INVESTIGATING MECHANISMS OF THE CMTX3 INSERTION USING INDUCED PLURIPOTENT STEM CELL (iPSC) DERIVED MOTOR NEURONS

Alexandra Jae Boyling

2024

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Faculty of Medicine and Health

University of Sydney

Primary Supervisor: Professor Marina L Kennerson

Auxiliary Supervisor: Dr Gonzalo Perez-Siles

The research reported in this thesis was supported by the award of a Research Training Program scholarship to the PhD Candidate

Statement of Originality

This is to certify that, to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that 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 acknowledged.

Alexandra Boyling

24th February 2024

Authorship Attribution Statement

Chapter 1, Part B of this thesis is a literature review published as:

Boyling, A., Perez-Siles, G. and Kennerson, M.L., 2022. Structural variation at a disease mutation hotspot: Strategies to investigate gene regulation and the 3D genome. Frontiers in Genetics, 13, p.842860. https://doi.org/10.3389/fgene.2022.842860

I reviewed the literature, wrote the drafts of the manuscript, and generated figures and tables. All authors reviewed the manuscript.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Alexandra Boyling

24th February, 2024

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Marina L. Kennerson

28th February, 2024

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Gonzalo Perez-Siles

28th February, 2024

TABLE OF CONTENTS

Statement of Originality..... II

Authorship Attribution Statement..... III

Acknowledgements..... X

Abstract..... XII

Publications..... XIV

Papers..... XIV

Conference Presentations..... XIV

List of Figures..... XVII

List of Tables..... XX

Abbreviations..... XXII

CHAPTER 1 | LITERATURE REVIEW **1**

LITERATURE REVIEW: PART A..... 1

1.1 | Charcot-Marie-Tooth Neuropathy..... 1

1.2 | Classification of CMT..... 2

1.3 | Genetic heterogeneity of CMT..... 2

1.4 | CMTX3..... 3

1.4.1 | Genetics of CMTX3..... 4

1.4.2 | Elucidating the pathomechanism of the CMTX3 insertion..... 6

1.5 | Project Rationale..... 7

1.6 | Hypothesis and Aims..... 8

LITERATURE REVIEW: PART B..... 9

Abstract..... 9

1 | Introduction..... 10

2 | Insertions at the Xq27.1 palindrome cause multiple disease phenotypes..... 12

3 | Structural variation can disrupt long-range regulation and cause genetic disease..... 16

3.1 | Cis-regulatory elements: dictating when, where and how much a gene is expressed. 17

3.2 | The role of the 3D genome in transcriptional regulation..... 19

4 | The disease-associated insertions at Xq27.1 can disrupt long-range regulation..... 21

5 | Difficulties in identifying the candidate gene(s) being dysregulated by the disease-associated insertions..... 25
6 | Importance of using disease-relevant tissue to investigate the disease-associated insertions at Xq27.1..... 28
7 | Cellular reprogramming technology provides access to disease-relevant tissue..... 30
8 | Experimental approaches for solving the pathomechanism of CMTX3..... 32
9 | Conclusion..... 37
References 38

CHAPTER 2 | GENERAL MATERIALS AND METHODS 49

2.1 | General materials..... 49
2.2 | Ethics statement..... 50
2.3 | Tissue Culture..... 50
2.3.1 | iPSC lines..... 50
2.3.2 | iPSC maintenance..... 51
2.3.3 | Differentiation of iPSC into sMN..... 51
2.4 | RNA extraction..... 55
2.5 | General immunofluorescence..... 56
2.6 | NanoString nCounter targeted gene expression analysis..... 57
SUPPLEMENTARY MATERIAL: Chapter 2..... 60

CHAPTER 3 | A NEURONAL MODEL FOR CMTX3 NEUROPATHY..... 62

3.1 | Introduction..... 62
3.2 | Methods..... 64
3.2.1 | Fibroblast maintenance..... 64
3.2.2 | Generation of iPSC..... 65
3.2.2.1 | iPSC reprogramming (FCDI)..... 65
3.2.2.2 | iPSC reprogramming (SCOF)..... 65
3.2.3 | DNA extraction..... 66
3.2.4 | CMTX3 genotyping PCR assay..... 66
3.2.5 | Agarose gel electrophoresis..... 67
3.2.6 | Quantitative real-time polymerase chain reaction (qPCR)..... 67
3.2.7 | NanoString nCounter analysis..... 68
3.2.8 | Immunofluorescence..... 68

3.2.9 | Quantification of MNX1 positive sMN..... 69
3.3 | Results..... 71
3.3.1 | Characterisation of CMTX3 patient-derived iPSC lines..... 71
3.3.1.1 CMTX3 iPSC are karyotypically normal and harbour the pathogenic 78 kb insertion..... 71
3.3.1.2 | CMTX3 iPSC lines express markers of pluripotency..... 72
3.3.2 | iPSC can be differentiated into OLIG2-positive motor neuron progenitors..... 74
3.3.3 | Differentiation of MNP gives rise to MNX1-positive spinal motor neurons..... 75
3.4 | Discussion..... 77
SUPPLEMENTARY MATERIAL: Chapter 3..... 81

CHAPTER 4 | EXPLORING THE PATHOMECHANISM OF THE CMTX3 STRUCTURAL VARIATION..... 83

4.1 | Introduction..... 83
4.1.1 Hypothesis 1: The pathomechanism involves the expression of the ARHGAP39 partial transcript..... 83
4.1.2 | Hypothesis 2: The CMTX3 insertion alters spatiotemporal regulation of nearby genes..... 85
4.2 | Methods..... 87
4.2.1 | RNA Sequencing..... 87
4.2.1.1 | Sample preparation..... 87
4.2.1.2 | RNA-Seq and data processing (Macrogen Oceania)..... 87
4.2.1.3 | Differential gene expression analysis..... 88
4.2.2 | NanoString targeted transcriptome analysis..... 88
4.2.3 | Western blot..... 89
4.2.3.1 | Protein extraction..... 89
4.2.3.2 | Protein quantification..... 89
4.2.3.3 | Sodium dodecylsulphate-polyacrylamide gel electrophoresis (SDS-PAGE)..... 90
4.2.3.4 | Protein transfer..... 90
4.2.3.5 | Antibody incubation and detection..... 90
4.2.3.6 | Western blot quantification..... 91
4.3 | Results..... 92
4.3.1 | Exploratory RNA-Seq analysis of iPSC-derived sMN..... 92
4.3.1.1 | Quality control metrics..... 92

4.3.1.2 Novel fusion transcripts involving the CMTX3 SV were not detected with RNA-Seq..... 92
4.3.1.3 | Arriba fusion transcript prediction..... 93
4.3.1.4 | deFuse fusion transcript prediction..... 95
4.3.1.5 | FusionCatcher fusion transcript prediction..... 95
4.3.1.6 Novel splice variants and novel transcript variants involving the CMTX3 SV were excluded using RNA-Seq..... 95
4.3.1.7 | Exploratory differential gene expression analysis..... 96
4.3.1.8 | Functional enrichment of DEG is not observed..... 99
4.3.2 | NanoString targeted transcriptome analysis..... 101
4.3.2.1 | NanoString assay excludes overexpression of the partially duplicated ARHGAP39 gene in CMTX3 patient samples..... 101
4.3.2.2 | NanoString analysis identifies spatiotemporally specific SOX3 dysregulation during CMTX3 sMN differentiation..... 103
4.3.2.3 | Exploring SOX3 expression throughout sMN differentiation..... 105
4.3.3 | RNA-Seq analysis of neuroepithelial progenitor (NEP) cells..... 106
4.3.4 | Differential expression analysis of CMTX3 patient and control NEP..... 107
4.3.4.1 | DESeq2 differential expression analysis..... 107
4.3.4.2 | edgeR differential expression analysis..... 108
4.3.4.3 | Functional enrichment of the DEG identified in NEP is not observed..... 109
4.3.5 | Preliminary western blot explores SOX3 protein expression..... 110
4.4 | Discussion..... 111
4.4.1 | Novel transcripts involving ARHGAP39 were not identified using RNA-Seq or NanoString..... 112
4.4.2 | FGF13 dysregulation was not observed in CMTX3 patient-derived neural cell types 112
4.4.3 | Transcriptomic analyses identify SOX3 as a high-priority candidate gene for CMTX3..... 114
SUPPLEMENTARY MATERIAL: Chapter 4..... 118

CHAPTER 5 | EXPLORING CHROMATIN INTERACTIONS AT THE CMTX3 LOCUS..... 127

5.1 | Introduction..... 127
5.1.1 | Evolution of 3C-based technologies..... 129
5.2 | Methods..... 133

5.2.1 | Cell culture and sample preparation..... 133
5.2.2 | Proximity ligation library preparation and target enrichment..... 134
5.2.3 | Next generation sequencing of enriched Micro-C libraries..... 135
5.2.4 | Computing resources..... 136
5.2.5 | Generating a CMTX3 patient-specific reference genome..... 136
5.2.5.1 | Creating a custom chromosome X file harbouring the unique CMTX3 SV..... 136
5.2.5.2 | Removing the duplicated sequence from chromosome 8q24.3..... 138
5.2.5.3 | Constructing a complete CMTX3 reference genome incorporating the modified chromosome files..... 140
5.2.5.4 | Editing coordinates of target capture sequences..... 140
5.2.6 | Indexing the reference genomes..... 140
5.2.7 | Analysis of Promoter Capture Micro-C data..... 141
5.3 | Results..... 142
5.3.1 | A custom-made patient-specific reference genome accurately recapitulates the CMTX3 SV..... 142
5.3.2 | Quality control of the Promoter Capture Micro-C Libraries..... 144
5.3.3 | Contact matrices enable visualisation of the effect of the CMTX3 SV upon local chromatin architecture..... 145
5.3.4 | Identification of statistically significant chromatin interactions..... 149
5.3.5 | Does the CMTX3 SV form ectopic chromatin contacts with genes in the CMTX3 locus?..... 151
5.3.6 | Investigating if the CMTX3 SV disturbs chromatin interactions on chromosome X... 154
5.3.7 | Investigating chromatin interactions involving the high-priority candidate gene SOX3..... 155
5.4 | Discussion..... 158
SUPPLEMENTARY MATERIAL: Chapter 5..... 163

CHAPTER 6 | GENERAL DISCUSSION..... 167
6.1 | SOX3 as a candidate gene for CMTX3 neuropathy..... 168
6.2 | Investigating the functional impact of SOX3 dysregulation..... 170
6.3 | Investigating the mechanism of gene dysregulation in CMTX3..... 173
6.4 | The importance of peripheral nerve development in CMT..... 174
6.5 | Conclusion..... 175

REFERENCES **177**

APPENDICES **205**

Appendix A 206

Appendix B 208

Appendix C 210

Appendix D 214

Appendix E 215

Appendix F 216

Appendix G 218

ACKNOWLEDGEMENTS

First, to Professor Marina Kennerson, I am incredibly grateful for the unwavering support and encouragement that you have provided throughout this entire PhD. Your dedication to your students is truly amazing. I have learnt so much from you throughout this process, and I consider myself very lucky to have ended up in the Kennerson lab, under your guidance. Thank you for every bit of technical, academic, and emotional support that you have shown, which has helped to make this PhD such a rewarding experience.

To Dr Gonzalo Perez-Siles, thank you for being a constant source of encouragement to me throughout this candidature, and for the invaluable technical assistance that you provided. Thank you for always bringing a positive and optimistic approach toward this project.

I am very grateful for the friendships I have formed within the Northcott Neuroscience Laboratory. I am truly going to miss all the lunchtime laughs, interesting conversations, and friendly faces. Thank you all for the technical support and words of encouragement that you have individually provided. To Dr Anthony Cutrupi, thank you for your technical assistance with iPSC, and for always answering my many questions with a smile. To Dr Ramesh Narayanan, thank you for being a wonderful friend and mentor. To Mel, thank you for your constant words of encouragement, and always cheering me up with a cat video. To Dr Bianca Grosz, Dora Yasar, Harrison Ross, and Janelle Yeo, thank you for your support and friendship.

I would also like to express my warmest thanks to Dr Georgina Samaha and Dr Hamid Alinejad-Rokny for providing invaluable assistance with the bioinformatics aspects of this project. Georgie, thank you for acquainting me with the world of bioinformatics. Developing this new skill was an incredibly rewarding experience, and I am truly grateful. A special thank you also to my colleagues at

the ANZAC Research Institute, Dr Gabrielle Pennings and Dr Maaike Kockx, for your technical assistance with western blots.

To my incredible family, I cannot possibly thank you enough for the love and support that you have shown me, and for encouraging me to pursue my dreams. Mum, you have been my rock throughout this journey, and this would not have been possible without you. To Dad, Maddy, Jack, Nan, and Pop, I am so grateful for your unwavering support throughout this candidature.

Thank you to my wonderful friends for being my biggest cheerleaders. A special thank you to Kyra, Trầm Hương, Emma, Sammy, and Jade, for all the motivational messages, phone calls, and catch-ups. I am truly grateful for your friendship. Kyra, I'm so glad we got to experience this PhD journey together. You have been such an enormous support for me, and I am forever grateful.

To my partner, Anthony, where do I begin. I am so incredibly thankful for your unconditional support. Thank you for always believing in me, and for reminding me to believe in myself. I could not have done this without you.

Finally, I extend my warmest thanks to all CMTX3 family members who have participated in this research.

This research was supported by the Australian Government Research Training Program scholarship, and a grant from the CMT Association Australia.

ABSTRACT

Charcot-Marie-Tooth neuropathy (CMT) is an incurable genetic condition characterised by the length dependent degeneration of motor and sensory nerves. CMTX3 (OMIM #302802) is a rare, X-linked form of CMT neuropathy. In 2016, our group identified a ~78 kb interchromosomal insertion as the causative mutation for CMTX3. The insertion sequence originating from chromosome 8q24.3 is duplicated and inserted into the CMTX3 locus at chromosome X (Xq26.3-q27.3). This genomic rearrangement occurs within a gene desert at Xq27.1 and therefore does not disrupt the sequence of known genes. Without any clear effect on coding genes, the pathogenic mechanism of this complex mutation is unknown.

Interpreting the functional consequences of pathogenic structural variation (SV) is challenging, as these large, complex mutations may cause disease via a range of different mechanisms. At present, there are two leading hypotheses proposed to explain the pathogenicity of the CMTX3 SV. The first postulates that trisomy of 8q24.3 may cause overexpression of the partially duplicated *ARHGAP39* transcript, which is contained within the insertion sequence. The second suggests the presence of the CMX3 SV causes transcriptional dysregulation of nearby genes. Given the tissue-specific nature of transcriptional regulation, it is essential that disease-relevant cell types are used when investigating this hypothesis. Technical and ethical barriers, however, can prevent access to human neuronal tissues. In this thesis, these challenges were circumvented using patient-derived induced pluripotent stem cells (iPSC) to generate disease-relevant neuronal tissue harbouring the complex SV. CMTX3 patient-derived fibroblasts were successfully reprogrammed into iPSC and subsequently differentiated into caudal neuroepithelial progenitors, motor neuron progenitors, and spinal motor neurons.

Using both targeted and global transcriptomic analyses, dysregulation of *SOX3* was observed in CMTX3 patient-derived iPSC and caudal neuroepithelial progenitors, but not in the later

developmental stages. This provided compelling evidence that the CMTX3 SV is capable of altering the regulation of nearby genes in a manner that is highly tissue-specific. In parallel, these transcriptomic analyses did not support the hypothesis of partial *ARHGAP39* transcript overexpression. Taken together, these results have led to the identification of *SOX3* as a high-priority candidate gene for CMTX3 neuropathy.

Promoter Capture Micro-C, a powerful technique to assess regulatory chromatin interactions, was then used to investigate the potential mechanisms driving the observed gene dysregulation in CMTX3 neuroepithelial progenitors. It was theorised that the CMTX3 SV may form ectopic chromatin interactions with genes localising to the CMTX3 linkage region. Although the results of this preliminary experiment were inconclusive, the study flagged potential interactions of interest that can be addressed in future studies.

The results of this thesis represent significant progress towards elucidating the pathomechanism of CMTX3 neuropathy and has generated an invaluable pre-clinical model that will continue to inform disease biology and guide future development of treatment therapies.

PUBLICATIONS

Papers included in this thesis

Boyling, A., Perez-Siles, G. and Kennerson, M.L., 2022. Structural variation at a disease mutation hotspot: Strategies to investigate gene regulation and the 3D genome. *Frontiers in Genetics*, *13*, p.842860.

Additional papers:

Cutrupi, A.N., Narayanan, R.K., Perez-Siles, G., Grosz, B.R., Lai, K., **Boyling, A**., Ellis, M., Lin, R.C., Neumann, B., Mao, D. and Uesugi, M., 2023. Novel gene–intergenic fusion involving ubiquitin E3 ligase UBE3C causes distal hereditary motor neuropathy. *Brain*, *146*(3), pp.880-897.

Narayanan, R.K., Perez-siles, G., Marzec, K.A., **Boyling, A**., Neumann, B., Menezes, M.P. and Kennerson, M.L., 2023. C. elegans model of riboflavin transporter deficiency (RTD) disorder reveals deficits in synaptic transmission and movement. *Genes & Diseases*.

Conference Presentations

Oral Presentations:

Boyling A, Cutrupi AN, Perez-Siles GA, Kennerson ML. Investigating gene dysregulation in CMTX3 motor neurons. The 59th Australian Society for Medical Research National Scientific Conference 2020, Virtual, November 18-19 (2020)

Boyling A, Cutrupi AN, Nicholson GA, Perez-Siles G, Kennerson ML. Using patient-derived motor neurons to investigate a complex structural variation causing CMT neuropathy. Australasian Society for Stem Cell Research Early Career Researcher Symposium Series NSW, Sydney, Australia, May 14 (2021)

Boyling A, Cutrupi AN, Nicholson GA, Perez-Siles G, Kennerson ML. Investigating the pathogenic mechanism of the CMTX3 structural variation. Peripheral Nerve Society Annual Meeting, Virtual, June 12-13 & 25-27 (2021)

Boyling A, Cutrupi AN, Nicholson GA, Perez-Siles G, Kennerson ML. Assessing long-range gene regulation in CMTX3 using targeted transcriptome profiling. EMBL Australia Postgraduate Symposium, Virtual, November 10-12 (2021)

Boyling A, Cutrupi AN, Nicholson GA, Perez-Siles G, Kennerson ML. Using patient-derived motor neurons to investigate a complex structural variation causing CMT neuropathy. Australasian Society for Stem Cell Research Early Career Researcher Award Symposium, Virtual, March 2 (2022)

Boyling A, Cutrupi AN, Nicholson GA, Vucic S, Perez-Siles G, Kennerson ML. Investigating the impact of the CMTX3 structural variation on local gene regulation during motor neuron differentiation. University of Sydney Faculty of Medicine and Health Higher Degree by Research Conference, Sydney, Australia, July 19-20 (2023)

Poster Presentations:

Boyling A, Cutrupi AN, Nicholson GA, Perez-Siles G, Kennerson ML. Exploring gene dysregulation in CMTX3 neuropathy using patient-derived peripheral nerve tissue. International Society for Stem Cell Research Annual Meeting, Virtual, June 21-25 (2021)

Boyling A, Cutrupi AN, Nicholson GA, Perez-Siles G, Kennerson ML. Investigating the pathomechanism of the intergenic structural variation causing CMTX3 neuropathy. Genetics Society of Australasia Conference, Virtual, October 6-8 (2021)

Boyling A, Cutrupi AN, Nicholson GA, Perez-Siles G, Kennerson ML. Investigating the pathomechanism of CMTX3 neuropathy; how does intergenic structural variation alter local gene expression? Hunter Cell Biology Meeting, Newcastle, Australia, May 16-19 (2022)

Boyling A, Cutrupi AN, Nicholson GA, Perez-Siles G, Kennerson ML. Investigating local gene dysregulation as a potential pathomechanism underlying CMTX3 neuropathy. The 61st Australian Society for Medical Research National Scientific Conference, Sydney, Australia, November 10-11 (2022)

Boyling A, Cutrupi AN, Nicholson GA, Vucic S, Perez-Siles G, Kennerson ML. Investigating the impact of the CMTX3 structural variation on gene regulation in patient neuronal tissue. Peripheral Nerve Society Annual Meeting, Copenhagen, Denmark, June 17-20 (2023)

Boyling A, Cutrupi AN, Nicholson GA, Vucic S, Samaha G, Perez-Siles G, Kennerson ML. Spatiotemporal dysregulation of SOX3 is observed during spinal motor neuron differentiation of CMTX3 patient-derived iPSC. Cellular and Molecular Advances in Neurodegeneration, Queensland, Australia, December 3-4 (2023).

Boyling A, Cutrupi AN, Nicholson GA, Vucic S, Samaha G, Perez-Siles G, Kennerson ML. The CMTX3 structural variation alters the spatiotemporal regulation of the neurodevelopmental transcription factor, SOX3, during spinal motor neuron differentiation. Australasian Neuroscience Society 41st Annual Scientific Meeting, Queensland, Australia, December 4-7 (2023).

Prizes and Awards

Best 'Short Talk' presented by a PhD Student, Australasian Society for Stem Cell Research Early Career Researcher Symposium Series NSW, Sydney, Australia, May 14 (2021)

Johnathon Pembroke Prize, Peripheral Nerve Society Annual Meeting, Virtual, June 12-13 & 25-27 (2021)

Oral Presentations Third Prize, EMBL Australia Postgraduate Symposium, Virtual, November 10-12 (2021)

Best Student Poster Presentation, The 61st Australian Society for Medical Research National Scientific Conference, Sydney, Australia, November 10-11 (2022)

Oral Presentation Winner (Neuroscience category), University of Sydney Faculty of Medicine and Health Higher Degree by Research Conference, Sydney, Australia, July 19-20 (2023)

Grants

Kennerson ML (PI), Perez-Siles G (Co-investigator), **Boyling A** (Co-investigator). Investigating the impact of a structural variation on the 3-dimensional (3D) genome and gene dysregulation causing CMTX3. Charcot Marie Tooth Association Australia, AUD 22,900 (2022)

LIST OF FIGURES

Figure 1.1: Overview of the interchromosomal insertion causing CMTX3 6

Figure 1.2: DNA-sequencing has revealed that the disease-associated insertions contain at least one breakpoint located toward the centre of the 180 bp quasi-palindrome at Xq27.1 15

Figure 1.3: Genetic landscape of the mutational hotspot at chromosome Xq27.1 16

Figure 1.4: The role of 3D genome organisation in long-range gene regulation 20

Figure 1.5: Experimental approach for uncovering the pathogenic mechanism of CMTX3 and other diseases associated with insertions (+/- deletions) at Xq27.1 33

Figure 2.1: Differentiation of iPSC into sMN 52

Figure 2.2: Overview of NanoString nCounter gene expression assay 59

Supplementary Figure 2.1: Brightfield microscope images showing optimal morphology of iPSC 60

Figure 3.1: Custom CellProfiler pipeline for quantifying MNX1-positive sMN..... 70

Figure 3.2: Genetic characterisation of CMTX3 patient-derived iPSC lines 72

Figure 3.3: CMTX3 patient-derived iPSC express markers of pluripotency at the mRNA and protein level 73

Figure 3.4: Generation of OLIG2-positive MNP 74

Figure 3.5: iPSC-derived sMN express markers of motor neuron identity 76

Supplementary Figure 3.1: Immunofluorescence confirms iPSC-derived MNP express OLIG2 81

Supplementary Figure 3.2: Immunofluorescence confirms the expression of differentiation markers in iPSC-derived sMN 82

Figure 4.1: Pathogenic rearrangement at theXq27.1 hotspot can alter the spatiotemporal regulation of nearby genes 86

Figure 4.2: Strategy for designing NanoString probes to target *ARHGAP39* 101

Figure 4.3: NanoString assay shows similar *ARHGAP39* expression between CMTX3 patients and controls. 102
Figure 4.4: Comparison of CMTX3 candidate gene expression between patient and control cells at different stages of sMN differentiation 104
Figure 4.5: *SOX3* expression at multiple timepoints throughout the differentiation of iPSC to sMN 106
Figure 4.6: Principle Component Analysis plot for RNA-Seq data 107
Figure 4.7: Preliminary western blot showing SOX3 expression in CMTX3 and control-derived NEP 111
Supplementary Figure 4.1: *SOX3* normalised counts from a preliminary NanoString nCounter assay on caudal NEP 126
Figure 5.1: Predicted regulatory elements contained within the CMTX3 insertion sequence..... 128
Figure 5.2: Overview of the 3C-based technologies 130
Figure 5.3: Overview of the CMTX3 genomic rearrangement 137
Figure 5.4: Rationale for designing the custom-built CMTX3 patient reference genome 139
Figure 5.5: Overview of the custom-built CMTX3 reference genome 143
Figure 5.6: Quality control metrics for the Promoter Capture Micro-C libraries 145
Figure 5.7: Data from Hi-C based analyses is commonly displayed as a contact matrix 146
Figure 5.8: Overview of the chromatin interaction landscape across a ~2 Mb window within the CMTX3 locus 148
Figure 5.9: Summary of CHiCAGO outputs 150
Figure 5.10: Overview of the *cis* chromatin interactions identified in the region surrounding the CMTX3 SV using the stringent CHiCAGO cutoff score of 5 152
Figure 5.11: Overview of the sub-threshold cis chromatin interactions identified in the region surrounding the CMTX3 insertion using the lenient CHiCAGO cutoff score of 3 153
Figure 5.12: Comparison of patient and control contact matrices at Xq27.1 (hg38) 155
Figure 5.13: Visualisation of pairwise interactions at the *SOX3* bait 157

Supplementary Figure 5.1: Modifications of the location of bait sequences ensures alignment with the custom-built CMTX3 reference genome 164
Supplementary Figure 5.2: Summary of the chromatin interactions identified by CHiCAGO at chromosome Xq27.1 166

LIST OF TABLES

Table 1: Disease phenotypes associated with insertions (+/- deletions) at the Xq27.1 mutational hotspot 14
Table 2.1: General consumables 49
Table 2.2: General equipment 50
Table 2.3: iPSC lines used in this thesis 51
Table 2.4: Composition of Neural Base Medium (NBM) 53
Table 2.5: Antibodies for immunofluorescence experiments 57
Supplementary Table 2.1: List of genes targeted in the custom NanoString nCounter Elements TagSet panel 60
Table 3.1: Primers for the CMTX3 genotyping PCR 67
Table 3.2: Taqman Gene Expression Assays used to assess pluripotency of iPSC 68
Table 4.1: List of fusion transcripts predicted by Arriba on chromosome X and chromosome 8 from CMTX3 patient-derived sMN 94
Table 4.2: Differential expression analysis results for the CMTX3 candidate genes 98
Table 4.3: Functional annotation clusters identified by DAVID enrichment analysis of the DEG identified in sMN 100
Table 4.4: List of DEG identified by DESeq2 analsysis of RNA-seq data obtained from CMTX3 and control NEP 108
Table 4.5: List of DEG identified by edgeR analysis of RNA-Seq data from CMTX3 and control NEP 109
Table 4.6: Functional annotation clusters identified by DAVID enrichment analysis of the DEG identified in NEP 110
Supplementary Table 4.1: Quality control metrics for sMN RNA-Seq libraries 118
Supplementary Table 4.2: Fusion transcripts, predicted by defuse, on chromosome X from CMTX3 patient-derived sMN 118
Supplementary Table 4.3: Fusion transcripts, predicted by defuse, on chromosome 8 from CMTX3 patient and control sMN 119

Supplementary Table 4.4: All fusion transcripts, predicted by FusionCatcher, with breakpoints spanning chromosome X and chromosome 8 within CMTX3 patient-derived sMN 120
Supplementary Table 4.5: Fusion transcripts, predicted by FusionCatcher, on chromosome X and chromosome 8 from CMTX3 patient-derived sMN 122
Supplementary Table 4.6: Novel transcript variants on chromosome 8 and X across CMTX3 and control-derived sMN 123
Supplementary Table 4.7: Novel splice variants at genomic regions of interest on chromosome X and 8 across CMTX3 and control-derived sMN 124
Supplementary Table 4.8: List of predicted DEG in CMTX3 sMN 125
Supplementary Table 4.9: Quality control metrics for NEP RNA-Seq libraries 126
Supplementary Table 5.1: GRCh38/hg38 CMTX3 SV coordinates 163
Supplementary Table 5.2: Quality control of Promoter Capture Micro-C Libraries 165

ABBREVIATIONS

ALK	Anaplastic Lymphoma Receptor Tyrosine Kinase
ARHGAP39	Rho GTPase activating protein 39
BCA	Bicinchoninic acid
BOFS	Branchio-oculo-facial syndrome
BSA	Bovine serum albumin
BSCL2	Berardinelli–Seip congenital lipodystrophy 2
cDNA	Complementary DNA
CHAT	Choline O-Acetyltransferase
ChIP-Seq	Chromatin immunoprecipitation sequencing
CNS	Central nervous system
CNV	Copy number variation
CMT	Charcot-Marie-Tooth neuropathy
CMT1	CMT Type 1
CMT2	CMT Type 2
CMTX1	X-linked Charcot-Marie-Tooth Type 1
CMTX3	X-linked Charcot-Marie-Tooth Type 3
CRE	Cis-regulatory element
Ct	Cycle threshold
CTCF	CCCTC-Binding Factor
DAPI	4’,6-diamidino-2-phenylindole
DEG	Differentially expressed genes
DHMN1	Distal Hereditary Motor Neuropathy Type 1
DNA	Deoxyribonucleic acid
DMEM	Dulbecco's Modified Eagle Medium
DMEM/F-12	Dulbecco's Modified Eagle Medium with Ham's F-12 Nutrient Mixture
DPBS	Dulbecco’s phosphate-buffered saline
$DPBS^{(-Ca\ -Mg)}$	Dulbecco’s phosphate-buffered saline, no calcium, no magnesium
ECL	Enhanced chemiluminescence
EDTA	Ethylenediaminetetraacetic acid
eQTL	expression quantitative trait loci

FC	Fold change
FCDI	FUJIFILM Cellular Dynamics, Inc
FDR	False discovery rate
FGF13	Fibroblast Growth Factor 13
F8	Coagulation factor VIII
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
GJB1	Gab junction beta 1
GRCh37/hg19	Genome Reference Consortium Human Build 37/human genome 19
GRCh38/hg38	Genome Reference Consortium Human Build 38/human genome 38
Hi-C	genome-wide chromosome conformation capture
HRP	Horseradish peroxidase
IgG	Immunoglobin G
IGV	Integrated Genomics Viewer
IPN	Inherited peripheral neuropathies
iPSC	induced pluripotent stem cells
ISL1	Islet-1
LASX	Leica Application Suite X
logFC	log(fold change)
MFN2	Mitofusin 2
Micro-C	Hi-C based methodology using micrococcal nuclease digestion
miRNA	micro ribonucleic acid
MNP	Motor neuron progenitors
MNX1	Motor neuron and pancreas homeobox 1
MNCV	Motor nerve conduction velocity
MPZ	Myelin protein zero
mRNA	messenger RNA
NANOG	Nanog homeobox
NEFL	Neurofilament light chain
NBM	Neural base media
NCC	Neural crest cells
NEP	Neuroepithelial progenitors (caudal-like)
NHEJ	Non-homologous end joining
NMM	Neuronal maturation media
OCT4	Octamer-binding transcription factor 4

OLIG2	Oligodendrocyte transcription factor 2
PCHiC	Promoter capture Hi-C
PCR	Polymerase chain reaction
PMP22	Peripheral myelin protein 22
PNS	Peripheral nervous system
PVDF	Polyvinylidene difluoride
QC	Quality control
RA	Retinoic acid
RIPA	Radioimmunoprecipitation assay
RNA	Ribonucleic acid
RT qPCR	Real-time quantitative polymerase chain reaction
SCOF	Stem Cell and Organoid Facility (Children's Medical Research Institute)
scRNA-Seq	single cell RNA sequencing
sgRNA	single guide RNA
sMN	spinal motor neurons
SNV	Single nucleotide variant
SOX2	SRY-box transcription factor 2
SOX3	SRY-box transcription factor 3
SV	Structural variation
TBS	Tris-buffered saline
TBS-T	Tris buffered saline, 0.1% Tween 20
TAD	Topologically associating domain
TAE	Tris acetate EDTA
TFAP2A	Transcription factor AP-2 alpha
TPM	Transcripts per kilobase million
TUBB3	Tubulin beta 3 class III
T2T-CHM13	Telomere-to-telomere reference genome
UBE3C	Ubiquitin protein ligase E3C
UBE3C-IF	UBE3C intergenic-fusion
WGS	Whole genome sequencing
XL-CGH	X-linked congenital generalized hypertrichosis
XL-HPT	X-linked hypoparathyroidism
3C	Chromosome conformation capture
4C	Chromosome conformation capture on-chip

5C	Chromosome conformation capture carbon copy
2D	2 dimensional
3D	3 dimensional

CHAPTER 1

Literature Review: Part A

1.1 Charcot-Marie-Tooth Neuropathy

Inherited peripheral neuropathies (IPN) are a group of diseases affecting the peripheral nervous system. The most common form of IPN is known as Charcot-Marie-Tooth neuropathy (CMT), which describes a group of genetic disorders characterised by the length-dependent degeneration of both motor and sensory nerves of the peripheral nervous system (PNS). CMT derives its name from the three neurologists who first formally described the disorder in 1886; Jean-Martin Charcot, Pierre Marie, and Howard Henry Tooth [1, 2]. With a prevalence of approximately 1 in 2,500 [3], it is the most common inherited neuromuscular disorder [4]. Patients experience a range of chronic symptoms including progressive distal muscle weakness, distal muscle atrophy, foot deformity (e.g. *pes cavus*, hammertoes) and distal sensory impairment [5, 6]. However, this disease is also known for its high degree of phenotypic variability, with marked differences in symptom severity often observed, including within a single kindred [4]. Less frequently, patients diagnosed with specific subtypes of CMT may experience additional symptoms such as optic atrophy [7, 8], sensorineural hearing loss [9, 10], cognitive impairment [11, 12], and laryngeal and respiratory dysfunction [13]. To date, there is neither a cure for this disease nor an effective pharmacological means of improving symptom severity. Instead, symptoms are often managed with rehabilitative therapy, corrective surgeries, and where necessary, pain medication [14, 15]. Ankle foot orthoses can improve mobility and stability [16], although in severe cases, the use of a wheelchair may be required.

1.2 Classification of CMT

Historically, CMT has been broadly classified into two major groups; CMT1 and CMT2 [17, 18]. These classifications reflect the primary cellular pathology underlying the disease. CMT1 (or 'demyelinating CMT') arises from defects within the myelin sheath, whereas CMT2 (or 'axonal CMT') is caused by damage to the axon itself [19]. The two types of CMT can be differentiated in the clinic via electrophysiological examination. CMT1 is characterised by markedly reduced motor nerve conduction velocities (MNCVs) (i.e. below 38 m/s), whereas patients with CMT2 exhibit near-to-normal MNCVs (i.e. above 38 m/s) [20]. Later, an additional type of CMT, intermediate CMT, was formally recognised [21]. As its name suggests, intermediate CMT displays features that overlap CMT1 and CMT2. MNCVs typically fall between 25-45 m/s [21], while histological analysis of peripheral nerve biopsies indicates both axonal and demyelinating pathologies [22].

As the understanding of CMT genetics improved, this classification system was continuously expanded to partition CMT into further subtypes with specific inheritance patterns and causative mutations. Due to the high genetic heterogeneity of CMT, this resulted in a complex classification system, that has prompted recommendations for a simplified nosology [23, 24].

1.3 Genetic heterogeneity of CMT

CMT is highly genetically heterogeneous, with an initial review on the history of CMT genetics describing more than 1,000 causative mutations across 80 discrete genes [25]. Currently, mutations in over 100 genes are known to cause CMT and related neuropathies [26, 27]. Many of these genes encode for proteins that perform critical functions within the peripheral nerve. This includes MPZ, the most abundant protein within the myelin-sheath, and NEFL, a component of the neuronal cytoskeleton, among others. Intriguingly, many of the mutations reported occur in ubiquitously expressed genes that can also give rise to a CMT-specific phenotype [25]. For example, pathogenic mutations have been identified in multiple aminoacyl-tRNA synthetase genes [28-33], as well as the

mitochondrial gene, *MT-ATP6* [34]. The multiple genes causing different forms of CMT represents a significant challenge for the development of effective therapeutics.

CMT1A is the most common form of CMT type 1 disease, accounting for ~40 % of CMT cases [6]. It is caused by the duplication of a ~1.4 Mb region on chromosome 17 containing the *PMP22* gene [35-39]. Mutations in the *GJB1* gene cause CMTX1, the second most common form of CMT [40], which accounts for ~7 to 18% of CMT [41]. This is followed by CMT1B and CMT2A, arising from mutations in *MPZ* [42] and *MFN2* [43], respectively. Together, mutations involving these four genes account for ~90% of genetically diagnosed CMT [44]. However, despite remarkable advancements in genetic sequencing technologies, ~30-40% of patients with CMT lack a genetic diagnosis [44-46].

Genetic testing services that are often used to diagnose CMT include individual gene testing, targeted gene panels, and whole exome sequencing [47, 48]; all of which focus largely upon coding DNA sequences. However, mutations within non-coding DNA, and structural variants (i.e. mutations affecting > 50 bp of DNA), are likely to account for a proportion of unsolved CMT, prompting recommendations to incorporate whole genome sequencing (WGS) into the diagnostic pipeline [26]. Using WGS coupled with the complete human genome reference (T2T-CHM13) released in 2022 [49], and the recent release of the draft human pangenome reference assembly [50], this will provide improved coverage across the genome to detect all classes of DNA variation as well as the genetic diversity for detecting pathogenic variants in the context of personalised medicine.

1.4 CMTX3

CMTX3 (OMIM #302802) is a rare, X-linked recessive form of CMT neuropathy. Clinically, CMTX3 presents as an early-onset disease, with symptoms often visible between birth and 5 years of age [51]. Affected males experience a moderate-to-severe peripheral neuropathy beginning in the lower limbs, with phenotypic variability observed between related individuals [52]. Female carriers are typically asymptomatic, although detailed clinical examination has revealed subtle signs of neuropathy

[52]. CMTX3 has been reported as an axonal form of CMT [53-55]. Although, other studies have observed a range of MNCVs consistent with an intermediate [52] or demyelinating [51] neuropathy.

1.4.1 Genetics of CMTX3

The CMTX3 locus was first reported in 1991 with the identification of 2 American CMT families displaying linkage to a ~31.2 Mb region on the long arm of chromosome X (Xq26-q28) [56]. In addition to a typical peripheral neuropathy phenotype, these patients also displayed pyramidal signs including spastic paraparesis and a positive Babinski sign [56]. In 2006, the CMTX3 locus was substantially refined to 5.7 Mb (Xq26.3-q27.1) after the identification of an additional family (CMT623) showing linkage to the same chromosomal region [52]. This large kindred, with family members in both the United Kingdom and New Zealand, showed some clinical differences from the American families, including a lack of central nervous system (CNS) involvement [52]. Similarly, CNS abnormalities were not observed in another CMTX3 family, CMT193-ext, a multi-generation Australian family sharing a common haplotype with CMT623 [57] and has since been shown to be directly related through identity-by-descent (IBD) analysis [58].

Mutation scanning methods were subsequently employed to identify the pathogenic mutation within the refined disease locus. This analysis excluded coding mutations in known genes located within the CMTX3 interval [52, 53, 57, 59], suggesting that the causative mutation was likely to reside within non-coding DNA [53]. During this time, one of the original American CMTX3 families was re-analysed using whole exome sequencing [60]. Surprisingly, this identified a known pathogenic mutation in the *BSCL2* gene located on chromosome 11 as the cause of disease in this family [60]. This missense mutation (p.N88S) can manifest clinically as a broad range of related phenotypes including distal hereditary motor neuropathy (DHMN), CMT, and hereditary spastic paraparesis [61], with the latter explaining the CNS involvement observed for this family [60]. Incomplete penetrance of this mutation has repeatedly been observed [61, 62], which may account for the apparent X-linked inheritance pattern initially reported in this small family [60].

In 2016, WGS was utilised to examine the CMTX3 linkage region for non-coding variants and structural variation (SV) mutations. This study identified a large inter-chromosomal insertion occurring within the CMTX3 locus that segregated fully with disease [63]. CMTX3 male patients and unaffected female carriers showed a ~78 kb insertion at chromosome Xq27.1, of DNA originating from chromosome 8q24.3 [der(X)dir ins(X;8) (q27.1;q24.3)] [63] (Figure 1.1A). The inserted sequence contained a portion of the *ARHGAP39* gene, including the first 7 of 12 exons that comprise transcript variant 1, as well as ~9.27 kb of upstream non-coding sequence [63] (Figure 1.1B). This genomic re-arrangement caused 8q24.3 trisomy. The mutation localised within a gene desert of the CMTX3 locus, with the closest neighbouring genes, *LOC389895* and *SOX3*, located ~329 kb and ~82 kb away from the proximal and distal re-arrangement breakpoints respectively [63] (Figure 1.1A). Interestingly, the insertion breakpoint localised to the centre of a 180 bp quasi-palindromic sequence, which is a known hotspot for complex genomic rearrangements [63]. A range of complex insertions occurring at this hotspot have previously been associated with multiple hereditary diseases, and are reviewed in Chapter 1, Part B [64]. Of note, an additional genetic disease caused by a complex SV at Xq27.1 (X-linked split-hand/foot malformation) was reported in 2023 [65], after the publication of the Boyling et al. (2022) review article [64].

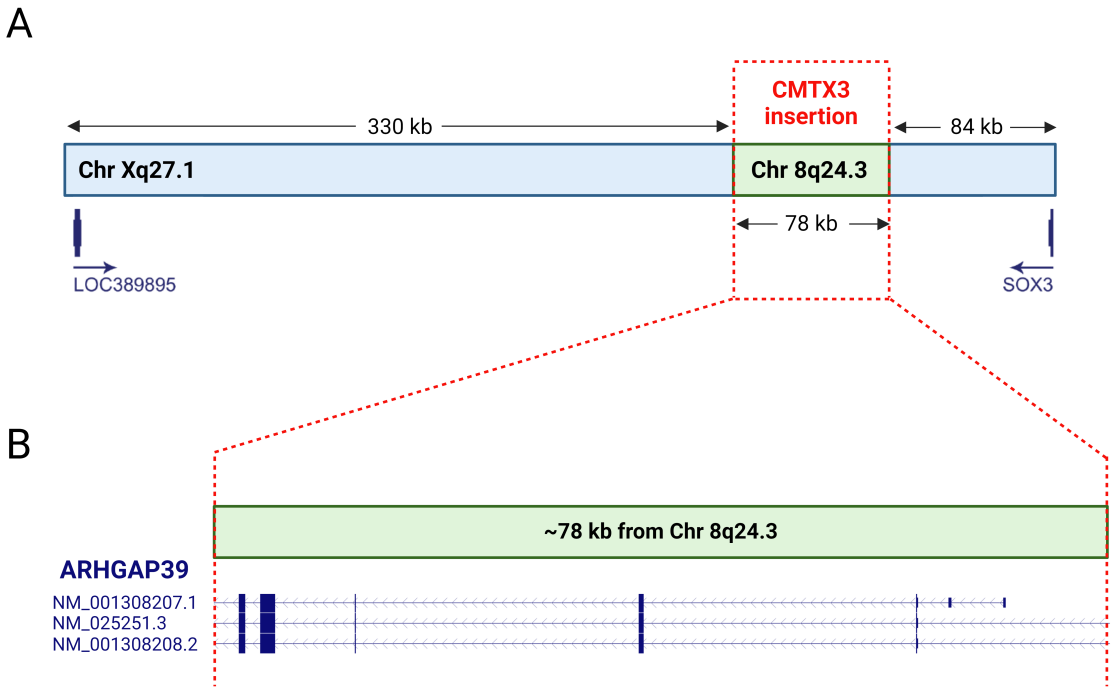


Figure 1.1: Overview of the interchromosomal insertion causing CMTX3. (A) A region of the CMTX3 Xq27.1 linkage region (blue) is shown. CMTX3 patients and unaffected female carriers show an insertion of ~78 kb of DNA from chromosome 8q24.3 (green). This insertion resides in a gene desert of the CMTX3 linkage region and is flanked by the *LOC389895* and *SOX3* genes. This figure has been adapted from Brewer et al., (2016) [63]. **(B)** Close-up schematic of the CMTX3 insertion. The ~78 kb fragment contains partial transcripts of *ARHGAP39*, which is transcribed from the negative strand. This includes the first 7 of 12 exons of the NM_001308207.1 transcript variant, as well as ~9.27 kb of upstream sequence. The insertion fragment terminates within intron 7 of *ARHGAP39*. Figure created using BioRender.com.

1.4.2 Elucidating the pathomechanism of the CMTX3 insertion

Despite an understanding of the pathogenic mutation at the DNA sequence level, the functional consequences of this genomic rearrangement remain poorly understood. Two hypotheses have been proposed to explain how the complex CMTX3 insertion causes peripheral neuropathy [63]. The first proposes that trisomy of 8q24.3 may cause overexpression of the partially duplicated *ARHGAP39* transcript [63]. The second suggests that the CMX3 insertion causes transcriptional dysregulation of nearby gene(s) [63].

Previous research has provided support for the second hypothesis. Gene expression analysis conducted on CMTX3 patient lymphoblasts revealed a ~3-fold upregulation of the nearby gene, *FGF13*, relative to controls [63]. However, it is important to test whether this dysregulation is recapitulated in disease-relevant cell types, or whether it is a benign bystander effect of the CMTX3 SV [63]. In contrast, patient and control samples showed similar expression of *ARGHAP39*, suggesting that overexpression of the partial transcript was unlikely to be the pathomechanism underlying CMTX3 [63]. However, this conclusion should also be validated in disease-relevant neuronal cell types. During this candidature, recent evidence emerged which reinforces the need to carefully consider whether the partially duplicated transcript may have a potential role in the disease pathomechanism. Distal hereditary motor neuropathy type 1 (DHMN1), an IPN affecting the peripheral motor neurons, is caused by a complex 1.35 Mb SV, leading to the duplication of five full length genes, and one partial gene [66]. The partially duplicated gene, *UBE3C*, forms a novel gene-intergenic fusion transcript in DHMN1 patient motor neurons [67]. This has led to speculation of whether a similar mechanism involving the partially duplicated *ARHGAP39* may occur in CMTX3 [68].

Taken together, there is evidence to suggest that the CMTX3 SV can alter the regulation of nearby genes, however, other disease mechanisms cannot be discounted.

1.5 Project Rationale

The goal of the current project is to experimentally investigate the complex CMTX3 SV for potential pathogenic mechanisms. This will require investigating the impact of the CMTX3 SV upon the transcriptional regulation of nearby genes within the CMTX3 linkage region. Given the spatiotemporally specific nature of gene regulation, it is essential that these investigations are performed using disease-relevant cell types. Induced pluripotent stem cells (iPSC) represent an attractive tool for gaining access to human neuronal tissue whilst preserving the genetic background of the CMTX3 complex SV. Gaining insight into the molecular pathogenic mechanisms driving CMTX3 neuropathy, using appropriate pre-clinical models, will inform the development of therapeutic

strategies to treat this debilitating condition, and provide a deeper understanding of how complex, intergenic SV can cause disease.

1.6 Hypothesis and Aims

Hypothesis:

The CMTX3 complex insertion causes neuropathy by dysregulating the transcription of one or more genes located within the CMTX3 locus at chromosome Xq27.1.

Overall aim:

To investigate the pathogenic mechanism of the CMTX3 complex insertion in disease-relevant cell types.

Specific aims:

- Use iPSC technology to develop a disease-relevant neuronal model for CMTX3 neuropathy.
- Perform gene expression analyses on CMTX3 patient and control neuronal tissue to characterise the impact of the pathogenic mutation on global and local transcriptional regulation.
- Investigate possible epigenetic mechanisms which may lead to the local gene dysregulation observed in CMTX3, by exploring long-range chromatin interactions across the CMTX3 locus.

CHAPTER 1

Literature Review: Part B

This section **(Literature Review: Part B)** has been published as: **Boyling, A**., Perez-Siles, G. and Kennerson, M.L., 2022. Structural variation at a disease mutation hotspot: Strategies to investigate gene regulation and the 3D genome. Frontiers in Genetics, 13, p.842860. (https://doi.org/10.3389/fgene.2022.842860)

Abstract

A rare form of X-linked Charcot-Marie-Tooth neuropathy, CMTX3, is caused by an interchromosomal insertion occurring at chromosome Xq27.1. Interestingly, eight other disease phenotypes have been associated with insertions (or insertion-deletions) occurring at the same genetic locus. To date, the pathogenic mechanism underlying most of these diseases remains unsolved, although local gene dysregulation has clearly been implicated in at least two phenotypes. The challenges of accessing disease-relevant tissue and modelling these complex genomic rearrangements has led to this research impasse. We argue that recent technological advancements can overcome many of these challenges, particularly induced pluripotent stem cells (iPSC) and their capacity to provide access to patient-derived disease-relevant tissue. However, to date these valuable tools have not been utilized to investigate the disease-associated insertions at chromosome Xq27.1. Therefore, using CMTX3 as a reference disease, we propose an experimental approach that can be

used to explore these complex mutations, as well as similar structural variants located elsewhere in the genome. The mutational hotspot at Xq27.1 is a valuable disease paradigm with the potential to improve our understanding of the pathogenic consequences of complex structural variation, and more broadly, refine our knowledge of the multifaceted process of long-range gene regulation. Intergenic structural variation is a critically understudied class of mutation, although it is likely to contribute significantly to unsolved genetic disease.

1. Introduction

Charcot-Marie-Tooth (CMT) neuropathy refers to a group of inherited conditions characterized by the length-dependent degeneration of both motor and sensory neurons within the peripheral nervous system. In the absence of a cure for this debilitating condition, individuals with CMT present with varying degrees of distal muscle wasting and sensory impairment, leading to chronic disability (Timmerman et al., 2014). This disease is highly genetically heterogeneous, with more than 1000 causative mutations reported in over 80 genes (Timmerman et al., 2014). These mutations range from single nucleotide changes to complex copy number variations, however, to date there has only been one interchromosomal insertion associated with CMT. A rare X-linked form of CMT, CMTX3, is caused by the insertion of 78 kb of DNA originating from chromosome 8q24.3 into an intergenic region of the CMTX3 locus on chromosome Xq27.1 (Brewer et al., 2016).

Interestingly, the CMTX3 insertion occurs within a quasi-palindromic sequence on chromosome Xq27.1 which is a recurring hotspot for genomic rearrangements. To date, eight additional disease phenotypes have been associated with complex insertions, or insertion-deletions, occurring at this specific locus (Bowl et al., 2005; Zhu et al., 2011; Chou, 2012; DeStefano et al., 2013; Bunyan et al., 2014; Haines et al., 2015; Taylor et al., 2015; Middelkamp et al., 2019; Si et al., 2019; Boschann et al., 2022). For simplicity, we will collectively refer to these mutations as ‘the disease-

associated insertions at Xq27.1' with Table 1 providing additional information regarding the complexity of these genomic rearrangements. This genomic hotspot resides within a gene desert, with approximately 82 kb of non-coding DNA separating the palindrome from the nearest coding gene (Zhu et al., 2011). Most researchers reporting these novel rearrangements hypothesize that they cause disease by disrupting the regulation of nearby genes (Zhu et al., 2011; Haines et al., 2015; Brewer et al., 2016; Middelkamp et al., 2019; Si et al., 2019; Boschann et al., 2022). However, it has proven very challenging to identify the specific gene(s) being dysregulated by the various insertions, and therefore, the pathogenic mechanisms underlying many of these diseases remain unsolved.

This challenge has largely arisen from our incomplete understanding of the non-coding genome, which has historically been overlooked and considered "junk DNA". Efforts of The ENCODE Project have provided fundamental resources to highlight the role of non-coding DNA in transcriptional regulation (ENCODE Project Consortium, 2012), however, there is still much to learn about the many complex processes controlling gene expression profiles at a cell-specific level. Technical limitations have also prohibited researchers from effectively investigating the disease-associated insertions at Xq27.1, including limited access to relevant patient tissue and the difficulties of cloning large genomic rearrangements. Tissue specific gene dysregulation has been the hypothesized pathomechanism driving these diseases, and therefore, it is essential to use the appropriate tissue types when investigating each condition. For CMTX3, the challenge is gaining access to spinal cord and peripheral nerve tissues which are affected in patients (Brewer et al., 2016). Although the disease-associated insertions at Xq27.1 have been well-characterized at the sequence level, traditional cloning techniques cannot be used to generate cellular or animal models of disease. Strategies are therefore needed to generate relevant tissues whilst retaining patient-specific genomic rearrangements.

Since the discovery of the first disease phenotype caused by a complex insertion at Xq27.1 (Bowl et al., 2005), various advancements in technology have provided effective tools for overcoming many of these challenges. Perhaps most notably, improvements in cellular reprogramming

technologies have made it possible to access patient-derived, disease-relevant tissue. Furthermore, improvements in transcriptome profiling and epigenome analyses can now facilitate a comprehensive interrogation of transcriptional organization and regulation in the appropriate tissue. To date, these valuable tools have not been utilized to investigate the different phenotypes associated with the insertions at Xq27.1.

With this in mind, the purpose of this review is to summarize the various phenotypes associated with complex rearrangements at Xq27.1 and highlight where recent technological advancements can be applied to gain insight into their underlying pathomechanisms. Using CMTX3 as the reference disease model, we propose an experimental approach that incorporates permutations of next-generation sequencing and induced pluripotent stem cell (iPSC) technologies to investigate how intergenic genomic rearrangements can produce disease. Understanding the contribution of intergenic structural variation to genetic disease is a growing area of interest, and the mutational hotspot at Xq27.1 provides a unique paradigm for increasing our knowledge of the complex, multi-faceted process of gene regulation.

2. Insertions at the Xq27.1 palindrome cause multiple disease phenotypes

Along with CMTX3, various other diseases are associated with complex insertions occurring at or near the palindrome sequence located at Xq27.1 (Table 1). The first of these, reported in 2005, was X-linked hypoparathyroidism (XL-HPT), arising from agenesis of the parathyroid glands (Bowl et al., 2005). This was caused by a complex genomic rearrangement involving a large insertion from chromosome 2p25.3 accompanied by a deletion at Xq27.1 (Bowl et al., 2005). Ten years later, another unrelated XL-HPT kindred was identified that also harbored an insertion from 2p25.3 and a deletion at Xq27.1 (Taylor et al., 2015). Intriguingly, despite both insertions originating from chromosome 2p25.3,

the inserted sequences were not overlapping between the two kindreds (Gaynor et al., 2020). However, the Xq27.1 deletions that occurred in both families did partially overlap although differed considerably in size (Gaynor et al., 2020). Insertions at Xq27.1 also cause X-linked congenital generalized hypertrichosis (XL-CGH); a rare condition characterized by excessive hair growth (Zhu et al., 2011; DeStefano et al., 2013). Three unrelated XL-CGH kindreds have been reported, all harboring interchromosomal insertions of varying sizes that originate from different autosomes (Zhu et al., 2011; DeStefano et al., 2013). A phenotype of isolated bilateral ptosis (drooping eyelids) is associated with an insertion from chromosome 1p21.3 (Bunyan et al., 2014), whilst a different chromosome 1 insertion from 1q25.2-q25.3 causes XX-male sex reversal (Haines et al., 2015). Interestingly, an XX-male sex reversal phenotype was also observed in a patient with a *de novo* insertion from chromosome 9q21 as part of a complex disease phenotype that also included bilateral anophthalmia, orbital teratoma, and neurodevelopmental anomalies (Chou, 2012). An insertion involving the pseudoautosomal region of the sex chromosomes was found to cause a rare, undiagnosed phenotype consisting of multiple limb abnormalities (genu varum, cubitus valgus) and everted lips (Si et al., 2019), whilst a chromosome 9p22.3-p22.2 insertion was identified as a 'variant of unknown significance' (VUS) in a patient with multiple congenital abnormalities (Middelkamp et al., 2019). More recently, a large insertion originating from 10q21.3 was identified as the likely cause of Plott syndrome; a rare congenital disease characterized by bilateral laryngeal abductor paralysis (vocal cord paralysis) (Boschann et al., 2022).

Table 1: Disease phenotypes associated with insertions (+/- deletions) at the Xq27.1 mutational hotspot. Genomic coordinates are based on the GRCh37/hg19 reference sequence.

Disease phenotype	Additional phenotypic features	Insertion sequence	Insertion size and orientation	Genes contained in insertion	Additional accompanying rearrangements at Xq27.1	Reference
X-linked hypoparathyroidism (OMIM: #307700)	N/A	2p25.3	~ 340 kb (Direct)	*SNTG2* (partial)	~25 kb deletion	Bowl et al., 2005
X-linked hypoparathyroidism (OMIM: #307700)	N/A	2p25.3 (Chr2: 856,644-903,578)	~ 47 kb (Direct)	Non-coding DNA only	~1.5 kb deletion 3 bp insertion	Taylor et al., 2015
X-linked congenital generalized hypertrichosis (OMIM: #307150)	N/A	4q31.22-q31.23 (Chr4: 148,464,548-148,764,583)	~ 300 kb (Inverted)	PRMT9 (full) TMEM184C (full) ARHGAP10 (partial) EDNRA (partial)	25 bp insertion 7 bp deletion*	Zhu et al., 2011
X-linked congenital generalized hypertrichosis (OMIM: #307150)	Scoliosis, spina bifida	5q35.3 (Chr5:177,831,590-177,957,166)	~126 kb (Direct)	COL23A1 (partial)	~1.3 kb deletion 124 bp insertion (from Xq28) 2 bp insertion	Zhu et al., 2011
X-linked congenital generalized hypertrichosis (OMIM: #307150)	Deafness, dental/palate anomalies	6p21.2 (Chr6:39,786,474-39,397,468)	~ 389 kb (Inverted)	KIF6 (partial) DAAM2 (partial)	56 bp insertion (from 3q21.2) 14 bp insertion 2 bp deletion	DeStefano et al., 2013
X-linked congenital bilateral ptosis (OMIM: %300245)	N/A	1p21.3 (Chr1:97,886,267-98,006,168)	~ 120 kb (Direct)	DPYD (partial)	7 bp duplication 427 bp duplication	Bunyan et al., 2014
Bilateral anophthalmia and XX sex reversal	Neurological abnormalities, orbital teratoma	9q21	640 kb (Direct)	TRPM3 (partial)	4 bp inversion 4 bp deletion	Chou, 2012
XX male sex reversal	N/A	1q25.2 – 1q25.3 (Chr1:180,243,986-181,017,783)	~ 774 kb (Direct)	ACBD6 (full) XPR1 (full) KIAA1614 (full) STX6 (full) OVAAL (full) MIR3121 (full) LHX4 (partial) MR1 (partial)	9 bp insertion 4 bp deletion	Haines et al., 2015
Charcot-Marie-Tooth neuropathy, X-linked recessive,3 (OMIM: #302802)	N/A	8q24.3 (Chr8:145,768,312–145,848,158)	~ 78 kb (Direct)	ARHGAP39 (partial)	19 bp insertion (from 12q13.12) 12 bp inversion 1 bp deletion 1 bp single nucleotide variant	Brewer et al., 2016
Rare X-linked recessive compound phenotype (genu varum, cubitus valgus, everted lips)	N/A	Xp22.33/Yp11.32 (ChrX:628,417-733,365/ChrY:578,417-683,365)	~ 105 kb (Direct)	Non-coding DNA only	N/A	Si et al., 2019
Multiple congenital abnormalities	N/A	9p22.3-9p22.2 (Chr9:16,489,097–16,659,203)	~ 170 kb (Inverted)	BNC2 (partial)	4 bp deletion*	Middelkamp et al., 2019
X-linked congenital bilateral laryngeal abductor paralysis (Plott syndrome, OMIM: 308850)	N/A	10q21.3 (Chr10: 65,948,754-66,352,777)	~ 404 kb (Inverted)	Non-coding DNA only	~10 kb deletion 59 bp insertion (from 10q21.3) 7 bp insertion 2 bp insertion	Boschann et al., 2022

**Deletion has been inferred based on the reported coordinates of the insertion breakpoints*

Clearly, this region of chromosome Xq27.1 is prone to genomic rearrangements, likely due to the presence of the 180 bp quasi-palindrome sequence (ChrX: 139,502,865 - 139,503,044; hg19) which is predicted to form an unstable hairpin loop structure (Brewer et al., 2016). Accordingly, for the disease-associated insertions that have been mapped with genetic sequencing, at least one breakpoint was found to localize toward the center of this palindrome (Figure 1.2) (Brewer et al., 2016). Interestingly, this mutational hotspot localizes within a ~410 kb gene desert with the nearest coding genes, *SOX3* and *LOC389895*, located approximately 82 kb and 329 kb away respectively (Figure 1.3) (Zhu et al., 2011; Brewer et al., 2016). Without any clear effect on coding DNA, defining the molecular mechanisms by which these intergenic structural variants produce disease remains a challenge.

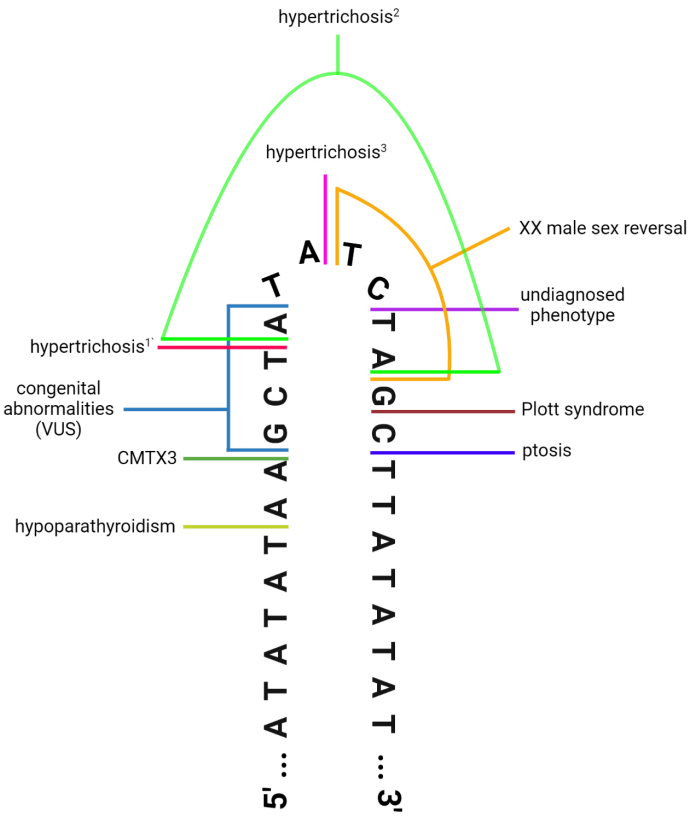


Figure 1.2: DNA-sequencing has revealed that the disease-associated insertions contain at least one breakpoint located toward the center of the 180 bp quasi-palindrome at Xq27.1. Illustration depicts the hairpin loop that is predicted to form at the center of the palindrome. Colored lines denote the position of the breakpoints for the insertions associated with hypoparathyroidism (Taylor et al., 2015), CMTX3 (Brewer et al.,

2016), hypertrichosis[1,2] (Zhu et al., 2011), hypertrichosis[3] (DeStefano et al., 2013), XX-male sex reversal (Haines et al., 2015), ptosis (Bunyan et al., 2014), Plott syndrome (Boschann et al., 2022), and a rare undiagnosed phenotype characterised by limb abnormalities and everted lips (Si et al., 2019). This schematic also depicts the breakpoints for a *de novo* insertion classified as a 'variant of unknown significance' in a patient with congenital abnormalities (Middelkamp et al., 2019). Note that some of the insertions display additional breakpoints located outside this region (Zhu et al., 2011; Taylor et al., 2015; Boschann et al., 2022); see Table 1 for details. Figure has been adapted from Brewer et al., (2016) (Figure 4), originally published in *PLOS Genetics*, to incorporate additional disease-associated insertions that have been identified in recent years. Image created with BioRender.com.

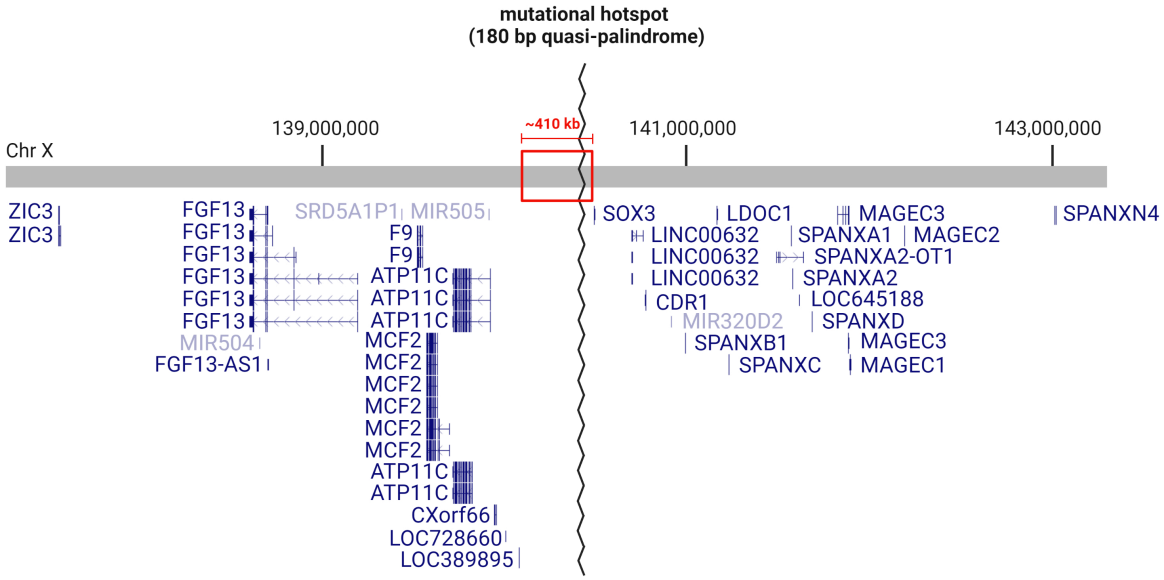


Figure 1.3: Genetic landscape of the mutational hotspot at chromosome Xq27.1. Diagram depicts the RefSeq curated genes located within ~ 3 Mb on either side of the quasi-palindrome sequence, which acts as a hotspot for genomic rearrangements. The quasi-palindrome is situated within a 410 kb gene desert (red box), with 82 kb of non-coding sequence separating it from the closest coding gene (*SOX3*). Image created with BioRender.com using information obtained from UCSC Genome Browser (GRCh38/hg38).

