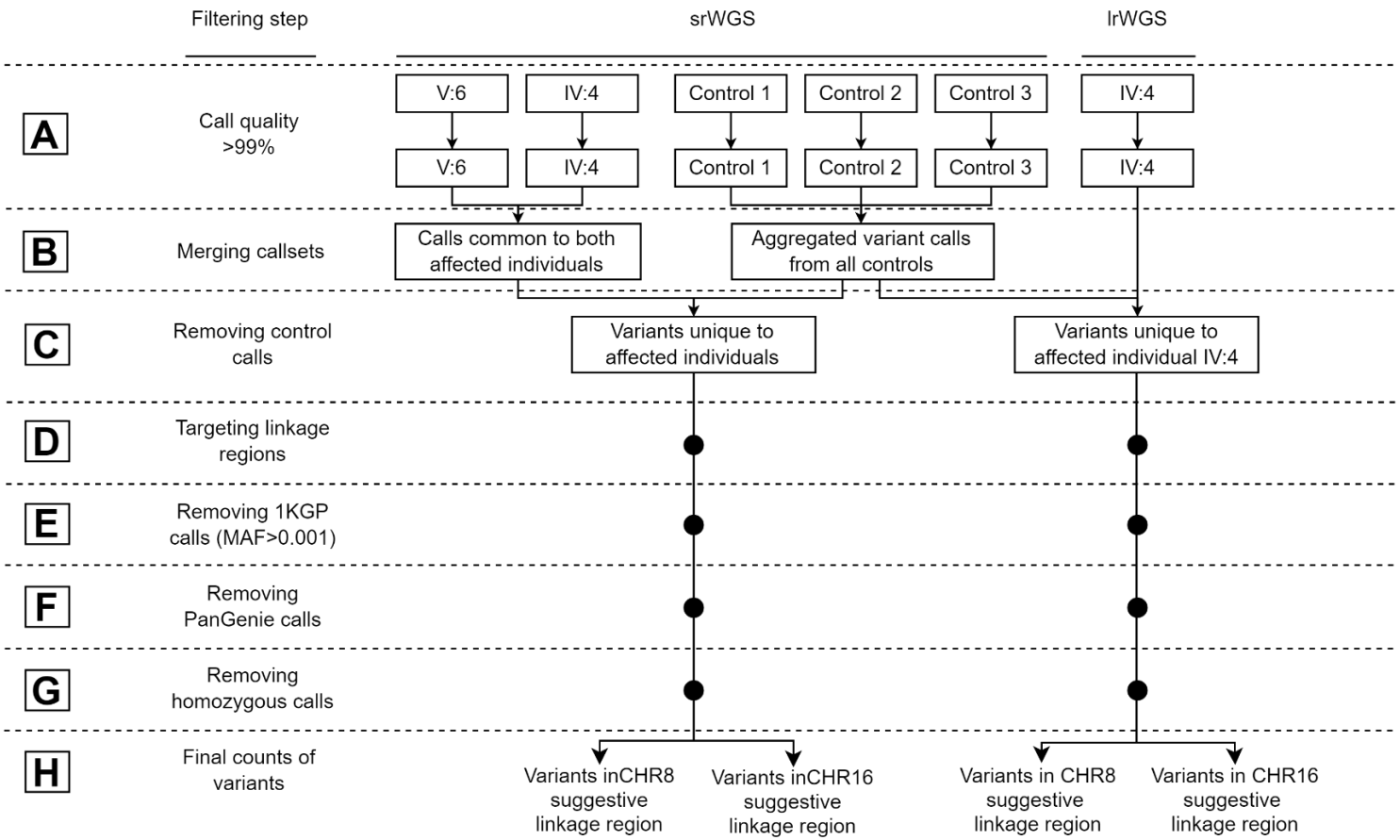


Figure 5.5: The workflow of variant filtering performed on the VCF files obtained from CMT720 patient srWGS and lrWGS T2T alignments. Each row corresponds to a step in the SNV and indel filtering pipeline used on the patient srWGS and lrWGS callsets as described in Section 5.6.3.3.

5.7.4. Manual SV filtering

The SVs identified in the lrWGS alignment of patient IV:4, were manually filtered using Microsoft Excel and visualizing the SV calls on Integrative Genomics Viewer (IGV) [187]. All the calls located outside the prioritised suggestive linkage regions were excluded. Homozygous calls and calls with low genotype quality and imprecise breakpoints were excluded for this study. The filtered SV calls from patient IV:4, were visualised [187] using two reference datasets: the merged genotyping calls generated by PanGenie [260], and the SV reference developed by sequencing 1019 samples from 1KGP using PacBio lrWGS [247]. SVs in the patient were excluded if an overlap of at least 70% with an SV of the same type in the reference was observed, based on commonly applied reciprocal overlap thresholds [245, 287]. Additionally, deletions in the patient that overlapped larger deletions in the two references were also excluded as described in the American Collage of Medical Genetics guidelines for assessing CNVs [288]. SV calls located in regions of poor coverage were excluded for this study.

5.7.5. Filtering RE calls

The filtering criteria for the output of the EHDN pipeline were provided by Dr. Chiara Folland through personal communication. Briefly, the RE calls that produced a p-value less than 0.05 following the case-control analysis were excluded. The RE calls that localised to the prioritised linkage regions on chromosomes 8 and 16 were identified. These RE calls were ordered by the rank given in the cohort level VCF based on the number of reads that support a call. The REs identified in both patients and ranked in the top 10 in the cohort based on the number of supporting read pairs were prioritised. If a call was not observed in both patients, it was excluded.

5.7.6. Variant selection and prioritisation

The workflow of variant identification, variant filtering, application of the multi-omics data and variant analysis described in this section has been summarised in Figure 5.6

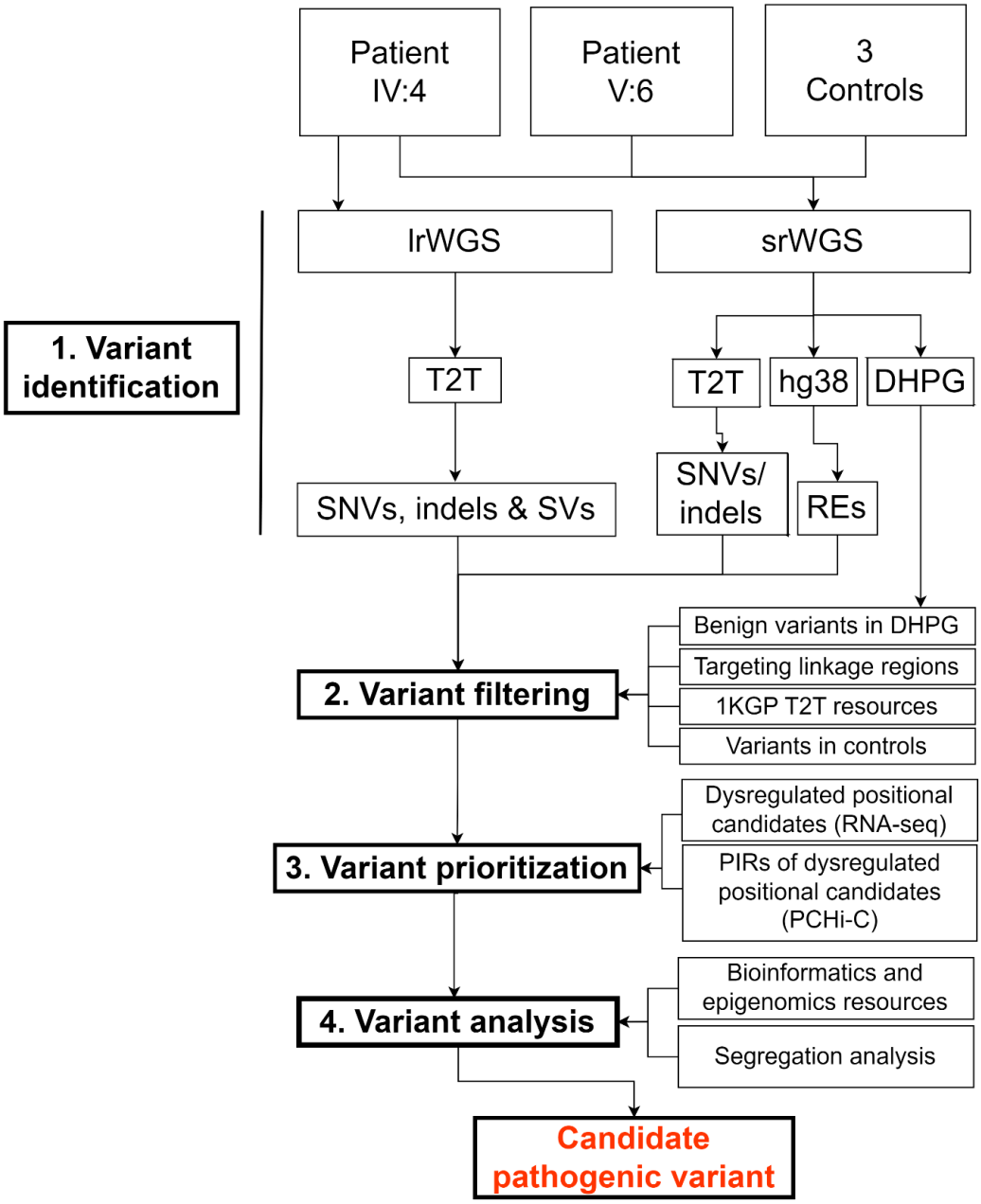


Figure 5.6: The summary of the multi-omics analysis that will be used in the current investigation of CMT720. The workflow diagram showing the tools, datasets and resources utilised at each step of the multi-omics variant analysis strategy. This will be used for identifying the candidate pathogenic variant in CMT720 that will be followed up with functional studies by the future investigations.

5.7.6.1. Using PCHi-C data and transcriptomic findings to prioritise all noncoding variants

PCHi-C data from adult human dorsolateral prefrontal cortex (DLPFC) and hippocampus tissues, aligned against the hg19 [274, 289] was accessed through the 3D Genome Browser [290]. The co-ordinates for each dysregulated positional candidate gene (Chapter 4) ± 50 kb of flanking sequence was queried to ensure the promoter of each gene was included in the specified interval. The co-ordinates of all 3D chromatin interactions anchored in ± 50 kb of each dysregulated positional candidate gene were obtained from the PCHi-C datasets, and lifted from hg19 to hg38 using the USCS LiftOver tool [192]. The lifted coordinates were queried through the Genome Browser (hg38) [192] using the ENCODE cCREs [273] and FANTOM5 [291] tracks to identify the candidate promoter elements and transcription start sites respectively. This was done to determine the PCHi-C bins that captured the promoter regions of each dysregulated positional candidate gene. The 3D chromatin interactions that did not involve the promoter regions of the dysregulated positional candidate genes were de-prioritised to obtain a refined set of functionally relevant PIR-promoter contacts. The PIR and promoter region coordinates identified in PCHi-C data were lifted from hg19 to T2T using the USCS LiftOver tool [192]. Using the PIR-promoter contact information, all patient variants that remained after filtering were then manually visualised on IGV [187]. The variants localizing to either the UTRs, introns or PIRs of the dysregulated positional candidate genes were prioritised for further bioinformatic analysis.

5.7.6.2. Additional criteria for prioritizing SVs and REs

As SVs or REs have not been previously investigated in CMT720, variants localising in exons of protein coding genes, will be prioritised for analysis. SVs can cause gene dysregulation by introducing ectopic CREs [292] or disrupting TAD boundaries [293].

Therefore, SVs within ±3 Mb of each dysregulated candidate gene will be prioritised for analysis, in addition to those localising to the PIRs identified by using PCHi-C data.

5.7.7. Bioinformatic variant analysis

All variants localising to the introns, UTRs and PIRs of dysregulated positional candidate genes, as well as SVs located within ± 3 Mb of each dysregulated positional candidate gene were analysed on UCSC Genome Browser [192] to identify matching reports on ClinVar [294] or population databases. Variants were excluded as candidates if they were reported as benign in ClinVar, or with a minor allele frequency (MAF) greater than 0.0001 in a population database, as this would exceed the expected prevalence of CMT2 [8]. For SVs and REs larger than 50 bp, Database of Genomic Variants (DGV) [287], dbVar [295] and GnomAD [296] were queried as the population databases. Since DGV and dbVar contain curated SVs from healthy individuals [295], any variants matching the patient were excluded. For all remaining variants, dbSNP [257] and GnomAD [296] were the queried population databases. The ENCODE cCREs [270], ENCODE Regulation [270] and ReMap CHiP-seq [297] tracks in Genome Browser were used to determine if the remaining candidate variants overlapped with CREs or transcription factor binding sites (TFBS) in neurons and fibroblasts. If Genome Browser analysis showed variants overlapping a TFBS, Factorbook [298] will be used to determine whether the variant localises to canonical TF binding motifs that are predicted based on the chromatin immunoprecipitation sequencing (CHiP-seq) data from ENCODE project [270]. This information will be used to determine whether a regulatory variant has the potential to disrupt TF binding and cause gene dysregulation and warrant further functional analysis.

5.7.8. Sanger sequencing

Prioritised variants that were determined to be unreported or with a MAF<0.0001 underwent Sanger sequencing in the remaining family members of CMT720 to perform segregation analysis. Reverse and forward primers were designed to amplify approximately ±300bp of a candidate variant as described (Chapter 2, section 5). Annealing temperatures and master mixes were optimised for the designed primers, and all available DNA samples from family CMT720 underwent the standard PCR protocol as previously described (Chapter 2, Section 6). Amplification products underwent gel electrophoresis as previously described (Chapter 2, Section 7) and were diluted by a factor of 20 in UltraPure™ nuclease-free distilled water (Invitrogen). DNA (10-30 ng) from each sample was mixed with 3.2 pmol of the forward and reverse primers in separate wells of a 96-well plate, dried down and sent to Garvan Institute of Medical Research (Australia) which was outsourced for Sanger sequencing.

5.8. Results

In this chapter, genomic variants in CMT720 patients have been identified using both srWGS and lrWGS. Variant filtering guided by the transcriptome profile of CMT720 patients was used for selecting candidate noncoding variants within the suggestive linkage regions chromosomes 8 and 16 and multi-omics was used for prioritising the variants for future functional studies. The quality metrics for the srWGS and lrWGS experiments are provided in Supplementary Section 5.9.1.

5.8.1. Querying PCHi-C data identifies PIRs for the promoters of the dysregulated positional candidate genes localising to the suggestive linkage regions

To anchor the analysis of candidate noncoding variants, biologically relevant regions of the genome predicted to control the genes dysregulated in the transcriptome data of

CMT720 was performed. Querying the 3D Genome Browser [290] for PCHi-C data from the dorsolateral prefrontal cortex (DLPFC) and hippocampus tissues [274] identified 590 3D chromatin interactions ±50 kb on either side of each dysregulated positional candidate gene within the suggestive linkage regions on chromosome 8 and 16 (Table 5.1)(Supplementary Tables 5.9-5.15). Each PCHi-C bin harbouring the promoter region of a dysregulated positional candidate gene was identified using the USCS Genome Browser (Figure 5.7). Using the suggestive linkage regions,189 PIRs contacting the promoter regions of the 4 dysregulated positional candidate genes identified in chromosome 8 (*ZNF704*) and 16 (*BCL7C, IRX6*, *PRRT2*), and were prioritised for further analysis (Table 5.1). The remaining 301 bins that did not interact with the promoter regions of the dysregulated candidate genes were deprioritised for further analysis. Notably, *ZNF704* and *PRRT2* showed substantial differences in the number of PIRs identified between hippocampus and DLPFC, whereas, *IRX6* had only one PIR in DLPFC and no PIR in hippocampus tissue (Table 5.1). The coordinates of the PIRs contacting the respective promoters of the dysregulated positional candidate genes were manually screened and visualised using IGV [187] and variants localizing to these regions were selected for analysis.

Table 5.1: The number of 3D chromatin contacts and PIRs interacting with the promoters of dysregulated positional candidates identified by accessing neuronal PCHi-C data.

Gene	Tissue	3D chromatin contacts	PIRs
ZNF704	DLPFC	4	4
	Hippocampus	147	68
BCL7C	DLPFC	84	19
	Hippocampus	76	12
IRX6	DLPFC	20	1
	Hippocampus	4	0
PRRT2	DLPFC	197	63
	Hippocampus	58	22
Total number of PIRs		590	189

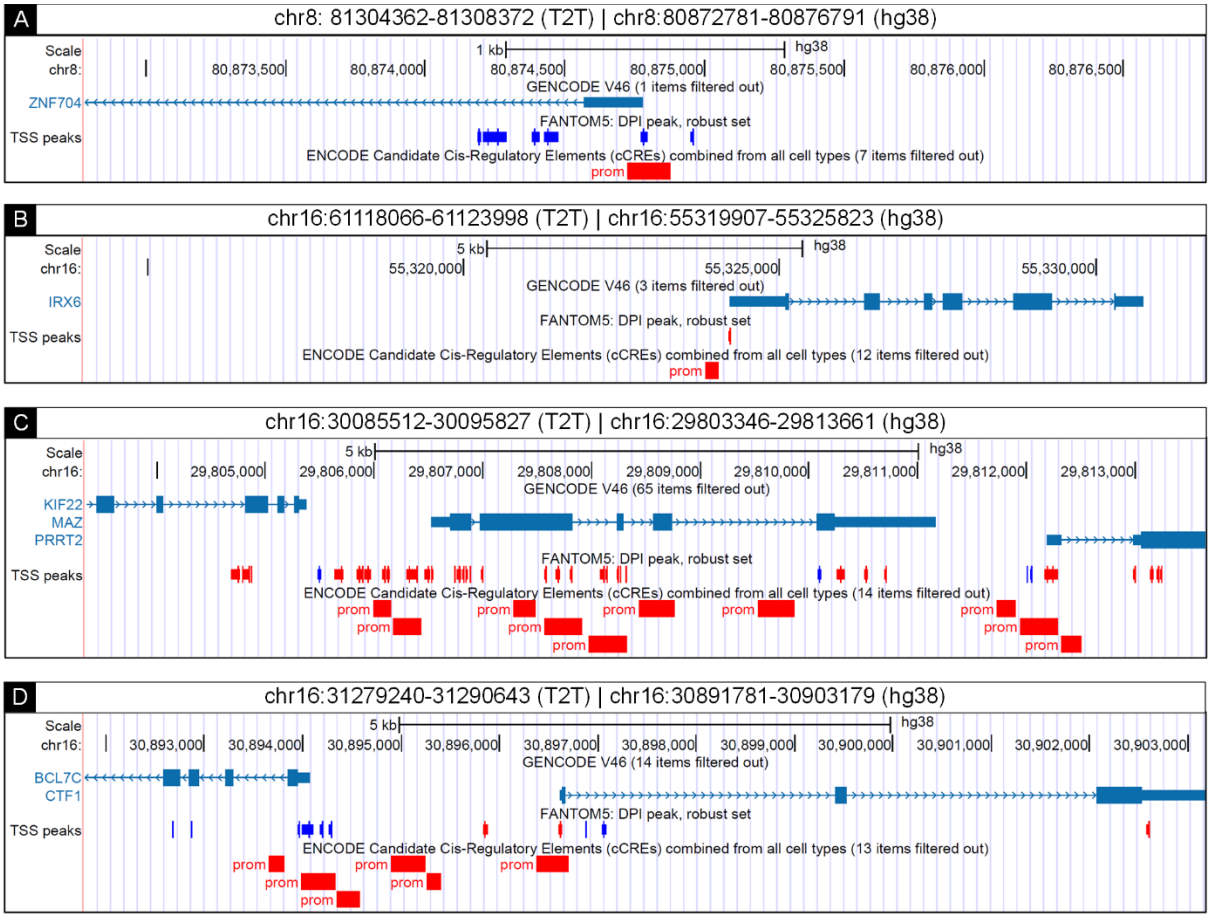


Figure 5.7: The Genome Browser screenshots showing the promoter regions of dysregulated positional candidate genes identified with the PCHi-C data. Each panel shows the coordinates of a bin in PCHi-C data that contain the 5' end of a dysregulated candidate gene overlapping the red promoter boxes. Coordinates of each interval with respect to the T2T and hg38 references are provided in the boxes above the panels. Promoter regions of *ZNF704* (**A**) and *IRX6* (**B**) only harbor the promoters of these dysregulated positional candidates. On the other hand, the promoter region of *PRRT2* also harbors the promoter of *MAZ* (**C**), and the promoter region of *BCL7C* additionally harbors the promoter of *CTF1* (**D**). The candidate promoter elements are represented by the red boxes in the ENCODE cCRE track [270]. The transcription start sites on the positive and negative strands are respectively represented by the red and blue marks on the FANTOM5 track [291]. All genomic features and coordinates are provided with respect to hg38, due to the current unavailability of these Genome Browser tracks for the T2T reference.

5.8.2. No candidate pathogenic RE is identified in CMT720 WGS alignments.

For the patients with srWGS alignments against the hg38 reference, EHDN predicted 2596 REs in patient V:6 and 2592 in patient IV:4 across the whole genome. No REs were identified that were likely to represent significant novel expansions ($p<0.05$) within the prioritised linkage regions on chromosome 8 and 16. Briefly, a total of 9 REs were predicted in the suggestive linkage region on chromosome 8 ($p>0.05$) (Table 5.2) and 16 RE calls on chromosome 16 ($p>0.05$) in patients IV:4 and V:6 (Table 5.3). The REs predicted did not reach the threshold of significance in the case control analysis. Interestingly, for patient V:6, one non-significant call ($p=0.1077$) for a 1716 bp RE within the suggestive linkage region on chromosome 16 was annotated as an expansion within an exon of the dysregulated positional candidate gene *IRX6* (Table 5.3). However, visualizing the patient srWGS and lrWGS alignments showed this call was a false positive, since no abnormality in patient alignments or an insertion call supporting the presence of a RE was observed in this region (Figure 5.8). Since the remaining REs within the suggestive linkage regions on chromosome 8 and 16 did not localise to the UTRs, introns, PIRs or exons of the dysregulated candidate genes, these variants were deprioritised as candidate variants for this study.

Table 5.2: The RE calls made by EHDN localising to the suggestive linkage region on chromosome 8.

Sample ID	CHR	Start	End	Size	Gene	Region	Repeat motif size	p-value	Sample rank
IV:4	8	75104358	75105834	1476	*CRISPLD1*(dist=69800) *CASC9*(dist=117283)	intergenic	5	0.50915	77
V:6	8	75104358	75105834	1476	*CRISPLD1*(dist=69800) *CASC9*(dist=117283)	intergenic	5	0.65173	93
V:6	8	69451532	69452851	1319	*LINC01603*(dist=3288) *SULF1*(dist=13773)	intergenic	5	0.05671	7
IV:4	8	83087942	83089359	1417	*LOC101927141*(dist=126018) *LINC01419*(dist=314399)	intergenic	14	0.05938	8
IV:4	8	72210956	72212174	1218	*LOC392232*	ncRNA_exonic	3	0.09166	18
V:6	8	72210956	72212174	1218	*LOC392232*	ncRNA_exonic	3	0.52744	79
IV:4	8	82023021	82024276	1255	*SNX16*(dist=180735) *LOC101927141*(dist=887828)	intergenic	4	0.33146	57
IV:4	8	81841837	81842948	1111	*SNX16*	UTR5	2	0.33146	57
V:6	8	81841837	81842948	1111	*SNX16*	UTR5	2	0.33146	57

Table 5.3: The RE calls made by EHDN localising to the suggestive linkage region on chromosome 16.