3. Structural variation can disrupt long-range gene regulation and cause genetic disease

The disease-associated insertions at Xq27.1 are classified as structural variation (SV) mutations. SV is a broad term used to describe mutations that affect a minimum of 50 bp of DNA at a time (Sudmant et al., 2015). SV can vary in size ranging from 50 to millions of base pairs and are therefore capable of altering chromatin organization (Zhang et al., 2018). SV can include deletions and duplications (also known as copy number variation), as well as inversions, insertions, and chromosomal translocations (Yang et al., 2013). Complex combinations of the various SV subtypes can also occur (Sudmant et al., 2015). SVs involving duplications/deletions can exert pathogenicity by altering the dosage of genes contained within the rearranged DNA. CMT1A, the most common inherited neuropathy, is the quintessential example of a pathogenic genomic duplication, involving trisomy of the peripheral myelin protein 22 (*PMP22*) gene (Lupski et al., 1991; Matsunami et al., 1992; Patel et al., 1992; Timmerman et al., 1992; Valentijn et al., 1992). However, in the case of balanced SV, and SV affecting the non-coding genome, defining how the DNA rearrangements contribute to genetic disease becomes more challenging. Often, the pathogenic consequence of these mutations involves disruption of regulatory genomic landscapes that, in turn, alters spatiotemporal gene expression (Spielmann et al., 2018; Sánchez-Gaya et al., 2020).

3.1 Cis-regulatory elements: dictating when, where and how much a gene is expressed

Cis-regulatory elements (CREs) are sequences of non-coding DNA controlling the expression of nearby genes (Li et al., 2015). Enhancers are the most abundant (Shen et al., 2012) and most well-studied of the distal CREs (Ray-Jones and Spivakov, 2021). Enhancers increase the expression of their target gene(s), often by recruiting transcriptional machinery and making physical contact with the target promoter (Gasperini et al., 2020; Panigrahi and O'Malley, 2021). Enhancer-promoter (EP) interactions can show marked spatiotemporal specificity and may span large genomic distances, with

reports of enhancers located up to 1.45 Mb away from their target gene (Long et al., 2020). This remarkable specificity was elegantly demonstrated in a series of experiments investigating the regulation of the developmental gene *SOX9*, which identified two novel enhancer clusters regulating *SOX9* expression within a narrow developmental window (Long et al., 2020). Strong chromatin contacts between the novel enhancers and the *SOX9* promoter were observed in cranial neural crest cells, but not in their embryonic stem cell precursors (Long et al., 2020). Furthermore, these enhancers were inactive in cranial chondrocytes which are directly derived from cranial neural crest cells, and interestingly, also display robust *SOX9* expression (Long et al., 2020). Accordingly, genes with tightly controlled expression profiles often utilize multiple enhancers that can act in a cooperative or even redundant manner (Spitz, 2016; Moorthy et al., 2017). Therefore, whilst we are now able to predict the genomic location of putative enhancers by the presence of distinct epigenetic signatures (Bulger and Groudine, 2011), inferring their target gene(s) and their spatiotemporal activity profile presents an enduring challenge (Pennacchio et al., 2013).

In contrast, silencers are regions of DNA that repress the expression of their target gene(s) (Maston et al., 2006) and represent an understudied and poorly understood class of CREs (Gisselbrecht et al., 2020; Pang and Snyder, 2020). However, by systematically screening genomic fragments for repressive activity, over 5,000 candidate human silencers have been recently discovered and were found to share many properties of enhancers such as tissue specificity and long-range communication (Pang and Snyder, 2020). Intriguingly, the use of a Drosophila model recently identified a surprisingly large number of 'bifunctional elements', i.e., regulatory sequences that can function as either enhancers or silencers, depending upon the spatiotemporal context, further attesting to the complexity of gene regulation (Gisselbrecht et al., 2020).

Another important class of distal CREs are insulator elements. These are often binding sites for 'insulator proteins' such as CCCTC-binding factor (CTCF) (Bell et al., 1999). Insulators have a key role in segmenting the genome into discrete structural units which largely define the genomic region over which regulatory elements can act (Bushey et al., 2008). This both facilitates the formation of

appropriate long-range regulatory interactions, whilst also preventing aberrant communication between CREs and non-target genes (Geyer and Clark, 2002; Kuhn and Geyer, 2003). Insulators can also function as barriers that prevent the spread of heterochromatin into adjacent euchromatic regions of DNA (Gaszner and Felsenfeld, 2006). Thus, insulators have a clear impact on the nuclear organization of chromatin (Phillips-Cremins and Corces, 2013).

3.2 The role of the 3D genome in transcriptional regulation

Transcriptional regulation heavily relies on the spatial organization of chromatin within the three-dimensional nucleus. Addressing the many complexities of 3D chromatin architecture is beyond the scope of this review, however, excellent articles provide in-depth discussion of this topic (Bonev and Cavalli, 2016; Stadhouders et al., 2019).

The functionality of CREs is intimately dependent upon chromatin structure (Kleinjan and van Heyningen, 2005). For example, enhancers located within inaccessible, nucleosome-dense regions of chromatin are deemed silent (Bozek and Gompel, 2020), while histone modifications can be used to infer whether an enhancer is in an active or poised state (Creyghton et al., 2010; Rada-Iglesias et al., 2011; Zentner et al., 2011). Furthermore, interactions between distal CREs and target genes can span regions of linear DNA that are very distant from one another (Lettice et al., 2003). These long-range interactions, essential for gene regulation, are largely mediated by the folding of chromatin into hierarchical 3D structures known as 'topologically associating domains' (TADs) (Figure 1.4A) (Dixon et al., 2012; Nora et al., 2012; Rao et al., 2014). Regions of DNA within the same TAD interact frequently with each other, whilst inter-TAD interactions are much less commonly observed (Dixon et al., 2012). TADs are largely conserved between cell types and across species (Dixon et al., 2012), although tissue specificity is observed at the level of sub-TADs; smaller and more dynamic domains visible at the sub-megabase scale (Phillips-Cremins et al., 2013). These structural domains represent local regulatory landscapes, with insulator elements often demarcating the boundary that separates two adjacent TADs (Figure 1.4B) (Dixon et al., 2012; Rao et al., 2014; Schoenfelder and Fraser, 2019).

Here, we have provided a simplified overview of long-range gene regulation, which is a highly complex, multi-faceted process and remains an active area of research. Despite significant progress in recent years, our current understanding of long-range gene regulation remains incomplete (Schoenfelder and Fraser, 2019). In some cases, for example, experimentally depleting TAD boundaries produces a surprisingly minimal impact on gene expression (Nora et al., 2017; Rao et al., 2017), indicating that additional mechanisms are at play. However, it is clear that SV are capable of altering local regulatory domains and inducing pathogenic changes in transcriptional regulation (Kleinjan and Coutinho, 2009; Lupiáñez et al., 2016; Spielmann et al., 2018).

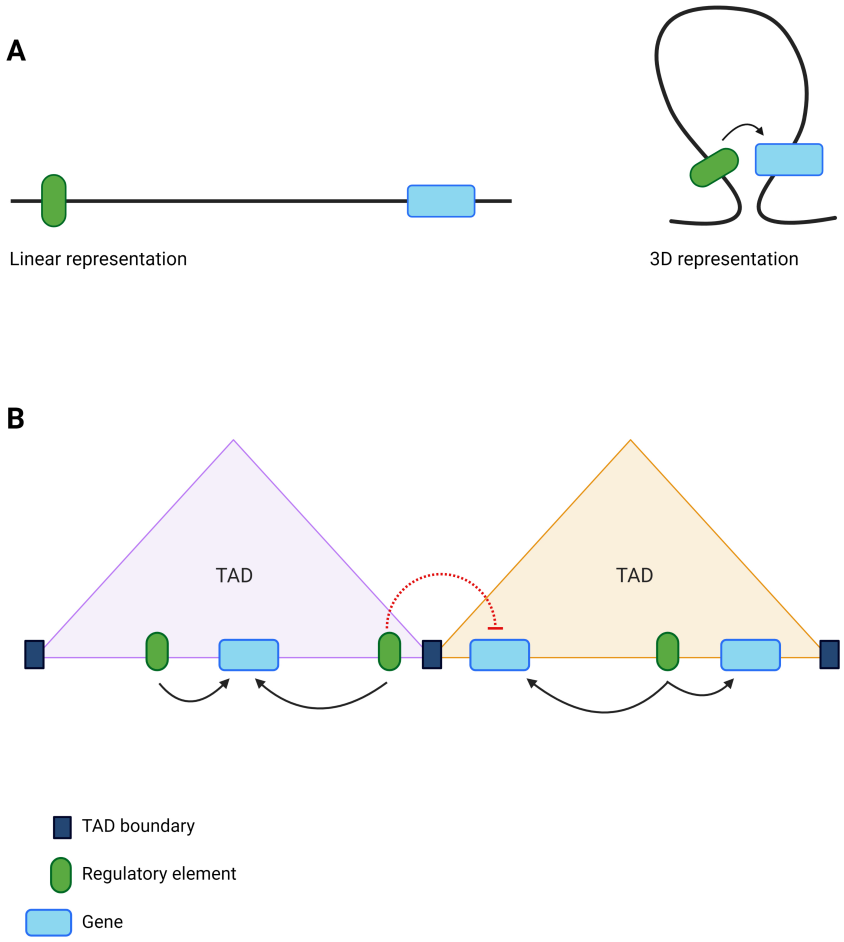


Figure 1.4: The role of 3D genome organisation in long-range gene regulation. (A) The folding of chromatin brings distant sequences into close proximity enabling the formation of long-range interactions between genes and distal regulatory elements. (B) The genome is organised into topologically associating domains (TADs), which act as local regulatory hubs. Sequences of DNA located within the same TAD can interact frequently with each other (black arrows), while chromatin interactions spanning TAD boundaries are much less common (red dotted

lines). This restricts the activity of regulatory elements to the appropriate genes. Image created with BioRender.com.

4. The disease-associated insertions at Xq27.1 can disrupt long-range gene regulation

The genomic landscape of the Xq27.1 locus has shown the disease-associated insertions occur within a gene desert. Therefore, the pathomechanism of these complex mutations does not involve disrupting the sequence of known coding genes located within this genomic region. However, the insertions introduced at Xq27.1 often contain regions of DNA that encode either partial (Bowl et al., 2005; Zhu et al., 2011; Chou, 2012; DeStefano et al., 2013; Bunyan et al., 2014; Brewer et al., 2016; Middelkamp et al., 2019) and/or full length (Zhu et al., 2011; Haines et al., 2015) genes, although 3 of the insertions consist entirely of non-coding sequence (Taylor et al., 2015; Si et al., 2019; Boschann et al., 2022). As the reported insertions are unbalanced, the over-expression of the duplicated genes (either full length or partial transcripts) may be a potential pathomechanism (Bowl et al., 2005; Bunyan et al., 2014; Brewer et al., 2016). This possibility has often been deemed unlikely for multiple reasons including (a) the insertion lacks an open reading frame (Bowl et al., 2005), (b) the duplicated genes have not been associated with the relevant disease phenotype (Haines et al., 2015), or (c) over-expression of the inserted gene(s) was not observed in quantitative gene expression experiments (Brewer et al., 2016; Gaynor et al., 2020).

Another possible pathomechanism may involve aberrant splicing between the inserted sequence and an adjacent gene, resulting in the formation of a fusion transcript (Bunyan et al., 2014; Brewer et al., 2016). In the case of CMTX3, this is unlikely as the inserted sequence and the nearest downstream gene are transcribed from opposite strands (Brewer et al., 2016). This mechanism has not been directly tested for the other insertions at Xq27.1, although in silico splice site prediction for the rearrangement associated with congenital ptosis suggests that it cannot be excluded as a possibility (Bunyan et al.,

2014). Mechanisms involving aberrant splicing are more likely for those insertions containing truncated transcripts with an intact transcription start site but lacking transcription termination signals. In a similar manner, the formation of aberrant read-through transcripts has been observed in other diseases where deletions remove the 3' portion of a gene, which results in transcriptional read-through into the neighbouring gene and epigenetic promoter silencing (Tufarelli et al., 2003; Ligtenberg et al., 2009).

Among researchers reporting the insertions at Xq27.1, the most common hypothesis proposed to explain the pathogenicity involves transcriptional dysregulation of nearby genes (Bowl et al., 2005; Zhu et al., 2011; Chou, 2012; DeStefano et al., 2013; Haines et al., 2015; Taylor et al., 2015; Brewer et al., 2016; Middelkamp et al., 2019; Si et al., 2019; Boschann et al., 2022). SV can pathologically alter regulatory landscapes via multiple mechanisms, with recent reviews available for a comprehensive commentary (Kleinjan and Coutinho, 2009; Spielmann and Mundlos, 2013; Krijger and De Laat, 2016; Lupiáñez et al., 2016; Cutrupi et al., 2018; Spielmann et al., 2018; D'haene and Vergult, 2020; Sánchez-Gaya et al., 2020).

One possibility is that the large insertions result in a position effect on nearby gene(s), altering spatiotemporal regulation by physically distancing the gene from its typical CREs (DeStefano et al., 2013). Initially, this was postulated as a potential pathomechanism for XL-CGH, with reports of 3 unique interchromosomal insertions ranging from 126 – 389 kb causing a common clinical phenotype of excessive hair growth (DeStefano et al., 2013). However, this view has repeatedly been challenged as additional pathogenic insertions were identified at the same genetic locus giving rise to very different clinical phenotypes. Some of the additional insertions reported were significantly larger and would therefore be expected to produce a similar position effect (Haines et al., 2015). Whilst the disconnection of genes and regulatory elements is a known pathomechanism of SV (Spielmann et al., 2018), the current evidence suggests this is not a common mechanism driving the various diseases associated with insertions into the Xq27.1 locus (Si et al., 2019).

Transcriptional dysregulation can also arise from mutations that disrupt TAD boundaries, given that long-range regulatory interactions typically occur within the constraints of these defined genomic domains (Lupiáñez et al., 2015). According to publicly available Hi-C data, the pathogenic insertions at Xq27.1 occur within a TAD and are therefore unlikely to disrupt boundary sites (Brewer et al., 2016; Gaynor et al., 2020). However, the disruption of sub-TAD boundaries may be a possibility and could contribute to the tissue-specificity of the disease phenotypes. Rather than disturbing pre-existing boundaries, Gaynor and colleagues (2020) suggested that the insertions could instead introduce ectopic boundary elements into Xq27.1 which may alter local regulatory interactions. Such a disease mechanism is relatively understudied, however a recent investigation by Willemin and colleagues (2021) has provided valuable insights. After generating a transgenic mouse line whereby a TAD boundary element usually located on chromosome 2 was inserted into a large TAD on chromosome 10, the authors observed splitting of the host TAD into two smaller regulatory domains which was associated with tissue specific changes in local gene expression (Willemin et al., 2021).

Enhancer (or silencer) deletions should also be considered as potential pathomechanisms for the diseases caused by complex genomic rearrangements, such as those observed in XL-HPT, where the insertions occur in combination with deletions at Xq27.1. Interchromosomal insertions from 2p25.3 accompanied by deletions at Xq27.1 were reported in two unrelated XL-HPT kindreds (Bowl et al., 2005; Taylor et al., 2015). The inserted fragments do not share common sequence despite originating from the same chromosome 2p25.3 sub-band, however the deletions overlap by ~ 1.5 kb (Gaynor et al., 2020). As the deleted sequence did not contain ultraconserved elements, which often act as tissue-specific enhancers (Pennacchio et al., 2006), CRE deletion was considered an unlikely disease mechanism (Gaynor et al., 2020). Ultraconserved elements, however, account for only a subset of CREs (Visel et al., 2008) with many enhancers displaying species specificity (Bulger and Groudine, 2011), and therefore this hypothesis should not be discounted.

One of the more prevalent hypotheses postulates that the insertions may introduce ectopic CREs into Xq27.1 which can interact with nearby gene(s) and alter spatiotemporal expression profiles (Zhu et al., 2011; Haines et al., 2015; Middelkamp et al., 2019; Si et al., 2019). Multiple lines of evidence lend support to this hypothesis. Firstly, the nature of the inserted sequence seems important in dictating the phenotype, since the various insertion breakpoints occur within the same palindromic sequence (Si et al., 2019), and insertions of comparable size are associated with different clinical phenotypes. For example, a ~126 kb insertion causes hypertrichosis (Zhu et al., 2011) while a ~120 kb insertion is associated with bilateral ptosis (Bunyan et al., 2014). Furthermore, by consulting publicly available annotations of the non-coding genome, multiple researchers have identified putative regulatory elements residing within the inserted fragments (Middelkamp et al., 2019; Si et al., 2019). Interestingly, a super-enhancer implicated in craniofacial development is contained within an insertion at Xq27.1 that has been classified as a VUS in a patient with a complex congenital phenotype which includes craniofacial anomalies (Middelkamp et al., 2019). Ectopic enhancer adoption is a disease mechanism that has been observed in similar genomic rearrangements. A recent study using the Hammer toe mouse model, which displays a syndactyly phenotype, identified a large interchromosomal insertion occurring upstream of the sonic hedgehog (*Shh*) gene which contains enhancer elements that drive ectopic *Shh* expression during limb development (Mouri et al., 2018).

Irrespective of the underlying molecular mechanism, transcriptional dysregulation of nearby genes has been observed in studies investigating 3 of the disease-associated insertions at Xq27.1 (DeStefano et al., 2013; Haines et al., 2015; Brewer et al., 2016). Lymphoblasts isolated from XX male sex reversal patients displayed ectopic expression of *SOX3* (Haines et al., 2015), whilst XL-CGH patient keratinocytes showed aberrant regulation of *FGF13* (DeStefano et al., 2013). Dysregulation of FGF13 was also observed in lymphoblasts derived from CMTX3 patients (Brewer et al., 2016). Clearly, the pathogenic rearrangements are capable of disrupting local regulatory landscapes and altering the expression profiles of nearby genes. It is interesting to note that both *SOX3* and *FGF13* are developmental regulator genes (Si et al., 2019) with important roles in mammalian embryogenesis (Rizzoti et al., 2004;

Puranam et al., 2015). In general, developmental regulator genes are disproportionately affected in 'cis-ruption diseases', i.e., genetic conditions caused by altered cis-regulation (Kleinjan and Coutinho, 2009). Developmental genes typically carry out multiple functions within a variety of cell types at specified time points, and as such, their expression is usually controlled by the action of numerous spatiotemporally specific CREs (Kleinjan and Coutinho, 2009; Spitz and Furlong, 2012; Spielmann et al., 2018). The marked complexity of their regulatory landscapes, combined with the dosage sensitivity of these developmental genes, may render them less tolerant to disruptions in the local regulatory region (Kleinjan and Coutinho, 2009).

It is clear that SV can act via multiple, intriguing pathomechanisms, including some that are not discussed here (e.g. disturbing non-coding RNA genes, reviewed in D'haene and Vergult, (2020)). Whilst we cannot assume that all of the insertions are acting in a similar manner, there is sufficient evidence to suggest that local gene dysregulation may be a recurrent mechanism amongst this group of diseases.

5. Difficulties in identifying the candidate gene(s) being dysregulated by the disease-associated insertions

Identifying candidate genes being dysregulated by the disease-associated insertions at Xq27.1 has proven a challenging task for many of the phenotypes studied. This is not an uncommon problem as predicting the consequences of complex SV is notoriously difficult and remains an active area of research. This is partly because large SV can simultaneously affect multiple genes and CREs. In some instances, the phenotype can be explained by the activity of a single dysregulated gene, however other phenotypes may arise from the combined effects of the mutation on several candidates (Weischenfeldt et al., 2013; Middelkamp et al., 2019). Long-range gene regulation can span large genomic distances and therefore, many genes could be considered as potential candidates. For

example, when investigating gene dysregulation for XL-CGH, the expression level of all genes located within 3 Mb either side of the causative SV were analyzed (DeStefano et al., 2013). Predicting the effect of complex SV is made more challenging by the incomplete annotation of the non-coding genome. Despite the concerted effort and progress in recent years, obtaining a comprehensive, genome-wide catalogue of CREs remains a challenging task (Pennacchio et al., 2013). Although bioinformatic tools are useful in predicting the presence of putative CREs, the data cannot adequately inform on important aspects of functionality, including (a) specific target gene(s) directly regulated by the CRE, and (b) the spatiotemporal context in which the CRE is active/inactive (Pennacchio et al., 2013).

In an attempt to overcome many of these challenges, Middelkamp and colleagues (2019) proposed a systematic pipeline for solving SV pathomechanisms that combines bioinformatics with experimental methods, whilst also considering the patient's unique phenotype. This approach was used to investigate a range of complex SV found in patients with genetically unsolved (neuro)developmental disorders. Briefly, potential candidate genes (within or 2 Mb adjacent to the SV), were ranked based on phenomatch scores calculated by comparing the patient's phenotype to known phenotypes already associated with each gene (Middelkamp et al., 2019). Transcriptome analyses were then performed using peripheral blood cells to assess candidate gene expression (Middelkamp et al., 2019). To account for the shortcomings of using blood cells to investigate gene dysregulation driving (neuro)developmental disorders, the authors utilized numerous computational tools to further assess candidate genes, including data from multiple cell types at every stage of analysis. Using this strategy, Hi-C datasets predicted genes most likely to be impacted by the mutation (Middelkamp et al., 2019). Epigenomic datasets identified nearby enhancers and predicted how many were being displaced by the SV for each candidate gene (Middelkamp et al., 2019). Promoter capture Hi-C (PCHi-C) data was then queried to detect known long-range interactions being disrupted by the mutation (Middelkamp et al., 2019). This comprehensive pipeline provides an innovative approach to solving this complex problem and is particularly well-suited to identifying pathomechanisms that

involve disruptions to known chromatin interactions (Middelkamp et al., 2019). However, it has been hypothesized that the disease-associated insertions at Xq27.1 may instead cause disease through introducing novel CREs into the local genomic region, resulting in aberrant expression of nearby gene(s) (Zhu et al., 2011; Haines et al., 2015; Si et al., 2019). If so, predicting which gene(s) form ectopic interactions with the novel CREs will be important in understanding the disease mechanism, yet is a daunting task given our limited understanding of long-range gene regulation. This pipeline makes great use of publicly available datasets to investigate local regulatory landscapes. However, some of the relevant technologies, such as PCHi-C, were only recently developed, with data available from a limited number of cell types. Although PCHiC data from 22 cell types was analyzed in this pipeline, 18 of these originate from a hematopoietic lineage. Since regulatory interactions can show marked cell-type specificity, the relevance of publicly available datasets to the diseases associated with insertions at Xq27.1 is questionable, and this data may need to be experimentally obtained.

We also encourage readers to consider that the phenotypic consequences of non-coding mutations can be distinct from coding mutations affecting the same gene (Kleinjan and van Heyningen, 2005). This is clearly observed for the developmental gene, *SHH*. Coding mutations in *SHH* cause holoprosencephaly (Roessler et al., 1996), whilst mutations disrupting an *SHH* enhancer cause the limb-specific phenotype of preaxial polydactyly (Lettice et al., 2003). Phenomatching is useful for prioritizing potential candidate genes, although using it as a filtering strategy may result in important candidates being overlooked.

Due to the multiple disease mechanisms that may underly SV, combined with the incomplete understanding of gene regulation, functional assays are still needed to validate the predictions that arise from these pipelines. Although not highly feasible for large patient cohorts, functional validation will require personalized, rather than high-throughput approaches to account for the patient genetic backgrounds and the specific cell-types impacted by each disease phenotype.

6. Importance of using disease-relevant tissue to investigate the disease-associated insertions at Xq27.1

The diseases caused by insertions at Xq27.1 show marked tissue-specificity. Patients with CMTX3, for example, display an isolated peripheral neuropathy phenotype (Huttner et al., 2006). Therefore, to properly address the prevailing hypothesis that gene dysregulation underlies these complex phenotypes, it is essential that functional studies are performed in the affected tissue type relevant to each disease.

This was clearly demonstrated by previous investigations into the ~ 389 kb insertion that causes XL-CGH (DeStefano et al., 2013). Dysregulated expression of *SOX3,* the closest gene to the pathogenic insertions, was initially considered the most likely mechanism underlying this disease (Zhu et al., 2011). However, transcriptomic analysis of patient skin biopsies identified the dysregulation of another nearby gene, *FGF13*, in XL-CGH patients relative to controls (DeStefano et al., 2013). Further analysis revealed that patient keratinocytes displayed a 6.7-fold reduction in *FGF13* mRNA, whereas expression in fibroblasts was comparable to controls (DeStefano et al., 2013). These findings from different tissues harboring the ~389 kb insertion clearly demonstrate that the structural variants at Xq27.1 can produce cell-type specific dysregulation. Whilst XL-CGH affects cell types that are easily obtainable from patients (skin cells), this is not true for many other diseases associated with insertions at Xq27.1. Without studying disease-relevant tissue, determining the effect of these mutations on the local gene regulation remains an enduring challenge.

When describing the respective insertions at Xq27.1, the *SOX3* gene is often proposed as the most likely candidate given its proximity to the pathogenic insertions (~82 kb) and its important role in developmental regulation (Bowl et al., 2005; Zhu et al., 2011; Taylor et al., 2015; Middelkamp et al., 2019; Boschann et al., 2022). However, experimentally investigating this hypothesis has proven very challenging. In many cases, researchers have been unable to detect *SOX3* expression when using

readily available tissue types, such as peripheral blood samples (Si et al., 2019), skin biopsies (DeStefano et al., 2013), and lymphoblasts (Brewer et al., 2016). This is perhaps unsurprising as this gene belongs to the SOXB1 family of transcription factors that are widely expressed in the early vertebrate embryo and developing nervous system (Uchikawa et al., 1999; Wood and Episkopou, 1999; Lefebvre et al., 2007). Access to disease-relevant tissue has not been possible for some of the phenotypes described, including XL-HPT, which affects embryonic parathyroid gland development (Gaynor et al., 2020). Furthermore, modelling these large and complex chromosomal rearrangements using available cloning technologies is currently not feasible. Thus, to investigate the role of *SOX3* in XL-HPT, Gaynor and colleagues (2020) generated knockout animal models with targeted deletions of either *SOX3* or a nearby ultraconserved element (uc482) presumed to regulate *SOX3* expression. For both transgenic lines, null hemizygotes displayed normal parathyroid gland development and function, suggesting that XL-HPT involves spatiotemporal disturbances in *SOX3* expression that cannot be recapitulated by a knockout model (Gaynor et al., 2020).

Overall, research into the disease-associated insertions at Xq27.1 has provided some valuable lessons for investigating the consequences of intergenic SV including: (1) it is important, although often difficult, to perform studies in the appropriate cellular context, which can depend on both cell-type and developmental stage, and (2) it is critical to utilize the patient's genetic background to retain the pathogenic rearrangement, as generic animal knockout models are not reflecting the regulatory disruptions causing these specific diseases. Fortunately, recent developments in cellular reprogramming technologies can now provide access to patient-derived disease-relevant tissue, thereby circumventing the major challenges that have previously led to a research impasse for these disorders.

7. Cellular reprogramming technology provides access to disease-relevant tissue

Induced pluripotent stem cells (iPSC) theoretically have the capacity to differentiate into any somatic cell population and can be generated from readily accessible sources such as fibroblasts (Takahashi et al., 2007), peripheral blood (Loh et al., 2009), and urine samples (Zhou et al., 2012). Robust protocols are available that utilize iPSC to generate cell types implicated in the pathogenesis of multiple diseases associated with insertions at Xq27.1. CMTX3, for example, can now be studied by differentiating patient-derived iPSC into the major cell types of the peripheral nervous system, including motor neurons (Du et al., 2015), sensory neurons (Saito-Diaz et al., 2021), and Schwann cells (Kim et al., 2017). Alternative reprogramming technologies may also be used to obtain neural tissue, such as 'direct lineage conversion' or 'direct reprogramming' approaches (e.g., Liu et al., 2016; Kim et al., 2020). In both approaches, the desired cell type is generated from mature somatic cells using a protocol that bypasses a pluripotent intermediate population (Prasad et al., 2016). Direct reprogramming produces multipotent progenitors, whilst direct lineage conversion does not (Prasad et al., 2016). Whilst these technologies appear an attractive and time-efficient option, there are advantages of using iPSC-based approaches for studying the diseases associated with insertions at Xq27.1.

iPSC reprogramming reverts mature cells back to a pluripotent embryonic state (Lapasset et al., 2011), and subsequent differentiation protocols attempt to mirror *in vivo* developmental processes to produce the desired cell type (Zhu and Huangfu, 2013). For example, motor neuron differentiation mimics signaling pathways that occur in the developing spinal cord, guiding iPSC through multiple, increasingly specialized precursor populations including caudal neuroepithelial cells and motor neuron progenitors (Du et al., 2015). Access to embryonic cell types is invaluable considering that many of the diseases associated with insertions at Xq27.1 display a clear developmental etiology. This includes, but

is not limited to, XX male sex reversal, characterized by abnormal gonad development (Haines et al., 2015), and XL-HPT, involving agenesis of the parathyroid glands (Bowl et al., 2005) (see Table 1 for additional developmental phenotypes). Fortunately, it is now possible to generate bipotential gonad/testis-like cells (Knarston et al., 2020) and parathyroid-like cells (Lawton et al., 2020) from human iPSC, providing unprecedented access to tissue types that are relevant for studying these complex developmental disorders. Cells can be analyzed at multiple timepoints throughout differentiation, enabling researchers to investigate the dynamic regulatory processes governing tissue development, and the impact that intergenic SV may have on this defined process. As previously discussed, *SOX3* is a high-priority candidate gene likely to be impacted by the insertions and has already been implicated in the pathogenesis of XX male sex reversal (Haines et al., 2015). Given that *SOX3* encodes a transcription factor involved in embryonic development, investigating developmental cell types will provide a valuable insight into how the disease-associated insertions may alter the spatiotemporal regulation of this interesting candidate gene.

Despite their utility, there are some inherent limitations to iPSC models that should be considered. Experimental variability can arise from multiple sources including genetic background of the donor (Rouhani et al., 2014), donor cell-type (Kim et al., 2011) and laboratory-specific practices (Volpato et al., 2018). iPSC lines can vary in their propensity to differentiate into certain lineages (Bock et al., 2011; Boulting et al., 2011). This must be considered when using iPSC models to investigate the disease-associated insertions at Xq27.1. Developmental anomalies are observed in several of these diseases, and the high-priority candidate gene, *SOX3,* is a transcription factor that has been shown to regulate differentiation in studies from numerous animal models (Bylund et al., 2003; Dee et al., 2008; Bergsland et al., 2011; McAninch et al., 2020). Aberrant differentiation may therefore be a cellular phenotype in patient-derived iPSC, and if so, it will be important to disentangle disease-related differences in differentiation from background line-line variability. Donor genetic background accounts for most of the variability observed between iPSC lines (Kajiwara et al., 2012; Rouhani et al., 2014; Kyttälä et al., 2016), and can be avoided through the use of isogenic controls. Unfortunately, the large,

complex nature of the pathogenic insertions at Xq27.1 limits the capacity to generate isogenic control lines, in which case, parental controls may be considered instead (Sánchez-Gaya et al., 2020). Several publications propose additional strategies for minimizing the effects of variability, including the use of relevant statistical approaches and a careful experimental design (Germain and Testa, 2017; Volpato and Webber, 2020). Although iPSC offer a unique opportunity to investigate cellular development *in vitro*, they cannot precisely mirror the complexity of embryogenesis (Sánchez-Gaya et al., 2020). However, co-cultures and 3D organoid systems are becoming increasingly useful for establishing cellular interactions and creating 3D microenvironments that more closely resemble *in vivo* development (Shi et al., 2017; Lee et al., 2020). Despite these limitations, iPSC technology is increasingly being recognized as a valuable tool for studying the pathogenic consequences of complex SV (Drakulic et al., 2020; Sánchez-Gaya et al., 2020). Recently, a patient-derived iPSC model was used to identify the specific regulatory disturbance causing branchio-oculo-facial disorder in a patient harboring a large chromosomal inversion (Laugsch et al., 2019). To date, this approach has not been applied to investigate any of the disease-associated insertions at Xq27.1, yet will likely be an important tool for uncovering the mechanistic complexities of these intriguing phenotypes.

8. Experimental approaches for solving the pathomechanism of CMTX3

After obtaining disease-relevant tissue and performing *in silico* predictions, experimental tools can be used to functionally explore the mechanisms underlying complex SV. Recent publications have clearly demonstrated the utility of iPSC-based approaches for revealing the pathomechanism of these types of mutations (Laugsch et al., 2019; Sánchez-Gaya et al., 2020). In this review, we will expand upon and modify some of these methodologies to create a pipeline suitable for investigating CMTX3

and the other disease-associated insertions at Xq27.1 (Figure 1.5). However, this approach could also be useful in guiding research into similar rearrangements located elsewhere in the genome.

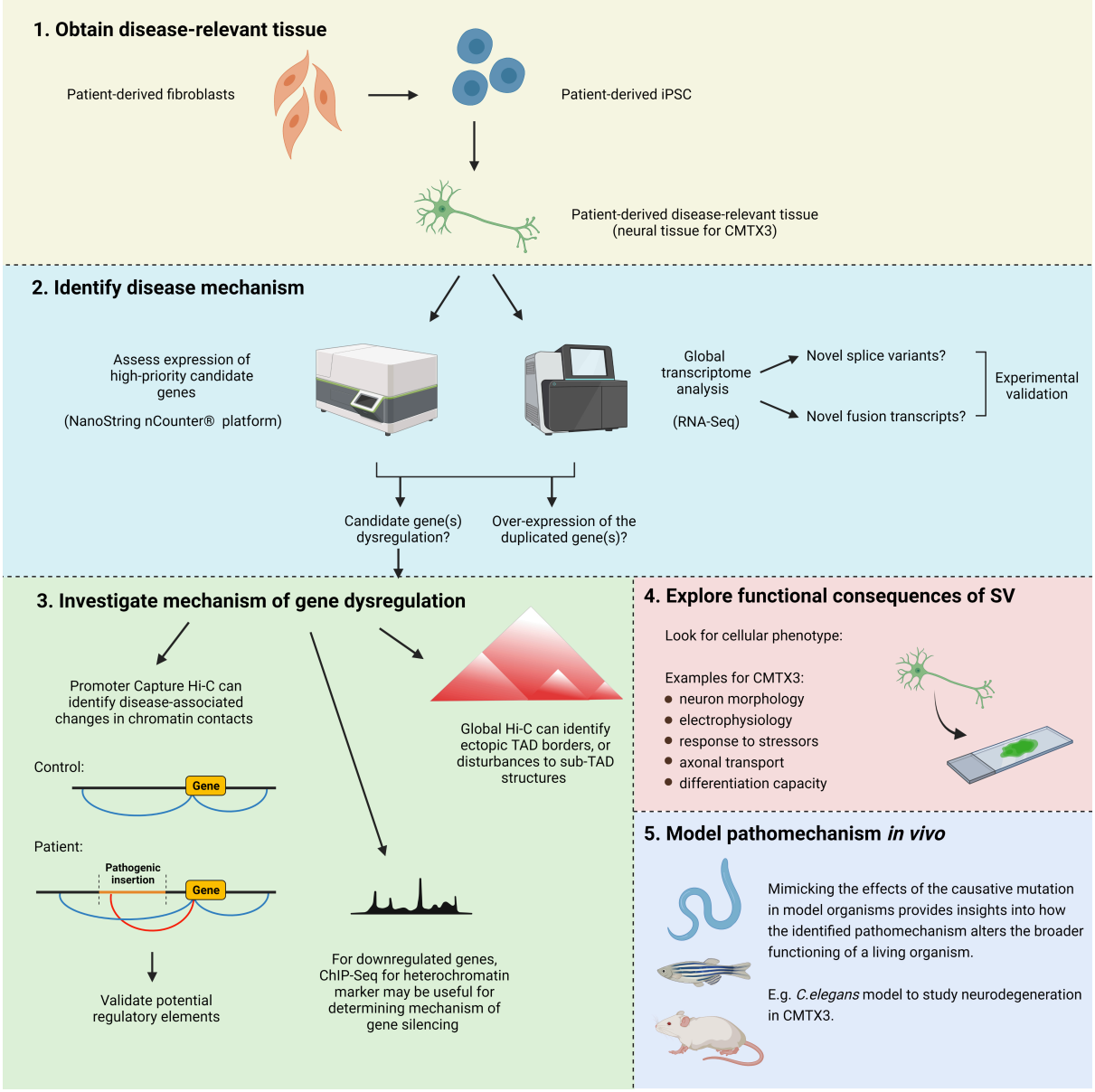


Figure 1.5: Experimental approach for uncovering the pathogenic mechanism of CMTX3 and the other diseases associated with insertions(+/-deletions) at Xq27.1. (1) induced pluripotent stem cells retaining the patient genetic background can be used to generate patient-derived disease-relevant tissue for downstream analysis. (2) The expression level of high-priority candidate genes can be specifically assessed through NanoString nCounter® analysis. Global transcriptomic analysis can be carried out using RNA-Seq. In addition to the predominant hypothesis of gene dysregulation, this approach directly addresses alternative pathomechanisms previously hypothesised to underly these diseases, such as abnormal splice variants, aberrant fusion transcripts, and overexpression of the duplicated genes. (3) If gene dysregulation is observed, experiments can help to uncover

the underlying mechanism. Promoter Capture Hi-C can identify alterations in long-range regulatory interactions. Global Hi-C can identify ectopic TAD boundaries that might be introduced by the insertions, as well as changes to tissue-specific sub-TAD organisation. If the disease mechanism involves gene repression, ChIP-Seq for a heterochromatin marker (H3K9me3) can identify whether this arises from mutation-induced alterations to the local chromatin environment rather than altered regulatory interactions (Laugsch et al., 2018). 4. A range of functional tests can be performed to look for a cellular phenotype in patient samples. 5. Potential pathomechanisms can be studied *in vivo* using model organisms. Image created with BioRender.com.

Local gene dysregulation is a potential pathomechanism driving CMTX3. Earlier research demonstrated that the causative insertion alters the regulation of *FGF13* in patient lymphoblasts (Brewer et al., 2016). However, since this was not assessed in neuronal tissue, it remains unclear whether aberrant *FGF13* regulation is directly implicated in the pathogenesis of CMTX3 or merely a bystander effect of this complex SV within lymphoblasts (Brewer et al., 2016). Therefore, the first step of our experimental pipeline will address gene expression analysis within patient-derived neuronal tissue.

The expression level of positional candidate genes (i.e., those contained within, or 3 Mb adjacent to, the pathogenic insertion) can be specifically assessed through targeted transcriptome profiling. By utilizing tools such as the NanoString Technologies nCounter® platform, custom panels can be designed to accurately quantify expression of high priority candidates at the mRNA and/or protein level. Importantly, by screening samples from progressive time-points throughout differentiation, it is possible to identify gene dysregulation that may be temporally restricted.

Global transcriptome profiling using RNA-seq should also be performed on CMTX3 neuronal tissue. This will provide an additional dataset to validate dysregulation of candidate genes, and through gene ontology analysis, can identify specific biological pathways impacted by the SV. Importantly, RNA-seq analysis will also detect fusion transcripts and/or novel transcript variants arising from the inserted fragment, which are alternative pathomechanisms that may cause CMTX3.

Once a strong candidate gene (or genes) has been identified, experimental tools can help uncover the molecular mechanisms driving the dysregulation. Using the candidate gene(s) promoter as an anchor, Capture Hi-C (Mifsud et al., 2015; Schoenfelder et al., 2015), can identify the specific regions of chromatin that interact with the target gene(s). Comparing the interaction profiles of patient and control samples will facilitate the identification of any disease-associated changes in the 3D chromatin contacts (i.e., gains or losses of DNA interactions). The genomic coordinates participating in the differential interactions should then be queried in genome browsers to gain an insight into potential disease mechanisms. For example, epigenomic datasets (e.g. Ensembl Regulatory Build; Zerbino et al., 2015) can be used to predict whether the region encodes putative CREs, which would suggest that altered regulatory interactions may underly the disease. Ideally, data from disease-relevant tissue should be utilized, although this may not be possible if the candidate gene dysregulation is spatially and temporally restricted. If the relevant tissue is not available, epigenomic data can instead be experimentally obtained. By using ChIP-Seq to profile various histone modifications, as well as ATAC-Seq to identify regions of open chromatin, novel enhancers regulating *SOX9* expression within iPSC-derived cranial neural crest cells were recently predicted (Long et al., 2020). Any potential CREs identified using these techniques should be functionally validated, which can be achieved using a variety of *in vitro* and/or *in vivo* reporter assays (Laugsch et al., 2019; Long et al., 2020). A global Hi-C approach may also be useful, particularly for determining whether ectopic TAD boundaries are being introduced by the pathogenic insertion, and for further characterizing the local regulatory landscape within the disease-relevant tissue. Micro-C, a modern variant of Hi-C technology, utilizes MNase rather than restriction enzymes to fragment the genome (Hsieh et al., 2015), allowing for a much higher resolution view of chromatin organization. With the capacity to detect ~3-5 fold more chromatin loops than conventional Hi-C, Micro-C improves the identification of individual regulatory interactions whilst providing a deeper insight into global chromatin architecture (Krietenstein et al., 2020). Gene dysregulation may also arise through position effect variegation, whereby a gene can be silenced if it is placed in the vicinity of heterochromatin (Kleinjan and Van

Heyningen, 1998). If the pathomechanism appears to involve candidate gene repression, ChIP-Seq analysis for the heterochromatin marker, H3K9me3, can be useful for determining whether the downregulation is arising from inappropriate heterochromatinization of the local region rather than altered regulatory interactions (Laugsch et al., 2019).

Morphological and functional characterization of patient-derived cells will provide deeper insight into the biological processes causing the disease. Having an understanding of the underlying pathomechanism will be useful for prioritizing the types of phenotypic tests to perform. Furthermore, assessing gene dysregulation in the context of global transcriptomic analysis can assist in identifying cellular pathogenic readouts. As an example, RNA-Seq analysis identified dysregulation of genes involved in cellular migration in samples obtained from a patient harboring a complex SV causing branchio-oculo-facial syndrome (Laugsch et al., 2019). Subsequently, in vitro migration assays were performed, revealing an impaired migratory capacity in patient-derived cranial neural crest cells (Laugsch et al., 2019). After identifying likely pathomechanisms, model organisms can be used to assess how these diseases more broadly affect animal function and behavior *in vivo*. The nematode *Caenorhabditis elegans* (*C. elegans*) is particularly useful for modelling neuronal diseases, including CMT, due to their simple and well-characterized nervous system. Recently, another form of X-linked CMT, CMTX6, was modelled using this approach, which was able to recapitulate the locomotion deficits and neurodegeneration presenting in patients (Narayanan et al., 2021). Whilst CMTX6 is caused by a point mutation (Kennerson et al., 2013), C. elegans is also suitable for modelling pathomechanisms associated with complex SV, such as gene dysregulation. Gene upregulation can be studied with an overexpression model, whereby tissue-specific promoters or inducible expression systems can confer spatiotemporal specificity. RNA interference of endogenous gene orthologues can model candidate gene downregulation, which may also be restricted to desired cell-types (Qadota et al., 2007) or developmental stages (Sin et al., 2014). Furthermore, alternative pathomechanisms such as novel splice variants or fusion transcripts can be studied via the delivery of transgenic constructs.

9. Conclusion

We have proposed a strategy for investigating the pathomechanism underlying CMTX3 and the other diseases associated with similar genomic rearrangements occurring at chromosome Xq27.1. We have incorporated the use of cutting-edge technologies that have a demonstrated capacity to revolutionize our understanding of complex SV, but have not yet been applied to study the disease-associated insertions (+/- deletions) at Xq27.1. This intriguing group of diseases may act as a unique paradigm for providing valuable insight into the pathogenic mechanisms underlying complex intergenic SV and more broadly, enhance our understanding of long-range gene regulation. Overall, SV involving the non-coding genome represents an understudied class of mutation contributing to unsolved genetic diseases that have been excluded for gene coding mutations. Understanding the molecular mechanisms will be essential for improving genetic diagnosis and guiding the development of future therapies for patients suffering from these debilitating conditions.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

M.L.K and A.B conceived the project and planned the review content. A.B wrote manuscript and prepared the figures. M.L.K and G.P-S corrected and edited manuscript. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by a National Health and Medical Research Council Ideas Grant (APP1186867) awarded to M.L.K and G.P.S. This research is also supported by an Australian Government Research Training Program (RTP) Scholarship awarded to A.B.

References

Bell, A.C., West, A.G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell 98(3), 387-396.

Bergsland, M., Ramsköld, D., Zaouter, C., Klum, S., Sandberg, R., and Muhr, J. (2011). Sequentially acting Sox transcription factors in neural lineage development. Genes & development 25(23), 2453-2464.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z.D., et al. (2011). Reference Maps of human ES and iPS cell variation enable high-throughput characterization of pluripotent cell lines. Cell 144(3), 439-452.

Bonev, B., and Cavalli, G. (2016). Organization and function of the 3D genome. Nature Reviews Genetics 17(11), 661-678.

Boschann, F., Moreno, D.A., Mensah, M.A., Sczakiel, H.L., Skipalova, K., Holtgrewe, M., et al. (2022). Xq27. 1 palindrome mediated interchromosomal insertion likely causes familial congenital bilateral laryngeal abductor paralysis (Plott syndrome). *Journal of Human Genetics***,** 1-6.

Boulting, G.L., Kiskinis, E., Croft, G.F., Amoroso, M.W., Oakley, D.H., Wainger, B.J., et al. (2011). A functionally characterized test set of human induced pluripotent stem cells. Nature biotechnology 29(3), 279-286.

Bowl, M.R., Nesbit, M.A., Harding, B., Levy, E., Jefferson, A., Volpi, E., et al. (2005). An interstitial deletion-insertion involving chromosomes 2p25. 3 and Xq27. 1, near SOX3, causes X-linked recessive hypoparathyroidism. The Journal of clinical investigation 115(10), 2822-2831.

Bozek, M., and Gompel, N. (2020). Developmental Transcriptional Enhancers: A Subtle Interplay between Accessibility and Activity: Considering Quantitative Accessibility Changes between Different Regulatory States of an Enhancer Deconvolutes the Complex Relationship between Accessibility and Activity. BioEssays 42(4), 1900188.

Brewer, M.H., Chaudhry, R., Qi, J., Kidambi, A., Drew, A.P., Menezes, M.P., et al. (2016). Whole genome sequencing identifies a 78 kb insertion from chromosome 8 as the cause of Charcot-Marie-Tooth neuropathy CMTX3. PLoS genetics 12(7).

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell 144(3), 327-339.

Bunyan, D.J., Robinson, D.O., Tyers, A.G., Huang, S., Maloney, V.K., Grand, F.H., et al. (2014). X-linked dominant congenital ptosis cosegregating with an interstitial insertion of a chromosome 1p21. 3 fragment into a quasipalindromic sequence in Xq27. 1. Open Journal of Genetics 4(06), 415.

Bushey, A.M., Dorman, E.R., and Corces, V.G. (2008). Chromatin insulators: regulatory mechanisms and epigenetic inheritance. Molecular cell 32(1), 1-9.

Bylund, M., Andersson, E., Novitch, B.G., and Muhr, J. (2003). Vertebrate neurogenesis is counteracted by Sox1–3 activity. Nature neuroscience 6(11), 1162-1168.

Chou, C.M. (2012). Mutations Involving RBP4 and SOX3 Underlie Two Novel Forms of Congenital Eye Malformations.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences 107(50), 21931-21936.

Cutrupi, A.N., Brewer, M.H., Nicholson, G.A., and Kennerson, M.L. (2018). Structural variations causing inherited peripheral neuropathies: a paradigm for understanding genomic organization, chromatin interactions, and gene dysregulation. Molecular genetics & genomic medicine 6(3), 422-433.

D'haene, E., and Vergult, S. (2020). Interpreting the impact of noncoding structural variation in neurodevelopmental disorders. Genetics in Medicine, 1-13.

Dee, C.T., Hirst, C.S., Shih, Y.-H., Tripathi, V.B., Patient, R.K., and Scotting, P.J. (2008). Sox3 regulates both neural fate and differentiation in the zebrafish ectoderm. Developmental biology 320(1), 289-301.

DeStefano, G.M., Fantauzzo, K.A., Petukhova, L., Kurban, M., Tadin-Strapps, M., Levy, B., et al. (2013). Position effect on FGF13 associated with X-linked congenital generalized hypertrichosis. Proceedings of the National Academy of Sciences 110(19), 7790-7795.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., et al. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485(7398), 376-380.

Drakulic, D., Djurovic, S., Syed, Y.A., Trattaro, S., Caporale, N., Falk, A., et al. (2020). Copy number variants (CNVs): a powerful tool for iPSC-based modelling of ASD. Molecular autism 11, 1-18.

Du, Z.-W., Chen, H., Liu, H., Lu, J., Qian, K., Huang, C.-L., et al. (2015). Generation and expansion of highly pure motor neuron progenitors from human pluripotent stem cells. Nature communications 6(1), 1-9.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414), 57.

Gasperini, M., Tome, J.M., and Shendure, J. (2020). Towards a comprehensive catalogue of validated and target-linked human enhancers. Nature Reviews Genetics, 1-19.

Gaszner, M., and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nature Reviews Genetics 7(9), 703-713.

Gaynor, K.U., Grigorieva, I.V., Mirczuk, S.M., Piret, S.E., Kooblall, K.G., Stevenson, M., et al. (2020). Studies of mice deleted for Sox3 and uc482: relevance to X-linked hypoparathyroidism. Endocrine Connections 9(2), 173-186.

Germain, P.-L., and Testa, G. (2017). Taming human genetic variability: transcriptomic meta-analysis guides the experimental design and interpretation of iPSC-based disease modeling. Stem Cell Reports 8(6), 1784-1796.

Geyer, P., and Clark, I. (2002). Protecting against promiscuity: the regulatory role of insulators. Cellular and Molecular Life Sciences CMLS 59(12), 2112-2127.

Gisselbrecht, S.S., Palagi, A., Kurland, J.V., Rogers, J.M., Ozadam, H., Zhan, Y., et al. (2020). Transcriptional silencers in Drosophila serve a dual role as transcriptional enhancers in alternate cellular contexts. Molecular Cell 77(2), 324-337. e328.

Haines, B., Hughes, J., Corbett, M., Shaw, M., Innes, J., Patel, L., et al. (2015). Interchromosomal insertional translocation at Xq26. 3 alters SOX3 expression in an individual with XX male sex reversal. The Journal of Clinical Endocrinology & Metabolism 100(5), E815-E820.

Hsieh, T.-H.S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., and Rando, O.J. (2015). Mapping nucleosome resolution chromosome folding in yeast by micro-C. *Cell* 162(1)**,** 108-119.

Huttner, I., Kennerson, M., Reddel, S., Radovanovic, D., and Nicholson, G. (2006). Proof of genetic heterogeneity in X-linked Charcot–Marie–Tooth disease. Neurology 67(11), 2016-2021.

Kajiwara, M., Aoi, T., Okita, K., Takahashi, R., Inoue, H., Takayama, N., et al. (2012). Donor-dependent variations in hepatic differentiation from human-induced pluripotent stem cells. Proceedings of the National Academy of Sciences 109(31), 12538-12543.

Kennerson, M.L., Yiu, E.M., Chuang, D.T., Kidambi, A., Tso, S.-C., Ly, C., et al. (2013). 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. Human molecular genetics 22(7), 1404-1416.

Kim, H.-S., Kim, J.Y., Song, C.L., Jeong, J.E., and Cho, Y.S. (2020). Directly induced human Schwann cell precursors as a valuable source of Schwann cells. Stem Cell Research & Therapy 11(1), 1-12.

Kim, H.-S., Lee, J., Lee, D.Y., Kim, Y.-D., Kim, J.Y., Lim, H.J., et al. (2017). Schwann cell precursors from human pluripotent stem cells as a potential therapeutic target for myelin repair. Stem cell reports 8(6), 1714-1726.

Kim, K., Zhao, R., Doi, A., Ng, K., Unternaehrer, J., Cahan, P., et al. (2011). Donor cell type can influence the epigenome and differentiation potential of human induced pluripotent stem cells. Nature biotechnology 29(12), 1117-1119.

Kleinjan, D.-J., and Coutinho, P. (2009). Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease. Briefings in Functional Genomics and Proteomics 8(4), 317-332.

Kleinjan, D.-J., and Van Heyningen, V. (1998). Position effect in human genetic disease. Human molecular genetics 7(10), 1611-1618.

Kleinjan, D.A., and van Heyningen, V. (2005). Long-range control of gene expression: emerging mechanisms and disruption in disease. The American Journal of Human Genetics 76(1), 8-32.

Knarston, I.M., Pachernegg, S., Robevska, G., Ghobrial, I., Er, P.X., Georges, E., et al. (2020). An In Vitro Differentiation Protocol for Human Embryonic Bipotential Gonad and Testis Cell Development. Stem cell reports 15(6), 1377-1391.

Krietenstein, N., Abraham, S., Venev, S.V., Abdennur, N., Gibcus, J., Hsieh, T.-H.S., et al. (2020). Ultrastructural details of mammalian chromosome architecture. *Molecular Cell*.

Krijger, P.H.L., and De Laat, W. (2016). Regulation of disease-associated gene expression in the 3D genome. Nature reviews Molecular cell biology 17(12), 771.

Kuhn, E.J., and Geyer, P.K. (2003). Genomic insulators: connecting properties to mechanism. Current opinion in cell biology 15(3), 259-265.

Kyttälä, A., Moraghebi, R., Valensisi, C., Kettunen, J., Andrus, C., Pasumarthy, K.K., et al. (2016). Genetic variability overrides the impact of parental cell type and determines iPSC differentiation potential. Stem cell reports 6(2), 200-212.

Lapasset, L., Milhavet, O., Prieur, A., Besnard, E., Babled, A., Aït-Hamou, N., et al. (2011). Rejuvenating senescent and centenarian human cells by reprogramming through the pluripotent state. Genes & development 25(21), 2248-2253.

Laugsch, M., Bartusel, M., Rehimi, R., Alirzayeva, H., Karaolidou, A., Crispatzu, G., et al. (2019). Modeling the pathological long-range regulatory effects of human structural variation with patient-specific hiPSCs. Cell Stem Cell 24(5), 736-752. e712.

Lawton, B.R., Martineau, C., Sosa, J.A., Roman, S., Gibson, C.E., Levine, M.A., et al. (2020). Differentiation of parathyroid hormone expressing cells from human pluripotent stem cells. Endocrinology.

Lee, K., Hawi, Z., Parkington, H., Parish, C., Kumar, P., Polo, J., et al. (2020). The application of human pluripotent stem cells to model the neuronal and glial components of neurodevelopmental disorders. Molecular psychiatry 25(2), 368-378.

Lefebvre, V., Dumitriu, B., Penzo-Méndez, A., Han, Y., and Pallavi, B. (2007). Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. The international journal of biochemistry & cell biology 39(12), 2195-2214.

Lettice, L.A., Heaney, S.J., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., et al. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Human molecular genetics 12(14), 1725-1735.

Li, Y., Chen, C.-y., Kaye, A.M., and Wasserman, W.W. (2015). The identification of cis-regulatory elements: A review from a machine learning perspective. Biosystems 138, 6-17.

Ligtenberg, M.J., Kuiper, R.P., Chan, T.L., Goossens, M., Hebeda, K.M., Voorendt, M., et al. (2009). Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3′ exons of TACSTD1. Nature genetics 41(1), 112-117.

Liu, M.-L., Zang, T., and Zhang, C.-L. (2016). Direct lineage reprogramming reveals disease-specific phenotypes of motor neurons from human ALS patients. Cell reports 14(1), 115-128.

Loh, Y.-H., Agarwal, S., Park, I.-H., Urbach, A., Huo, H., Heffner, G.C., et al. (2009). Generation of induced pluripotent stem cells from human blood. Blood, The Journal of the American Society of Hematology 113(22), 5476-5479.

Long, H.K., Osterwalder, M., Welsh, I.C., Hansen, K., Davies, J.O., Liu, Y.E., et al. (2020). Loss of extreme long-range enhancers in human neural crest drives a craniofacial disorder. Cell Stem Cell 27(5), 765-783. e714.

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., et al. (2015). Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161(5), 1012-1025.

Lupiáñez, D.G., Spielmann, M., and Mundlos, S. (2016). Breaking TADs: how alterations of chromatin domains result in disease. Trends in Genetics 32(4), 225-237.

Lupski, J.R., de Oca-Luna, R.M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B.J., et al. (1991). DNA duplication associated with Charcot-Marie-Tooth disease type 1A. Cell 66(2), 219-232.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet. 7, 29-59.

Matsunami, N., Smith, B., Ballard, L., Lensch, M.W., Robertson, M., Albertsen, H., et al. (1992). Peripheral myelin protein–22 gene maps in the duplication in chromosome 17p11. 2 associated with Charcot–Marie–Tooth 1A. Nature genetics 1(3), 176-179.

McAninch, D., Mäkelä, J.-A., La, H.M., Hughes, J.N., Lovell-Badge, R., Hobbs, R.M., et al. (2020). SOX3 promotes generation of committed spermatogonia in postnatal mouse testes. Scientific reports 10(1), 1-13.

Middelkamp, S., Vlaar, J.M., Giltay, J., Korzelius, J., Besselink, N., Boymans, S., et al. (2019). Prioritization of genes driving congenital phenotypes of patients with de novo genomic structural variants. Genome medicine 11(1), 1-15.

Mifsud, B., Tavares-Cadete, F., Young, A.N., Sugar, R., Schoenfelder, S., Ferreira, L., et al. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nature genetics 47(6), 598.

Moorthy, S.D., Davidson, S., Shchuka, V.M., Singh, G., Malek-Gilani, N., Langroudi, L., et al. (2017). Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. Genome research 27(2), 246-258.

Mouri, K., Sagai, T., Maeno, A., Amano, T., Toyoda, A., and Shiroishi, T. (2018). Enhancer adoption caused by genomic insertion elicits interdigital Shh expression and syndactyly in mouse. Proceedings of the National Academy of Sciences 115(5), 1021-1026.

Narayanan, R.K., Brewer, M.H., Perez-Siles, G., Ellis, M., Ly, C., Burgess, A., et al. (2021). Charcot–Marie–tooth disease causing mutation (p. R158H) in pyruvate dehydrogenase kinase 3 (PDK3) affects synaptic transmission, ATP production and causes neurodegeneration in a CMTX6 C. elegans model. Human Molecular Genetics.

Nora, E.P., Goloborodko, A., Valton, A.-L., Gibcus, J.H., Uebersohn, A., Abdennur, N., et al. (2017). Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization. Cell 169(5), 930-944. e922.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., et al. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485(7398), 381-385.

Pang, B., and Snyder, M.P. (2020). Systematic identification of silencers in human cells. Nature genetics 52(3), 254-263.

Panigrahi, A., and O'Malley, B.W. (2021). Mechanisms of enhancer action: the known and the unknown. Genome biology 22(1), 1-30.

Patel, P.I., Roa, B.B., Welcher, A.A., Schoener-Scott, R., Trask, B.J., Pentao, L., et al. (1992). The gene for the peripheral myelin protein PMP–22 is a candidate for Charcot–Marie–Tooth disease type 1A. Nature genetics 1(3), 159-165.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature 444(7118), 499-502.

Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A., and Bejerano, G. (2013). Enhancers: five essential questions. Nature Reviews Genetics 14(4), 288-295.

Phillips-Cremins, J.E., and Corces, V.G. (2013). Chromatin insulators: linking genome organization to cellular function. Molecular cell 50(4), 461-474.

Phillips-Cremins, J.E., Sauria, M.E., Sanyal, A., Gerasimova, T.I., Lajoie, B.R., Bell, J.S., et al. (2013). Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell 153(6), 1281-1295.

Prasad, A., Manivannan, J., Loong, D.T., Chua, S.M., Gharibani, P.M., and All, A.H. (2016). A review of induced pluripotent stem cell, direct conversion by trans-differentiation, direct reprogramming and oligodendrocyte differentiation. Regenerative medicine 11(2), 181-191.

Puranam, R.S., He, X.P., Yao, L., Le, T., Jang, W., Rehder, C.W., et al. (2015). Disruption of Fgf13 causes synaptic excitatory–inhibitory imbalance and genetic epilepsy and febrile seizures plus. Journal of Neuroscience 35(23), 8866-8881.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature 470(7333), 279-283.

Rao, S.S., Huang, S.-C., St Hilaire, B.G., Engreitz, J.M., Perez, E.M., Kieffer-Kwon, K.-R., et al. (2017). Cohesin loss eliminates all loop domains. Cell 171(2), 305-320. e324.

Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., et al. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7), 1665-1680.

Ray-Jones, H., and Spivakov, M. (2021). Transcriptional enhancers and their communication with gene promoters. Cellular and Molecular Life Sciences, 1-33.

Rizzoti, K., Brunelli, S., Carmignac, D., Thomas, P.Q., Robinson, I.C., and Lovell-Badge, R. (2004). SOX3 is required during the formation of the hypothalamo-pituitary axis. Nature genetics 36(3), 247-255.

Roessler, E., Belloni, E., Gaudenz, K., Jay, P., Berta, P., Scherer, S.W., et al. (1996). Mutations in the human Sonic Hedgehog gene cause holoprosencephaly. Nature genetics 14(3), 357-360.

Rouhani, F., Kumasaka, N., de Brito, M.C., Bradley, A., Vallier, L., and Gaffney, D. (2014). Genetic background drives transcriptional variation in human induced pluripotent stem cells. PLoS genetics 10(6), e1004432.

Saito-Diaz, K., Street, J.R., Ulrichs, H., and Zeltner, N. (2021). Derivation of Peripheral Nociceptive, Mechanoreceptive, and Proprioceptive Sensory Neurons from the same Culture of Human Pluripotent Stem Cells. Stem cell reports 16(3), 446-457.

Sánchez-Gaya, V., Mariner-Faulí, M., and Rada-Iglesias, A. (2020). Rare or overlooked? Structural disruption of regulatory domains in human neurocristopathies. Frontiers in Genetics 11.

Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. Nature Reviews Genetics, 1.

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., et al. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. Genome research 25(4), 582-597.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature 488(7409), 116-120.

Shi, Y., Inoue, H., Wu, J.C., and Yamanaka, S. (2017). Induced pluripotent stem cell technology: a decade of progress. Nature reviews Drug discovery 16(2), 115-130.

Si, N., Meng, X., Zhao, Z., Xia, W., and Zhang, X. (2019). A 105 kb interstitial insertion in the Xq27. 1 palindrome from pseudoautosomal region PAR1 causes a novel X-linked recessive compound phenotype. Journal of translational medicine 17(1), 138.

Sin, O., Michels, H., and Nollen, E.A. (2014). Genetic screens in Caenorhabditis elegans models for neurodegenerative diseases. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1842(10), 1951-1959.

Spielmann, M., Lupiáñez, D.G., and Mundlos, S. (2018). Structural variation in the 3D genome. Nature Reviews Genetics 19(7), 453-467.

Spielmann, M., and Mundlos, S. (2013). Structural variations, the regulatory landscape of the genome and their alteration in human disease. Bioessays 35(6), 533-543.

Spitz, F. (Year). "Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles", in: Seminars in cell & developmental biology: Elsevier), 57-67.

Spitz, F., and Furlong, E.E. (2012). Transcription factors: from enhancer binding to developmental control. Nature reviews genetics 13(9), 613-626.

Stadhouders, R., Filion, G.J., and Graf, T. (2019). Transcription factors and 3D genome conformation in cell-fate decisions. Nature 569(7756), 345-354.

Sudmant, P.H., Rausch, T., Gardner, E.J., Handsaker, R.E., Abyzov, A., Huddleston, J., et al. (2015). An integrated map of structural variation in 2,504 human genomes. Nature 526(7571), 75-81.

Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., et al. (2007). Induction of pluripotent stem cells from adult human fibroblasts by defined factors. cell 131(5), 861-872.

Taylor, J.C., Martin, H.C., Lise, S., Broxholme, J., Cazier, J.-B., Rimmer, A., et al. (2015). Factors influencing success of clinical genome sequencing across a broad spectrum of disorders. Nature genetics 47(7), 717-726.

Timmerman, V., Nelis, E., Van Hul, W., Nieuwenhuijsen, B., Chen, K., Wang, S., et al. (1992). The peripheral myelin protein gene PMP–22 is contained within the Charcot–Marie–Tooth disease type 1A duplication. Nature genetics 1(3), 171-175.