Sample ID	CHR	Start	End	Size	Gene	Region	Repeat motif size	pvalue	Sample rank
IV:4	16	27138118	27139943	1825	*C16orf82*(dist=68952) *KDM8*(dist=63583)	intergenic	3	0.85954	123
V:6	16	27138362	27139943	1581	*C16orf82*(dist=69196) *KDM8*(dist=63583)	intergenic	3	0.39154	64
V:6	16	52750467	52751564	1097	*CASC16*(dist=143492) *LOC105371267*(dist=284126)	intergenic	5	0.16764	34
V:6	16	55830168	55831218	1050	*CES1*	intronic	4	0.13048	27
IV:4	16	34096041	34098052	2011	*ENPP7P13*(dist=311766) *MIR9901*(dist=65526)	intergenic	5	0.90056	132
V:6	16	34096041	34098052	2011	*ENPP7P13*(dist=311766) *MIR9901*(dist=65526)	intergenic	5	0.08794	17
IV:4	16	22980112	22981168	1056	*HS3ST2*(dist=63774) *USP31*(dist=80239)	intergenic	4	0.49085	75
V:6	16	22980112	22981168	1056	*HS3ST2*(dist=63774) *USP31*(dist=80239)	intergenic	4	0.4817	74
IV:4	16	25872919	25874189	1270	*HS3ST4*	intronic	4	0.85436	122
V:6	16	25872919	25874189	1270	*HS3ST4*	intronic	4	0.85436	122

IV:4	16	26809172	26810631	1459	*HS3ST4*(dist=671487) *C16orf82*(dist=256296)	intergenic	4	0.73968	104
V:6	16	26809172	26810631	1459	*HS3ST4*(dist=671487) *C16orf82*(dist=256296)	intergenic	4	0.93786	143
IV:4	16	26809328	26810802	1474	*HS3ST4*(dist=671643) *C16orf82*(dist=256125)	intergenic	20	0.18547	37
V:6	16	55324747	55326463	1716	*IRX6*	exonic	4	0.1077	22
V:6	16	54439552	54440878	1326	*LINC02140*(dist=68853) *LOC101927480*(dist=406373)	intergenic	3	0.34827	59
V:6	16	24832877	24834357	1480	*TNRC6A*(dist=6661) *SLC5A11*(dist=11506)	intergenic	4	0.37404	62

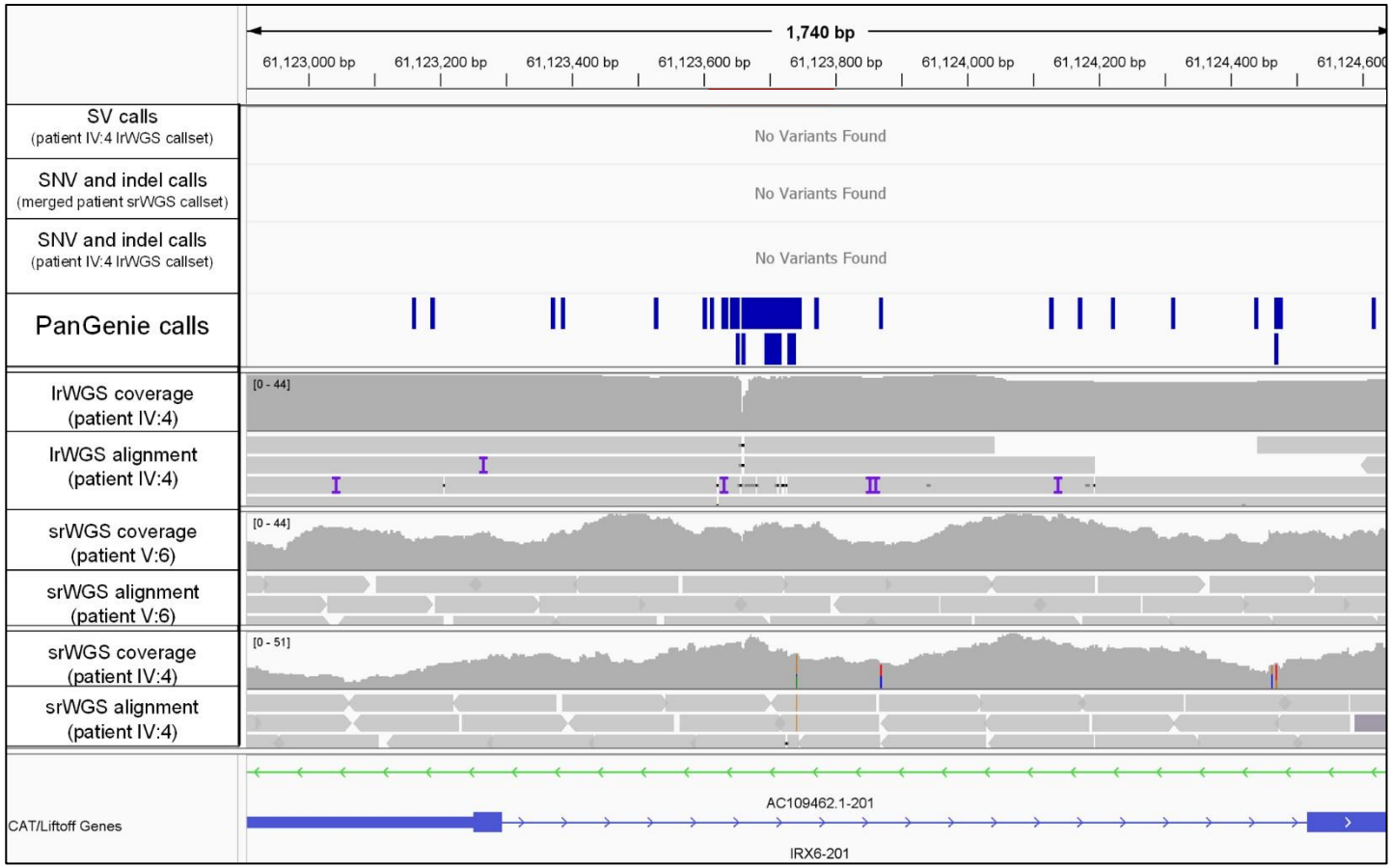


Figure 5.8: IGV display of the RE call in *IRX6* from patient V:6 visualised in the T2T alignments. The coordinates of the RE call were lifted to T2T using the UCSC LiftOver tool [192] and were visualised on IGV. Indel or SV calls were not observed in any patient alignment. No insertions (<50pb or >50bp) were present in this region and all indel calls originating from this intronic region were eliminated during variant filtering, suggesting that these indels are common or not shared between the affected individuals.

5.8.3. No candidate pathogenic SVs are identified in CMT720 patient WGS aligned to the T2T reference.

A total of 26,721 SVs were identified for patient IV:4 lrWGS aligned to the T2T reference (Figure 5.9). Targeting the linkage regions provided powerful filtering efficiency by eliminating 26,283 of these SV calls (98.4%) (Figure 5.9). Following the removal of variants with low genotype quality (Figure 5.9), variant filtering was then performed by manual visual analysis using the IGV viewing tool [187]. SV calls in patient IV:4 were further reduced for analysis based on removing SV calls present in the 1KGP [247] and PanGenie callsets [260]. In particular, SVs identified by PanGenie genotyping eliminated an additional 14 variants that were not represented by the 1KGP SV reference. This resulted in the identification of 17 candidate noncoding SVs in the suggestive linkage regions on chromosome 8, and 30 candidate noncoding SVs in the suggestive linkage region chromosome 16. An example demonstrating how of the overlap between the SVs identified in patients and healthy individuals can be assessed by manual variant filtering is shown (Figure 5.10).

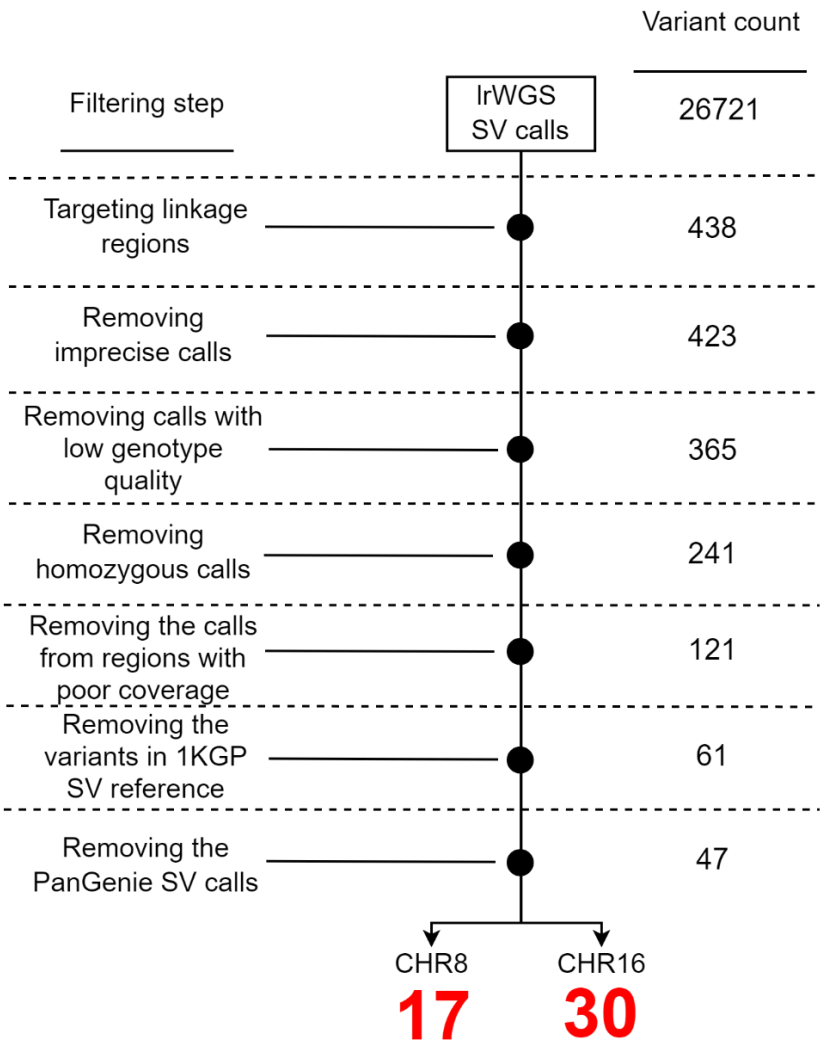


Figure 5.9: Flowchart diagram showing the results of variant filtering performed on the SVs identified in CMT720 patient IV:4. The counts of SVs that remain in the lrWGS callset after each filtering step are indicated in the respective rows. The number of SV calls that remain within the suggestive linkage regions on chromosome 8 (n=17) and 16 (n=30) following SV filtering are highlighted in red at the bottom.

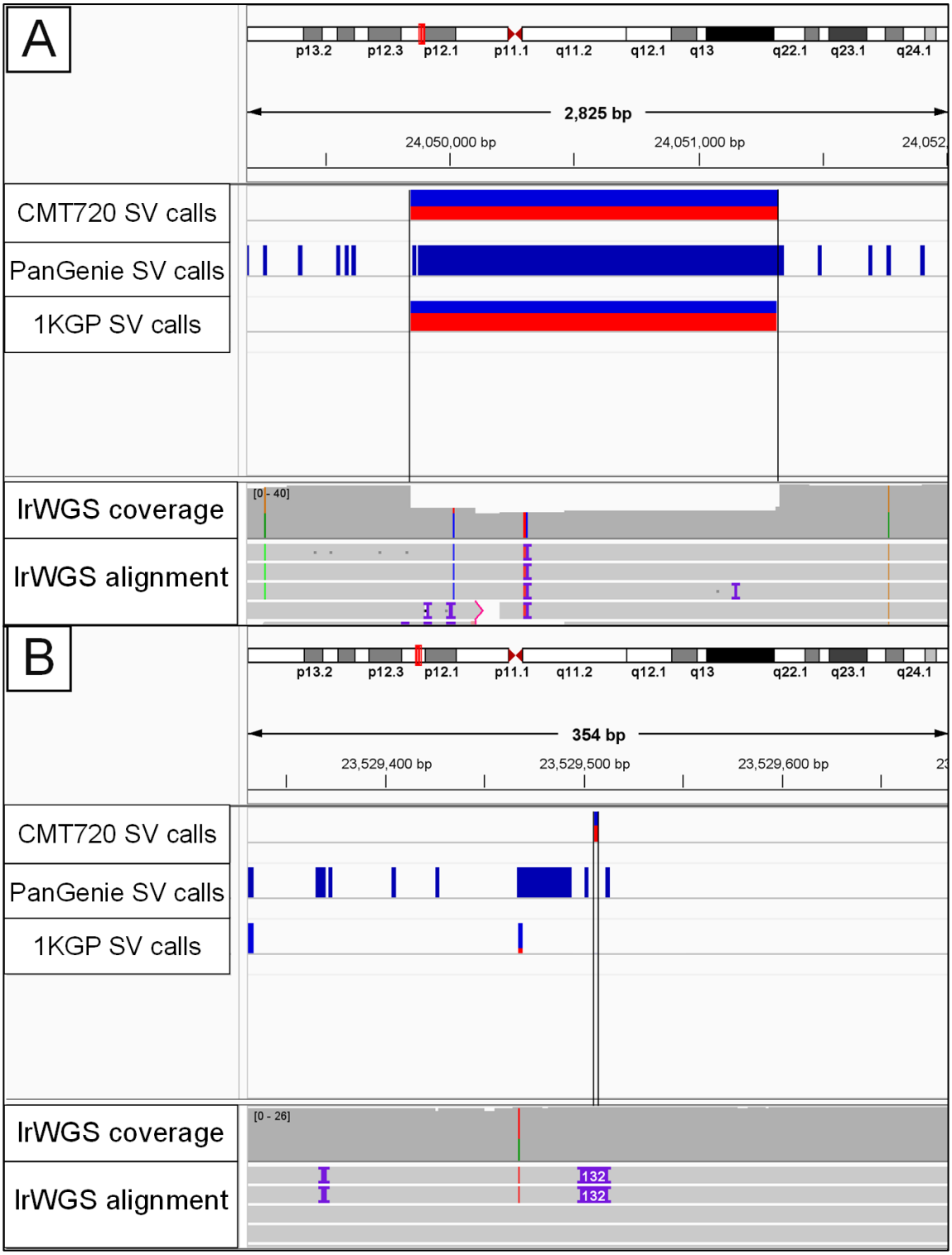


Figure 5.10: Manual SV filtering by visualizing patient, PanGenie and 1KGP calls in IGV. (**A**) A deletion called in patient IV:4 which was filtered due to near complete overlap with a benign deletion genotyped by PanGenie and complete overlap with a benign deletion identified in the T2T alignments of 1KGP lrWGS data [247]. (**B**) An insertion called in patient IV:4 which was retained as no overlap with an SV of the same size and type in the reference callsets was present.

SVs overlapping the exons of protein coding positional candidate genes were not detected during manual filtering of the 365 SV calls from the suggestive linkage regions on chromosome 8 and 16 (Figure 5.9), strongly suggesting that the pathogenic mutation in CMT720 is not an SV disrupting coding sequences. Furthermore, no translocation event was identified among the SVs localising to the suggestive linkage regions on chromosome 8 and 16. This suggests the interchromosomal translocation calls predicted by FusionCatcher for the RNA-seq analysis in Chapter 4 most likely represent false positives and CMT720 is unlikely to be caused by a fusion transcript.

SV calls localising to regions of poor coverage on chromosome 16 (n=120) were removed since these calls were likely to include high numbers of false positives (Figure 5.9)[299]. The regions of poor coverage in the suggestive linkage region on chromosome 16 included centromeric and pericentromeric repeat arrays on both the p and q arms [142, 300] (Supplementary Figure 5.1). No SV calls from the suggestive linkage region on chromosome 8 localised to regions of poor coverage.

No SVs were identified within the UTRs, introns or PIRs of the dysregulated positional candidates in the suggestive linkage regions on chromosome 8 and 16. In total, 6 SVs were identified within ∓3 Mb of the dysregulated positional candidates *BCL7C*, *IRX6* and *PRRT2* on chromosome 16 (Table 5.4). The coordinates of these SVs were lifted to hg38 using the UCSC LiftOver tool [192], to determine whether these SVs were previously reported in the hg38 reference. The SVs localising ∓3 Mb on either side of the dysregulated positional candidates were either reported with MAF>0.0001 or were identified in healthy individuals indicated by matching reports in the DGV database [287] (Table 5.4), and were excluded as candidate variants.

Table 5.4: SV calls prioritised based on proximity to dysregulated candidate genes and subsequently excluded for pathogenic role in CMT720.

Associated positional candidate (Prioritisation criteria)	Coordinates (T2T)	SV type	Size (bp)	Overlapping genes	Coordinates (hg38)	Accession number	AC (AF)	Database
PRRT2 (-3 Mb)	chr16:28084164-28084325	Deletion	161	*GSG1L* (Intron 6/6)	chr16:27805269-27805429	nsv1131111	2/2 (N/A)	DGV
						nssv3763955	1/1 (N/A)	DGV
PRRT2, BCL7C (-3 Mb)	chr16:29580865-29581064	Deletion	199	Intergenic	chr16:29298662-29298862	gssvL43294	2/97 (0.02061)	DGV
PRRT2, BCL7C (-3 Mb)	chr16:29580987-29581067	Deletion	80	Intergenic	chr16:29298662-29298862	gssvL43294	2/97 (0.02061)	DGV
PRRT2, BCL7C (-3 Mb)	chr16:29995734-29995734	Insertion	137	Intergenic	chr16:29713469-29713469	rs1967789126	3/11556 (0.00026)	GnomAD
IRX6 (-3 Mb)	chr16:58601665-58601665	Insertion	55	Intergenic	chr16:52803805-52803848	esv2714494	6/96 (0.06250)	DGV
IRX6 (-3 Mb)	chr16:60104736-60104736	Insertion	85	Intergenic	chr16:54306747-54306747	rs1180686555	13/110348 (0.00012)	GnomAD

***AC=allele count, AF= allele frequency**

5.8.4. SNVs and indels

5.8.4.1. Variant filtering eliminates >99.7% of SNV and indel calls identified in CMT720 patient srWGS and lrWGS

Alignment of patient srWGS to the T2T reference, identified a total of number of 10,620,340 calls for patient V:6 and 7,601,181 calls for patient IV:4 (Figure 5.11). Variants remaining after the removal of QUAL scores <20 were 3,817,693 (patient V:6) and 4,112,195 (patient IV:4) respectively (Figure 5.11). By aggregating information from the patient and control srWGS variant calls, a total of 629,769 variants were common to the affected individuals. Targeting variants within the suggestive linkage regions provided the most effective filtering efficiency (98.3%) yielding 10,582 prioritised candidate variants (Figure 5.11). Within the linkage regions on chromosome 8 and 16 the collective power of using the 1KGP and PanGenie variant calls to filter variants resulted in a total of 691 calls remaining (Figure 5.11). As the disease inheritance was assumed to be autosomal dominant the removal of homozygous calls resulted in 114 and 318 respectively in the chromosome 8 and 16 suggestive linkage regions for further analysis.

The lrWGS alignment of patient IV:4 to the T2T reference identified 5,184,001 SNVs and indels (Figure 5.11). Aggregating the srWGS calls from controls with the lrWGS calls of patient V:4 identified 2,803,524 unique to the affected patient. Targeting the prioritised linkage regions provided the highest filtering efficiency (97.9%) by removing 2,746,270 SNVs and indels to give a final variant count of 57,254 (Figure 5.11). In total, 19,193 SNVs and indels identified in the lrWGS alignment of patient IV:4 localised within the suggestive linkage regions on chromosome 8 and 16 following filtering against the calls from 1KGP, PanGenie and removal of homozygous calls (Figure 5.11). As SNV and indel calls with high call quality are highly concordant between the PacBio HiFi lrWGS and Illumina srWGS alignments [243, 253],

for this study the SNVs and indels that were identified in both patient srWGS and lrWGS alignments were prioritised for this study. All remaining deprioritised variants will be addressed in future studies if needed.

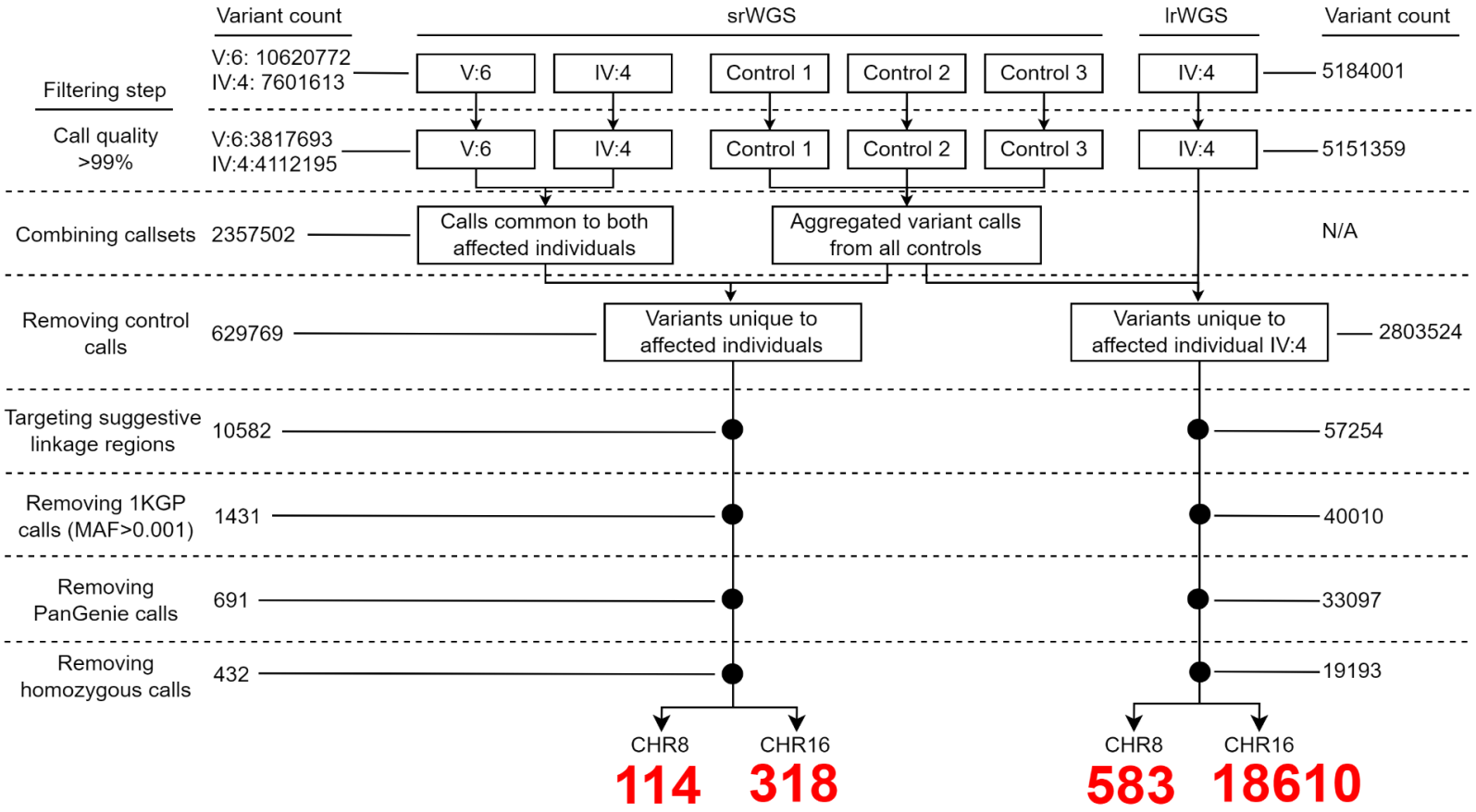


Figure 5.11: Variant filtering strategy to prioritise SNVs and indels identified from patient srWGS and lrWGS aligned to the T2T reference. The flowchart shows the number of variant counts after each step of the SNV and indel filtering pipeline. The variant counts on the left-hand side indicate the aggregated number of SNVs and indels that remain in the srWGS callsets, while the counts on the right-hand side indicate the number of SNVs and indels remaining in the lrWGS callset. The aggregated counts of SNVs and indels within each suggestive linkage region remaining after variant filtering in patient srWGS and lrWGS callsets are shown in red.

5.8.4.2. SNVs and indels localising to GDAP1 and the genes unique to T2T were deprioritised by bioinformatic analysis.

GDAP1 is a known CMT2 gene located within the suggestive linkage region on chromosome 8. Although no evidence for gene dysregulation or abnormal splicing in *GDAP1* was obtained during the transcriptomic analysis on patient fibroblasts, variants localising to *GDAP1* were prioritised and analysed as a conservative measure against missing etiologically important transcriptomic abnormalities by using non-neuronal tissue. A 14 bp duplication (GDAP1:c.695-33907_695-33894dup), 2 bp deletion (GDAP1:c.694+18724_694+18725del) and a single nucleotide substitution (GDAP1: c.165+46327G>C) were identified in the introns 5 and 7 of *GDAP1* (Table 5.5). All 3 variants were previously reported in population databases with MAF>0.0001 (Table 5.5). Although, the *GDAP1*: c.165+46327G>C was reported in 17 healthy individuals included in the GnomAD v4.1.0 database [301], the MAF of this variant (0.00017) is close to the reported prevalence of CMT2 (1/10000) [9] (Table 5.5). We have accessed the srWGS data of CMT720 patients uploaded to Seqr [169] to determine whether this variant was identified in the local cohort of 7598 clinically evaluated individuals, which resulted in identification of only a healthy adult carrying this intronic mutation, indicating an AF=0.00013 and an AC=7598 in the local Seqr cohort. The potential impact of this intronic mutation on splicing was analysed using SpliceAI [302] as a conservative measure. SpliceAI produced a score of 0.00 for *GDAP1*: c.165+46327G>C, indicating that this variant is not predicted to impact splicing. Accordingly, all variants identified in *GDAP1* were deprioritised as unlikely candidates and excluded from variant analysis in this study based on the absence of transcriptomic abnormalities in *GDAP1*, high frequency of the variants in the healthy population and no predicted impact on splicing.

Table 5.5: SNVs and indels identified in *GDAP1* and the genes unique to the T2T reference located in the suggestive linkage regions

Associated positional candidate	Position	Prioritisation criteria	Location	REF	ALT	rsID	Max. AC (Max. AF)	Database
GDAP1	chr8:74811077-74811078	Within known CMT2 gene	Intron 5/7	AAA	A	rs59246526	3032/142094 (0.021)	GnomAD
GDAP1	chr8:74826964	Within known CMT2 gene	Intron 5/7	G	C	rs1389424159	2/11862 (0.00017)	dbSNP
GDAP1	chr8:74884038	Within known CMT2 gene	Intron 7/7	C	CCAAATAT AATGATA	rs141075428	27718/64412 (0.43)	GnomAD
DUSP22	chr16:34992660	Within a gene unique to T2T	Intron 3/7	G	A	N/A	N/A	N/A
DUSP22	chr16:34994749-34994750	Within a gene unique to T2T	Intron 3/7	GC	G	N/A	N/A	N/A
DUSP22	chr16:35007871-35007873	Within a gene unique to T2T	Intron 2/7	AGT	A	N/A	N/A	N/A

***Max. AC: Maximum allele count, Max. AF: Maximum allele frequency**

Since the novel T2T reference was used in the current investigation of CMT720, the possibility of identifying variants localising to the genes unique to the T2T alignment was assessed. The list of novel genes called with the T2T alignment was queried using the supplementary material provided in the original report of the T2T reference [142]. The novel genes mapping to the suggestive linkage regions on chromosome 8 and 16 were assessed for candidate pathogenic variants. No protein coding gene unique to T2T was identified in the chromosome 8 suggestive linkage region. In the suggestive linkage region on chromosome 16, *BOLA2B, SLX1B* and *DUSP22* were identified as paralogous genes unique to T2T. The *BOLA2B* and *SLX1B genes* did not harbour any variants, while *DUSP22* harboured one intronic base substitution and 2 intronic deletions (Table 5.5). The *DUSP22* paralog identified in the suggestive linkage region on chromosome 16 localised to a segmental duplication corresponding to a gap in the hg38 reference, and there are currently no variant reports available from this region. Therefore, the population frequency for the 3 intronic variants identified in this gene could not be interrogated, and they remain variants of unknown significance. Notably, the patient srWGS alignments showed poor coverage at this region and there were no variant calls in either short read alignment (Figure 5.12). A review of the literature indicated this gene is a known paralog of *DUSP22* that is transcriptionally inactive [303]. Accordingly, the intronic variants in this *DUSP22* paralog were deprioritised in the current study.