Timmerman, V., Strickland, A.V., and Züchner, S. (2014). Genetics of Charcot-Marie-Tooth (CMT) disease within the frame of the human genome project success. Genes 5(1), 13-32.

Tufarelli, C., Stanley, J.A.S., Garrick, D., Sharpe, J.A., Ayyub, H., Wood, W.G., et al. (2003). Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nature genetics 34(2), 157-165.

Uchikawa, M., Kamachi, Y., and Kondoh, H. (1999). Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken. Mechanisms of development 84(1-2), 103-120.

Valentijn, L., Bolhuis, P., Zorn, I., Hoogendijk, J., Van den Bosch, N., Hensels, G., et al. (1992). The peripheral myelin gene PMP–22/GAS–3 is duplicated in Charcot–Marie–Tooth disease type 1A. Nature genetics 1(3), 166-170.

Visel, A., Prabhakar, S., Akiyama, J.A., Shoukry, M., Lewis, K.D., Holt, A., et al. (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nature genetics 40(2), 158-160.

Volpato, V., Smith, J., Sandor, C., Ried, J.S., Baud, A., Handel, A., et al. (2018). Reproducibility of molecular phenotypes after long-term differentiation to human iPSC-derived neurons: a multi-site omics study. Stem cell reports 11(4), 897-911.

Volpato, V., and Webber, C. (2020). Addressing variability in iPSC-derived models of human disease: guidelines to promote reproducibility. Disease models & mechanisms 13(1), dmm042317.

Weischenfeldt, J., Symmons, O., Spitz, F., and Korbel, J.O. (2013). Phenotypic impact of genomic structural variation: insights from and for human disease. Nature Reviews Genetics 14(2), 125-138.

Willemin, A., Lopez-Delisle, L., Bolt, C.C., Gadolini, M.-L., Duboule, D., and Rodriguez-Carballo, E. (2021). Induction of a chromatin boundary in vivo upon insertion of a tad border. PLoS genetics 17(7), e1009691.

Wood, H.B., and Episkopou, V. (1999). Comparative expression of the mouse Sox1, Sox2 and Sox3 genes from pre-gastrulation to early somite stages. Mechanisms of development 86(1-2), 197-201.

Yang, L., Luquette, L.J., Gehlenborg, N., Xi, R., Haseley, P.S., Hsieh, C.-H., et al. (2013). Diverse mechanisms of somatic structural variations in human cancer genomes. Cell 153(4), 919-929.

Zentner, G.E., Tesar, P.J., and Scacheri, P.C. (2011). Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. Genome research 21(8), 1273-1283.

Zerbino, D.R., Wilder, S.P., Johnson, N., Juettemann, T., and Flicek, P.R. (2015). The ensembl regulatory build. *Genome biology* 16(1)**,** 1-8.

Zhang, X., Zhang, Y., Zhu, X., Purmann, C., Haney, M.S., Ward, T., et al. (2018). Local and global chromatin interactions are altered by large genomic deletions associated with human brain development. Nature communications 9(1), 1-15.

Zhou, T., Benda, C., Dunzinger, S., Huang, Y., Ho, J.C., Yang, J., et al. (2012). Generation of human induced pluripotent stem cells from urine samples. Nature protocols 7(12), 2080-2089.

Zhu, H., Shang, D., Sun, M., Choi, S., Liu, Q., Hao, J., et al. (2011). X-linked congenital hypertrichosis syndrome is associated with interchromosomal insertions mediated by a human-specific palindrome near SOX3. The American Journal of Human Genetics 88(6), 819-826.

Zhu, Z., and Huangfu, D. (2013). Human pluripotent stem cells: an emerging model in developmental biology. Development 140(4), 705-717.

CHAPTER 2

General Materials and Methods

2.1 General materials

Sterilisation of plasticware and glassware was performed using an autoclave. Unless otherwise stated, solutions were prepared in Milli-Q UltraPure water obtained from a Milli-Q EQ 7000 Ultrapure Water Purification System (Millipore). General consumables used in the present study are outlined in Table 2.1, while general equipment is listed in Table 2.2.

Table 2.1: General consumables

Product	Manufacturer
Aerosol Barrier Tips (10 µL, 10 XL, 20 µL, 200 µL, 1000 µL)	Interpath Services
High grade polypropylene pipet tips (200 µL, 1000 µL)	Greiner Bio-One
Freload tips 1000 µL	Interpath Services
Costar Stripette serological pipettes (5 mL, 10 mL, 25 mL)	Corning
1.5 mL Snapseal microcentrifuge tubes	Scientific Specialties Inc.
15 mL clear polypropylene centrifuge tubes	Corning
50 mL clear polypropylene centrifuge tubes	Corning
Costar TC-treated 6-well plates	Corning
Costar ultra-low attachment 6-well plates	Corning
Costar 35 mm TC-treated culture dish	Corning
PhenoPlate 96-well, black, optically clear flat-bottom, TC-treated microplates	Perkin Elmer
50 mL disposable reagent reservoir	Biotix
Internal thread cryogenic vials	Corning
Countess cell counting chamber slides	Invitrogen

Table 2.2: General equipment

Product	Manufacturer
Pipetman pipettes (P2, P20, P200, P1000)	Gilson
Acura 855 Multichannel Pipette 20-200 µL	Sococrex
Reax Top Vortex Mixer	Heidolph
Heracell VIOS 160i CO2 Incubator	Thermo Fisher Scientific
Heraeus Megafuge 8 Centrifuge	Thermo Fisher Scientific
Centrifuge 5417C	Eppendorf
ThermoMixer Compact 5350	Eppendorf
Centrifuge 5415D	Eppendorf
VACUSAFE Vacuum Aspiration System	Integra Biosciences
NanoDrop ND-1000 Spectrophotometer	Thermo Fisher Scientific
Countess II Automated Cell Counter	Thermo Fisher Scientific

2.2 Ethics statement

The protocols utilised in the current study have been approved by the Sydney Local Health District Human Ethics Review Committee (2019/ETH07839). Informed consent was obtained from study participants.

2.3 Tissue culture

Tissue culture was performed using a class II biosafety cabinet (Gelaire) and aseptic techniques. All cell lines were maintained at 37 ℃ in humidified air and 5% CO_2. Cells were cryopreserved in CryoStor CS10 cryopreservation media (STEMCELL Technologies) following the manufacturer's guidelines. For long term storage, frozen cells were placed in liquid nitrogen.

2.3.1 iPSC lines

An overview of the iPSC lines utilised in this thesis is provided in Table 2.3.

Table 2.3: iPSC lines used in this thesis

Cell ID	Disease Status	Sex
CMTX3 P1 (2 clones)	CMTX3	Male
CMTX3 P2	CMTX3	Male
Control 1	Control	Male
Control 2	Control	Male
Control 3	Control	Male

2.3.2 iPSC maintenance

iPSC were cultured in 6-well plates (Corning) coated with Matrigel Growth Factor Reduced Basement Membrane Matrix (Corning) (0.167 mg/mL). Cells were grown in TeSR-E8 medium (STEMCELL Technologies), which was replaced daily. Each day, iPSC morphology was examined using an EVOS XL Core Imaging System (Thermo Fisher Scientific). Cells were deemed ready for passage once the iPSC appeared as large, densely packed colonies with optimal cellular morphology (i.e. round, high nucleus:cytoplasm ratio) (Supplementary Figure 2.1). iPSC were passaged as small clumps (~50 – 200 µm) using an enzyme-free passaging protocol. Briefly, cells were washed with $DPBS^{(-Ca\ -Mg)}$ (Gibco) and then incubated in 0.5 mM UltraPure EDTA (Invitrogen) at room temperature for ~5-8 min. The EDTA was aspirated, cells were collected in DMEM/F-12 (Gibco), and pelleted by centrifugation at 200 x g for 4 min. After aspirating the supernatant, cells were resuspended in TeSR-E8 and re-seeded onto Matrigel-coated 6-well plates. Cells were maintained at 37 ℃ in humidified air and 5% CO_2.

2.3.3 Differentiation of iPSC into sMN

The differentiation protocol used during this candidature is based on the protocol published by Du et al., (2015), which utilises a combination of small molecules to generate spinal motor neurons (sMN) from human iPSC [69]. Minor modifications to this protocol have been incorporated by the Northcott Neuroscience Laboratory to improve the purity of sMN cultures, as outlined in Cutrupi et al., (2023) [67]. A graphical overview of the differentiation protocol is provided in Figure 2.1.

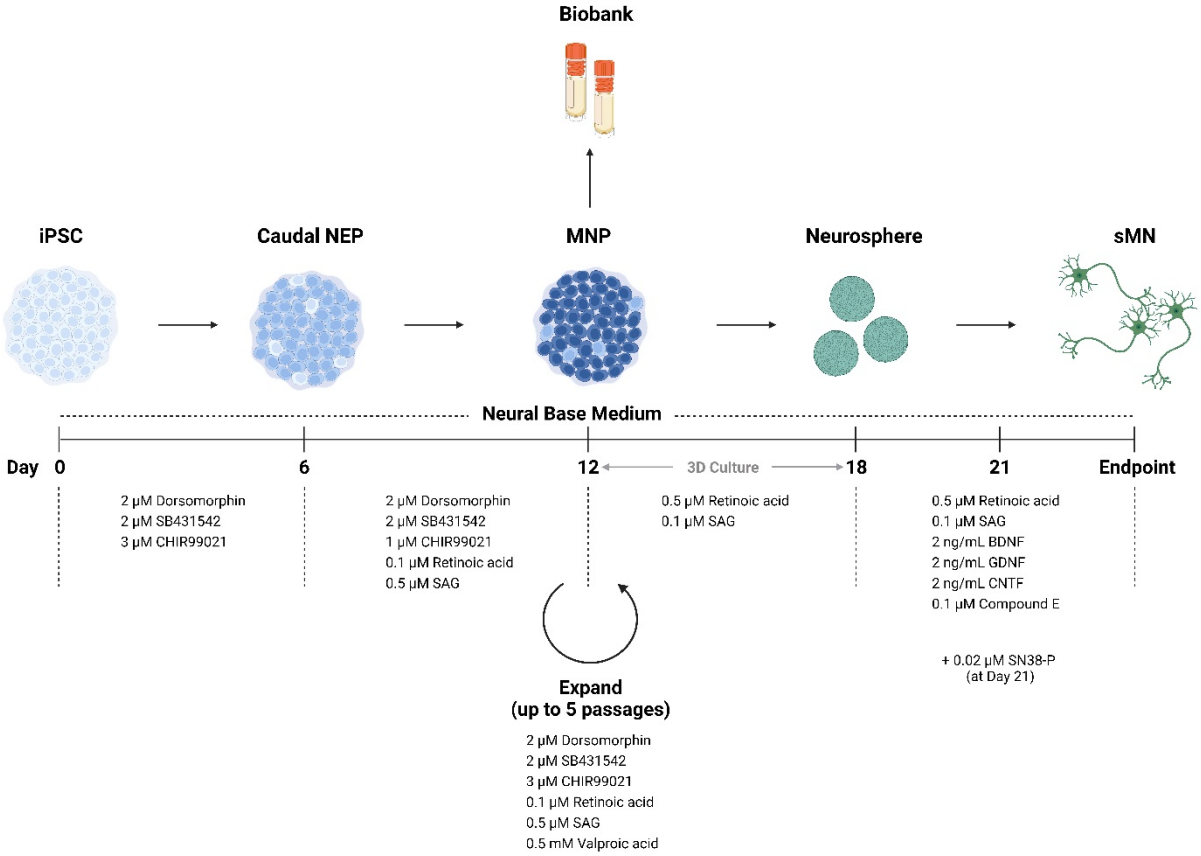


Figure 2.1: Differentiation of iPSC into sMN. iPSC are cultured for 6 days in the presence of dual SMAD inhibitors (dorsomorphin and SB431542) and a WNT activator (CHIR99021) to generate caudalised NEP. NEP are further specified to produce ventrally patterned MNP through the addition of retinoic acid and SAG to the culture media. MNP may be cryopreserved or expanded for up to 5 passages. To generate sMN, MNPs are dissociated and cultured in suspension for 6 days in the presence of retinoic acid and SAG, forming 3D neurospheres. Neurospheres are then dissociated into single cells and maintained under 2D culturing conditions. sMN are matured in media supplemented with retinoic acid, SAG, neurotrophic factors, and the NOTCH inhibitor, compound E. The addition of SN38-P at day 21 decontaminates the neuronal culture of remaining proliferative progenitors. iPSC; induced pluripotent stem cells, NEP; neuroepithelial progenitors, MNP: motor neuron progenitors, sMN: spinal motor neurons, SAG: smoothened agonist, BDNF: brain-derived neurotrophic factor, GDNF: glial cell-derived neurotrophic factor, CNTF: ciliary neurotrophic factor.

iPSC were cultured on 6-well Matrigel-coated plates as previously described. iPSC were maintained for 3-4 days in TeSR-E8 until colonies displayed an optimal cellular morphology. Media was then replaced with Neural Base Medium (NBM) (Table 2.4) supplemented with 2 µM SB431542 (STEMCELL Technologies), 3 µM CHIR99021 (Sigma Aldrich), and 2 µM Dorsomorphin (STEMCELL Technologies). Cells were cultured under these conditions for a total of 6 days to induce the formation

of caudalised neuroepithelial progenitors (NEP). Media was changed after the first 24 h and then every 48 h thereafter.

Table 2.4: Composition of Neural Base Medium (NBM)

Reagent	Concentration	Manufacturer
DMEM/F-12, GlutaMAX supplement	50%	Gibco
Neurobasal Medium	50%	Gibco
N-2 Supplement (100X)	0.5X	Gibco
B-27 Supplement (50X)	0.5X	Gibco
L-ascorbic acid, suitable for cell culture	0.1 mM	Sigma Aldrich
Penicillin/Streptomycin	100 U/mL	Gibco

NEP were dissociated using Dispase (STEMCELL Technologies). Briefly, Dispase was prepared to a final concentration of 1 U/mL by diluting in DMEM/F-12 (Gibco). The solution was filtered using a sterile 0.22 µM Millex-GP syringe filter unit (Millipore) attached to a sterile 10 mL Luer lock syringe (Terumo). NEP were washed with $DPBS^{(-Ca\ -Mg)}$ (Gibco) and incubated in 1 mL Dispase solution for 1 min at room temperature. The solution was then aspirated to leave a thin film and cells were incubated at 37℃ for 5-6 min. Dispase was aspirated and cells were collected by washing the plate with a 1:1 (v/v) solution of Neurobasal Medium (Gibco) and DMEM/F-12 GlutaMAX Supplement (Gibco). Aggregates were centrifuged at 200 x g for 4 min. The supernatant was carefully aspirated and cells were resuspended in NBM supplemented with 1 µM CHIR99021, 2 µM Dorsomorphin, 2 µM SB431542, 0.1 µM retinoic acid (RA; Sigma Aldrich), and 0.5 µM Smoothened Agonist (SAG; STEMCELL Technologies). The resuspended NEP were seeded at 1:6 on 6-well Matrigel-coated plates. Cells were maintained in this media for a total of 6 days to induce differentiation into motor neuron progenitors (MNP), with media changed after the first 24 h and every 48 h thereafter.

At this point in the protocol, MNP can be either (a) cryopreserved, (b) expanded, or (c) differentiated into sMN. Cryopreservation was carried out as described in Section 2.3. For expansion of MNP, cells were first dissociated using Dispase as outlined above, with a modification to increase the enzyme concentration to 3 U/mL to improve detachment of this highly clumped and adherent cell type. Following cellular detachment, MNP clumps were washed with $DPBS^{(-Ca\ -Mg)}$, pelleted, and re-plated onto 6-well Matrigel-coated plates at a seeding density of 1:6. MNP cultured for expansion were

maintained in MNP Expansion Media, comprised of NBM supplemented with 3 µM CHIR99021, 2 µM Dorsomorphin, 2 µM SB431542, 0.1 µM RA, 0.5 µM SAG, and 0.5 mM valproic acid (Sigma Aldrich). MNPs can be maintained for 5 passages in this defined media without notable loss of OLIG2 expression [69].

For terminal differentiation of MNP into sMN, MNP were detached from the plate as described above and resuspended in Neuronal Medium, comprised of NBM supplemented with 0.5 µM RA and 0.1 µM SAG. The cell solution was then repeatedly passed through a P1000 pipette tip to gently dissociate the large cell aggregates into smaller, homogenous clusters. Cells were then re-plated in ultra-low attachment 6-well plates (Corning) and cultured in suspension as 3D neurospheres for a total of 6 days. Media was changed after the first 24 h and every 48 h thereafter. Media changes were performed by collecting neurospheres in 15 mL falcon tubes and centrifuging at 100 x g for 1 min under reduced acceleration and deceleration velocities.

Neurospheres were then dissociated into a single cell suspension by incubating in a 1:1 (v/v) mixture of Accumax (Invitrogen) and 0.25% trypsin-EDTA (Gibco). Briefly, media was removed and cells washed in 5 mL $DPBS^{(-Ca\ -Mg)}$. Cells were then incubated in the enzyme solution for ~15 min at room temperature with regular agitation, after which time, cells were repeatedly passed through a P1000 pipette tip to obtain a single cell suspension. NBM (5 mL) was added to the solution and cells were centrifuged at 200 x g for 4 min. Supernatant was aspirated, cells were washed in 5 mL of $DPBS^{(-Ca\ -Mg)}$ and centrifuged at 200 x g for 4 min. DPBS was aspirated and the cell pellet resuspended in Neuronal Maturation Media (NMM); NBM supplemented with 0.5 µM RA, 0.1 µM SAG, 2 ng/mL brain-derived neurotrophic factor (BDNF; Thermo Fisher Scientific), 2 ng/mL glial cell-derived neurotrophic factor (GDNF; Thermo Fisher Scientific), 2 ng/mL ciliary neurotrophic factor (CNTF; STEMCELL Technologies), and 0.1 µM Compound E (STEMCELL Technologies). Neurons were counted and assessed for viability using a Countess II Automated Cell Counter (Thermo Fisher Scientific), according to the manufacturer's guidelines. Cells were then seeded onto either 96-well or 6-well Matrigel-coated plates depending on end-point experiment requirements. Neurons were left to settle for ~72 h prior to the first media

change to prevent physical disruption of the fragile neuronal network. Media was then replaced with NMM supplemented with 0.02 µM SN38-P; a compound used to remove contaminating iPSC and neural progenitor populations from iPSC-derived neuron cultures [70]. Media was replaced every 2 days.

2.4 RNA extraction

Cells were cultured in 6-well plates as previously described. Adherent cells were detached from the plate using the appropriate dissociation reagent for each cell type, i.e. 0.5 mM EDTA for iPSC (Section 2.3.2), 1 U/mL Dispase for NEPs, and 3 U/mL Dispase for MNP (Section 2.3.3). sMN were manually detached by gentle tapping of the culture vessel. Cells were then pelleted by centrifugation at 200 x g for 4 min (iPSC, NEP, MNP and sMN), or 100 x g for 1 min with reduced acceleration and deceleration velocities (neurospheres). Supernatant was aspirated and cells were snap frozen in liquid nitrogen. Frozen cell pellets were temporarily stored at -80℃.

In preparation for RNA isolation, all equipment and surfaces were decontaminated with RNaseZap (Sigma Aldrich). Total RNA was then extracted using the RNeasy Mini Kit (Qiagen), following manufacturer's guidelines. Briefly, samples were lysed in Buffer RLT (350 µL for < 5 x 10^6 cells, or 600 µL for > 5 x 10^6 cells) and homogenised by repeatedly passing through a sterile Agani hypodermic 20G needle (Terumo) fitted to a sterile 10 mL Luer lock syringe (Terumo). An equal volume of 70% (v/v) ethanol, prepared in UltraPure DNase/RNase-free distilled water (Invitrogen), was added to the homogenised lysate and thoroughly mixed by repeated pipetting. The resulting solution was placed in an RNeasy spin column fitted with a 2 mL collection tube and centrifuged for 15 s at 8,000 x g. After discarding the flow-through, Buffer RW1 (700 µL) was added to the column, which was subsequently centrifuged for 15 s at 8,000 x g. The flow-through was discarded, Buffer RPE (500 µL) was added to the column, and it was centrifuged for 15 s at 8,000 x g. The flow-through was discarded and Buffer RPE (500 µL) was again added to the column, however, centrifugation was performed for 2 min at 8,000 x g to wash and dry the membrane. Flow-through waste was discarded. To ensure minimal

carryover of buffers or flow-through waste, the empty spin column was then placed in a new 2 mL collection tube and centrifuged at 8,000 x g for 1 min. The column was then transferred to a sterile 1.5 mL collection tube and 30 µL of RNase-free water was added directly onto the spin column membrane. The RNA was then eluted from the column by centrifuging at 8,000 x g for 1 min. The concentration and purity of the resulting RNA was assessed using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific). RNA samples were stored at -80˚C.

2.5 General immunofluorescence

Cells were cultured as previously described (Section 2.3), in black, optically clear PhenoPlate 96-well microplates (Perkin Elmer) coated in Matrigel. Culture media was aspirated, and cells were fixed by incubating in 4% (w/v) paraformaldehyde (PFA; Sigma Aldrich) for 20 min at room temperature. PFA was aspirated and cells were washed in DPBS (Gibco). Permeabilisation was performed by incubating samples in DPBS containing 0.2% (v/v) Triton X-100 (Calbiochem) for 30 min at room temperature. Samples were then blocked in DPBS containing 5% (w/v) bovine serum albumin (Sigma Aldrich) (i.e. blocking buffer) for 1 h at room temperature. Primary antibodies were diluted in blocking buffer according to the concentrations outlined in Table 2.5. Cells were incubated in primary antibody overnight at 4˚C and then washed twice with DPBS. Samples were then incubated in Alexa-Fluor conjugated secondary antibodies for 1 h at room temperature. Table 2.5 lists all antibodies used in the current study. Cells were then washed with DPBS and counterstained with 4',6-diamidino-2-phenylindole (DAPI; Invitrogen), which was diluted in blocking buffer to achieve a final concentration of 0.2 µg/mL. Following a 10 min incubation period, cells were washed with DPBS and then placed in mounting media comprised of 75% (v/v) glycerol (ICN Biomedicals, Inc) and 20 mM Tris (AppliChem GmbH).

Images were acquired as a Z-stack (20X) with a Leica SP8 confocal microscope (Leica Microsystems) using Leica Application Suite X (LAS X) software (Leica Microsystems). Images were processed using FIJI (version 2.14.0) [71] to convert Leica Image Files (LIF) into Tag Image File Format

(TIFF) for downstream analysis. This was achieved by splitting each LIF file into its individual colour channels, which were then used to create max-intensity Z-projections. Additional details on image processing and downstream analysis are provided in Chapter 3, Section 3.2.9.

Table 2.5: Antibodies for immunofluorescence experiments

Antibody	Dilution	Company	Catalogue Number
Anti-SOX2	1:400	Cell Signalling	3579
Anti-NANOG	1:400	Cell Signalling	4903
Anti-OCT-4A	1:400	Cell Signalling	2840
Anti-OLIG2	1:100	Millipore	MABN50
Anti-MNX1	1:500	Atlas Antibodies	HPA071717
Anti-TUBB3	1:1000	Sigma Aldrich	T2220
Goat anti-rabbit IgG Alexa Fluor 488	1:500	Invitrogen	A11008
Goat anti-rabbit IgG Alexa Fluor 555	1:500	Invitrogen	A21428
Goat anti-mouse IgG Alexa Fluor 488	1:500	Invitrogen	A11001
Goat anti-mouse IgG Alexa Fluor 555	1:500	Invitrogen	A21422

2.6 NanoString nCounter targeted gene expression analysis

A custom NanoString nCounter Elements TagSet panel was designed to detect 48 RNA targets (Supplementary Table 2.1), including 21 genes within and flanking the CMTX3 linkage region, 1 gene contained within the CMTX3 insertion sequence, 11 genes implicated in PNS differentiation, and 14 housekeeping genes. Housekeeping genes were selected in accordance with the NanoString technical guidelines [72] which recommend that a combination of genes with low, medium, and high expression levels be included within each custom panel. Probe sequences were designed by NanoString Technologies (Washington, USA) and synthesised by Integrated DNA Technologies (Iowa, USA).

Total RNA was extracted as previously described (Section 2.4) and quantified using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific). RNA samples were diluted in nuclease-free water to a final concentration of 14.286 ng/µL, to fulfil the sample input requirements for the assay (i.e. 100 ng of RNA in 7 µL volume). Samples were assessed with the TapeStation system (Agilent) to confirm RNA integrity. NanoString nCounter Elements gene expression assays were performed by the Westmead Institute for Medical Research (Sydney, Australia) according to the recommended protocol

[73]. Briefly, RNA samples were mixed with sequence-specific oligonucleotide probes, biotinylated capture probes, and reporter probes containing target-specific fluorescence barcodes (Figure 2.2A). Hybridisation was performed by incubating the solution for 16 hr at 65 ℃. Excess probes were then removed from the solution using the nCounter preparation station. Purified probe:target complexes were immobilised and electrostretched onto the imaging surface. Fluorescence barcodes were imaged and quantified using the nCounter Digital Analyzer (Figure 2.2B). NanoString internal positive and negative controls are included in each assay for quality control.

Raw data was analysed using nSolver Analysis software (version 4.0) (NanoString Technologies, Inc); a customised software package specifically designed for analysis of nCounter data. Normalisation and statistical testing was performed according to manufacturer's guidelines [74]. This involves normalising to both positive spike-in control probes as well as the endogenous housekeeping genes. For analyses involving multiple different cell types, housekeeping expression levels were found to be quite variable. In these cases, a selection of housekeeping genes showing minimal variability were utilised in the analysis. A Welch's t-test was used for statistical testing of differential gene expression between patient and control samples, using the log2-transformed normalised count data as input and a significance threshold of $p < 0.05$.

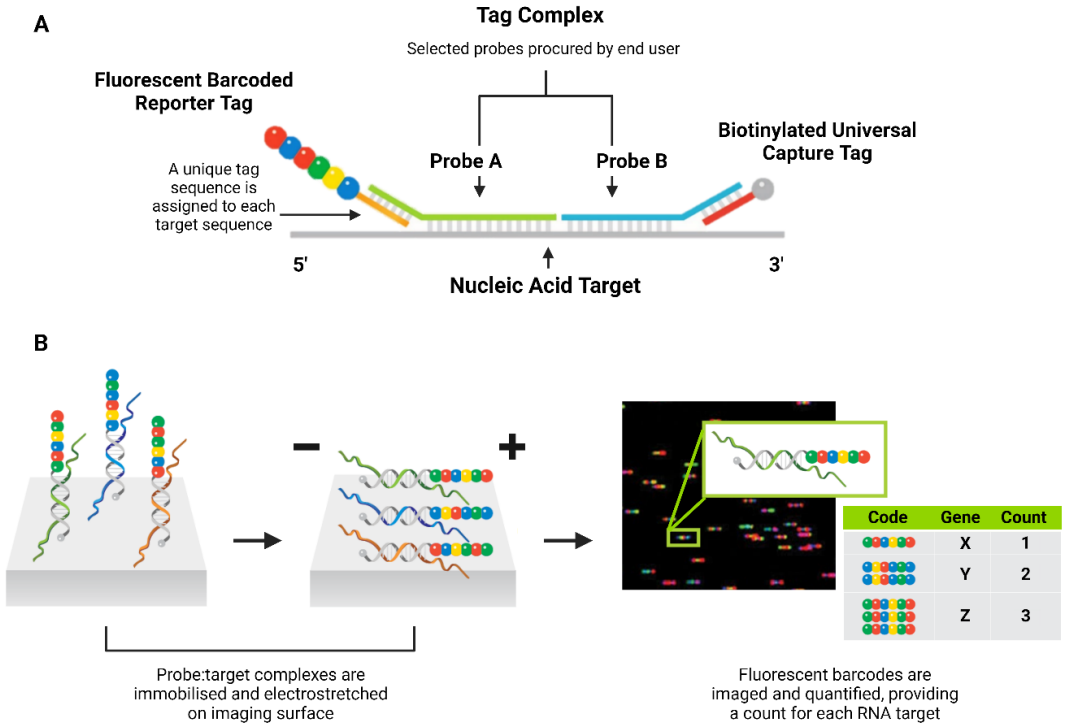
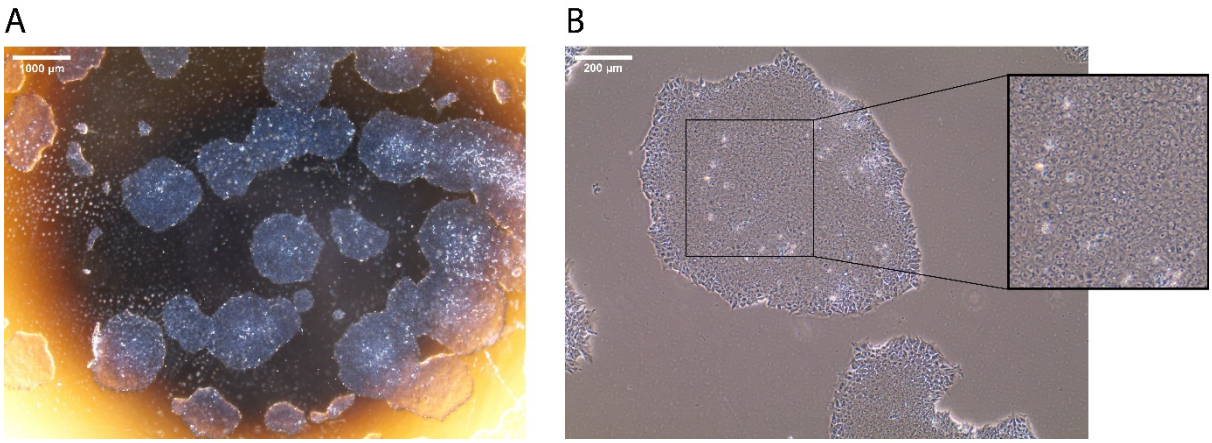


Figure 2.2: Overview of NanoString nCounter gene expression assay. (A) Two sequence-specific oligonucleotide probes (Probe A and Probe B) bind to each RNA target. Probe A hybridises to a reporter probe, which is tagged with a fluorescent barcode. Each RNA target is assigned a unique fluorescent barcode. Probe B hybridises to a capture tag which is labelled with biotin. Diagram adapted from NanoString Technologies, Inc [75]. (B) The biotinylated capture probe allows for immobilisation of the probe:target complex onto the imaging surface, which is coated with streptavidin. Complexes are electrostretched on the imaging surface and then a fluorescence microscope within the nCounter Digital Analyzer images the fluorescent barcodes, allowing for absolute quantification of each RNA target of interest. Images adapted from NanoString Technologies, Inc [76, 77].

Supplementary Material: Chapter 2



Supplementary Figure 2.1: Brightfield microscope images showing optimal morphology of iPSC (A) Healthy iPSC grow as tightly packed, round colonies with clearly defined borders. (B) When viewing images at higher magnification, characteristic features of iPSC can be clearly observed, including the dense packing of cells within colonies and the high nucleus to cytoplasm ratio within individual cells. Nuclei appear as dark blue spots within the circular cells.

Supplementary Table 2.1: List of genes targeted in the custom NanoString nCounter Elements TagSet panel.

Gene Target	Class	Isoform coverage (#hit/#total)
ARHGAP39*	Endogenous	6/6
ARHGAP39*	Endogenous	6/6
ATP11C	Endogenous	18/18
CDR1	Endogenous	1/1
CHAT	Endogenous	7/7
CXorf66	Endogenous	1/1
EGR2	Endogenous	6/6
F9	Endogenous	3/3
FGF13	Endogenous	8/8
FGF13-AS1	Endogenous	1/1
GAP43	Endogenous	3/3
ISL1	Endogenous	2/2
LDOC1	Endogenous	1/1
LINC00632	Endogenous	1/3
LOC389895	Endogenous	1/1
LOC645188	Endogenous	1/1
LOC728660	Endogenous	1/1
MAGEC1	Endogenous	2/2
MAGEC2	Endogenous	1/1

MAGEC3	Endogenous	4/4
MCF2	Endogenous	12/12
MNX1	Endogenous	2/2
MPZ	Endogenous	3/3
NES	Endogenous	1/1
NOTCH1	Endogenous	2/2
OLIG2	Endogenous	2/2
PMP22	Endogenous	9/9
SOX2	Endogenous	1/1
SOX3	Endogenous	1/1
SPANX family**	Endogenous	
SPANXA2-OT1	Endogenous	1/1
SPANXN4	Endogenous	1/3
SRD5A1P1	Endogenous	1/1
ZIC3	Endogenous	2/2
CLTC	Housekeeping	2/2
G6PD	Housekeeping	3/3
GUSB	Housekeeping	16/16
HPRT1	Housekeeping	1/1
PGK1	Housekeeping	1/1
POLR1B	Housekeeping	10/10
POLR2A	Housekeeping	1/1
RPL30	Housekeeping	1/1
RPLP0	Housekeeping	2/2
SDHA	Housekeeping	8/8
TBP	Housekeeping	2/2
TFRC	Housekeeping	10/10
TUBB	Housekeeping	7/7
YWHAZ	Housekeeping	11/13

* Two probes for ARHGAP39 were included in the custom panel; one probe targeting sequence contained within the CMTX3 SV, and one probe located outside of this region.
**High sequence homology between SPANX family gene members prohibited unique probe design. A singular probe was designed to target SPANXA1, SPANXA2, SPANXB1, SPANXC and SPANXD @ >95%.

CHAPTER 3

A Neuronal Model for CMTX3 Neuropathy

3.1 Introduction

The genetic cause of CMTX3 neuropathy was initially reported in 2016 [63], however, the pathomechanism of this complex mutation remains unknown. This knowledge gap ultimately stems from the lack of appropriate model systems for investigating CMTX3 neuropathy. To date, patient-derived lymphoblasts have been the only model available for investigating this disease. Lymphoblasts provide an accessible source of patient material harbouring the complex SV, however there are clear limitations in their capacity to model a phenotype that exclusively affects peripheral nerves. Nonetheless, previous research using CMTX3 patient lymphoblasts has shown promising evidence to support the hypothesis that spatiotemporal gene dysregulation may be an underlying mechanism for this neuropathy [63]. These previous studies reinforce the importance of obtaining a disease-relevant model of CMTX3, as spatiotemporal gene regulation is intricately dependent upon the appropriate cell type being investigated in the disease context.

Modelling neurological diseases is inherently challenging due to the ethical and technical difficulties in accessing human neural tissue. Historically, nerve biopsies have provided valuable insights into the pathology of CMT [78-80], however, these invasive procedures are rarely performed in the current IPN research setting. Post-mortem tissue represents an alternative source of human neuronal material and has provided key insights into the pathology of many neurodegenerative

conditions [81] such as Alzheimer's Disease (e.g. [82]), Parkinson's Disease (e.g. [83]), and Huntington's Disease (e.g. [84]) among others. However, this is not a feasible option for CMTX3 given the ultra-rare nature of this disease.

Transgenic animal models have been widely adopted for investigating various forms of CMT neuropathy (reviewed in [85, 86]). Mutant animals often display molecular and clinical phenotypes that mimic those reported in CMT patients, providing a preclinical model for investigating disease mechanisms *in vivo* and testing therapeutic approaches [19]. For example, the CMT1A rat model, harbouring three additional copies of the mouse *Pmp22* gene, shows histological features of CMT, such as segmental demyelination and onion-bulb formation [87]. This is accompanied by phenotypic deficits including unsteady gait and reduced muscle strength, as well as reduced nerve conduction velocities, which is a hallmark of CMT type 1 [87]. However, as with any model system, there are caveats and limitations originating from the innate biological differences between animals and humans [19]. For example, the slowly progressing, length-dependent axonal subtypes of CMT have been particularly difficult to model [88], with reports of mutant mice showing less severe phenotypes than patients [89, 90]. The shorter lifespan and reduced axonal length of these animals, among other factors, have been suggested to contribute to these phenotypic discrepancies [90]. CMT has also been modelled in Drosophila (e.g. [91, 92], zebrafish (e.g. [93-96]) and C. elegans (e.g. [97-99], which can offer substantial economic and temporal advantages compared to murine models. However, the suitability of these model organisms varies depending upon the evolutionary conservation of the candidate gene being investigated. Further, Drosophila and C. elegans do not harbour myelinated axons and are therefore not appropriate for modelling demyelinating neuropathies [100]. Despite the clear utility of these types of models, transgenic animals are not a practical option for modelling CMTX3 neuropathy, as the causative SV is difficult to recreate with traditional cloning technologies. Recent advances in CRISPR/Cas gene editing technologies have improved strategies for modelling SV, such as the CRISVar (CRISPR/Cas-induced structural variants) method [101], which has been used for inducing deletions, inversions, and duplications at targeted genomic loci [101-103]. Whilst this is promising, certain types

of complex SV cannot be efficiently engineered with such approaches [104]. However, with the ever-expanding repertoire of CRISPR/Cas technologies (e.g. [105], this may become a feasible option.

The rise and increasing utilisation of induced pluripotent stem cells (iPSC) [106] has provided a powerful alternative to the use of animal models for modelling IPN. With the capacity to be differentiated into a range of specialised cell-types, iPSC models are providing unprecedented access to disease-relevant neuronal tissue for investigating disease mechanisms in the human context. A major advantage of iPSC derived from patients is being able to retain the specific genetic background, which is especially attractive for investigating complex SV [67, 107]. iPSC can be differentiated into the major cell types of the PNS, with protocols available for generating sMN [69] sensory neurons [108] and Schwann cells [109]. These cells can be generated in isolation, or as part of more complex 3D organoids comprised of multiple PNS cell populations [110]. Accordingly, this breadth of iPSC technologies has been rapidly utilised within the CMT field (e.g. [67, 110-117]). Saporta and colleagues (2015) were the first group to report an iPSC-based model of CMT, using sMN derived from CMT2A and CMT2E patients [113]. Abnormalities in mitochondrial trafficking, as well as neuronal excitability, were observed in sMN derived from both patients, while CMT2E patient sMN also displayed cytoskeletal abnormalities that were mirrored in a murine model of this disease [113]. This seminal example illustrates the clear utility of iPSC technologies for establishing preclinical models of CMT. Therefore, the goal of this chapter is to generate and characterise an iPSC-derived sMN model of CMTX3 neuropathy, which will improve our capacity to investigate the molecular mechanisms driving this disease.

3.2 Methods

3.2.1 Fibroblast maintenance

Primary fibroblasts were thawed in a 37 ℃ water bath and seeded in sterile T-25 flasks (Sarstedt). Fibroblasts were maintained in DMEM (Gibco) supplemented with 10% (v/v) fetal bovine

serum (SAFC Biosciences), 1% (v/v) Penicillin/Streptomycin (Gibco) and 1% (v/v) L-glutamine (Gibco). Cells were maintained at 37 ˚C in humidified air containing 5% CO_2. Media was changed every 2-3 days.

3.2.2 Generation of iPSC

Fibroblasts from two CMTX3 patients were reprogrammed into iPSC. This service was outsourced to two external companies; FUJIFILM Cellular Dynamics, Inc (FCDI) (Wisconsin, USA) and the Stem Cell and Organoid Facility (SCOF) (Children's Medical Research Institute, Westmead, Australia). Both facilities utilised non-integrative, episomal methods of reprogramming, summarised below. All iPSC used in this study were derived from males (Chapter 2, Table 2.3). iPSC maintenance is described in Chapter 2, Section 2.3.2.

3.2.2.1 iPSC reprogramming (FCDI)

iPSC were generated from fibroblasts utilising a proprietary episomal method of reprogramming as previously described [118]. This involves transfecting the fibroblasts with oriP/EBNA-1 (Epstein-Barr virus nuclear antigen 1) based episomal vectors [119], using the Amaxa Human Dermal Fibroblasts Nucleofector Kit (Lonza). Cells were cultured in reprogramming media for 7 days and then Essential 8 (E8) medium (Gibco) for 2 weeks. Culturing was performed in Matrigel-coated plates. Resulting iPSC colonies were picked and expanded in E8. iPSC underwent immunofluorescence staining to confirm pluripotency, and cellular identity was matched to the original fibroblast line (FUJIFILM Cellular Dynamics). iPSC were confirmed to be karyotypically normal using G-banding karyotype analysis (450-500 resolution) performed by WiCell (Wisconsin, USA).

3.2.2.2 iPSC reprogramming (SCOF)

Fibroblasts were nucleofected with episomal plasmids [120]; pCXLE-hUL (Addgene plasmid #27080; http://www.addgene.org/27080/; RRID:Addgene_27080), pCXLE-hSK (Addgene plasmid #27078; http://www.addgene.org/27078/; RRID:Addgene_27078) and pCXLE-hOCT3/4-shp53-F (Addgene plasmid # 27077; http://www.addgene.org/27077/; RRID:Addgene_27077), using the Amaxa 4D nucleofection system and P2 Primary Cell 4D-Nucleofector kit (Lonza). pCXLE-hUL, pCXLE-hSK and pCXLE-hOCT3/4-shp53-F were gifts from Shinya Yamanaka. Cells were transferred to 6-well

plates coated in Geltrex (Gibco) and cultured in DMEM (Gibco) containing 10% (v/v) fetal bovine serum (i.e. fibroblast medium). After two days, media was replaced with E8 and fibroblast medium (1:1). Two days later, cells were transitioned to E8 and cultured for 3-4 weeks. iPSC colonies were picked and cultured in E8 and 10 µM Y-27632 (STEMCELL Technologies). iPSC were expanded in E8 in Geltrex-coated plates. iPSC were confirmed to be karyotypically normal using digital SNP karyotyping with the Illumina Infinium GSA-24 v3.0 microarray (Victorian Clinical Genetics Services, Melbourne, Australia). The SCOF performed in-house immunofluorescence to confirm pluripotency. Short tandem repeat cell line authentication was performed by CellBank Australia (Westmead, Australia).

3.2.3 DNA extraction

Patient and control iPSC lines were cultured in 6-well plates as previously described (Chapter 2, Section 2.3.2). iPSC underwent DNA extraction using the QuickExtract DNA Extraction Solution 1.0 (Epicenter Bio). Briefly, cells were washed in $DPBS^{(-Ca\ -Mg)}$ (Gibco), and then 500 µL of QuickExtract solution was added to each sample. The cell solution was transferred to a sterile 1.5 mL microcentrifuge tube and vortexed for 15 s. Tubes were heated to 65°C for 6 min then vortexed for 15 s. Samples were incubated at 95°C for 2 min and then stored at -20°C until required.

3.2.4 CMTX3 genotyping PCR assay

A multiplex PCR was used to genotype the CMTX3 complex insertion as previously described [63]. Validated primer sequences are listed in Table 3.1. Primers were synthesised by Sigma Aldrich. PCR reactions were comprised of; 50 ng template DNA, 1X MyTaq HS Red Mix (Bioline), 8 pmol of X-Forward primer, 8 pmol of X-Reverse primer, 2 pmol of 8-Forward primer, 4 pmol of 8-Reverse primer. Milli-Q water was used to make up the reactions to final volume of 15 µL.

Table 3.1: Primers for the CMTX3 gentotyping PCR

Primer	Sequence 5' to 3'
X-Forward	CTCCAGCTTTGTTCTTTGGAC
X-Reverse	TCACCAACATTTCCAATCTCC
8-Forward	CAAACCCAATTCAGGTCCAG
8-Reverse	GCCTAGGAGGTGTCCCTTTC

Thermocycling was carried out in a Mastercycler Pro S (Eppendorf). Initial denaturation was performed at 95°C for 15 min. This was followed by 10 cycles of a touchdown protocol consisting of (1) denaturation at 95°C for 30 s, (2) annealing starting at 60°C for 30 s and decreasing by 1.5°C per cycle, and (3) extension at 68°C for 40 s. Samples then underwent 25 cycles of (1) denaturation at 95°C for 30 s, (2) annealing at 50°C for 30 s, and (3) extension at 68°C for 40 s, followed by a final extension at 68°C for 5 min.

3.2.5 Agarose gel electrophoresis

PCR products were size fractionated using agarose gel electrophoresis. Agarose (Bioline) was dissolved in 1X TAE buffer (Astral Scientific) to prepare a 2% (w/v) gel. SYBR Safe DNA Gel Stain (Invitrogen) (1X) was added to agarose while in the molten state. Electrophoresis was performed at ~ 5 V/cm for 55 min using a multiSUB Choice Horizontal Electrophoresis System (Cleaver Scientific) connected to a Model 3000Xi Electrophoresis Power Supply (Bio-Rad). HyperLadder 100 bp (Bioline) was used as the size standard. DNA was visualised using a Safe Imager Blue-Light Transilluminator (Invitrogen) and images captured with a PhotoShot S5-IS digital camera with Hoya O(G) filter (Canon).

3.2.6 Quantitative real-time polymerase chain reaction (qPCR)

RNA was extracted from iPSC and fibroblasts using the RNeasy Mini Kit (Qiagen) according to manufacturer's guidelines. The iScript cDNA Synthesis Kit (Bio-Rad) was used to generate cDNA from an RNA template, following the recommended protocol. A probe-based qPCR was carried out using Taqman Gene Expression Assays (Applied Biosystems) to assess the expression of a range of pluripotency genes and the endogenous control gene, *GAPDH*. Taqman probes utilised in this study are listed in Table 3.2.

Table 3.2: Taqman Gene Expression Assays used to assess pluripotency of iPSC

Gene ID	Taqman Assay ID
LIN28A	Hs00702808_s1
NANOG	Hs02387400_g1
NODAL	Hs00415443_m1
FOXD3	Hs00255287_s1
CDH1	Hs00170423_m1
TDGF1	Hs02339499_g1
GAPDH	Hs99999905_m1

qPCR reactions were prepared in triplicate in a MicroAmp Fast Optical 96-well reaction plate (Applied Biosystems). Each 20 µL reaction contained 1X Taqman Gene Expression Master Mix (Applied Biosystems), 1X Taqman Gene Expression Assay, template cDNA (25 ng) and Milli-Q water. 'No Template Control' reactions were performed using Milli-Q water to replace template cDNA. Thermal cycling was carried out using a StepOnePlus Real-Time PCR System (Applied Biosystems). Initial denaturation was performed at 95°C for 10 min, followed by 40 cycles of amplification (95°C for 15 s; 60°C for 1 min). Data analysis was performed using the StepOne Plus Software (version 2.1). GraphPad Prism (version 10) was used to plot the raw Ct values, enabling a qualitative representation of the expression of pluripotency genes between iPSC and fibroblasts [111, 112].

3.2.7 NanoString nCounter analysis

NanoString nCounter gene expression assays were performed as previously described (Chapter 2, Section 2.6), using RNA samples extracted from the following timepoints; iPSC, MNP (differentiation day 12), and mature sMN (differentiation day 28). For each timepoint, RNA was extracted from 2 patient and 3 control cell lines.

3.2.8 Immunofluorescence

iPSC were cultured in Nunc Lab-Tek 8-well chamber slides (Thermo Fisher Scientific) for 2-3 days in TeSR-E8. Immunofluorescence was then performed as described in Chapter 2, Section 2.5, utilising the following primary antibodies; anti-NANOG (Cell Signalling, #4903, 1:400), anti-OCT4A (Cell Signalling, #2840, 1:400) and anti-SOX2 (Cell Signalling, #3579, 1:400). MNP were plated in optically clear PhenoPlate 96-well microplates (Perkin Elmer) coated in Matrigel. After 3 days of expansion, cells

underwent immunofluorescence staining as previously described (Chapter 2, Section 2.5), however, permeabilisation was performed for 30 min using 0.3% (v/v) Triton in DPBS to account for the high degree of clumping observed in this cell type. MNPs were incubated with the anti-OLIG2 primary antibody (Millipore, MABN50 1:100). For sMN staining, neurons were plated in optically clear, Matrigel-coated plates at a density of 30,000 cells/well. After 8 days in culture, the immunofluorescence protocol was performed as previously described (Chapter 2, Section 2.5), using the following primary antibodies; anti-MNX1 (Sigma, HPA071717, 1:500) and anti-TUBB3 (Sigma, T2220, 1:1000). Images were acquired with a Leica SP8 confocal microscope (Leica Microsystems) and LASX software (Leica Microsystems). Images were processed using Fiji (version 2.14.0) [71].

3.2.9 Quantification of MNX1 positive sMN

FIJI (version 2.14.0) was used to generate maximum intensity Z-projection images for downstream analysis. Images were obtained from 3 separate rounds of differentiation. A custom CellProfiler (version 4.2.5) [121] pipeline was used to calculate the percentage of sMN positive for MNX1. First, the 'IdentifyPrimaryObjects' module was used to identify DAPI-stained nuclei. A global minimum cross-entropy threshold method was employed. To determine the lower threshold limit, images were manually inspected using the 'pixel intensity' tool, to provide intensity values for background fluorescence. To determine the nuclei diameter range, the 'measure length' tool was used to manually assess the typical diameter of nuclei within the fluorescence images. Optimal segmentation of nuclei was achieved by de-clumping objects based upon intensity. However, the presence of small, bright DAPI speckles introduced the risk of false-positive nuclei identification, which could have significant effects on downstream analysis. To overcome this issue, the 'FindMaxima' module was used to identify high-intensity DAPI speckles, which were then processed using the 'ConvertImageToObjects' module and subsequently excluded from the image using the 'MaskObjects' function. Masking settings were optimised to retain partially masked objects if they display high overlap with the masking region. This setting minimises the risk of nuclei being discarded in cases where they co-localise with the DAPI speckles. Figure 3.1A summarises this process of speckle

identification and removal to obtain the final nuclei count. To identify the number of MNX1-positive cells, the MNX1-stained images were processed using the 'IdentifyPrimaryObjects' module, as described above. The 'RelateObjects' module was used to associate the two sets of identified objects (i.e. the DAPI objects and the MNX1 objects) with each other. Nuclei were then filtered based on the presence of MNX1-staining using the 'FilterObjects' module (Figure 3.1B). The number of MNX1-positive nuclei was divided by the total number of nuclei to calculate the proportion of sMN showing positive expression of MNX1. Proportion data underwent arcsine square root transformation prior to statistical testing. Statistical analysis was performed in GraphPad Prism (version 10) using an unpaired two-tailed t-test with a significance threshold of $p < 0.05$.

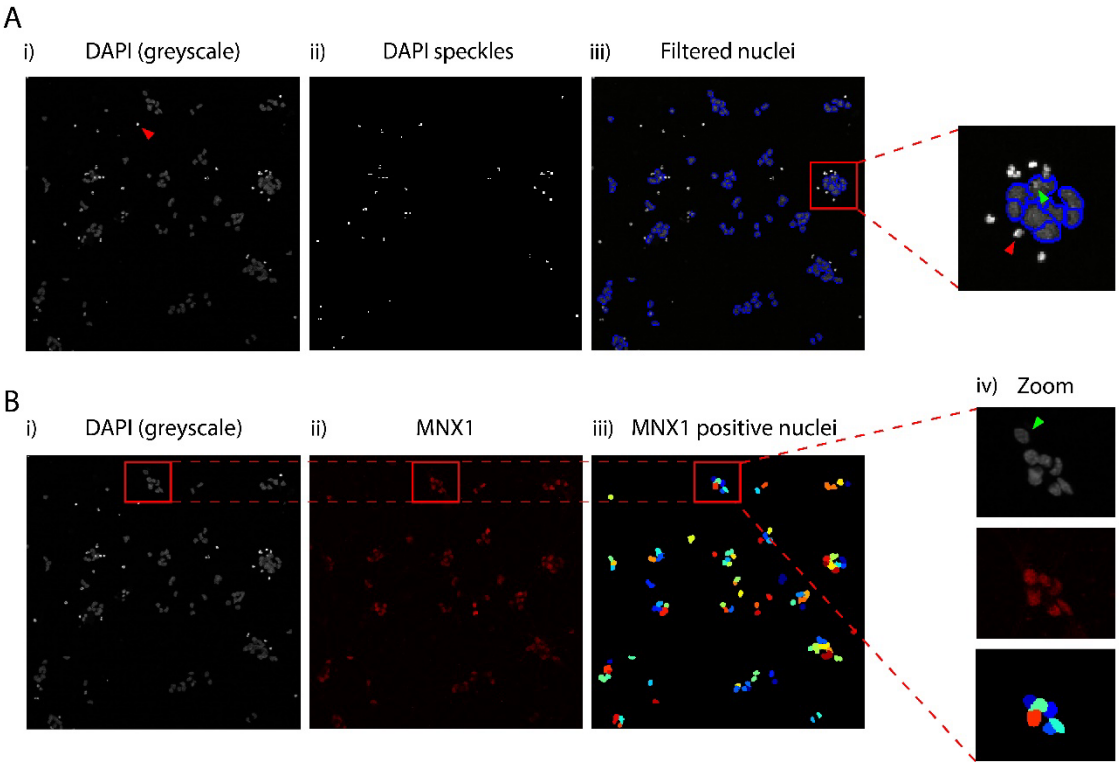


Figure 3.1: Custom CellProfiler pipeline for quantifying MNX1-positive sMN. (A) (i) sMN stained with DAPI to identify total cell nuclei. Intensely stained speckles are visible on the microscope images (e.g. red arrow) and must be excluded to minimise errors in nuclei identification and quantification. (ii) DAPI speckles were isolated based on their heightened pixel intensity. These speckles were filtered out of the image before proceeding with nuclei quantification. (iii) Image demonstrating the final cell nuclei detected by the custom pipeline, which are shown here as blue outlines overlying the original DAPI-stained image. Objects falling outside of the blue outlines (e.g. red arrow) were discarded from the analysis. Masking settings were optimised such that any speckles showing a high degree of overlap with nuclei were retained (e.g. green arrow), to prevent excluding underlying

cells from downstream analysis. (B) Representative images showing (i) DAPI staining and (ii) MNX1 staining of sMN cultures. (iii) Image showing the nuclei that were identified as MNX1 positive by the custom CellProfiler pipeline. Each coloured object represents an MNX1-positive nucleus. (iv) A zoomed in section of the three images confirms that the pipeline is successfully able to identify the nuclei that are positive and negative (green arrow) for MNX1.

3.3 Results

3.3.1 Characterisation of CMTX3 patient-derived iPSC lines

3.3.1.1 CMTX3 iPSC are karyotypically normal and harbour the pathogenic 78 kb insertion

It is well-documented that the process of reprogramming somatic cells into iPSC may result in the acquisition of chromosomal abnormalities [122-124]. Therefore, the CMTX3 patient-derived iPSC underwent karyotypic analysis to assess genomic integrity using either G-banding analysis or digital SNP karyotyping, which confirmed that both cell lines display a normal, 46, XY karyotype (Figure 3.2A). Furthermore, a genotyping PCR was performed to ensure that the ~78 kb CMTX3 SV was retained after cellular reprogramming. Using primers that span the breakpoints of the CMTX3 insertion, a previously designed multiplex PCR (Figure 3.2B-C) was used to genotype the iPSC samples [63]. As seen in Figure 3.2D, the presence of the CMTX3 insertion was confirmed in the male patient-derived iPSC lines. As expected, the CMTX3 structural variant was absent from the three control iPSC lines used in this study (Figure 3.2D).

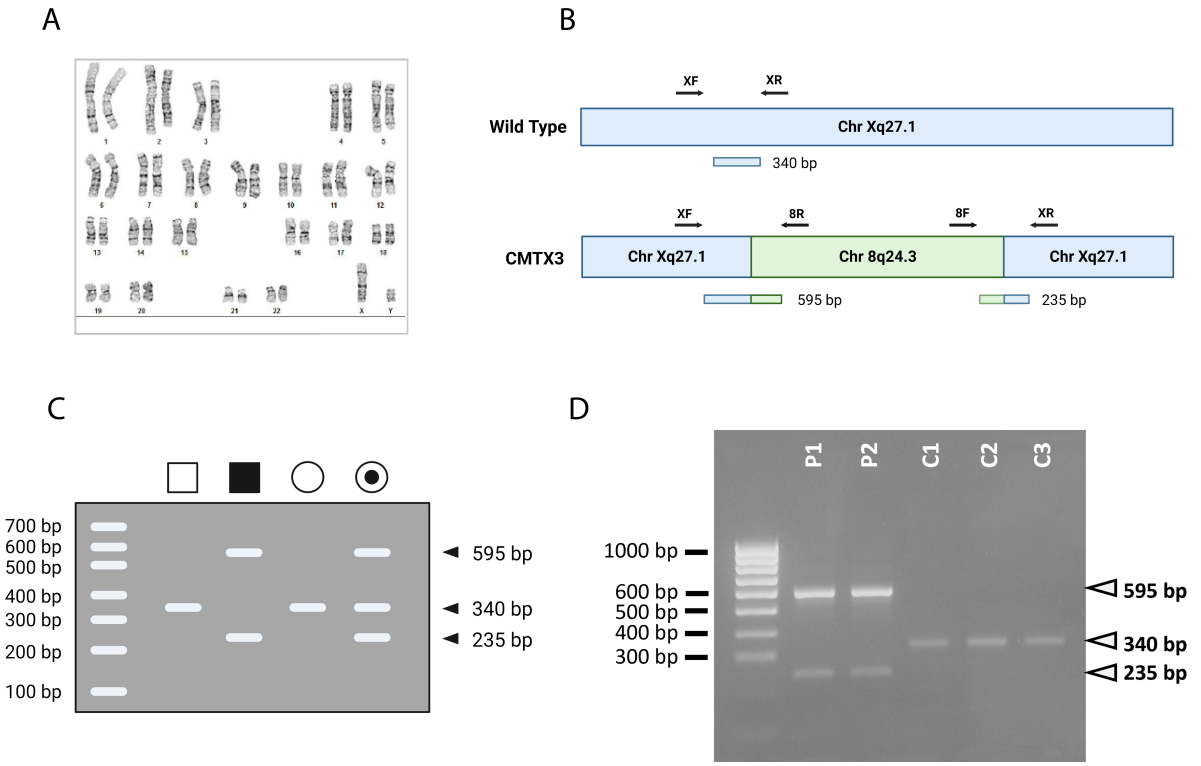


Figure 3.2: Genetic characterisation of CMTX3 patient-derived iPSC lines. (A) G-banding analysis performed on one of the CMTX3 iPSC lines confirms a normal 46, XY karyotype. The second iPSC line was confirmed as karyotypically normal using digital SNP karyotyping (data not shown). (B) The location of the primers (F = forward, R = reverse) used in the CMTX3 multiplex genotyping PCR assay, and the sizes of the PCR amplicons from either the wild type or CMTX3 X-chromosomes. Adapted from Brewer et al., (2016) [63]. (C) Representation of the possible amplicons generated by the CMTX3 multiplex PCR genotyping assay; unaffected male (open square), affected male (shaded square), unaffected female (open circle), female carrier (dotted circle). Adapted from Brewer et al., (2016) [63]. (D) CMTX3 genotyping assay confirms that patient-derived iPSC lines (P1-2) display two amplicons (595 bp and 235 bp) which span the proximal and distal breakpoints of the CMTX3 complex insertion, respectively. Control iPSC lines (C1-3) show a single PCR product (340 bp) arising from amplification of the wild type X-chromosome. Size ladder (lane 1) HyperLadder 100 bp (Bioline) DNA ladder.

3.3.1.2 CMTX3 iPSC lines express markers of pluripotency

To validate the pluripotency of the patient-derived iPSC lines, the expression of pluripotency markers were assessed at both the RNA and protein level using RT qPCR and immunofluorescence respectively. Compared to fibroblasts, patient-derived iPSC expressed higher levels of the six pluripotency-associated genes (*LIN28A, NANOG, NODAL, FOXD3, CDH1,* and *TDGF1*), as observed by lower Ct values (Figure 3.3A). In contrast, the expression level of the housekeeping gene, *GAPDH*, was

comparable between both cell types (Figure 3.3A). Data has been represented as raw Ct values rather than relative gene expression, as this calculation was not always possible due to undetectable expression levels of pluripotent genes in fibroblasts. Immunofluorescence staining confirmed that patient-derived iPSC lines express the pluripotency proteins, SOX2, NANOG and OCT4 (Figure 3.3B).

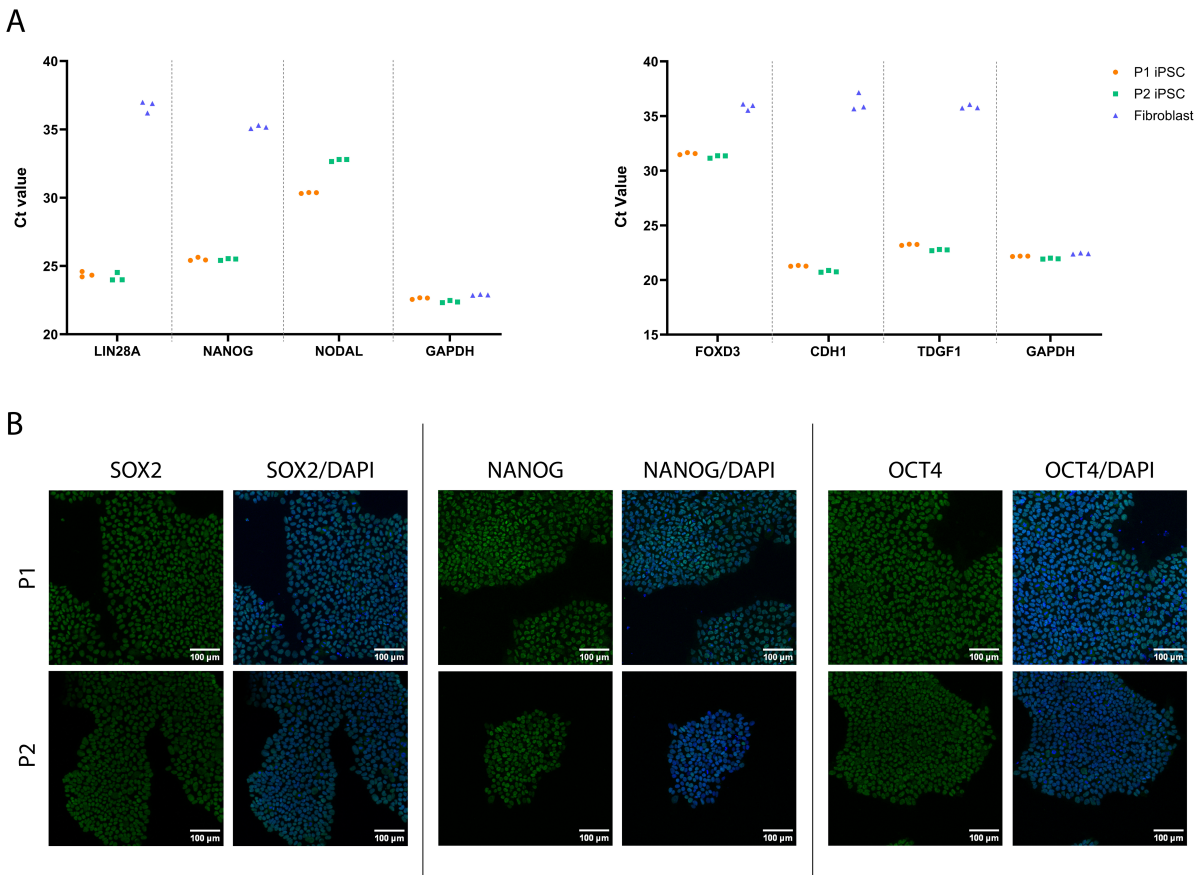


Figure 3.3: CMTX3 patient-derived iPSC express markers of pluripotency at the mRNA and protein level. (A) CMTX3 patient-derived iPSC lines (P1, P2) show increased expression levels of the pluripotency genes, *LIN28A*, *NANOG*, *NODAL*, *FOXD3*, *CDH1* and *TDGF1* compared to fibroblasts, as measured using RT qPCR. Note *NODAL* was undetectable in fibroblasts. *GAPDH* expression level is included as an endogenous control. (B) Immunofluorescence staining demonstrates that the two patient-derived iPSC lines express the common molecular markers of pluripotency, SOX2, NANOG and OCT4 (green). DAPI staining (blue) is overlayed to identify nuclei. Scale bar = 100 µm.

3.3.2 iPSC can be differentiated into OLIG2-positive motor neuron progenitors

Validated iPSC lines underwent directed differentiation into sMN to establish a disease-relevant cellular model of CMTX3. The differentiation protocol, originally published by Du et al., (2015) [69], utilises a combination of small molecules to promote the induction and caudalisation of neuroepithelial progenitors, which are then further specified to produce OLIG2-positive MNP. OLIG2 is a transcription factor expressed in the motor neuron progenitor domain (pMN) of the developing spinal cord, which ultimately gives rise to spinal motor neurons [125, 126]. The MNP generated by this protocol can either be further differentiated into mature sMN, or undergo expansion and cryopreservation for approximately five passages to generate frozen stocks of this intermediate cell type [69]. Immunofluorescence staining confirmed that the patient and control MNP lines used in this study express the MNP marker, OLIG2 (Figure 3.4A, Supplementary Figure 3.1). Similarly, NanoString gene expression data demonstrated a marked increase of OLIG*2* mRNA expression levels in MNP compared to undifferentiated iPSC lines, which do not express detectable levels of this gene (Figure3.4B(i)). Statistical analysis of NanoString gene expression data reveals similar levels of *OLIG2* expression between patient and control MNP ($p > 0.05$) (Figure 3.4B(ii)).