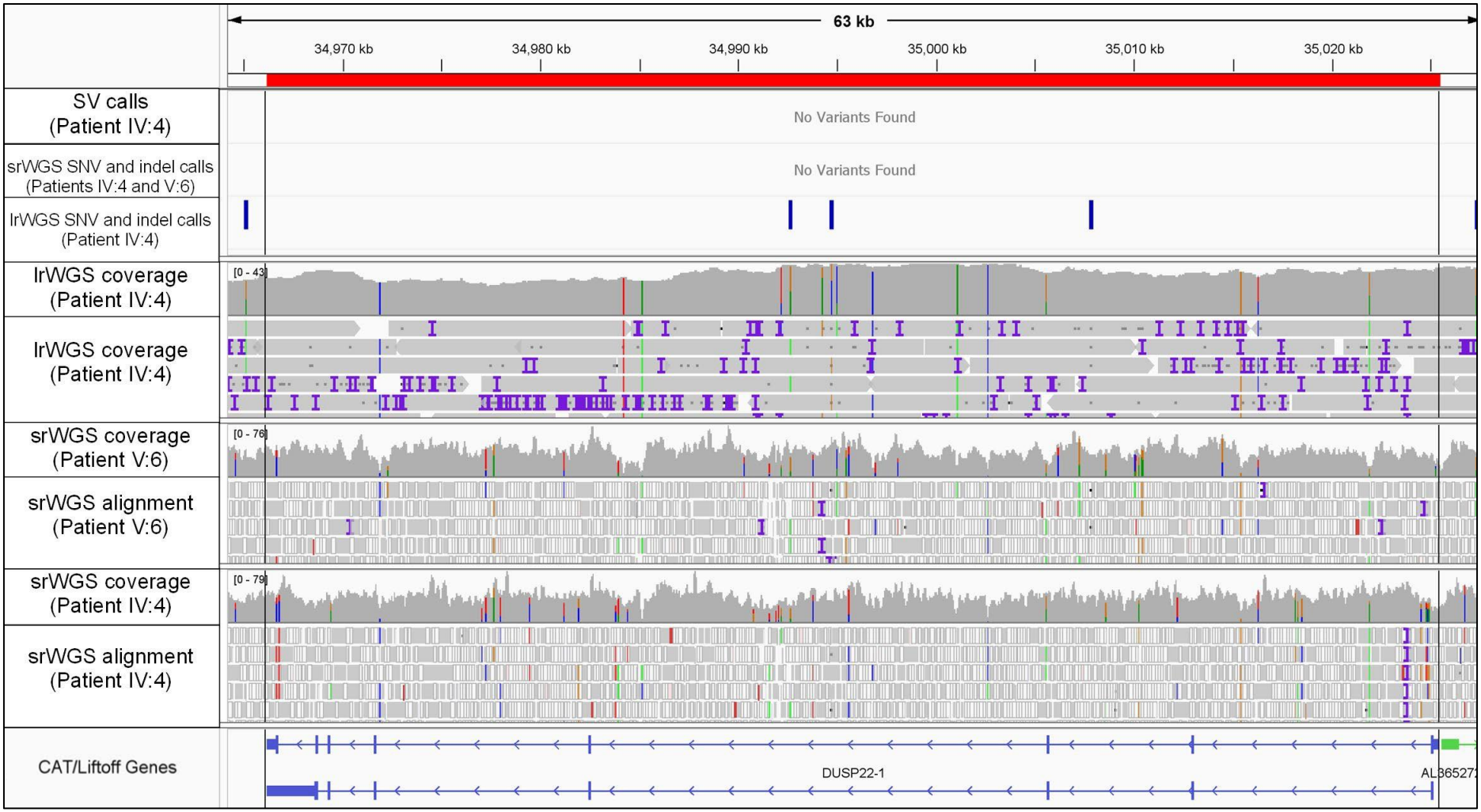


Figure 5.12: IGV panel showing the intronic variants identified in the *DUSP22* paralog unique to T2T on chromosome 16 in CMT720 patient alignments. The 3 intronic variants are only present in the lrWGS alignment of patient IV:4. The srWGS alignment of patients IV:4 and V:6 show fluctuations in

coverage and reads with MQ<0 as indicated by white boxes in the respective alignment tracks. The red bar and the vertical guidelines delineate the span of *DUSP22*.

5.8.4.3. PCHi-C data prioritises 2 very rare SNVs as candidate pathogenic variants

Screening the UTRs, introns and PIRs of the dysregulated positional candidate genes identified 3 noncoding base substitutions localising to the prioritised linkage region on chromosome 16 (Table 5.6). These candidate variants underwent further bioinformatic analysis to determine allele frequency in the population databases, based on the hg38 reference. The BCL7C:c.529-24798C>T candidate variant localised to intron 5 and had a MAF of 0.000057 in the GnomAD v4.1.0 database [301](Table 5.6), however, this variant was reported in the homozygous state in an individual, and had MAF above the threshold of 0.0001 in South Asian (MAF= 0.00021), Middle Eastern (MAF= 0.0068) and Admixed American (MAF= 0.00020) populations. These findings suggest that this variant is unlikely to be the pathogenic mutation, therefore, BCL7C:c.529-24798C>T was deprioritised, and excluded from further analysis in the current investigation. The remaining prioritised candidate SNVs g.31037657G>A and g.31287661G>C localised to the PIRs of the *PRRT2* and *BCL7C* genes respectively as determined by the PCHi-C data from DLPFC (Table 5.6). Both variants had a MAF<0.0001 in all databases queried during analysis across all populations (Table 5.6) and were only found in CMT720 patients IV:4 and V:6 in the local Seqr cohort. The g.31287661G>C candidate noncoding variant was reported in 3/68046 non-Finnish European individuals in GnomAD v4.1.0 with no homozygous carriers of the variant being reported. The g.31037657G>A candidate noncoding variant was absent in GnomAD v4.1.0, and reported in 2/264690 individuals in the TOPMED database [305] (Table 5.6) with no homozygous carriers. These candidate variants were therefore selected for further bioinformatic analysis to predict functional impacts by determining if these noncoding SNVs localise within CREs or TFBS predicted to regulate either the *PRRT2* or *BCL7C* genes.

Table 5.6: Prioritised SNVs identified in CMT720 patients

Associated positional candidate	Position	Prioritisation criteria	Location	REF	ALT	rsID	Max. AC (Max. AF)	Database
PRRT2	chr16:31037657	PIR:31034877-31037876	Intergenic	G	A	rs1364535320	2/264690 (0.0000075)	TOPMED
BCL7C	chr16:31287661	PIR:31279240-31290643	Intergenic	G	C	rs1031290770	3/152214 (0.000020)	GnomAD
BCL7C	chr16:31247435	Intronic	Intron 5	C	T	rs960149910	8/140282 (0.000057)	GnomAD

***Max. AC: Maximum allele count, Max. AF: Maximum allele frequency**

5.8.5. Bioinformatic analysis of the prioritised candidate SNVs g.31287661G>C and g.31037657G>A

5.8.5.1. The g.31287661G>C does not localise to CREs or TFBS, and unlikely to be a functionally impactful noncoding variant

To determine whether the rare candidate SNVs g.31287661G>C and g.31037657G>A can potentially disrupt regulatory elements, the overlap of these variants with CREs and TFBS was assessed by querying the CRE annotations and CHiP-seq data available through the UCSC Genome Browser (Supplementary Table 5.16) [192]. Since these features are currently not annotated in the T2T reference, variant coordinates were lifted to hg38 and the bioinformatic analysis was conducted based on the CRE and TFBS annotations in hg38. For clarity, we will refer to the prioritised candidate SNVs by prefixing the Human Genome Variation Society identifier [306] of each variant with the appropriate reference genome from this point onwards. G.31287661G>C will be referred to as T2Tg.31287661G>C against the T2T reference and hg38g.30900198G>C against the hg38 reference, whereas g.31037657G>A will be represented by T2Tg.31037657G>A and hg38g.30650690G>A, respectively.

The hg38g.30900198G>C localises to the intron 2 of the *CTF1* gene, and is 6.1 kb upstream to the 5' end of *BCL7C* on the negative strand (Figure 5.13). The hg38g.30900198G>C does not overlap any cCREs displayed in the ENCODE cCREs [270] track or TF CHiP-seq peaks in the ENCODE Regulation [270] and ReMap CHiP-seq tracks [297] (Figure 5.13). These findings suggest the hg38g.30900198G>C is unlikely to cause the downregulation observed for the *BCL7C* gene, and was subsequently deprioritised as a candidate noncoding pathogenic variant for this study.

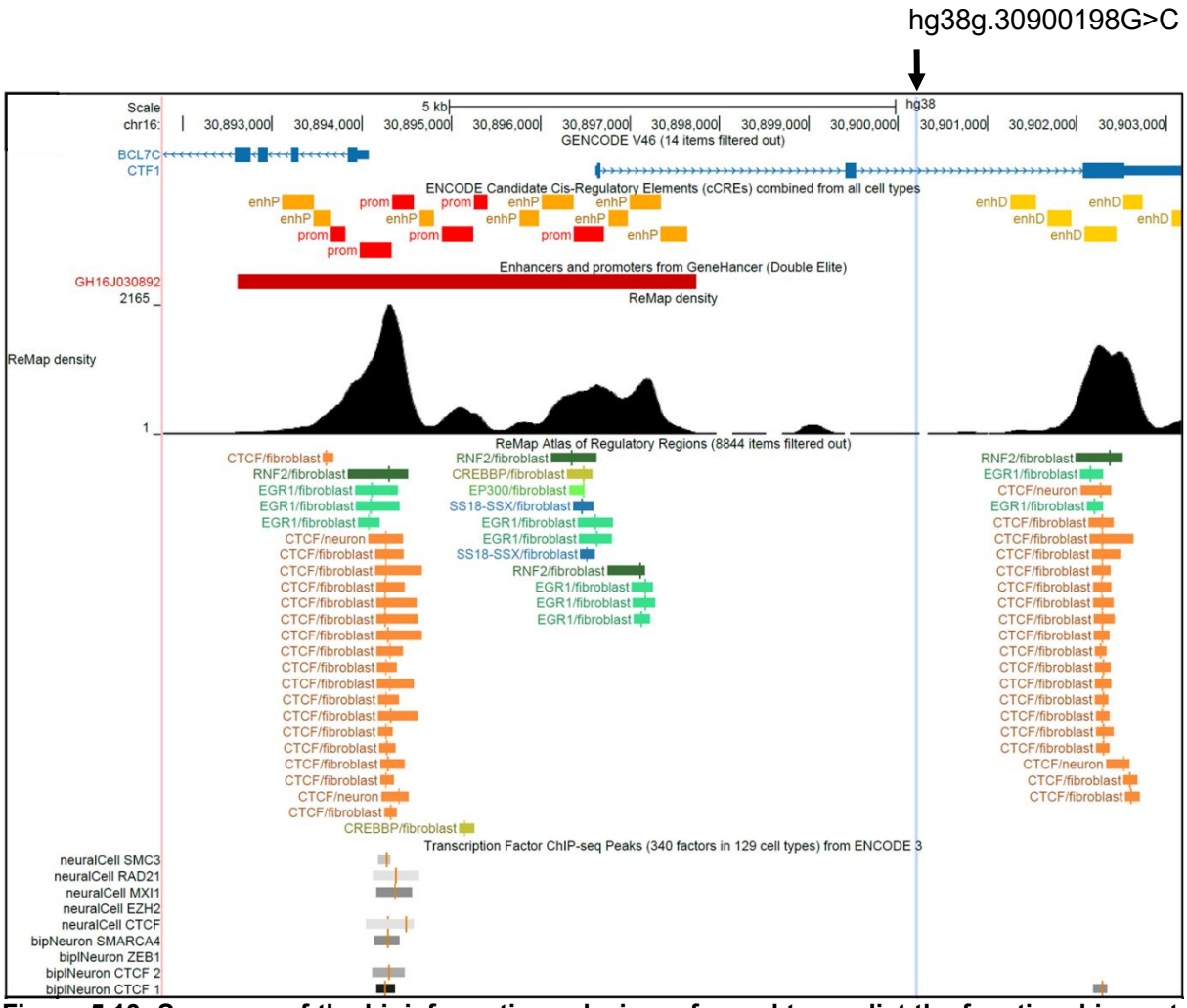


Figure 5.13: Summary of the bioinformatic analysis performed to predict the functional impact of hg38g.30900198G>C. The SNV hg38g.30900198G>C localises to the intron 2 of *CTF1* and does not overlap any of the elements displayed in the ENCODE [270] ReMap Atlas, Transcription Factor CHIP-seq [299], or the candidate Cis regulatory element (cCRE) tracks [270]. Blue vertical line indicates the position of hg38g.30900198G>C.

5.8.5.2. Hg38g.30650690G>A is a TFBS variant with high potential for pathogenicity

Variant hg38g.30650690G>A is located 163 bp upstream to the 5' end of the *PRR14* gene (Figure 5.14B) in a PIR of *PRRT2* (Figure 5.14A). ENCODE cCREs track [270] predicted that hg38g.30650690G>A localises to a candidate proximal enhancer with the ENCODE accession number EH38E1811591 (Figure 5.14B) which could potentially interact with the *PRRT2* promoter region (Figure 5.14A). This variant may therefore potentially disrupt

a CRE that may regulate the expression of *PRRT2*. In contrast, no variants were identified in proximal enhancer EH38E1811590 located in the same PIR as hg38g.30650690G>A (Figure 5.14B).

Accessing the TF CHiP-seq data on Genome Browser showed that hg38g.30650690G>A overlaps 2050 TF CHiP-seq peaks identified across all cell types analysed by the ReMap project (Figure 5.14B) [297], indicating that this region is a highly active TF binding site. Analysing tissue-specific TF CHiP-seq data revealed that the hg38g.30650690G>A noncoding variant intersects with 3 TF CHiP-seq peaks identified in neurons and 17 CHiP-seq peaks in fibroblasts (Figure 5.14B). Among the intersected CHiP-seq peaks, the candidate noncoding variant was closest to the peak summits of CTCF in both tissues with a perfect overlap in fibroblasts and a distance of 30bp in neurons (Figure 5.14B, Supplementary Table 5.17). This suggests the hg38g.30650690G>A variant most likely localises to a CTCF binding site [307, 308]. The nucleotide altered at hg38g.30650690G>A is conserved with a phyloP score of ~1.31 [309], and surrounded by a small cluster of highly conserved nucleotides with phyloP scores ranging from 2.04 to 3.25 (Figure 5.14C). PhyloP scores of >1 in regulatory regions indicate highly conserved TF binding sites [310]. The robust CHiP-seq signal and the degree of conservation observed at this region strongly suggest that hg38g.30650690G>A is located in a TFBS, and may potentially disrupt a canonical TF binding motif [311, 312].

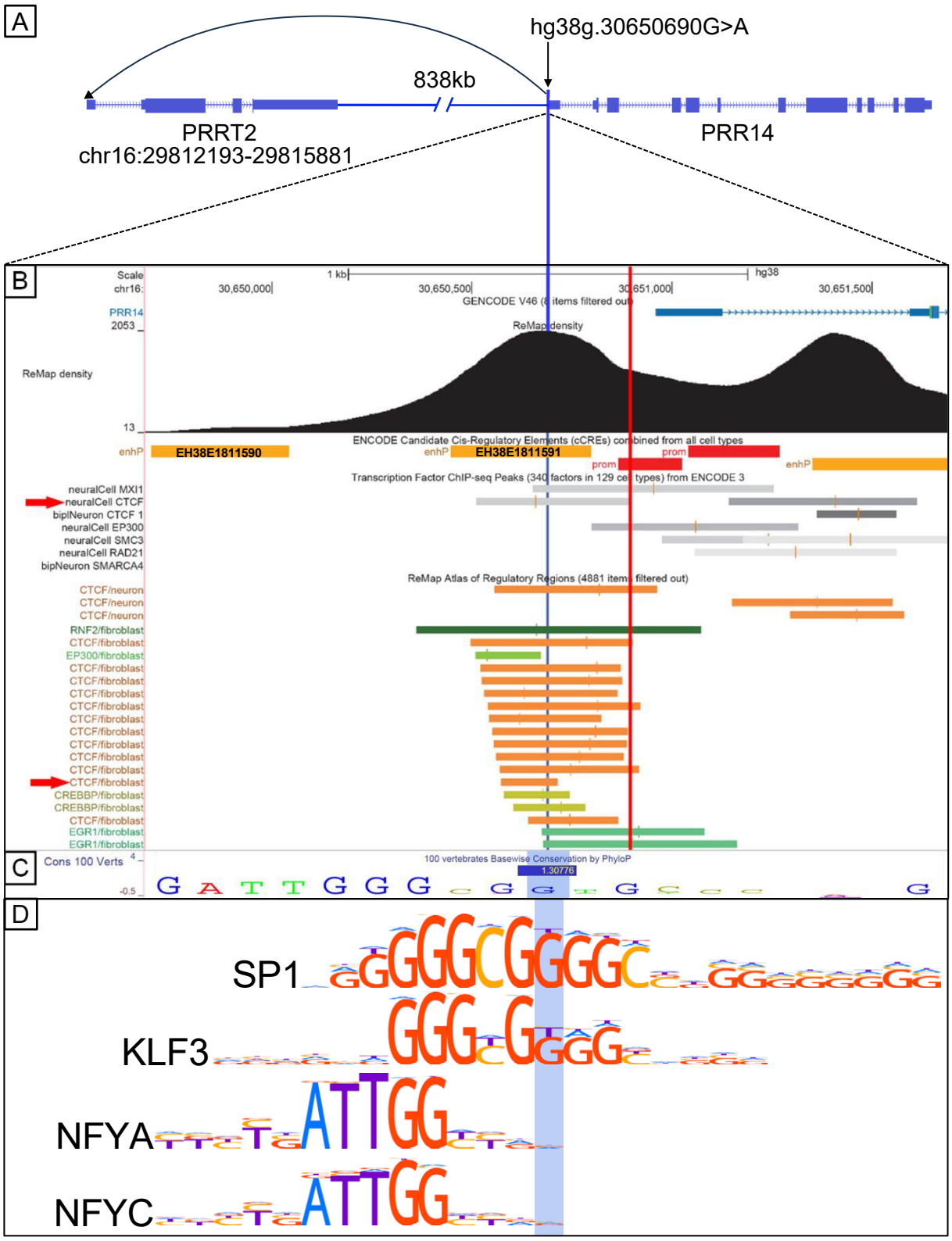


Figure 5.14: Summary of the bioinformatic analysis performed to predict the functional impact of hg38g.30650690G>A. (**A**) The SNV hg38g.30650690G>A is located 270 bp upstream to the 5' end of *PRR14* within the PIR that forms a 3D chromatin interaction with the promoter region of *PRRT2* spanning 838 kb in the hg38 reference. (**B**) The Genome Browser screenshot showing the region ±1kb on either side of hg38g.30650690G>A. The candidate variant intersects a predicted proximal enhancer

element indicated by a yellow box in the ENCODE cCREs [270] track and 2050 TF CHiP-seq peaks as shown in the density plot of ReMap CHiP-seq track [297]. Tissue specific TF CHiP-seq peaks in neurons and fibroblasts are indicated by the bars in the ENCODE Regulation [270] and ReMap CHiP-seq tracks, where the CHiP-seq summits are denoted by the vertical lines within each bar. Red arrows point towards the CHiP-seq summits closest to hg38g.30650690G>A in neurons and fibroblasts to highlight the most likely TFBS the candidate variant may overlap. The 3' end of the PIR harboring hg38g.30650690G>A is indicated by the red vertical line at chr16:30650909. (**C**) The sequence logo produced by the Conservation track, where the height of each letter indicates the corresponding phyloP score and the degree of conservation for a nucleotide among 100 vertebrate species [309]**.** The position of hg38g.30650690G>A is indicated by the blue highlight (**D**) Factorbook predicted that hg38g.30650690G>A overlaps the canonical binding motifs of SP1, KLF3, NFYA and NFYC at the positions indicated by the blue highlight. The height of each letter represents the frequency of a nucleotide observed at the binding sites of indicated TF across the human genome. The logos for canonical TF binding motifs were obtained from HOCOMOCO database [313].

To determine whether hg38g.30650690G>A is located in a canonical TF binding motif, the variant annotator function of Factorbook was used [298]. This tool predicted that hg38g.30650690G>A noncoding variant is likely (FDR>0.05) to overlap the canonical binding motifs of the transcription factors KLF3, NFYA, NFYC and SP1 (Supplementary Table 5.18). These predictions indicate that hg38g.30650690G>A is most likely to disrupt the core binding motifs of SP1 and KLF3 since the variant alters highly overrepresented nucleotides in the predicted binding motifs (Figure 5.14D). In contrast, the hg38g.30650690G>A variant intersects the NFYA and NFYC motifs at highly underrepresented nucleotides, suggesting that the variant is less likely to disrupt the binding of NFYA and NFYC (Figure 5.14D). Overall, the CHiP-seq data and motif analysis suggests the noncoding hg38g.30650690G>A variant is likely to be in a TF binding motif where CTCF, SP1 or KLF3 binding may occur.

Overall, bioinformatic analysis to predict the functional impact of hg38g.30650690G>A has revealed that this candidate variant has high potential for pathogenicity, and could potentially cause the downregulation of *PRRT2* observed in the CMT720 patients by disrupting a TF binding motif in a candidate enhancer element.

5.8.6. Segregation of the hg38g.30650690G>A variant suggests reduced disease penetrance in family CMT720

Since the bioinformatic analysis on Genome Browser databases suggested that hg38g.30650690G>A is potentially pathogenic, the available family members from CMT720 (17 in total) were genotyped to analyse the segregation of the candidate noncoding variant (Figure 5.15). Representative chromatograms showing variant validation are provided from selected CMT720 samples (Figure 5.16). Segregation analysis revealed the hg38g.30650690G>A is present in all affected individuals (III:4, III:8, IV:4, IV:6, IV:9, IV:11 and V:6) and individual V:4 who shows mild signs of the disease phenotype (Figure 5.15). Individuals IV:2, V:1 and V:5 who were coded as unknown phenotype for an “affected only” linkage analysis also carry the hg38g.30650690G>A, which could suggest the reduced disease penetrance for the hg38g.30650690G>A variant given the mild disease course and variability in phenotype as demonstrated by patient V:4 [168]. The Sanger sequencing also validates that the affected individuals IV:4 and V:6 carry the hg38g.30650690G>A noncoding variant by an orthogonal method in addition to srWGS and lrWGS.

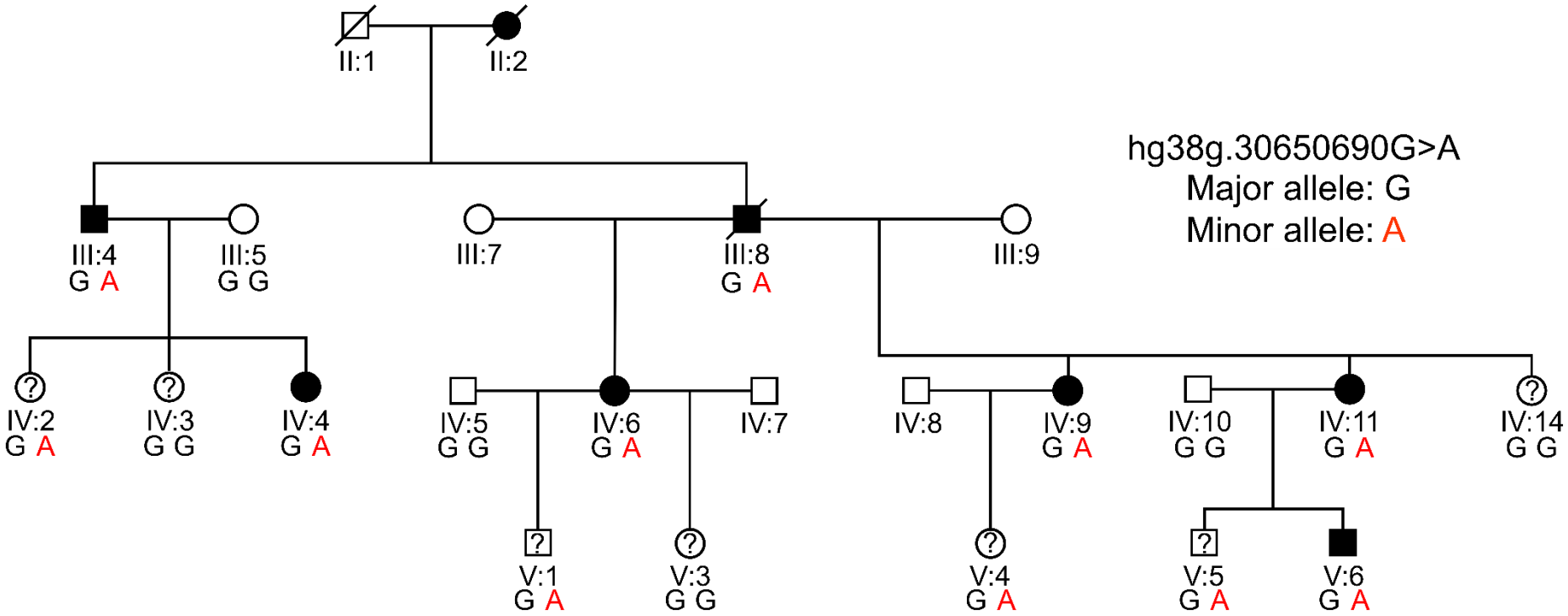


Figure 5.15: The pedigree of CMT720 showing the segregation of the candidate TFBS variant hg38g.30650690G>A. The candidate noncoding pathogenic variant is highlighted in red. This pedigree was adapted from the original publication by Kochanski *et al.* [168] and the individual identifiers were retained from the original pedigree. Question marks denote "unknown" phenotype. Unfilled symbols denote normal phenotype and closed symbols denote affected phenotype. Squares denote males and circles denote females. Diagonal line denotes deceased individuals.

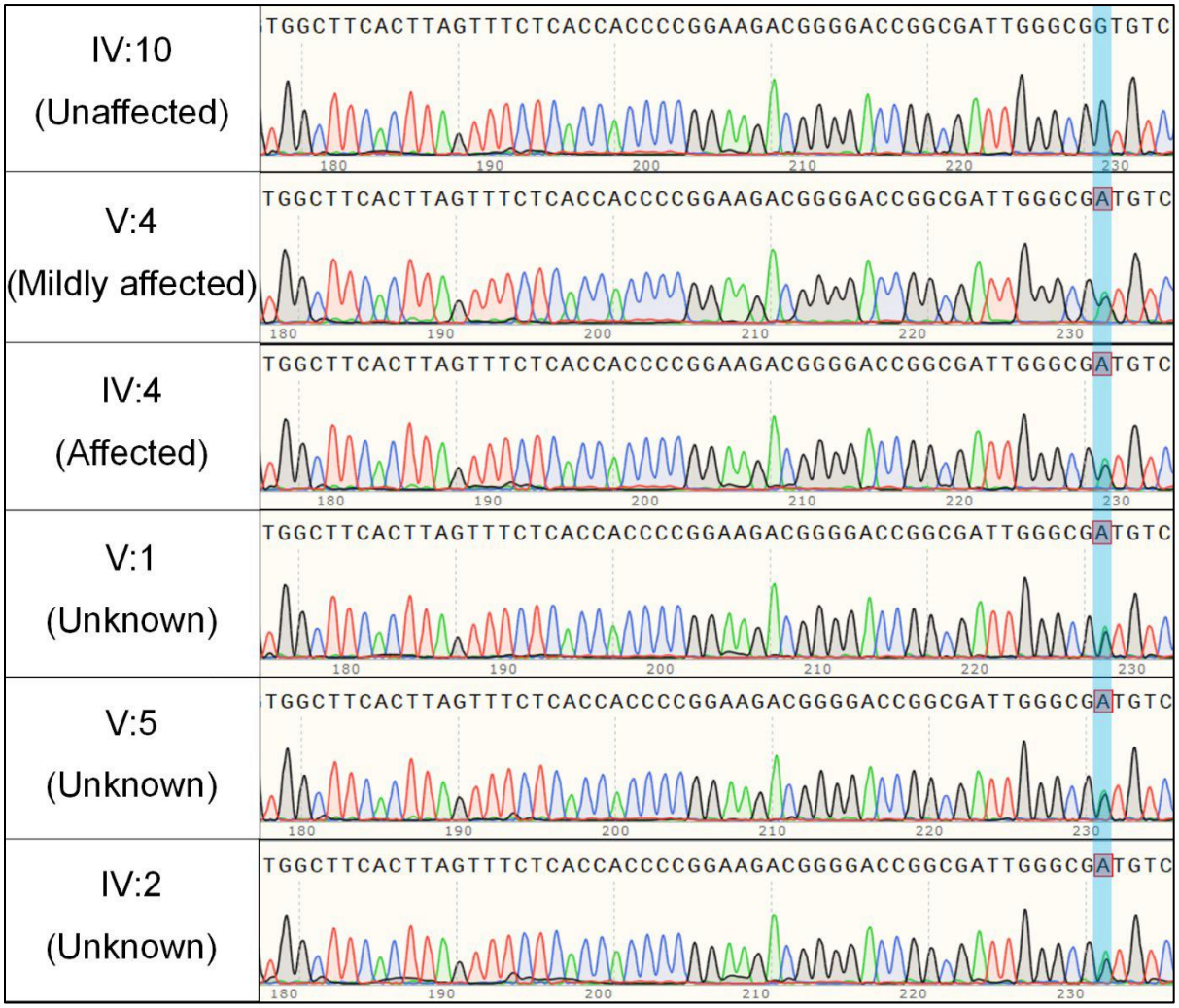


Figure 5.16: Sanger sequencing chromatograms of the CMT720 family members genotyped for segregation analysis of the noncoding variant hg38g.30650690G>A. Examples of a married-in unaffected individual homozygous for the reference allele (IV:10), the mildly affected individual V:4 and an affected individual (IV:4) being heterozygous for the candidate noncoding variant are shown. Chromatograms of the individuals with "unknown" phenotype (V:1, V:5 and IV:2) that are heterozygous for the candidate noncoding variant are also provided. The vertical blue line represents the position of the variant and the potentially pathogenic allele is indicated by the red boxes

5.9. Discussion

In this chapter, the full spectrum of genomic variants in CMT720 patients within the suggestive linkage regions on chromosome 8 and 16 were identified and prioritised using a combination of srWGS and lrWGS, variant filtering and multi-omics strategy. This approach facilitated the efficient selection and analysis of noncoding variants worthy of follow up functional experiments to test pathogenicity.

Using lrWGS aligned to the T2T reference increased the sensitivity and accuracy of variant detection, and allowed analysis of variants located in the challenging regions of the genome that remained inaccessible with srWGS technologies [248, 314]. The use of PacBio lrWGS detected approximately 27,000 SVs in patient IV:4. This number was similar to those obtained in other population studies (19,000-24,000 calls) utilising PacBio lrWGS [247]. The detection rate in this study was approximately 45% and 83% higher when compared to two studies [136, 315] utilising srWGS alignments against the hg38 reference respectively. These findings suggest the lrWGS alignment to the T2T reference substantially improved the variant call rate of SVs in the current investigation of CMT720. Similarly, a 21% increase in the number of high quality SNVs and indels, was observed in the T2T aligned lrWGS of patient IV:4 compared to the srWGS T2T alignment from the same individual. Utilising the T2T reference in combination with lrWGS also allowed detection of variation in a *DUSP22* paralog which localises to a gap in the hg38 reference [142], and was poorly covered by the patient srWGS reads. The *DUSP22* paralog is not expressed and unlikely to be biologically relevant [303], however the variation identified in this gene paralog demonstrated the ability of PacBio lrWGS to resolve some of the most challenging regions in the genome that were previously inaccessible with srWGS alignments used in the current investigation.

Although lrWGS is providing a powerful tool to identify candidate noncoding variants in family CMT720, the coverage across the large centromeric and pericentromeric repeat

arrays in the suggestive linkage region on chromosome 16 was poor in the T2T alignment of patient IV:4. Large centromeric repeats remain inaccessible to any single sequencing technology available today [316], and resolving these regions requires *de novo* telomere-to-telomere assembly by using the lrWGS data from both PacBio and ONT platforms [239]. Such studies can cost as much as 55,000 USD per genome [106], which can be prohibitive in a research setting with limited budget. Further sequencing to cover the centromeric regions on chromosome 16 will be required if these regions need to be revisited for candidate pathogenic SVs. Alternatively, combining the output from multiple SV calling algorithms may allow elimination of false positive calls [317], which are likely to account for a substantial proportion of the SVs predicted in lrWGS alignments across large repetitive regions [299]. In this investigation, Sniffles2 [283] was used to call the SVs. For exploratory studies such as this one, it is useful to run multiple callers to maximise variant detection. The re-analysis of CMT720 lrWGS alignments with additional detection algorithms and selecting for the commonly supported calls could help to eliminate the false positive calls around the centromeric repeat arrays in the suggestive linkage region on chromosome 16.

The EHDN pipeline did not call any novel REs in the patient srWGS alignments that are likely to represent real expansions within the suggestive linkage regions. Notably, no RE localised to the coding exons or intronic regions of genes, which suggests this previously unaddressed class of pathogenic variant is unlikely to cause CMT720. The potential involvement of REs in CMT720 however, cannot be excluded based on these findings alone, since poor alignment quality of reads from srWGS in repetitive regions, is known to cause high rates of false positive and false negative RE calls [318, 319], as well as inaccurate sizing for expansions [320]. In agreement with these reports, the EHDN algorithm produced a RE call in the coding region of the dysregulated positional candidate gene *IRX6*. This RE was manually excluded based on the absence of a corresponding variant in the patient's srWGS and lrWGS alignments. Due to the time constraints in the current project a RE detection pipeline

compatible with lrWGS could not be developed, and REs were detected by utilising available srWGS data instead. To achieve higher accuracy, a caller compatible with PacBio lrWGS could be utilised in future studies to identify the REs in CMT720 [321]. For example, the tool Straglr is capable of detecting novel REs and can identify expansions in PacBio alignments with >99% accuracy and sensitivity in benchmark studies [321]. Accordingly, reanalysis of patient lrWGS alignment with a compatible RE detection algorithm has the potential to reveal expansions that were missed by the analysis of patient srWGS alignments.

The variant filtering strategies implemented in this study have substantially reduced the burden of variant analysis by deprioritising the noncoding variants that are unlikely to be pathogenic. Targeting the suggestive linkage loci on chromosomes 8 and 16 provided the highest filtering efficiency by eliminating >98% of SNV, indel and SV calls, and >99% of RE calls respectively. Notably, targeting the suggestive linkage loci was instrumental for obtaining a number of SVs amenable to manual analysis. These findings demonstrate similar levels of filtering efficiency with the previous studies targeting linkage loci using NGS to reduce the number of candidate variants [75, 100]. The benign variants identified by PanGenie eliminated 23% of the remaining SV calls and 48% of the remaining SNV and indel calls that could not be removed by comparison to the 1KGP references [142, 247]. Accordingly, we have been able to demonstrate that the DHPG is a powerful tool for facilitating the analysis of the noncoding genome in unsolved IPN families by providing a diverse and bias free reference of the benign polymorphisms in the human population. While these resources were useful for aiding the interpretation of most variants, those located within the gaps completed in T2T remain poorly annotated in hg38 and the intronic variants in *DUSP22* which remains as VUS exemplify this shortcoming of working with the T2T reference. While these resources are still in their infancy, this study has shown that incorporating the genomic resources developed for the T2T reference in combination with the Pangenome will continue to improve the filtering efficiency and facilitate a more thorough analysis of the noncoding genome for family CMT720

in future investigations.

Despite the efficiency of our filtering strategy for srWGS, 19,139 SNVs and indels were retained in the lrWGS callset of patient IV:4. A second individual with a lrWGS aligned to the T2T reference could have improved the filtering, however, this resource was not available for the current project. A second lrWGS alignment would also be effective for selecting the SVs that are common in both patients. This would facilitate SV identification and eliminate the false positive calls located in the centromeric repeats in the suggestive linkage region on chromosome 16. While the SNV and indel calls obtained from srWGS and lrWGS alignments show high concordance in non-repetitive regions, repeats still pose major challenges for detection of SNVs and indels with srWGS [141, 243]. Comparative benchmarking studies have shown srWGS misses approximately 40% of the SNVs and indels within repetitive regions detected by lrWGS using DeepVariant [243]. lrWGS also detects more than 10 times the number of SNVs and indels that can be detected by srWGS across segmental duplications with high identity [141]. Therefore, sequencing a second individual with lrWGS will be a helpful future experiment for increasing filtering efficiency and improving variant detection.

In this investigation, online PCHi-C data was used to help select noncoding variants that could explain the dysregulation of the 4 positional candidate genes identified using transcriptomic analysis of CMT720 patient fibroblasts (Chapter 4). This multi-omics strategy identified 3 high priority noncoding variants among SNV, indel and SV calls that remained after standard filtering. The epigenomic evidence suggests the very rare noncoding candidate variant hg38g.30650690G>A could potentially disrupt an enhancer mediating long- range regulatory chromatin interactions with the dysregulated positional candidate gene *PRRT2*. It should be noted that, although the cCRE harboring hg38g.30650690G>A was classified as a candidate proximal enhancer, this classification does not indicate a functional relationship with the nearest genes, but is rather attributed to all candidate elements within +/-2kb of a transcription start site by ENCODE [270]. Despite the strengths of our multi-omics strategy,

the promoter regions of both *PRRT2* and *MAZ* were captured in the same interaction bin due to the limited resolution of the PCHi-C data utilised in this study [274], which prevents determining whether the candidate enhancer harbouring hg38g.30650690G>A specifically interacts with the promoter of *PRRT2.* Therefore, this regulatory interaction needs further validation by performing chromatin conformation assays with higher resolution on CMT720 patient neuronal tissue in the future investigations. Nevertheless, given that no other candidate variants were found in the introns, UTRs or PIRs of *PRRT2,* and the *MAZ* gene showed no dysregulation during transcriptomic analysis, disruption of a long-range enhancer-promoter contact by hg38g.30650690G>A remains the most well-supported predicted mechanism explaining the downregulation in *PRRT2*.

Accessing additional CHiP-seq data available from ENCODE [270] and ReMap [297] projects predicted hg38g.30650690G>A is likely to localise to a CTCF binding site. Motif analysis performed by Factorbook tool which is based on experimental TF binding data [298], predicted the variant alters highly overrepresented nucleotides at SP1 and KLF3 binding sites. Disruption of SP1 binding by g hg38g.30650690G>A may explain the downregulation in *PRRT2* as SP1 is a transcriptional activator that can mediate long-range enhancer-promoter contacts [322-324]. SNVs that disrupt SP1 binding sites are known to cause a variety of Mendelian disorders through the downregulation of gene expression [325-328]. Similarly, CTCF sites are a well-established mediator of long-range enhancer-promoter contacts [329] and noncoding SNVs disrupting CTCF binding can cause monogenic disorders due to gene downregulation [330]. Disruption of KLF3 binding is unlikely to explain the *PRRT2* downregulation in patient fibroblasts since this protein is a fundamental repressor with no known role as a transcriptional activator [331-333]. Due to the predictive evidence to suggest hg38g.30650690G>A is a variant impacting a TFBS that may cause the downregulation of *PRRT2*, this variant remains a high-ranking candidate that should be followed up by functional studies.

Overall, our findings in this chapter have provided a proof of concept for using lrWGS, improved genomics resources and multi-omics analysis as a powerful strategy for efficiently analysing noncoding variants in this unsolved family with axonal CMT.

5.10. Supplementary material

Supplementary Table 5.1: Preprocessing of the 1KGP T2T reference VCF file containing SNV and indel calls.

	# *Preprocessing the* 1KGP T2T reference chromosome 8 VCF
1	bcftools norm -m-any 1KGP.CHM13v2.0.chr8.recalibrated.snp_indel.pass.vcf.gz > 1KGP_CHR8_split.vcf
2	bgzip 1KGP_CHR8_split.vcf
3	tabix -p vcf 1KGP_CHR8_split.vcf.gz
4	bcftools view -q 0.001:minor 1KGP_CHR8_split.vcf.gz > 1KGP_CHR8_split_AF0.001.vcf
5	bgzip 1KGP_CHR8_split_AF0.001.vcf
6	tabix -p vcf 1KGP_CHR8_split_AF0.001.vcf.gz
	# *Preprocessing the* 1KGP T2T reference chromosome 8 VCF
7	bcftools norm -m-any 1KGP.CHM13v2.0.chr16.recalibrated.snp_indel.pass.vcf.gz > 1KGP_CHR16_split.vcf
8	bgzip 1KGP_CHR16_split.vcf
9	tabix -p vcf 1KGP_CHR16_split.vcf.gz
10	bcftools view -q 0.001:minor 1KGP_CHR16_split.vcf.gz > 1KGP_CHR16_split_AF0.001.vcf
11	bgzip 1KGP_CHR16_split_AF0.001.vcf
12	tabix -p vcf 1KGP_CHR16_split_AF0.001.vcf.gz

Supplementary Table 5.2: Preprocessing the PanGenie VCF files.

	# Merging the vcf files produced by genotyping the control and patient samples using PanGenie
1	bcftools isec -n+1 -w1 -c none -p Pangenome_all_merged P1_genotyping_biallelic.vcf.gz P2_genotyping_biallelic.vcf.gz C1_genotyping_biallelic.vcf.gz C2_genotyping_biallelic.vcf.gz C3_genotyping_biallelic.vcf.gz

Supplementary Table 5.3: Removing low quality calls.

	# Removing the low quality variant calls in control srWGS VCF files
1	bcftools view -i 'QUAL >= 20' C1.vcf > C1_QUAL_20.vcf

2	bgzip C1_QUAL_20.vcf
3	tabix -p vcf C1_QUAL_20.vcf.gz
4	bcftools view -i 'QUAL >= 20' C2.vcf > C2_QUAL_20.vcf
5	bgzip C2_QUAL_20.vcf
6	tabix -p vcf C2_QUAL_20.vcf.gz
7	bcftools view -i 'QUAL >= 20' C3.vcf > C3_QUAL_20.vcf
8	bgzip C3_QUAL_20.vcf
9	tabix -p vcf C3_QUAL_20.vcf.gz
	# Removing the low quality variant calls in patient srWGS VCF files (patients IV:4 and IV:6)
10	bcftools view -i 'QUAL >= 20' P1.vcf > P1_QUAL_20.vcf
11	bgzip P1_QUAL_20.vcf
12	tabix- p vcf P1_QUAL_20.vcf.gz
13	bcftools view -i 'QUAL >= 20' P2.vcf > P2_QUAL_20.vcf
14	bgzip P2_QUAL_20.vcf
15	tabix- p vcf P2_QUAL_20.vcf.gz
	# Removing the low quality variant calls in the patient lrWGS VCF file (patient IV:4)
16	bcftools view -f 'PASS' MARKEN_CMT.CMT720_combined_PB.chm13.SNPs_Indels.phased.vcf.gz > PacBio_T2T_SNV_Indel_PASS.vcf
17	bgzip PacBio_T2T_SNV_Indel_PASS.vcf
18	tabix -p vcf PacBio_T2T_SNV_Indel_PASS.vcf.gz

Supplementary Table 5.4: Merging the control VCF files.

	# Combining all calls from controls into a single VCF
1	bcftools isec -n+1 -w1 -c none -p MERGED_CONTROL C1_QUAL_20.vcf.gz C2_QUAL_20.vcf.gz C3_QUAL_20.vcf.gz
2	bgzip MERGED_CONTROL.vcf
3	tabix- p vcf MERGED_CONTROL.vcf.gz

Supplementary Table 5.5: Variant filtering pipeline for the SNVs and indels identified in the CMT720 srWGS data aligned against the T2T reference.

	# Merging the VCF files from both affected individuals to identify the common calls
1	bcftools isec -p affected_shared -n=2 -w1 -c none P1_QUAL_20.vcf.gz P2_QUAL_20.vcf.gz
2	bgzip affected_shared.vcf
3	tabix-p vcf affected_shared.vcf.gz
4	bcftools stats affected_shared.vcf.gz > affected_shared.txt
	# Removing the variants identified in controls
5	bcftools isec -C -Oz -p CMT720_Unique affected_shared.vcf.gz MERGED_CONTROL.vcf.gz
6	bcftools stats CMT720_Unique.vcf.gz > CMT720_Unique.txt

```
   # Splitting the multiallelic calls into biallelic calls
7  bcftools norm -m-any CMT720_Unique.vcf.gz > CMT720_Unique_split.vcf
8  bgzip CMT720_Unique_split.vcf
9  tabix-p vcf CMT720_Unique_split.vcf.gz
10 bcftools stats CMT720_Unique_split.vcf.gz > CMT720_Unique_split.txt
   # Filtering the calls from outside the prioritised linkage regions
11 bcftools view -r chr8:69547472-89287890 CMT720_Unique_split.vcf.gz -o
   CMT720_Unique_split_CHR8LR.vcf
12 bgzip CMT720_Unique_split_CHR8LR.vcf
13 tabix-p vcf CMT720_Unique_split_CHR8LR.vcf.gz
14 bcftools view -r chr16:23128937-61435132 CMT720_Unique_split.vcf.gz -o
   CMT720_Unique_split_CHR16LR.vcf
15 bgzip CMT720_Unique_split_CHR16LR.vcf
16 tabix-p vcf CMT720_Unique_split_CHR16LR.vcf.gz
17 bcftools isec  -n+1 -w1 -c none -p CMT720_Unique_split_LR
   CMT720_Unique_split_CHR8LR.vcf.gz CMT720_Unique_split_CHR16LR.vcf.gz
18 bgzip CMT720_Unique_split_LR.vcf
19 tabix-p vcf CMT720_Unique_split_LR.vcf.gz
20 bcftools stats CMT720_Unique_split_LR.vcf.gz > CMT720_Unique_split_LR.txt
   # Filtering out the SNVs and indels with MAF>0.001 in the 1KGP T2T reference
21 bcftools isec -C -Oz -p CMT720_1KGP_CHR8LR CMT720_Unique_split_LR.vcf.gz
   1KGP_CHR8LR_split_AF0.001.vcf.gz
22 bcftools isec -C -Oz -p CMT720_1KGP_LR CMT720_1KGP_CHR8LR.vcf.gz
   1KGP_CHR16LR_split_AF0.001.vcf.gz
23 bcftools stats CMT720_1KGP_LR.vcf.gz > CMT720_1KGP_LR.txt
   # Filtering out the biallelic variants genotyped by PanGenie
24 bcftools isec -C -Oz -p CMT720_1KGP_PG_LR CMT720_1KGP_LR.vcf.gz
   pangenome_all_merged.vcf.gz
25 bcftools stats CMT720_1KGP_PG_LR.vcf.gz > CMT720_1KGP_PG_LR.txt
   # Filtering out the homozygous calls
26 bcftools view -i 'GT[*]="0/1"' CMT720_1KGP_PG_LR.vcf.gz -o
   CMT720_1KGP_PG_LR_HET.vcf
27 bgzip CMT720_1KGP_PG_LR_HET.vcf
28 tabix – p vcf CMT720_1KGP_PG_LR_HET.vcf.gz
   # Splitting the final output of the pipeline into individual VCF files containing the
   variant calls from each prioritised linkage region
29 bcftools view -r chr8:69547472-89287890 CMT720_1KGP_PG_LR_HET.vcf.gz  -o
   CMT720_1KGP_PG_HET_CHR8LR.vcf
30 bgzip CMT720_1KGP_PG_HET_CHR8LR.vcf
31 bcftools stats CMT720_1KGP_PG_HET_CHR8LR.vcf.gz >
   CMT720_1KGP_PG_HET_CHR8LR.txt
32 bcftools view -r chr16:23128937-61435132 CMT720_1KGP_PG_LR_HET.vcf.gz  -o
   CMT720_1KGP_PG_HET_CHR16LR.vcf
33 bgzip CMT720_1KGP_PG_HET_CHR16LR.vcf
34 bcftools stats CMT720_1KGP_PG_HET_CHR16LR.vcf.gz >
   CMT720_1KGP_PG_HET_CHR16LR.txt
```

Supplementary Table 5.6: Variant filtering pipeline for the SNVs and indels identified in the CMT720 lrWGS data aligned against the T2T reference.

	# Removing the variants identified in controls
1	bcftools isec -C -Oz -p CMT720_PacBio_Unique PacBio_T2T_SNV_Indel_PASS.vcf.gz MERGED_CONTROL.vcf.gz
2	bcftools stats CMT720_PacBio_Unique.vcf.gz > CMT720_PacBio_Unique.txt
	# Splitting the multiallelic calls into biallelic calls
3	bcftools norm -m-any CMT720_PacBio_Unique.vcf.gz > CMT720_PacBio_Unique_split.vcf.gz
4	bgzip CMT720_PacBio_Unique_split.vcf
5	tabix-p vcf CMT720_PacBio_Unique_split.vcf.gz
6	bcftools stats CMT720_PacBio_Unique_split.vcf.gz > CMT720_PacBio_Unique_split.txt
	# Filtering the calls from outside the prioritised linkage regions
7	bcftools view -r chr8:69547472-89287890 CMT720_PacBio_Unique_split.vcf.gz -o CMT720_PacBio_Unique_split_CHR8LR.vcf
8	bgzip CMT720_PacBio_Unique_split_CHR8LR.vcf
9	tabix-p vcf CMT720_PacBio_Unique_split_CHR8LR.vcf.gz
10	bcftools view -r chr16:23128937-61435132 CMT720_PacBio_Unique_split.vcf.gz -o CMT720_PacBio_Unique_split_CHR16LR.vcf
11	bgzip CMT720_PacBio_Unique_split_CHR16LR.vcf
12	tabix-p vcf CMT720_PacBio_Unique_split_CHR16LR.vcf.gz
13	bcftools isec -n+1 -w1 -c none -p CMT720_PacBio_Unique_split_LR CMT720_PacBio_Unique_split_CHR8LR.vcf.gz CMT720_PacBio_Unique_split_CHR16LR.vcf.gz
14	bgzip CMT720_PacBio_Unique_split_LR.vcf
15	tabix-p vcf CMT720_PacBio_Unique_split_LR.vcf.gz
16	bcftools stats CMT720_PacBio_Unique_split_LR.vcf.gz > CMT720_PacBio_Unique_split_LR.txt
	# Filtering out the SNVs and indels with MAF>0.001 identified in the realignment of 1KGP srWGS data
17	bcftools isec -C -Oz -p CMT720_PacBio_1KGP_CHR8LR CMT720_PacBio_Unique_split_LR.vcf.gz 1KGP_CHR8LR_split_AF0.001.vcf.gz
18	bcftools isec -C -Oz -p CMT720_PacBio_1KGP_LR CMT720_PacBio_1KGP_CHR8LR.vcf.gz 1KGP_CHR16LR_split_AF0.001.vcf.gz
19	bcftools stats CMT720_PacBio_1KGP_LR.vcf.gz > CMT720_PacBio_1KGP_LR.txt
	# Filtering out the biallelic variants genotyped by PanGenie
20	bcftools isec -C -Oz -p CMT720_PacBio_1KGP_PG_LR CMT720_PacBio_1KGP_LR.vcf.gz pangenome_all_merged.vcf.gz
21	bcftools stats CMT720_PacBio_1KGP_PG_LR.vcf.gz > CMT720_PacBio_1KGP_PG_LR.txt
	# Filtering out homozygous calls
22	bcftools view -i 'GT[*]="0/1"' CMT720_PacBio_1KGP_PG_LR.vcf.gz -o CMT720_PacBio_1KGP_PG_LR_HET.vcf
23	bgzip CMT720_PacBio_1KGP_PG_LR_HET.vcf
24	tabix -p vcf CMT720_PacBio_1KGP_PG_LR_HET.vcf.gz
	# Splitting the final output of the pipeline into individual vcf files containing the variant calls from each prioritised linkage region
25	bcftools view -r chr8:69547472-89287890 CMT720_PacBio_1KGP_PG_LR_HET.vcf.gz -o CMT720_PacBio_1KGP_PG_HET_CHR8LR.vcf

```
26 bgzip CMT720_PacBio_1KGP_PG_HET_CHR8LR.vcf
27 bcftools stats CMT720_PacBio_1KGP_PG_ HET_CHR8LR.vcf.gz >
   CMT720_PacBio_1KGP_PG_ HET_CHR8LR.txt
28 bcftools view -r chr16:23128937-61435132 CMT720_PacBio_1KGP_PG_LR_HET.vcf.gz -
   o CMT720_PacBio_1KGP_PG_HET_CHR16LR.vcf
29 bgzip CMT720_PacBio_1KGP_PG_HET_CHR16LR.vcf
30 bcftools stats CMT720_PacBio_1KGP_PG_HET_CHR16LR.vcf.gz >
   CMT720_PacBio_1KGP_PG_HET_CHR16LR.vcf.txt
```

Supplementary Section 5.9.1: Quality analysis for srWGS and lrWGS T2T alignments

PacBio HiFi lrWGS generated a total of 105.91 Gb of sequence distributed across 8,203,446 reads, with an average read length of 12,911 bp. The average Phred score per base was 38.6, indicating 99.99% accuracy. 0.39% of reads had mapping quality (MQ) <0.

On average, Illumina srWGS generated a total of 126.09 Gb of sequence distributed across 841,080,092 reads with a median read length of 151 bp. The average Phred score per base across all patient and control alignments was 35.9, indicating 99.97% accuracy. On average, 6.36% of the reads had MQ<0. Comparing the srWGS and lrWGS data from patient IV:4 aligned against the T2T reference indicates an increase from 35.0 to 38.5 in the average Phred score per base, which corresponds to an increase 0.02% in base calling accuracy. Notably, the proportion of reads with MQ<0 is reduced to 0.39% in the PacBio HiFi lrWGS alignment from 6.46% in the Illumina srWGS alignment, indicating an increase of 6.07% in mapping accuracy. These findings suggest that using PacBio HiFi to perform lrWGS has substantially increased the quality of alignment when compared to the srWGS alignment of CMT720 patient IV:4.