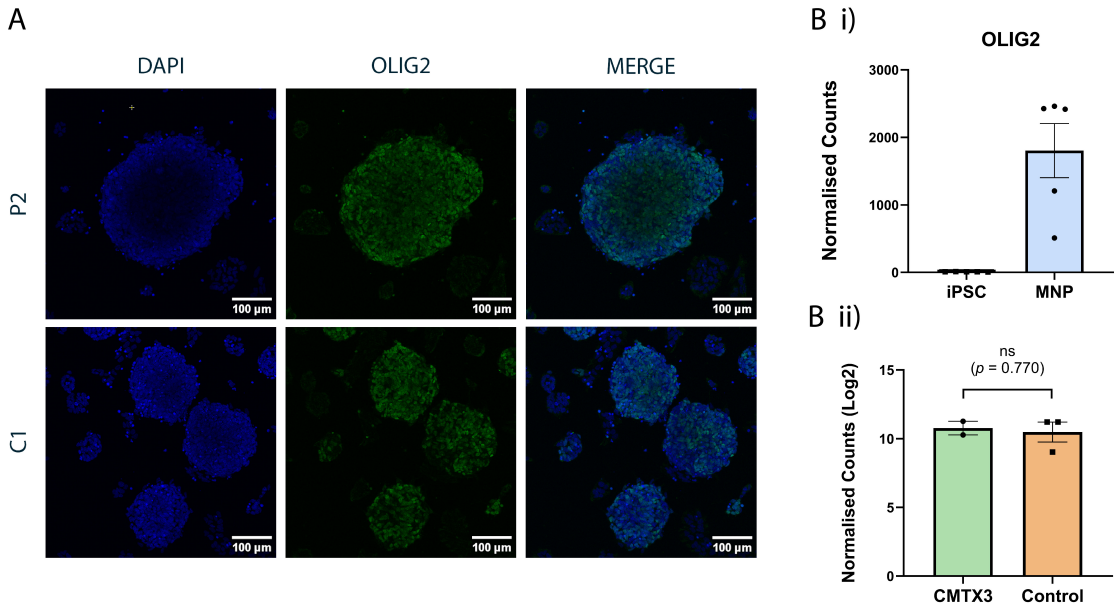


Figure 3.4: Generation of OLIG2-positive MNP. (A) Immunofluorescence staining confirms the expression of OLIG2 (green) within the population of MNPs that are generated from the differentiation of iPSC lines. DAPI staining (blue) is overlayed to identify nuclei. Representative images from one patient (P2) and one control (C1) are displayed. (B) (i) NanoString gene expression data demonstrating the normalised counts of *OLIG2* mRNA in iPSC (n=5) compared to MNP (n=5). (ii) Comparison of *OLIG2* mRNA expression levels between CMTX3 patient-derived MNP (*n*=2) and control MNP (*n*=3). Error bars represent mean ± SEM. Data is from one experimental replicate. Statistical analysis was performed using a t-test with Welch's correction (ns = non-significant) on log2-transformed normalised count values.

3.3.3 Differentiation of MNP gives rise to MNX1-positive spinal motor neurons

OLIG2-positive MNP were then dissociated and cultured in suspension as 3D neurospheres, under defined chemical conditions, to facilitate differentiation into sMN. After six days, the neurospheres were dissociated into single cells and maintained under adherent conditions in neuronal media supplemented with neurotrophic factors and a NOTCH inhibitor (Compound E), which improves maturation of the sMN [69]. To confirm cellular identity, the resulting sMN cultures underwent immunofluorescence staining for MNX1, also known as HB9; an early marker of post-mitotic spinal motor neurons [127]. As seen in Figure 3.5A, both patient and control-derived sMN display robust expression of MNX1 within cellular nuclei. These cells also express TUBB3, a general marker of the neuronal cytoskeleton (Figure 3.5A). Supplementary Figure 3.2 provides representative microscope images from all five cell lines used in this thesis. A custom CellProfiler pipeline was used to quantify the percentage of nuclei positively labelled for MNX1. Statistical analysis indicated the percentage of MNX1-positive cells is similar between the patient (76.53% ± 3.966 SEM) and control (69.99% ± 1.418 SEM) sMN cultures ($p > 0.05$) (Figure 3.5B). To further investigate sMN differentiation, the expression level of three motor neuron markers, *MNX1*, *CHAT* and *ISL1*, were assessed using a NanoString nCounter gene expression assay. Like MNX1, ISL1 is a transcription factor expressed in post-mitotic motor neurons [128]. ChAT is an enzyme required for acetylcholine synthesis and is commonly used as a marker for mature motor neurons [129]. All three differentiation markers show robust tissue-specific expression within the sMN cultures, but not in the undifferentiated iPSC lines nor the MNPs (Figure

3.5C(i)). Patient and control sMN display similar expression levels of all three differentiation markers tested ($p > 0.05$) (Figure 3.5C(ii)).

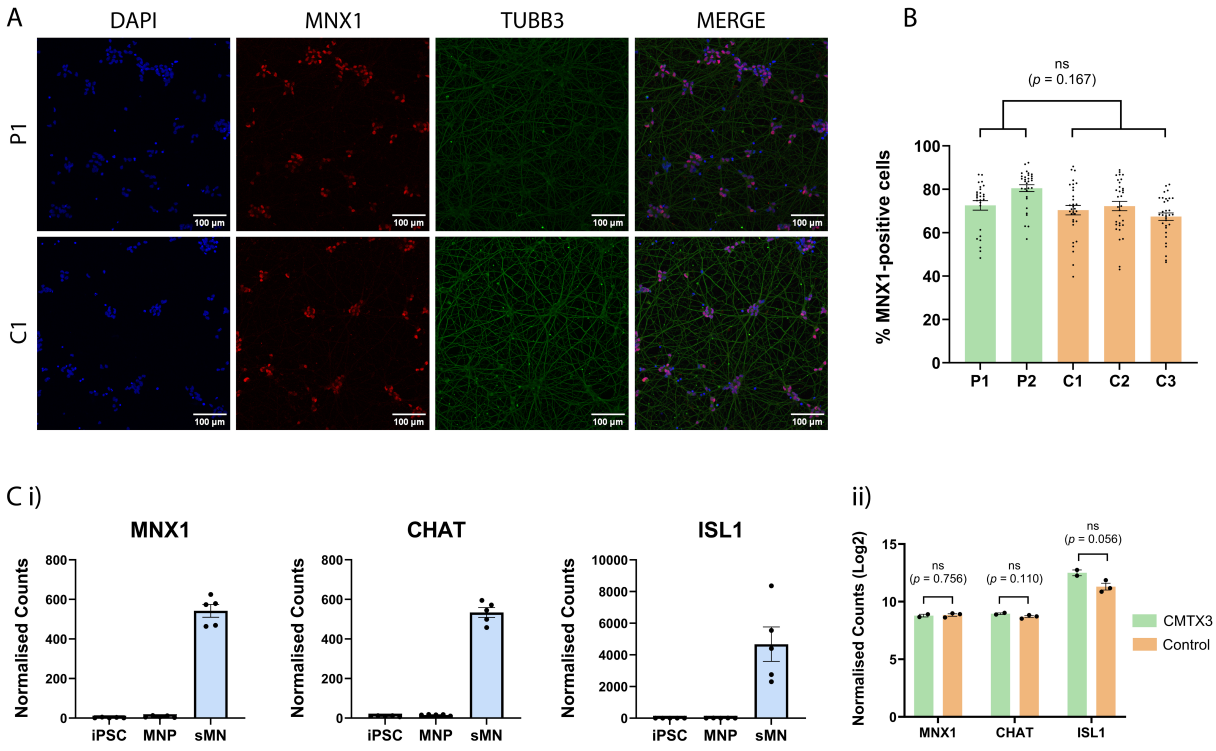


Figure 3.5: iPSC-derived sMN express markers of motor neuron identity. (A) Representative immunofluorescence images from patient sample (P1) and control sample (C2) confirm the expression of MNX1 (red) in the nuclei (blue) of sMN, as well as TUBB3 (green) in the cytoplasm. (B) Quantification of the percentage of sMN nuclei showing positive immunostaining of MNX1 for patient (n=2) and control (n=3) sMN cultures. Immunofluorescence data was aggregated from three independent rounds of differentiation to obtain the mean value for each sample, with a total of 27-32 images analysed per cell line. Column graph displays the mean ± SEM. An unpaired two-tailed Student's t-test was used to compare patient and control means (ns = non-significant). (C) (i) NanoString gene expression data demonstrating the expression levels of the three motor neuron markers, *MNX1*, *CHAT* and *ISL1*, across the three increasingly specialised cell types; iPSC (n=5), MNP (n=5) and sMN (n=5). (ii) Comparison of the NanoString gene expression levels of sMN markers between patient (n=2) and control (n=3) sMN. Statistical analysis was performed on the log2-transformed normalised count data using a t-test with Welch's correction and a significance threshold of $p < 0.05$ (ns = not significant).

3.4 Discussion

In this chapter, the generation and characterisation of the first neuronal model for CMTX3, through leveraging the available expertise of the team at the Northcott Neuroscience Laboratory for establishing iPSC-derived sMN models of IPN is described [67, 111, 112]. Initially, the major barrier preventing progress for investigating the pathomechanism of the CMTX3 SV was the lack of access to disease-relevant cell types. This was due to the technical and ethical challenges associated with obtaining patient-derived neural tissue which would be essential for maintaining the patient genetic background and causative mutation of the disease. The CMTX3 iPSC-derived sMN generated and described in this chapter resolved this limitation, making available a valuable patient neuronal tissue resource for use throughout this study to characterise the molecular consequences of the CMTX3 SV in a disease-relevant context.

Characterisation of the iPSC generated from re-programmed fibroblasts showed the CMTX3 patient derived lines retained the ~78 kb complex SV, were successfully differentiated into OLIG2-positive MNP and subsequently, MNX1-positive sMN. NanoString assays performed across multiple timepoints throughout the differentiation protocol confirmed the expected temporal expression pattern of differentiation markers. The transition from undifferentiated iPSC to MNP was marked by a clear increase in expression of the MNP marker, *OLIG2* (Figure 3.4B(i)). In addition, the three markers of spinal motor neurons used in this study, *MNX1, CHAT,* and *ISL1*, were undetectable in the iPSC and MNP cultures, but robustly expressed in all 5 sMN lines (Figure 3.5C(i)). As the relevant gold standard markers were robustly expressed, this indicated the differentiation protocol implemented for this study successfully generated the desired cell type, mirroring the gene expression patterns previously observed in iPSC-derived sMN models generated within the Northcott Neuroscience Laboratory [67].

Inspection of the normalised NanoString count data indicated a degree of variability in the expression levels of these markers between cell lines, which was particularly evident in the *OLIG2* and *ISL1* datasets. This was not surprising, as it is well-documented that iPSC lines can vary in their propensity to differentiate into target cell types [130-132], including iPSC that are derived from the

same donor cell line [133]. This inherent variability represents a limitation that is common across iPSC-based disease-models [134]. An important finding in this study, however, was that the patient and control cohorts did not significantly differ in the expression levels of the differentiation markers tested, at both the mRNA (*MNX1, CHAT, ISL1*) and protein (MNX1) level. It is possible to overcome the impact of genetic variability by using isogenic control lines [135]. Isogenic controls are genetically identical to the parental cell line, except that a selected mutation may be introduced into the genome at a specific site of interest [136]. In the case of CMTX3, this would represent a correction of the pathogenic ~78 kb SV. While it is difficult to genetically engineer the complex CMTX3 SV, current CRISPR/Cas gene editing technologies can be used to generate large genomic deletions [101, 137], which could be applied to excise the ~78 kb CMTX3 insertion. In brief, a pair of sgRNA are used to induce double stranded DNA breaks at two specific sites, which can result in deletion of the intervening region of DNA through the non-homologous end joining (NHEJ) repair pathway [101, 137]. This strategy has been used to achieve genomic deletions of 1.5 Mb and larger [101, 102]. Of note, the NHEJ pathway is error prone and often generates indels at the site of the double stranded break, and the efficiency of genome engineering can vary across genomic loci [101].

The initial publication outlining the sMN differentiation protocol implemented in this study reported a yield of > 90% MNX1-positive sMN [69]. In the current study, an average yield of 73% was observed across the 5 cell lines (Figure 3.5B). It is likely that this discrepancy is partly explained by the well-reported line-to-line variability of iPSC as outlined above, with differences in genetic background accounting for the greatest source of variability among iPSC lines [130, 138]. However, technical sources of variability, such as inter-laboratory variability [139], are also likely to be contributing. It is interesting that the purity of sMN obtained in the present study are slightly reduced from previous studies within our laboratory in which common control iPSC lines have been used for both projects (~80% MNX1-positive) [67]. This could suggest it may be worthwhile evaluating the stability of the small molecules utilised in the differentiation process, to ensure optimal potency for cellular differentiation. For example, the current protocol incorporates SN38-P to chemically decontaminate

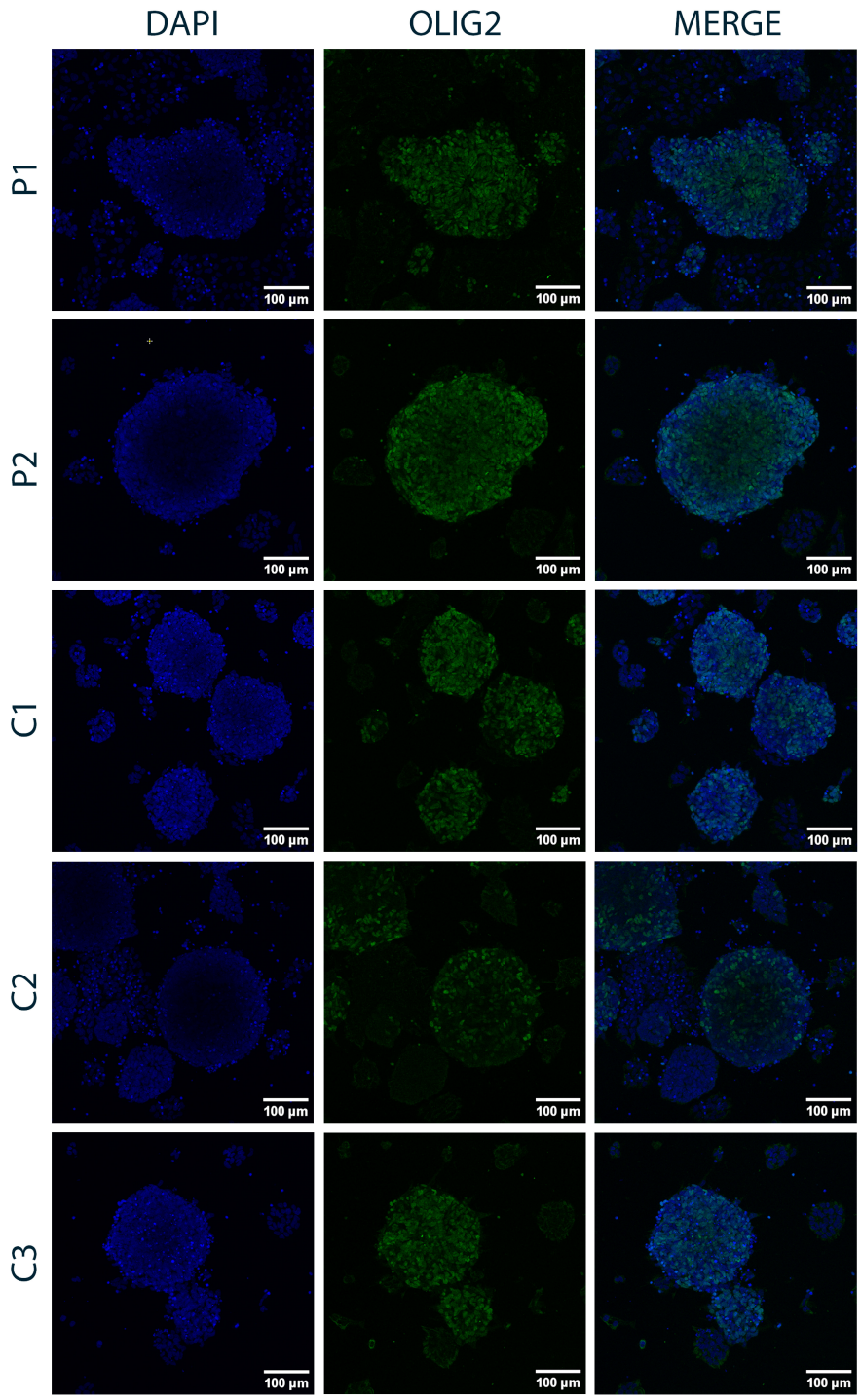
the sMN of residual proliferative progenitor cells [70], thereby improving the purity of the resulting sMN cultures [111]. Stocks of this compound have not been regularly replenished, as the current suppliers of SN38-P through our collaborator have discontinued in-house synthesising of this compound. Despite this limitation, the purity of MNX1-positive sMN obtained in this study falls within the range that has previously been reported among various other iPSC-derived sMN models of CMT neuropathy (~50-80%) [67, 111, 112, 140]. It should be noted, however, that not all publications reporting iPSC-derived sMN for modelling CMT explicitly state the differentiation efficiencies [113-115], making comparisons between models difficult and unreliable.

A benefit of iPSC-based model systems is the capacity to produce large amounts of patient-derived cellular material for downstream experiments and analysis. From the experiences in this study, the differentiation of one 35 mm well of iPSC produced up to 6 wells of MNP, which themselves could be passaged further due to the highly expandable nature of this cell type [69]. As noted by Acemel and colleagues (2019)[141], this property makes iPSC a valuable resource for modelling complex SV where altered regulatory mechanisms are suspected. This is because the experimental techniques required to probe the epigenome, such as chromatin-conformation capture assays, or ChIP-Seq, often require a large amount of cells as starting material. Further, the capacity for MNP to be cryopreserved substantially reduces the time course required to generate sMN compared to iPSC differentiation. An added benefit of iPSC models is the ability to provide unprecedented access to transient, developmental cell-types [107]. This will be particularly advantageous for this project as multiple candidate genes within the CMTX3 linkage region have known roles in regulating embryonic development. This patient model will therefore allow a thorough investigation into the impact of the CMTX3 SV on local gene regulation in a variety of spatiotemporal contexts of neuronal tissue.

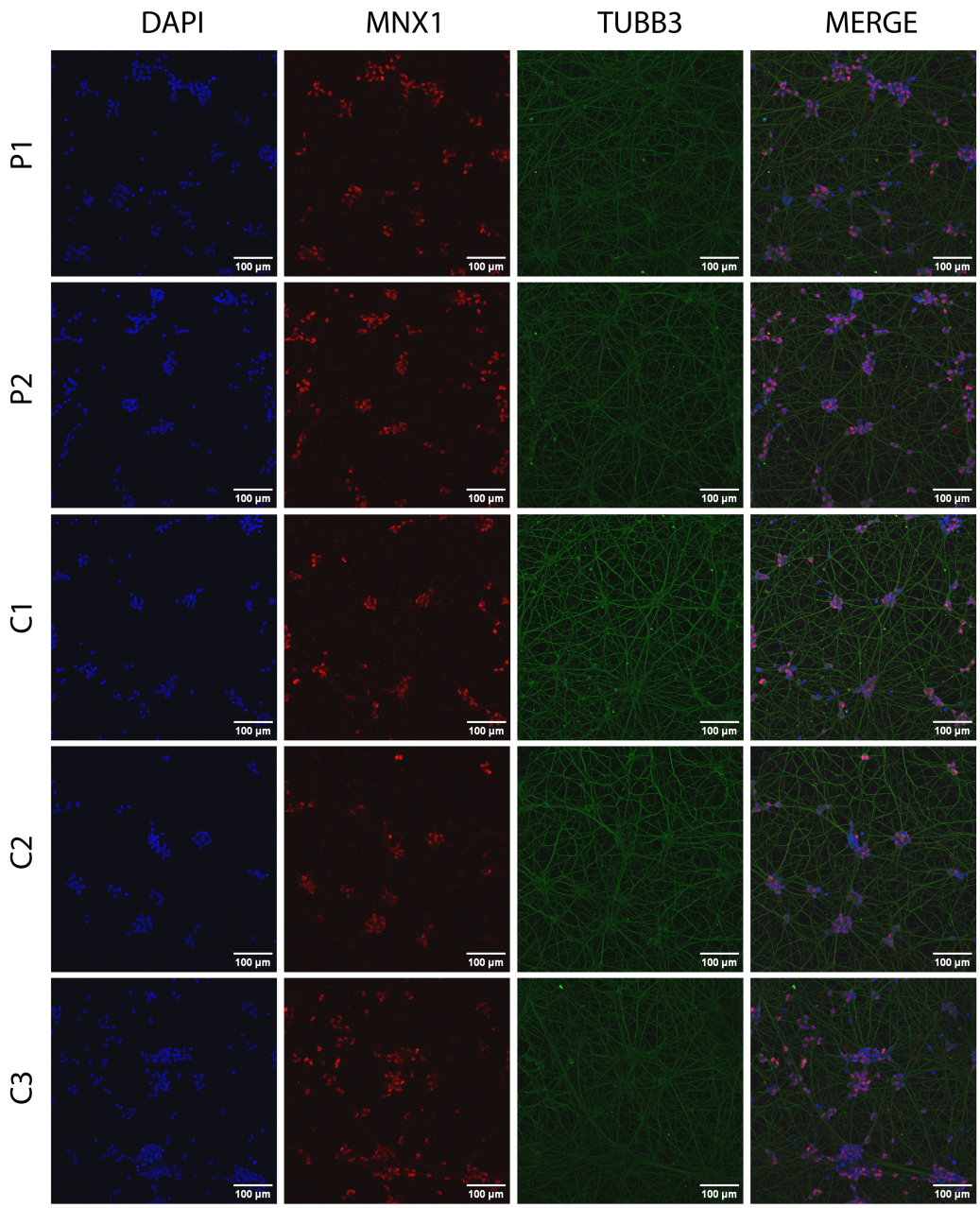
In summary, a neuronal *in vitro* model for CMTX3 has been generated and to the best of our knowledge, this represents the first iPSC-based model generated for any of the diseases caused by complex SV at the Xq27.1 disease hotspot (reviewed in Chapter 1, Part B; [64]). This model offers a valuable tool for exploring the effects of pathogenic SV at this interesting genomic locus. In the current

study, a differentiation protocol was adopted to generate sMN [69]. However, sMN are not the only cell type affected in CMT neuropathy. Patients also exhibit pathology within sensory fibres and Schwann cells, in cases of demyelinating and intermediate forms of CMT. Thus, an additional advantage of the iPSC resource generated and described in this chapter is the flexibility for generating additional cell types, such as sensory neurons [108, 142], Schwann cells [109], or PNS organoids [110], depending on future biological questions that may arise from this work. Therefore, the utility of this model will extend far beyond the scope of the current project.

Supplementary Material: Chapter 3



Supplementary Figure 3.1: Immunofluorescence confirms iPSC-derived MNP express OLIG2. Fluorescence microscopy images of the iPSC-derived MNP lines (passage 2), used in this thesis and stained with OLIG2 (green). The lines include two patient (P1-2) and three control (C1-3) samples. Nuclei are co-stained with DAPI (blue).



Supplementary Figure 3.2: Immunofluorescence confirms the expression of differentiation markers in iPSC-derived sMN. Representative images demonstrating the expression of MNX1 (red) and TUBB3 (green) in the 5 iPSC-derived sMN lines used in this thesis, including 2 patient and 3 control samples. Nuclei are co-stained with DAPI (blue).

CHAPTER 4

Exploring the Pathomechanism of the CMTX3 Structural Variation

4.1 Introduction

At present, there are two leading hypotheses which require careful consideration when investigating the pathomechanism underlying the CMTX3 interchromosomal insertion. The first relates to the presence of the *ARHGAP39* partial transcript localising within the inserted DNA fragment. Overexpression of this partial transcript, or the formation of a novel transcript arising from aberrant splicing events, have been proposed as possible mechanisms underlying this disease [63]. The second hypothesis postulates the presence of the ~78 kb CMTX3 SV alters the expression of one or more nearby genes [63]. Access to a neuronal model of CMTX3 will enable testing the two hypotheses in a disease-relevant context.

4.1.1 Hypothesis 1: The pathomechanism involves the expression of the ARHGAP39 partial transcript

When the CMTX3 mutation was first reported in 2016, gene expression analysis was performed to assess the expression level of *ARHGAP39*. Using a probe-based qPCR, lymphoblasts isolated from a CMTX3 patient showed similar levels of *ARHGAP39* mRNA when compared with control samples [63]. This suggested that overexpression of the partial transcript is not likely to be the pathomechanism driving CMTX3 [63], although validating this conclusion would require assessment in

disease-relevant cell types. Aberrant splicing was also deemed an unlikely mechanism as the closest downstream gene, when considering the direction of *ARHGAP39* transcription, is transcribed from the opposing strand [63]. However, aberrant splicing involving the partial *ARHGAP39* transcript and non-coding regions of DNA could also be a possibility. Whilst rarely reported in the literature, genomic rearrangements can give rise to fusion transcripts which incorporate pseudo-exons (or cryptic exons) derived from non-coding sequences of DNA [143-148]. For example, a rearrangement between an intergenic region on chromosome 21 and an intron of the *ALK* gene was observed in tumour cells from a patient with lung adenocarcinoma, which produced a stable fusion transcript including a novel exon derived entirely from the intergenic sequence [148]. In another study, a patient with haemophilia A was found to harbour a complex SV within an intron of the *F8* gene, which generated a novel fusion transcript involving *F8* and a pseudo-exon containing non-coding sequence from within the SV [143]. Recent research within the field of IPN further reinforces the importance of considering gene-intergenic fusion transcripts as a potential pathomechanism. A study published by our laboratory in 2023 identified the presence of a novel fusion transcript within iPSC-derived sMN from patients with Distal Hereditary Motor Neuropathy Type 1 (DHMN1) [67]. This rare condition, characterised by selective degeneration of lower motor neurons, is caused by a complex genomic rearrangement [66]. In DHMN1 patients, 1.35 Mb of chromosome 7q36.3 has been duplicated, inverted, and inserted into the DHMN1 linkage locus located approximately 2.3 Mb proximal to the duplicated sequence [66]. The large insertion encodes 5 full-length genes and 1 partial transcript of the *UBE3C* gene, including the first 10 exons [66]. Similar to CMTX3, the SV breakpoint occurs within an intron of the partially duplicated gene. RNA-Seq analysis detected a fusion transcript specific to the DHMN1 patient samples, that comprised of the first 10 exons of *UBE3C* followed by a terminal pseudo-exon from splicing of non-coding DNA within the DHMN1 linkage region [67]. Although the precise pathomechanism of this novel *UBE3C*-intergenic fusion (UBE3C-IF) transcript remains unclear, a reduction in the expression of wild type UBE3C was observed in DHMN1 sMN, and a C. elegans model overexpressing UBE3C-IF showed synaptic transmission deficits and susceptibility to heat stress [67]. Considering the similarities

between the DHMN1 and CMTX3 genomic rearrangements, this raises the question of whether a similar molecular mechanism may be driving these two conditions [68].

4.1.2 Hypothesis 2: The CMTX3 insertion alters spatiotemporal regulation of nearby genes

As discussed in chapter 1, there is compelling evidence in the literature demonstrating that SV can cause disease by rewiring interactions between genes and distal cis-regulatory elements, leading to altered gene regulation (reviewed in [68, 149, 150]). For the numerous diseases that are associated with complex insertions occuring at the Xq27.1 mutational hotspot (reviewed in Chapter 1, Part B), local gene dysregulation has been clearly observed in three; XX-male sex reversal [151], X-linked congenital generalised hypertrichosis [152], and CMTX3 [63] (Figure 4.1). Robust ectopic expression of *SOX3* was observed in lymphoblasts isolated from a patient with XX-male sex reversal, while *SOX3* mRNA was undetectable in controls [151]. Duplications involving *SOX3*, as well as genomic rearrangements upstream of *SOX3*, have previously been reported in 46, XX-male patients [153-155], whilst transgenic mice with ectopic *SOX3* expression within the developing gonad show an XX-male phenotype [153]. Together, this provides strong evidence suggesting that the XX-male sex reversal phenotype arises from altered regulation of *SOX3* induced by the pathogenic SV [151]. In X-linked hypertrichosis, patient hair follicles were found to show markedly reduced expression of another nearby gene, *FGF13*, when compared with controls [152]. Whilst *FGF13* has not been previously linked to this clinical phenotype, this spatiotemporal dysregulation was observed within disease-relevant tissue, and analysis of *FGF13* expression during murine embryogenesis suggests it may play a role in hair follicle morphogenesis [152]. *FGF13* dysregulation was also observed within the context of CMTX3, with patient lymphoblasts showing upregulation of this gene [63]. Neither *FGF13*, nor any of the genes within the CMTX3 linkage region, have previously been linked with a CMT phenotype. *FGF13* is an interesting candidate gene for CMTX3, as it performs multiple functions within neuronal cells (reviewed in [156]). For example, FGF13 was shown to regulate polarisation and migration of developing cortical neurons through its action as a microtubule stabilising protein [157]. Its expression is maintained in the adult peripheral nervous system, where it co-localises with voltage-gated sodium

channels in afferent fibres [158]. However, given that gene regulation has not yet been assessed within a disease-relevant cellular context, it is unclear whether the dysregulation of *FGF13* is implicated in the pathogenesis of CMTX3 or merely a bystander effect of this mutation [63]. Nevertheless, this finding clearly demonstrates the capacity of the CMTX3 SV to alter the expression profile of a nearby gene, even across a genomic distance spanning 1.2 Mb, and underscores the importance of conducting a thorough assessment of local gene regulation within neuronal tissue.

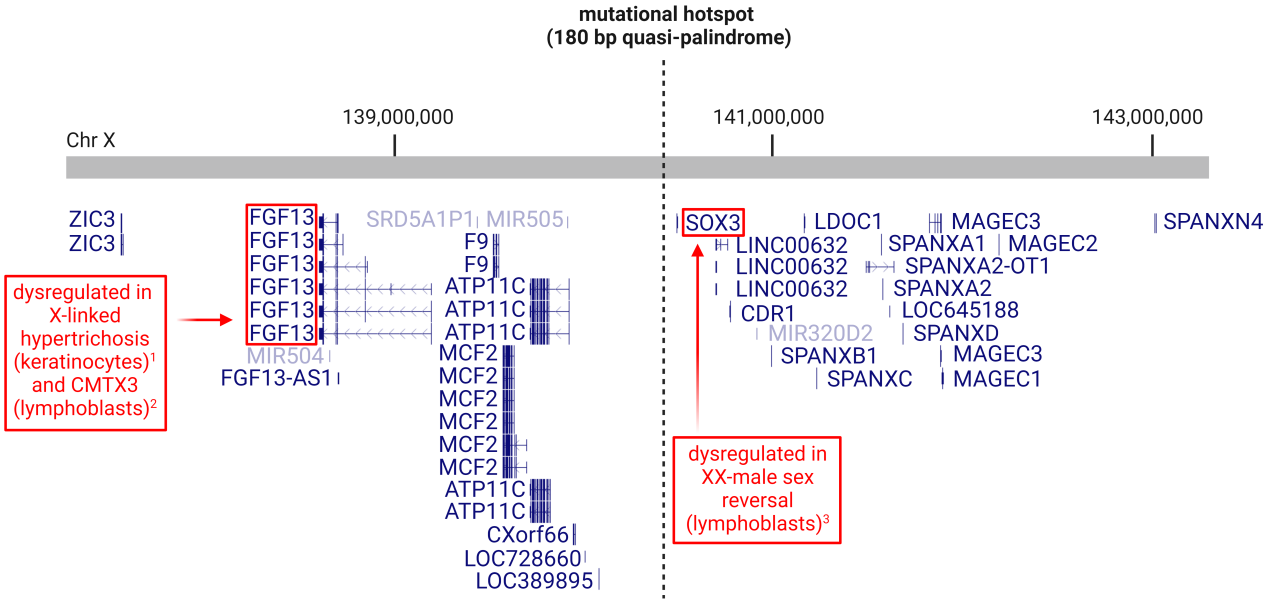


Figure 4.1: Pathogenic rearrangements at the Xq27.1 hotspot can alter the spatiotemporal regulation of nearby genes. An overview of the genomic region surrounding the Xq27.1 quasi-palindrome, which is prone to genomic rearrangements. Multiple pathogenic SV occurring at this hotspot alter the regulation of two nearby genes, *FGF13* and *SOX3* (red boxes), located ~1.2 Mb and ~82 kb from the hotspot, respectively. For visual representation, all isoforms are outlined, however, dysregulation was observed at the gene level (i.e. not individual isoforms). A ~389 kb interchromosomal insertion causing X-linked congenital generalised hypertrichosis drives downregulation of *FGF13* expression in patient keratinocytes[1] (DeStefano et al., 2013). A ~78 kb interchromosomal insertion causing CMTX3 neuropathy results in reduced *FGF13* expression in patient lymphoblasts[2] (Brewer et al., 2016). A ~774 kb interchromosomal insertion causing 46, XX-male sex reversal drives ectopic expression of *SOX3* within patient lymphoblasts[3] (Haines et al., 2015).

The purpose of this chapter is to utilise the iPSC-derived sMN model of CMTX3 to address the two hypotheses outlined above, which have not yet been adequately explored due to a lack of

appropriate models. To do so, we draw upon the experimental pipeline proposed in Chapter 1, Part A, which incorporates two key technologies, RNA-Seq and NanoString, to investigate the transcriptional consequences of the CMTX3 SV.

4.2 Methods

4.2.1 RNA Sequencing

4.2.1.1 Sample preparation

iPSC were cultured and differentiated as previously described (Chapter 2, Section 2.3). Total RNA was isolated using the RNeasy Mini Kit (Qiagen), according to manufacturer's guidelines. RNA was extracted from cells at the following timepoints; NEP (differentiation day 6) (n=2 patients, n=3 controls) and sMN (differentiation day 28) (n=1 patient, n=3 controls). Note that two clonal cell lines were available for the CMTX3 patient sMN RNA. RNA was quantified using a NanoDrop ND-1000 spectrophotometer (Thermo Fisher Scientific).

4.2.1.2 RNA-Seq and data processing (Macrogen Oceania)

RNA-Seq and data analysis was performed by Macrogen Oceania (Seoul, South Korea). Briefly, mRNA libraries were prepared using the TruSeq Stranded mRNA Library Prep Kit (Illumina), which selects for RNA species containing poly(A) tails. Short read (151 bp) paired-end sequencing was performed using a NovaSeq platform (Illumina). The quality of raw data was assessed using the FastQC program (version 0.11.7) [159]. This was followed by trimming of adaptor sequences and low-quality bases using the bioinformatics tool, Trimmomatic (version 0.38) [160]. The resulting trimmed reads were then aligned to the GRCh38/hg38 reference genome using HISAT2 (version 2.1.0) [161], which is a tool enabling spliced alignment of RNA-Seq reads. Aligned reads were assembled into transcripts and quantified using StringTie (version 2.1.3b) [162]. This provides information on known and novel transcripts which were then classified using GffCompare [163]. Three separate programs were used to

detect the presence of fusion genes; Arriba (version 1.2.0) [164], deFuse (version 0.8.1) [165], and FusionCatcher (version 1.00) [166].

4.2.1.3 Differential gene expression analysis

To identify genes that show differential expression between patients and controls, either edgeR [167] or DESeq2 [168] analysis packages were employed, depending on the experimental design. edgeR was utilised for the sMN data, in which biological replicates for patient RNA-Seq data was not available, as this tool can analyse datasets when only one experimental group is replicated [167]. Differential expression analysis was conducted using the edgeR Exact Test, according to the edgeR user guide [169] (Appendix A). The two CMTX3 clones were separately compared to the control cohort, as treating genetically identical cell lines as replicates violates the assumption that replicates are independent of one another [134]. The edgeR default parameters were used for detecting differentially expressed genes (DEG), including log2 fold change ≠ 0 and FDR-adjusted p value < 0.05. Both edgeR and DESeq2 were used to identify differentially expressed genes from the NEP dataset, as biological replicates were available at this timepoint. Default parameters were used for the DESeq2 analysis, which identifies differentially expressed genes between patients and controls as those with log2 fold change ≠ 0 and an FDR-adjusted p value < 0.1. DESeq2 code is available in Appendix B, adapted from a pipeline previously utilised by our laboratory [67].

To explore the biological functions of the DEG, the bioinformatics tool, DAVID [170, 171] was used to perform a functional enrichment analysis. The background gene list used for this analysis included all genes that underwent differential expression analysis (i.e. after low count filtering was performed). The 'Functional Annotation Clustering' tool was used, which groups similar annotation terms together into clusters showing related biology.

4.2.2 NanoString targeted transcriptome analysis

RNA was extracted as previously described (Chapter 2, Section 2.4) from cells isolated at the following timepoints; iPSC, caudal NEP (differentiation day 6), MNP (differentiation day 12), neurospheres (differentiation day 18) and mature sMN (differentiation day 28). For each timepoint,

RNA was extracted from 2 patient and 3 control cell lines. NanoString nCounter gene expression assays were performed and data analysed as outlined in Chapter 2, Section 2.6. Graphs were plotted with GraphPad Prism (version 10) and Microsoft Excel.

4.2.3 Western blot

4.2.3.1 Protein extraction

To obtain total protein, iPSC were cultured in 6-well plates and differentiated into caudal NEP as previously described (Chapter 2, Section 2.3.3). Cells were lysed using 1X RIPA lysis buffer (Merck) supplemented with 1X cOmplete Mini EDTA-free protease inhibitor cocktail (Roche). Briefly, cells were washed twice with chilled $DPBS^{(-Ca\ -Mg)}$ (Gibco) then collected in 500 µL of chilled lysis buffer. Lysates were transferred to 1.5 mL microcentrifuge tubes (Eppendorf) and agitated at 4°C for 30 min using an SB3 Rotator (Stuart). Samples were centrifuged at 16,000 x g for 10 min at 4°. Supernatant was collected and stored at -80°C until required.

4.2.3.2 Protein quantification

A bicinchoninic acid (BCA) assay was performed to determine the protein concentration within cell lysates using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific). Briefly, a working reagent was prepared through combining Reagent A and Reagent B at a ratio of 50:1. Standardised protein solutions, ranging from 0 – 2000 µg/mL, were prepared through serial dilutions of a bovine serum albumin (BSA) standard solution (supplied in kit). Based on previous observations, experimental samples were diluted in lysis buffer (1:2), to maximise the likelihood of values falling within the linear range of the standard curve. Reactions were then prepared by combining 200 µL of working reagent with 10 µL of each sample (unknown and BSA standards) in one well of a Nunclon MicroWell plate (Merck). Triplicate wells were prepared for each sample. After gently pipetting up and down to mix reaction components, the plate was incubated at 37°C for 30 min. The EnSpire II Multimode Plate Reader (Perkin Elmer) was used to measure absorbance at 562 nm. A standard curve was generated by plotting the absorbance values (blank-corrected) against protein concentration for each of the BSA

standard solutions. This curve was used as a reference to calculate the protein concentrations of each experimental sample.

4.2.3.3 Sodium dodecylsulphate-polyacrylamide gel electrophoresis (SDS-PAGE)

Cell lysates were mixed with Laemmli sample buffer (1X) (Bio-Rad) and 2-mercaptoethanol (2.5% v/v) (Sigma-Aldrich) then heated at 90°C for 5 min. For each sample, 20 µg of protein was loaded onto a 4-15% Mini-PROTEAN TGX Precast Protein Gel (Bio-Rad) using Prot/Elec pipet tips (Bio-Rad). To determine molecular weight, the Precision Plus Protein Kaleidoscope Prestained Protein Standards (Bio-Rad) ladder was loaded alongside experimental samples. Electrophoresis was performed using a Mini-PROTEAN Tetra Vertical Electrophoresis Cell (Bio-Rad) and a pre-mixed Tris/Glycine/SDS running buffer (Bio-Rad) containing 25 mM Tris, 192 mM glycine and 0.1% (v/v) SDS. Samples were size fractionated at 110 V for 70 min at an ambient temperature of 4°C.

4.2.3.4 Protein transfer

The size-fractionated proteins were then transferred onto an Immobilon-P polyvinylidene difluoride (PVDF) membrane (Millipore) using a Mini Trans-Blot Electrophoretic Transfer Cell (Bio-Rad). Electroblotting was performed at 70 V for 70 min, using a pre-mixed Tris/Glycine buffer (Bio-Rad) comprised of 25 mM Tris and 192 mM glycine. To prevent overheating, an icepack was inserted into the transfer tank and electroblotting was performed at 4°C with continuous magnetic stirring. To confirm successful protein transfer, PVDF membranes were briefly rinsed in Milli-Q water and immersed in Ponceau S solution (Sigma Aldrich) for 10 min, which enabled visualisation of protein bands.

4.2.3.5 Antibody incubation and detection

PVDF membranes were blocked in 5% (w/v) skim milk (Oxoid) in 1X Tris Buffered Saline (TBS), comprised of 20 mM Tris base (AppliChem GmbH) and 150 mM NaCl (Sigma Aldrich) dissolved in Milli-Q water with a final pH of 7.6. Blocking was carried out for 1 h at room temperature under constant agitation. Membranes were then washed (3 x 10 min) in 1X TBS containing 0.1% (v/v) Tween 20 detergent (Calbiochem) (TBS-T). Primary antibodies were prepared by diluting the following antibodies

in 'antibody dilution solution' comprised of 5% (w/v) BSA (Sigma Aldrich) dissolved in 1X TBS-T; rabbit anti-SOX3 polyclonal antibody (1:1000, PA5-35983, Invitrogen) and mouse anti-α-Tubulin monoclonal antibody (1:5000, T5168, Sigma-Aldrich). Membranes were incubated in the primary antibody solution overnight at a temperature of 4°C. To remove excess primary antibody, membranes were washed in 1X TBS-T (3 x 10 min) under constant agitation. Membranes were then incubated with anti-rabbit IgG horseradish peroxidase conjugated antibody (A0545, Sigma-Aldrich) diluted in antibody dilution solution (1:5000) for a total of 90 min at room temperature. Excess secondary antibody was removed by washing in 1X TBS-T (3 x 10 min). Membranes were rinsed and subsequently washed in 1X TBS to prepare the membrane for imaging. Membranes were then immersed for 5 min in Clarity Western ECL Substrate, comprised of a mix (1:1) of Clarity Western Peroxide Reagent and Clarity Western Luminol/Enhancer Reagent (Bio-Rad). Images were acquired with a ChemiDoc MP Imaging System (Bio-Rad). Membranes were then washed in 1X TBS-T (3 x 10 min) and incubated with anti-mouse IgG horseradish peroxidase conjugated antibody (1:5000, ab97023, Abcam) for 90 min at room temperature. Excess secondary antibody was removed by washing in 1X TBS-T (3 x 10 min) and imaging was performed as described above.

4.2.3.6 Western blot quantification

Densitometric quantification of western blot images was performed using FIJI [71]. Briefly, a region of interest (ROI) was drawn around the target protein band and an intensity plot generated. The same ROI selection frame was used to define protein bands across all samples. The line selection tool was used to delineate the boundaries of the intensity peaks corresponding to protein bands, and the area of each peak was then measured. Normalisation factors were calculated for each lane based on the loading control (α-tubulin) signals, and this value was then used to calculate the normalised signal for the target protein (SOX3). GraphPad Prism (version 10) was used to generate graphs and calculate statistical significance, using an unpaired two-tailed t-test and a significance threshold of $p < 0.05$.

4.3 Results

4.3.1 Exploratory RNA-Seq analysis of iPSC-derived sMN

After obtaining iPSC-derived sMN harbouring the CMTX3 SV, an RNA-Seq experiment was performed. This powerful technology provides a starting point for exploring the pathomechanism of the CMTX3 SV as the method can simultaneously address both hypotheses outlined in Chapter 4, Section 4.1 by (a) quantifying gene expression levels and (b) detecting novel features, including fusion transcripts and aberrant splice variants.

4.3.1.1 Quality control metrics

Poly(A) selected RNA-Seq was performed on sMN RNA from one patient (two clonal cell lines) and three controls. RNA was assessed using an Agilent Bioanalyzer system to calculate the RNA Integrity Number (RIN). All samples showed RIN values between 9.3-9.8, exceeding the minimum cutoff score of 8 that is typically recommended for RNA-Seq [172]. Libraries were prepared, sequenced, and processed by Macrogen Oceania (Seoul, South Korea). Raw sequencing reads were assessed for GC content and Phred scores, which reflect the quality of base calling. QC metrics were similar across all 5 libraries (Supplementary Table 4.1). Following the trimming of adaptor sequences, PCR duplicates, and low-quality reads, transcripts were aligned to hg38, where the average number of mapped reads was 123,680,405 per sample (Supplementary Table 4.1). Mapped reads accounted for, on average, 98.84% of trimmed reads (Supplementary Table 4.1).

4.3.1.2 Novel fusion transcripts involving the CMTX3 SV were not detected with RNA-Seq

To explore whether the CMTX3 SV gives rise to a novel fusion transcript, RNA-Seq data was analysed using three bioinformatics packages designed for detecting fusion transcripts; Arriba [164], deFuse [165], and FusionCatcher [166].

4.3.1.3 Arriba fusion transcript prediction

Table 4.1 lists the fusion transcripts predicted by Arriba [164] which involve DNA sequence originating from chromosome X or chromosome 8 in either of the CMTX3 clones. Fusion transcripts spanning these two chromosomes were not identified in either sample. While 15 fusion transcripts were detected on chromosome X, they did not involve sequence from within the CMTX3 linkage region (Table 4.1A). Similarly, the 16 predicted fusions on chromosome 8 did not involve the ~78 kb fragment of 8q24.3 that is duplicated in CMTX3 patients (Table 4.1B).

Table 4.1: List of fusion transcripts, predicted by Arriba, on chromosome X and chromosome 8 from CMTX3 patient-derived sMN. (A) Data from chromosome X. (B) Data from chromosome 8.

A

Sample	gene1	gene2	strand1 (gene/fusion)	strand2 (gene/fusion)	breakpoint1	breakpoint2	confidence
CMTX3 Clone A	LOC102723546	IGSF1	-/-	-/-	X:131389017	X:131286741	low
CMTX3 Clone A	LOC102723546	IGSF1	-/-	-/-	X:131389017	X:131289275	high
CMTX3 Clone A	RAI2(23395),LINC01456(85432)	RAI2	./-	-/-	X:17884741	X:17802034	low
CMTX3 Clone A	PRRG1	LANCL3	+/+	+/+	X:37426000	X:37655688	low
CMTX3 Clone A	XK	CYBB	+/+	+/+	X:37694548	X:37782088	low
CMTX3 Clone A	XK	CYBB	+/+	+/+	X:37694548	X:37791975	low
CMTX3 Clone A	NLGN4X	LOC101928228(118365),NLGN4X(45525)	-/-	./-	X:5903077	X:5844501	low
CMTX3 Clone B	LOC102723546	IGSF1	-/-	-/-	X:131364749	X:131286741	low
CMTX3 Clone B	LOC102723546	IGSF1	-/-	-/-	X:131364749	X:131289275	low
CMTX3 Clone B	LOC102723546	IGSF1	-/-	-/-	X:131389017	X:131286741	low
CMTX3 Clone B	LOC102723546	IGSF1	-/-	-/-	X:131389017	X:131289275	low
CMTX3 Clone B	PTCHD1-AS(40131),PTCHD1(1092)	PTCHD1-AS	./-	-/-	X:23333277	X:22389720	high
CMTX3 Clone B	PRRG1	PRRG1(34526),FTH1P19(136)	+/+	./+	X:37426000	X:37491821	low
CMTX3 Clone B	NLGN4X	LOC101928228(118365),NLGN4X(45525)	-/-	./-	X:5903077	X:5844501	low
CMTX3 Clone B	LOC107985664	ARL5AP5(89490),LOC107985665(23275)	+/+	./+	X:75624888	X:76121419	medium

B

Sample	gene1	gene2	strand1 (gene/fusion)	strand2 (gene/fusion)	breakpoint1	breakpoint2	confidence
CMTX3 Clone A	LOC105375704	EIF3E	-/-	-/-	8:108309336	8:108229195	low
CMTX3 Clone A	LOC105375704	EIF3E	-/-	-/-	8:108309336	8:108241913	low
CMTX3 Clone A	UTP23	LOC112268030	+/+	+/+	8:116766791	8:116786503	low
CMTX3 Clone A	UTP23	LOC112268030	+/+	+/+	8:116770366	8:116786503	low
CMTX3 Clone A	ZFP41	GLI4	+/+	+/+	8:143251340	8:143274704	low
CMTX3 Clone A	BNIP3L	PNMA2	+/+	-/+	8:26408376	8:26505307	low
CMTX3 Clone A	FAM110B	LOC107986945	+/+	-/+	8:58075623	8:58197831	low
CMTX3 Clone B	PINX1	SOX7	-/-	-/-	8:10826152	8:10726666	low
CMTX3 Clone B	UTP23	LOC112268030	+/+	+/+	8:116766791	8:116786503	low
CMTX3 Clone B	ZFP41	GLI4	+/+	+/+	8:143251340	8:143269360	low
CMTX3 Clone B	ZFP41	GLI4	+/+	+/+	8:143251340	8:143274704	low
CMTX3 Clone B	CLN8	ARHGEF10	+/+	+/+	8:1771597	8:1843353	low
CMTX3 Clone B	BNIP3L	PNMA2	+/+	-/+	8:26408376	8:26505307	low
CMTX3 Clone B	FAM110B	FAM110B(40534),LOC107986945(2002)	+/+	./+	8:58075623	8:58189318	medium
CMTX3 Clone B	FAM110B	LOC107986945	+/+	-/+	8:58075623	8:58197831	medium
CMTX3 Clone B	LINC01414	LOC105375876(9859),LINC01289(67741)	-/-	./-	8:64133448	8:63701689	medium

4.3.1.4 deFuse fusion transcript prediction

Similar to the Arriba prediction tool, deFuse [165] did not detect any fusion transcripts with breakpoints spanning chromosome X and chromosome 8. The 32 predicted fusions on chromosome X did not show breakpoints within the CMTX3 linkage region (Supplementary Table 4.2). A novel fusion transcript on 8q24.3 was predicted in both patient clones, with one breakpoint located within the *ARHGAP39* gene, which is partially duplicated in the CMTX3 insertion (Supplementary Table 4.3). However, this was excluded as a possible pathogenic candidate due to (1) the second breakpoint being located outside of the CMTX3 insertion sequence, and (2) the fusion transcript being present in all 3 control samples (Supplementary Table 4.3).

4.3.1.5 FusionCatcher fusion transcript prediction

Analysis of the two patient clones with FusionCatcher [166] predicted a total of 111 fusion transcripts with one breakpoint in chromosome 8 and the other in chromosome X (Supplementary Table 4.4). All were excluded as possible pathogenic candidates since their genomic breakpoints localised outside the CMTX3 linkage region (Supplementary Table 4.4). The 16 fusion transcripts predicted to have both breakpoints localise to chromosome X did not align to sequence in the CMTX3 locus (Supplementary Table 4.5A). The 120 fusion transcripts with both breakpoints on chromosome 8 did not originate from the duplicated ~78 kb region of 8q24.3 inserted into the CMTX3 locus (Supplementary Table 4.5B).

4.3.1.6 Novel splice variants and novel transcript variants involving the CMTX3 SV were excluded using RNA-Seq

StringTie [162] was used to assemble the mapped RNA-Seq reads into transcripts, including known and novel transcripts, which were then classified with GffCompare [163].

Across all 5 samples, 15 novel transcripts were predicted on chromosome 8, which did not involve sequence from the CMTX3 complex insertion (Supplementary Table 4.6A). Similarly, the 13 novel transcripts identified on chromosome X did not localise to the CMTX3 locus at Xq27.1

(Supplementary Table 4.6B). Eight novel splice variants were detected within the CMTX3 linkage region but were not unique to the CMTX3 samples (Supplementary Table 4.7A). No splice variants were identified within the CMTX3 complex SV (Supplementary Table 4.7B).

4.3.1.7 Exploratory differential gene expression analysis

To explore whether the CMTX3 SV causes transcriptional dysregulation of nearby genes within sMN, differential gene expression analysis was performed. Predicted differentially expressed genes (DEG) then underwent functional enrichment analysis to investigate broader biological pathways that may be perturbed in CMTX3 sMN. This experiment was considered exploratory, as biological replicates were not available, and care was taken to not over interpret the datasets. However, the analysis was performed using the bioinformatics package, edgeR, which can perform differential expression analysis when at least one cohort contains biological replication [167].

We performed two separate analyses to compare each of the genetically identical CMTX3 clones to the group of 3 unrelated controls (1 vs 3), using an FDR < 0.05 to identify DEG. Only those genes predicted to be differentially expressed in both CMTX3 clones, showing the same direction of log(fold change), were retained for downstream analysis. A total of 51 DEG were identified, of which, 27 were downregulated and 24 were upregulated in the CMTX3 sample compared to controls (Supplementary Table 4.8). The 51 DEG did not map to the CMTX3 linkage region on chromosome Xq27.1, or the ~78 kb duplicated region of chromosome 8q24.3 inserted into the CMTX3 locus (Supplementary Table 4.8).

The log(fold change) values were then assessed for all genes located ~3 Mb on either side of the CMTX3 SV breakpoint at Xq27.1, as well as *ARHGAP39*, the gene that is partially duplicated due to the CMTX3 complex insertion. Inspecting the fold change values assessed the relative expression level of each candidate gene between the patient and controls, which in this analysis did not reach statistical significance. For the 29 positional candidate genes, 17 were excluded by edgeR based on very low expression levels within the iPSC-derived sMN (Table 4.2). The remaining 12 genes were ranked according to the average log(fold change) of the two CMTX3 clones (Table 4.2). The majority of the 12

candidate genes show average log(fold change) values close to 0, indicating that gene expression levels are similar between patients and controls. Two genes, *SOX3* and *ZIC3*, show average log(fold change) values that are larger in magnitude, however, *ZIC3* expression is highly variable between the two CMTX3 clones (Table 4.2). In contrast, *SOX3* displays consistent log(fold change) values between clone A (-1.383) and clone B (-1.341). Assessment of the TPMs (transcripts per kilobase million) for *SOX3* indicate low levels of expression in sMN tissue, in both the patient sMN clones (average = 0.565 TPM) and the control lines (average = 1.477 TPM).

Table 4.2: Differential expression analysis results for the CMTX3 candidate genes. Differential expression analysis was performed with edgeR. Table lists the RefSeq curated genes contained within the CMTX3 SV, and those located within the 3 Mb flanking the CMTX3 SV breakpoint. LogFC represents patient vs control. LogFC = Log(fold change); TPM = transcripts per kilobase million; FDR = false discovery rate; NA= not applicable (low count genes excluded from analysis).

	CMTX3 Clone A				CMTX3 Clone B				
Gene	**logFC**	**TPM**	**P-Value**	**FDR**	**logFC**	**TPM**	**P-Value**	**FDR**	**Average LogFC**
SOX3	-1.383	0.555	0.021	0.580	-1.341	0.580	0.025	0.696	-1.362
ZIC3	-2.185	0.199	0.019	0.557	-0.094	0.904	0.974	1.000	-1.139
FGF13-AS1	-0.406	0.190	0.920	1.000	-0.517	0.182	0.881	1.000	-0.462
LDOC1	-0.147	57.074	0.341	1.000	-0.235	54.791	0.134	1.000	-0.191
LINC00632	-0.045	21.621	0.964	1.000	-0.275	18.838	0.590	1.000	-0.160
ATP11C	-0.135	26.080	0.435	1.000	-0.140	26.060	0.432	1.000	-0.138
ARHGAP39	-0.062	33.932	0.732	1.000	-0.191	31.664	0.292	1.000	-0.126
FGF13	0.032	28.670	0.821	1.000	-0.210	25.448	0.176	1.000	-0.089
MCF2	-0.161	9.560	0.560	1.000	0.128	12.045	0.593	1.000	-0.016
MAGEC3	0.029	0.277	0.975	1.000	0.026	0.436	0.951	1.000	0.028
LOC389895	0.114	6.007	0.754	1.000	0.063	5.930	0.854	1.000	0.088
SPANXA2-OT1	-0.079	0.126	1.000	1.000	0.988	0.290	0.288	1.000	0.455
CDR1	NA	0.066	NA	NA	NA	0.160	NA	NA	NA
CXorf66	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
F9	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
LOC645188	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
LOC728660	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
MAGEC1	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
MAGEC2	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
MIR320D2	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
MIR504	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
MIR505	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
SPANXA1	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
SPANXA2	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
SPANXB1	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
SPANXC	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
SPANXD	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
SPANXN4	NA	0.000	NA	NA	NA	0.000	NA	NA	NA
SRD5A1P1	NA	0.005	NA	NA	NA	0.065	NA	NA	NA

4.3.1.8 Functional enrichment of DEG is not observed

Functional enrichment analysis was performed on the DEG using DAVID [170, 171], to explore whether the DEG are enriched for specific biological processes which may provide insights into the pathology of CMTX3. Using the 'Functional Annotation Clustering' tool, four annotation clusters were observed, however, none of the annotation terms within any of the four clusters show statistically significant enrichment (FDR adjusted *p*-value > 0.05) (Table 4.3). Gene ontology terms represented within these clusters relate to (1) the immune response, (2) cellular organelles, (3) cellular membrane, and (4) DNA transcription (Table 4.3).

Table 4.3: Functional annotation clusters identified by DAVID enrichment analysis of the DEG identified in sMN. FDR = false discovery rate; % = percentage of genes associated with the term/total genes that were annotated by DAVID.

Annotation Cluster 1	Enrichment Score: 1.60653185					
Category	**Term**	**Count**	**%**	**PValue**	**Fold Enrichment**	**FDR**
GOTERM_BP_DIRECT	positive regulation of T cell mediated cytotoxicity	3	6.25	0.0008	68.2032	0.2651
GOTERM_MF_DIRECT	peptide antigen binding	3	6.25	0.0008	66.7407	0.0767
GOTERM_BP_DIRECT	adaptive immune response	3	6.25	0.0147	15.5007	0.4414
UP_KW_BIOLOGICAL_PROCESS	Adaptive immunity	3	6.25	0.0174	13.8069	0.4187
GOTERM_BP_DIRECT	immune response	3	6.25	0.0420	8.7891	0.6958
UP_KW_BIOLOGICAL_PROCESS	Immunity	4	8.33	0.0441	4.6788	0.5287
GOTERM_CC_DIRECT	cell surface	4	8.33	0.0589	4.3065	0.6310
REACTOME_PATHWAY	Adaptive Immune System	3	6.25	0.2560	2.8602	1.0000
REACTOME_PATHWAY		3	6.25	0.7280	1.1503	1.0000

Annotation Cluster 2	Enrichment Score: 0.82755241					
Category	**Term**	**Count**	**%**	**PValue**	**Fold Enrichment**	**FDR**
UP_KW_BIOLOGICAL_PROCESS	Immunity	4	8.33	0.0441	4.6788	0.5287
GOTERM_CC_DIRECT	Golgi membrane	4	8.33	0.1070	3.3360	0.7331
UP_KW_CELLULAR_COMPONENT	Endosome	3	6.25	0.3147	2.5426	1.0000
UP_KW_CELLULAR_COMPONENT	Cytoplasmic vesicle	3	6.25	0.3299	2.4548	1.0000

Annotation Cluster 3	Enrichment Score: 0.82496031					
Category	**Term**	**Count**	**%**	**PValue**	**Fold Enrichment**	**FDR**
GOTERM_CC_DIRECT	integral component of membrane	12	25.00	0.0555	1.7004	0.6310
UP_KW_CELLULAR_COMPONENT	Cell membrane	9	18.75	0.0776	1.8601	0.7758
UP_KW_CELLULAR_COMPONENT	Membrane	15	31.25	0.1526	1.3162	1.0000
GOTERM_CC_DIRECT	membrane	8	16.67	0.3173	1.3829	1.0000
GOTERM_CC_DIRECT	plasma membrane	9	18.75	0.3601	1.2824	1.0000

Annotation Cluster 4	Enrichment Score: 0.23076551					
Category	**Term**	**Count**	**%**	**PValue**	**Fold Enrichment**	**FDR**
GOTERM_BP_DIRECT	negative regulation of transcription from RNA polymerase II promoter	4	8.33	0.1757	2.6374	0.9881
GOTERM_MF_DIRECT	DNA binding	4	8.33	0.2540	2.1862	1.0000
REACTOME_PATHWAY	Generic Transcription Pathway	3	6.25	0.5430	1.5848	1.0000
REACTOME_PATHWAY	RNA Polymerase II Transcription	3	6.25	0.6022	1.4312	1.0000
REACTOME_PATHWAY	Gene expression (Transcription)	3	6.25	0.6724	1.2687	1.0000
UP_KW_MOLECULAR_FUNCTION	DNA-binding	3	6.25	0.7000	1.2059	1.0000
UP_KW_BIOLOGICAL_PROCESS	Transcription	4	8.33	0.8244	0.9077	1.0000
UP_KW_BIOLOGICAL_PROCESS	Transcription regulation	3	6.25	0.9375	0.7018	1.0000
GOTERM_CC_DIRECT	nucleus	7	14.58	0.9452	0.7221	1.0000
UP_KW_CELLULAR_COMPONENT	Nucleus	6	12.50	0.9814	0.6187	1.0000

4.3.2 NanoString targeted transcriptome analysis

In addition to RNA-Seq, the NanoString nCounter assay was a key tool for investigating the pathomechanism of the CMTX3 SV as it provides a robust, sensitive, and reproducible method for transcriptomic profiling of target genes [173].

4.3.2.1 NanoString assay excludes overexpression of the partially duplicated *ARHGAP39* gene in CMTX3 patient samples

One of the key hypotheses of this project was whether the CMTX3 SV leads to the formation of a novel transcript involving the partially duplicated copy of the *ARHGAP39* gene. To test this hypothesis, two probes targeting *ARHGAP39* were included on a custom designed NanoString panel. One probe targeted exon 6 of transcript variant 1 (probe 1), which is duplicated in CMTX3, and the other targeted the non-duplicated exon 9 (probe 2) (Figure 4.2). Based on the designed assay, evidence supporting this hypothesis would predict differential expression of probe 1, but not probe 2, in CMTX3 patients.

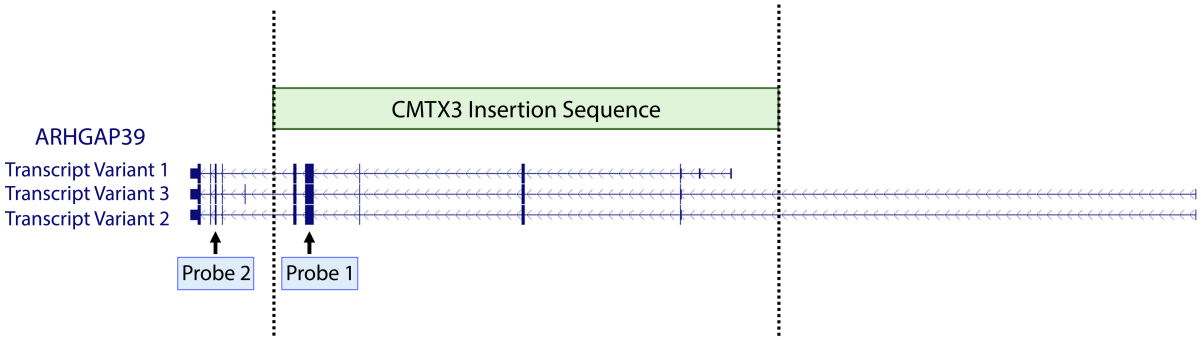


Figure 4.2: Strategy for designing NanoString probes to target *ARHGAP39*. Depiction of RefSeq curated transcript variants for *ARHGAP39*, viewed using UCSC Genome Browser. Vertical blue bars depict exons, horizontal lines depict introns, and arrows denote direction of transcription. Green box indicates the ~78 kb region that is duplicated in CMTX3 patients. The location of NanoString probes are specified below, with one targeting a duplicated exon (exon 6 of transcript variant 1) and one targeting a non-duplicated exon (exon 9 of transcript variant 1). Figure created with BioRender.

Differential expression of *ARHGAP39* (probe 1) was not observed between patients and controls across the 5 timepoints tested (Figure 4.3). As expected, probe 2 identified similar levels of *ARHGAP39* mRNA between patients and controls across all 5 timepoints (Figure 4.3).