The patient srWGS alignment against the hg38 reference was generated by our collaborator Dr. Chiara Folland to be used for the EHDN pipeline and we did not have access to the resulting the statistics on BAM quality.

Supplementary Table 5.7: Quality control statistics on aligned patient lrWGS data.

Sample	Total bases (Gb)	Total reads	Reads with MQ<0 (%)	Average Phred score	Average read length
IV:4	105.91	8203446	31785 (0.39)	38.6	12911

MQ: Mapping quality

Supplementary Table 5.8: Quality control statistics on aligned patient and control srWGS data.

Sample	Total bases (Gb)	Total reads	Reads with MQ<0 (%)	Average Phred score	Average read length
IV:4	129.23	872102038	56369673 (6.46)	35	151
V:6	111.76	740155156	50505233 (6.82)	35.1	148
C1	151.44	1016871100	953896943 (6.20)	34.8	148
C2	119.16	789166738	743335028 (5.81)	38.4	151
C3	118.85	787105428	735911171 (6.51)	36.2	151

MQ: Mapping quality

Supplementary Figure 5.1: The regions showing poor coverage in the PacBio lrWGS T2T alignment of patient IV:4 for the prioritised linkage regions on chromosome 8 and 16. Red bars below the coordinate axis in both panels represent the regions with poor coverage. (**A**) Within the prioritised linkage region on chromosome 16, the repetitive centromeric regions cause abnormal changes in the coverage of PacBio lrWGS alignment, which starts at 35,850 kb

mark on chromosome 16 with the beginning of centromeric repeats. These fluctuations in read coverage coincide with high frequency of base mismatches indicated by colored lines in the coverage track and clusters of SV calls with questionable accuracy, which were excluded during filtering (**B**) A 5.27 Mb long representative section of the repeat arrays of chromosome 16q11.2 showing abnormal fluctuations in read coverage and complete mismatching of base calls indicated by the colored lines in the coverage track. Chromosome 16q11.2 is a large pericentromeric repeat array [300] that spans approximately 14.5 Mb in the T2T reference. Due to the unreliable base calls and high repetitiveness of this locus, the SV calls localizing to chromosome 16q11.2 were deprioritised for further analysis in this study.

Supplementary Table 5.9: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of *ZNF704* in DLPFC.

Promoter region			PIRs		
Chromosome	Start coordinates	End coordinates	Chromosome	Start coordinates	End coordinates
chr8	81304362	81308372	chr8	80456238	80465277
chr8	81304362	81308372	chr8	80915981	80919793
chr8	81304362	81308372	chr8	80915981	80919793
chr8	81304362	81308372	chr8	80456238	80465277

Supplementary Table 5.10: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of *ZNF704* in hippocampus.

Promoter region			PIRs		
Chromosome	Start coordinates	End coordinates	Chromosome	Start coordinates	End coordinates
chr8	81304362	81308372	chr8	79432644	79436994
chr8	81304362	81308372	chr8	80133275	80136972
chr8	81304362	81308372	chr8	80365544	80375982
chr8	81304362	81308372	chr8	80476155	80479704
chr8	81304362	81308372	chr8	80508251	80511324
chr8	81304362	81308372	chr8	80528569	80535968
chr8	81304362	81308372	chr8	80565807	80570998
chr8	81304362	81308372	chr8	80579215	80584876
chr8	81304362	81308372	chr8	80601128	80605215
chr8	81304362	81308372	chr8	80912807	80915980
chr8	81304362	81308372	chr8	80915981	80919793
chr8	81304362	81308372	chr8	80915981	80919793
chr8	81308373	81316054	chr8	80915981	80919793
chr8	81304362	81308372	chr8	80919794	80926452
chr8	81304362	81308372	chr8	80926453	80932251
chr8	81304362	81308372	chr8	80936369	80940039
chr8	81304362	81308372	chr8	80943074	80947248
chr8	81304362	81308372	chr8	80947249	80951287
chr8	81304362	81308372	chr8	80951288	80956249
chr8	81304362	81308372	chr8	80997063	81001472
chr8	81304362	81308372	chr8	81001473	81004584
chr8	81304362	81308372	chr8	81051196	81055595
chr8	81304362	81308372	chr8	81055596	81060626
chr8	81304362	81308372	chr8	81069147	81074533
chr8	81304362	81308372	chr8	81086715	81089907
chr8	81304362	81308372	chr8	81089908	81093917
chr8	81304362	81308372	chr8	81108880	81113267

chr8	81304362	81308372	chr8	81113268	81116293
chr8	81304362	81308372	chr8	81116294	81120437
chr8	81304362	81308372	chr8	81120438	81126225
chr8	81304362	81308372	chr8	81126226	81132391
chr8	81304362	81308372	chr8	81132392	81137155
chr8	81304362	81308372	chr8	81137156	81143813
chr8	81304362	81308372	chr8	81143814	81148548
chr8	81304362	81308372	chr8	81148549	81151749
chr8	81304362	81308372	chr8	81151750	81160025
chr8	81304362	81308372	chr8	81160026	81170275
chr8	81304362	81308372	chr8	81170276	81174891
chr8	81304362	81308372	chr8	81174892	81179258
chr8	81304362	81308372	chr8	81182714	81185924
chr8	81304362	81308372	chr8	81185925	81189926
chr8	81304362	81308372	chr8	81212626	81217323
chr8	81304362	81308372	chr8	81217324	81223808
chr8	81304362	81308372	chr8	81234298	81237886
chr8	81304362	81308372	chr8	81237887	81242991
chr8	81304362	81308372	chr8	81242992	81246902
chr8	81304362	81308372	chr8	81267716	81273021
chr8	81304362	81308372	chr8	81273022	81281120
chr8	81304362	81308372	chr8	81281121	81286031
chr8	81304362	81308372	chr8	81324464	81330333
chr8	81304362	81308372	chr8	81330334	81336897
chr8	81304362	81308372	chr8	81343904	81346942
chr8	81304362	81308372	chr8	81346943	81350666
chr8	81304362	81308372	chr8	81350667	81353751
chr8	81304362	81308372	chr8	81353752	81357709
chr8	81304362	81308372	chr8	81357710	81362128
chr8	81304362	81308372	chr8	81365522	81369267
chr8	81304362	81308372	chr8	81369268	81378162
chr8	81304362	81308372	chr8	81394168	81399147
chr8	81304362	81308372	chr8	81399148	81403475
chr8	81304362	81308372	chr8	81414135	81423719
chr8	81304362	81308372	chr8	81431740	81435837
chr8	81304362	81308372	chr8	81435838	81441506
chr8	81304362	81308372	chr8	81455837	81462090
chr8	81304362	81308372	chr8	81533043	81536918
chr8	81304362	81308372	chr8	81541086	81548638
chr8	81304362	81308372	chr8	81541086	81548638
chr8	81304362	81308372	chr8	81664493	81667696

Supplementary Table 5.11: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of *PRRT2* in DLPFC.

Promoter region			PIRs		
Chromosome	Start coordinates	End coordinates	Chromosome	Start coordinates	End coordinates
chr16	30085512	30095827	chr16	28201220	28214065
chr16	30085512	30095827	chr16	29099207	29110039
chr16	30085512	30095827	chr16	29099207	29110039
chr16	30085512	30095827	chr16	29119167	29146304
chr16	30085512	30095827	chr16	29119167	29146304
chr16	30085512	30095827	chr16	29165015	29176962
chr16	30085512	30095827	chr16	29227305	29236284
chr16	30085512	30095827	chr16	29227305	29236284
chr16	30085512	30095827	chr16	29244234	29247454
chr16	30085512	30095827	chr16	29253039	29283043
chr16	30085512	30095827	chr16	29253039	29283043
chr16	30085512	30095827	chr16	29378096	29381537
chr16	30085512	30095827	chr16	29395470	29402327
chr16	30085512	30095827	chr16	29486858	29491250
chr16	30085512	30095827	chr16	29509868	29518959
chr16	30085512	30095827	chr16	29856073	29883065
chr16	30085512	30095827	chr16	29893803	29902696
chr16	30085512	30095827	chr16	29980200	29991103
chr16	30085512	30095827	chr16	30104554	30109538
chr16	30085512	30095827	chr16	30112930	30117275
chr16	30085512	30095827	chr16	30134770	30144944
chr16	30085512	30095827	chr16	30144945	30151476
chr16	30085512	30095827	chr16	30144945	30151476
chr16	30085512	30095827	chr16	30174697	30184912
chr16	30085512	30095827	chr16	30174697	30184912
chr16	30085512	30095827	chr16	30190699	30197677
chr16	30085512	30095827	chr16	30197678	30215906
chr16	30085512	30095827	chr16	30197678	30215906
chr16	30085512	30095827	chr16	30223742	30232858
chr16	30085512	30095827	chr16	30238856	30243044
chr16	30085512	30095827	chr16	30243045	30250050
chr16	30085512	30095827	chr16	30250051	30255939
chr16	30085512	30095827	chr16	30255940	30266631
chr16	30085512	30095827	chr16	30255940	30266631
chr16	30085512	30095827	chr16	30266632	30279731
chr16	30085512	30095827	chr16	30266632	30279731
chr16	30085512	30095827	chr16	30279732	30283736
chr16	30085512	30095827	chr16	30283737	30293506
chr16	30085512	30095827	chr16	30293507	30324117
chr16	30085512	30095827	chr16	30293507	30324117

chr16	30085512	30095827	chr16	30324118	30333323
chr16	30085512	30095827	chr16	30333324	30335132
chr16	30085512	30095827	chr16	30333324	30335132
chr16	30085512	30095827	chr16	30336298	30343035
chr16	30085512	30095827	chr16	30343036	30360103
chr16	30085512	30095827	chr16	30343036	30360103
chr16	30085512	30095827	chr16	30360104	30361310
chr16	30085512	30095827	chr16	30361311	30392175
chr16	30085512	30095827	chr16	30361311	30392175
chr16	30085512	30095827	chr16	30401134	30433408
chr16	30085512	30095827	chr16	30401134	30433408
chr16	30085512	30095827	chr16	30733944	30745286
chr16	30085512	30095827	chr16	30733944	30745286
chr16	30085512	30095827	chr16	30748537	30753369
chr16	30085512	30095827	chr16	30753370	30760051
chr16	30085512	30095827	chr16	30753370	30760051
chr16	30085512	30095827	chr16	30928989	30934007
chr16	30085512	30095827	chr16	31004066	31015042
chr16	30085512	30095827	chr16	31004066	31015042
chr16	30085512	30095827	chr16	31034877	31037876
chr16	30085512	30095827	chr16	31034877	31037876
chr16	30085512	30095827	chr16	31078929	31090977
chr16	30085512	30095827	chr16	31078929	31090977

Supplementary Table 5.12: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of *PRRT2* in hippocampus.

Promoter region			PIRs		
Chromosome	Start coordinate	End coordinate	Chromosome	Start coordinate	End coordinate
chr16	30085512	30095827	chr16	30948995	30966907
chr16	30085512	30095827	chr16	29509868	29518959
chr16	30085512	30095827	chr16	30361311	30392175
chr16	30085512	30095827	chr16	30343036	30360103
chr16	30085512	30095827	chr16	30293507	30324117
chr16	30085512	30095827	chr16	30279732	30283736
chr16	30085512	30095827	chr16	30266632	30279731
chr16	30085512	30095827	chr16	30255940	30266631
chr16	30085512	30095827	chr16	30243045	30250050
chr16	30085512	30095827	chr16	30190699	30197677
chr16	30085512	30095827	chr16	30174697	30184912
chr16	30085512	30095827	chr16	30157388	30164189
chr16	30085512	30095827	chr16	30144945	30151476
chr16	30085512	30095827	chr16	30104554	30109538

chr16	30085512	30095827	chr16	30144945	30151476
chr16	30085512	30095827	chr16	30174697	30184912
chr16	30085512	30095827	chr16	30255940	30266631
chr16	30085512	30095827	chr16	30266632	30279731
chr16	30085512	30095827	chr16	30293507	30324117
chr16	30085512	30095827	chr16	30343036	30360103
chr16	30085512	30095827	chr16	30361311	30392175
chr16	30085512	30095827	chr16	30948995	30966907

Supplementary Table 5.13: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of *BCL7C* in DLPFC.

Promoter region			PIRs		
Chromosome	Start coordinate	End coordinate	Chromosome	Start coordinate	End coordinate
chr16	31279240	31290643	chr16	30401134	30433408
chr16	31279240	31290643	chr16	30401134	30433408
chr16	31279240	31290643	chr16	30786484	30792342
chr16	31279240	31290643	chr16	30828481	30838417
chr16	31279240	31290643	chr16	31096110	31100231
chr16	31279240	31290643	chr16	31100232	31105260
chr16	31279240	31290643	chr16	31130569	31139690
chr16	31279240	31290643	chr16	31130569	31139690
chr16	31279240	31290643	chr16	31144177	31153395
chr16	31279240	31290643	chr16	31144177	31153395
chr16	31279240	31290643	chr16	31158581	31163310
chr16	31279240	31290643	chr16	31163311	31168904
chr16	31279240	31290643	chr16	31185322	31192962
chr16	31279240	31290643	chr16	31263234	31266411
chr16	31279240	31290643	chr16	31301195	31312773
chr16	31279240	31290643	chr16	31301195	31312773
chr16	31279240	31290643	chr16	31326565	31331612
chr16	31279240	31290643	chr16	31771818	31776072
chr16	31279240	31290643	chr16	31847048	31851756

Supplementary Table 5.14: T2T coordinates of the PIRs forming chromatin contacts with the promoter region of *BCL7C* in hippocampus.

Promoter region			PIRs		
Chromosome	Start coordinate	End coordinate	Chromosome	Start coordinate	End coordinate
chr16	31279240	31290643	chr16	31840696	31847047
chr16	31279240	31290643	chr16	31144177	31153395

chr16	31279240	31290643	chr16	31301195	31312773
chr16	31279240	31290643	chr16	31840696	31847047
chr16	31279240	31290643	chr16	30996471	31004065
chr16	31279240	31290643	chr16	31015043	31019425
chr16	31279240	31290643	chr16	31019426	31026985
chr16	31279240	31290643	chr16	31090978	31096109
chr16	31279240	31290643	chr16	31105261	31109940
chr16	31279240	31290643	chr16	31144177	31153395
chr16	31279240	31290643	chr16	31163311	31168904
chr16	31279240	31290643	chr16	31301195	31312773

Supplementary Table 5.15: T2T coordinates of the PIRs contacting the promoter region of *IRX6* in DLPFC.

Promoter region			PIRs		
Chromosome	Start	End	Chromosome	Start	End
chr16	61118066	61123998	chr16	60114248	60121476

Supplementary Table 5.16: The TF CHiP-seq data accessed during variant analysis.

ENCODE Regulation track [270] was used to access the TF CHiP-seq data obtained from bipolar neurons transdifferentiated from the GM23338 cell line and neural cells transdifferentiated from H1 human embryonic stem cells. ReMap CHiP-seq track [297] was used to access the TF CHiP-seq data by selecting the “fibroblast” and “neuron” biotypes, which displays the TF CHiP-seq data from all cultured cell lines that are categorised under the indicated biotypes. All available TF CHiP-seq experiments performed on the fibroblasts and neuronal cell lines by the ENCODE and ReMap projects were accessed during analysis.

Transcriptional regulator	ReMap CHiP-Seq		ENCODE Regulatory	
	Fibroblasts	Neurons	Bipolar neurons	Neural cells
AR	✓			
CREBBP	✓			
CTCF	✓	✓	✓	✓
EGR1	✓			
EP300	✓			
EZH2	✓	✓		✓
MXI1				✓
PCGF2	✓			

PML	✓			
POU5F1	✓	✓		
PRKDC	✓			
RAD21				✓
RB1	✓			
RNF2	✓			
SMARCA4			✓	
SMC3				✓
ZEB1		✓	✓	

* Tick marks denote the availability of a TF CHiP-seq experiment from a tissue of interest in the indicated Genome Browser tracks.

Supplementary Table 5.17: The distance between hg38g.30650690G>A and the summits of the CHiP-seq peaks intersected by this candidate variant.

Database	Cell type	Transcription factor	Summit distance (bp)
ENCODE	Neural cell	CTCF	30
ENCODE	Neural cell	MXI1	162
ReMap	Neuron	CTCF	128
ReMap	Fibroblast	CTCF	0
ReMap	Fibroblast	CREBBP	13
ReMap	Fibroblast	EGR1	165
ReMap	Fibroblast	RNF2	19

Supplementary Table 5.18: The canonical TF binding motifs hg38g.30650690G>A is predicted to overlap by Factorbook

SNP	SNP coordinates	Best database match	FDR corrected p-value
rs1364535320	chr16:30650689-30650690	NFYA_HUMAN.H11MO.0.A (HOCOMOCO)	0.0142
rs1364535320	chr16:30650689-30650690	NFYA_HUMAN.H11MO.0.A (HOCOMOCO)	0.0215
rs1364535320	chr16:30650689-30650690	SP1_HUMAN.H11MO.0.A (HOCOMOCO)	0.0227
rs1364535320	chr16:30650689-30650690	NFYA_HUMAN.H11MO.0.A (HOCOMOCO)	0.0244
rs1364535320	chr16:30650689-30650690	SP1_HUMAN.H11MO.0.A (HOCOMOCO)	0.0284
rs1364535320	chr16:30650689-30650690	KLF3_HUMAN.H11MO.0.B (HOCOMOCO)	0.0286
rs1364535320	chr16:30650689-30650690	NFYC_HUMAN.H11MO.0.A (HOCOMOCO)	0.0318
rs1364535320	chr16:30650689-	NFYC_HUMAN.H11MO.0.A	0.0361

30650690	(HOCOMOCO)

Chapter 6
Final Discussion

Despite the widespread use of high-throughput sequencing technologies, approximately 40% of CMT cases remain unsolved following interrogation of the exome with particularly poorer diagnostic rates observed for CMT2 [68, 117, 118]. In CMT cases for which coding mutations have been excluded, noncoding variants constitute an important class of pathogenic mutations that may explain the missing heritability [125, 130, 132]. Since family CMT720 remains unsolved following exclusion of all coding variants within the 5 previously established suggestive linkage loci, analysing noncoding variants to identify the pathogenic mutations was a feasible strategy to genetically solve the family. Despite the clinical importance of noncoding variants, it remains prohibitively difficult to identify pathogenic noncoding mutations [334]. This is partly due to the structural complexity and localisation of noncoding variants to the genomic regions that are challenging to sequence with the commonly used NGS technologies [245]. The combined use of lrWGS technologies and the complete T2T genome reference was reported as an effective strategy for resolving the repetitive and previously inaccessible parts of the genome, improving the detection of all forms of variants [141]. Even after successful detection of noncoding variants, their abundance and a lack of computational methods for accurately predicting their impact adds enormous burden on variant analysis [196, 335]. A novel genomics resource that aids variant interpretation is the human pangenome, which provides an unbiased reference of “normal” variation in the human population [144, 336, 337]. To address the challenges in interpreting the impact of noncoding variants, multi-omics approaches have been increasingly employed for prioritising noncoding variants based on their functional impacts in genetic studies of unsolved Mendelian disorders [128, 158]. As a monogenic disorder that remains unsolved following exclusion of

pathogenic coding variants, CMT720 was a suitable family for exploring the utility of multi-omics variant analysis, advanced sequencing technologies and novel genomics tools to identify a candidate noncoding mutation to be validated by future functional studies.

Linkage analysis is a fundamental tool to carry out targeted analysis of Mendelian disorders with NGS [96, 100]. Prioritizing the suggestive linkage regions on chromosome 8 and 16 by performing fine mapping linkage analysis in Chapter 1 was instrumental for reducing the number of positional candidate genes and noncoding mutations. In Chapter 4, we performed the first transcriptomic analysis for family CMT720 using patient fibroblasts. Our findings indicated that CMT720 is unlikely to be caused by aberrant transcripts and revealed robust dysregulation in *BCL7C*, *IRX6*, *PRRT2* and *ZNF704*. This provided the functional evidence of pathogenicity we used for prioritizing noncoding variants that may impact gene expression and facilitated identification of a candidate noncoding variant with high potential for pathogenicity. In this study, using srWGS, lrWGS and the complete T2T reference facilitated detecting all forms of noncoding variants in CMT720. Our findings confirmed the utility of combining lrWGS and T2T for improving variant detection across the challenging regions in the genome that remained as gaps in the preceding hg38 reference [141, 338]. This approach substantially increased the number of variants detected compared to an average srWGS alignment [147, 315]. By complementary use of the online PCHi-C data to identify the PIRs of dysregulated positional candidate genes, we selected hg38g.30650690G>A as a potential CRE variant, which localised to a TFBS in a candidate enhancer as predicted by utilising additional epigenomics data during bioinformatic analysis. This SNV remains as a high- ranking candidate associated with the downregulated positional candidate gene *PRRT2*, which is known to cause a broad range of neurological disorders with substantial aetiological similarity to CMT2.

6.1. Linkage analysis for facilitating the interrogation of the noncoding variants

Linkage analysis is a highly effective method for facilitating genetic diagnosis of Mendelian disorders by reducing the number of positional candidates and the candidate variants [75, 96]. In this investigation, we have confirmed suggestive linkage to previously identified loci on chromosome 8 and 16 by performing fine mapping linkage analysis based on an autosomal dominant disease model. Targeting these loci provided the highest filtering efficiency across all variant types identified in CMT720 patient WGS alignments and allowed identification of 4 dysregulated positional candidates among 629 DEGs predicted by transcriptomic analysis of CMT720 patient fibroblasts. The variant filtering power and selection of a small number of dysregulated positional candidates provided by linkage analysis was key to conduct an effective transcriptome guided examination of the noncoding genome. In line with the earlier multi-omics analysis of Mendelian disorders unsolved by screening the exome [229], our findings confirm that combining linkage analysis and transcriptome directed variant selection is a powerful tool for facilitating investigations aiming to identify pathogenic noncoding variants.

The candidate variant hg38g.30650690G>A segregated with all confirmed affected patients however the variant was observed in 3 “at risk” individuals (IV:2, V:1 and V:5) as well as the mildly affected individual V:4. These findings would provide evidence against a fully penetrant dominant mutation causing a severe phenotype, however, the mild clinical course of CMT720 [168] combined with the presence of a very mildly affected individual in this family raises the possibility of reduced penetrance. This is not unprecedented since an increasing number of dominant pathogenic mutations are being recognised in challenging unsolved CMT cases where carriers may show very mild symptoms that remain well below the clinical threshold of detection [193]. For instance, dominant missense mutations in *ITPR3* were reported to cause CMT1 with a broad range of phenotypic severity, where patients may only

display mildly reduced NCVs with no clinical symptoms and show neurological symptoms in as late as the 7th decade of life [193]. This has caused some families with *ITPR3* mutations to remain undiagnosed for over 30 years due to incorrect attribution of affected status in segregation analyses [193]. It is interesting to note that the penetrance of the paroxysmal movement disorders caused by *PRRT2* is 61% [339], which is very similar to the proportion of asymptomatic and mildly affected individuals (64%) carrying hg38g.30650690G>A in CMT720. Accordingly, it is possible for hg38g.30650690G>A to potentially cause a CMT2 phenotype with incomplete penetrance. In the event that a dominant causative mutation cannot be identified in this family, an alternative genetic model could be autosomal pseudodominant inheritance. This has been reported in CMT2 cases caused by the compound heterozygous mutations in *GDAP1* [340] and recessive repeat expansions in *RFC1* in patients with CANVAS [341]. As clinical examination of some members from family CMT720 has not been possible since the initial study published in 2005 [168], it is not known whether the individuals with the "unknown" phenotype are currently displaying symptoms, or how severely is individual V:4 is affected. Clinical re-examination to determine whether individuals carrying hg38g.30650690G>A show symptoms can substantially aid the interpretating the pathogenicity of this candidate variant in CMT720, and allow considerations of alternative modes of inheritance in future studies.

6.2. An exploratory investigation of CMT720 transcriptome to identify candidate noncoding variants

In multi-omics analyses performed on rare diseases, transcriptomics is the most commonly used functional genomics method due to its informativeness for detecting a wide range of coding and noncoding mutations [128, 158, 197]. A major advantage of transcriptomic analysis is that, in most cases, it can reveal molecular abnormalities associated with disease aetiology by using easily accessible patient tissues [158]. In Chapter 4, fibroblasts provided

easily accessible patient-derived cell lines as a substitute for the neuronal tissue to conduct the initial transcriptomic profiling of CMT720. Identification of the 4 dysregulated positional candidate genes implicated gene dysregulation as the most probable disease mechanism for CMT720, while splicing and fusion transcripts are unlikely to underly this condition. This is in agreement with preceding transcriptome guided variant analyses performed in Mendelian disorders, where gene dysregulation is most commonly observed transcriptomic readout that guides the analysis towards the pathogenic variant [158, 208]. Notably, downregulation of the high-ranking positional candidate *PRRT2* has allowed the selection of the hg38g.30650690G>A noncoding variant as a regulatory high-ranking candidate in combination with the PCHi-C data [274]. While transcriptomic analyses have been carried out by previous investigations of CMT to validate or analyse the molecular mechanism of known pathogenic variants [342], this study has highlighted the application of transcriptomics for guiding variant analysis and demonstrated the power of this approach for identifying potentially pathogenic noncoding variants in unsolved CMT cases.

The most significant limitation of this study was the use of patient fibroblasts to model the transcriptome of the neurological phenotype of CMT720. Earlier investigations were able to successfully detect differential gene expression and aberrant splicing caused by known pathogenic variants by using fibroblasts derived from CMT2 patients [209, 210], which demonstrates the utility of this cell type as a substitute for the disease relevant neuronal tissue. Nevertheless, fibroblast display weak or no expression (<1TPM) in 20% of the genes that are associated with neurological phenotypes in OMIM [36], including neuropathy [207], and show substantial differences in alternative splicing [208]. Therefore, in this study, it was not possible to account for potentially overlooking transcriptomic abnormalities in the genes expressed or spliced in a tissue-specific manner in neurons. A growing number of investigations have been reprogramming patient-derived fibroblasts to induced pluripotent stem cells (iPSC), to obtain patient-derived MNs for modelling axonal IPNs [343]. This method could provide a robust

disease-relevant tissue for performing transcriptomic analysis of CMT720 and yield transcriptomic findings that reflect the molecular pathology underlying CMT720 more accurately. Our group has previously demonstrated the power of performing RNA-seq on patient-derived MNs for identifying disease relevant transcriptomic changes by detecting a gene-intergenic fusion transcript causing DHMN1 using patient-derived MNs [131]. On the other hand, producing iPSC-derived MNs often requires lengthy optimisation and culturing processes [344], and was not feasible within the timeframe of the current project. Accordingly, the main priority of the future investigations remains to generate data from disease relevant tissue by reprogramming patient-derived iPSC into MNs and perform a second transcriptomic analysis for family CMT720. This will allow verification of the dysregulated positional candidates obtained in this investigation or direct the analysis towards the newly identified dysregulated positional candidates if the current candidates do not show dysregulation in neurons.