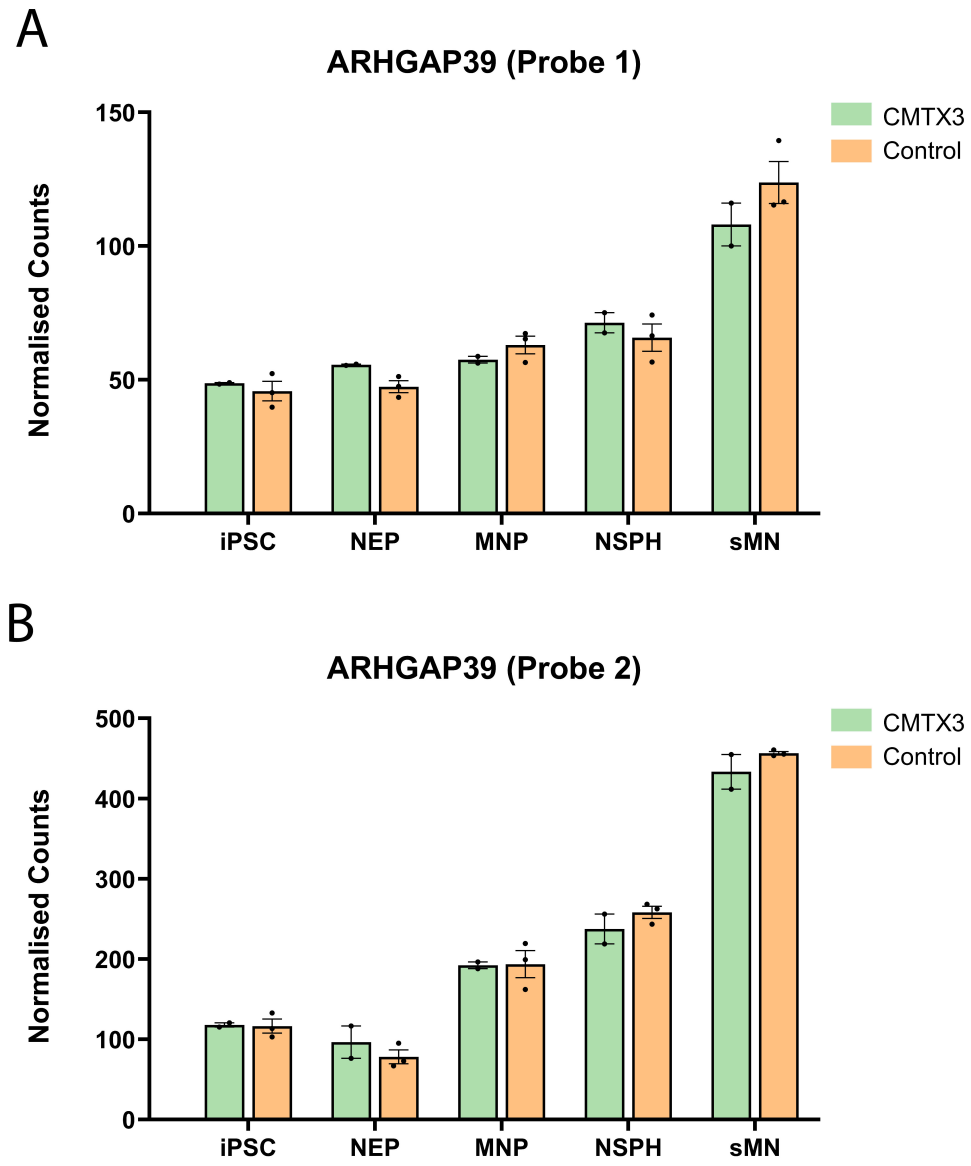


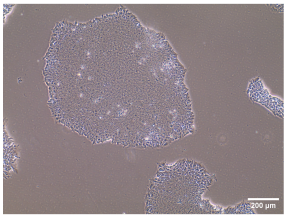
Figure 4.3: NanoString assay shows similar *ARHGAP39* expression between CMTX3 patients and controls. Normalised count data from (A) *ARHGAP39* probe 1, which targets an exon contained witin the CMTX3 complex SV, and (B) *ARHGAP39* probe 2, which targets an exon not affected by the ~78 kb CMTX3 complex SV.. The NanoString nCounter assay was performed on CMTX3 patient (n=2) and control (n=3) samples across 5 timepoints during sMN differentiation. Differential expression analysis was performed with nSolver software. Error bars represent mean ± SEM. iPSC; induced pluripotent stem cells, NEP; caudal neuroepithelial progenitors, MNP; motor neuron progenitors, NSPH; neurospheres, sMN; spinal motor neurons.

4.3.2.2 NanoString analysis identifies spatiotemporally specific *SOX3* dysregulation during CMTX3 sMN differentiation

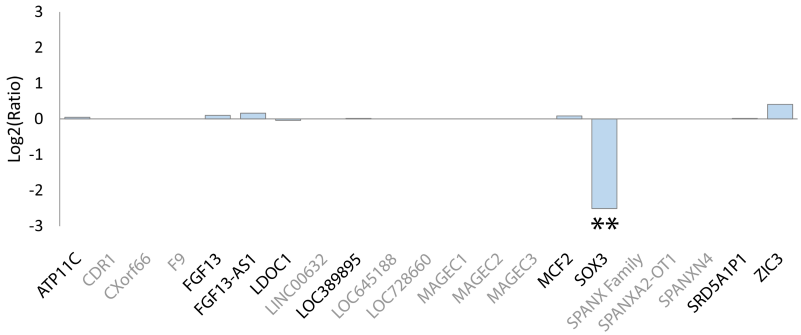
To investigate the hypothesis that the CMTX3 SV results in local transcriptional dysregulation, a custom NanoString panel was designed containing 21 probes to target genes located within 3 Mb flanking the CMTX3 SV breakpoint at Xq27.1. This panel targets all RefSeq curated genes within this genomic region, except for three microRNA genes (MIR504, MIR505 and MIR320D2). A single probe was designed to collectively target five SPANX family genes (SPANXA1, SPANXA2, SPANXB1, SPANXC, SPANXD), as the high sequence homology prohibited unique probe design. To assess spatiotemporal gene expression, the NanoString assay was performed on RNA isolated from 5 different timepoints throughout the sMN differentiation process; iPSC, caudal NEP, MNP, neurospheres, and mature sMN. At each timepoint, two patients and three controls were analysed.

For the iPSC timepoint, 12 of the 21 candidates did not show detectable levels of expression across all 5 samples (Figure 4.4A). For the remaining 9 candidates, *SOX3* was the only gene showing statistically significant differential expression between CMTX3 and controls (p = 0.0012) (Figure 4.4A). The log2(Ratio) of *SOX3* was -2.51, indicating a ~5.7-fold reduction in *SOX3* mRNA levels between CMTX3 and control iPSC (Figure 4.4A). For caudal NEP, the candidate genes did not show statistically significant differential expression between CMTX3 and control samples (Figure 4.4B). There is, however, a clear trend of reduced *SOX3* mRNA levels within the CMTX3 cells, with a log2(Ratio) of -1.75 (p = 0.157) (Figure 4.4B). Over the next 3 progressive timepoints during sMN differentiation, no CMTX3 candidate genes showed significant differences in expression levels between CMTX3 and control samples (Figure 4.4C-E). The high priority candidate gene, *FGF13*, previously shown to be dysregulated in lymphoblasts isolated from CMTX3 patients [63], was not dysregulated at any of the timepoints tested (Figure 4.4A-E).

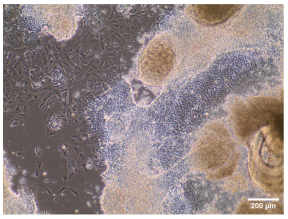
A i) iPSC



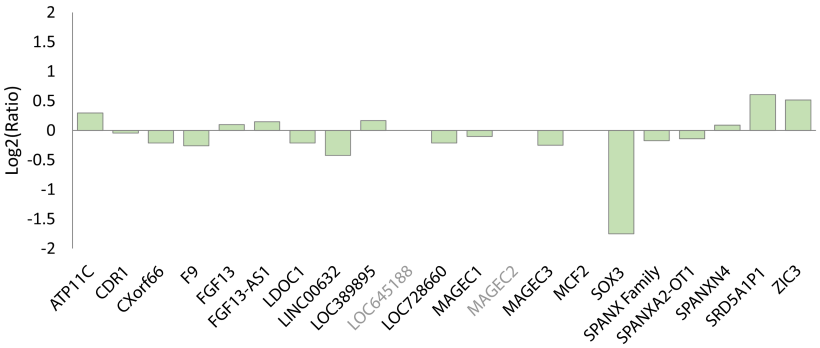
ii)



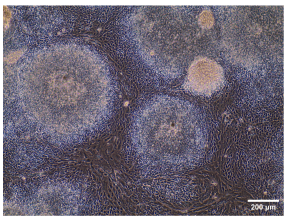
B i) Caudal NEP



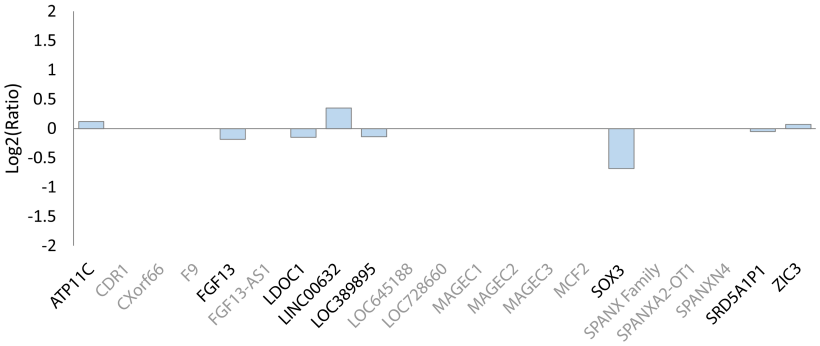
ii)



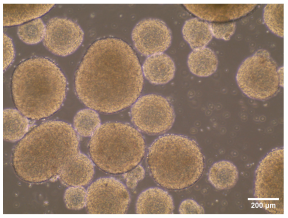
C i) MNP



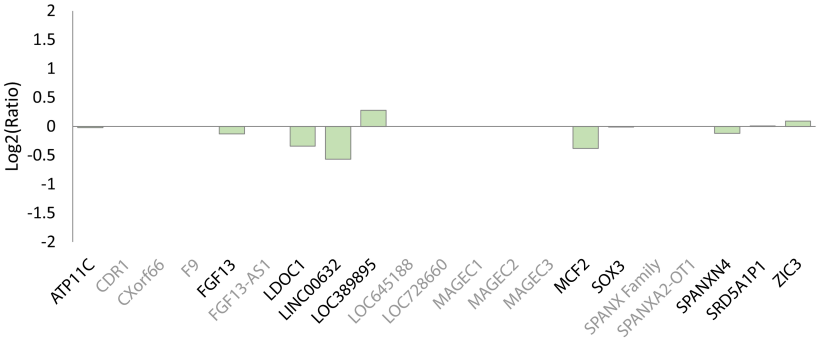
ii)



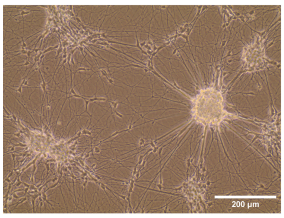
D i) Neurosphere



ii)



E i) sMN



ii)

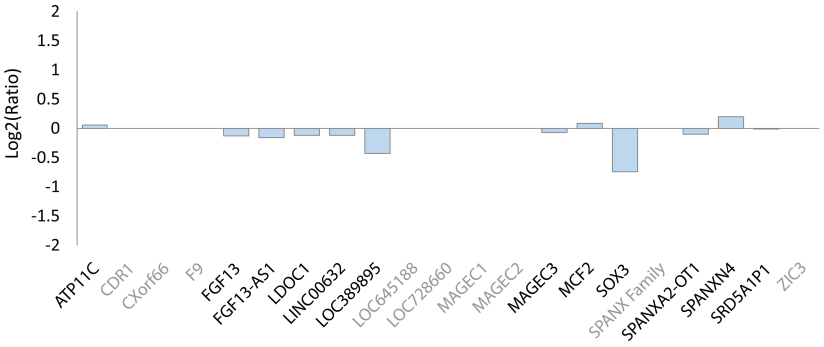


Figure 4.4: Comparison of CMTX3 candidate gene expression between patient and control cells at different stages of sMN differentiation. RNA from CMTX3 (n=2) and control (n=3) samples was isolated from 5 timepoints throughpout sMN differentiation; (A) iPSC; (B) caudal NEP, (C) MNP, (D) neurospheres, and (E) sMN. Expression of CMTX3 candidate genes was assessed using a custom NanoString nCounter panel. (i) Brightfield microscope images for each timepoint, showing the morphological changes that occur throughout sMN differentiation (scalebar = 200 µm). (ii) Relative expression of each candidate gene as the log2(Ratio) of CMTX3 vs control. Candidate genes with undetectable expression levels in all 5 samples are represented with grey font. Statistical analysis was performed using nSolver Analysis Software. $^{**}p < 0.01$.

4.3.2.3 Exploring *SOX3* expression throughout sMN differentiation

To further investigate the spatiotemporal expression of *SOX3*, we assessed the normalised count data obtained from the NanoString nCounter analyses. Figure 4.5 combines count data from all 5 timepoints tested, providing an insight into the regulation of *SOX3* across these discrete developmental stages during sMN differentiation. Here, the *SOX3* downregulation in CMTX3-derived iPSC can be clearly observed (Figure 4.5). A trend of downregulation is evident within the CMTX3 NEP population, and although statistical significance is not reached, the trend is consistent across the two patient-derived samples (Figure 4.5). It should also be noted that a preliminary experiment conducted on NEP RNA isolated from 1 patient and 3 controls displayed a similar trend of downregulation within the CMTX3 sample, suggesting this is a reproducible observation (Supplementary Figure 4.1). MNP and neurosphere cultures show the highest levels of *SOX3* expression, which is followed by a drastic reduction of *SOX3* mRNA as cells differentiate into mature sMN (Figure 4.5).

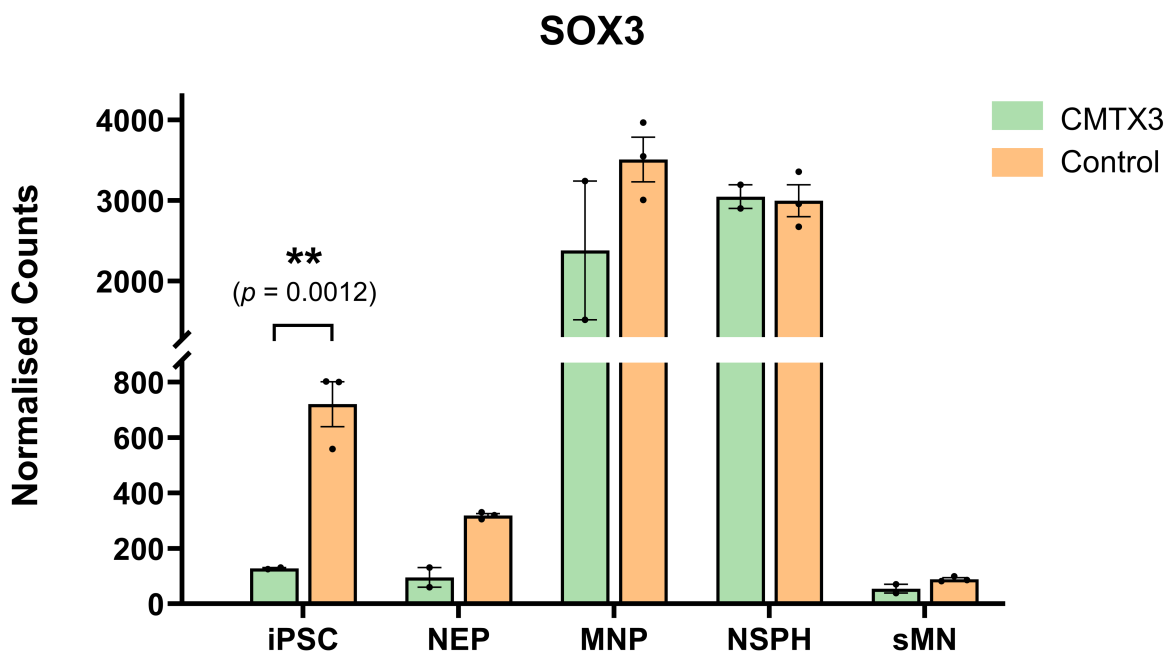


Figure 4.5: *SOX3* expression at multiple timepoints throughout the differentiation of iPSC to sMN. Normalised count data for *SOX3* in CMTX3 patients (n=2) and controls (n=3) across 5 specific timepoints during sMN differentiation. Count data was obtained using the NanoString nCounter platform and differential expression analysis performed with nSolver software. ** $p < 0.01$.

4.3.3 RNA-Seq analysis of neuroepithelial progenitor (NEP) cells

Investigation into the function of *SOX3* revealed that this gene encodes an embryonic transcription factor belonging to the SOXB1 family of transcription factors [174]. *SoxB1* genes are widely expressed within the developing central nervous system [175]. SOX3 is typically considered a transcriptional activator [176, 177], although transcriptional repression has also been reported [178]. This gene is predominately expressed within neural precursors, where it maintains cells in an undifferentiated, progenitor state thereby preventing premature differentiation [177]. RNA-Seq analysis was therefore performed to investigate the transcriptome of neural progenitors in the context of CMTX3, where the spatiotemporal regulation of *SOX3* is perturbed. RNA was extracted from caudal NEP derived from CMTX3 patients (n=2) and controls (n=3). Library preparation, sequencing, and data-processing was outsourced to Macrogen Oceania (Seoul, South Korea). A summary of quality control metrics for each library are provided in Supplementary Table 4.9.

4.3.4 Differential expression analysis of CMTX3 patient and control NEP

Many bioinformatics packages are available to perform differential expression analysis on RNA-Seq datasets and can show substantial variability in their detection of DEG [179]. There is currently no gold standard approach for this type of analysis [180]. Two tools were used for exploring the datasets and included DESeq2 [168] and edgeR [167], which are popular choices for datasets with few biological replicates [181]. Both tools are parametric approaches that assume a negative binomial distribution [167, 168].

4.3.4.1 DESeq2 differential expression analysis

A Principal Components Analysis (PCA) plot was used to visually assess the relationship between all samples. Figure 4.6 clearly shows that the samples do not cluster based on their disease status, indicating that this does not account for the majority of variance between samples.

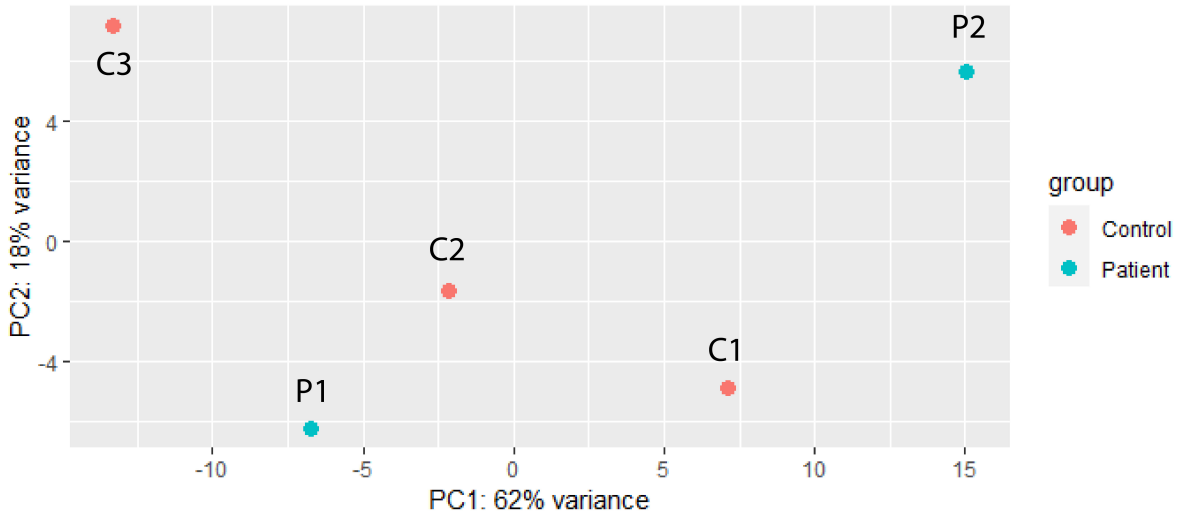


Figure 4.6: Principle Component Analysis plot for RNA-Seq data. PCA plot generated using DESeq2 to visualise overall similarity between gene expression data across the 5 samples. Samples do not cluster based on their disease status.

After filtering out low count genes, 19,867 genes were included in the differential gene expression analysis. A total of 14 DEG were identified, including three upregulated and 11 downregulated in the CMTX3 NEP (Table 4.4). Only one of the 14 DEG localised to chromosome X (Table 4.4). This was the *SOX3* gene, confirming the previous trend of reduced *SOX3* expression within

patient-derived NEP (Figure 4.4). In this analysis, the CMTX3 NEP show a ~ 3.89-fold reduction (log2(Fold change) = -1.96) in *SOX3* mRNA levels relative to controls (*p*-adj = 0.0015) (Table 4.4). The NanoString assay and the RNA-Seq analysis were performed on RNA isolated from separate rounds of differentiation, suggesting this is a reproducible finding.

Table 4.4: List of DEG identified by DESeq2 analysis of RNA-Seq data obtained from CMTX3 and control NEP. Log2FoldChange was calculated using 'control' as the reference group, with the default significance threshold of *p*-adjusted < 0.1. The CMTX3 candidate gene, *SOX3*, is highlighted in green. Genes are ranked according to adjusted *p*-value.

Gene	log2FoldChange	Adjusted *p* -value	Chr
LINC01139	3.671	1.30861E-11	1
EIF2S3B	3.781	1.58463E-06	12
TMX2-CTNND1	3.398	9.82779E-06	11
PLOD2	-1.151	1.25649E-05	3
SOX3	-1.962	0.00149	X
REC8	-1.948	0.00303	14
TRNY	-1.127	0.00367	MT
AHRR	-9.207	0.01918	5
TMEM132C	-1.343	0.01954	12
PLEKHA2	-1.277	0.02150	8
TRNS1	-1.137	0.03343	MT
LOC112268219	-7.415	0.03343	1
MIR600HG	-4.269	0.03622	9
CASKIN1	-0.910	0.06108	16

4.3.4.2 edgeR differential expression analysis

After applying the recommended low-count filtering method within edgeR, 19,975 genes were included in the differential expression analysis. A total of 16 DEG were detected using the default significance threshold (FDR < 0.05), and included 7 downregulated and 9 upregulated in CMTX3 NEP relative to controls (Table 4.5). Whilst *SOX3* was not included in this list, inspection of the results indicate that the FDR only slightly exceeded the significance threshold (FDR = 0.0506). Eight of the DEG identified by edgeR were also detected by DESeq2. Interestingly, four of the DEG identified in the NEP dataset (*EIF2S3B*, *TMEM132C, RPL10P9, HLA-DRB1*) were also predicted to be differentially

expressed in the exploratory RNA-Seq performed on sMN (Supplementary Table 4.8), with consistent log2(FoldChange) directionality.

Table 4.5: List of DEG identified by edgeR analysis of RNA-Seq data from CMTX3 and control NEP. Log2FoldChange was calculated using 'control' as the reference group, with the default cutoff FDR < 0.05. Genes ranked according to FDR. FDR; false discovery rate.

Gene	log2FoldChange	FDR	Chr
LINC01139	3.660	2.33E-06	1
TEKT4P2	11.275	1.33E-05	21
EIF2S3B	3.765	5.94E-05	12
HLA-DRB1	7.946	9.09E-05	6
TMX2-CTNND1	3.392	0.00043	11
LOC112268219	-7.960	0.00064	1
PLOD2	-1.156	0.00328	3
BORCS7-ASMT	-7.637	0.00594	10
REC8	-1.954	0.00594	14
TMX2P1	6.849	0.01027	9
MRPL23-AS1	-7.212	0.01368	11
CCDC38	6.805	0.01416	12
TMEM132C	-1.348	0.02414	12
TRNY	-1.132	0.02414	MT
TDGF1	6.514	0.03495	3
RPL10P9	6.588	0.03839	5

4.3.4.3 Functional enrichment of the DEG identified in NEP is not observed

To explore possible biological processes or functions that may be impacted in the CMTX3-derived NEP, a functional enrichment analysis was performed on the DEG. This analysis was done using the DAVID Bioinformatics Resource, pooling the genes identified from both DESeq2 and edgeR. The 'Functional Annotation Clustering' tool identified one annotation cluster, however, none of the associated annotation terms showed statistically significant enrichment (FDR adjusted *p*-value > 0.05) (Table 4.6). Gene ontology terms within this cluster relate to the nucleus and protein binding (Table 4.6).

Table 4.6: Functional annotation clusters identified by DAVID enrichment analysis of the DEG identified in NEP. FDR = false discovery rate; % = percentage of genes associated with the term/total genes that were annotated by DAVID.

Annotation Cluster 1	Enrichment Score: 0.12318188					
Category	**Term**	**Count**	**%**	**PValue**	**Fold Enrichment**	**FDR**
UP_KW_CELLULAR_COMPONENT	Nucleus	4	21.05	0.6429	1.1677	1.0000
GOTERM_MF_DIRECT	protein binding	6	31.58	0.7529	1.0023	1.0000
GOTERM_CC_DIRECT	nucleus	3	15.79	0.8822	0.8507	1.0000

Although not the focus of this section, we did inspect the NEP RNA-Seq datasets for the presence of fusion transcripts, novel transcript variants, and novel splice variants. Concordant with the previous findings (Section 4.3.1.2), no novel features involving the genomic regions implicated in the CMTX3 SV were detected.

4.3.5 Preliminary western blot explores SOX3 protein expression

Western blot was used to assess SOX3 expression at the protein level. Total protein was extracted from NEP derived from CMTX3 patients (n=2) and controls (n=3). Figure 4.7 shows preliminary data in which a trend of reduced SOX3 expression is detected in CMTX3 NEP relative to controls, although this did not reach statistical significance (p = 0.1009). Due to technical issues, the replication of this experiment was not possible within the timeline of the current project. Therefore, validating this promising result will be required.

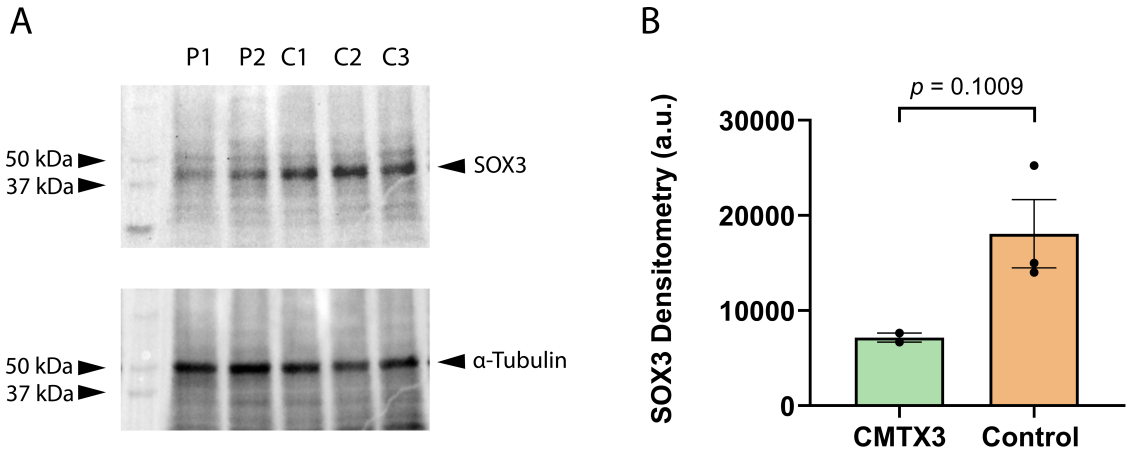


Figure 4.7: Preliminary western blot showing SOX3 expression in CMTX3 and control-derived NEP. (A) Western blots show the expression of SOX3 (~ 45 kDa) and the loading control protein, a-Tubulin (~50 kDa). Protein was extracted from two CMTX3 patients (P1-2) and three control (C1-3) NEP, with 20 µg loaded per lane. (B) Quantification of western blot images using FIJI. SOX3 levels were normalised to α-Tubulin. Means were compared using an unpaired, two-tailed t-test. Error bars represent mean ± SEM. Results are from one experimental replicate.

4.4 DISCUSSION

At the commencement of this project, there were two primary hypotheses proposed for testing the underlying pathogenic mechanism of the CMTX3 SV. The first posits that the CMTX3 genomic rearrangement gives rise to a novel transcript, while the second suggests that the pathogenic SV may alter the transcriptional regulation of nearby genes [63]. In this chapter, we addressed both hypotheses by using multiple methods, including RNA-Seq and NanoString nCounter assays, to assess the transcriptional consequences of the CMTX3 SV using disease-relevant cell types. A similar experimental pipeline was recently utilised to investigate the pathomechanism of the DHMN1 SV, which led to the discovery of a novel gene-intergenic fusion transcript arising from the genomic rearrangement [67].

4.4.1 Novel transcripts involving ARHGAP39 were not identified using RNA-Seq or NanoString

Fusion genes have been well-documented within the context of cancer [182], however, the recent discovery of the DHMN1 gene-intergenic fusion represents the first within the field of IPN [67]. The 1.35 Mb SV causing DHMN1 harbours a partial duplication of *UBE3C*, which was shown to produce a fusion transcript by incorporating sequence from adjacent non-coding DNA [67]. We set out to explore whether a similar mechanism might occur in CMTX3, involving the partially duplicated copy of *ARHGAP39*. Although three independent fusion prediction algorithms were utilised to assess the sMN RNA-Seq data, patient-specific fusion transcripts involving *ARHGAP39* were not detected. In contrast, two of the three programs used in this study were successful in predicting the DHMN1 SV [67], demonstrating their capacity to detect complex gene-intergenic fusions. Our search was expanded to include other novel features, such as novel splice variants or novel transcript variants, that may theoretically arise from the presence of the CMTX3 SV. Patient-specific novel transcripts involving *ARHGAP39* were not detected.

The results of the NanoString assays are consistent with this finding. Similar levels of *ARHGAP39* mRNA were observed between patients and controls, despite using a probe to selectively target an exon that is duplicated in CMTX3 patients. This result was consistently observed across the 5 discrete timepoints that were tested throughout the sMN differentiation process (iPSC, NEP, MNP, neurospheres, sMN).

In summary, this chapter presents multiple lines of evidence which independently do not support the hypothesis that the partially duplicated *ARHGAP39* gives rise to an aberrant transcript in CMTX3. This is a key finding for progressing our understanding of the pathomechanism of CMTX3 neuropathy.

4.4.2 FGF13 dysregulation was not observed in CMTX3 patient-derived neural cell types

FGF13 initially presented as a high-priority candidate gene for CMTX3 which warranted careful investigation, as it was previously shown to be dysregulated in lymphoblasts isolated from a CMTX3

patient [63]. Furthermore, rodent studies have demonstrated the important role of *FGF13* in neural development and neuron excitability (reviewed in [156]). However, a lack of disease-relevant models for CMTX3 precluded further investigation into this candidate gene, and it remained unclear whether *FGF13* dysregulation is implicated in CMTX3 pathogenesis or merely a bystander effect of the mutation [63]. The present study represents the first investigation of *FGF13* expression levels within patient-derived disease-relevant tissue.

NanoString assays showed similar *FGF13* expression levels between patient and control samples isolated from five discrete timepoints throughout sMN development. RNA-Seq data was available for two of these timepoints (NEP, sMN), which confirmed these results. Thus, the *FGF13* upregulation observed in patient lymphoblasts was not recapitulated in the neuronal cell types tested here. This suggests that the CMTX3 SV impacts spatiotemporal gene regulation in a manner that is highly dependent upon cellular context. Tissue-specific gene dysregulation was also observed in X-linked congenital generalised hypertrichosis, which is caused by a ~389 kb inter-chromosomal insertion occurring within the Xq27.1 quasi-palindrome [152]. The authors showed that patients harbouring the complex SV display reduced *FGF13* mRNA within skin biopsies, and demonstrated that this downregulation selectively affects keratinocytes and not fibroblasts [152]. Together, these results reinforce the importance of using disease-relevant tissue to explore the impact of complex SV upon local gene regulation.

A limitation of the present study is being unable to distinguish between the various transcript variants of *FGF13*. While differential expression analyses are often conducted at the gene-level, bioinformatics tools are available for assessing isoform-level differences (e.g. [183, 184], which could be useful for the *FGF13* analysis. However, obtaining accurate isoform-level information from short-read RNA-Seq is inherently limited by the ambiguity of mapping the short reads to the appropriate isoform [185, 186]; an issue that can be overcome with long-read sequencing technologies [185, 187]. Of note, the initial report of *FGF13* dysregulation in CMTX3 derived lymphoblasts was detected at the

gene-level [63], as was the *FGF13* dysregulation observed in X-linked congenital generalised hypertrichosis [152].

4.4.3 Transcriptomic analyses identify SOX3 as a high-priority candidate gene for CMTX3

Transcriptomic analyses revealed *SOX3* dysregulation in CMTX3 patient-derived samples, which was only observed within a discrete developmental window throughout the differentiation of iPSC into mature sMN. The NanoString assay identified a ~5.7-fold reduction in *SOX3* mRNA in patient-derived iPSC relative to controls. Although not statistically significant, a trend of *SOX3* downregulation was also observed in patient-derived caudal NEP. RNA-Seq analysis validated this trend, with patient NEP showing ~3.9-fold reduced expression of this gene. However, in subsequent developmental timepoints (MNP, neurospheres, sMN), *SOX3* expression was similar between patients and controls. Importantly, no other candidate genes within the CMTX3 linkage region showed differential expression between patient and control samples, at any of the 5 timepoints tested. Taken together, these results suggest that the CMTX3 SV perturbs the spatiotemporal control of *SOX3* throughout sMN differentiation and thereby identifies *SOX3* as a high-priority candidate for further investigation. A preliminary western blot analysis revealed a similar trend of SOX3 downregulation within CMTX3 patient-derived NEP, although this did not reach statistical significance. A priority for future studies will therefore be to validate this downregulation at the protein level.

SOX3 is the closest gene to the CMTX3 SV, located ~ 82 kb telomeric to the insertion breakpoint, and is a member of the SOX (SRY-related HMG-box) family of transcription factors, which are important developmental regulators [188, 189]. SOX transcription factors can behave as classical transcription factors as well as pioneer factors, which can bind silent genes and predispose them for activation at later developmental stages [190, 191]. Functionally, this gene represents an interesting candidate, as the SOXB1 family of transcription factors, comprised of SOX1, SOX2 and SOX3 [176], have a known role in neuronal differentiation (reviewed in [191]). Specifically, Sox3 has been shown to maintain neural precursors in the progenitor state, preventing premature neurogenesis [177]. Evidence also suggests SOX3 may perform a similar role to prevent premature differentiation of

astrocytes within the developing spinal cord [192]. Given the important role of the SOX3 transcription factor within neural precursor populations, we performed RNA-Seq on samples from the NEP timepoint, to explore the global transcriptional consequences of *SOX3* downregulation in CMTX3. A total of 22 genes showed differential expression between patients and controls; 14 identified by DESeq2 [168], 16 identified by edgeR [167], with 8 being common to both pipelines. This may appear a surprisingly low number, given that SOX3 has previously been reported to bind more than 8,000 genomic sites in murine neural precursor cells [193]. However, a similar observation was made by Rogers and colleagues (2014) when investigating Sox3-null murine neural precursors, with only 19 DEG identified between Sox3-null and control cells [194]. Given the functional redundancy between SOXB1 transcription factors [177, 195], the authors postulate that SOX1/SOX2 may therefore compensate for the loss of SOX3 at many, but not all, genomic loci [194]. However, further studies are necessary to investigate whether the DEG identified in the current study are direct targets of the SOX3 transcription factor. This may include ChIP-Seq analysis, to investigate SOX3 binding profiles within the caudal NEP used in this study.

Functional enrichment analysis was performed on the DEG with the aim of identifying biological pathways or processes that may be perturbed within CMTX3 neuronal cell types. This type of analysis can provide valuable clues into the functional consequences of complex SV, as seen in a recent study investigating an 89 Mb inversion causing branchio-oculo-facial syndrome (BOFS) [107]. Genes showing downregulation in BOFS patients were enriched for terms associated with cell adhesion and migration [107]. This was subsequently explored using *in vitro* migration assays, which uncovered a phenotype of migratory defects within BOFS patient iPSC-derived neural crest cells [107]. Within the present study, the DEG identified from both the sMN and NEP RNA-Seq datasets were not significantly enriched for functional annotation terms. This may relate to the relatively small sizes of the gene lists being analysed in this study, particularly in the NEP DEG list (22 genes; 19 annotated by DAVID), as the statistical power of DAVID is increased with larger lists [170]. In such cases, DAVID analysis may be used in an exploratory capacity to examine the annotations associated with the genes of interest [170]. It is

interesting to note that general terms such as 'immune response', 'membrane' and 'DNA transcription' emerged as annotation clusters in the sMN DEG analysis. However, considering that the sMN analysis was conducted using data from a single CMTX3 patient, the generalisability of these results is unknown. Increasing biological replicates for RNA-Seq will improve sensitivity to detect DEG [181, 196], and may therefore also improve statistical power for a functional enrichment analysis. One study recommends using a minimum of six biological replicates when conducting differential expression analyses [181], although, this is a challenging task within the context of rare genetic disease.

A limitation of the present study was the inability to assess the expression level of the miRNA genes within the CMTX3 linkage region (*MIR504*, *MIR505*, *MIR320D2*). This study design was optimised for the accurate quantification of coding genes, and long non-coding genes, within the CMTX3 locus. This included the use of a poly(A) selection approach for RNA-Seq, which enriches RNA-Seq libraries for polyadenylated transcripts (mRNA and a subset of lncRNA) [197]. Similarly, the custom NanoString panel was designed to assay coding and lncRNA gene expression within the region. Whilst this represents a logical approach for conducting an initial investigation into the pathomechanism of CMTX3, we cannot exclude the possibility that nearby miRNA genes may be dysregulated by the complex insertion.

Overall, the present study has assessed the impact of the CMTX3 SV upon the local and global transcriptome within disease-relevant neuronal cell types. The results of the RNA-Seq and NanoString assays collectively do not support the hypothesis that the CMTX3 SV gives rise to an aberrant transcript. However, both assays independently identified *SOX3* downregulation in CMTX3 patient samples, supporting the hypothesis that local gene dysregulation may underly CMTX3 neuropathy. Local gene dysregulation has previously been reported in XX-male sex reversal [151] and X-linked congenital generalised hypertrichosis [152], both caused by complex inter-chromosomal insertions occurring at the Xq27.1 quasi-palindrome, suggesting this may represent a common pathomechanism at this disease hotspot. Interestingly, the CMTX3 SV appears to alter the spatiotemporal control of *SOX3* within a discrete developmental window as cells differentiate from iPSC to sMN. Further

characterising and validating the timeline of *SOX3* dysregulation, and exploring its functional impact, will be important in determining whether this gene dysregulation drives the neuropathy phenotype observed in CMTX3 patients.

Supplementary Material: Chapter 4

Supplementary Table 4.1: Quality control metrics for sMN RNA-Seq libraries. GC = guanine-cytosine content; Q20 = ratio of bases with Phred quality score ≥ 20; Q30 = ratio of bases with Phred quality score ≥ 30.

	Raw Statistics				Pre-processing & mapping statistics		
Sample ID	**Total reads**	**GC (%)**	**Q20 (%)**	**Q30 (%)**	**Trimmed reads**	**Mapped reads (%)**	**Unmapped reads (%)**
CMTX3 Clone A	121,740,578	49.66	98.34	95.1	120,554,518	98.67%	1.33%
CMTX3 Clone B	126,533,932	49.69	98.34	95.08	125,349,620	98.91%	1.09%
Control 1	131,312,724	49.94	98.24	94.93	129,723,764	98.93%	1.07%
Control 2	121,070,090	49.79	98.28	94.97	119,826,344	98.72%	1.28%
Control 3	131,694,598	49.6	98.25	94.9	130,283,856	98.89%	1.11%

Supplementary Table 4.2: Fusion transcripts, predicted by deFuse, on chromosome X from CMTX3 patient-derived sMN.

Sample	Gene1	Gene2	Gene1 Chr	Gene2 Chr	Genomic Strand1	Genomic Strand2	Genomic Break Position1	Genomic Break Position2	Probability
CMTX3 Clone A	LINC01201	ARHGAP36	X	X	+	-	131031253	131067378	0.9421
CMTX3 Clone A	SPRY3	TMLHE	X	X	-	+	155767959	155612644	0.7459
CMTX3 Clone A	ARMCX4	ARMCX1	X	X	+	-	101532012	101550971	0.9540
CMTX3 Clone A	RF00019	RAB28P5	X	X	+	-	136820821	136847805	0.7846
CMTX3 Clone A	RANP4	RF00019	X	X	+	-	136807922	136814865	0.9453
CMTX3 Clone A	GPRASP2	BHLHB9	X	X	+	-	102714266	102720534	0.8684
CMTX3 Clone A	ARHGAP36	LINC01201	X	X	-	+	131081760	131051650	0.8998
CMTX3 Clone A	PUDP	STS	X	X	+	-	7148081	7190878	0.8148
CMTX3 Clone A	STS	PUDP	X	X	-	+	7253197	7148084	0.8167
CMTX3 Clone A	ARHGAP36	LINC01201	X	X	-	+	131081759	131051649	0.9166
CMTX3 Clone A	FAAH2	AL022157.1	X	X	-	+	57287016	57127056	0.6835
CMTX3 Clone A	PDZD4	IDH3G	X	X	-	+	153805097	153790851	0.9719
CMTX3 Clone A	HDAC6	ERAS	X	X	+	-	48823588	48829081	0.8740
CMTX3 Clone A	RF00019	RAB28P5	X	X	+	-	136815112	136847805	0.8792
CMTX3 Clone A	SLC25A6	IL3RA	X	X	-	+	1386681	1378811	0.5493
CMTX3 Clone A	LINC01201	ARHGAP36	X	X	+	-	131039856	131067381	0.9317
CMTX3 Clone A	LINC01201	ARHGAP36	X	X	+	-	130981856	131067378	0.9414
CMTX3 Clone A	ARHGAP36	LINC01201	X	X	-	+	131081759	131051649	0.5631
CMTX3 Clone B	SLC25A6	IL3RA	X	X	-	+	1386683	1378813	0.6244
CMTX3 Clone B	HDAC6	ERAS	X	X	+	-	48823588	48829081	0.7247
CMTX3 Clone B	LINC01201	ARHGAP36	X	X	+	-	130981859	131067381	0.6261
CMTX3 Clone B	MTMR8	ASB12	X	X	-	+	64348660	64225674	0.7892
CMTX3 Clone B	ARHGAP36	LINC01201	X	X	-	+	131081759	131051649	0.7660
CMTX3 Clone B	STS	PUDP	X	X	-	+	7253197	7148084	0.6232
CMTX3 Clone B	RANP4	RF00019	X	X	+	-	136807923	136814866	0.7704
CMTX3 Clone B	RF00019	RAB28P5	X	X	+	-	136820821	136847805	0.8673
CMTX3 Clone B	ARHGAP36	LINC01201	X	X	-	+	131081759	131051649	0.8930
CMTX3 Clone B	TMSB15B	AC234783.1	X	X	+	+	103962723	104062987	0.6209
CMTX3 Clone B	SPRY3	TMLHE	X	X	-	+	155767962	155612647	0.7699
CMTX3 Clone B	LINC01201	ARHGAP36	X	X	+	-	131039856	131067381	0.6798
CMTX3 Clone B	ARHGAP36	LINC01201	X	X	-	+	131058444	131051649	0.6282
CMTX3 Clone B	ARMCX5-GPRASP2	BHLHB9	X	X	+	-	102714266	102720534	0.7965

Supplementary Table 4.3: Fusion transcripts, predicted by deFuse, on chromosome 8 from CMTX3 patient and control sMN. A fusion transcript involving *ARHGAP39*, which is partially duplicated in CMTX3, was identified yet is found in controls as well as patient samples.

Sample	Gene1	Gene2	Gene1 Chr	Gene2 Chr	Genomic Strand1	Genomic Strand2	Genomic Break Position1	Genomic Break Position2	Probability
CMTX3 Clone A	NEFL	AC107373.1	8	8	+	-	24962147	25008393	0.836
CMTX3 Clone A	CCNE2	INTS8	8	8	-	+	94894235	94876489	0.884
CMTX3 Clone A	ARHGAP39	RF00017	8	8	+	-	144605696	144685687	0.581
CMTX3 Clone A	AC120193.1	AF106564.1	8	8	+	-	24884125	24912878	0.662
CMTX3 Clone A	RNU6-663P	AC090204.1	8	8	+	-	32928093	33041635	0.872
CMTX3 Clone A	RNU6-663P	AC090204.1	8	8	+	-	32928093	32996201	0.872
CMTX3 Clone B	TUSC3	PPM1AP1	8	8	+	-	15748465	15806363	0.951
CMTX3 Clone B	TSTA3	AC067930.3	8	8	+	-	143618314	143632523	0.715
CMTX3 Clone B	AC021321.1	AC021321.1	8	8	+	-	67344302	67373715	0.766
CMTX3 Clone B	AC107373.2	AC107373.1	8	8	+	-	24962147	25008393	0.746
CMTX3 Clone B	ARHGAP39	RF00017	8	8	+	-	144605695	144685686	0.954
CMTX3 Clone B	CCNE2	INTS8	8	8	-	+	94894235	94876489	0.894
CMTX3 Clone B	RNU6-663P	AC090204.1	8	8	+	-	32928093	33041635	0.807
CMTX3 Clone B	TDRP	FBXO25	8	8	-	+	494494	457100	0.954
Control 1	RNU6-663P	AC090204.1	8	8	+	-	32928093	32996201	0.919
Control 1	INTS8	CCNE2	8	8	+	-	94876489	94894235	0.987
Control 1	NEFL	AC107373.1	8	8	+	-	24962148	25008394	0.939
Control 1	RNU6-663P	AC090204.1	8	8	+	-	32928093	33041635	0.948
Control 1	ARHGAP39	RF00017	8	8	+	-	144605695	144685686	0.984
Control 2	RNU6-663P	AC090204.1	8	8	+	-	32928093	32996201	0.913
Control 2	RNU6-663P	AC090204.1	8	8	+	-	32928093	33041635	0.895
Control 2	INTS8	CCNE2	8	8	+	-	94876489	94894235	0.977
Control 2	AC107373.2	AC107373.1	8	8	+	-	24962147	25008393	0.840
Control 2	ARHGAP39	RF00017	8	8	+	-	144605695	144685686	0.966
Control 3	AC021321.1	AC021321.1	8	8	+	-	67344302	67373715	0.912
Control 3	NEFL	AC107373.1	8	8	+	-	24962148	25008394	0.851
Control 3	TUSC3	PPM1AP1	8	8	+	-	15748463	15806361	0.517
Control 3	CCNE2	INTS8	8	8	-	+	94894235	94876489	0.966
Control 3	TDRP	FBXO25	8	8	-	+	494494	457100	0.839
Control 3	ARHGAP39	RF00017	8	8	+	-	144605695	144685686	0.981
Control 3	AC120193.1	AF106564.1	8	8	+	-	24884125	24912878	0.524
Control 3	XPO7	NPM2	8	8	+	-	22004030	22025436	0.828
Control 3	RNU6-663P	AC090204.1	8	8	+	-	32928093	33041635	0.914
Control 3	NRBP2	PUF60	8	8	-	+	143835544	143830101	0.908
Control 3	LY6E	C8orf31	8	8	+	-	143023875	143031937	0.932
Control 3	RNU6-663P	AC090204.1	8	8	+	-	32928093	32996201	0.951

Supplementary Table 4.4: All fusion transcripts, predicted by FusionCatcher, with breakpoints spanning chromosome X and chromosome 8 within CMTX3 patient-derived sMN.

Sample	Gene1	Gene2	Fusion point (gene1)	Fusion point (gene2)
CMTX3 Clone A	NEFL	BEX1	8:24953569:-	X:103062974:-
CMTX3 Clone A	NEFL	BEX1	8:24953618:-	X:103063284:-
CMTX3 Clone A	NEFL	BEX4	8:24951554:-	X:103216685:+
CMTX3 Clone A	NEFL	BEX4	8:24952573:-	X:103216778:+
CMTX3 Clone A	NEFL	BEX3	8:24956465:-	X:103377554:+
CMTX3 Clone A	STMN2	DCX	8:79666118:+	X:111294611:-
CMTX3 Clone A	DPYSL2	DCX	8:26656942:+	X:111296650:-
CMTX3 Clone A	NEFL	DCX	8:24952570:-	X:111297205:-
CMTX3 Clone A	NEFL	DCX	8:24956405:-	X:111299248:-
CMTX3 Clone A	NEFL	DCX	8:24955531:-	X:111299901:-
CMTX3 Clone A	NEFM	DCX	8:24917509:+	X:111300904:-
CMTX3 Clone A	YWHAZ	DCX	8:100919927:-	X:111301503:-
CMTX3 Clone A	NEFM	DCX	8:24914544:+	X:111301599:-
CMTX3 Clone A	NEFL	TMSB4X	8:24955612:-	X:12976275:+
CMTX3 Clone A	SARAF	TMSB4X	8:30063573:-	X:12976824:+
CMTX3 Clone A	SARAF	TMSB4X	8:30063184:-	X:12976926:+
CMTX3 Clone A	NEFL	TMSB4X	8:24953619:-	X:12976938:+
CMTX3 Clone A	NEFL	GLRA2	8:24955743:-	X:14607131:+
CMTX3 Clone A	NEFL	GLRA2	8:24954180:-	X:14609149:+
CMTX3 Clone A	NEFL	IDS	8:24953505:-	X:149479488:-
CMTX3 Clone A	NEFL	IDS	8:24956447:-	X:149479694:-
CMTX3 Clone A	NEFL	PDZD4	8:24952025:-	X:153804538:-
CMTX3 Clone A	NEFL	PDZD4	8:24956024:-	X:153808188:-
CMTX3 Clone A	NEFL	PDZD4	8:24955647:-	X:153830298:-
CMTX3 Clone A	STMN4	L1CAM	8:27240011:-	X:153862787:-
CMTX3 Clone A	NEFM	L1CAM	8:24918300:+	X:153867434:-
CMTX3 Clone A	NEFL	L1CAM	8:24952837:-	X:153868972:-
CMTX3 Clone A	STMN4	L1CAM	8:27236486:-	X:153875893:-
CMTX3 Clone A	NEFL	GDI1	8:24956459:-	X:154440049:+
CMTX3 Clone A	DPYSL2	GDI1	8:26657052:+	X:154442754:+
CMTX3 Clone A	NEFL	TSPAN7	8:24955695:-	X:38674226:+
CMTX3 Clone A	NEFL	USP11	8:24951167:-	X:47247904:+
CMTX3 Clone A	NEFL	SYP	8:24956009:-	X:49189047:-
CMTX3 Clone A	NEFL	SYP	8:24951923:-	X:49199012:-
CMTX3 Clone A	NEFL	NUDT10	8:24953677:-	X:51336934:+
CMTX3 Clone A	NEFL	RPS4X	8:24951447:-	X:72273346:-
CMTX3 Clone A	NEFL	RLIM	8:24956194:-	X:74588938:-
CMTX3 Clone A	DCX	PABPC1	X:111294496:-	8:100706776:-
CMTX3 Clone A	DCX	PABPC1	X:111295974:-	8:100722068:-
CMTX3 Clone A	HS6ST2	YWHAZ	X:132626913:-	8:100918941:-
CMTX3 Clone A	L1CAM	NEFM	X:153863883:-	8:24914288:+
CMTX3 Clone A	EIF1AX	NEFM	X:20127018:-	8:24914288:+
CMTX3 Clone A	EIF1AX	NEFM	X:20127844:-	8:24914402:+
CMTX3 Clone A	MAGED1	NEFM	X:51896821:+	8:24917089:+
CMTX3 Clone A	IDS	NEFM	X:149482782:-	8:24917235:+
CMTX3 Clone A	TMSB4X	NEFM	X:12977051:+	8:24918208:+
CMTX3 Clone A	GDI1	NEFL	X:154440895:+	8:24951256:-
CMTX3 Clone A	BEX1	NEFL	X:103063182:-	8:24951281:-
CMTX3 Clone A	ARMCX3	NEFL	X:101625765:+	8:24951494:-
CMTX3 Clone A	L1CAM	NEFL	X:153862087:-	8:24951602:-
CMTX3 Clone A	TMEM35A	NEFL	X:101096206:+	8:24952020:-
CMTX3 Clone A	DCX	NEFL	X:111297332:-	8:24952081:-
CMTX3 Clone A	GLRA2	NEFL	X:14730545:+	8:24952249:-
CMTX3 Clone A	TMSB4X	NEFL	X:12977040:+	8:24952923:-

Sample	Gene1	Gene2	Fusion point (gene1)	Fusion point (gene2)
CMTX3 Clone A	SMS	NEFL	X:21978028:+	8:24952952:-
CMTX3 Clone A	DCX	NEFL	X:111299591:-	8:24952954:-
CMTX3 Clone A	RPS4X	NEFL	X:72275062:-	8:24953717:-
CMTX3 Clone A	DCX	NEFL	X:111401104:-	8:24953771:-
CMTX3 Clone A	RTL8C	NEFL	X:135032693:+	8:24955599:-
CMTX3 Clone A	TMSB4X	NEFL	X:12977041:+	8:24955734:-
CMTX3 Clone A	ARMCX3	NEFL	X:101626083:+	8:24955742:-
CMTX3 Clone A	PGRMC1	NEFL	X:119243996:+	8:24955745:-
CMTX3 Clone A	SMS	NEFL	X:21984352:+	8:24956016:-
CMTX3 Clone A	BEX3	NEFL	X:103377078:+	8:24956116:-
CMTX3 Clone A	DCX	NEFL	X:111301359:-	8:24956585:-
CMTX3 Clone A	BEX1	NEFL	X:103063140:-	8:24956596:-
CMTX3 Clone A	GDI1	STMN4	X:154442605:+	8:27236581:-
CMTX3 Clone A	IDS	STMN4	X:149482413:-	8:27241117:-
CMTX3 Clone A	BEX3	STMN2	X:103378015:+	8:79641447:+
CMTX3 Clone A	HUWE1	STMN2	X:53629516:-	8:79664815:+
CMTX3 Clone A	RPS4X	STMN2	X:72273263:-	8:79664908:+
CMTX3 Clone A	DCX	STMN2	X:111296616:-	8:79664937:+
CMTX3 Clone A	RPS4X	STMN2	X:72273234:-	8:79664938:+
CMTX3 Clone B	NEFL	ARMCX3	8:24952412:-	X:101623991:+
CMTX3 Clone B	STMN2	ARMCX3	8:79665555:+	X:101627452:+
CMTX3 Clone B	NEFL	CLCN4	8:24952653:-	X:10208133:+
CMTX3 Clone B	NEFL	BEX1	8:24952221:-	X:103063260:-
CMTX3 Clone B	YWHAZ	DCX	8:100924930:-	X:111298960:-
CMTX3 Clone B	PABPC1	DCX	8:100702923:-	X:111299220:-
CMTX3 Clone B	STMN2	DCX	8:79665494:+	X:111410117:-
CMTX3 Clone B	MTMR9	DCX	8:11324592:+	X:111410174:-
CMTX3 Clone B	STMN2	DCX	8:79665486:+	X:111410314:-
CMTX3 Clone B	NEFM	TMSB4X	8:24918010:+	X:12976264:+
CMTX3 Clone B	STMN2	TMSB4X	8:79654976:+	X:12977024:+
CMTX3 Clone B	NEFL	TMSB4X	8:24952277:-	X:12977072:+
CMTX3 Clone B	NEFM	SLC9A6	8:24917219:+	X:136010493:+
CMTX3 Clone B	NEFL	GLRA2	8:24955529:-	X:14730932:+
CMTX3 Clone B	NEFM	IDS	8:24917105:+	X:149480599:-
CMTX3 Clone B	NEFL	HMGB3	8:24953612:-	X:150988617:+
CMTX3 Clone B	NEFL	HMGB3	8:24951803:-	X:150988667:+
CMTX3 Clone B	STMN2	L1CAM	8:79665946:+	X:153865382:-
CMTX3 Clone B	STMN2	GDI1	8:79665856:+	X:154438574:+
CMTX3 Clone B	NEFL	GDI1	8:24953752:-	X:154441094:+
CMTX3 Clone B	NEFM	GDI1	8:24914184:+	X:154442770:+
CMTX3 Clone B	STMN2	CNKSR2	8:79665701:+	X:21601282:+
CMTX3 Clone B	NEFM	DDX3X	8:24914718:+	X:41347750:+
CMTX3 Clone B	NEFL	NONO	8:24951987:-	X:71296963:+
CMTX3 Clone B	DCX	NEFM	X:111294302:-	8:24917061:+
CMTX3 Clone B	DCX	NEFM	X:111297547:-	8:24917450:+
CMTX3 Clone B	GDI1	NEFM	X:154442214:+	8:24918289:+
CMTX3 Clone B	PDZD4	NEFL	X:153803688:-	8:24955717:-
CMTX3 Clone B	DCX	NEFL	X:111296947:-	8:24955822:-
CMTX3 Clone B	RBMX	NEFL	X:136873734:-	8:24955907:-
CMTX3 Clone B	UBA1	NEFL	X:47210084:+	8:24955915:-
CMTX3 Clone B	TMSB4X	NEFL	X:12977041:+	8:24956170:-
CMTX3 Clone B	L1CAM	NEFL	X:153867998:-	8:24956334:-
CMTX3 Clone B	ATP6AP2	NEFL	X:40605770:+	8:24956590:-
CMTX3 Clone B	GPRASP2	NEFL	X:102716128:+	8:24956593:-
CMTX3 Clone B	SLC25A6	STMN2	X:1389487:-	8:79665066:+
CMTX3 Clone B	IDS	STMN2	X:149481832:-	8:79665295:+
CMTX3 Clone B	PGK1	LAPTM4B	X:78126050:+	8:97851751:+

Supplementary Table 4.5: Fusion transcripts, predicted by FusionCatcher, on chromosome X and chromosome 8 from CMTX3 patient-derived sMN. (A) Data from chromosome X. (B) Subset of data from chromosome 8. Due to the high number of predicted transcripts involving chromosome 8, only those with a breakpoint located within 3 Mb of the genomic region of interest (chr8:144,542,928 – 144,620,773) are shown.

A

Sample	Gene1	Gene2	Fusion point (gene1)	Fusion point (gene2)
CMTX3 Clone A	ARMCX3	TMSB4X	X:101626142:+	X:12976889:+
CMTX3 Clone A	WDR13	VN1R110P	X:48604390:+	X:48636418:+
CMTX3 Clone A	HDAC6	ERAS	X:48823588:+	X:48829081:+
CMTX3 Clone A	HDAC6	ERAS	X:48823785:+	X:48829081:+
CMTX3 Clone A	AL022157.1	FAAH2	X:57127058:+	X:57290251:+
CMTX3 Clone A	AL022157.1	FAAH2	X:57127058:+	X:57292498:+
CMTX3 Clone B	AC234782.2	ZCCHC18	X:103918409:-	X:104115252:+
CMTX3 Clone B	TMSB15B	AC234783.1	X:103962723:+	X:104062987:-
CMTX3 Clone B	TMSB15B	AC234783.1	X:103964622:+	X:104062987:-
CMTX3 Clone B	TMSB15B	AC234783.1	X:103966452:+	X:104062987:-
CMTX3 Clone B	FRMPD3	PRPS1	X:107598142:+	X:107639295:+
CMTX3 Clone B	TMEM164	RTL9	X:110171520:+	X:110445115:+
CMTX3 Clone B	CFP	SYN1	X:47623542:-	X:47607198:-
CMTX3 Clone B	HDAC6	ERAS	X:48823588:+	X:48829081:+
CMTX3 Clone B	HDAC6	ERAS	X:48823785:+	X:48829081:+
CMTX3 Clone B	HDAC6	ERAS	X:48824696:+	X:48829081:+

B

Sample	Gene1	Gene2	Fusion point (gene1)	Fusion point (gene2)
CMTX3 Clone A	NEFL	RPL8	8:24951691:-	8:144792137:-
CMTX3 Clone B	RPL8	NEFL	8:144789926:-	8:24956127:-
CMTX3 Clone B	RPL8	NEFL	8:144791814:-	8:24953706:-
CMTX3 Clone B	NEFL	EEF1D	8:24951803:-	8:143581128:-
CMTX3 Clone B	NEFL	ZNF623	8:24951189:-	8:143653266:+
CMTX3 Clone B	NEFL	RPL8	8:24952829:-	8:144791419:-
CMTX3 Clone B	NEFL	RPL8	8:24955547:-	8:144791463:-
CMTX3 Clone B	NEFL	RPL8	8:24955668:-	8:144792137:-

Supplementary Table 4.6: Novel transcript variants on chromosomes 8 and X across CMTX3 and control-derived sMN. (A) Data from chromosome 8. (B) Data from chromosome X. Reads represented as TPM (transcripts per kilobase million).

A

Transcript_ID	Gene_ID	Transcript_Locus	Strand	Class Code	CMTX3 Clone A TPM	CMTX3 Clone B TPM	Control 1 TPM	Control 2 TPM	Control 3 TPM
MSTRG.21773.1	MSTRG.21773	chr8:299132-300663	-	u	9.746	11.078	9.362	9.919	10.907
MSTRG.21774.1	MSTRG.21774	chr8:303680-306934	-	u	15.108	16.079	7.423	0.000	4.392
MSTRG.21774.5	MSTRG.21774	chr8:303777-306934	-	u	0.000	0.000	9.434	0.000	8.633
MSTRG.21774.4	MSTRG.21774	chr8:303777-306934	-	u	0.000	0.000	0.000	0.000	8.921
MSTRG.22168.1	MSTRG.22168	chr8:53713271-53714469	+	u	63.217	45.597	45.953	54.156	48.732
MSTRG.22197.1	MSTRG.22197	chr8:59969965-59970317	+	u	45.944	29.969	24.109	17.160	19.005
MSTRG.22224.1	MSTRG.22224	chr8:63585704-63589380	+	u	0.255	0.283	0.363	0.084	0.108
MSTRG.22298.1	MSTRG.22298	chr8:74346154-74350072	-	u	5.108	4.606	4.858	3.207	5.522
MSTRG.22497.4	MSTRG.22497	chr8:102995623-103002487	+	u	0.136	0.011	0.550	0.000	0.179
MSTRG.22519.1	MSTRG.22519	chr8:104289776-104290085	-	u	5.483	5.650	10.074	4.685	0.000
MSTRG.22540.1	MSTRG.22540	chr8:114029500-114030028	-	u	98.711	83.626	100.373	93.892	99.637
MSTRG.22656.3	MSTRG.22656	chr8:140504374-140508744	-	u	0.000	0.283	0.000	0.000	0.000
MSTRG.22686.6	MSTRG.22686	chr8:143304620-143308344	-	u	0.000	0.000	0.710	0.000	1.162
MSTRG.22713.1	MSTRG.22713	chr8:144081028-144082529	-	u	0.675	0.625	0.235	0.417	0.155
MSTRG.22713.2	MSTRG.22713	chr8:144081558-144082529	-	u	0.184	0.275	0.164	0.051	0.253

B

Transcript_ID	Gene_ID	Transcript_Locus	Strand	Class Code	CMTX3 Clone A TPM	CMTX3 Clone B TPM	Control 1 TPM	Control 2 TPM	Control 3 TPM
MSTRG.23823.1	MSTRG.23823	chrX:326995-328241	-	u	2.564	2.825	2.896	2.679	3.370
MSTRG.23854.12	MSTRG.23854	chrX:3811721-3816452	-	u	0.000	0.000	0.353	0.124	0.331
MSTRG.23975.1	MSTRG.23975	chrX:26206921-26207295	-	u	34.964	26.195	29.630	29.816	27.321
MSTRG.23998.1	MSTRG.23998	chrX:40014976-40015681	+	u	0.424	0.235	1.042	0.604	0.911
MSTRG.24015.1	MSTRG.24015	chrX:42473440-42475077	-	u	0.006	0.016	0.021	0.004	0.011
MSTRG.24139.1	MSTRG.24139	chrX:56618232-56627186	+	u	10.389	11.323	6.491	15.812	9.901
MSTRG.24296.1	MSTRG.24296	chrX:98573848-98982589	+	u	1.108	1.654	2.728	0.000	0.545
MSTRG.24296.2	MSTRG.24296	chrX:98573848-98982589	+	u	0.312	0.000	0.193	0.084	0.492
MSTRG.24296.3	MSTRG.24296	chrX:98573854-98982589	+	u	4.422	4.095	7.173	0.895	0.000
MSTRG.24307.2	MSTRG.24307	chrX:101341174-101344622	-	u	0.000	0.000	0.000	0.333	0.000
MSTRG.24307.5	MSTRG.24307	chrX:101341703-101344622	-	u	2.587	0.917	1.314	0.844	1.356
MSTRG.24570.1	MSTRG.24570	chrX:153075614-153077718	-	u	33.800	32.015	19.785	25.290	22.637
MSTRG.24619.1	MSTRG.24619	chrX:154639391-154640885	-	u	11.360	10.578	10.966	10.356	9.770

Supplementary Table 4.7: Novel splice variants at genomic regions of interest on chromosomes X and 8 across CMTX3 and control-derived sMN. (A) Predicted splice variants located within 3 Mb either side of the breakpoint of the CMTX3 SV on Xq27.1. (B) Predicted splice variants falling within a genomic window extending 100 kb either side of the region of chromosome 8q24.3 that is duplicated in CMTX3 (chr8:144,542,928 – 144,620,773). Reads represented as TPM (transcripts per kilobase million).