Beyond providing the transcriptomic data for variant prioritisation, RNA-seq analysis identified *PRRT2* as high-ranking dysregulated positional candidate with significant relevance to CMT720. This gene was previously reported in a case of hereditary spastic paraplegia showing clinical symptoms of polyneuropathy [236]. Interestingly, *PRRT2* is also associated with neurological movement disorders such as paroxysmal kinesigenic dyskinesia with seizures (PKD) [234] and ataxia [235]. 80% of the PKD cases are caused by a single nucleotide duplication in *PRRT2* (c.649dupC) leading to haploinsufficiency by forming a premature stop codon [234, 345], implicating that *PRRT2* downregulation observed in the current study can potentially cause a neurological phenotype. While hg38g.30650690G>A is a noncoding variant that may potentially reduce *PRRT2* expression by a regulatory mechanism, it will be important to validate the reduced *PRRT2* expression in CMT720 patient-derived MN RNA and protein, to investigate *PRRT2* as a potential candidate gene susceptible to loss of function in future studies. Association of *PRRT2* with sudden infantile death due to epilepsy

[346] provides further support for this gene as a high-ranking positional candidate. Three individuals in CMT720 were reported to have died in infancy prior to the initial clinical investigation in 2005 (personal communication of clinician Andrzej Kochanski) one of which was due to a seizure at the age of 3 (not included in the pedigree). Accordingly, the clinical reports from the literature on the broad spectrum of *PRRT2*-associated neurological disorders and the infantile deaths with a reported seizure in CMT720 family supports *PRRT2* as a high-ranking potential candidate gene that is worthy of follow up in future investigations.

6.3. Improved sequencing technologies and genomic references for effectively analysing noncoding variants in challenging regions

The combined use of lrWGS, T2T and the DHPG demonstrated that the enhanced coverage provided by lrWGS and T2T, and the representation of genetic diversity contributed by the DHPG substantially facilitates the interrogation of the noncoding genome in genetic studies of this unsolved CMT case. We have demonstrated the power of DHPG as a filtering tool for analysing the noncoding genome in CMT which can eliminate a substantial proportion of variants that are not represented by the 1KGP population reference [142]. Although the *DUSP22* paralog encountered in this study is most likely not expressed [303], intronic SNV calls in this gene demonstrated the improvements in the detection of noncoding variants provided by combining the T2T reference and lrWGS to access the challenging regions that remain gaps in the hg38 reference [142]. Long read WGS resolved the *DUSP22* intronic variants despite the high identity of this paralog (99.9%) to its parent gene [303], whereas most short reads were poorly mapped as reported by previous investigations [141, 347]. Power to discriminate biologically inactive gene copies is an important contribution to variant detection and interpretation, as demonstrated by our pervious investigation on *SORD1* mutations in a cohort of CMT2 [348]. In this study, ONT lrWGS platform identified a pathogenic

coding SNV in the *SORD1* parent gene that could not be accurately detected by using srWGS due to short reads mismapping to the inactive *SORD2P* pseudogene with high sequence identity [348]. Segmental duplications may also harbour expressed copies of pathogenic genes that are associated with CMT, such as *NOTCH2NLC* associated with intermediate CMT in Asian populations [146]. *NOTCH2NLC* localises to a segmental duplication showing 98% sequence identity to 4 other paralogous genes and harbours pathogenic REs in 5'UTR region associated with CMT [146], thus the clinical investigation of this gene have only been made possible by the advent of lrWGS [349, 350]. Interestingly, an extra copy of *NOTCH2NLC* is represented in a haplotype included in the DHPG [144], which demonstrates the need for interrogating the challenging parts of the genome with an adequate representation of genomic diversity to aid variant interpretation. Therefore, these reports support the applicability of variant identification and filtering approaches utilised in this study for interrogating the structurally complex and repetitive parts of the genome implicated in CMT.

On the other hand, utilising the full potential of lrWGS was not possible in this study due to the limited filtering efficiency of SNVs and indels by the unavailability of a second patient lrWGS alignment against T2T. Although an increased number of high quality variants are expected to be identified by lrWGS [240, 243, 253], the total number of variants retained within the suggestive linkage loci on chromosome 8 and 16 following filtering was ~44 fold higher in the patient lrWGS callset compared to the srWGS callset, which cannot be explained by the improved detection due to lrWGS [141, 245]. Considering that both callsets underwent the same filtering pipeline that only differed by selection of the variants common to both patient callsets, lrWGS should be performed on patient IV:9 and the resulting variant callset should be incorporated into the variant filtering pipeline to eliminate the high number of SNVs and indels that are not common to both patients. A second lrWGS alignment will also be effective for eliminating false positives by selecting the SVs that are common in both patients.

Despite the advantages provided by the improved genomics tools in the current study, the scarcity of bioinformatic resources and genome annotations developed for the T2T imposed limitations to interpreting variant pathogenicity. We could effectively perform bioinformatic analysis of most variants using the hg38 annotations, whereas, the *DUSP22* intronic variants within the regions unique to T2T remained as variants of unknown significance. It is currently not possible to reliably determine the frequency of these variants using the available references, since the 1KGP SNV and indel callset is obtained by realigning srWGS data against the T2T reference [142] and would not be expected to contain accurate calls in such challenging regions as observed in this study [141]. The current transition of online genomics resources from the hg38 to the T2T reference is anticipated to occur over years [141], therefore, the vast collection of annotations produced for the hg38 are likely to remain useful.

While the Pangenome can provide a powerful filtering tool, the human pangenome reference is still in a draft state and comprises only 47 individuals, which is currently not capturing the full spectrum of polymorphic diversity in humans [144]. The complete human pangenome is planned to include 700 haplotypes from 350 individuals by the end of 2024 [71, 144], which will provide substantially more power for excluding the benign polymorphic variants in the human population. The most complete coverage of the genome in WGS alignments can currently be achieved only by the using lrWGS and the T2T reference [141], as confirmed by our findings. Therefore, lrWGS, T2T and pangenome references should also be utilised by future investigation of CMT720 while simultaneously incorporating the resources that will be developed for the T2T and pangenome references as well as those available for the hg38 reference to maximise the detection and aid interpretation of noncoding variants.

6.4. Utility of multi-omics for prioritisation and analysis of noncoding variants

Epigenomics is an essential component in multi-omics analyses that aim to select regulatory variants [351, 352]. Identification of the potentially pathogenic regulatory variant hg38g.30650690G>A demonstrated that integration of epigenomics was a very effective way of targeting noncoding mutations with known disease mechanisms implicated in CMT [164]. Similar to the mechanism proposed in this investigation for hg38g.30650690G>A, SNVs in a long- range (~150 kb) enhancer of *SH3TC2* are known to disrupt a conserved SOX10 binding motif, leading to 80% reduction in *SH3TC2* and increased severity of the CMT1 phenotype caused by the coding variants in this gene [164]. A 10.7 kb deletion in a downstream (~40kb) enhancer of *EGR2* that overlaps a highly conserved TFBS for multiple TFs with essential roles in Schwann cell development is associated with congenital CMT1 [353]. These examples demonstrate that integration of epigenomics data to the transcriptome guided analysis has the potential to detect noncoding pathogenic variants in CMT with varying structural complexities. Notably, further utilisation of epigenomics during bioinformatics analysis is important to realise the pathogenic potential of a candidate variant since disruption of conserved TFBS is a common mechanism among pathogenic regulatory variants observed in Mendelian disorders, including CMT [127, 164, 271].

Although PCHi-C data provided a powerful tool for selecting regulatory variants, the promoter regions of both *PRRT2* and *MAZ* were captured in the same bin due to the limited resolution of the PCHi-C data utilised in this study [274]. This raises questions against the specificity of the chromatin contact between the candidate enhancer harbouring hg38g.30650690G>A and the promoter of *PRRT2*. Since the bin size (resolution) in 3C based experiments such as PCHi-C is determined by the restriction enzyme used during library preparation [278, 354], this limitation can be addressed in future studies by validating the 3D regulatory interactions between the candidate enhancer overlapping hg38g.30650690G>A

and the promoter of *PRRT2* by performing high resolution 3C [355] on CMT720 patient MNs. Utilising micrococcal nuclease during library preparation in these 3C experiments can achieve resolutions beyond a single nucleosome (~147bp) [355], which would allow separation of *PRRT2* and *MAZ* promoters into distinct bins, and determining whether the enhancer overlapping hg38g.30650690G>A and the promoter of *PRRT2* form a 3D chromatin contact.

Due to the low resolution of CHiP-seq at determining the canonical TF binding motifs [356, 357] and the degeneracy in motifs that are recognised by TFs [358], predicting the impact of a noncoding variant in a TFBS is challenging [359, 360]. Accordingly, the predicted binding of SP1 and CTCF must be confirmed by functional studies in CMT720 patient neurons to delineate the mechanism of hg38g.30650690G>A. Since disruption of predicted SP1 or CTCF binding by the candidate variant may potentially explain the downregulation in *PRRT2*, capturing all regions of DNA bound to SP1 and CTCF with chromatin immunoprecipitation followed by the targeted PCR amplification (CHiP-PCR) of the candidate enhancer element harbouring hg38g.30650690G>A can allow quantification of TF occupancy at this region [361].

6.5. Future investigations of CMT720

Since *PRRT2* was identified as a promising candidate gene associated with the variant hg38g.30650690G>A, the main priority of the future studies of CMT720 will be to validate these findings. In this study, transcriptomic analysis provided the patient-derived dataset used to direct the variant prioritisation during the multi-omics analysis. Therefore, fibroblasts available from patients IV:4 and IV:9 will be reprogrammed into iPSC, which will then be differentiated into MNs. Patient MNs will be used for performing RNA-seq to obtain the most accurate transcriptomic profile of CMT720 in a disease relevant tissue, which will permit targeting the variants that could not be prioritised due to the tissue-specific differences in

fibroblasts. If downregulation of *PRRT2* is validated by RNA-seq studies, using a quantitative method such as Western blot to recapitulate gene dysregulation at the level of protein expression can support further evidence of pathogenicity, since loss of function in *PRRT2* is a well-established mechanism in neurological movement disorders [345, 362].

To address the limitations in variant filtering observed in this study, lrWGS will be performed on patient IV:9 and the resulting data will be aligned against the T2T reference. Including the SNVs and indels identified in the lrWGS alignment of patient IV:9 in the filtering pipeline, will reduce the number of SNV, indel and SV to a more manageable number. All available population variant references developed for the T2T reference will be included in variant filtering. Additionally, the most comprehensive graph VCF should be utilised during genome inference with PanGenie [260] to maximise the number of normal variants to eliminate during variant filtering. These combined measures are expected to aid the interpretation and detection of noncoding variants across a larger portion of the suggestive linkage loci on chromosome 8 and 16.

Use of epigenomic data in a complementary manner to transcriptomic analysis was essential in the current investigation to select and analyse regulatory variants by revealing the function of the noncoding regions these variants localised to. If *PRRT2* is validated by RNA-seq analysis of patient MNs, the epigenomic findings providing the evidence of pathogenicity for the candidate variant hg38g.30650690G>A should be recapitulated in patient MNs. Micro-C can reveal chromatin interactions below the resolution of a single nucleosome (~147bp) [167] and can accurately determine whether the enhancer element harbouring the candidate variant hg38g.30650690G>A forms 3D chromatin interactions with the promoter of *PRRT2*. SP1 or CTCF binding at this enhancer element can be analysed by CHiP-PCR to support the predicted molecular mechanism of hg38g.30650690G>A. If differential TF binding is detected, changing the nucleotide at this position to the reference allele by the guided CRISPR–Cas9 genome editing in patient MNs can be a robust method for determining the functional impact

of this candidate variant [352]. Restoration of endogenous levels of *PRRT2* expression after CRISPR editing would provide strong evidence for pathogenicity for this candidate variant [127].

If patient MN RNA-seq experiments do not support the dysregulation of *PRRT2*, this candidate gene will be deprioritised, and other transcriptomic abnormalities observed in patient MNs will be incorporated into the same transcriptome guided analysis strategy described in this study. While aberrant gene splicing in the patient fibroblasts was not detected, abnormal splicing in neurons may not have been recapitulated due to potential tissue-specific differences in splicing [208]. During RNA-seq analysis of patient MNs, multiple algorithms should be utilised to detect aberrantly spliced transcripts to accurately detect splicing abnormalities [230]. Any intronic variants in the genes showing aberrant splicing in patient MNs should be prioritised for analysis. Artificial intelligence-based pipelines have been increasingly used in multi- omics analysis for rapidly integrating the findings from all omics datasets [363]. These technologies can be implemented in the future investigations to accelerate the identification of the candidate noncoding variants by integrating the evidence from transcriptomic abnormalities observed in patient MNs with online PCHi-C and CHiP-seq datasets to perform multi-omics analysis on patient lrWGS callsets.

6.6. Conclusion

Although noncoding variants are known to cause CMT2 and other forms of IPN, identifying pathogenic noncoding mutations remains highly challenging. In this project, a strategy was developed to effectively identify, prioritise and analyse all forms of noncoding mutations detected in CMT720 patients, to select a candidate with high potential for pathogenicity. Incorporating linkage analysis in our method provided a targeted area of search for the pathogenic mutation and gene, and facilitated all subsequent analyses. This study

contributed to the molecular characterisation of CMT720 by performing the first transcriptomic analysis of this disease on patient fibroblasts. The findings from our transcriptomic analysis revealed dysregulation of positional candidate genes and suggested that alternative pathomechanisms such as aberrant splicing or gene fusion transcripts were unlikely to contribute to the pathogenesis of CMT720. Utilising lrWGS in combination with the novel T2T reference substantially improved the detection of noncoding variants including those identified in structurally complex and repetitive loci. The DHPG and 1KGP resources were used as collections of common variants to perform highly efficient variant filtering. Multi-omics analysis was performed on all remaining candidates where epigenomics data was used for selecting the noncoding mutations that are most likely to cause gene dysregulation for variant analysis. This strategy has allowed us to identify hg38g.30650690G>A as a regulatory variant that may dysregulate the high-ranking positional candidate *PRRT2* by disrupting a TF binding site in a distal enhancer element. In conclusion, the variant identification and prioritisation strategy outlined in this project is a robust method for identifying noncoding mutations with high potential for pathogenicity, and will facilitate interrogation of the noncoding genome for CMT cases which remain unsolved following exclusion of all coding variants.

References

1. Reilly, M.M., *Classification and diagnosis of the inherited neuropathies.* Ann Indian Acad Neurol, 2009. **12**(2): p. 80-8.

2. Liu, X., et al., *Molecular analysis and clinical diversity of distal hereditary motor neuropathy.* European Journal of Neurology, 2020. **27**(7): p. 1319-1326.

3. Axelrod, F.B. and G. Gold-von Simson, *Hereditary sensory and autonomic neuropathies: types II, III, and IV.* Orphanet J Rare Dis, 2007. **2**: p. 39.

4. Kuhlenbäumer, G., et al., *Clinical features and molecular genetics of hereditary peripheral neuropathies.* J Neurol, 2002. **249**(12): p. 1629-50.

5. Harding, A.E. and P.K. Thomas, *The clinical features of hereditary motor and sensory neuropathy types I and II.* Brain, 1980. **103**(2): p. 259-80.

6. Charcot, J.M., *Sur une forme paticuliere d'atrophie musuculaire progressive souvent familial, debutante par les pieds et les jambes et atteignant plus tard les mains.* Rev. Med Fr, 1886. **6**: p. 97-138.

7. Tooth, H., *The peroneal type of progressive muscular atrophy [Thesis].* London: University of London, 1886.

8. Barreto, L.C., et al., *Epidemiologic Study of Charcot-Marie-Tooth Disease: A Systematic Review.* Neuroepidemiology, 2016. **46**(3): p. 157-65.

9. Braathen, G.J., *Genetic epidemiology of Charcot–Marie–Tooth disease.* Acta Neurologica Scandinavica, 2012. **126**(s193): p. iv-22.

10. Rudnik-Schöneborn, S., M. Auer-Grumbach, and J. Senderek, *Charcot-Marie-Tooth disease and hereditary motor neuropathies – Update 2020.* 2020. **32**(3): p. 207-219.

11. Dyck, P.J. and E.H. Lambert, *Lower Motor and Primary Sensory Neuron Diseases With Peroneal Muscular Atrophy: II. Neurologic, Genetic, and Electrophysiologic Findings in Various Neuronal Degenerations.* Archives of Neurology, 1968. **18**(6): p. 619-625.

12. Pareyson, D., V. Scaioli, and M. Laurà, *Clinical and electrophysiological aspects of Charcot-Marie-Tooth disease.* Neuromolecular Med, 2006. **8**(1-2): p. 3-22.

13. Berciano, J., et al., *Clinico-electrophysiological correlation of extensor digitorum brevis muscle atrophy in children with charcot-marie-tooth disease 1A duplication.* Neuromuscul Disord, 2000. **10**(6): p. 419-24.

14. Dyck, P.J., J.L. Karnes, and E.H. Lambert, *Longitudinal study of neuropathic deficits and nerve conduction abnormalities in hereditary motor and sensory neuropathy type 1.* Neurology, 1989. **39**(10): p. 1302-8.

15. Sahenk, Z., et al., *NT-3 promotes nerve regeneration and sensory improvement in CMT1A mouse models and in patients.* Neurology, 2005. **65**(5): p. 681.

16. Sahenk, Z. and B. Ozes, *Gene therapy to promote regeneration in Charcot-Marie- Tooth disease.* Brain Res, 2020. **1727**: p. 146533.

17. Rossor, A.M., P.J. Tomaselli, and M.M. Reilly, *Recent advances in the genetic neuropathies.* Curr Opin Neurol, 2016. **29**(5): p. 537-48.

18. Pipis, M., et al., *Natural history of Charcot-Marie-Tooth disease type 2A: a large international multicentre study.* Brain, 2020. **143**(12): p. 3589-3602.

19. Feely, S.M., et al., *MFN2 mutations cause severe phenotypes in most patients with CMT2A.* Neurology, 2011. **76**(20): p. 1690-6.

20. Davis, C.J., W.G. Bradley, and R. Madrid, *The peroneal muscular atrophy syndrome: clinical, genetic, electrophysiological and nerve biopsy studies. I. Clinical, genetic and electrophysiological findings and classification.* J Genet Hum, 1978. **26**(4): p. 311-49.

21. Buchthal, F. and F. Behse, *Peroneal muscular atrophy (PMA) and related disorders. I. Clinical manifestations as related to biopsy findings, nerve conduction and electromyography.* Brain, 1977. **100 Pt 1**: p. 41-66.

22. Bradley, W.G., R. Madrid, and C.J.F. Davis, *The peroneal muscular atrophy syndrome: Clinical, genetic, electrophysiological and nerve biopsy studies part 3. Clinical, electrophysiological and pathological correlations.* Journal of the Neurological Sciences, 1977. **32**(1): p. 123-136.

23. *2nd Workshop of the European CMT Consortium: 53rd ENMC International Workshop on Classification and Diagnostic Guidelines for Charcot-Marie-Tooth Type 2 (CMT2- HMSN II) and Distal Hereditary Motor Neuropathy (distal HMN-Spinal CMT) 26-28 September 1997, Naarden, The Netherlands.* Neuromuscul Disord, 1998. **8**(6): p. 426-31.

24. Madrid, R., W.G. Bradley, and C.J.F. Davis, *The peroneal muscular atrophy syndrome: Clinical, genetic, electrophysiological and nerve biopsy studies part 2. Observations on pathological changes in sural nerve biopsies.* Journal of the Neurological Sciences, 1977. **32**(1): p. 91-122.

25. Nicholson, G. and S. Myers, *Intermediate forms of Charcot-Marie-Tooth neuropathy: a review.* Neuromolecular Med, 2006. **8**(1-2): p. 123-30.

26. Murphy, S.M., et al., *Charcot-Marie-Tooth disease: frequency of genetic subtypes and guidelines for genetic testing.* J Neurol Neurosurg Psychiatry, 2012. **83**(7): p. 706-10.

27. Charlotte, F., et al., *Charcot–Marie–Tooth disease in Northern England.* Journal of Neurology, Neurosurgery & Psychiatry, 2012. **83**(5): p. 572.

28. Calvo, J., et al., *Genotype-phenotype correlations in Charcot-Marie-Tooth disease type 2 caused by mitofusin 2 mutations.* Arch Neurol, 2009. **66**(12): p. 1511-6.

29. Bouhouche, A., et al., *Autosomal recessive axonal Charcot-Marie-Tooth disease (ARCMT2): phenotype-genotype correlations in 13 Moroccan families.* Brain, 2007. **130**(Pt 4): p. 1062-75.

30. Pitceathly, R.D.S., et al., *Genetic dysfunction of MT-ATP6 causes axonal Charcot- Marie-Tooth disease.* Neurology, 2012. **79**(11): p. 1145-1154.

31. Fay, A., et al., *A Mitochondrial tRNA Mutation Causes Axonal CMT in a Large Venezuelan Family.* Ann Neurol, 2020. **88**(4): p. 830-842.

32. Kwon, H. and B.-O. Choi, *Analyzing clinical and genetic aspects of axonal Charcot– Marie-Tooth disease.* Journal of Genetic Medicine, 2021. **18**: p. 83-93.

33. Pipis, M., et al., *Next-generation sequencing in Charcot-Marie-Tooth disease: opportunities and challenges.* Nat Rev Neurol, 2019. **15**(11): p. 644-656.

34. Martin, P.B., et al., *Overlapping spectrums: The clinicogenetic commonalities between Charcot-Marie-Tooth and other neurodegenerative diseases.* Brain Research, 2020. **1727**: p. 146532.

35. Kenneth, R., *CMT-Associated Genes: The Definitive Guide* [Internet]. 1st ed. Detroit: Charcot-Marie-Tooth Association. 2021 [27/07/2023]. 109.Available from:https://www.cmtausa.org/understanding-cmt/cmt-associated-genes-the-definitive-guide/

36. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.* Nucleic Acids Res, 2005. **33**(Database issue): p. D514-7.

37. Latour, P., et al., *A major determinant for binding and aminoacylation of tRNA(Ala) in cytoplasmic Alanyl-tRNA synthetase is mutated in dominant axonal Charcot-Marie- Tooth disease.* Am J Hum Genet, 2010. **86**(1): p. 77-82.

38. Lassuthova, P., et al., *Mutations in ATP1A1 Cause Dominant Charcot-Marie-Tooth Type 2.* Am J Hum Genet, 2018. **102**(3): p. 505-514.

39. Rebelo, A.P., et al., *A CADM3 variant causes Charcot-Marie-Tooth disease with marked upper limb involvement.* Brain, 2021. **144**(4): p. 1197-1213.

40. Xu, W.Y., et al., *A nonsense mutation in DHTKD1 causes Charcot-Marie-Tooth disease type 2 in a large Chinese pedigree.* Am J Hum Genet, 2012. **91**(6): p. 1088-94.

41. Fabrizi, G.M., et al., *Two novel mutations in dynamin-2 cause axonal Charcot-Marie- Tooth disease.* Neurology, 2007. **69**(3): p. 291-5.

42. Weedon, M.N., et al., *Exome sequencing identifies a DYNC1H1 mutation in a large pedigree with dominant axonal Charcot-Marie-Tooth disease.* Am J Hum Genet, 2011. **89**(2): p. 308-12.

43. Antonellis, A., et al., *Glycyl tRNA synthetase mutations in Charcot-Marie-Tooth disease type 2D and distal spinal muscular atrophy type V.* Am J Hum Genet, 2003. **72**(5): p. 1293-9.

44. Mendoza-Ferreira, N., et al., *De Novo and Inherited Variants in GBF1 are Associated with Axonal Neuropathy Caused by Golgi Fragmentation.* The American Journal of Human Genetics, 2020. **107**(4): p. 763-777.

45. Cuesta, A., et al., *The gene encoding ganglioside-induced differentiation-associated protein 1 is mutated in axonal Charcot-Marie-Tooth type 4A disease.* Nature Genetics, 2002. **30**(1): p. 22-25.

46. Safka Brozkova, D., et al., *Loss of function mutations in HARS cause a spectrum of inherited peripheral neuropathies.* Brain, 2015. **138**(Pt 8): p. 2161-72.

47. Evgrafov, O.V., et al., *Mutant small heat-shock protein 27 causes axonal Charcot- Marie-Tooth disease and distal hereditary motor neuropathy.* Nat Genet, 2004. **36**(6):

p. 602-6.

48. Tang, B.S., et al., *Small heat-shock protein 22 mutated in autosomal dominant Charcot-Marie-Tooth disease type 2L.* Hum Genet, 2005. **116**(3): p. 222-4.

49. Weterman, M.A., et al., *A frameshift mutation in LRSAM1 is responsible for a dominant hereditary polyneuropathy.* Hum Mol Genet, 2012. **21**(2): p. 358-70.

50. Sullivan, J.M., et al., *Dominant mutations of the Notch ligand Jagged1 cause peripheral neuropathy.* J Clin Invest, 2020. **130**(3): p. 1506-1512.

51. Gonzalez, M., et al., *Exome sequencing identifies a significant variant in methionyl- tRNA synthetase (MARS) in a family with late-onset CMT2.* J Neurol Neurosurg Psychiatry, 2013. **84**(11): p. 1247-9.

52. Züchner, S., et al., *Mutations in the mitochondrial GTPase mitofusin 2 cause Charcot- Marie-Tooth neuropathy type 2A.* Nature Genetics, 2004. **36**(5): p. 449-451.

53. Higuchi, Y., et al., *Mutations in MME cause an autosomal-recessive Charcot-Marie- Tooth disease type 2.* Ann Neurol, 2016. **79**(4): p. 659-72.

54. Albulym, O.M., et al., *MORC2 mutations cause axonal Charcot-Marie-Tooth disease with pyramidal signs.* Ann Neurol, 2016. **79**(3): p. 419-27.

55. Sevilla, T., et al., *Mutations in the MORC2 gene cause axonal Charcot-Marie-Tooth disease.* Brain, 2016. **139**(Pt 1): p. 62-72.

56. Tétreault, M., et al., *Adult-onset painful axonal polyneuropathy caused by a dominant NAGLU mutation.* Brain, 2015. **138**(Pt 6): p. 1477-83.

57. Beijer, D., et al., *Dominant NARS1 mutations causing axonal Charcot–Marie–Tooth disease expand NARS1-associated diseases.* Brain Communications, 2024. **6**(2): p. fcae070.

58. Rebelo, A.P., et al., *Cryptic Amyloidogenic Elements in the 3′ UTRs of Neurofilament Genes Trigger Axonal Neuropathy.* Am J Hum Genet, 2016. **98**(4): p. 597-614.

59. Mersiyanova, I.V., et al., *A new variant of Charcot-Marie-Tooth disease type 2 is probably the result of a mutation in the neurofilament-light gene.* Am J Hum Genet, 2000. **67**(1): p. 37-46.

60. Verhoeven, K., et al., *Mutations in the small GTP-ase late endosomal protein RAB7 cause Charcot-Marie-Tooth type 2B neuropathy.* Am J Hum Genet, 2003. **72**(3): p. 722- 7.

61. Løseth, S., et al., *Late-onset sensory-motor axonal neuropathy, a novel SLC12A6- related phenotype.* Brain, 2023. **146**(3): p. 912-922.

62. Landouré, G., et al., *Mutations in TRPV4 cause Charcot-Marie-Tooth disease type 2C.* Nat Genet, 2010. **42**(2): p. 170-4.

63. McCray, B. and S. Scherer, *Axonal Charcot-Marie-Tooth Disease: from Common Pathogenic Mechanisms to Emerging Treatment Opportunities.* Neurotherapeutics, 2021. **18**.

64. Beijer, D., et al., *Defects in Axonal Transport in Inherited Neuropathies.* J Neuromuscul Dis, 2019. **6**(4): p. 401-419.

65. Irobi, J., et al., *Unraveling the genetics of distal hereditary motor neuronopathies.* Neuromolecular Med, 2006. **8**(1-2): p. 131-46.

66. Moss, K.R. and A. Höke, *Targeting the programmed axon degeneration pathway as a potential therapeutic for Charcot-Marie-Tooth disease.* Brain Research, 2020. **1727**: p. 146539.

67. Cashman, C.R. and A. Höke, *Mechanisms of distal axonal degeneration in peripheral neuropathies.* Neuroscience Letters, 2015. **596**: p. 33-50.

68. Bis-Brewer, D.M., S. Fazal, and S. Züchner, *Genetic modifiers and non-Mendelian aspects of CMT.* Brain Research, 2020. **1726**: p. 146459.

69. Lee, D.C., et al., *Yield of next-generation neuropathy gene panels in axonal neuropathies.* J Peripher Nerv Syst, 2019. **24**(4): p. 324-329.

70. Bis-Brewer, D.M., et al., *A network biology approach to unraveling inherited axonopathies.* Sci Rep, 2019. **9**(1): p. 1692.

71. Parmar, J.M., et al., *Genetics of inherited peripheral neuropathies and the next frontier: looking backwards to progress forwards.* Journal of Neurology, Neurosurgery & Psychiatry, 2024: p. jnnp-2024-333436.

72. Berta, E.-A., et al., *Genetic approaches and pathogenic pathways in the clinical management of Charcot-Marie-Tooth disease.* J Transl Genet Genom, 2022. **6**(3): p. 333-352.

73. Higuchi, Y. and H. Takashima, *Clinical genetics of Charcot-Marie-Tooth disease.* J Hum Genet, 2023. **68**(3): p. 199-214.

74. Stavrou, M., et al., *Charcot–Marie–Tooth neuropathies: Current gene therapy advances and the route toward translation.* Journal of the Peripheral Nervous System, 2023. **28**(2): p. 150-168.

75. Smith, K.R., et al., *Reducing the exome search space for Mendelian diseases using genetic linkage analysis of exome genotypes.* Genome Biology, 2011. **12**(9): p. R85.

76. Teare, M.D. and M.F. Santibañez Koref, *Linkage analysis and the study of Mendelian disease in the era of whole exome and genome sequencing.* Briefings in Functional Genomics, 2014. **13**(5): p. 378-383.

77. Kennerson, M.L., et al., *A new locus for X-linked dominant Charcot-Marie-Tooth disease (CMTX6) is caused by mutations in the pyruvate dehydrogenase kinase isoenzyme 3 (PDK3) gene.* Hum Mol Genet, 2013. **22**(7): p. 1404-16.

78. Collins, F.S., *Positional cloning: Let's not call it reverse anymore.* Nature Genetics, 1992. **1**(1): p. 3-6.

79. Morgan, T.H., *Random Segregation Versus Coupling in Mendelian Inheritance.* Science, 1911. **34**(873): p. 384-384.

80. Dueker, N.D. and M.A. Pericak-Vance, *Analysis of Genetic Linkage Data for Mendelian Traits.* Current Protocols in Human Genetics, 2014. **83**(1): p. 1.4.1-1.4.31.

81. Botstein, D., et al., *Construction of a genetic linkage map in man using restriction fragment length polymorphisms.* Am J Hum Genet, 1980. **32**(3): p. 314-31.

82. Ott, J. and J. Terwilliger, *Handbook of Human Genetic Linkage*. 1994, Baltimore, MD: Johns Hopkins University Press.

83. Morton, N.E., *Sequential tests for the detection of linkage.* Am J Hum Genet, 1955. **7**(3): p. 277-318.

84. Evans, D.M. and L.R. Cardon, *Guidelines for Genotyping in Genomewide Linkage Studies: Single-Nucleotide–Polymorphism Maps Versus Microsatellite Maps.* The American Journal of Human Genetics, 2004. **75**(4): p. 687-692.

85. Payseur, B.A., P. Jing, and R.J. Haasl, *A genomic portrait of human microsatellite variation.* Mol Biol Evol, 2011. **28**(1): p. 303-12.

86. Kruglyak, L., *The use of a genetic map of biallelic markers in linkage studies.* Nature Genetics, 1997. **17**(1): p. 21-24.

87. Leal, S.M., *Genetic maps of microsatellite and single-nucleotide polymorphism markers: are the distances accurate?* Genet Epidemiol, 2003. **24**(4): p. 243-52.

88. Sellick, G.S., et al., *Genomewide linkage searches for Mendelian disease loci can be efficiently conducted using high-density SNP genotyping arrays.* Nucleic Acids Research, 2004. **32**(20): p. e164-e164.

89. Ulgen, A. and W. Li, *Comparing single-nucleotide polymorphism marker-based and microsatellite marker-based linkage analyses.* BMC Genetics, 2005. **6**(1): p. S13.

90. Lathrop, G.M., et al., *Strategies for multilocus linkage analysis in humans.* Proc Natl Acad Sci U S A, 1984. **81**(11): p. 3443-6.

91. Ott, J., *Analysis of human genetic linkage*. 3rd ed. 1991, Baltimore, MD: The Johns Hopkins University Press.

92. Edwards, J.H., *Exclusion mapping.* Journal of Medical Genetics, 1987. **24**(9): p. 539-543.

93. Lander, E. and L. Kruglyak, *Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results.* Nature Genetics, 1995. **11**(3): p. 241-247.

94. Nyholt, D.R., *All LODs are not created equal.* Am J Hum Genet, 2000. **67**(2): p. 282-8.

95. Levinson, D.F., *Power to detect linkage with heterogeneity in samples of small nuclear families.* American Journal of Medical Genetics, 1993. **48**(2): p. 94-102.

96. Ott, J., J. Wang, and S.M. Leal, *Genetic linkage analysis in the age of whole-genome sequencing.* Nat Rev Genet, 2015. **16**(5): p. 275-84.

97. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.

98. Frankish, A., et al., *GENCODE: reference annotation for the human and mouse genomes in 2023.* Nucleic Acids Res, 2023. **51**(D1): p. D942-d949.

99. Wang, X., et al., *Genome-wide linkage scan of a pedigree with familial hypercholesterolemia suggests susceptibility loci on chromosomes 3q25-26 and 21q22.* PLoS One, 2011. **6**(10): p. e24838.

100. Sobreira, N.L.M., et al., *Whole-Genome Sequencing of a Single Proband Together with Linkage Analysis Identifies a Mendelian Disease Gene.* PLOS Genetics, 2010. **6**(6): p. e1000991.

101. Green, E.D., et al., *Charting a course for genomic medicine from base pairs to bedside.* Nature, 2011. **470**(7333): p. 204-213.

102. Bamshad, M.J., D.A. Nickerson, and J.X. Chong, *Mendelian Gene Discovery: Fast and Furious with No End in Sight.* The American Journal of Human Genetics, 2019. **105**(3): p. 448-455.

103. Rossor, A.M., et al., *Clinical implications of genetic advances in Charcot–Marie–Tooth disease.* Nature Reviews Neurology, 2013. **9**(10): p. 562-571.

104. Timmerman, V., A.V. Strickland, and S. Züchner, *Genetics of Charcot-Marie-Tooth (CMT) Disease within the Frame of the Human Genome Project Success.* Genes (Basel), 2014. **5**(1): p. 13-32.

105. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA.* Genomics, 2016. **107**(1): p. 1-8.

106. Logsdon, G.A., M.R. Vollger, and E.E. Eichler, *Long-read human genome sequencing and its applications.* Nature Reviews Genetics, 2020. **21**(10): p. 597-614.

107. Levy, S.E. and B.E. Boone, *Next-Generation Sequencing Strategies.* Cold Spring Harb Perspect Med, 2019. **9**(7).

108. Wang, H. and R. Chen, *Whole-exome sequencing and whole-genome sequencing*, in *Genetics and Genomics of Eye Disease*, X.R. Gao, Editor. 2020, Academic Press. p. 27-39.

109. Wenger, A.M., et al., *Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome.* Nat Biotechnol, 2019. **37**(10): p. 1155-1162.

110. Hu, T., et al., *Next-generation sequencing technologies: An overview.* Human Immunology, 2021. **82**(11): p. 801-811.

111. Jain, M., et al., *Nanopore sequencing and assembly of a human genome with ultra- long reads.* Nature Biotechnology, 2018. **36**(4): p. 338-345.

112. Nikopoulos, K., et al., *Next-Generation Sequencing of a 40 Mb Linkage Interval Reveals TSPAN12 Mutations in Patients with Familial Exudative Vitreoretinopathy.* The American Journal of Human Genetics, 2010. **86**(2): p. 240-247.

113. Satam, H., et al., *Next-Generation Sequencing Technology: Current Trends and Advancements.* Biology (Basel), 2023. **12**(7).

114. Ng, S.B., et al., *Targeted capture and massively parallel sequencing of 12 human exomes.*

Nature, 2009. **461**(7261): p. 272-6.

115. Backman, J.D., et al., *Exome sequencing and analysis of 454,787 UK Biobank participants.* Nature, 2021. **599**(7886): p. 628-634.

116. Smedley, D., et al., *A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease.* Am J Hum Genet, 2016. **99**(3): p. 595-606.

117. Fridman, V., et al., *CMT subtypes and disease burden in patients enrolled in the Inherited Neuropathies Consortium natural history study: a cross-sectional analysis.* Journal of Neurology, Neurosurgery & Psychiatry, 2015. **86**(8): p. 873.

118. Walsh, M., et al., *Diagnostic and cost utility of whole exome sequencing in peripheral neuropathy.* Ann Clin Transl Neurol, 2017. **4**(5): p. 318-325.

119. Cutrupi, A.N., et al., *Structural variations causing inherited peripheral neuropathies: A paradigm for understanding genomic organization, chromatin interactions, and gene dysregulation.* Molecular Genetics & Genomic Medicine, 2018. **6**(3): p. 422-433.

120. Drew, A.P., et al., *A 1.35 Mb DNA fragment is inserted into the DHMN1 locus on chromosome 7q34-q36.2.* Hum Genet, 2016. **135**(11): p. 1269-1278.

121. Kim, Y.G., et al., *Whole-genome sequencing in clinically diagnosed Charcot-Marie- Tooth disease undiagnosed by whole-exome sequencing.* Brain Commun, 2023. **5**(3): p. fcad139.

122. Ewans, L.J., et al., *Whole exome and genome sequencing in mendelian disorders: a diagnostic and health economic analysis.* European Journal of Human Genetics, 2022. **30**(10): p. 1121-1131.

123. van der Sanden, B.P.G.H., et al., *The performance of genome sequencing as a first- tier test for neurodevelopmental disorders.* European Journal of Human Genetics, 2023. **31**(1): p. 81-88.

124. Record, C.J., et al., *Whole genome sequencing increases the diagnostic rate in Charcot-Marie-Tooth disease.* Brain, 2024: p. awae064.

125. Brewer, M.H., et al., *Whole Genome Sequencing Identifies a 78 kb Insertion from Chromosome 8 as the Cause of Charcot-Marie-Tooth Neuropathy CMTX3.* PLOS Genetics, 2016. **12**(7): p. e1006177.

126. DiVincenzo, C., et al., *The allelic spectrum of Charcot-Marie-Tooth disease in over 17,000 individuals with neuropathy.* Mol Genet Genomic Med, 2014. **2**(6): p. 522-9.

127. Ellingford, J.M., et al., *Recommendations for clinical interpretation of variants found in non-coding regions of the genome.* Genome Medicine, 2022. **14**(1): p. 73.

128. Yépez, V.A., et al., *Clinical implementation of RNA sequencing for Mendelian disease diagnostics.* Genome Med, 2022. **14**(1): p. 38.

129. Pantera, H., et al., *Regulation of the neuropathy-associated Pmp22 gene by a distal super-enhancer.* Hum Mol Genet, 2018. **27**(16): p. 2830-2839.

130. Grosz, B.R., et al., *A deep intronic variant in MME causes autosomal recessive Charcot-Marie-Tooth neuropathy through aberrant splicing.* J Peripher Nerv Syst, 2024.

131. Cutrupi, A.N., et al., *Novel gene-intergenic fusion involving ubiquitin E3 ligase UBE3C causes distal hereditary motor neuropathy.* Brain, 2023. **146**(3): p. 880-897.

132. Cassini, T.A., et al., *Whole genome sequencing reveals novel IGHMBP2 variant leading to unique cryptic splice-site and Charcot-Marie-Tooth phenotype with early onset symptoms.* Mol Genet Genomic Med, 2019. **7**(6): p. e00676.

133. Sullivan, R., et al., *RFC1 repeat expansion analysis from whole genome sequencing data simplifies screening and increases diagnostic rates.* medRxiv, 2024: p. 2024.02.28.24303510.

134. Cortese, A., et al., *Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-*

onset ataxia. Nat Genet, 2019. **51**(4): p. 649-658.

135. Escaramís, G., E. Docampo, and R. Rabionet, *A decade of structural variants: description, history and methods to detect structural variation.* Briefings in Functional Genomics, 2015. **14**(5): p. 305-314.

136. Sudmant, P.H., et al., *An integrated map of structural variation in 2,504 human genomes.* Nature, 2015. **526**(7571): p. 75-81.

137. Depienne, C. and J.L. Mandel, *30 years of repeat expansion disorders: What have we learned and what are the remaining challenges?* Am J Hum Genet, 2021. **108**(5): p. 764-785.

138. Pagnamenta, A.T., et al., *An ancestral 10-bp repeat expansion in VWA1 causes recessive hereditary motor neuropathy.* Brain, 2021. **144**(2): p. 584-600.

139. van Blitterswijk, M., et al., *Association between repeat sizes and clinical and pathological characteristics in carriers of C9ORF72 repeat expansions (Xpansize-72): a cross-sectional cohort study.* Lancet Neurol, 2013. **12**(10): p. 978-88.

140. Li, H., et al., *Exome variant discrepancies due to reference-genome differences.* Am J Hum Genet, 2021. **108**(7): p. 1239-1250.

141. Aganezov, S., et al., *A complete reference genome improves analysis of human genetic variation.* Science, 2022. **376**(6588): p. eabl3533.

142. Nurk, S., et al., *The complete sequence of a human genome.* Science, 2022. **376**(6588): p. 44-53.

143. Schneider, V.A., et al., *Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly.* Genome Res, 2017. **27**(5): p. 849-864.

144. Liao, W.-W., et al., *A draft human pangenome reference.* Nature, 2023. **617**(7960): p.312-324.

145. Taylor, D.J., et al., *Beyond the Human Genome Project: The Age of Complete Human Genome Sequences and Pangenome References.* Annu Rev Genomics Hum Genet, 2024.

146. Ando, M., et al., *Clinical phenotypic diversity of NOTCH2NLC-related disease in the largest case series of inherited peripheral neuropathy in Japan.* J Neurol Neurosurg Psychiatry, 2023. **94**(8): p. 622-630.

147. Auton, A., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.

148. Pedersen, B.S., et al., *Effective variant filtering and expected candidate variant yield in studies of rare human disease.* npj Genomic Medicine, 2021. **6**(1): p. 60.

149. Coonrod, E.M., et al., *Clinical analysis of genome next-generation sequencing data using the Omicia platform.* Expert Review of Molecular Diagnostics, 2013. **13**(6): p. 529-540.

150. Austin-Tse, C.A., et al., *Best practices for the interpretation and reporting of clinical whole genome sequencing.* npj Genomic Medicine, 2022. **7**(1): p. 27.

151. Schwarz, J.M., et al., *MutationTaster evaluates disease-causing potential of sequence alterations.* Nature Methods, 2010. **7**(8): p. 575-576.

152. Sefid Dashti, M.J. and J. Gamieldien, *A Practical Guide To Filtering and Prioritizing Genetic Variants.* BioTechniques, 2017. **62**(1): p. 18-30.

153. Carrion-Castillo, A., et al., *Whole-genome sequencing identifies functional noncoding variation in SEMA3C that cosegregates with dyslexia in a multigenerational family.* Human Genetics, 2021. **140**(8): p. 1183-1200.

154. Chakravarti, A., *Genomic contributions to Mendelian disease.* Genome Res, 2011. **21**(5): p. 643-4.

155. Wojcik, M.H., et al., *Beyond the exome: what's next in diagnostic testing for Mendelian*

conditions. ArXiv, 2023.

156. Lee, P.H., et al., *Principles and methods of in-silico prioritization of non-coding regulatory variants.* Human Genetics, 2018. **137**(1): p. 15-30.

157. Wang, Z., et al., *Performance Comparison of Computational Methods for the Prediction of the Function and Pathogenicity of Non-coding Variants.* Genomics, Proteomics & Bioinformatics, 2023. **21**(3): p. 649-661.

158. Murdock, D.R., et al., *Transcriptome-directed analysis for Mendelian disease diagnosis overcomes limitations of conventional genomic testing.* J Clin Invest, 2021. **131**(1).

159. Lunke, S., et al., *Integrated multi-omics for rapid rare disease diagnosis on a national scale.* Nature Medicine, 2023. **29**(7): p. 1681-1691.

160. Hasin, Y., M. Seldin, and A. Lusis, *Multi-omics approaches to disease.* Genome Biology, 2017. **18**(1): p. 83.

161. Colin, E., et al., *Stepwise use of genomics and transcriptomics technologies increases diagnostic yield in Mendelian disorders.* Front Cell Dev Biol, 2023. **11**: p. 1021920.

162. Hafstað, V., et al., *Improved detection of clinically relevant fusion transcripts in cancer by machine learning classification.* BMC Genomics, 2023. **24**(1): p. 783.

163. Wittkopp, P.J. and G. Kalay, *Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence.* Nature Reviews Genetics, 2012. **13**(1): p. 59-69.

164. Brewer, M.H., et al., *Haplotype-specific modulation of a SOX10/CREB response element at the Charcot-Marie-Tooth disease type 4C locus SH3TC2.* Hum Mol Genet, 2014. **23**(19): p. 5171-87.

165. Sun, W., et al., *Altered chromatin topologies caused by balanced chromosomal translocation*

lead to central iris hypoplasia. Nature Communications, 2024. **15**(1): p. 5048.

166. Lee, A.J., et al., *Characterization of altered molecular mechanisms in Parkinson's disease through cell type–resolved multiomics analyses.* Science Advances, 2023. **9**(15): p. eabo2467.

167. Lee, B.H., Z. Wu, and S.K. Rhie, *Characterizing chromatin interactions of regulatory elements and nucleosome positions, using Hi-C, Micro-C, and promoter capture Micro-C.* Epigenetics & Chromatin, 2022. **15**(1): p. 41.

168. Kochanski, A., et al., *Mild early onset axonal Charcot-Marie-Tooth disease not linked to other axonal Charcot-Marie-Tooth loci.* Neurology, 2005. **64**(3): p. 533-5.

169. Scherer, S.S. and K.A. Kleopa, *X-linked Charcot-Marie-Tooth disease.* J Peripher Nerv Syst, 2012. **17 Suppl 3**(0 3): p. 9-13.

170. Vieland, V.J., K. Wang, and J. Huang, *Power to Detect Linkage Based on Multiple Sets of Data in the Presence of Locus Heterogeneity: Comparative Evaluation of Model- Based Linkage Methods for Affected Sib Pair Data.* Human Heredity, 2001. **51**(4): p. 199-208.

171. Pais, L.S., et al., *seqr: A web-based analysis and collaboration tool for rare disease genomics.* Human Mutation, 2022. **43**(6): p. 698-707.

172. Tey, S., et al., *Linkage analysis and whole exome sequencing reveals AHNAK2 as a novel genetic cause for autosomal recessive CMT in a Malaysian family.* neurogenetics, 2019. **20**.

173. Ye, J., et al., *Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.* BMC Bioinformatics, 2012. **13**: p. 134.

174. Apte, A. and S. Daniel, *PCR primer design.* Cold Spring Harb Protoc, 2009. **2009**(3): p. pdb.ip65.

175. Robertson, J.M. and J. Walsh-Weller, *An Introduction to PCR Primer Design and Optimization* 215

of Amplification Reactions, in *Forensic DNA Profiling Protocols*, P.J. Lincoln and J. Thomson, Editors. 1998, Humana Press: Totowa, NJ. p. 121-154.

176. Krueger, K.A., et al., *SNP haplotype mapping in a small ALS family.* PLoS One, 2009. **4**(5): p. e5687.

177. Wiltshire, S., et al., *How useful is the fine-scale mapping of complex trait linkage peaks? Evaluating the impact of additional microsatellite genotyping on the posterior probability of linkage.* Genet Epidemiol, 2005. **28**(1): p. 1-10.

178. Li, X., et al., *Two-Stage Genome-Wide Linkage Scan in Keratoconus Sib Pair Families.* Investigative Ophthalmology & Visual Science, 2006. **47**(9): p. 3791-3795.

179. Gonzalez-Neira, A., et al., *Genomewide high-density SNP linkage analysis of non- BRCA1/2 breast cancer families identifies various candidate regions and has greater power than microsatellite studies.* BMC Genomics, 2007. **8**: p. 299.

180. Park, M.H., et al., *Allelic frequencies and heterozygosities of microsatellite markers covering the whole genome in the Korean.* Journal of Human Genetics, 2008. **53**(3): p. 254-266.

181. Guo, X. and R.C. Elston, *Linkage information content of polymorphic genetic markers.* Hum Hered, 1999. **49**(2): p. 112-8.

182. Brusse, E., et al., *A novel 16p locus associated with BSCL2 hereditary motor neuronopathy: a genetic modifier?* neurogenetics, 2009. **10**(4): p. 289-297.

183. Schüle, R., et al., *Autosomal dominant spastic paraplegia with peripheral neuropathy maps to chr12q23-24.* Neurology, 2009. **72**(22): p. 1893-1898.

184. Senderek, J., et al., *Mutations in a gene encoding a novel SH3/TPR domain protein cause autosomal recessive Charcot-Marie-Tooth type 4C neuropathy.* Am J Hum Genet, 2003. **73**(5): p. 1106-19.

185. Bailey-Wilson, J.E., *Parametric versus nonparametric and two-point versus multipoint: controversies in gene mapping*, in *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*. 2005.

186. Kruglyak, L., et al., *Parametric and nonparametric linkage analysis: a unified multipoint approach.* Am J Hum Genet, 1996. **58**(6): p. 1347-63.

187. Thorvaldsdóttir, H., J.T. Robinson, and J.P. Mesirov, *Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration.* Brief Bioinform, 2013. **14**(2): p. 178-92.

188. Broman, K.W., et al., *Comprehensive human genetic maps: individual and sex-specific variation in recombination.* Am J Hum Genet, 1998. **63**(3): p. 861-9.

189. Yang, Z., et al., *A novel locus on 19q13 associated with autosomal-dominant macular dystrophy in a large Greek family.* J Med Genet, 2006. **43**(12): p. e57.

190. Paluru, P., et al., *New locus for autosomal dominant high myopia maps to the long arm of chromosome 17.* Invest Ophthalmol Vis Sci, 2003. **44**(5): p. 1830-6.

191. Hearne, C.M., S. Ghosh, and J.A. Todd, *Microsatellites for linkage analysis of genetic traits.* Trends Genet, 1992. **8**(8): p. 288-94.

192. Nassar, L.R., et al., *The UCSC Genome Browser database: 2023 update.* Nucleic Acids Research, 2022. **51**(D1): p. D1188-D1195.

193. Beijer, D., et al., *A recurrent missense variant in ITPR3 causes demyelinating Charcot- Marie-Tooth with variable severity.* Brain, 2024: p. awae206.

194. Morales, J., et al., *A joint NCBI and EMBL-EBI transcript set for clinical genomics and research.* Nature, 2022. **604**(7905): p. 310-315.

195. Zimoń, M., et al., *Dominant GDAP1 mutations cause predominantly mild CMT phenotypes.*

Neurology, 2011. **77**(6): p. 540-8.

196. Liu, L., et al., *Biological relevance of computationally predicted pathogenicity of noncoding variants.* Nature Communications, 2019. **10**(1): p. 330.

197. Kerr, K., et al., *A scoping review and proposed workflow for multi-omic rare disease research.* Orphanet Journal of Rare Diseases, 2020. **15**(1): p. 107.

198. Cummings, B.B., et al., *Improving genetic diagnosis in Mendelian disease with transcriptome sequencing.* Science Translational Medicine, 2017. **9**(386): p. eaal5209.

199. Zhang, J., et al., *INTEGRATE: gene fusion discovery using whole genome and transcriptome data.* Genome Res, 2016. **26**(1): p. 108-18.

200. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

201. Kukurba, K.R. and S.B. Montgomery, *RNA Sequencing and Analysis.* Cold Spring Harb Protoc, 2015. **2015**(11): p. 951-69.

202. Marco-Puche, G., et al., *RNA-Seq Perspectives to Improve Clinical Diagnosis.* Frontiers in Genetics, 2019. **10**.

203. Byron, S.A., et al., *Translating RNA sequencing into clinical diagnostics: opportunities and challenges.* Nature Reviews Genetics, 2016. **17**(5): p. 257-271.