A

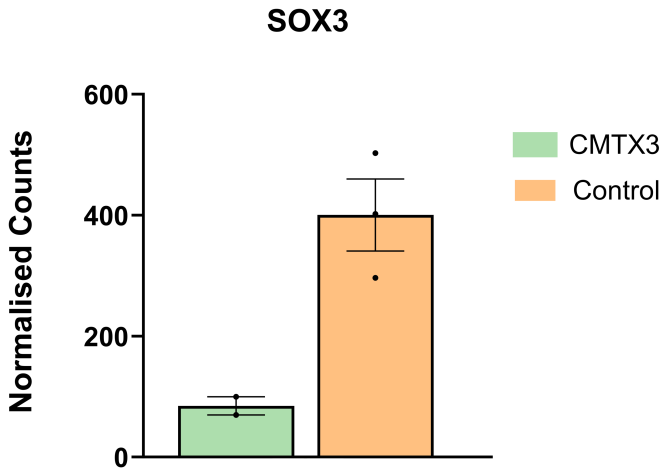
Nearest refTranscript Name	stringtie Transcript Name	Gene	Transcript Locus	Strand	Class Code	CMTX3 Clone A TPM	CMTX3 Clone B TPM	Control 1 TPM	Control 2 TPM	Control 3 TPM
XM_017029441	MSTRG.24523.2	ATP11C	chrX:139726160-139933100	-	j	8.444	6.842	7.968	4.633	5.613
XM_017029441	MSTRG.24523.1	ATP11C	chrX:139726160-139933100	-	j	7.879	5.902	7.788	7.969	6.986
XM_017029441	MSTRG.24523.3	ATP11C	chrX:139726346-139933000	-	j	2.487	2.650	3.363	3.056	4.820
gene-HNRNPA3P3	MSTRG.24526.1	.	chrX:140033518-140033857	+	p	37.601	27.032	23.961	27.230	30.862
NR_104228	MSTRG.24529.1	LINC00632	chrX:140709747-140788723	+	j	7.049	6.169	5.055	7.316	9.096
NR_104228	MSTRG.24529.2	LINC00632	chrX:140709755-140788723	+	j	0.471	0.000	0.870	0.908	1.909
NR_104228	MSTRG.24529.6	LINC00632	chrX:140709763-140788723	+	j	0.338	1.321	2.746	0.560	0.421
NM_012317	MSTRG.24531.1	LDOC1	chrX:141141598-141177189	-	k	46.727	24.615	50.995	45.481	46.258

B

Nearest refTranscript Name	stringtieTranscript Name	Gene	Transcript Locus	Strand	Class Code	CMTX3 Clone A TPM	CMTX3 Clone B TPM	Control 1 TPM	Control 2 TPM	Control 3 TPM
NR_162109	MSTRG.22731.1	MIR10400	chr8:144447347-144465570	-	k	14.371	14.679	14.465	16.219	15.381
NM_138496	MSTRG.22731.2	CYHR1	chr8:144447371-144465712	-	j	1.404	2.427	1.933	2.335	2.915
NM_001330618	rna-MIR10400	CYHR1	chr8:144462485-144462518	-	c	0.000	0.000	0.000	0.000	0.000
NR_162109	NR_162109	MIR10400	chr8:144462485-144462972	-	k	0.000	0.000	0.001	0.001	0.000
rna-MIR10400-2	rna-MIR10400-2	MIR10400	chr8:144462519-144462972	-	k	0.001	0.001	0.001	0.001	0.000
NM_138496	MSTRG.22731.8	CYHR1	chr8:144462973-144465712	-	j	1.808	1.606	1.299	1.947	0.998
NM_001369769	MSTRG.22732.3	KIFC2	chr8:144466032-144474435	+	j	0.000	13.515	0.000	0.979	0.000
NM_001369769	MSTRG.22732.1	KIFC2	chr8:144466032-144474435	+	j	66.869	37.321	43.750	63.577	74.361
NM_001369769	MSTRG.22732.4	KIFC2	chr8:144466032-144474435	+	j	2.707	6.420	5.706	1.524	9.364
NM_001369769	MSTRG.22732.2	KIFC2	chr8:144466032-144474435	+	j	41.338	47.979	52.041	48.113	46.802
NM_001369769	MSTRG.22732.5	KIFC2	chr8:144466041-144474435	+	j	2.465	0.000	0.000	2.258	3.062
NM_001369769	MSTRG.22732.6	KIFC2	chr8:144466046-144474435	+	j	0.000	13.315	0.000	2.329	0.000
NM_001382664	MSTRG.22737.1	.	chr8:144507128-144508198	-	x	3.679	4.333	4.048	5.203	5.810
XM_011516806	MSTRG.22738.5	MFSD3	chr8:144509070-144511228	+	j	7.041	5.265	5.230	8.339	5.711
NM_004260	MSTRG.22739.1	RECQL4	chr8:144511288-144529111	-	j	8.981	3.374	11.578	12.106	11.718
NM_004260	MSTRG.22739.2	RECQL4	chr8:144511288-144529111	-	j	3.822	1.690	6.388	2.980	3.358
NM_004260	MSTRG.22739.21	RECQL4	chr8:144511304-144529111	-	j	6.224	2.698	2.205	4.769	3.536
NM_004260	MSTRG.22739.22	RECQL4	chr8:144511446-144529111	-	j	0.000	9.206	0.000	0.000	0.000
NM_001272036	MSTRG.22740.5	LRRC14	chr8:144517977-144525172	+	j	3.748	3.641	2.862	7.785	3.143
XR_001746144	MSTRG.22742.3	LOC100996662	chr8:144700617-144713423	+	j	0.000	0.040	1.666	1.254	0.000
XR_001746144	MSTRG.22742.4	LOC100996662	chr8:144700635-144713423	+	j	0.000	0.000	3.356	0.000	0.000
XR_001746144	MSTRG.22742.5	LOC100996662	chr8:144700650-144713423	+	j	0.000	0.000	0.711	1.581	0.000

Supplementary Table 4.8: List of predicted DEG in CMTX3 sMN. DEG were identified using edgeR with FDR cutoff < 0.05. Each CMTX3 clone was individually compared to a cohort of 3 controls. Only those DEG identified in both clones, showing the same direction of logFC, are listed. Genes are listed alphabetically. logFC = log2(fold change); FDR = false discovery rate.

	CMTX3 Clone A		CMTX3 Clone B		
Gene	logFC	FDR	logFC	FDR	Chr
AGT	1.933	3.25E-06	2.280	3.25667E-10	1
BMS1P14	3.324	5.09E-05	3.660	8.51378E-07	9
CACNA1A	-1.028	0.021104445	-1.228	0.001603346	19
COL8A1	-1.005	0.011387034	0.820	0.008979526	3
CTBP2P11	-3.058	0.0022776	-2.883	0.003469296	21
DDX12P	-1.709	0.041438048	-1.967	0.008236152	12
DNAJC25-GNG10	4.753	0.000327944	4.861	8.60989E-05	9
EIF2S3B	3.990	3.41E-06	3.941	1.57584E-06	12
EZR	1.481	0.027973536	1.672	0.00328736	6
FAR2P1	-1.911	9.60E-06	-2.559	4.47331E-10	2
FOXG1	6.993	0.003490571	7.001	0.002405563	14
GRAMD1B	0.773	0.005211177	0.705	0.019543207	11
GYG2P1	1.910	0.034626143	2.141	0.005039234	Y
HCG24	-4.279	0.000373939	-3.725	0.001305872	6
HDAC9	0.638	0.03432631	0.946	9.59408E-06	7
HLA-DRA	4.911	0.000203273	4.089	0.009096045	6
HLA-DRB1	8.693	1.50E-07	8.969	2.53159E-09	6
HTATSF1P2	-8.133	0.000471579	-8.161	0.000214111	6
IDI2-AS1	-8.638	0.000373939	-8.665	0.000174928	10
KYAT3	-0.989	0.020326933	-1.040	0.007918176	1
LOC100130701	8.806	4.93E-08	8.420	6.39659E-07	3
LOC101927610	2.123	0.015591267	2.442	0.000761506	7
LOC101929599	5.080	0.001224275	4.578	0.007761304	5
LOC102724219	-1.748	0.019614775	-1.659	0.026211798	21
LOC102724528	-2.556	0.021329703	-2.486	0.021771625	1
LOC102724701	-1.163	0.001641453	-1.105	0.002278872	21
LOC102724843	-1.410	0.000194649	-1.329	0.000272401	21
LOC105377762	7.606	0.000235823	7.640	9.34497E-05	5
LOC105379005	-3.687	0.02816672	-7.873	0.000973928	5
LOC107984489	-7.922	0.004739163	-7.950	0.002818204	12
LOC441081	6.382	1.60E-09	5.878	1.1372E-07	5
LOC642846	-3.210	1.97E-06	-3.571	6.26642E-08	12
MED15P9	-2.471	0.027974033	-9.753	1.41531E-08	2
MICB	2.862	0.028068753	3.072	0.007918176	6
OPTN	-0.741	0.000471579	-0.679	0.001572468	10
PLCXD3	0.724	0.000929537	0.683	0.002818204	5
POLR2J2	1.091	1.45E-05	1.010	7.47275E-05	7
POTEF	-3.161	0.011269764	-5.337	4.45397E-06	2
PRDX6-AS1	5.032	0.019609235	5.507	0.001820063	1
RBFOX1	-2.436	3.50E-21	-2.470	4.87659E-22	16
RIMBP2	-0.725	0.001193665	-0.782	0.000164651	12
RPL10P16	2.179	1.42E-06	1.973	2.12077E-05	19
RPL10P9	8.133	3.57E-139	7.858	3.8367E-116	5
SAXO1	-7.107	0.037371354	-7.134	0.029706207	9
SDHDP6	-3.370	0.013232303	-8.515	3.49661E-05	1
SLITRK2	-0.679	0.031596354	-0.858	0.001025725	X
ST8SIA6	1.371	0.027974033	1.884	2.2354E-05	10
TACR1	-2.323	0.001253761	-2.863	1.55689E-05	2
TMEM132C	-2.457	5.44E-29	-1.731	1.85959E-15	12
TMEM132D	-2.064	3.12E-15	-1.997	1.9235E-14	12
VSX2	-1.540	0.030309056	-2.171	0.000180492	14
ZNF736P9Y	8.505	8.49E-07	8.420	6.39659E-07	Y



Supplementary Figure 4.1: *SOX3* normalised counts from a preliminary NanoString nCounter assay on caudal NEP. NEP RNA was isolated from one CMTX3 patient (with 2 clonal cell lines) and 3 controls. Statistical analysis could not be performed due to lack of biological replicates within the CMTX3 group, although a trend of reduced SOX3 expression in the patient sample is evident. Error bars represent mean ± SEM.

Supplementary Table 4.9: Quality control metrics for NEP RNA-Seq libraries. GC = guanine-cytosine content;

	Raw Statistics				Pre-processing & mapping statistics		
Sample ID	**Total reads**	**GC (%)**	**Q20 (%)**	**Q30 (%)**	**Trimmed reads**	**Mapped reads (%)**	**Unmapped reads (%)**
CMTX3 P1	111,278,298	50.27	97.61	93.63	109,170,560	98.46%	1.54%
CMTX3 P2	122,126,074	49.44	97.36	93.15	119,465,858	98.17%	1.83%
Control 1	115,012,680	49.36	97.36	93.11	112,343,280	97.64%	2.36%
Control 2	103,088,466	48.93	97.46	93.38	100,805,200	97.99%	2.01%
Control 3	111,440,304	49.65	97.26	92.89	108,969,670	98.04%	1.96%

CHAPTER 5

Exploring Chromatin Interactions at the CMTX3 Locus

5.1 Introduction

In chapter 4, we showed that CMTX3 patient-derived cells show spatiotemporally specific dysregulation of *SOX3*, which encodes a developmental transcription factor. *SOX3* downregulation was observed in patient-derived iPSC as well as caudal NEPs, but not in MNP, neurospheres, or sMN. This is the first report of a disease-relevant cell type showing local gene dysregulation when comparing CMTX3 patient and control tissue and provides compelling evidence for *SOX3* to be a high priority candidate gene for this neuropathy. However, the precise mechanism of how the CMTX3 insertion leads to altered spatiotemporal regulation of *SOX3* remains unknown. Gaining an insight into the underlying molecular mechanisms will progress our understanding of the pathogenesis of CMTX3 and inform on disorders caused by similar genomic rearrangements, such as the pathogenic insertions (+/- deletions) occurring at the Xq27.1 hotspot (reviewed in Chapter 1, Part B).

It has been hypothesised that the pathogenic Xq27.1 insertions may cause transcriptional dysregulation by altering interactions between genes and distal cis-regulatory elements (CREs) [63]. One possible hypothesis is the inserted DNA sequences may introduce ectopic CREs into the local genomic landscape, which could alter the regulation of nearby genes [63, 151, 198-201]. Indeed, examination of the CMTX3 insertion sequence on UCSC Genome Browser (http://genome.ucsc.edu)

[202] suggests that multiple predicted regulatory elements are contained within this region. Figure 5.1 displays the regulatory elements predicted by various databases, including ENCODE project [203], ORegAnno [204], and GeneHancer [205]. Alternatively, the pathogenic SVs may disrupt existing regulatory interactions, separating genes from their typical distal CREs [152]. Experimental investigations into the effect of these numerous pathogenic insertions at Xq27.1 upon local regulatory interactions are clearly lacking.

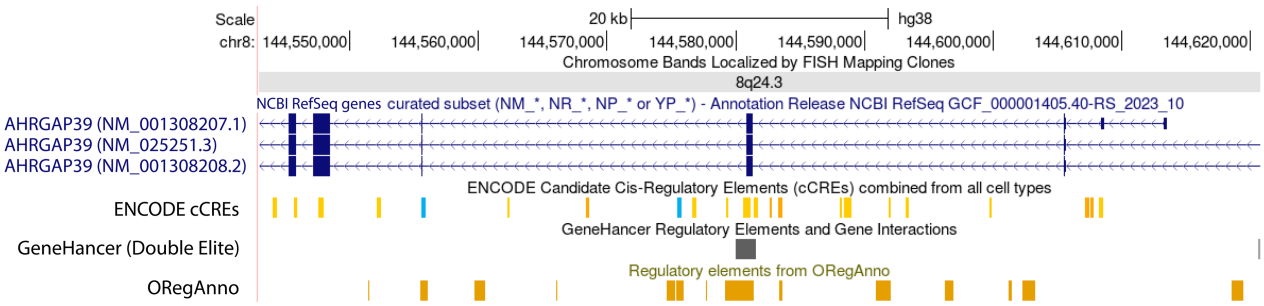


Figure 5.1: Predicted regulatory elements contained within the CMTX3 insertion sequence. Schematic of the CMTX3 insertion sequence (~ 78 kb of chromosome 8q24.3), visualised using the UCSC Genome Browser. The 3 partial transcript variants of *ARHGAP39* contained within the insertion sequence are displayed using the UCSC RefSeq Genes track. Candidate cis-regulatory elements (cCREs) predicted by the ENCODE Project Consortium [203] are shown below. Yellow and orange blocks denote cCREs which display enhancer-like signatures, while blue blocks represent cCREs displaying CTCF-binding. Immediately underneath is the GeneHancer [205] track. The dark grey block denotes an enhancer element with a high confidence score. The final track shows regulatory elements identified by ORegAnno [204], where orange blocks represent transcription factor binding sites.

As discussed in Chapter 1, CREs act to control the transcription of target genes to ensure correct expression levels within the appropriate spatiotemporal context. Enhancers are the most well-studied of the distal CREs and are thought to activate gene expression through physically interacting with the target gene's promoter [206]. However, enhancers and their target genes are often separated by large genomic distances, in some cases exceeding 1 Mb [207, 208]. Communication between these distant genomic elements is largely facilitated by the 3D organisation of chromatin within the nucleus, which includes the formation of 'chromatin loops' that bring an enhancer in to the proximity of a target

promoter [209]. Although there is limited research on silencer elements compared to enhancers, long-range chromatin interactions connecting silencer elements with the promoters of distal target genes have also been experimentally observed [210, 211]. It is well-documented that SV mutations can cause disease by re-wiring regulatory chromatin interactions and altering transcriptional output [102, 149, 212-215]. For example, SVs may cause ectopic interactions between genes and regulatory elements that drive gene misexpression [102, 212]. SVs may also disconnect genes from their cognate enhancers, resulting in reduced expression of the target gene [107].

The main experimental tools for investigating 3D chromatin architecture include imaging-based approaches and chromosome conformation capture (3C) technologies [216]. 3C-derived methods are particularly well-suited to interrogating long-range regulatory interactions as they can offer a high-throughput, high-resolution, and genome-wide insight into chromatin organisation [217]. 3C technology, developed by Dekker and colleagues in 2002, is used to assess the frequency of physical interactions between genomic loci [218]. Since its inception, this technology has been modified and expanded upon to serve as the basis for a wide range of derivative methods, such as 4C, 5C, Hi-C, Capture Hi-C and Micro-C, which are reviewed below.

5.1.1 Evolution of 3C-based technologies

The basic workflow of 3C is outlined in Figure 5.2A. This involves (a) formaldehyde fixation of DNA within the nucleus, which preserves 3D chromatin architecture, (b) restriction-enzyme digestion of the DNA which generates crosslinked DNA fragments, (c) 'proximity ligation' of DNA fragments, which preferentially ligates fragments located in close physical proximity, and (d) detection and quantification of these chimeric DNA fragments [218]. The abundance of each chimeric fragment serves as a proxy for the frequency of interaction between the two genomic loci [218].

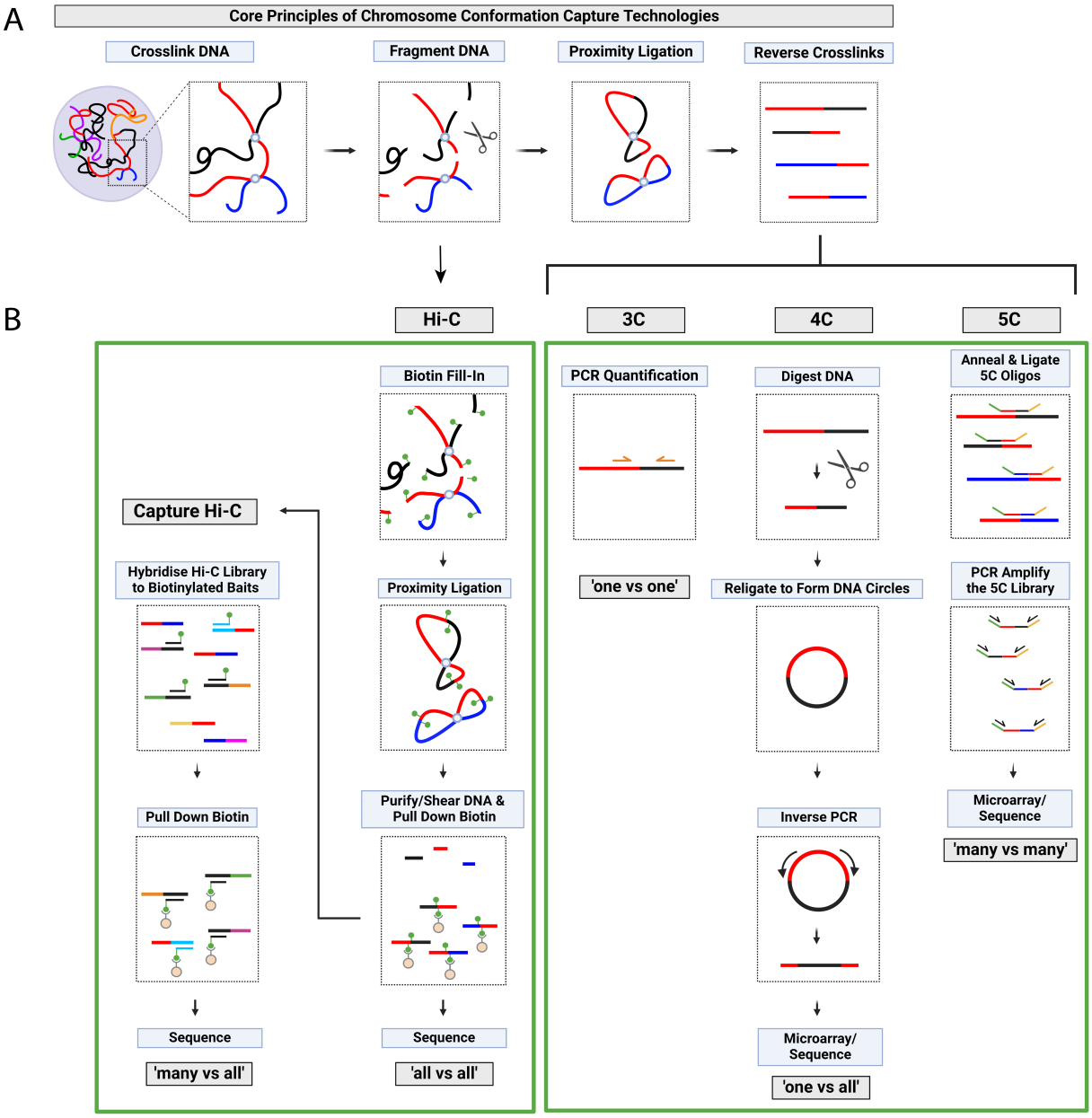


Figure 5.2: Overview of the 3C-based technologies. (A) 3C-derived technologies share the same core workflow. DNA is crosslinked inside the nucleus and digested into fragments. Ligation is performed using experimental conditions favouring ligation of fragments located in close physical proximity. Crosslinks are reversed and DNA is purified. Ligated fragments can be detected using a range of techniques. The more frequently a specific ligation event is observed, the more likely it is that those two genomic loci interact within the nucleus. (B) In 3C, individual chimeric fragments are detected using PCR. In 4C, DNA circles are generated, and inverse PCR is performed with primers against a target sequence. Products are analysed with microarray/sequencing. 5C can detect multiple ligation products simultaneously. Primers (with common tail sequences) are designed to target specific ligation products. Primers that anneal across a ligation junction are ligated together, creating 'copies' of the ligation

junctions, which are then amplified by PCR and analysed. In Hi-C, DNA overhangs are filled in with biotin-labelled nucleotides (green circle) prior to proximity ligation. This allows selective pull-down of ligation junctions with streptavidin-beads (white circles). Hi-C library is analysed with next-generation sequencing. Capture Hi-C enriches the Hi-C libraries for target regions using a hybridization reaction. Figure made with BioRender, using information adapted from de Wit & de Laat (2012) [219] and Denker & de Laat (2016) [220].

In the classic 3C workflow, each specific chimeric fragment is amplified by its own set of PCR primers, designed to target the two, pre-selected genomic loci [218] (Figure 5.2B). PCR products are run on an agarose gel and semi-quantitative PCR is used to calculate the cross-linking frequency between the two sites [218]. Thus, 3C has been termed a “one-vs-one” approach and can only be used to interrogate known sequences at a limited throughput [221].

The subsequent development of 4C (chromosome conformation capture-on-chip) technology enabled a higher-throughput, “one-vs-all” approach [222] (Figure 5.2B). 4C follows the standard 3C workflow however, after proximity ligation, chimeric fragments undergo an additional DNA digest and are re-ligated to generate DNA circles [222]. Primers targeting a specific sequence of interest are used to amplify the DNA circles by inverse PCR [222]. PCR products are analysed with microarray [222] or high-throughput sequencing [223], to identify all genomic loci interacting with the target sequence. In the same year, the term 4C was also used to describe a similar experimental technique, termed ‘circular chromosome conformation capture’, which also relies on the formation of circular DNA and PCR amplification of unknown sequences interacting with a known bait [224].

5C, or ‘chromosome conformation capture carbon copy’, enables simultaneous detection of multiple ligation fragments [225] (Figure 5.2B). 5C is based upon multiplex ligation-mediated amplification [225]. Briefly, primers are designed to target multiple possible ligation junctions, and can be combined into a single reaction. Once a set of primers anneals across a ligation junction, the primers are ligated together, creating a ‘copy’ of the original chimeric fragment. However, the tails of all the primers contain common sequence, permitting PCR amplification of the entire copied library, which is subsequently analysed using microarray or high-throughput sequencing [225].

The above techniques revolutionised understanding of the spatial organisation of chromatin in the nucleus, however, they all rely on a pre-existing knowledge of either one (4C) or both (3C and 5C) halves of the ligation products. The development of Hi-C technology offered an unbiased and genome-wide assessment of chromatin organisation by combining the principles of 3C with massively parallel DNA sequencing [226] (Figure 5.2B). In this method, cells are fixed, and the DNA is digested as described above. However, unlike traditional 3C, the 5' overhangs are filled in with nucleotides, including one that is biotin-labelled. The blunt ends are joined by proximity ligation and the DNA is purified. Biotin is removed from the ends of un-ligated fragments using an exonuclease, and the DNA fragments are sheared by sonication. The ligation junctions, marked with biotin, can then be selectively pulled down using streptavidin beads. The resulting library of chimeric fragments is then sequenced using high-throughput sequencing, providing a genome-wide, "all-vs-all", insight into chromatin interactions [226]. However, this powerful technology brings its own unique issues for consideration. Deep sequencing is required to detect 'true' chromatin contacts from within the highly complex, noisy Hi-C libraries, and the data resolution is critically dependent on the sequencing depth [227]. For example, one article recommends using approximately 100 million valid junction reads to achieve a data resolution of 40 kb [228]. A resolution as high as 1 kb has been achieved, however, required analysis of 4.9 billion valid pairwise contacts [229]. Ultimately, due to the prohibitive cost of sequencing and high computational burden of data analysis, obtaining a data resolution sufficient for mapping fine-scale chromatin interactions is beyond the resources of smaller research laboratories.

Sequencing burden can be reduced by including a hybridisation capture step, to enrich Hi-C libraries for targeted regions of interest prior to high-throughput sequencing [230]. This approach, termed Capture Hi-C, provides high-resolution data for targeted genomic loci [230] (Figure 5.2B). Promoter Capture Hi-C is a variation of this technique which utilises baits designed to target known promoter sequences, enabling the identification of specific DNA sequences that interact with gene promoters [231, 232]. This technology has provided valuable insights into the principles of long-range gene regulation, particularly when coupled with additional functional data, such as information on a

gene's level of transcriptional activity (e.g. RNA-Seq), and the histone modification signatures of promoters and their interacting partners [232, 233].

Another factor restricting the resolution of conventional Hi-C relates to the use of restriction enzymes to fragment the DNA prior to performing proximity ligation. Restriction enzyme sites are unequally distributed across the genome, resulting in fragments of varying sizes which limits data resolution [234, 235]. To address this shortcoming, Hsieh et al., (2015) developed the Micro-C technique which replaces restriction enzymes with micrococcal nuclease (MNase), to fragment the genome into mononucleosomes [235, 236]. This produces many more DNA fragments which also show more regular spacing across the genome, resulting in higher resolution data [236]. Accordingly, Micro-C has been shown to detect ~3-5 times more looping interactions than conventional Hi-C [237]. Thus, Micro-C technology facilitates deeper investigation into the fine-scale chromatin interactions that play an important role in gene regulation yet are difficult to interrogate using conventional Hi-C methods [238].

In this chapter, we performed a preliminary investigation to explore the effect of the CMTX3 SV upon local chromatin interactions. A Promoter Capture Micro-C approach was adopted to specifically focus on interactions anchored at gene promoters. The experimental service to construct the Micro-C Pan Promoter library for analysis was outsourced to Cantata Bio (Scotts Valley, CA, USA). The tissue used for the library was caudal NEP samples, as this cell-type showed local gene dysregulation in CMTX3 patients when compared to controls.

5.2 Methods

5.2.1 Cell culture and sample preparation

Patient (n=1) and control (n=1) iPSC were cultured in 6-well plates, as described (Chapter 2, Section 2.3.2). Once optimal iPSC morphology was achieved, neuronal differentiation was commenced (Chapter 2, Section 2.3.3). Cells were maintained in differentiation media for 6 days to induce the

formation of caudal NEP. Cells were gently washed with DPBS(-Ca -Mg) (Gibco) and dissociated into a single cell suspension using Accumax™ cell aggregate dissociation medium (Thermo Fisher Scientific). Briefly, NEP underwent a 5 min incubation in a solution of Accumax and DPBS(-Ca -Mg) (1:1 v/v) at 37°C. The cell suspension was collected in a sterile 15 mL falcon tube and residual cells were washed from the plate using DPBS(-Ca -Mg). Cells were pelleted by centrifugation at 200 x g for 4 min. After removal of the supernatant, cells were resuspended in 1 mL DPBS(-Ca -Mg). The solution was gently pipetted ~5 times, using a P1000 pipette, to assist in cellular dissociation. To calculate the cell number and viability, 10 µL of the cell solution was combined with 10 µL of trypan blue and then pipetted into a Countess Cell Counting Chamber Slide (Thermo Fisher Scientific). Cell counting was performed using a Countess II Automated Cell Counter (Thermo Fisher Scientific). Cells were partitioned into sterile 1.5 mL microcentrifuge tubes ($1x10^6$ cells/tube). Cell suspensions were centrifuged at 3,000 x g for 5 min, and the supernatant was gently removed. Cell pellets were snap frozen in liquid nitrogen and stored at -80°C.

5.2.2 Proximity ligation library preparation and target enrichment

Preparation of the proximity ligation libraries, as well as target enrichment, was outsourced to Cantata Bio (Scotts Valley, CA, USA). This involved the generation of single-indexed, Micro-C libraries using the Dovetail Micro-C Kit (Cantata Bio). A comprehensive workflow is outlined in the Dovetail Micro-C Kit User Guide [239]. To summarise, frozen cell pellets were thawed at room temperature and fixed using a dual crosslinking method, which utilises both formaldehyde and disuccinimidyl glutarate. *In situ* nuclease digestion was performed by incubating the cells in MNase for 15 min under agitation. Cells were then lysed using SDS. QC assessment was performed to quantify the lysate and confirm successful MNase digestion. Chromatin was bound to Chromatin Capture Beads and an end-polishing reaction was performed. Biotinylated bridge sequences were ligated to the ends of the DNA fragments, and proximity ligation was performed by incubating samples in Intra-Aggregate Ligation Enzyme Mix solution for 1 h at 22 ̊C with constant agitation. Samples were then incubated in Crosslink Reversal Buffer and Proteinase K and the DNA was purified using SPRIselect beads (Beckman Coulter). To

prepare libraries for Illumina sequencing, fragments underwent end-repair, adaptor sequences were ligated, and the resulting DNA was purified. Streptavidin beads were used to capture DNA fragments containing biotinylated ligation junctions, which then underwent PCR amplification and purification. Purified NGS libraries underwent hybridisation capture using the Dovetail Human Pan Promoter Enrichment Kit (Cantata Bio) to enrich the libraries for DNA fragments containing known human promoter sequences. More than 98% of human promoter regions are targeted in this panel, which captures over 84,000 promoters across more than 27,000 genes (> 19,000 coding, > 7,000 non-coding). The workflow is comprehensively described in the Dovetail Target Enrichment User Guide [240]. To briefly summarise, the two, single-indexed libraries were first pooled and dried. For the hybridisation reaction, Probe solution (i.e. Pan Promoter Capture probes and Hybridisation Mix) was combined with the libraries which had been resuspended in a mixture of Blocker Solution and Universal Blockers. Following hybridisation, the solution was mixed with Streptavidin Binding Beads, which enables magnetic selection of the hybridised targets. Captured DNA fragments then underwent PCR amplification and DNA purification. Enriched Micro-C libraries were assessed using Cantata Bio's in-house QC pipeline. This involves shallow sequencing of the libraries to assess various metrics, including library complexity (i.e. the number of unique read pairs).

5.2.3 Next generation sequencing of enriched Micro-C libraries

Next generation sequencing of the Promoter Capture Micro-C libraries was outsourced to Macrogen Oceania (Seoul, South Korea). Library quality was first assessed using a 2100 Bioanalyzer system (Agilent). The pooled libraries underwent paired-end (150 bp) sequencing using one lane of a HiSeq X Ten system (Illumina) to generate a minimum of 150 M read pairs per sample. A PhiX spike-in of 1% was used in accordance with recommendations from Cantata Bio. Raw FASTQ files were received by the laboratory.

5.2.4 Computing resources

Analysis of Micro-C data is a computationally intensive process which requires access to high performance computing infrastructure. The analyses described in this thesis were performed using both the Nimbus cloud service, provided by Pawsey Supercomputing Centre, as well as the University of Sydney's High Performance Computing cluster, Artemis. We are grateful to Dr Georgina Samaha (Australian BioCommons and Sydney Informatics Hub, University of Sydney) and Dr Hamid Alinejad-Rokny (University of New South Wales) for providing technical assistance and advice for this analysis.

5.2.5 Generating a CMTX3 patient-specific reference genome

The purpose of running a Micro-C experiment is to explore the effect of the CMTX3 SV upon local chromatin interactions. Standard analysis protocols, which align sequencing reads to a control human reference genome (e.g. hg38), are therefore not suitable to address this biological question, since the reference genome does not harbour the patient-specific SV. As a result, any reads arising from the CMTX3 insertion fragment will be aligned to the original position on chromosome 8q24.3, rather than the actual location on chromosome Xq27.1. Therefore, it is important to carefully consider the experimental design prior to analysing the Micro-C data. Below outlines the bioinformatic approach used in this study, which involved generating a custom-built genome harbouring the CMTX3 SV which was then used for alignment of patient Micro-C reads.

5.2.5.1 Creating a custom chromosome X file harbouring the unique CMTX3 SV

The first step was to accurately characterise the CMTX3 SV sequence using the information provided in Brewer et al., (2016) [63]. As seen in Figure 5.3, the proximal breakpoint is well-defined and simple, while the distal breakpoint contains additional sequence complexity, including a short sequence showing homology to chromosome 12q13.12, and a 12 bp inversion of chromosome Xq27.1 [63]. The adjacent region on chromosome X also harbours a single nucleotide variant as well as a single nucleotide deletion [63] (Figure 5.3). The mutation coordinates, initially reported using GRCh37/hg19, were converted to the updated GRCh38/hg38 reference using the UCSC LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver) (Supplementary Table 5.1). The next step was to

generate a custom FASTA file containing the ~78 kb CMTX3 insertion sequence as well as the adjacent, modified region of chromosome Xq27.1. This process was performed using a custom script (Appendix C). To summarise, this involved downloading the 77,846 bp fragment of chromosome 8q24.3 that constitutes the CMTX3 insertion, as well as the adjacent 9 bp of sequence originating from chromosome 12q13.12. The first 12 bp of chromosome Xq27.1 located at the distal breakpoint of the CMTX3 insertion were downloaded and inverted. The adjacent 16 bp of chromosome Xq27.1 was also downloaded and manually edited to incorporate the single nucleotide variant and single nucleotide deletion observed in CMTX3 patients. All DNA sequences were downloaded from the Ensembl database [241]. Specifically, Ensembl Release 97 was used to match the reference genome (GRCh38.p12) used in Dovetail's Promoter Panel data analysis guidelines (https://dovetail-capture.readthedocs.io/en/latest/index.html).

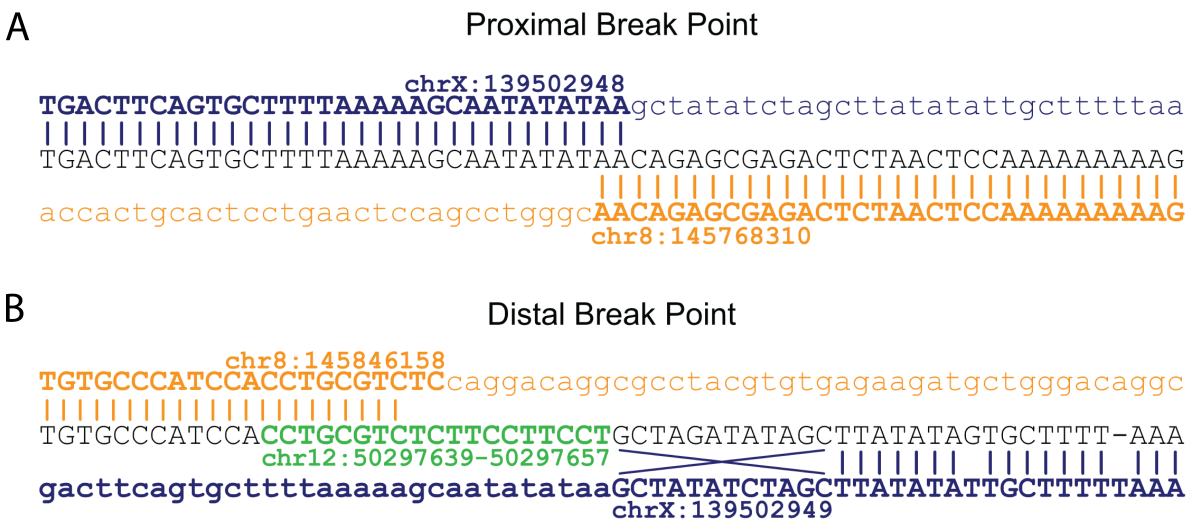


Figure 5.3: Overview of the CMTX3 genomic rearrangement. (A) The proximal breakpoint of the CMTX3 SV. Sequence from the CMTX3 patient is represented in black. Blue and orange text represents the reference sequences for chromosome X and chromosome 8, respectively. (B) The distal breakpoint of the CMTX3 SV. The chromosome 8 insertion (yellow) is followed by a small insertion from chromosome 12 (green), a 12 bp inversion of chromosome X, a single nucleotide variant, and a single nucleotide deletion. Images sourced from Brewer et al., (2016) [63]. Genomic coordinates are from hg19.

The individual text files were then concatenated to create a continuous sequence and converted into FASTA format. The FASTA file was visually inspected using SnapGene viewer (version 6.2.1) (GSL Biotech LLLC) to confirm that the correct breakpoint sequences were created. The bioinformatics tool, Reform [242], was then used to insert the custom DNA sequence into the appropriate location within the hg38 reference chromosome X FASTA file. Using this tool, the 28 bp of chromosome Xq27.1 at the distal breakpoint was also deleted, as this region was included within our custom sequence.

5.2.5.2 Removing the duplicated sequence from chromosome 8q24.3

The custom-built, patient-specific chromosome X file provides an appropriate reference sequence for mapping Micro-C reads derived from CMTX3 samples. However, the unbalanced nature of the CMTX3 insertion, which causes trisomy of 8q24.3, introduces the possibility for additional alignment errors to occur. Theoretically, any read originating from within the duplicated fragment of chromosome 8q24.3 could either be aligned to (a) the CMTX3 SV, or (b) the original sequence located at 8q24.3 (Figure 5.4A). Splitting reads over these two locations could hinder the capacity to detect chromatin interactions involving the CMTX3 SV, as the number of each read pair is critical for determining the strength of chromatin contacts. To minimise this risk, the custom patient genome was further modified by removing the 77,846 bp insertion sequence from its original location at 8q24.3 (Figure 5.4B). Therefore, any reads originating from this sequence will only map to the CMTX3 SV at Xq27.1. Whilst this will result in the inaccurate mapping of reads originating from the ‘true’ 8q24.3, these can be readily identified post-hoc.

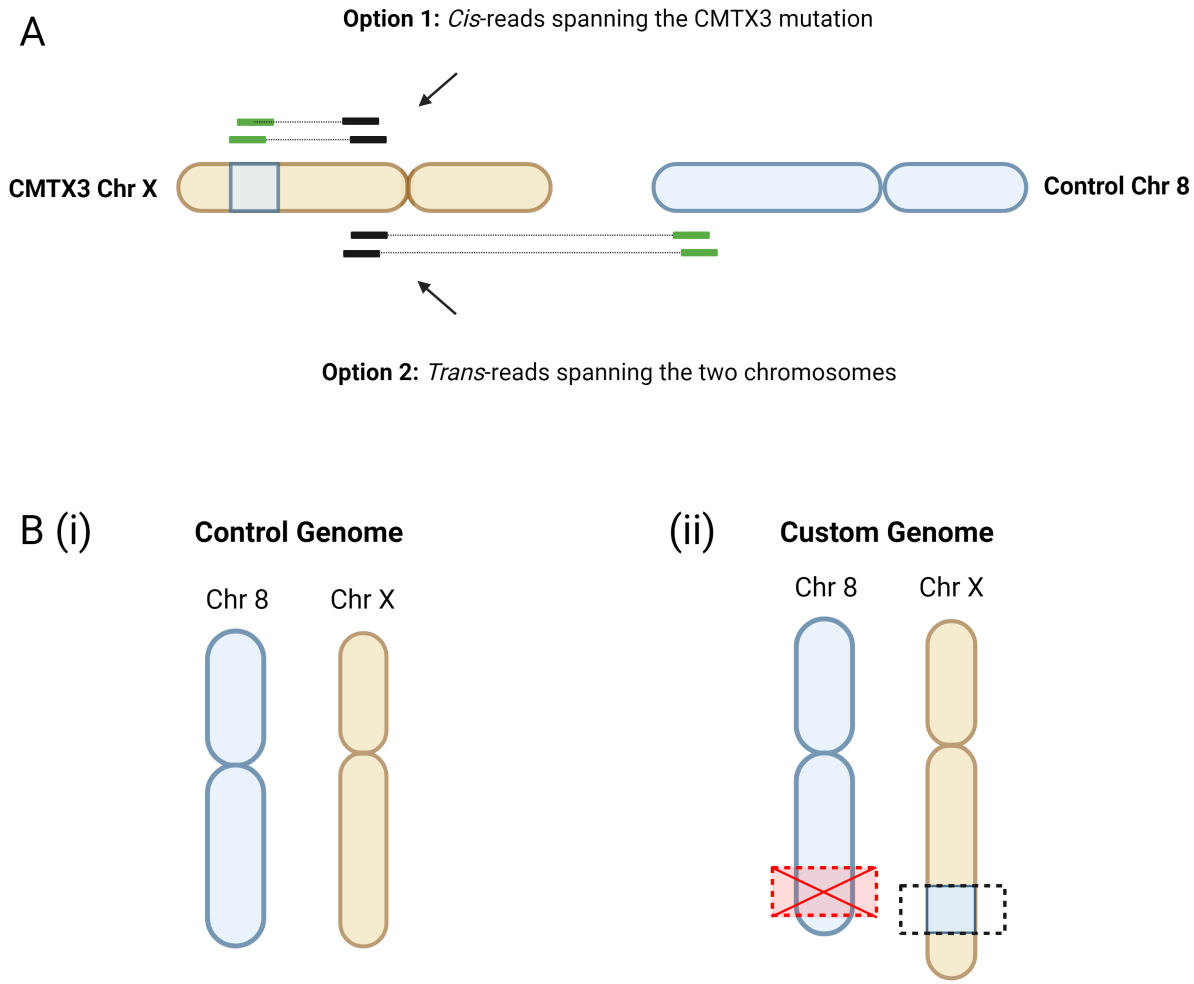


Figure 5.4: Rationale for designing the custom-built CMTX3 patient reference genome. (A) Possible alignment errors that may arise due to the presence of the unbalanced CMTX3 insertion which contains sequence that is identical to a ~78 kb region located on chromosome 8 (blue). Examples of read pairs are shown, which could theoretically be mapped in one of two ways. Option 1: Read 1 (green) maps within the CMTX3 insertion, while Read 2 (black) maps to adjacent sequence on chromosome X (yellow). This mapping configuration is a '*cis*'-read pair, likely originating from chromatin contacts occurring between the CMTX3 SV and nearby DNA on chromosome X. Option 2: Read 1 (green) instead maps to the corresponding sequence on chromosome 8q24.3, whilst Read 2 maps to chromosome X. This mapping configuration is a '*trans*'-read pair, which arises from inter-chromosomal interactions. The frequency with which each read pair is observed distinguishes 'valid' chromatin contacts from background noise. Dispersing reads across the two genomic loci may increase the risk of diluting the signal originating from the CMTX3 SV, and therefore impact the sensitivity to identify valid contacts involving this sequence. (B) (i) Control reference chromosomes; chromosome 8 and chromosome X. (ii) Modifications made to generate the custom CMTX3 reference genome. The ~78 kb region on chromosome 8 has been inserted into the long arm of chromosome X and deleted from the original location. (Images not to scale).

The hg38 reference chromosome 8 sequence was downloaded from Ensembl [241] Release 97 and edited using a custom script (Appendix D). This involved first deleting the target sequence from chromosome 8 and then re-formatting the modified file to comply with FASTA formatting standards.

5.2.5.3 Constructing a complete CMTX3 reference genome incorporating the modified chromosome files

Reference FASTA files for the remaining 22 human chromosomes as well as mitochondrial DNA were downloaded from Ensembl [241] Release 97 and combined with the custom chromosomes, in the correct order, to generate a complete, patient-specific hg38 reference genome. This was performed using a custom script, available in Appendix E.

5.2.5.4 Editing coordinates of target capture sequences

Dovetail Genomics supplied '.bed' files specifying the genomic location of the target capture baits and probes used in the Pan Promoter Enrichment Panel (https://dovetail-capture.readthedocs.io/en/latest/data_sets.html). The location of these sequences required modification to align with the coordinates of the custom-built CMTX3 reference genome. Baits/probes located distal to the CMTX3 SV breakpoint on chromosome X (chrX: 140,420,783) were shifted telomeric by 77,854 bp (reflecting the net gain of the CMTX3 SV). Conversely, baits/probes located distal to the deletion on chromosome 8 (chr8:144,542,928-144,620,773) were shifted centromeric by 77,846 bp (reflecting the size of the deletion). Baits/probes contained within the CMTX3 insertion sequence were moved from the original location on chromosome 8q24.3 to the updated location on chromosome Xq27.1. Code for editing the capture bait/probe coordinates is available in Appendix F.

5.2.6 Indexing the reference genomes

Indexing of patient and control reference genomes was performed using both BWA (version 0.7.17) [243] and SAMtools (version 1.15.1) [244], through the 'IndexReferenceFasta-nf' pipeline (https://github.com/Sydney-Informatics-Hub/IndexReferenceFasta-nf).

5.2.7 Analysis of Promoter Capture Micro-C data

Data analysis was performed according to the instructions outlined in Dovetail Genomics' Promoter Panel data analysis documentation (https://dovetail-capture.readthedocs.io/en/latest/index.html). This pipeline has been optimized for the analysis of data generated using Dovetail Genomics' Pan Promoter Enrichment panel (Cantata Bio). A summary of the key steps within this pipeline is provided below, whilst annotated code is available in Appendix G.

Raw FASTQ files were aligned to the appropriate reference genome using the BWA-MEM algorithm (BWA version 0.7.17) [243]. The control sample was aligned to hg38. The patient sample was aligned to both the custom CMTX3 reference genome and the control hg38 reference genome, to extract maximum information from this exploratory analysis. Pairtools (version 1.0) [245] was then used to process the Micro-C data, performing a range of tasks including; detecting ligation junctions, sorting the contact pairs, eliminating reads arising from PCR duplication, generating QC statistics, and producing a BAM file for downstream analysis. SAMtools (version 1.17) [244] was then used to process and convert the BAM file into a format required for downstream analysis. A library QC report was generated by running the QC script ('get_qc.py') supplied in the Dovetail analysis documentation (https://github.com/dovetail-genomics/capture.git). SAMtools [244] and Mosdepth (version 0.3.3) [246] were then used to assess QC of the target enrichment process. Significant chromatin interactions were then detected using CHiCAGO (version 1.22.0) [247]. Input files for CHiCAGO were prepared using the chicagoTools scripts (https://github.com/dovetail-genomics/chicago). First, 'chinput' files were generated using the 'bam2chicago.sh' script. 'Design files', which contain information on bait sequences and genome fragmentation, were supplied by Dovetail Genomics (https://dovetail-capture.readthedocs.io/en/latest/data_sets.html#id2). However, custom design files were generated for the patient-specific reference genome, according to the instructions outlined in the analysis documentation. The recommended binsize of 10 kb was used. The CHiCAGO [247] analysis generates output files in '.ibed' format, which were subsequently converted into '.bedpe' files to enable

visualization of contact loops using Integrated Genomics Viewer (IGV, version 2.16.1) [248]. Contact matrices were created by converting the '.pairs' files into '.hic' files using JuicerTools [249], and visualised using Juicebox software (version 1.11.08) [250].

5.3 Results

Given the high cost associated with this technology and availability of a modest budget, a 1 versus 1 experimental design was employed. In the absence of biological replicates, it is recognised that the results reported are exploratory in nature and should therefore be interpreted with caution.

5.3.1 A custom-made patient-specific reference genome accurately recapitulates the CMTX3 SV

The experimental approach taken in this chapter involved aligning patient Micro-C reads to a custom-built reference genome harbouring the CMTX3 SV. A patient-specific reference genome was created by modifying the chromosome Xq27.1 reference sequence to contain the ~78 kb CMTX3 inter-chromosomal insertion. The inserted sequence was also deleted from its original location on chromosome 8q24.3, to overcome challenges that may arise from multi-mapping reads. To confirm the desired genomic rearrangements were successfully incorporated into the custom-built genome, the DNA sequence was reviewed using SnapGene Viewer (GSL Biotech LLC). Figure 5.5A-B demonstrates the sequences at both the proximal and distal breakpoints of the CMTX3 insertion are identical to the sequences originally reported by Brewer and colleagues upon discovery of this pathogenic mutation [63]. Figure 5.5C confirms the two edited chromosomes were modified by the appropriate sizes, and that all other chromosome sizes were equivalent to the control hg38 reference. For downstream analysis of the Promoter Capture Micro-C data, the location of the sequences used for target enrichment are required as input, and therefore had to be adjusted to align with the custom-built genome. The coordinates of all affected baits were modified, and the resulting files were assessed

using IGV [248], which confirmed that the updated coordinates accurately represent the genomic rearrangements observed in the custom-built genome (Supplementary Figure 5.1).

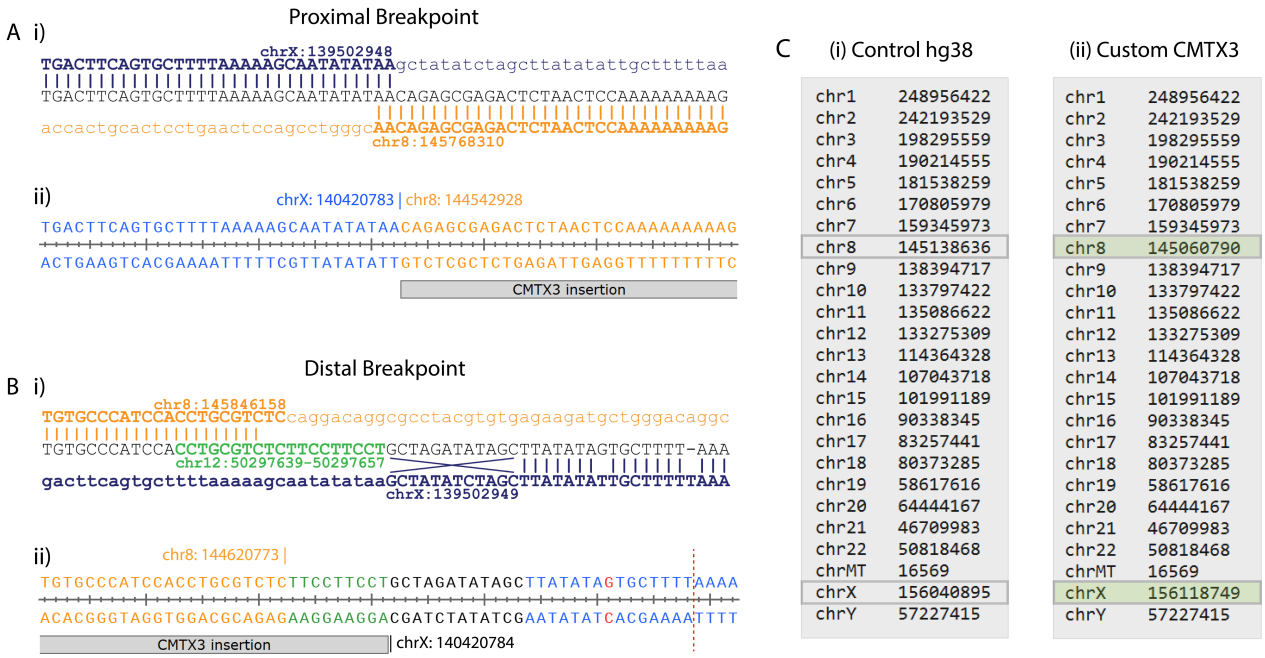


Figure 5.5: Overview of the custom-built CMTX3 reference genome. (A) (i) Schematic of the DNA sequence (hg19) at the proximal breakpoint of the CMTX3 SV. Image sourced from Brewer et al., (2016) [63]. Sequence from the CMTX3 patient is represented in black. Blue and orange text represents the reference sequence for chromosome X and chromosome 8, respectively. (ii) Double-stranded DNA sequence of the custom-built CMTX3 chromosome X FASTA used in the current study, visualised using SnapGene Viewer. Proximal breakpoint sequence is identical to (i), showing a clear junction between chromosome X (blue) and the chromosome 8 insertion (orange). Updated hg38 coordinates were obtained using the UCSC LiftOver tool. (B) (i) Schematic of the DNA sequence (hg19) at the distal breakpoint of the CMTX3 SV. Image sourced from Brewer et al., (2016) [63]. The chromosome 8 insertion (orange) is followed by a small insertion from chromosome 12 (green), a 12 bp inversion of chromosome X, a single nucleotide variant and a single nucleotide deletion. (ii) Double-stranded DNA sequence of the custom-built CMTX3 chromosome X FASTA, visualised using SnapGene Viewer. Distal breakpoint sequence incorporates all the genetic rearrangements outlined in (i). Red dotted line marks the position of the single nucleotide deletion. Updated hg38 coordinates are shown. (C) A list of the sizes of the chromosomes contained within the control hg38 reference genome (i) and the custom-built CMTX3 reference genome (ii), obtained from the indexed FASTA files. The custom-built genome is characterised by a 77,846 bp loss from chromosome 8 and a 77,854 bp gain at chromosome X.

5.3.2 Quality control of the Promoter Capture Micro-C libraries

The Promoter Capture Micro-C libraries underwent QC assessment using the bioinformatics tool, Pairtools [245]. This tool provides a summary of various QC metrics, including the total number of read pairs, the number of mapped and unmapped read pairs, the number of PCR duplicates, and the number of total valid ligation events. A valid ligation event occurs when paired sequencing reads align to different chromosomes (i.e. trans reads), or to distinct regions (separated by ≥ 1 kb) located on the same chromosome (i.e. cis reads). These are the desired read types obtained from proximity ligation assays. In comparison, cis read pairs that map within 1 kb of each other can arise from the re-ligation of adjacent sequences of DNA, and are classified as 'invalid' as they confer no useful information regarding 3D chromatin organisation [251]. Figure 5.6 provides an overview of the QC metrics for each of the two libraries analysed, with additional summary statistics available in Supplementary Table 5.2. Whilst the total number of mapped read pairs is elevated in the control library, the proportion of non-duplicate reads and PCR duplicates are similar between the two samples, as is the proportion of valid read pairs which account for 89.74% and 87.43% of the non-duplicate mapped reads for the patient and control libraries, respectively (Figure 5.6). Cis reads accounted for 52.32% of valid read pairs in the patient library, and 44.98% in the control library (Figure 5.6). To assess the performance of the target enrichment process, the coverage depths at target and non-target regions of the genome were compared using the bioinformatics tool, Mosdepth (Pedersen and Quinlan, 2018). For the patient Micro-C library, the mean coverage depth at captured regions was 323.13 compared to a value of 12.2 at non-captured regions, indicating 26.5 times higher coverage at targeted regions of the genome. Similar values were observed within the control library, which showed 27.8 times higher coverage at targeted regions, with mean coverage depths of 395.28 and 14.21 for targeted and non-targeted regions, respectively.

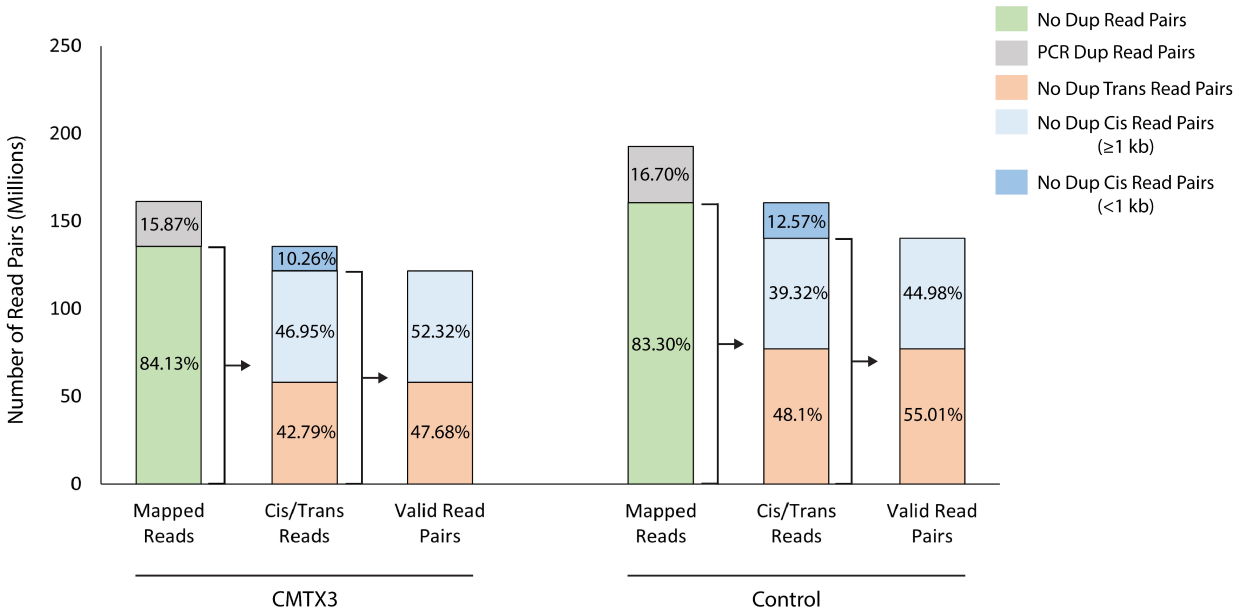


Figure 5.6: Quality control metrics for the Promoter Capture Micro-C libraries. Mapped reads represent the total number of read pairs exceeding the minimum mapping quality (MAPQ) score threshold (MAPQ > 40). Mapped reads are then labelled as either PCR duplicates (PCR Dup) or non-duplicate reads (No-Dup). No Dup reads are further characterised as trans read pairs, mapping to different chromosomes, or cis read pairs, mapping to the same chromosome. Cis reads are sub-classified based on a distance threshold of 1 kb. Trans read pairs and cis read pairs (≥ 1 kb) are collectively defined as 'valid' ligation events.

5.3.3 Contact matrices enable visualisation of the effect of the CMTX3 SV upon local chromatin architecture

Contact matrices are used to visualise the complex, large datasets generated from Hi-C and related technologies, providing a graphical overview of the many pairwise interactions identified in these analyses. In a contact matrix, genomic coordinates are plotted on both the X and Y axes and pairwise chromatin interactions are displayed as a heatmap, where the colour intensity of each point in the matrix represents the frequency of interactions that were observed between the two corresponding genomic loci (Figure 5.7).

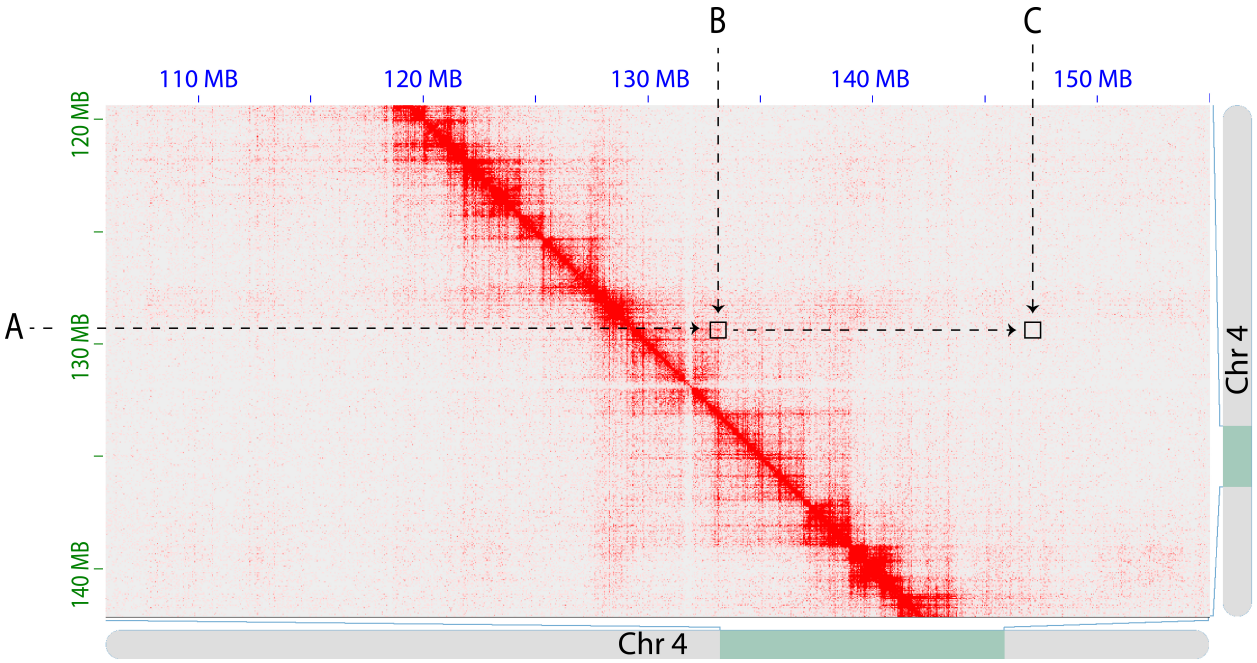


Figure 5.7: Data from Hi-C based analyses is commonly displayed as a contact matrix. Example of the contact matrix format, using data obtained from a Promoter Capture Micro-C experiment. Data is from a defined region of chromosome 4, with coordinates marked on the X and Y axes. Note that the matrix is symmetrical about the diagonal. Pairwise interactions between two genomic loci are represented as red dots, where the intensity of the dot increases with the number of interactions that are observed. For example, the intersection of loci 'A' and 'B' shows a much higher intensity than the intersection of points 'A' and 'C'. This indicates that loci 'A' and 'B' engage in more frequent chromatin contacts than loci 'A' and 'C'.

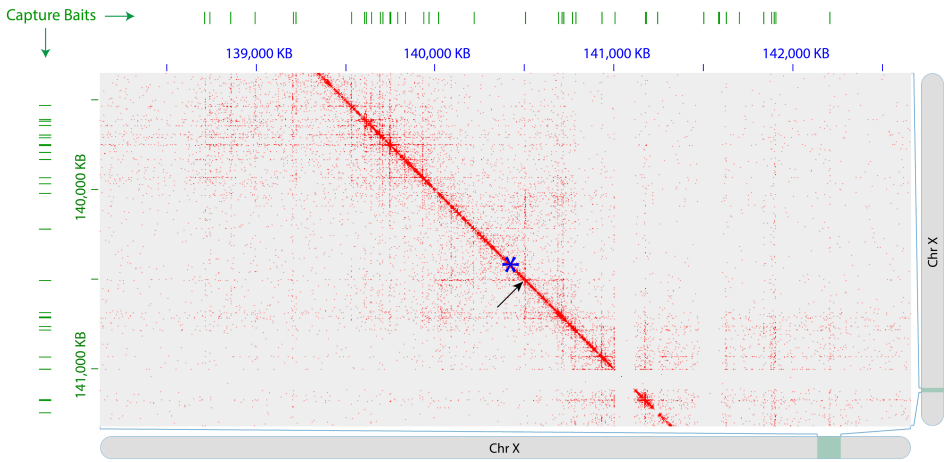
Contact matrices were generated to gain insight into the chromatin interactions occurring at the CMTX3 locus, and to visually compare the interactions observed in the wild-type state with those observed in the presence of the pathogenic CMTX3 SV. It is important to note that raw, unnormalized contact matrices were used for the purposes of this exploratory study, as standard Hi-C normalisation options do not account for the unique statistical properties of Capture Hi-C/Micro-C datasets [252].

Figure 5.8A represents the raw contact matrix obtained from the control sample, focusing on the region of Xq27.1 immediately surrounding the location of the CMTX3 SV. Visual assessment of the pairwise interactions provide a general overview of how the different bait sequences are interacting with nearby chromatin. For example, all pairwise interactions that span the location of the CMTX3 SV are easily observed. One notable feature is a 'line' of pairwise interactions connecting the *SOX3* bait

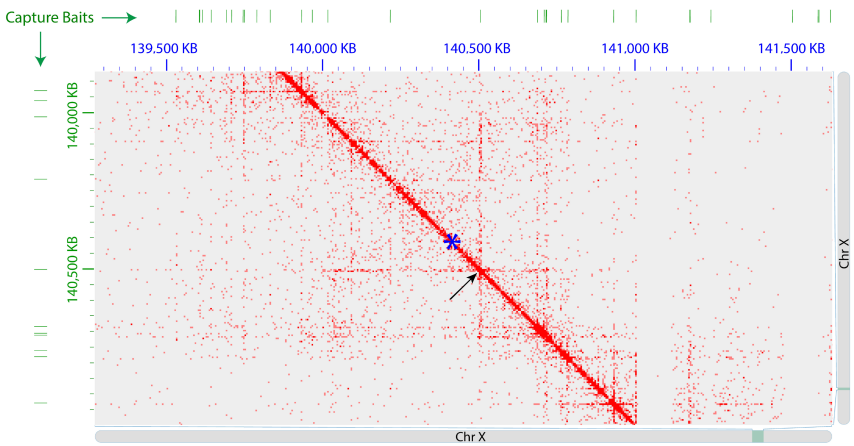
with proximal regions of DNA, with many of these interactions spanning the site of the CMTX3 SV breakpoint (Figure 5.8A). Interactions between the *SOX3* bait and distal DNA sequences are also visible although appear to be confined within a relatively smaller genomic window (Figure 5.8A).