204. Van Lent, J., et al., *Advances and challenges in modeling inherited peripheral neuropathies using iPSCs.* Experimental & Molecular Medicine, 2024. **56**(6): p. 1348- 1364.

205. Manisha, J., et al., *Challenges in modelling the Charcot-Marie-Tooth neuropathies for therapy development.* Journal of Neurology, Neurosurgery & Psychiatry, 2019. **90**(1): p. 58.

206. Kisiel, M.A. and A.S. Klar, *Isolation and Culture of Human Dermal Fibroblasts*, in *Skin Tissue Engineering: Methods and Protocols*, S. Böttcher-Haberzeth and T. Biedermann, Editors. 218

2019, Springer New York: New York, NY. p. 71-78.

207. Li, S., et al., *The clinical utility and diagnostic implementation of human subject cell transdifferentiation followed by RNA sequencing.* The American Journal of Human Genetics, 2024. **111**(5): p. 841-862.

208. Wagner, N., et al., *Aberrant splicing prediction across human tissues.* Nature Genetics, 2023. **55**(5): p. 861-870.

209. Stergachis, A.B., et al., *Full-length isoform sequencing for resolving the molecular basis of Charcot-Marie-Tooth 2A.* bioRxiv, 2023.

210. Sancho, P., et al., *Characterization of molecular mechanisms underlying the axonal Charcot–Marie–Tooth neuropathy caused by MORC2 mutations.* Human Molecular Genetics, 2019. **28**(10): p. 1629-1644.

211. Everaert, C., et al., *Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data.* Scientific Reports, 2017. **7**(1): p. 1559.

212. Babraham Bioinformatics, *FastQC: A quality control tool for high throughput sequence data*. 2010, [Online]: from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

213. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-2120.

214. Kim, D., et al., *Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.* Nature Biotechnology, 2019. **37**(8): p. 907-915.

215. Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.* Nature Biotechnology, 2015. **33**(3): p. 290-295.

216. Uhrig, S., et al., *Accurate and efficient detection of gene fusions from RNA sequencing data.* Genome Res, 2021. **31**(3): p. 448-460.

217. McPherson, A., et al., *deFuse: an algorithm for gene fusion discovery in tumor RNA- Seq data.* PLoS Comput Biol, 2011. **7**(5): p. e1001138.

218. Nicorici, D., et al., *FusionCatcher – a tool for finding somatic fusion genes in paired- end RNA-sequencing data.* bioRxiv, 2014: p. 011650.

219. Costa-Silva, J., D. Domingues, and F.M. Lopes, *RNA-Seq differential expression analysis: An extended review and a software tool.* PLoS One, 2017. **12**(12): p. e0190152.

220. Finotello, F. and B. Di Camillo, *Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis.* Briefings in Functional Genomics, 2014. **14**(2): p. 130-142.

221. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12): p. 550.

222. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.

223. Tarazona, S., et al., *Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package.* Nucleic Acids Research, 2015. **43**(21): p. e140-e140.

224. Bullard, J.H., et al., *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.* BMC Bioinformatics, 2010. **11**(1): p. 94.

225. Zhao, Y., et al., *TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient- Derived Models Repository.* J Transl Med, 2021. **19**(1): p. 269.

226. McIntyre, L.M., et al., *RNA-seq: technical variability and sampling.* BMC Genomics, 2011. **12**(1): p. 293.

227. Sun, L., et al., *DiVenn: An Interactive and Integrated Web-Based Visualization Tool for Comparing Gene Lists.* Frontiers in Genetics, 2019. **10**.

228. Livak, K.J. and T.D. Schmittgen, *Analysis of relative gene expression data using real- time quantitative PCR and the 2(-Delta Delta C(T)) Method.* Methods, 2001. **25**(4): p. 402-8.

229. Evrony, G.D., et al., *Integrated genome and transcriptome sequencing identifies a noncoding mutation in the genome replication factor DONSON as the cause of microcephaly-micromelia syndrome.* Genome Res, 2017. **27**(8): p. 1323-1335.

230. Jiang, M., et al., *A comprehensive benchmarking of differential splicing tools for RNA- seq analysis at the event level.* Briefings in Bioinformatics, 2023. **24**(3).

231. Carrara, M., et al., *State-of-the-art fusion-finder algorithms sensitivity and specificity.* Biomed Res Int, 2013. **2013**: p. 340620.

232. Pipis, M., et al., *Post-transcriptional microRNA repression of PMP22 dose in severe Charcot-Marie-Tooth disease type 1.* Brain, 2023. **146**(10): p. 4025-4032.

233. Catela, C., et al., *The Iroquois (Iro/Irx) homeobox genes are conserved Hox targets involved in motor neuron development.* bioRxiv, 2024.

234. Chen, W.J., et al., *Exome sequencing identifies truncating mutations in PRRT2 that cause paroxysmal kinesigenic dyskinesia.* Nat Genet, 2011. **43**(12): p. 1252-5.

235. Labate, A., et al., *Homozygous c.649dupC mutation in PRRT2 worsens the BFIS/PKD phenotype with mental retardation, episodic ataxia, and absences.* Epilepsia, 2012. **53**(12): p. e196-e199.

236. Wang, Z., et al., *Association of an insertion mutation in PRRT2 with hereditary spastic paraplegia accompanied by polyneuropathy.* J Clin Lab Anal, 2021. **35**(6): p. e23772.

237. Hoyt, S.J., et al., *From telomere to telomere: The transcriptional and epigenetic state of human*

repeat elements. Science, 2022. **376**(6588): p. eabk3112.

238. Barber, J.C.K., et al., *16p11.2–p12.2 duplication syndrome; a genomic condition differentiated from euchromatic variation of 16p11.2.* European Journal of Human Genetics, 2013. **21**(2): p. 182-189.

239. Nurk, S., et al., *HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads.* Genome Res, 2020. **30**(9): p. 1291-1305.

240. Dolzhenko, E., et al., *ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data.* Genome Biology, 2020. **21**(1): p. 102.

241. Mills, R.E., et al., *Mapping copy number variation by population-scale genome sequencing.* Nature, 2011. **470**(7332): p. 59-65.

242. Aganezov, S., et al., *Comprehensive analysis of structural variants in breast cancer genomes using single-molecule sequencing.* Genome Res, 2020. **30**(9): p. 1258-1273.

243. Kosugi, S. and C. Terao, *Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data.* Human Genome Variation, 2024. **11**(1): p. 18.

244. Liu, Q., et al., *Interrogating the "unsequenceable" genomic trinucleotide repeat disorders by long-read sequencing.* Genome Medicine, 2017. **9**(1): p. 65.

245. English, A.C., et al., *Truvari: refined structural variant comparison preserves allelic diversity.* Genome Biology, 2022. **23**(1): p. 271.

246. Sedlazeck, F.J., et al., *Accurate detection of complex structural variations using single-molecule sequencing.* Nature Methods, 2018. **15**(6): p. 461-468.

247. Schloissnig, S., et al., *Long-read sequencing and structural variant characterization in 1,019*

samples from the 1000 Genomes Project. bioRxiv, 2024: p. 2024.04.18.590093.

248. Ebbert, M.T.W., et al., *Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight.* Genome Biology, 2019. **20**(1): p. 97.

249. Mitsuhashi, S. and N. Matsumoto, *Long-read sequencing for rare human genetic diseases.* Journal of Human Genetics, 2020. **65**(1): p. 11-19.

250. Oehler, J.B., et al., *The application of long-read sequencing in clinical settings.* Human Genomics, 2023. **17**(1): p. 73.

251. Delahaye, C. and J. Nicolas, *Sequencing DNA with nanopores: Troubles and biases.* PLoS One, 2021. **16**(10): p. e0257521.

252. Abdelwahab, O., F. Belzile, and D. Torkamaneh, *Performance analysis of conventional and AI-based variant callers using short and long reads.* BMC Bioinformatics, 2023. **24**(1): p. 472.

253. ten Berk de Boer, E., et al., *Long-read sequencing and optical mapping generates near T2T assemblies that resolves a centromeric translocation.* Scientific Reports, 2024. **14**(1): p. 9000.

254. Ziaei Jam, H., et al., *A deep population reference panel of tandem repeat variation.* Nature Communications, 2023. **14**(1): p. 6711.

255. Zhao, X., et al., *Expectations and blind spots for structural variation detection from long-read assemblies and short-read genome sequencing technologies.* The American Journal of Human Genetics, 2021. **108**(5): p. 919-928.

256. Cerdán-Vélez, D. and M.L. Tress, *The T2T-CHM13 reference assembly uncovers essential WASH1 and GPRIN2 paralogues.* Bioinformatics Advances, 2024. **4**(1): p. vbae029.

257. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. **29**(1): p. 308-11.

258. Reis, A.L.M., et al., *The landscape of genomic structural variation in Indigenous Australians.*

Nature, 2023. **624**(7992): p. 602-610.

259. Pan, B., et al., *Similarities and differences between variants called with human reference genome HG19 or HG38.* BMC Bioinformatics, 2019. **20**(2): p. 101.

260. Ebler, J., et al., *Pangenome-based genome inference allows efficient and accurate genotyping across a wide spectrum of variant classes.* Nature Genetics, 2022. **54**(4): p. 518-525.

261. Paten, B., et al., *Genome graphs and the evolution of genome inference.* Genome Res, 2017. **27**(5): p. 665-676.

262. Häntze, H. and P. Horton, *Effects of spaced k-mers on alignment-free genotyping.* Bioinformatics, 2023. **39**(Supplement_1): p. i213-i221.

263. Chaisson, M.J.P., et al., *Multi-platform discovery of haplotype-resolved structural variation in human genomes.* Nature Communications, 2019. **10**(1): p. 1784.

264. Rao, S.S., et al., *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, 2014. **159**(7): p. 1665-80.

265. Freire-Pritchett, P., et al., *Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells.* eLife, 2017. **6**: p. e21926.

266. Andersson, R. and A. Sandelin, *Determinants of enhancer and promoter activities of regulatory elements.* Nature Reviews Genetics, 2020. **21**(2): p. 71-87.

267. Pang, B. and M.P. Snyder, *Systematic identification of silencers in human cells.* Nature Genetics, 2020. **52**(3): p. 254-263.

268. Doni Jayavelu, N., et al., *Candidate silencer elements for the human and mouse genomes.* Nature Communications, 2020. **11**(1): p. 1061.

269. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-380.

270. Abascal, F., et al., *Expanded encyclopaedias of DNA elements in the human and mouse genomes.* Nature, 2020. **583**(7818): p. 699-710.

271. Gurnett, C.A., et al., *Two novel point mutations in the long-range SHH enhancer in three families with triphalangeal thumb and preaxial polydactyly.* American Journal of Medical Genetics Part A, 2007. **143A**(1): p. 27-32.

272. Sanyal, A., et al., *The long-range interaction landscape of gene promoters.* Nature, 2012. **489**(7414): p. 109-113.

273. Dunham, I., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

274. Jung, I., et al., *A compendium of promoter-centered long-range chromatin interactions in the human genome.* Nat Genet, 2019. **51**(10): p. 1442-1449.

275. Han, J., Z. Zhang, and K. Wang, *3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering.* Molecular Cytogenetics, 2018. **11**(1): p. 21.

276. Bonev, B. and G. Cavalli, *Organization and function of the 3D genome.* Nature Reviews Genetics, 2016. **17**(11): p. 661-678.

277. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-93.

278. Schoenfelder, S., et al., *Promoter Capture Hi-C: High-resolution, Genome-wide Profiling of Promoter Interactions.* J Vis Exp, 2018(136).

279. Mifsud, B., et al., *Mapping long-range promoter contacts in human cells with high- resolution capture Hi-C.* Nature Genetics, 2015. **47**(6): p. 598-606.

280. Li, G., et al., *Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application.* BMC Genomics, 2014. **15**(12): p. S11.

281. Li, H., *Minimap2: pairwise alignment for nucleotide sequences.* Bioinformatics, 2018. **34**(18): p. 3094-3100.

282. Zheng, Z., et al., *Symphonizing pileup and full-alignment for deep learning-based long- read variant calling.* Nature Computational Science, 2022. **2**(12): p. 797-803.

283. Smolka, M., et al., *Detection of mosaic and population-level structural variants with Sniffles2.* Nature Biotechnology, 2024.

284. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* Gigascience, 2021. **10**(2).

285. Poplin, R., et al., *A universal SNP and small-indel variant caller using deep neural networks.* Nature Biotechnology, 2018. **36**(10): p. 983-987.

286. Danecek, P., et al., *The variant call format and VCFtools.* Bioinformatics, 2011. **27**(15): p. 2156-8.

287. MacDonald, J.R., et al., *The Database of Genomic Variants: a curated collection of structural variation in the human genome.* Nucleic Acids Res, 2014. **42**(Database issue): p. D986-92.

288. Riggs, E.R., et al., *Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen).* Genet Med, 2020. **22**(2): p. 245-257.

289. Church, D.M., et al., *Modernizing reference genome assemblies.* PLoS Biol, 2011. **9**(7): p. e1001091.

290. Wang, Y., et al., *The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions.* Genome Biology, 2018. **19**(1): p. 151.

291. Forrest, A.R.R., et al., *A promoter-level mammalian expression atlas.* Nature, 2014. **507**(7493): p. 462-470.

292. DeStefano, G.M., et al., *Position effect on FGF13 associated with X-linked congenital generalized hypertrichosis.* Proc Natl Acad Sci U S A, 2013. **110**(19): p. 7790-5.

293. Lupiáñez, D.G., et al., *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.* Cell, 2015. **161**(5): p. 1012-1025.

294. Landrum, M.J., et al., *ClinVar: improving access to variant interpretations and supporting evidence.* Nucleic Acids Research, 2018. **46**(D1): p. D1062-D1067.

295. Lappalainen, I., et al., *DbVar and DGVa: public archives for genomic structural variation.* Nucleic Acids Res, 2013. **41**(Database issue): p. D936-41.

296. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.

297. Hammal, F., et al., *ReMap 2022: a database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an integrative analysis of DNA-binding sequencing experiments.* Nucleic Acids Res, 2022. **50**(D1): p. D316-d325.

298. Pratt, H.E., et al., *Factorbook: an updated catalog of transcription factor motifs and candidate regulatory motif sites.* Nucleic Acids Research, 2021. **50**(D1): p. D141-D149.

299. Chen, Y., et al., *Deciphering the exact breakpoints of structural variations using long sequencing reads with DeBreak.* Nature Communications, 2023. **14**(1): p. 283.

300. Liehr, T., *Types of CG-CNVs*, in *Benign & Pathological Chromosomal Imbalances*, T. Liehr, Editor. 2014, Academic Press: Oxford. p. 37-119.

301. Chen, S., et al., *A genomic mutational constraint map using variation in 76,156 human genomes.* Nature, 2024. **625**(7993): p. 92-100.

302. Jaganathan, K., et al., *Predicting splicing from primary sequence with deep learning.* Cell, 2019. **176**(3): p. 535-548. e24.

303. Mélard, P., et al., *Molecular alterations and tumor suppressive function of the DUSP22 (Dual Specificity Phosphatase 22) gene in peripheral T-cell lymphoma subtypes.* Oncotarget, 2016. **7**(42): p. 68734-68748.

304. Tryka, K.A., et al., *NCBI's Database of Genotypes and Phenotypes: dbGaP.* Nucleic Acids Res, 2014. **42**(Database issue): p. D975-9.

305. Taliun, D., et al., *Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program.* Nature, 2021. **590**(7845): p. 290-299.

306. den Dunnen, J.T., et al., *HGVS Recommendations for the Description of Sequence Variants: 2016 Update.* Human Mutation, 2016. **37**(6): p. 564-569.

307. Zhang, Y., et al., *Model-based Analysis of ChIP-Seq (MACS).* Genome Biology, 2008. **9**(9): p. R137.

308. Kharchenko, P.V., M.Y. Tolstorukov, and P.J. Park, *Design and analysis of ChIP-seq experiments for DNA-binding proteins.* Nature Biotechnology, 2008. **26**(12): p. 1351- 1359.

309. Pollard, K.S., et al., *Detection of nonneutral substitution rates on mammalian phylogenies.* Genome Res, 2010. **20**(1): p. 110-21.

310. Vierstra, J., et al., *Global reference mapping of human transcription factor footprints.* Nature, 2020. **583**(7818): p. 729-736.

311. Whitfield, T.W., et al., *Functional analysis of transcription factor binding sites in human promoters.* Genome Biology, 2012. **13**(9): p. R50.

312. Guo, Y., S. Mahony, and D.K. Gifford, *High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints.* PLoS Comput Biol, 2012. **8**(8): p. e1002638.

313. Vorontsov, I.E., et al., *HOCOMOCO in 2024: a rebuild of the curated collection of binding*

models for human and mouse transcription factors. Nucleic Acids Research, 2024. **52**(D1): p. D154-D163.

314. Lee, H. and M.C. Schatz, *Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score.* Bioinformatics, 2012. **28**(16): p. 2097- 105.

315. Collins, R.L., et al., *A structural variation reference for medical and population genetics.* Nature, 2020. **581**(7809): p. 444-451.

316. Espejo Valle-Inclan, J., et al., *A multi-platform reference for somatic structural variation detection.* Cell Genomics, 2022. **2**(6): p. 100139.

317. Kosugi, S., et al., *Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing.* Genome Biology, 2019. **20**(1): p. 117.

318. Tankard, R.M., et al., *Detecting Expansions of Tandem Repeats in Cohorts Sequenced with Short-Read Sequencing Data.* Am J Hum Genet, 2018. **103**(6): p. 858-873.

319. Rajan-Babu, I.-S., et al., *Genome-wide sequencing as a first-tier screening test for short tandem repeat expansions.* Genome Medicine, 2021. **13**(1): p. 126.

320. Ibañez, K., et al., *Whole genome sequencing for the diagnosis of neurological repeat expansion disorders in the UK: a retrospective diagnostic accuracy and prospective clinical validation study.* Lancet Neurol, 2022. **21**(3): p. 234-245.

321. Chiu, R., et al., *Straglr: discovering and genotyping tandem repeat expansions using whole genome long-read sequences.* Genome Biology, 2021. **22**(1): p. 224.

322. Deshane, J., et al., *Sp1 regulates chromatin looping between an intronic enhancer and distal promoter of the human heme oxygenase-1 gene in renal cells.* J Biol Chem, 2010. **285**(22): p. 16476-86.

323. Su, W., et al., *DNA looping between sites for transcriptional activation: self-association of DNA-*

bound Sp1. Genes Dev, 1991. **5**(5): p. 820-6.

324. Letovsky, J. and W.S. Dynan, *Measurement of the binding of transcription factor Sp1 to a single GC box recognition sequence.* Nucleic Acids Research, 1989. **17**(7): p. 2639-2653.

325. Plumitallo, S., et al., *Functional analysis of a novel ENG variant in a patient with hereditary hemorrhagic telangiectasia (HHT) identifies a new Sp1 binding-site.* Gene, 2018. **647**: p. 85-92.

326. Salowsky, R., et al., *Basal transcription activity of the dyskeratosis congenita gene is mediated by Sp1 and Sp3 and a patient mutation in a Sp1 binding site is associated with decreased promoter activity.* Gene, 2002. **293**(1): p. 9-19.

327. Carew, J.A., et al., *Severe Factor VII Deficiency Due to a Mutation Disrupting an Sp1 Binding Site in the Factor VII Promoter.* Blood, 1998. **92**(5): p. 1639-1645.

328. Keith, W.N., et al., *A mutation in a functional Sp1 binding site of the telomerase RNA gene (hTERC) promoter in a patient with Paroxysmal Nocturnal Haemoglobinuria.* BMC Hematology, 2004. **4**(1): p. 3.

329. Kubo, N., et al., *Promoter-proximal CTCF binding promotes distal enhancer-dependent gene activation.* Nature Structural & Molecular Biology, 2021. **28**(2): p. 152-161.

330. Liu, Q., et al., *Disruption of a −35 kb Enhancer Impairs CTCF Binding and MLH1 Expression in Colorectal Cells.* Clinical Cancer Research, 2018. **24**(18): p. 4602-4611.

331. Eaton, S.A., et al., *A network of Krüppel-like Factors (Klfs). Klf8 is repressed by Klf3 and activated by Klf1 in vivo.* J Biol Chem, 2008. **283**(40): p. 26937-47.

332. Pearson, R.C.M., A.P.W. Funnell, and M. Crossley, *The mammalian zinc finger transcription factor Krüppel-like factor 3 (KLF3/BKLF).* IUBMB Life, 2011. **63**(2): p. 86-93.

333. Turner, J. and M. Crossley, *Cloning and characterization of mCtBP2, a co-repressor that*

associates with basic Krüppel-like factor and other mammalian transcriptional regulators. The EMBO Journal, 1998. **17**(17): p. 5129-5140.

334. Chen, Y., et al., *Finding Needles in the Haystack: Strategies for Uncovering Noncoding Regulatory Variants.* Annu Rev Genet, 2023. **57**: p. 201-222.

335. Castro, C.P., A.G. Diehl, and A.P. Boyle, *Challenges in screening for de novo noncoding variants contributing to genetically complex phenotypes.* HGG Adv, 2023. **4**(3): p. 100210.

336. Groza, C., et al., *Pangenome graphs improve the analysis of structural variants in rare genetic diseases.* Nature Communications, 2024. **15**(1): p. 657.

337. Miga, K.H. and T. Wang, *The Need for a Human Pangenome Reference Sequence.* Annual Review of Genomics and Human Genetics, 2021. **22**(Volume 22, 2021): p. 81-102.

338. Vollger, M.R., et al., *Segmental duplications and their variation in a complete human genome.* Science, 2022. **376**(6588): p. eabj6965.

339. van Vliet, R., et al., *PRRT2 phenotypes and penetrance of paroxysmal kinesigenic dyskinesia and infantile convulsions.* Neurology, 2012. **79**(8): p. 777-84.

340. van Paassen, B.W., et al., *Pseudodominant inheritance pattern in a family with CMT2 caused by GDAP1 mutations.* J Peripher Nerv Syst, 2017. **22**(4): p. 464-467.

341. Beijer, D., et al., *RFC1 repeat expansions: A recurrent cause of sensory and autonomic neuropathy with cough and ataxia.* European Journal of Neurology, 2022. **29**(7): p. 2156-2161.

342. Peeters, K., et al., *Charcot–Marie–Tooth disease type 2G redefined by a novel mutation in LRSAM1.* Annals of Neurology, 2016. **80**(6): p. 823-833.

343. Van Lent, J., et al., *Induced pluripotent stem cell-derived motor neurons of CMT type 2 patients reveal progressive mitochondrial dysfunction.* Brain, 2021. **144**(8): p. 2471-2485.

344. Castillo Bautista, C.M. and J. Sterneckert, *Progress and challenges in directing the*

differentiation of human iPSCs into spinal motor neurons. Front Cell Dev Biol, 2022. **10**: p. 1089970.

345. Fruscione, F., et al., *PRRT2 controls neuronal excitability by negatively modulating Na+ channel 1.2/1.6 activity.* Brain, 2018. **141**(4): p. 1000-1016.

346. Labate, A., et al., *Mutations in PRRT2 result in familial infantile seizures with heterogeneous phenotypes including febrile convulsions and probable SUDEP.* Epilepsy Research, 2013. **104**(3): p. 280-284.

347. Hijikata, A., et al., *Exome-wide benchmark of difficult-to-sequence regions using short- read next-generation DNA sequencing.* Nucleic Acids Research, 2024. **52**(1): p. 114-124.

348. Grosz, B.R., et al., *Long read sequencing overcomes challenges in the diagnosis of SORD neuropathy.* Journal of the Peripheral Nervous System, 2022. **27**(2): p. 120- 126.

349. Tian, Y., et al., *Expansion of Human-Specific GGC Repeat in Neuronal Intranuclear Inclusion Disease-Related Disorders.* Am J Hum Genet, 2019. **105**(1): p. 166-176.

350. Sone, J., et al., *Long-read sequencing identifies GGC repeat expansions in NOTCH2NLC associated with neuronal intranuclear inclusion disease.* Nature Genetics, 2019. **51**(8): p. 1215-1221.

351. Weedon, M.N., et al., *Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis.* Nature Genetics, 2014. **46**(1): p. 61-64.

352. Claussnitzer, M., et al., *FTO Obesity Variant Circuitry and Adipocyte Browning in Humans.* New England Journal of Medicine. **373**(10): p. 895-907.

353. Funalot, B., et al., *Homozygous deletion of an EGR2 enhancer in congenital amyelinating neuropathy.* Annals of Neurology, 2012. **71**(5): p. 719-723.

354. Sahlén, P., et al., *Genome-wide mapping of promoter-anchored interactions with close to*

single-enhancer resolution. Genome Biology, 2015. **16**(1): p. 156.

355. Hamley, J.C., et al., *Determining chromatin architecture with Micro Capture-C.* Nature Protocols, 2023. **18**(6): p. 1687-1711.

356. Zhang, Q., et al., *Systematic evaluation of the impact of ChIP-seq read designs on genome coverage, peak identification, and allele-specific binding detection.* BMC Bioinformatics, 2016. **17**(1): p. 96.

357. Hu, M., et al., *On the detection and refinement of transcription factor binding sites using ChIP-Seq data.* Nucleic Acids Research, 2010. **38**(7): p. 2154-2167.

358. Holloway, D.T., M. Kon, and C. DeLisi, *Integrating genomic data to predict transcription factor binding.* Genome Inform, 2005. **16**(1): p. 83-94.

359. Steinhaus, R., P.N. Robinson, and D. Seelow, *FABIAN-variant: predicting the effects of DNA variants on transcription factor binding.* Nucleic Acids Research, 2022. **50**(W1): p. W322-W329.

360. Jayaram, N., D. Usvyat, and A.C. R. Martin, *Evaluating tools for transcription factor binding site prediction.* BMC Bioinformatics, 2016. **17**(1): p. 547.

361. Mukhopadhyay, A., et al., *Chromatin immunoprecipitation (ChIP) coupled to detection by quantitative real-time PCR to study transcription factor binding to DNA in Caenorhabditis elegans.* Nature Protocols, 2008. **3**(4): p. 698-709.

362. Tan, G.-H., et al., *PRRT2 deficiency induces paroxysmal kinesigenic dyskinesia by regulating synaptic transmission in cerebellum.* Cell Research, 2018. **28**(1): p. 90-110.

363. Pammi, M., N. Aghaeepour, and J. Neu, *Multiomics, artificial intelligence, and precision medicine in perinatology.* Pediatric Research, 2023. **93**(2): p. 308-315.