Figure 5.8B represents the raw contact matrix obtained from the patient sample, which shares many similarities with the control. One obvious difference is the presence of the ~78 kb CMTX3 insertion sequence, denoted by a dark red square along the diagonal of the heatmap (Figure 5.8B). The increased colour intensity at these genomic coordinates is a consequence of how the custom-built patient reference genome was designed, which involved removing the ~78 kb duplicated region from the original location at 8q24.3 to prevent multi-mapping reads. This approach forced any reads involving this sequence to map to the CMTX3 insertion at Xq27.1, leading to an increased signal at this locus. The ~78 kb inserted fragment contains one bait sequence which is annotated on the heatmap (Figure 5.8B). Diagonal 'lines' extending from this locus denote pairwise interactions between the captured sequence and nearby regions on chromosome X (Figure 5.8B). While some of these interactions appear to involve the nearby *SOX3* bait (Figure 5.8B(ii)), it should be noted that bait-to-bait interactions are expected to show higher background signal than interactions with non-baited 'other-ends', since both ends are selectively enriched in the library [247]. Similar to the control contact matrix in Figure 5.8A, the *SOX3* bait appears to participate in more pairwise interactions with proximal, rather than distal, DNA sequences (Figure 5.8B).

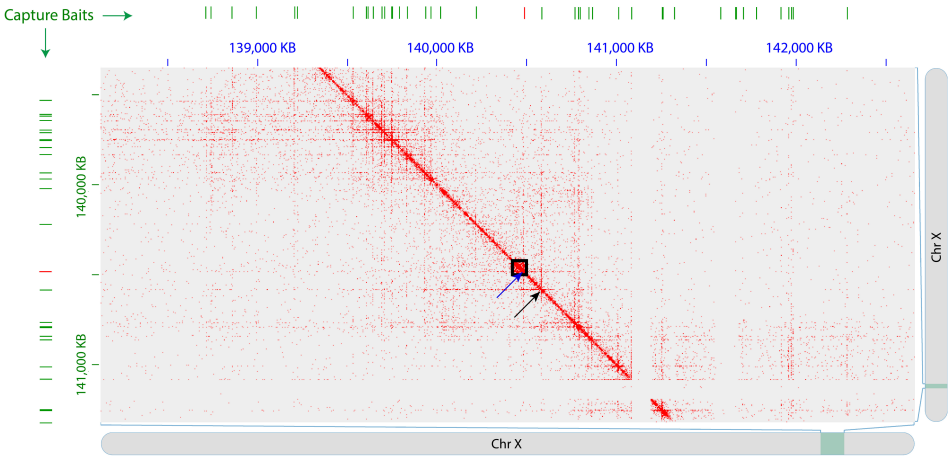
A i) Control



A ii) Control zoom



B i) CMTX3 (Custom Genome)



B ii) CMTX3 (Custom Genome) zoom

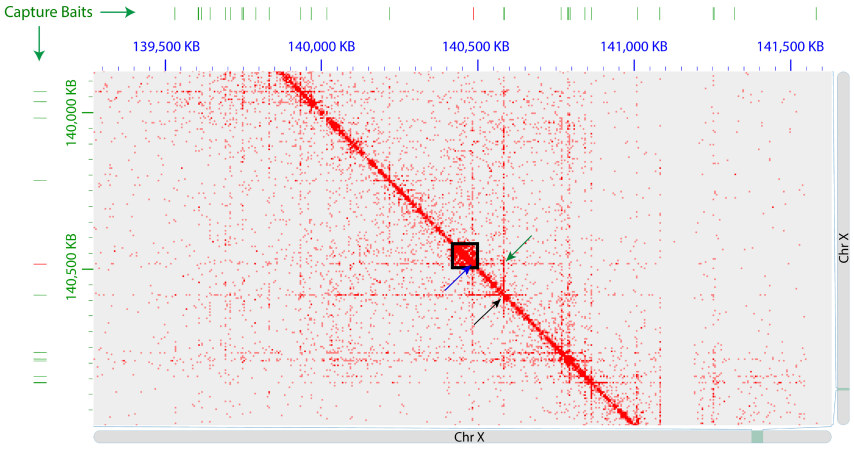


Figure 5.8: Overview of the chromatin interaction landscape across a ~2 Mb window within the CMTX3 locus. Raw contact matrices for the Promoter Capture Micro-C libraries were generated using Juicebox software and visualised at 5 kb resolution. (A) Contact matrix of control sample aligned to hg38. Key features have been marked, including a blue asterisk which highlights the location of the CMTX3 breakpoint, and a black arrow denoting the position of the *SOX3* bait. (B) Contact matrix of the patient sample aligned to the custom-built CMTX3 reference genome. The CMTX3 SV is outlined by a black box. Black arrow denotes the position of the *SOX3* bait and the blue arrow highlights the location of the *ARHGAP39* bait contained within the CMTX3 SV. Green arrow points to apparent pairwise interactions between the two baits. The location of the bait sequences used for target enrichment are marked above each heatmap in green. The bait contained within the CMTX3 insertion is marked in red.

5.3.4 Identification of statistically significant chromatin interactions

The contact matrices in Figure 5.8 provide a general visual overview of the local chromatin landscape with and without the CMTX3 insertion. However, to identify statistically significant chromatin interactions amongst the many millions of pairwise contacts, additional analyses are required. To detect significant interactions, we utilised the bioinformatics package, CHiCAGO [247], which was specifically developed for the analysis of Capture Hi-C data. This tool accounts for the unique sources of bias that arise from the target enrichment step which are absent from typical Hi-C approaches. Significant chromatin interactions are defined as those displaying a higher frequency of contacts than would be expected for read pairs located at a given genomic distance [247]. Typically, interactions are considered statistically significant if the 'CHiCAGO score' (soft-thresholded, weighted -log *P*-value) exceeds the stringent cutoff value of 5 [247, 253]. However, given the exploratory nature of this study, the data was also analysed using a less stringent cutoff score of 3. Interactions were identified using a binsize of 10 kb for 'other-end' fragments. Supplementary Figure 5.2 provides an overview of the chromatin interactions that were identified across the CMTX3 locus at Xq27.1 using both threshold values.

The absolute number of genome-wide chromatin interactions identified by the CHiCAGO analysis was notably similar between the patient and control samples. Using a cutoff CHiCAGO score of 5, a total of 30,404 chromatin interactions were identified in the control sample and 29,466 in the patient (Figure 5.9A(i)). As expected, many more interactions were detected using a cutoff value of 3,

totalling 188,700 and 190,749 within the control and patient libraries, respectively (Figure 5.9B(i)). At both cutoff values, approximately 18% of the interactions identified in the control library were also present in the patient library, whilst the remainder were sample-specific (Figure 5.9A(i), 5.9B(i)). The patient and control libraries showed a comparable proportion of cis and trans interactions at both cutoff values (Figure 5.9A(ii), 5.9B(ii)). It should be noted that these summary statistics have been calculated by aligning the patient library to hg38, as the altered coordinates of the custom-built reference genome were found to hinder this type of comparative analysis.

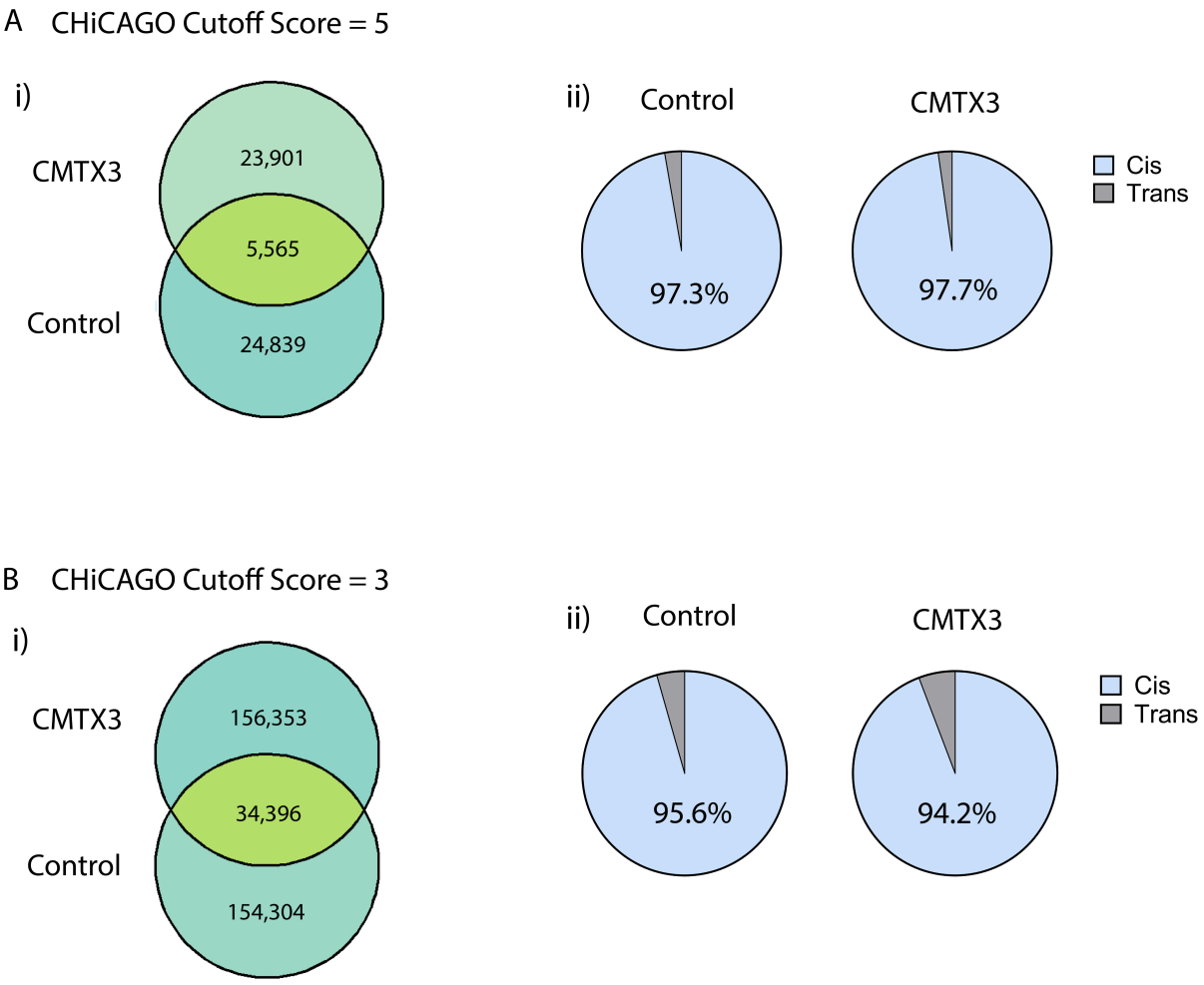


Figure 5.9: Summary of CHiCAGO outputs. Overview of chromatin interactions identified from the patient and control Promoter Capture Micro-C libraries with the CHiCAGO pipeline, using a stringent cutoff score of 5 (A) and a lenient cutoff score of 3 (B). (i) Venn diagrams show the total number of interactions detected by the analysis, including the number of interactions unique to each sample, and those that are shared between the two. (ii) Pie charts display the percentage of interactions that are intra-chromosomal (cis) or inter-chromosomal (trans).

5.3.5 Does the CMTX3 SV form ectopic chromatin contacts with genes in the CMTX3 locus?

One hypothesis that we aimed to address within the present study was whether the CMTX3 SV forms ectopic chromatin interactions with genes in the CMTX3 linkage region, as this could provide a plausible explanation for the gene dysregulation observed in CMTX3. The following question was therefore posed; "Do any of the significant interactions detected in the patient sample show sequence from the CMTX3 linkage region interacting with sequence from the CMTX3 SV?" To facilitate this investigation, the data was represented as 'contact loops' rather than contact matrices, in which regions of DNA sequence showing significant chromatin interactions are joined with an arc. This simple visual format is particularly useful for conducting a fine-scale investigation of specific genomic loci.

When the stringent CHiCAGO cutoff score of 5 was used to detect significant interactions, ectopic interactions involving the CMTX3 SV could not be observed (Figure 5.10). The more lenient cutoff score of 3 was then used to further investigate the local genomic landscape by identifying the 'sub-threshold' interactions across the region. Substantially more chromatin loops were observed with the lower threshold (Figure 5.11A). The subset of loops where either end of the interaction (i.e. 'bait' or 'other-end' fragment) was located within the CMTX3 SV was then isolated (Figure 5.11B). This identified four loops connecting the CMTX3 SV with nearby sequence on chromosome X (Figure 5.11B). Interestingly, none of these loops connected gene promoters (baits) from the CMTX3 locus to 'other-end' fragments residing within the CMTX3 SV. Instead, all four loops were anchored at the bait for the *ARHGAP39* promoter, which is contained within the CMTX3 SV. Two of these loops connected the bait with proximal regions of DNA, spanning distances of ~ 280 kb (CHiCAGO score = 3.67) and ~ 430 kb (score = 4.18), respectively (Figure 5.11B). The other two loops interacted with distally located sequence within the CMTX3 linkage region. One spans a distance of ~ 90 kb (CHiCAGO score = 4.06) while the other spans ~ 460 kb (CHiCAGO score = 4.25) (Figure 5.11B). Interestingly, one of these loops involved an 'other-end' fragment located near *SOX3* (Figure 5.11B).

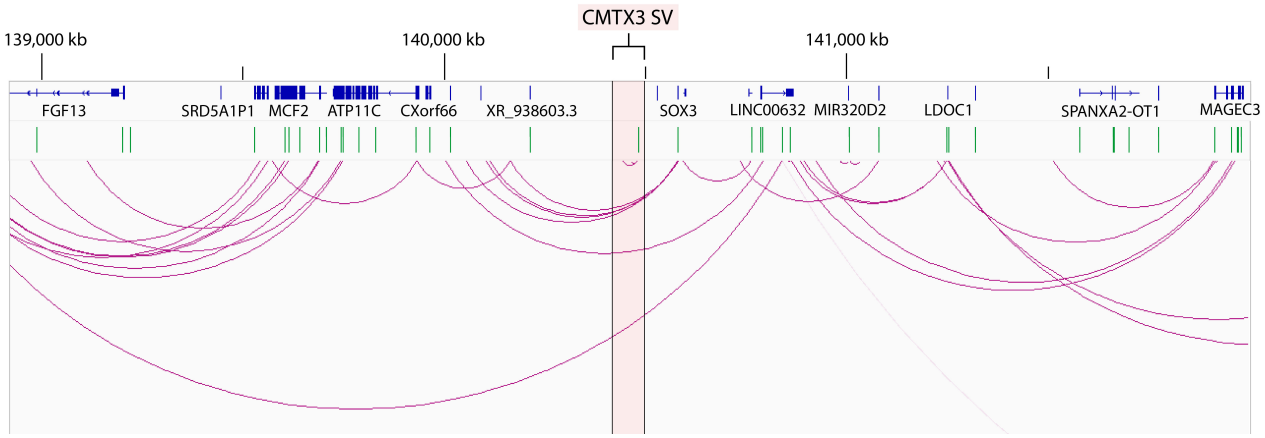


Figure 5.10: Overview of the cis chromatin interactions identified in the region surrounding the CMTX3 SV, using the stringent CHiCAGO cutoff score of 5. A portion of the CMTX3 linkage region is shown, including 1.5 Mb either side of the CMTX3 SV (highlighted in red). An overview of the RefSeq genes contained within the region is shown (blue). The location of bait sequences used in the Pan Promoter Enrichment Panel are marked with green bars. Significant chromatin contacts (i.e. those with a CHiCAGO score ≥ 5) are represented as magenta loops. The CHiCAGO analysis was conducted using a genomic bin size of 10 kb. Data was visualised using the IGV software. Chromatin loops between the CMTX3 SV and the host Xq27.1 are not observed. Two, short-range loops are contained entirely within the boundaries of the CMTX3 SV itself. These are likely an artefact of the unbalanced nature of the CMTX3 SV; i.e. more reads than expected are mapping at this locus.

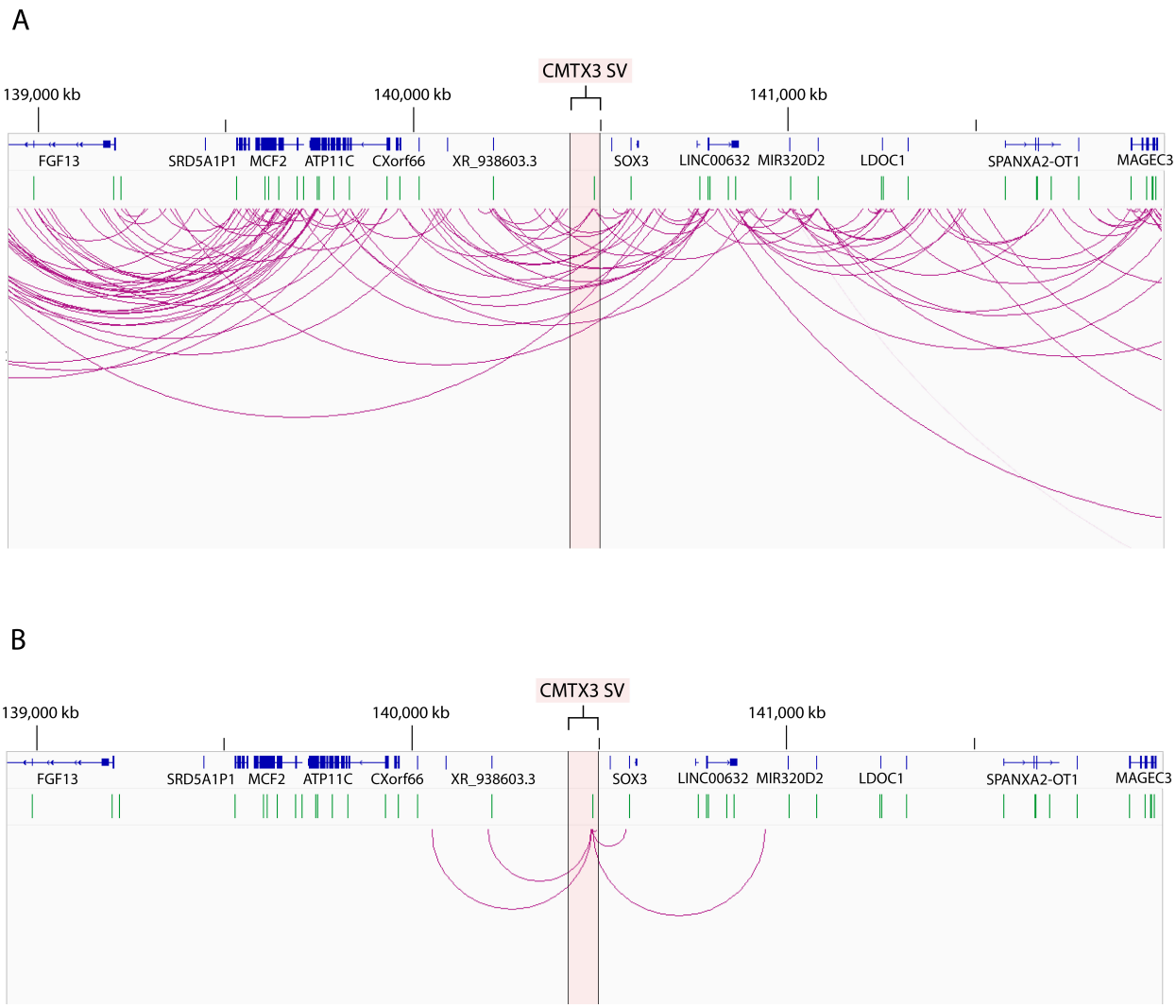


Figure 5.11: Overview of the sub-threshold cis chromatin interactions identified in the region surrounding the CMTX3 SV, using the lenient CHiCAGO cutoff score of 3. Figure depicts a region of Xq27.1, including 1.5 Mb either side of the CMTX3 SV. Genomic coordinates are shown, and the ~78 kb CMTX3 insertion sequence is highlighted in red. An overview of the RefSeq genes contained within the region are shown in blue, and the position of bait sequences used in the Pan Promoter Enrichment Panel are marked with green bars. CHiCAGO analysis was conducted using a genomic bin size of 10 kb. (A) All sub-threshold cis chromatin contacts (i.e. those with a CHiCAGO score ≥ 3 but < 5) within the region are represented as magenta loops. (B) Figure depicts only those interactions where either end of the contact loop (i.e. 'bait' or 'other-end' fragment) is localised within the boundaries of the CMTX3 SV. All the loops identified are anchored at the bait sequence contained within the CMTX3 SV. One of these loops involves the 10 kb bin located adjacent to the *SOX3* bait. Another loop involves the 10 kb bin adjacent to the *LOC105373343* (XR_938603.3) bait. Data was visualised using the IGV software.

5.3.6 Investigating if the CMTX3 SV disturbs chromatin interactions on chromosome X

Another possible hypothesis is the CMTX3 SV may dysregulate genes by disrupting interactions with their canonical CREs. To investigate this possibility, the patient library was aligned to the hg38 reference. When considering pairwise interactions where both reads map to chromosome X, alignment to the hg38 reference provided an easier method for comparing patient and controls, as the genomic coordinates were identical between the two samples. If using the patient-specific reference genome, the coordinates of genomic DNA located distal to the CMTX3 SV would be shifted by ~78 kb. Detecting differential interactions between Capture Micro-C datasets requires the use of specialised bioinformatics tools, including Chicdiff [254]. However, this type of analysis also requires biological replicates, which were unavailable in this study, which limited the capacity to address this hypothesis. Despite this limitation and in the absence of replicates, the patient and control contact matrices were visually compared to search for any possible trends in the data (Figure 5.12). This exploratory analysis was inconclusive, reflecting the challenges of discriminating differences in fine-scale chromatin interaction profiles (Figure 5.12). The significant chromatin interactions that are marked on the contact maps provide a visual overview of the region (Figure 5.12). They should however, not be used for directly comparing interactions between the two samples. Comparative analyses should consider the score of each interaction, not simply those interactions that crossed the significance threshold, due to the undersampled nature of Capture Hi-C datasets [247, 255].

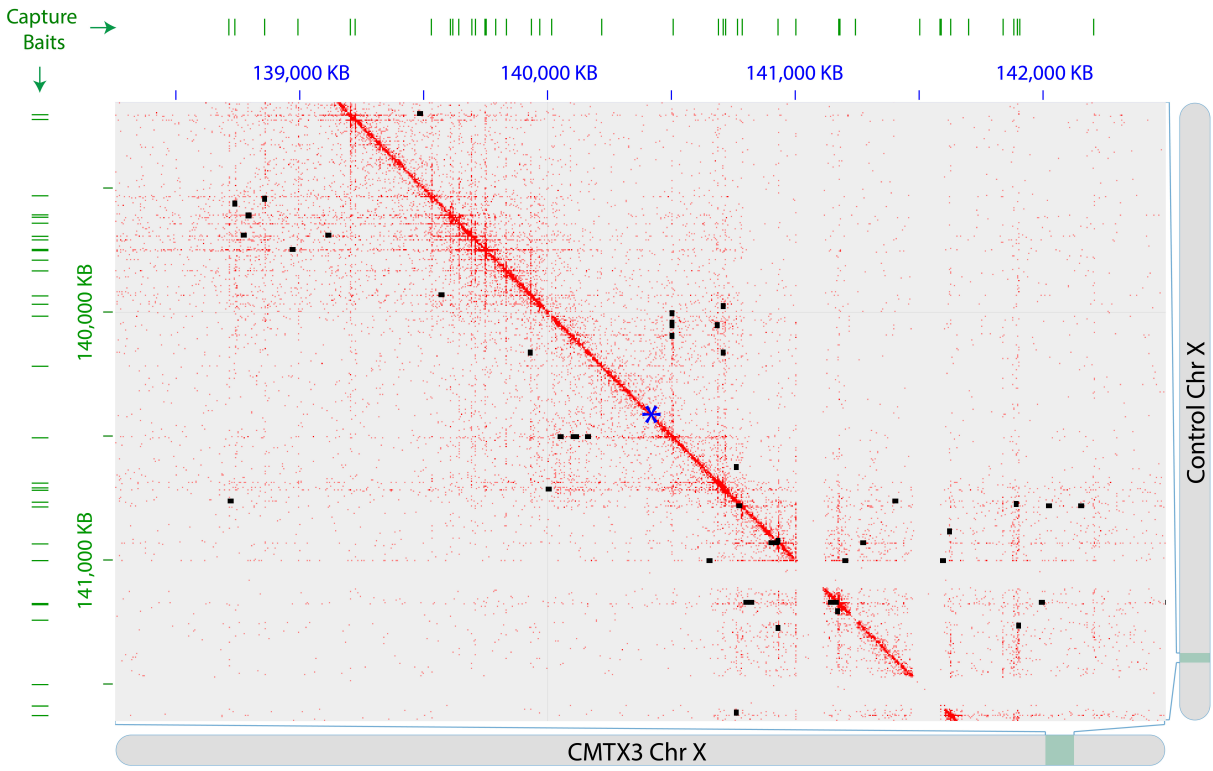


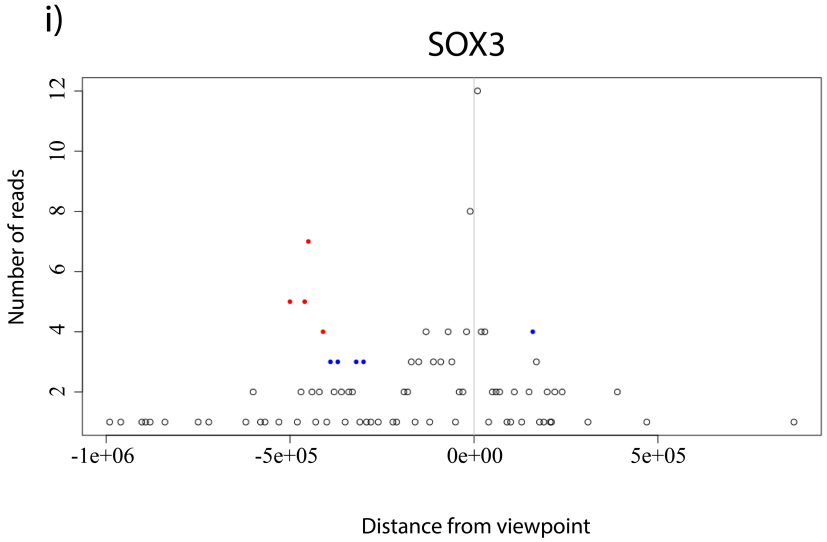
Figure 5.12: Comparison of patient and control contact matrices at Xq27.1 (hg38). The diagonal line splits the contact matrix so that data from the patient sample is displayed on the bottom half, and the control sample on the top half. Raw contact maps are shown, with both samples aligned to hg38. Significant chromatin interactions (CHiCAGO score ≥ 5) are marked as black dots. Location of baits for Promoter Capture are marked with green bars. Blue asterisk denotes the location of the breakpoint for the CMTX3 SV. Many general similarities between the samples can be observed. Visually discerning differences in fine-scale interaction patterns is challenging. Heatmaps were visualised with Juicebox software using a resolution of 5 kb.

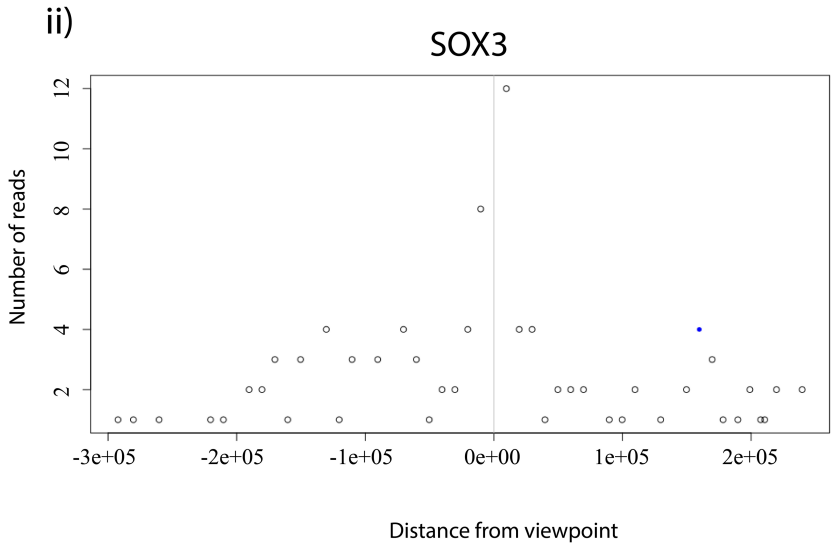
5.3.7 Investigating chromatin interactions involving the high-priority candidate gene SOX3

In Chapter 4, CMTX3-derived NEP were reported to show dysregulated expression of *SOX3*. This provided a high-priority candidate gene to focus the analysis on specific chromatin interactions of interest. To gain a deeper insight into the pairwise interactions occurring at the *SOX3* promoter, the number of read pairs anchored at the *SOX3* bait were plotted against the linear genomic distance (Figure 5.13). The distribution of interactions can be observed, including those mapping within the CMTX3 insertion (Figure 5.13B). Both patients and controls show a higher proportion of interactions mapping proximal (centromeric) to the bait, rather than distal (telomeric) (Figure 5.13). Patients

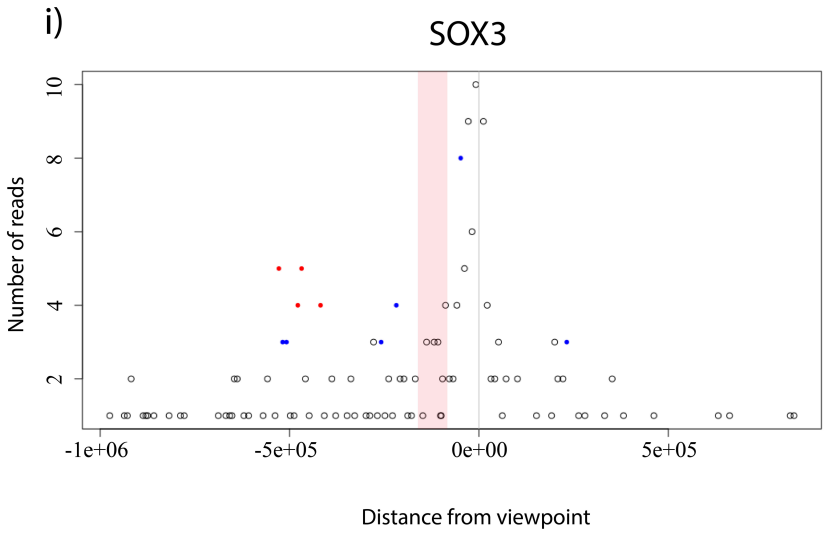
showed a higher number of 'sub-threshold' interactions involving the *SOX3* bait and the proximal 300,000 kb region than controls (Figure 5.13A(ii), 5.13B(ii)).

A Control (hg38)





B CMTX3 (Custom Genome)



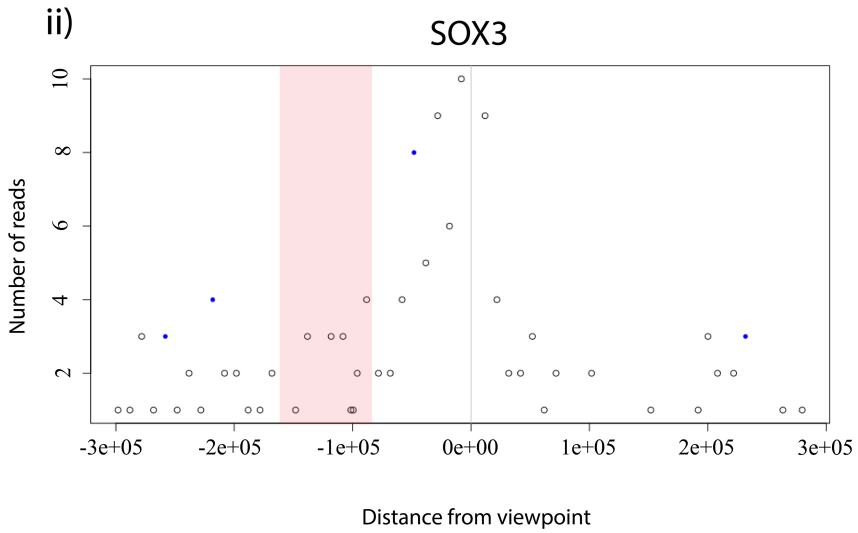


Figure 5.13: Visualisation of pairwise interactions at the *SOX3* bait. Plots of the pairwise interactions anchored at the *SOX3* promoter bait (i.e. the viewpoint) against the linear genomic distance. 'Other-end' fragment size is 10 kb. Interactions with a CHiCAGO score exceeding 3 are blue, and those exceeding 5 are red. (A) Data from the control sample is aligned to hg38. (B) Data from the patient sample is aligned to the custom-built CMTX3 reference genome. Therefore, coordinates do not align when comparing the two samples. Red shading denotes the location of the ~78 kb CMTX3 SV. Interactions are shown for distances up to 1 Mb (i) or 300,000 kb (ii) from the *SOX3* bait. Plots generated using the 'plotBaits' function of CHiCAGO.

5.4 Discussion

When considering the various pathogenic insertions reported at the Xq27.1 hotspot (reviewed in Chapter 1, Part B), a common hypothesis is the complex SV may cause disease by introducing additional regulatory elements into the Xq27.1 region, which may interact with nearby genes and drive misexpression [63, 65, 198, 199]. However, to date, no studies have experimentally assessed the effect of the pathogenic SV upon local regulatory interactions. In this chapter, we employed a 3C-based technology to explore 3D chromatin architecture in the context of the CMTX3 complex insertion. Specifically, we used the Dovetail Pan Promoter Panel Micro-C approach to gain insights into promoter-specific chromatin interactions, given their importance in long-range gene regulation (e.g. enhancer-promoter contacts, promoter-promoter contacts) [232, 256]. The analysis was performed on NEP, as this cell-type showed clear differences of *SOX3* expression levels between patient and control samples (Chapter 4, Table 4.2).

Key to this analysis was the construction of a custom patient-specific reference genome harbouring the CMTX3 SV, to provide the appropriate genetic background to investigate chromatin interactions. The custom modifications to the chromosome X reference faithfully recapitulated the known sequence of the CMTX3 SV [63], and removal of the ~78 kb duplicated fragment from its original location on chromosome 8 was successfully achieved to minimise the risk of multi-mapping reads. Patient-specific reference genomes have previously been used to study the effects of other pathogenic SV, such as the 1.35 Mb genomic rearrangement causing DHMN1 [67]. This led to the identification of a neo-TAD within the DHMN1 locus of patient-derived sMN [67], demonstrating the utility of this

approach for detecting aberrant chromatin interactions involving disease-specific DNA rearrangements.

A notable observation was that the percentage of trans read pairs was higher than expected in both the patient and control Micro-C libraries. Typically, trans read pairs account for ~20-30% of the total read pairs generated from a Dovetail Genomics Pan Promoter Encrichment Micro-C library, after the removal of PCR duplicates (https://dovetail-capture.readthedocs.io/en/latest/library_qc.html#library-qc). In the present study, values between ~43-48% were observed. A high ratio of trans reads suggests high rates of random ligations, which contribute to technical noise within the library [257, 258]. Thus, the cis/trans ratio is often used to infer library quality [228]. Reassuringly, despite the elevated trans read pairs, the two libraries used in this study successfully passed Cantata Bio's in-house QC assessment and were similar with respect to overall QC metrics.

After generating contact matrices to visualise how the CMTX3 SV fits into the broader genomic landscape, analysis to identify 'significant' chromatin interactions was performed. A particular focus of this analysis was to determine whether any chromatin loops connected promoter 'baits' from the CMTX3 linkage region to 'other-end' fragments residing within the ~78 kb inserted fragment. These types of interactions would support the hypothesis that the CMTX3 SV forms ectopic interactions with nearby genes. No such loops were observed within our dataset (Figure 5.10). In this analysis, 'significant' interactions were defined as those with a CHiCAGO score ≥ 5, the default cutoff value for the program [247] which is common practice (e.g. [259-261]. However, this cutoff has been described as 'stringent' [253], and in certain applications, researchers have utilised a more lenient threshold score of 3 [262, 263]. The data was analysed for chromatin interactions using this less stringent cutoff, and also did not identify loops connecting gene promoters on chromosome X to the CMTX3 insertion (Figure 5.11B). However, multiple interactions were observed between the *ARHGAP39* promoter, contained within the CMTX3 SV, and neighbouring regions of DNA on chromosome X. It is interesting that one of these interactions involves sequence near the *SOX3* gene (Figure 5.11B). While the use of

a cutoff score is valuable for the prioritisation of candidate chromatin interactions, it should not be used to directly infer whether a given interaction is present/absent within a sample [253]. Accordingly, the results presented here have flagged potential interactions between the CMTX3 SV and nearby genomic loci, but further validation will be needed including the use of additional biological replicates, and ideally, using complementary experimental approaches (addressed below).

When interpreting the results in this chapter, it is important to distinguish between the 'statistical significance' of a chromatin interaction and its 'biological relevance'. Significant chromatin interactions are typically identified as those showing more frequent pairwise interactions than expected for any two genomic loci [247, 264]. Lajoie and colleagues (2014) [228] recommend carefully evaluating interactions that arise from these assumptions. It is possible that interactions may be overlooked between proximal DNA sequences, which are expected to show a high level of interaction but may still be functionally important. Similarly, not all significant interactions identified within these complex sequencing libraries with noisy background signals, will be biologically relevant [265].

Due to the lack of biological replicates, the present study could not adequately address the hypothesis that the CMTX3 SV may disrupt existing chromatin interactions within the CMTX3 locus. Ideally, this hypothesis should be tested by conducting a differential interaction analysis using the Chicdiff tool, [254], however, biological replicates are required. For the purposes of investigating the data from this Promoter Capture Micro-C experiment, we included a visual comparison of the patient and control contact matrices of the region immediately flanking the site of the CMTX3 SV. Whilst no striking visible differences between the samples were observed, no conclusions regarding this hypothesis can be made without appropriate statistical testing, as discerning differences in fine-scale chromatin interactions is challenging.

Given that the CMTX3-derived NEP showed dysregulation of *SOX3* expression, chromatin interactions occurring at the *SOX3* promoter were of particular interest. Plots were generated to visually assess the distribution of pairwise interactions involving this promoter. When considering interactions between the *SOX3* promoter and centromeric regions of DNA, the patient sample appears

to show more reads supporting shorter-range pairwise interactions than the control. However, this result should be considered with caution as these observations were derived from one biological replicate. A benefit of the current experimental approach was having the capacity to survey promoter-specific interactions across the CMTX3 linkage region and investigating the complete disease locus. However, for future studies, 4C-Seq [222, 223] would be a more appropriate methodology to provide the resolution required for performing a detailed characterisation of *SOX3*-specific interactions. This 'one vs all' method is used to investigate all interactions involving one specific site of interest at superior resolution. 4C-Seq has proven to be a valuable tool for detecting aberrant chromatin interactions arising from pathogenic SV, including instances of ectopic enhancer adoption [102, 212, 214, 215], as well as enhancer disconnection [107, 212].

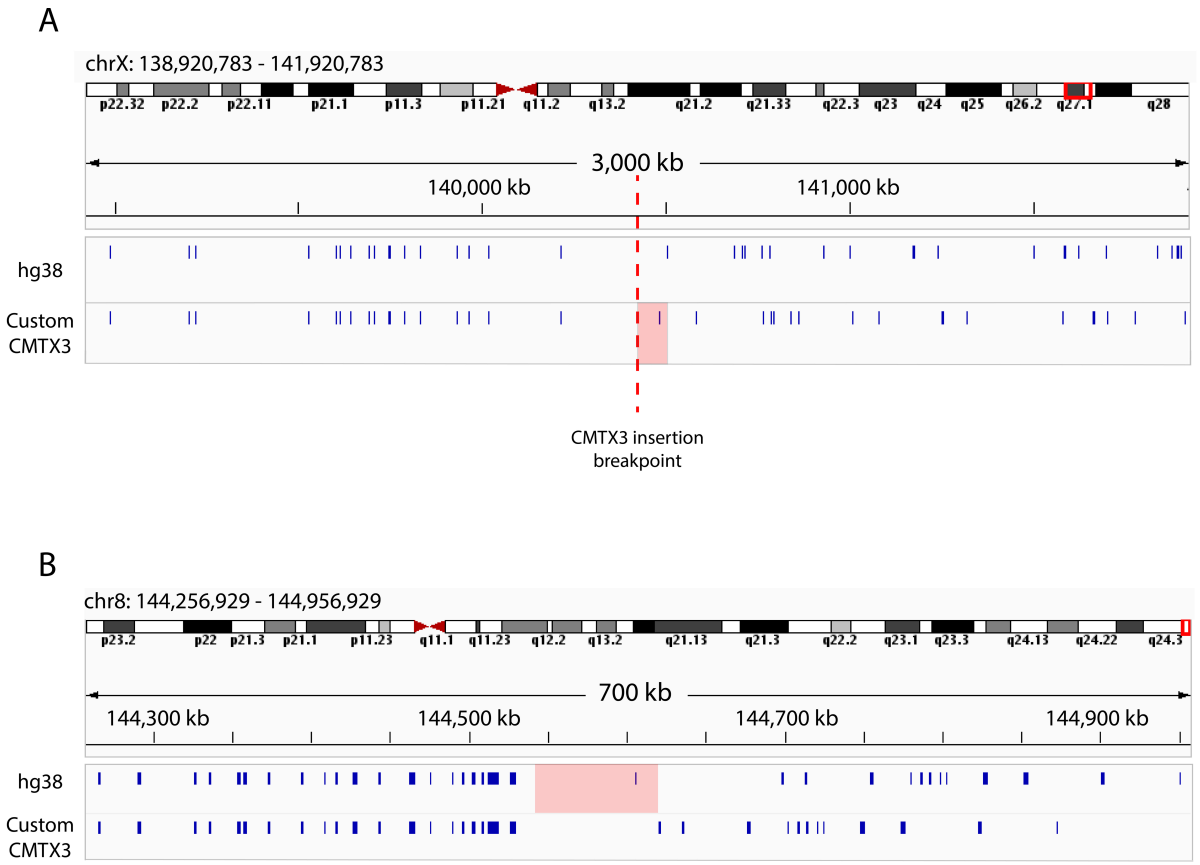
For this study, a Promoter Capture Micro-C approach was utilised, as it offered some key benefits particularly well-suited for the current project. Firstly, Micro-C [235] displays improved data resolution over conventional Hi-C, in which inclusion of the Promoter Capture step further increases capacity to detect fine-scale, promoter-specific interactions [260]. Secondly, this experimental approach was offered as a service by Cantata Bio. This was an important consideration as these sophisticated techniques require specialised equipment and technical expertise beyond the scope of our research laboratory. Lastly, Dovetail Genomics (Cantata Bio) have published detailed analysis documentation with clear instructions for analysing and interpreting these complex datasets, opening this technology up to command-line beginners. Despite these clear benefits, there are some caveats that must be addressed. The genome-wide nature of the promoter capture approach limited the sequencing depth, and therefore resolution, that could be achieved across our specific region of interest. Future studies will benefit from a more targeted approach which can produce higher resolution data focusing exclusively on the CMTX3 locus, such as the recently developed method of Region Capture Micro-C [266]. This approach enriches Micro-C libraries for target genomic loci using capture probes that are tiled across the region, providing impressively high-resolution insights into fine-scale chromatin interactions at the targeted locus [266].

In summary, this chapter represents a preliminary investigation into the effects of the CMTX3 SV upon local promoter-specific chromatin interactions. Due to the lack of biological replicates, and the limited data resolution across the region of interest, it is unclear whether the CMTX3 SV leads to the formation of ectopic chromatin interactions, or whether it may disturb existing chromatin interactions. It is expected that the alternative methodologies proposed above will provide valuable insights into these biological questions. However, any chromatin interactions identified from such analyses will require functional evaluation before the impact upon local gene regulation can be inferred. For example, public datasets can be queried to assess whether the interacting fragments overlap with predicted CREs or eQTLs [264, 265, 267]. Further insights can be obtained through (a) ATAC-Seq, to assess chromatin accessibility, (b) ChIP-Seq, to probe for chromatin markers associated with regulatory features, and (c) reporter assays, to test the impact of candidate sequences upon gene expression levels [107, 207].

Supplementary Material: Chapter 5

Supplementary Table 5.1: **GRCh38/hg38 CMTX3 SV coordinates.** Coordinates were converted from hg19 to hg38 using the UCSC LiftOver tool.

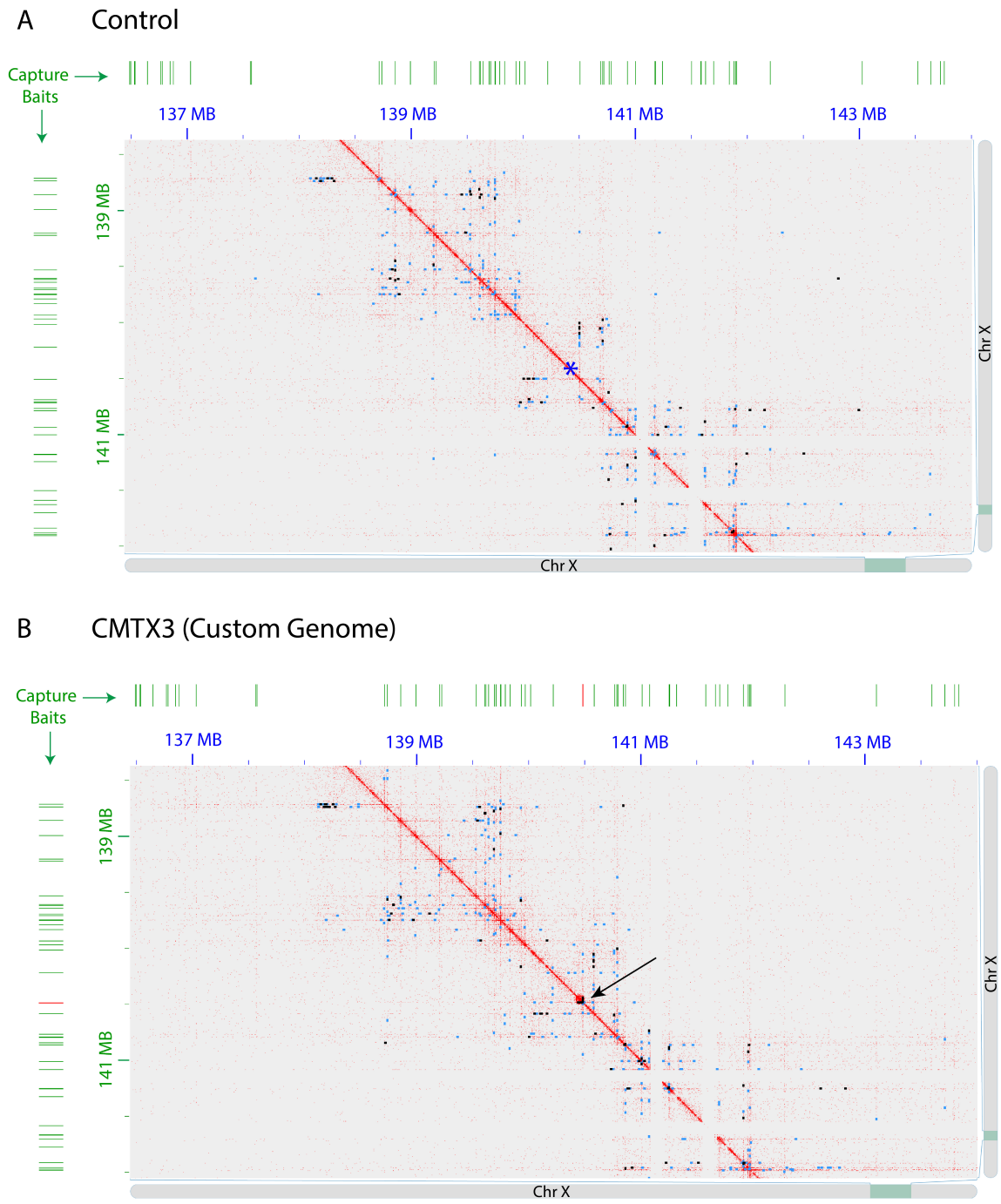
Feature description	hg19 coordinates	hg38 coordinates
CMTX3 SV breakpoint at Xq27.1	chrX:139,502,948–139,502,949	chrX:140,420,783-140,420,784
Inserted sequence (~ 78 kb) from 8q24.3	chr8:145,768,312-145,846,158	chr8:144,542,928-144,620,773
Insertion (9 bp) from 12q13.12	chr12:50,297,649-50,297,657	chr12:49,903,866-49,903,874
Inversion (12 bp) at Xq27.1	chrX:139,502,949-139,502,960	chrX:140,420,784-140,420,795
Single nucleotide variant at Xq27.1	chrX:139,502,968	chrX:140,420,803
Single nucleotide deletion at Xq27.1	chrX:139,502,976	chrX:140,420,811



Supplementary Figure 5.1: Modifications to the location of bait sequences ensures alignment with the custom-built CMTX3 reference genome. (A) Location of the bait sequences used in the Pan Promoter Enrichment Panel, within a 3 Mb region of the CMTX3 locus. Top panel shows bait location for hg38. Bottom panel shows bait location for the custom-built CMTX3 reference genome. At the breakpoint of the CMTX3 SV (red dotted line), CMTX3 patients show a gain of ~78 kb of DNA (red shading), which includes the sequence for one capture bait. The coordinates of baits located distal to the breakpoint were shifted (telomeric) by ~78 kb to account for the genomic rearrangement. (B) Location of bait sequences within a 700 kb region of chromosome 8q24.3. Top panel shows bait location for hg38. The ~78 kb region that is duplicated in CMTX3 is highlighted (red shading). Bottom panel shows bait location for the custom-built CMTX3 reference genome, where the ~78 kb fragment has been deleted from 8q24.3. Therefore, the coordinates of all baits located distal to the deletion have been shifted (centromeric) to align with the custom CMTX3 reference genome. Data visualised using IGV.

Supplementary Table 5.2: Quality control of Promoter Capture Micro-C Libraries.

	Control Sample		Patient sample *(aligned to custom CMTX3 genome)*		Patient sample *(aligned to hg38)*	
	# Read Pairs	Percentage	# Read Pairs	Percentage	# Read Pairs	Percentage
Total Read Pairs	257,368,931	100%	212,593,670	100%	212,593,670	100%
Unmapped Read Pairs	15,265,283	5.93%	11,747,186	5.53%	11,747,186	5.53%
Mapped Read Pairs	192,826,700	74.92%	161,256,274	75.85%	161,256,283	75.85%
PCR Dup Read Pairs	32,194,258	12.51%	25,585,614	12.03%	25,585,614	12.03%
No-Dup Read Pairs	160,632,442	62.41%	135,670,660	63.82%	135,670,669	63.82%
No-Dup Cis Read Pairs	83,367,596	51.90%	77,618,966	57.21%	77,620,853	57.21%
No-Dup Trans Read Pairs	77,264,846	48.10%	58,051,694	42.79%	58,049,816	42.79%
No-Dup Valid Read Pairs (Cis >= 1kb + Trans)	140,433,038	87.43%	121,751,658	89.74%	121,751,662	89.74%
No-Dup Cis Read Pairs < 1kb	20,199,404	12.57%	13,919,002	10.26%	13,919,007	10.26%
No-Dup Cis Read Pairs >= 1kb	63,168,192	39.32%	63,699,964	46.95%	63,701,846	46.95%
No-Dup Cis Read Pairs >= 10kb	48,300,721	30.07%	50,782,628	37.43%	50,784,492	37.43%



Supplementary Figure 5.2: Summary of the chromatin interactions identified by CHiCAGO at chromosome Xq27.1. (A) Raw contact matrix for the control sample aligned to hg38. Blue asterisk highlights the region on chromosome Xq27.1 where the CMTX3 breakpoint occurs. (B) Raw contact matrix for the patient sample, aligned to the custom-built CMTX3 reference genome. Black arrow depicts the CMTX3 SV. For both heatmaps, black dots represent significant chromatin interactions identified by CHiCAGO analysis (CHiCAGO score ≥ 5), and blue dots indicate 'sub-threshold' interactions (CHiCAGO score ≥ 3 but < 5). Interactions were called using 10 kb bin sizes. Location of bait sequences used in the Pan Promoter Enrichment Panel is shown with green bars while the single bait contained within the CMTX3 insertion is marked in red. Heatmaps were visualised with Juicebox software using a resolution of 5 kb.

CHAPTER 6

General Discussion

Despite the many advancements in genomic sequencing technologies, challenges remain with the discovery of novel pathogenic mutations underlying Mendelian disease, with many patients still lacking a genetic diagnosis [268]. This is particularly pertinent to the CMT field. More than 1,000 mutations across 100+ genes have been reported to cause CMT and related neuropathies [25-27], however it is estimated that ~40% of CMT patients remain genetically undiagnosed [46]. SV is a class of mutation that is suggested to contribute to this missing heritability in inherited peripheral neuropathies [66, 269]. The burden of detecting these complex mutations has been greatly reduced with developments in long-range sequencing, 3D genomic technologies, a completed reference genome (T2T-CHM13) [49], and the development of the Human Pangenome to provide genetic diversity [50]. Once a causative SV is identified however, the real challenge lies in interpreting the functional consequences of these complex mutations. This is exemplified in the case of CMTX3 neuropathy, caused by a ~78 kb interchromosomal insertion [63]. Although the causative mutation was reported in 2016, without any clear effect on coding genes, and the lack of an appropriate model for studying this disease, the pathomechanism has remained unsolved. Understanding the molecular mechanisms of the CMTX3 SV will be essential for guiding the development of therapeutic strategies for this debilitating condition.

iPSC-based models are an attractive tool for investigating the CMTX3 pathogenic SV, enabling access to disease-relevant cellular material harbouring the patient's complex genetic mutation.

Accordingly, several recent publications have leveraged iPSC technology to elucidate the functional consequences of complex SV [67, 107, 270]. iPSC strategies have also been increasingly popular within the IPN field to overcome the traditional barriers of accessing nerve tissue and for generating preclinical models for therapy development. An important outcome of this thesis was the generation of a neuronal model of CMTX3 neuropathy, using patient-derived iPSC to culture sMN. Access to disease-relevant tissue harbouring the complex SV was critical for testing potential pathogenic mechanisms of this mutation. Chapter 3 demonstrated that CMTX3 patient iPSC can successfully be differentiated into mature sMN, that display a comparable differentiation efficiency to control iPSC lines. The observed similarity between patient and control sMN cultures is reassuring, as differences in differentiation propensities can complicate downstream transcriptomic analyses, and potentially mask biologically meaningful differences [136]. Whilst a degree of variability between iPSC lines is unavoidable, this can be substantially reduced by the use of isogenic control lines [135]. It is possible to engineer large deletions using CRISPR/Cas9 technology [101, 137], which could be applied to correct the pathogenic CMTX3 SV. The future generation of an isogenic CMTX3 control line would provide a valuable resource to complement the CMTX3 neuronal model generated in this thesis.

6.1 *SOX3* as a candidate gene for CMTX3 neuropathy

Based on previous findings [63], transcriptional dysregulation was identified as a plausible pathomechanism of CMTX3 neuropathy. However, other disease mechanisms also required careful consideration, particularly with the recent discovery that the DHMN1 SV results in the formation of a novel gene-intergenic fusion transcript [67]. A crucial aspect of this project was elucidating the transcriptional consequences of the CMTX3 SV. Through a combination of global and targeted transcriptomic analyses, spatiotemporal dysregulation of the *SOX3* gene was identified in CMTX3 patient samples. In parallel, RNA-Seq and NanoString experiments independently excluded the hypothesis that the CMTX3 SV gives rise to an aberrant transcript. Taken together, this project provided additional evidence to the initial suggestion by Brewer et al. (2016) that gene dysregulation

may underly CMTX3 neuropathy [63]. Importantly, this study has led to the identification of *SOX3* as a novel, high-priority candidate gene for CMTX3.

SOX3 mutations have been shown to cause a range of disease phenotypes, although none have been associated with IPN. Mutations involving *SOX3* have been reported in X-linked hypopituitarism patients with variable severities, from isolated growth hormone deficiency to panhypopituitarism [271]. This includes duplications of *SOX3* [271-273], where patients may also present with a range of additional phenotypes such as intellectual disability [271], neural tube defects [271], or structural brain abnormalities [271, 272]. Both expansions [272, 274, 275] and deletions [276, 277] within the polyalanine tract of *SOX3* have also been reported in X-linked hypopituitarism. *In vitro* assays demonstrated that these mutations alter the transcriptional activation activity of SOX3 [272, 276], with the expansion causing impaired SOX3 nuclear localisation [272]. Whilst rare, pathogenic *SOX3* point mutations have been observed in patients with hypopituitarism of varying severities, and are often associated with additional phenotypic features [278]. For example, Jelsig and colleagues (2018) identified a missense mutation in a patient with a complex phenotype including hypopituitarism as well as microcephaly, eye abnormalities, facial dysmorphology, and dental anomalies [278]. Genomic rearrangements that affect *SOX3* regulation have been associated with XX-male sex reversal [151, 153]. This includes *SOX3* duplications [153, 154], a deletion upstream of *SOX3* [153], and an inter-chromosomal insertion occurring downstream of *SOX3* [151]. Taken together, *SOX3* mutations are implicated in a range of disease phenotypes, although to date, have not been linked to peripheral neuropathy. Whilst the current study has identified *SOX3* as a promising candidate for CMTX3, future experimental work is necessary to explore the functional consequences of *SOX3* dysregulation, and whether this is directly responsible for the neuropathy observed in patients.

6.2 Investigating the functional impact of *SOX3* dysregulation

SOX3 encodes a transcription factor with important roles in the developing nervous system. Although we have not provided evidence for a causal role of *SOX3* dysregulation in CMTX3 neuropathy, there is a precedent for investigating the contribution of transcription factors in the context of CMT. For example, mutations in *EGR2*, which encodes a transcription factor that regulates the expression of key myelination genes such as *PMP22*, *MPZ* and *GJB1* [279], are a known cause of both demyelinating [280-282] and axonal [283, 284] CMT. Therefore, one possibility is that the *SOX3* dysregulation observed in CMTX3 may lead to transcriptional defects during nervous system development. To explore the impact of *SOX3* dysregulation on the transcriptome, we performed RNA-Seq on NEP from CMTX3 patients and controls. This identified a small number of genes (22) showing differential expression between the two cohorts (Chapter 4, Section 4.3.4). None of the identified DEG are associated with IPN, according to the Inherited Neuropathy Variant Browser (https://neuropathybrowser.zuchnerlab.net/#/). Furthermore, functional enrichment analysis did not identify any biological pathways or processes that were significantly enriched among the DEG. Thus, the RNA-Seq analysis has not provided an overt biological link between *SOX3* dysregulation and IPN, and further investigation of the identified DEG is required. This may include performing ChIP-Seq for SOX3, to confirm whether these genes are direct targets of this transcription factor.

RNA-Seq has previously been used to characterise the functional consequences that arise from the dysregulation of developmental transcription factors. This is clearly illustrated in a recent study by Laugsch and colleagues (2019), which explores the dysregulation of *TFAP2A*, encoding an important regulator of neural crest differentiation, in a patient with branchio-oculo-facial syndrome (BOFS) [107]. Interestingly, similar to CMTX3, the causative mutation is a complex SV (~89 Mb inversion), and leads to reduced expression of *TFAP2A* in BOFS patient-derived neural crest cells [107]. RNA-Seq identified genes showing differential expression between patient and control neural crest cells. Interestingly, genes that were downregulated in BOFS cells showed enrichment for genes associated with typical BOFS phenotypes, and functional enrichment analysis identified biologically relevant cellular pathways

being impacted in this disease [107]. Furthermore, ChIP-Seq was used to demonstrate that the downregulated genes are likely direct targets of the TFAP2A transcription factor, and that BOFS patient cells display reduced TFAP2A binding at these sites [107]. Therefore, in this example, RNA-Seq provided key insights into the transcriptional defects that arise from the dysregulation of the TFAP2A transcription factor and how this may contribute to the BOFS phenotype. It is important to note in this study, that the neural crest cells were shown to be a relatively pure, homogenous cell population by scRNA-Seq analysis [107]. scRNA-Seq is a powerful technology that provides transcriptomic data at the individual cell level and can be used to identify heterogeneity within cell cultures and distinguish discrete cell populations [136]. It is worth considering whether intra-culture heterogeneity within the caudal NEP used in the present study may have hindered the capacity to detect differentially expressed genes. A known limitation of iPSC differentiation protocols is that they are not 100% efficient and can result in heterogenous cell populations [285, 286]. The future use of scRNA-Seq to assess the heterogeneity within the caudal NEP would provide valuable insights into the compositions of these cultures. Further, given the role of *SOX3* in maintaining neural progenitor identity and preventing premature differentiation [177], it would be interesting to assess whether *SOX3* dysregulation may impact the proportions of cellular subtypes that are generated during the process of neuronal differentiation.

CMTX3 has been reported as an axonal CMT [53, 54]. In another study, CMTX3 was classified as an intermediate CMT, due to both axonal and demyelinating MNCVs presenting within the one family [52]. However, a recent study on children from an Australian CMTX3 family suggested this may be a primary demyelinating neuropathy, although an axonal neuropathy was observed in one individual [51]. Therefore, assessing the spatiotemporal regulation of SOX3 within patient-derived Schwann cells will be important for further investigating the pathomechansim of the CMTX3 SV, particularly given that *SOX3* dysregulation was observed in the iPSC (Chapter 4, Figure 4.4, 4.5), preceding the specification of neuronal or Schwann cell lineages. Obtaining Schwann cells from iPSC has proven to be a challenging task. Although published differentiation protocols are available, they

often result in limited purity, lack of Schwann cell maturity, and require extended culturing time periods [287]. For example, in an iPSC-derived Schwann cell model of CMT1A developed by Shi et al., (2018), the generation of Schwann cells took ~ 3 months and required fluorescence-activated cell sorting to purify the neural crest cell cultures [116]. Recently published protocols, however, can overcome some of these shortcomings. Kim and colleagues (2017) report the successful differentiation of (immature) Schwann cells from iPSC in ~ 4 weeks, reducing the timeline required to generate this cell type [109]. Another protocol developed by Majd and colleagues (2023) produced Schwann cells with gene expression patterns similar to adult Schwann cells that are capable of *in vitro* and *in vivo* myelination [287]. Recently, Van Lent and colleagues (2023) developed an iPSC-derived organoid system for modelling the PNS, which may also be considered for modelling CMTX3 neuropathy [110]. Modelling organoids for CMTX3 would provide an experimental system containing the major cell types of the PNS, including myelinating Schwann cells [110]. This would offer an advantage over monolayer culture systems as complex organisation and interactions between cell types within the nervous system can be mimicked [110]. As well as providing a platform for phenotypic testing, such as the detection of myelination deficits, the PNS organoids may provide a valuable tool for determining whether *SOX3* dysregulation affects PNS development. scRNA-Seq or immunofluorescence experiments can be used to assess whether CMTX3 patient and control organoids differ in cellular composition, which may indicate inherent defects in differentiation. For example, dorsal forebrain organoids derived from a patient with the neurodevelopmental disorder, Rett syndrome, appear to show premature neuronal differentiation and reduction of the neural progenitor pool [288].

In addition, an *in vivo* system would offer valuable insights into how *SOX3* dysregulation may impact nerve development and function within an intact nervous system. Zebrafish (*Danio rerio*) represent a suitable choice for modelling the CMTX3 *SOX3* dysregulation. Firstly, an orthologue of *SOX3* is present in the zebrafish, and unlike various other model organisms such as *C.* elegans and Drosophila, zebrafish contain myelinated nerves. Further, zebrafish show a relatively simple nervous system, and optically transparent embryos, which enables visualisation of nervous system development [289, 290].

Accordingly, zebrafish models have previously been generated to investigate pathogenic mutations causing various forms of CMT (e.g. [93-96]). Achieving the precise spatiotemporal *SOX3* dysregulation observed in CMTX3 neuropathy will likely be challenging. Testing *SOX3* expression at additional time points and cell types will help inform a spatiotemporal profile mimicking the CMTX3 cellular phenotype. A number of approaches may be considered for modulating *SOX3* expression in zebrafish. For example, transient gene knockdown can be achieved using morpholinos, which are antisense oligonucleotides designed against a target gene. With this approach, the possibility of non-specific and off target effects associated with this technology would need to be assessed and appropriate controls included [291]. Alternatively, the auxin-inducible degron system [292] is gaining popularity for performing conditional knockdown experiments within a range of model organisms [293-295]. This system can be used to achieve efficient and rapid depletion of a target protein by tagging it for degradation by the proteasome [292]. To implement the auxin-inducible degron system, transgenic cells/animals must be generated so that they co-express the protein of interest fused with an 'auxin-inducible degron', and the plant E3 ligase adapter, TIR1 [292, 296]. When cells/animals are exposed to the plant hormone, auxin, the fused protein of interest is tagged with ubiquitin and subsequently degraded by the proteosome [292]. This degradation is reversible and can be modulated by altering the concentration of auxin [292]. Therefore, this system represents an ideal approach for mimicking the *SOX3* dysregulation observed in CMTX3 neuropathy, which is restricted to a discrete temporal window, and does not involve the complete loss of *SOX3* expression.

6.3 Investigating the mechanism of gene dysregulation in CMTX3

In chapter 5 of this thesis, we aimed to investigate how the CMTX3 SV causes the gene dysregulation observed in NEP from CMTX3 patients. We hypothesised that this may involve the rewiring of regulatory chromatin interactions, as this is a well-documented pathomechanism of complex SV (reviewed in [68, 149]). A preliminary Promoter Capture Micro-C experiment was

performed to assess the promoter-specific chromatin interactions across the CMTX3 locus. Of particular interest was exploring whether the patient sample might display ectopic interactions between the CMTX3 SV and nearby DNA from within the CMTX3 linkage region. Analysis of the Micro-C datasets did not detect statistically significant chromatin contacts involving the CMTX3 SV and the CMTX3 linkage region. However, there were several 'sub-threshold' interactions detected between the *ARHGAP39* promoter, contained within the CMTX3 insertion, and nearby sequence from Xq27.1. Ideally, data from additional biological replicates and a significantly increased sequencing depth of coverage will be needed to determine whether these are reproducible interactions. This hypothesis however, should not be discounted. Due to the lack of biological replicates, testing if the CMTX3 SV may disrupt existing chromatin interactions within the CMTX3 locus could not be adequately addressed. A benefit of the Promoter Capture approach utilised in this exploratory study, was the overview provided for promoter-specific chromatin interactions across the entire CMTX3 linkage region, thereby allowing visualisation of the CMTX3 SV interactions in the context of the broader regulatory landscape. However, given the identification of *SOX3* as a high priority candidate gene for CMTX3, future studies may benefit from a more targeted approach to deeply characterise chromatin interactions involving this gene of interest. In this respect, 4C seq represents a valuable tool as it would provide high resolution interaction data for selected genomic sites [220]. This technology has been instrumental in defining how pathogenic SV can rewire chromatin interactions (e.g. [102, 107, 214, 215]).

6.4 The importance of peripheral nerve development in CMT

SOX3 encodes a developmental transcription factor and was dysregulated during early developmental cell types in CMTX3 patient-derived cells. This raises the question of whether the underlying cause of CMTX3 neuropathy could arise from abnormal peripheral nerve development. With this in mind, it is interesting that CMTX3 can present as an early onset neuropathy [51]. Clinical examination of children with CMTX3 revealed that male patients show an age of onset ranging from

birth to 5 years, with most children displaying symptoms within the first 2 years of life [51]. Developmental defects have been reported in other forms of CMT. For example, a CMT1A rat model shows perturbed Schwann cell differentiation that is evident from the early postnatal period, causing a reduction in the number of myelinated axons when compared to wild-type controls [297]. Interestingly, defective Schwann cell differentiation was also observed in an iPSC model of CMT1A [116]. CMT1A and control iPSC showed similar capacity to differentiate into neural crest cells. However, subsequent differentiation toward the Schwann cell lineage was strikingly impaired in CMT1A cells, which preferentially differentiated into endoneurial fibroblasts rather than Schwann cells [116]. The authors therefore suggested that the PMP22 dosage effect in CMT1A may cause a developmental switch during Schwann cell differentiation [116]. Taken together, there is strong evidence to suggest that developmental abnormalities can give rise to CMT neuropathy. CMTX3 neuropathy may provide a useful paradigm for exploring this possibility further.

6.5 Conclusion

Interpreting the pathogenic mechanism of SV remains a challenging task in Mendelian diseases such as CMT. This complex class of mutation can cause disease via a range of different mechanisms, including altered gene dosage, local transcriptional dysregulation, and the formation of aberrant transcripts [68]. Accordingly, the pathomechanism for many reported complex SV have remained unsolved. In Chapter 1, Part B, an experimental pipeline was proposed in which multiple experimental strategies could be used to investigate how the ~78 kb CMTX3 interchromosomal insertion causes disease. For this thesis, the pipeline was systematically applied and has led to the generation of important data informing further insights into the functional consequences of this complex mutation. Through the generation of a disease-relevant model for CMTX3, we have discovered altered regulation of *SOX3* during the differentiation of CMTX3-derived iPSC into sMN, which is now considered a high-priority candidate gene for future studies. This finding provides further evidence to suggest that gene dysregulation may be the underlying disease mechanism for

CMTX3 neuropathy. At the same time, we did not observe evidence to support alternative pathomechanisms, such as the formation of novel fusion transcripts or splice variants. Taken together, this body of work represents major progress towards elucidating the pathomechanism of CMTX3 neuropathy and has generated an invaluable pre-clinical model that will continue to inform disease biology and guide future development of treatment therapies.

REFERENCES

1. Charcot, J.M. and Marie, P., *Sur une forme paticuliere d'atrophie musuculaire progressive souvent familial, debutante par les pieds et les jambes et atteignant plus tard les mains.* Revue de Médecine, 1886. **6**: p. 97-138.
2. Tooth, H., *The peroneal type of progressive muscular atrophy*. 1886, University of Cambridge.
3. Skre, H., *Genetic and clinical aspects of Charcot-Marie-Tooth's disease.* Clinical Genetics, 1974. **6**(2): p. 98-118.
4. Barisic, N., et al., *Charcot-Marie-Tooth disease: a clinico-genetic confrontation.* Annals of Human Genetics, 2008. **72**(3): p. 416-441.
5. Reilly, M.M., Murphy, S.M., and Laurá, M., *Charcot-Marie-Tooth disease.* Journal of the Peripheral Nervous System, 2011. **16**(1): p. 1-14.
6. Szigeti, K. and Lupski, J.R., *Charcot–Marie–Tooth disease.* European Journal of Human Genetics, 2009. **17**(6): p. 703-710.
7. Sommer, C. and Schröder, J.M., *Hereditary motor and sensory neuropathy with optic atrophy: ultrastructural and morphometric observations on nerve fibers, mitochondria, and dense-cored vesicles.* Archives of Neurology, 1989. **46**(9): p. 973-977.
8. Züchner, S., et al., *Axonal neuropathy with optic atrophy is caused by mutations in mitofusin 2.* Annals of Neurology, 2006. **59**(2): p. 276-281.
9. Stojkovic, T., et al., *Sensorineural deafness in X-linked Charcot-Marie-Tooth disease with connexin 32 mutation (R142Q).* Neurology, 1999. **52**(5): p. 1010-1010.
10. Papadakis, C.E., et al., *Bilateral sudden sensorineural hearing loss caused by Charcot-Marie-Tooth disease.* The Journal of Laryngology & Otology, 2003. **117**(5): p. 399-401.

11. Cowchock, F.S., et al., *X-linked motor-sensory neuropathy type-II with deafness and mental retardation: a new disorder.* American Journal of Medical Genetics, 1985. **20**(2): p. 307-315.

12. Ando, M., et al., *SLC12A6 heterozygous variants identified in patients of Charcot-Marie-Tooth disease with cognitive impairment.* Journal of the Neurological Sciences, 2017. **381**: p. 378.

13. Sevilla, T., et al., *Vocal cord paresis and diaphragmatic dysfunction are severe and frequent symptoms of GDAP1-associated neuropathy.* Brain, 2008. **131**(11): p. 3051-3061.

14. Pareyson, D. and Marchesi, C., *Diagnosis, natural history, and management of Charcot–Marie–Tooth disease.* The Lancet Neurology, 2009. **8**(7): p. 654-667.

15. Grandis, M. and Shy, M.E., *Current therapy for Charcot-Marie-Tooth disease.* Current Treatment Options in Neurology, 2005. **7**(1): p. 23-31.

16. Phillips, M., Radford, K., and Wills, A., *Ankle foot orthoses for people with Charcot Marie Tooth disease–views of users and orthotists on important aspects of use.* Disability and Rehabilitation: Assistive Technology, 2011. **6**(6): p. 491-499.

17. Dyck, P.J. and Lambert, E.H., *Lower motor and primary sensory neuron diseases with peroneal muscular atrophy: I. Neurologic, genetic, and electrophysiologic findings in hereditary polyneuropathies.* Archives of Neurology, 1968. **18**(6): p. 603-618.

18. Dyck, P.J. and Lambert, E.H., *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.

19. Juneja, M., et al., *Challenges in modelling the Charcot-Marie-Tooth neuropathies for therapy development.* Journal of Neurology, Neurosurgery & Psychiatry, 2019. **90**(1): p. 58-67.

20. Harding, A. and Thomas, P., *The clinical features of hereditary motor and sensory neuropathy types I and II.* Brain, 1980. **103**(2): p. 259-280.

21. Davis, C., Bradley, W., and Madrid, R., *The peroneal muscular atrophy syndrome: clinical, genetic, electrophysiological and nerve biopsy studies. I. Clinical, genetic and*

electrophysiological findings and classification. Journal de Genetique Humaine, 1978. **26**(4): p. 311-349.

22. Madrid, R., Bradley, W., and Davis, C., *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.
23. Magy, L., et al., *Updating the classification of inherited neuropathies: results of an international survey.* Neurology, 2018. **90**(10): p. e870-e876.
24. Mathis, S., et al., *Charcot–Marie–Tooth diseases: an update and some new proposals for the classification.* Journal of Medical Genetics, 2015. **52**(10): p. 681-690.
25. Timmerman, V., Strickland, A.V., and Züchner, S., *Genetics of Charcot-Marie-Tooth (CMT) disease within the frame of the human genome project success.* Genes, 2014. **5**(1): p. 13-32.
26. Pipis, M., et al., *Next-generation sequencing in Charcot–Marie–Tooth disease: opportunities and challenges.* Nature Reviews Neurology, 2019. **15**(11): p. 644-656.
27. Rossor, A.M., Tomaselli, P.J., and Reilly, M.M., *Recent advances in the genetic neuropathies.* Current Opinion in Neurology, 2016. **29**(5): p. 537.
28. Antonellis, A., et al., *Glycyl tRNA synthetase mutations in Charcot-Marie-Tooth disease type 2D and distal spinal muscular atrophy type V.* The American Journal of Human Genetics, 2003. **72**(5): p. 1293-1299.
29. Jordanova, A., et al., *Disrupted function and axonal distribution of mutant tyrosyl-tRNA synthetase in dominant intermediate Charcot-Marie-Tooth neuropathy.* Nature Genetics, 2006. **38**(2): p. 197-202.
30. Latour, P., et al., *A major determinant for binding and aminoacylation of tRNAAla in cytoplasmic alanyl-tRNA synthetase is mutated in dominant axonal Charcot-Marie-Tooth disease.* The American Journal of Human Genetics, 2010. **86**(1): p. 77-82.

31. Gonzalez, M., et al., *Exome sequencing identifies a significant variant in methionyl-tRNA synthetase (MARS) in a family with late-onset CMT2.* Journal of Neurology, Neurosurgery & Psychiatry, 2013. **84**(11): p. 1247-1249.

32. McLaughlin, H.M., et al., *Compound heterozygosity for loss-of-function lysyl-tRNA synthetase mutations in a patient with peripheral neuropathy.* The American Journal of Human Genetics, 2010. **87**(4): p. 560-566.

33. Safka Brozkova, D., et al., *Loss of function mutations in HARS cause a spectrum of inherited peripheral neuropathies.* Brain, 2015. **138**(8): p. 2161-2172.

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

35. Lupski, J.R., et al., *DNA duplication associated with Charcot-Marie-Tooth disease type 1A.* Cell, 1991. **66**(2): p. 219-232.

36. Patel, P.I., et al., *The gene for the peripheral myelin protein PMP–22 is a candidate for Charcot–Marie–Tooth disease type 1A.* Nature Genetics, 1992. **1**(3): p. 159-165.

37. Matsunami, N., et al., *Peripheral myelin protein–22 gene maps in the duplication in chromosome 17p11. 2 associated with Charcot–Marie–Tooth 1A.* Nature Genetics, 1992. **1**(3): p. 176-179.

38. Timmerman, V., et al., *The peripheral myelin protein gene PMP–22 is contained within the Charcot–Marie–Tooth disease type 1A duplication.* Nature Genetics, 1992. **1**(3): p. 171-175.

39. Valentijn, L., et al., *The peripheral myelin gene PMP–22/GAS–3 is duplicated in Charcot–Marie–Tooth disease type 1A.* Nature Genetics, 1992. **1**(3): p. 166-170.

40. Bergoffen, J., et al., *Connexin mutations in X-linked Charcot-Marie-Tooth disease.* Science, 1993. **262**(5142): p. 2039-2042.

41. Kleopa, K.A., *The role of gap junctions in Charcot-Marie-Tooth disease.* Journal of Neuroscience, 2011. **31**(49): p. 17753-17760.

42. Hayasaka, K., et al., *Charcot–Marie–Tooth neuropathy type 1B is associated with mutations of the myelin P0 gene.* Nature Genetics, 1993. **5**(1): p. 31-34.

43. 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.

44. Murphy, S.M., et al., *Charcot–Marie–Tooth disease: frequency of genetic subtypes and guidelines for genetic testing.* Journal of Neurology, Neurosurgery & Psychiatry, 2012. **83**(7): p. 706-710.

45. Saporta, A.S., et al., *Charcot-Marie-Tooth disease subtypes and genetic testing strategies.* Annals of Neurology, 2011. **69**(1): p. 22-33.

46. 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. 873-878.

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

48. Montenegro, G., et al., *Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family.* Annals of Neurology, 2011. **69**(3): p. 464-470.

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

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

51. Kanhangad, M., et al., *Unique clinical and neurophysiologic profile of a cohort of children with CMTX3.* Neurology, 2018. **90**(19): p. e1706-e1710.

52. Huttner, I., et al., *Proof of genetic heterogeneity in X-linked Charcot–Marie–Tooth disease.* Neurology, 2006. **67**(11): p. 2016-2021.

53. Chaudhry, R.K., *Molecular analysis of the Xlinked CMT (CMTX3) disease locus using NGS technologies.* 2014, University of Sydney.

54. Brewer, M.H., et al., *X-linked CMT: genes and gene loci in an Australian cohort.* Neurogenetics, 2010. **11**: p. 267-269.

55. Nicholson, G., et al., *Genotypes & sensory phenotypes in 2 new X-linked neuropathies (CMTX3 and dSMAX) and dominant CMT/HMN overlap syndromes.* Inherited Neuromuscular Diseases: Translation from Pathomechanisms to Therapies, 2009: p. 201-206.

56. Ionasescu, V., et al., *Heterogeneity in X-linked recessive Charcot-Marie-Tooth neuropathy.* American Journal of Human Genetics, 1991. **48**(6): p. 1075.

57. Brewer, M., et al., *Evidence of a founder haplotype refines the X-linked Charcot–Marie-Tooth (CMTX3) locus to a 2.5 Mb region.* Neurogenetics, 2008. **9**: p. 191-195.

58. Henden, L., et al., *Identity-by-descent analysis of CMTX3 links three families through a common founder.* Journal of Human Genetics, 2023. **68**(1): p. 47-49.

59. Brewer, M.H., *Molecular genetics of X-linked Charcot-Marie-Tooth neuropathy (CMTX3)*. 2011, University of Sydney.

60. Chaudhry, R., et al., *Re-analysis of an original CMTX3 family using exome sequencing identifies a known BSCL2 mutation.* Muscle & Nerve, 2013. **47**(6): p. 922-924.

61. Auer-Grumbach, M., et al., *Phenotypes of the N88S Berardinelli–Seip congenital lipodystrophy 2 mutation.* Annals of Neurology, 2005. **57**(3): p. 415-424.

62. Irobi, J., et al., *The phenotype of motor neuropathies associated with BSCL2 mutations is broader than Silver syndrome and distal HMN type V.* Brain, 2004. **127**(9): p. 2124-2130.

63. 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.

64. Boyling, A., Perez-Siles, G., and Kennerson, M.L., *Structural variation at a disease mutation hotspot: Strategies to investigate gene regulation and the 3D genome.* Frontiers in Genetics, 2022. **13**: p. 842860.

65. de Boer, E., et al., *A complex structural variant near SOX3 causes X-linked split-hand/foot malformation.* Human Genetics and Genomics Advances, 2023.

66. Drew, A.P., et al., *A 1.35 Mb DNA fragment is inserted into the DHMN1 locus on chromosome 7q34–q36. 2.* Human Genetics, 2016. **135**: p. 1269-1278.

67. 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.

68. Allou, L. and Mundlos, S., *Disruption of regulatory domains and novel transcripts as disease-causing mechanisms.* BioEssays, 2023. **45**(10): p. 2300010.

69. Du, Z.-W., et al., *Generation and expansion of highly pure motor neuron progenitors from human pluripotent stem cells.* Nature Communications, 2015. **6**(1): p. 6626.

70. Mao, D., et al., *Chemical decontamination of iPS cell-derived neural cell mixtures.* Chemical Communications (Cambridge, England), 2018. **54**(11): p. 1355-1358.

71. Schindelin, J., et al., *Fiji: an open-source platform for biological-image analysis.* Nature Methods, 2012. **9**(7): p. 676-682.

72. NanoString Technologies, I. *Reference Genes for Normalization of Expression Data*. 2009; Available from: https://nanostring.com/wp-content/uploads/TN_Normalization_of_Expression_Data.pdf.

73. Goytain, A. and Ng, T., *NanoString nCounter technology: high-throughput RNA validation.* Chimeric RNA: methods and protocols, 2020: p. 125-139.

74. NanoString Technologies, I. *Gene Expression Data Analysis Guidelines* MAN-C0011-04 2017; Available from: https://nanostring.com/wp-content/uploads/Gene_Expression_Data_Analysis_Guidelines.pdf.

75. NanoString Technologies, I. *NANOSTRING USER MANUAL PlexSet™ Reagents for Gene Expression* MAN-10040-06 2019; Available from: https://nanostring.com/wp-content/uploads/MAN-10040-06_PlexSet_Reagents_User_Manual.pdf.

76. NanoString Technologies, I. *Explore the Biology of Model Organisms with nCounter® Gene Expression Analysis*. 2019; Available from: https://nanostring.com/wp-content/uploads/FL_MK1781_CustomCodeset_ORG_r2_FINAL.pdf.

77. NanoString Technologies, I. *nCounter® Analysis System Grant Application Package*. 2022; Available from: https://nanostring.com/wp-content/uploads/FL_MK2295_nCounter_Grant_Support_R5.pdf.

78. Gabreëls-Festen, A., et al., *Early morphological features in dominantly inherited demyelinating motor and sensory neuropathy (HMSN type I).* Journal of the Neurological Sciences, 1992. **107**(2): p. 145-154.

79. Gabreëls-Festen, A., et al., *Charcot-Marie-Tooth disease type 1A: morphological phenotype of the 17p duplication versus PMP22 point mutations.* Acta Neuropathologica, 1995. **90**: p. 645-649.

80. Robertson, A., et al., *Comparison of a new pmp22 transgenic mouse line with other mouse models and human patients with CMT1A.* Journal of Anatomy, 2002. **200**(4): p. 377-390.

81. Kretzschmar, H., *Brain banking: opportunities, challenges and meaning for the future.* Nature Reviews Neuroscience, 2009. **10**(1): p. 70-78.

82. Masters, C.L., et al., *Amyloid plaque core protein in Alzheimer disease and Down syndrome.* Proceedings of the National Academy of Sciences, 1985. **82**(12): p. 4245-4249.

83. Spillantini, M.G., et al., *α-Synuclein in Lewy bodies.* Nature, 1997. **388**(6645): p. 839-840.

84. DiFiglia, M., et al., *Aggregation of huntingtin in neuronal intranuclear inclusions and dystrophic neurites in brain.* Science, 1997. **277**(5334): p. 1990-1993.

85. Fledrich, R., Stassart, R.M., and Sereda, M.W., *Murine therapeutic models for Charcot-Marie-Tooth (CMT) disease.* British Medical Bulletin, 2012. **102**(1): p. 89-113.

86. Bouhy, D. and Timmerman, V., *Animal models and therapeutic prospects for Charcot–Marie–Tooth disease.* Annals of Neurology, 2013. **74**(3): p. 391-396.

87. Sereda, M., et al., *A transgenic rat model of Charcot-Marie-Tooth disease.* Neuron, 1996. **16**(5): p. 1049-1060.

88. McCray, B.A. and Scherer, S.S., *Axonal Charcot-Marie-Tooth disease: from common pathogenic mechanisms to emerging treatment opportunities.* Neurotherapeutics, 2021. **18**(4): p. 2269-2285.

89. Bouhy, D., et al., *A knock-in/knock-out mouse model of HSPB8-associated distal hereditary motor neuropathy and myopathy reveals toxic gain-of-function of mutant Hspb8.* Acta Neuropathologica, 2018. **135**: p. 131-148.

90. Shackleford, G.G., et al., *A new mouse model of Charcot-Marie-Tooth 2J neuropathy replicates human axonopathy and suggest alteration in axo-glia communication.* PLoS Genetics, 2022. **18**(11): p. e1010477.

91. Storkebaum, E., et al., *Dominant mutations in the tyrosyl-tRNA synthetase gene recapitulate in Drosophila features of human Charcot–Marie–Tooth neuropathy.* Proceedings of the National Academy of Sciences, 2009. **106**(28): p. 11782-11787.

92. Niehues, S., et al., *Impaired protein translation in Drosophila models for Charcot–Marie–Tooth neuropathy caused by mutant tRNA synthetases.* Nature Communications, 2015. **6**(1): p. 7520.

93. Chapman, A.L., et al., *Axonal transport defects in a mitofusin 2 loss of function model of Charcot-Marie-Tooth disease in zebrafish.* PLoS One, 2013. **8**(6): p. e67276.

94. Ponomareva, O.Y., Eliceiri, K.W., and Halloran, M.C., *Charcot-Marie-Tooth 2b associated Rab7 mutations cause axon growth and guidance defects during vertebrate sensory neuron development.* Neural Development, 2016. **11**(1): p. 1-14.

95. Hong, Y.B., et al., *DGAT2 Mutation in a Family with Autosomal-Dominant Early-Onset Axonal Charcot-Marie-Tooth Disease.* Human Mutation, 2016. **37**(5): p. 473-480.

96. Vettori, A., et al., *Developmental defects and neuromuscular alterations due to mitofusin 2 gene (MFN2) silencing in zebrafish: a new model for Charcot-Marie-Tooth type 2A neuropathy.* Neuromuscular Disorders, 2011. **21**(1): p. 58-67.

97. Soh, M.S., et al., *Disruption of genes associated with Charcot-Marie-Tooth type 2 lead to common behavioural, cellular and molecular defects in Caenorhabditis elegans.* PLoS One, 2020. **15**(4): p. e0231600.

98. Narayanan, R.K., et al., *Charcot–Marie–Tooth disease causing mutation (p. R158H) in pyruvate dehydrogenase kinase 3 (PDK3) affects synaptic transmission, ATP production and causes neurodegeneration in a CMTX6 C. elegans model.* Human Molecular Genetics, 2022. **31**(1): p. 133-145.

99. Johnson, D. and Nehrke, K., *Mitochondrial fragmentation leads to intracellular acidification in Caenorhabditis elegans and mammalian cells.* Molecular Biology of the Cell, 2010. **21**(13): p. 2191-2201.

100. Chung, K.W., Kim, J.S., and Lee, K.S., *A database of Caenorhabditis elegans locomotion and body posture phenotypes for the peripheral neuropathy model.* Molecules and Cells, 2020. **43**(10): p. 880.

101. Kraft, K., et al., *Deletions, inversions, duplications: engineering of structural variants using CRISPR/Cas in mice.* Cell Reports, 2015. **10**(5): p. 833-839.

102. Lupiáñez, D.G., et al., *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.* Cell, 2015. **161**(5): p. 1012-1025.

103. Kragesteen, B.K., et al., *Dynamic 3D chromatin architecture contributes to enhancer specificity and limb morphogenesis.* Nature Genetics, 2018. **50**(10): p. 1463-1473.

104. Sánchez-Gaya, V., Mariner-Faulí, M., and Rada-Iglesias, A., *Rare or overlooked? Structural disruption of regulatory domains in human neurocristopathies.* Frontiers in Genetics, 2020. **11**: p. 688.

105. Yarnall, M.T., et al., *Drag-and-drop genome insertion of large sequences without double-strand DNA cleavage using CRISPR-directed integrases.* Nature Biotechnology, 2023. **41**(4): p. 500-512.

106. Takahashi, K. and Yamanaka, S., *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.* Cell, 2006. **126**(4): p. 663-676.

107. Laugsch, M., et al., *Modeling the pathological long-range regulatory effects of human structural variation with patient-specific hiPSCs.* Cell Stem Cell, 2019. **24**(5): p. 736-752. e12.

108. Chambers, S.M., et al., *Combined small-molecule inhibition accelerates developmental timing and converts human pluripotent stem cells into nociceptors.* Nature Biotechnology, 2012. **30**(7): p. 715-720.

109. Kim, H.-S., et al., *Schwann cell precursors from human pluripotent stem cells as a potential therapeutic target for myelin repair.* Stem Cell Reports, 2017. **8**(6): p. 1714-1726.

110. Van Lent, J., et al., *Downregulation of PMP22 ameliorates myelin defects in iPSC-derived human organoid cultures of CMT1A.* Brain, 2023. **146**(7): p. 2885-2896.

111. Perez-Siles, G., et al., *Modelling the pathogenesis of X-linked distal hereditary motor neuropathy using patient-derived iPSCs.* Disease Models & Mechanisms, 2020. **13**(2): p. dmm041541.

112. Perez-Siles, G., et al., *Energy metabolism and mitochondrial defects in X-linked Charcot-Marie-Tooth (CMTX6) iPSC-derived motor neurons with the p. R158H PDK3 mutation.* Scientific Reports, 2020. **10**(1): p. 9262.

113. Saporta, M.A., et al., *Axonal Charcot–Marie–Tooth disease patient-derived motor neurons demonstrate disease-specific phenotypes including abnormal electrophysiological properties.* Experimental Neurology, 2015. **263**: p. 190-199.

114. Sainio, M.T., et al., *Absence of NEFL in patient-specific neurons in early-onset Charcot-Marie-Tooth neuropathy.* Neurology Genetics, 2018. **4**(3).

115. 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.

116. Shi, L., et al., *Modeling the pathogenesis of Charcot-Marie-Tooth disease type 1A using patient-specific iPSCs.* Stem Cell Reports, 2018. **10**(1): p. 120-133.

117. Feliciano, C.M., et al., *Allele-specific gene editing rescues pathology in a human model of Charcot-Marie-Tooth disease type 2E.* Frontiers in Cell and Developmental Biology, 2021. **9**: p. 723023.

118. Yu, J., et al., *Efficient feeder-free episomal reprogramming with small molecules.* PLoS One, 2011. **6**(3): p. e17557.

119. Yu, J., et al., *Human induced pluripotent stem cells free of vector and transgene sequences.* Science, 2009. **324**(5928): p. 797-801.

120. Okita, K., et al., *A more efficient method to generate integration-free human iPS cells.* Nature Methods, 2011. **8**(5): p. 409-412.

121. Carpenter, A.E., et al., *CellProfiler: image analysis software for identifying and quantifying cell phenotypes.* Genome Biology, 2006. **7**: p. 1-11.

122. Mayshar, Y., et al., *Identification and classification of chromosomal aberrations in human induced pluripotent stem cells.* Cell Stem Cell, 2010. **7**(4): p. 521-531.

123. Hussein, S.M., et al., *Copy number variation and selection during reprogramming to pluripotency.* Nature, 2011. **471**(7336): p. 58-62.

124. Liu, X., et al., *Chromosomal aberration arises during somatic reprogramming to pluripotent stem cells.* Cell Division, 2020. **15**(1): p. 1-13.

125. Lee, S.-K. and Pfaff, S.L., *Transcriptional networks regulating neuronal identity in the developing spinal cord.* Nature Neuroscience, 2001. **4**(Suppl 11): p. 1183-1191.

126. Masahira, N., et al., *Olig2-positive progenitors in the embryonic spinal cord give rise not only to motoneurons and oligodendrocytes, but also to a subset of astrocytes and ependymal cells.* Developmental Biology, 2006. **293**(2): p. 358-369.

127. Arber, S., et al., *Requirement for the homeobox gene Hb9 in the consolidation of motor neuron identity.* Neuron, 1999. **23**(4): p. 659-674.

128. Liang, X., et al., *Isl1 is required for multiple aspects of motor neuron development.* Molecular and Cellular Neuroscience, 2011. **47**(3): p. 215-222.

129. Shimojo, D., et al., *Rapid, efficient, and simple motor neuron differentiation from human pluripotent stem cells.* Molecular Brain, 2015. **8**: p. 1-15.

130. Kyttälä, A., et al., *Genetic variability overrides the impact of parental cell type and determines iPSC differentiation potential.* Stem Cell Reports, 2016. **6**(2): p. 200-212.

131. Mills, J.A., et al., *Clonal genetic and hematopoietic heterogeneity among human-induced pluripotent stem cell lines.* Blood, 2013. **122**(12): p. 2047-2051.

132. Kajiwara, M., et al., *Donor-dependent variations in hepatic differentiation from human-induced pluripotent stem cells.* Proceedings of the National Academy of Sciences, 2012. **109**(31): p. 12538-12543.

133. Hu, B.-Y., et al., *Neural differentiation of human induced pluripotent stem cells follows developmental principles but with variable potency.* Proceedings of the National Academy of Sciences, 2010. **107**(9): p. 4335-4340.

134. Germain, P.-L. and Testa, G., *Taming human genetic variability: transcriptomic meta-analysis guides the experimental design and interpretation of iPSC-based disease modeling.* Stem Cell Reports, 2017. **8**(6): p. 1784-1796.

135. Kim, H.S., et al., *Genomic editing tools to model human diseases with isogenic pluripotent stem cells.* Stem Cells and Development, 2014. **23**(22): p. 2673-2686.

136. Volpato, V. and Webber, C., *Addressing variability in iPSC-derived models of human disease: guidelines to promote reproducibility.* Disease Models & Mechanisms, 2020. **13**(1): p. dmm042317.

137. Canver, M.C., et al., *Characterization of genomic deletion efficiency mediated by clustered regularly interspaced palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells.* Journal of Biological Chemistry, 2014. **289**(31): p. 21312-21324.

138. Rouhani, F., et al., *Genetic background drives transcriptional variation in human induced pluripotent stem cells.* PLoS Genetics, 2014. **10**(6): p. e1004432.

139. Volpato, V., et al., *Reproducibility of molecular phenotypes after long-term differentiation to human iPSC-derived neurons: a multi-site omics study.* Stem Cell Reports, 2018. **11**(4): p. 897-911.

140. Faye, P.-A., et al., *Optimized protocol to generate spinal motor neuron cells from induced pluripotent stem cells from Charcot Marie Tooth patients.* Brain Sciences, 2020. **10**(7): p. 407.

141. Acemel, R.D. and Gómez-Skarmeta, J.L., *Reprogramming nuclear architecture: just a TAD.* Cell Stem Cell, 2019. **24**(5): p. 679-681.

142. Young, G.T., et al., *Characterizing human stem cell–derived sensory neurons at the single-cell level reveals their ion channel expression and utility in pain research.* Molecular Therapy, 2014. **22**(8): p. 1530-1543.

143. Chatron, N., et al., *Severe hemophilia A caused by an unbalanced chromosomal rearrangement identified using nanopore sequencing.* Journal of Thrombosis and Haemostasis, 2019. **17**(7): p. 1097-1103.

144. Waddell, L.B., et al., *WGS and RNA studies diagnose noncoding DMD variants in males with high creatine kinase.* Neurology Genetics, 2021. **7**(1).

145. Gonçalves, A., et al., *Exonization of an intronic LINE-1 element causing Becker muscular dystrophy as a novel mutational mechanism in dystrophin gene.* Genes, 2017. **8**(10): p. 253.

146. Hoogstrate, Y., et al., *Fusion transcripts and their genomic breakpoints in polyadenylated and ribosomal RNA–minus RNA sequencing data.* Gigascience, 2021. **10**(12): p. giab080.

147. Cmero, M., et al., *MINTIE: identifying novel structural and splice variants in transcriptomes using RNA-seq data.* Genome Biology, 2021. **22**(1): p. 1-26.

148. Zhang, J., et al., *A novel linc00308/D21S2088E intergenic region ALK fusion and its enduring clinical responses to crizotinib.* Journal of Thoracic Oncology, 2020. **15**(6): p. 1073-1077.

149. Spielmann, M., Lupiáñez, D.G., and Mundlos, S., *Structural variation in the 3D genome.* Nature Reviews Genetics, 2018. **19**(7): p. 453-467.

150. Ibrahim, D.M. and Mundlos, S., *Three-dimensional chromatin in disease: What holds us together and what drives us apart?* Current Opinion in Cell Biology, 2020. **64**: p. 1-9.

151. Haines, B., et al., *Interchromosomal insertional translocation at Xq26. 3 alters SOX3 expression in an individual with XX male sex reversal.* The Journal of Clinical Endocrinology & Metabolism, 2015. **100**(5): p. E815-E820.

152. DeStefano, G.M., et al., *Position effect on FGF13 associated with X-linked congenital generalized hypertrichosis.* Proceedings of the National Academy of Sciences, 2013. **110**(19): p. 7790-7795.

153. Sutton, E., et al., *Identification of SOX3 as an XX male sex reversal gene in mice and humans.* The Journal of Clinical Investigation, 2011. **121**(1): p. 328-341.

154. Moalem, S., et al., *XX male sex reversal with genital abnormalities associated with a de novo SOX3 gene duplication.* American Journal of Medical Genetics Part A, 2012. **158**(7): p. 1759-1764.

155. Oroz, M., et al., *The smallest dislocated microduplication of Xq27. 1 harboring SOX3 gene associated with XX male phenotype.* Journal of Pediatric Endocrinology and Metabolism, 2023. **36**(1): p. 86-90.

156. Zhang, X., et al., *Roles of intracellular fibroblast growth factors in neural development and functions.* Science China Life Sciences, 2012. **55**: p. 1038-1044.

157. Wu, Q.-F., et al., *Fibroblast growth factor 13 is a microtubule-stabilizing protein regulating neuronal polarization and migration.* Cell, 2012. **149**(7): p. 1549-1564.

158. Wittmack, E.K., et al., *Fibroblast growth factor homologous factor 2B: association with Nav1.6 and selective colocalization at nodes of Ranvier of dorsal root axons.* Journal of Neuroscience, 2004. **24**(30): p. 6765-6775.

159. Andrews, S. *FastQC: A quality control tool for high throughput sequence data.* 2010; Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

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

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

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

163. Pertea, G. and Pertea, M., *GFF utilities: GffRead and GffCompare.* F1000Research, 2020. **9**.

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

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

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

167. Robinson, M.D., McCarthy, D.J., and Smyth, G.K., *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-140.

168. Love, M.I., Huber, W., and Anders, S., *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12): p. 1-21.

169. Chen, Y., et al., *edgeR: differential analysis of sequence read count data User's Guide.* 2020; Available from:

https://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf.

170. Huang da, W., Sherman, B.T., and Lempicki, R.A., *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nature Protocols, 2009. **4**(1): p. 44-57.

171. Sherman, B.T., et al., *DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update).* Nucleic Acids Research, 2022. **50**(W1): p. W216-w221.

172. Chatterjee, A., et al., *A guide for designing and analyzing RNA-Seq data.* Gene Expression Analysis: methods and protocols, 2018: p. 35-80.

173. Veldman-Jones, M.H., et al., *Evaluating robustness and sensitivity of the NanoString technologies nCounter platform to enable multiplexed gene expression analysis of clinical samples.* Cancer Research, 2015. **75**(13): p. 2587-2593.

174. Schepers, G.E., Teasdale, R.D., and Koopman, P., *Twenty pairs of sox: extent, homology, and nomenclature of the mouse and human sox transcription factor gene families.* Developmental Cell, 2002. **3**(2): p. 167-170.

175. Collignon, J., et al., *A comparison of the properties of Sox-3 with Sry and two related genes, Sox-1 and Sox-2.* Development, 1996. **122**(2): p. 509-520.

176. Uchikawa, M., Kamachi, Y., and Kondoh, H., *Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken.* Mechanisms of Development, 1999. **84**(1-2): p. 103-120.

177. Bylund, M., et al., *Vertebrate neurogenesis is counteracted by Sox1–3 activity.* Nature Neuroscience, 2003. **6**(11): p. 1162-1168.

178. Shih, Y.-H., et al., *SoxB1 transcription factors restrict organizer gene expression by repressing multiple events downstream of Wnt signalling.* Development, 2010. **137**(16): p. 2671-2681.

179. Seyednasrollah, F., Laiho, A., and Elo, L.L., *Comparison of software packages for detecting differential expression in RNA-seq studies.* Briefings in Bioinformatics, 2015. **16**(1): p. 59-70.

180. Lamarre, S., et al., *Optimization of an RNA-Seq differential gene expression analysis depending on biological replicate number and library size.* Frontiers in Plant Science, 2018. **9**: p. 108.

181. Schurch, N.J., et al., *How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use?* RNA, 2016. **22**(6): p. 839-851.

182. Mertens, F., et al., *The emerging complexity of gene fusions in cancer.* Nature Reviews Cancer, 2015. **15**(6): p. 371-381.

183. Leng, N., et al., *EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments.* Bioinformatics, 2013. **29**(8): p. 1035-1043.

184. Trapnell, C., et al., *Differential analysis of gene regulation at transcript resolution with RNA-seq.* Nature Biotechnology, 2013. **31**(1): p. 46-53.

185. Dong, X., et al., *The long and the short of it: unlocking nanopore long-read RNA sequencing data with short-read differential expression analysis tools.* NAR Genomics and Bioinformatics, 2021. **3**(2): p. lqab028.

186. Liu, X., et al., *Improving RNA-Seq expression estimation by modeling isoform-and exon-specific read sequencing rate.* BMC Bioinformatics, 2015. **16**: p. 1-13.

187. Stark, R., Grzelak, M., and Hadfield, J., *RNA sequencing: the teenage years.* Nature Reviews Genetics, 2019. **20**(11): p. 631-656.

188. Pevny, L.H. and Lovell-Badge, R., *Sox genes find their feet.* Current Opinion in Genetics & Development, 1997. **7**(3): p. 338-344.

189. Bowles, J., Schepers, G., and Koopman, P., *Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators.* Developmental Biology, 2000. **227**(2): p. 239-255.

190. Bergsland, M., et al., *Sequentially acting Sox transcription factors in neural lineage development.* Genes & Development, 2011. **25**(23): p. 2453-2464.

191. Stevanovic, M., et al., *SOX transcription factors as important regulators of neuronal and glial differentiation during nervous system development and adult neurogenesis.* Frontiers in Molecular Neuroscience, 2021. **14**: p. 654031.

192. Klum, S., et al., *Sequentially acting SOX proteins orchestrate astrocyte-and oligodendrocyte-specific gene expression.* EMBO Reports, 2018. **19**(11): p. e46635.

193. McAninch, D. and Thomas, P., *Identification of highly conserved putative developmental enhancers bound by SOX3 in neural progenitors using ChIP-Seq.* PLoS One, 2014. **9**(11): p. e113361.

194. Rogers, N., McAninch, D., and Thomas, P., *Dbx1 is a direct target of SOX3 in the spinal cord.* PLoS One, 2014. **9**(4): p. e95356.

195. Graham, V., et al., *SOX2 functions to maintain neural progenitor identity.* Neuron, 2003. **39**(5): p. 749-765.

196. Rapaport, F., et al., *Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.* Genome Biology, 2013. **14**(9): p. 1-13.

197. Zhao, S., et al., *Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion.* Scientific Reports, 2018. **8**(1): p. 4781.

198. Zhu, H., et al., *X-linked congenital hypertrichosis syndrome is associated with interchromosomal insertions mediated by a human-specific palindrome near SOX3.* The American Journal of Human Genetics, 2011. **88**(6): p. 819-826.

199. Si, N., et al., *A 105 kb interstitial insertion in the Xq27. 1 palindrome from pseudoautosomal region PAR1 causes a novel X-linked recessive compound phenotype.* Journal of Translational Medicine, 2019. **17**(1): p. 1-10.

200. Boschann, F., et al., *Xq27. 1 palindrome mediated interchromosomal insertion likely causes familial congenital bilateral laryngeal abductor paralysis (Plott syndrome).* Journal of Human Genetics, 2022. **67**(7): p. 405-410.

201. Middelkamp, S., et al., *Prioritization of genes driving congenital phenotypes of patients with de novo genomic structural variants.* Genome Medicine, 2019. **11**(1): p. 1-15.

202. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Research, 2002. **12**(6): p. 996-1006.

203. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57.

204. Griffith, O.L., et al., *ORegAnno: an open-access community-driven resource for regulatory annotation.* Nucleic Acids Research, 2007. **36**(suppl_1): p. D107-D113.

205. Fishilevich, S., et al., *GeneHancer: genome-wide integration of enhancers and target genes in GeneCards.* Database, 2017. **2017**: p. bax028.

206. Bulger, M. and Groudine, M., *Functional and mechanistic diversity of distal transcription enhancers.* Cell, 2011. **144**(3): p. 327-339.

207. Long, H.K., et al., *Loss of extreme long-range enhancers in human neural crest drives a craniofacial disorder.* Cell Stem Cell, 2020. **27**(5): p. 765-783. e14.

208. Bahr, C., et al., *A Myc enhancer cluster regulates normal and leukaemic haematopoietic stem cell hierarchies.* Nature, 2018. **553**(7689): p. 515-520.

209. Bonev, B. and Cavalli, G., *Organization and function of the 3D genome.* Nature Reviews Genetics, 2016. **17**(11): p. 661-678.

210. Cai, Y., et al., *H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions.* Nature Communications, 2021. **12**(1): p. 719.

211. Xu, J., et al., *Subtype-specific 3D genome alteration in acute myeloid leukaemia.* Nature, 2022. **611**(7935): p. 387-398.

212. Gröschel, S., et al., *A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in leukemia.* Cell, 2014. **157**(2): p. 369-381.

213. Franke, M., et al., *Formation of new chromatin domains determines pathogenicity of genomic duplications.* Nature, 2016. **538**(7624): p. 265-269.

214. Socha, M., et al., *Position effects at the FGF8 locus are associated with femoral hypoplasia.* The American Journal of Human Genetics, 2021. **108**(9): p. 1725-1734.

215. Franke, M., et al., *Duplications disrupt chromatin architecture and rewire GPR101-enhancer communication in X-linked acrogigantism.* The American Journal of Human Genetics, 2022. **109**(4): p. 553-570.

216. Aljogol, D., et al., *Comparison of capture Hi-C analytical pipelines.* Frontiers in Genetics, 2022. **13**: p. 786501.

217. McCord, R.P., Kaplan, N., and Giorgetti, L., *Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function.* Molecular Cell, 2020. **77**(4): p. 688-708.

218. Dekker, J., et al., *Capturing chromosome conformation.* Science, 2002. **295**(5558): p. 1306-1311.

219. De Wit, E. and De Laat, W., *A decade of 3C technologies: insights into nuclear organization.* Genes & Development, 2012. **26**(1): p. 11-24.

220. Denker, A. and De Laat, W., *The second decade of 3C technologies: detailed insights into nuclear organization.* Genes & Development, 2016. **30**(12): p. 1357-1382.

221. Han, J., Zhang, Z., and Wang, K., *3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering.* Molecular Cytogenetics, 2018. **11**(1): p. 1-10.

222. Simonis, M., et al., *Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4C).* Nature Genetics, 2006. **38**(11): p. 1348-1354.

223. Van De Werken, H.J., et al., *Robust 4C-seq data analysis to screen for regulatory DNA interactions.* Nature Methods, 2012. **9**(10): p. 969-972.

224. Zhao, Z., et al., *Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions.* Nature Genetics, 2006. **38**(11): p. 1341-1347.

225. Dostie, J., et al., *Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements.* Genome Research, 2006. **16**(10): p. 1299-1309.

226. Lieberman-Aiden, E., et al., *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, 2009. **326**(5950): p. 289-293.

227. Zhang, Y., et al., *Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus.* Nature Communications, 2018. **9**(1): p. 750.

228. Lajoie, B.R., Dekker, J., and Kaplan, N., *The Hitchhiker's guide to Hi-C analysis: practical guidelines.* Methods, 2015. **72**: p. 65-75.

229. 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. 1665-1680.

230. Dryden, N.H., et al., *Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C.* Genome Research, 2014. **24**(11): p. 1854-1868.

231. Schoenfelder, S., et al., *The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements.* Genome Research, 2015. **25**(4): p. 582-597.

232. 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.

233. Freire-Pritchett, P., et al., *Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells.* Elife, 2017. **6**: p. e21926.

234. Ma, W., et al., *Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution.* Methods, 2018. **142**: p. 59-73.

235. Hsieh, T.-H.S., et al., *Mapping nucleosome resolution chromosome folding in yeast by micro-C.* Cell, 2015. **162**(1): p. 108-119.

236. Hsieh, T.-H.S., et al., *Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome.* Nature Methods, 2016. **13**(12): p. 1009-1011.

237. Krietenstein, N., et al., *Ultrastructural details of mammalian chromosome architecture.* Molecular Cell, 2020. **78**(3): p. 554-565. e7.

238. Hsieh, T.-H.S., et al., *Resolving the 3D landscape of transcription-linked mammalian chromatin folding.* Molecular Cell, 2020. **78**(3): p. 539-553. e8.

239. Cantata Bio, L., *Dovetail® Micro-C Kit USER GUIDE VERSION 1.2.* 2022; Available from: https://cantatabio.com/wp-content/uploads/2023/12/Dovetail%C2%AE-Micro-C-Kit-User-Guide-Version-1.2.pdf.

240. Cantata Bio, L. *Dovetail® Target Enrichment USER GUIDE VERSION 1.0*. 2022; Available from: https://cantatabio.com/wp-content/uploads/2023/12/Dovetail%C2%AE-Target-Enrichment-User-Guide-Version-1.0.pdf.

241. Martin, F.J., et al., *Ensembl 2023.* Nucleic Acids Research, 2023. **51**(D1): p. D933-D941.

242. Khalfan, M., et al., *Modifying Reference Sequence and Annotation Files Quickly and Reproducibly with ref orm.* 2021.

243. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* arXiv preprint arXiv:1303.3997, 2013.

244. Li, H., et al., *The sequence alignment/map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.

245. Open2C, A.N., et al., *Pairtools: from sequencing data to chromosome contacts.* bioRxiv, 2023.

246. Pedersen, B.S. and Quinlan, A.R., *Mosdepth: quick coverage calculation for genomes and exomes.* Bioinformatics, 2018. **34**(5): p. 867-868.

247. Cairns, J., et al., *CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data.* Genome Biology, 2016. **17**(1): p. 1-17.

248. Robinson, J.T., et al., *Integrative genomics viewer.* Nature Biotechnology, 2011. **29**(1): p. 24-26.

249. Durand, N.C., et al., *Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments.* Cell Systems, 2016. **3**(1): p. 95-98.

250. Durand, N.C., et al., *Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom.* Cell Systems, 2016. **3**(1): p. 99-101.

251. Wingett, S., et al., *HiCUP: pipeline for mapping and processing Hi-C data.* F1000Research, 2015. **4**.

252. Ben Zouari, Y., et al., *ChiCMaxima: a robust and simple pipeline for detection and visualization of chromatin looping in Capture Hi-C.* Genome Biology, 2019. **20**(1): p. 1-19.

253. Freire-Pritchett, P., et al., *Detecting chromosomal interactions in Capture Hi-C data with CHiCAGO and companion tools.* Nature Protocols, 2021. **16**(9): p. 4144-4176.

254. Cairns, J., et al., *Chicdiff: a computational pipeline for detecting differential chromosomal interactions in Capture Hi-C data.* Bioinformatics, 2019. **35**(22): p. 4764-4766.

255. Cairns J, F.-P.P., Spivakov M, *CHiCAGO Vignette.* 2016; Available from: https://www.bioconductor.org/packages/devel/bioc/vignettes/Chicago/inst/doc/Chicago.html.

256. Li, G., et al., *Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation.* Cell, 2012. **148**(1): p. 84-98.

257. Nagano, T., et al., *Comparison of Hi-C results using in-solution versus in-nucleus ligation.* Genome Biology, 2015. **16**(1): p. 1-13.

258. Hansen, P., et al., *Computational processing and quality control of Hi-C, capture Hi-C and Capture-C data.* Genes, 2019. **10**(7): p. 548.

259. Novo, C.L., et al., *Long-range enhancer interactions are prevalent in mouse embryonic stem cells and are reorganized upon pluripotent state transition.* Cell Reports, 2018. **22**(10): p. 2615-2627.

260. Lee, B.H., Wu, Z., and Rhie, S.K., *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. 1-20.

261. Tomás-Daza, L., et al., *Low input capture Hi-C (liCHi-C) identifies promoter-enhancer interactions at high-resolution.* Nature Communications, 2023. **14**(1): p. 268.

262. Rennie, S., et al., *Transcriptional decomposition reveals active chromatin architectures and cell specific regulatory interactions.* Nature Communications, 2018. **9**(1): p. 487.

263. Chovanec, P., et al., *Widespread reorganisation of pluripotent factor binding and gene regulatory interactions between human pluripotent states.* Nature Communications, 2021. **12**(1): p. 2098.

264. Holgersen, E.M., et al., *Identifying high-confidence capture Hi-C interactions using CHiCANE.* Nature Protocols, 2021. **16**(4): p. 2257-2285.

265. Alinejad-Rokny, H., et al., *MaxHiC: A robust background correction model to identify biologically relevant chromatin interactions in Hi-C and capture Hi-C experiments.* PLoS Computational Biology, 2022. **18**(6): p. e1010241.

266. Goel, V.Y., Huseyin, M.K., and Hansen, A.S., *Region Capture Micro-C reveals coalescence of enhancers and promoters into nested microcompartments.* Nature Genetics, 2023: p. 1-9.

267. Liu, N., et al., *Seeing the forest through the trees: prioritising potentially functional interactions from Hi-C.* Epigenetics & Chromatin, 2021. **14**(1): p. 1-17.

268. Marwaha, S., Knowles, J.W., and Ashley, E.A., *A guide for the diagnosis of rare and undiagnosed disease: beyond the exome.* Genome Medicine, 2022. **14**(1): p. 1-22.

269. 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.

270. de Bruijn, S.E., et al., *Structural variants create new topological-associated domains and ectopic retinal enhancer-gene contact in dominant retinitis pigmentosa.* The American Journal of Human Genetics, 2020. **107**(5): p. 802-814.

271. Arya, V.B., et al., *Xq27. 1 duplication encompassing SOX3: Variable phenotype and smallest duplication associated with hypopituitarism to date–A large case series of unrelated patients and a literature review.* Hormone Research in Paediatrics, 2020. **92**(6): p. 382-389.

272. Woods, K.S., et al., *Over-and underdosage of SOX3 is associated with infundibular hypoplasia and hypopituitarism.* The American Journal of Human Genetics, 2005. **76**(5): p. 833-849.

273. Du, C., et al., *Xq26. 3-q27. 1 duplication including SOX3 gene in a Chinese boy with hypopituitarism: case report and two years treatment follow up.* BMC Medical Genomics, 2022. **15**(1): p. 19.

274. Laumonnier, F., et al., *Transcription factor SOX3 is involved in X-linked mental retardation with growth hormone deficiency.* The American Journal of Human Genetics, 2002. **71**(6): p. 1450-1455.

275. Wright, E.M.B., et al., *X-linked isolated growth hormone deficiency: expanding the phenotypic spectrum of SOX3 polyalanine tract expansions.* Clinical Dysmorphology, 2009. **18**(4): p. 218.

276. Alatzoglou, K.S., et al., *Increased Transactivation Associated with SOX 3 Polyalanine Tract Deletion in a Patient with Hypopituitarism.* The Journal of Clinical Endocrinology & Metabolism, 2011. **96**(4): p. E685-E690.

277. Takagi, M., et al., *A novel mutation in SOX3 polyalanine tract: a case of Kabuki syndrome with combined pituitary hormone deficiency harboring double mutations in MLL2 and SOX3.* Pituitary, 2014. **17**: p. 569-574.

278. Jelsig, A.M., et al., *A complex phenotype in a family with a pathogenic SOX3 missense variant.* European Journal of Medical Genetics, 2018. **61**(3): p. 168-172.

279. Nagarajan, R., et al., *EGR2 mutations in inherited neuropathies dominant-negatively inhibit myelin gene expression.* Neuron, 2001. **30**(2): p. 355-368.

280. Bellone, E., et al., *A novel mutation (D305V) in the early growth response 2 gene is associated with severe Charcot-Marie-Tooth type 1 disease.* Human Mutation, 1999. **14**(4): p. 353-354.

281. Mikešová, E., et al., *Novel EGR2 mutation R359Q is associated with CMT type 1 and progressive scoliosis.* Neuromuscular Disorders, 2005. **15**(11): p. 764-767.

282. Warner, L.E., et al., *Mutations in the early growth response 2 (EGR2) gene are associated with hereditary myelinopathies.* Nature Genetics, 1998. **18**(4): p. 382-384.

283. Sevilla, T., et al., *The EGR2 gene is involved in axonal Charcot– Marie– Tooth disease.* European Journal of Neurology, 2015. **22**(12): p. 1548-1555.

284. Tozza, S., et al., *A novel family with axonal Charcot-Marie-Tooth disease caused by a mutation in the EGR2 gene.* Journal of the Peripheral Nervous System, 2019. **24**(2): p. 219-223.

285. Schriebl, K., et al., *Selective removal of undifferentiated human embryonic stem cells using magnetic activated cell sorting followed by a cytotoxic antibody.* Tissue Engineering Part A, 2012. **18**(9-10): p. 899-909.

286. Yamanaka, S., *Pluripotent stem cell-based cell therapy—promise and challenges.* Cell Stem Cell, 2020. **27**(4): p. 523-531.

287. Majd, H., et al., *Deriving Schwann cells from hPSCs enables disease modeling and drug discovery for diabetic peripheral neuropathy.* Cell Stem Cell, 2023. **30**(5): p. 632-647. e10.

288. Gomes, A.R., et al., *Modeling Rett syndrome with human patient-specific forebrain organoids.* Frontiers in Cell and Developmental Biology, 2020. **8**: p. 610427.

289. Chia, K., et al., *Zebrafish as a model organism for neurodegenerative disease.* Frontiers in Molecular Neuroscience, 2022. **15**: p. 940484.

290. Saleem, S. and Kannan, R.R., *Zebrafish: an emerging real-time model system to study Alzheimer's disease and neurospecific drug discovery.* Cell Death Discovery, 2018. **4**(1): p. 45.

291. Eisen, J.S. and Smith, J.C., *Controlling morpholino experiments: don't stop making antisense.* Development, 2008. **135**(10):1735-43.

292. Nishimura, K., et al., *An auxin-based degron system for the rapid depletion of proteins in nonplant cells.* Nature Methods, 2009. **6**(12): p. 917-922.

293. Yesbolatova, A., et al., *The auxin-inducible degron 2 technology provides sharp degradation control in yeast, mammalian cells, and mice.* Nature Communications, 2020. **11**(1): p. 5701.

294. Ashley, G.E., et al., *An expanded auxin-inducible degron toolkit for Caenorhabditis elegans.* Genetics, 2021. **217**(3): p. iyab006.

295. Negishi, T., et al., *The auxin-inducible degron 2 (AID2) system enables controlled protein knockdown during embryogenesis and development in Caenorhabditis elegans.* Genetics, 2022. **220**(2): p. iyab218.

296. Zhang, L., et al., *The auxin-inducible degradation (AID) system enables versatile conditional protein depletion in C. elegans.* Development, 2015. **142**(24): p. 4374-4384.

297. Fledrich, R., et al., *Soluble neuregulin-1 modulates disease pathogenesis in rodent models of Charcot-Marie-Tooth disease 1A.* Nature Medicine, 2014. **20**(9): p. 1055-1061.

APPENDICES

(commence on following page)

Appendix A

Differential Gene Expression Analysis using edgeR

This analysis was performed in RStudio using EdgeR (Robinson et al., 2010) (3.40.2), and based on the basic workflow outlined in the edgeR User Guide: `https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf` (Chen et al., 2014).

Load the relevant packages:

```
library(edgeR)
library(dplyr)
```

Read in the table of counts:

```
Counts <- read.table("counts.csv", header = TRUE, sep = ",", row.names = 1)
Counts <- as.matrix(Counts)
```

Set up the cohorts; 'Patient' and 'Control':

```
condition <- factor(c(rep("Patient", 1), rep("Control", 3)))
```

Make a 'DGEList' data object:

```
y <- DGEList(counts = Counts, group = condition)
```

Perform filtering to discard low-count genes:

```
keep <- filterByExpr(y)
y <- y[keep, , keep.lib.sizes=FALSE]
```

Calculate scaling factors for normalising library sizes:

```
y <- calcNormFactors(y)
```

Set up the design matrix:

```
design <- model.matrix(~condition)
```

Estimate dispersions:

```
y <- estimateDisp(y, design)
```

Explore the data:

```
plotMDS(y)
plotBCV(y)
y$common.dispersion
```

Test for differential gene expression using the ExactTest:

```
et <- exactTest(y)
topTags(et)
summary(decideTests(et))
```

Export results:

```
AllGenesData <- as.data.frame(topTags(et, n=30000))
write.table(AllGenesData, "AllGenes_patient_vs_control.txt", sep = "\t", col.names = NA,
quote = FALSE)
plotMD(et)
GenesFDR0.05 <- AllGenesData %>% subset(FDR < 0.05)
write.table(X3.211_FDR0.05, "X3.211_FDR_Less_0.05.txt", sep = "\t", col.names = NA,
quote = FALSE)
```

Appendix B

Differential Gene Expression Analysis using DESeq2

This analysis was performed in RStudio using DESeq2 (Love et al., 2014) (1.38.3), based on the instructions in the DESeq2 Vignette: `https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#quick-start` (Love et al., 2024).

This pipeline has previously been used by our laboratory (Cutrupi et al., 2023).

Load the relevant packages:

```
library(DESeq2)
```

Read in the table of counts:

```
counts <- read.table("counts.csv", sep = ",", header = TRUE, row.names = 1)
counts <- as.matrix(counts)
```

Set up cohorts; 'Patient' and 'Control':

```
condition <- factor(c(rep("Patient", 2), rep("Control", 3)))
samples <- colnames(counts)
coldata <- data.frame(row.names=colnames(counts), condition, samples)
```

Make a DESeqDatSet:

```
dds <- DESeqDataSetFromMatrix(countData = counts, colData = coldata, design = ~ condition)
```

Filter low-count genes (counts < 10 in less than two samples):

```
keep <- rowSums(counts(dds) >= 10) >=2
dds <- dds[keep,]
```

Set 'Control' as the reference group:

```
dds$condition <- relevel(dds$condition, ref = "Control")
```

Perform differential expression analysis and explore results:

```
ddsDEG <- DESeq(dds)
res <- results(ddsDEG)
summary(res)
```

Include extra information in the results (outliers, low counts):

```
MaxCooks <- apply(assays(ddsDEG)[["cooks"]], 1, max)
LowCounts <- res$baseMean < metadata(res)$filterThreshold
res$Outlier = res$baseMean > 0 & is.na(res$pvalue)
res$MaxCooks = MaxCooks
res$LowCounts = LowCounts
```

Order results by the adjusted p-values:

```
resOrdered <- res[order(res$pvalue),]
summary(resOrdered)
resSig <- subset(resOrdered, padj < 0.1)
```

write results files:

```
write.csv(as.data.frame(resOrdered), file="Patient_vs_control_all_res_default_alpha_Ordered_Pad
j.csv")
write.csv(as.data.frame(resSig), file="Patient_vs_control_sig_res_default_alpha_0.1.csv")
```

Tools for exploring the data:

1. PCA plot

```
rld <- rlog(dds, blind = FALSE)
plotPCA(rld, intgroup = c("condition", "samples"))
plotPCA(rld, intgroup = c("condition"))
```

2. Boxplot of Cook's distances

```
par(mar=c(8,5,2,2))
boxplot(log10(assays(ddsDEG)[["cooks"]]), range=0, las=2)
```

3. MA Plot

```
plotMA(res, ylim=c(-2,2))
```

4. Dispersion estimates

```
plotDispEsts(ddsDEG)
```

Appendix C

Generating the Custom CMTX3 Chromosome X Reference

Step 1: Downloading Chr8 insertion sequence

Download the Chr8 reference sequence from Ensembl:

```
wget https://ftp.ensembl.org/pub/release-97/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.8.fa.gz
gunzip Homo_sapiens.GRCh38.dna.chromosome.8.fa.gz
```

Print the first few lines of the FASTA to determine how many characters are in the header line:

```
head -n 5 Homo_sapiens.GRCh38.dna.chromosome.8.fa
```

The header contains **55 characters** which must be factored in when specifying the starting position for the Chr8 insertion.

- The insertion starts at chr8:144,542,928 (i.e. character 144,542,983 within the FASTA), encompassing *77,846* bp in total

Extract the relevant sequence and print output to text file:

```
cat Homo_sapiens.GRCh38.dna.chromosome.8.fa | tr -d '\n' | awk '{print substr($0,144542983,77846)}' > Chr8_Insert_Region_Release97.txt
```

Step 2: Downloading the small Chr12 insertion at distal breakpoint

Download the Chr12 reference sequence from Ensembl:

```
wget https://ftp.ensembl.org/pub/release-97/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.12.fa.gz
gunzip Homo_sapiens.GRCh38.dna.chromosome.12.fa.gz
```

Print the first few lines of the FASTA to determine how many characters are in the header line:

```
head -n 5 Homo_sapiens.GRCh38.dna.chromosome.12.fa
```

The header contains **57 characters**, which must be factored in when specifying the starting position for the Chr12 insertion

- The insertion sequence starts at chr12:49,903,866 (i.e. character 49,903,923 within the FASTA), encompassing *9 bp* in total

Extract the relevant sequence and print output to text file:

```
cat Homo_sapiens.GRCh38.dna.chromosome.12.fa | tr -d '\n' | awk '{print substr($0,49903923,9)}' > Chr12_Insert_Seq.txt
```

Step 3: Downloading ChrX sequence at distal breakpoint

Download the ChrX reference sequence from Ensembl:

```
wget https://ftp.ensembl.org/pub/release-97/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.chromosome.X.fa.gz
gunzip Homo_sapiens.GRCh38.dna.chromosome.X.fa.gz
```

Print the first few lines of the FASTA to determine how many characters are in the header line:

```
head -n 5 Homo_sapiens.GRCh38.dna.chromosome.X.fa
```

The header contains **55 characters**, which must be factored in when specifying the location of target sequences on ChrX

- The ChrX inversion starts at chrX:140,420,784 (i.e. character 140,420,839 within the FASTA), encompassing *12 bp* in total
- The SNV occurs at position 140,420,803 (i.e. character 140,420,858 within the FASTA)
- The single nucleotide deletion occurs at position 140,420,811 (i.e. character 140,420,866 within the FASTA)

First, extract the *12 bp* involved in the ChrX inversion:

```
cat Homo_sapiens.GRCh38.dna.chromosome.X.fa | tr -d '\n' | awk '{print substr($0,140420839,12)}'
> ChrX_Region_For_Inversion.txt
```

Then, use Nano Text Editor to change this sequence to its **reverse complement** (and change file name):

```
nano ChrX_Region_For_Inversion.txt
mv ChrX_Region_For_Inversion.txt ChrX_Inversion_Complete.txt
```

Then, extract the *7 bp* of control ChrX sequence separating the small inversion and the downstream SNV:

```
cat Homo_sapiens.GRCh38.dna.chromosome.X.fa | tr -d '\n' | awk '{print substr($0,140420851,7)}'
> Normal_ChrX_Between_Inv_and_SNV.txt
```

Use Nano Text Editor to add the SNV *('G')* immediately after the last nucleotide:

```
nano Normal_ChrX_Between_Inv_and_SNV.txt
```

Extract the *7 bp* of control ChrX sequence separating the SNV and the downstream deletion:

```
cat Homo_sapiens.GRCh38.dna.chromosome.X.fa | tr -d '\n' | awk '{print substr($0,140420859,7)}'
> Normal_ChrX_Between_SNV_and_Del.txt
```

Step 4: Join mutation sequences into a single file

String together the above files, in the correct order, to create a file representing the entire CMTX3 SV sequence:

```
cat Chr8_Insert_Region_Release97.txt Chr12_Insert_Seq.txt ChrX_Inversion_Complete.txt Normal_ChrX_Between_Inv_and_SNV.txt Normal_ChrX_Between_SNV_and_Del.txt | tr -d '\n' > TOTAL_CMTX3_REGION.txt
```

Use Nano Text Editor to manually add a header line, to comply with FASTA format:

```
nano TOTAL_CMTX3_REGION.txt
```

Convert the text file to FASTA format:

```
mv TOTAL_CMTX3_REGION.txt TOTAL_CMTX3_REGION.fa
```

Step 5: Running the *Reform* tool to insert CMTX3 sequence within ChrX

Reform requires users to provide FASTA files containing the sequences that flank the novel insertion sequence as input

Extract the *3,000 bp* upstream and downstream of the CMTX3 SV (incorporating the 55 character offset from the header):

```
cat Homo_sapiens.GRCh38.dna.chromosome.X.fa | tr -d '\n' | awk '{print substr($0,140417838,300
0)}' > Upstream_ChrX.txt
cat Homo_sapiens.GRCh38.dna.chromosome.X.fa | tr -d '\n' | awk '{print substr($0,140420867,300
0)}' > Downstream_ChrX.txt
```

Manually add a header line to each file to comply with FASTA format, using Nano:

```
nano Upstream_ChrX.txt
nano Downstream_ChrX.txt
```

Convert files to FASTA format:

```
mv Upstream_ChrX.txt Upstream_ChrX.fa
mv Downstream_ChrX.txt Downstream_ChrX.fa
```

Reform also requires GFF files for ChrX as input

Download the reference ChrX GFF file from Ensembl:

```
wget https://ftp.ensembl.org/pub/release-97/gff3/homo_sapiens/Homo_sapiens.GRCh38.97.chromosome.
X.gff3.gz
gunzip Homo_sapiens.GRCh38.97.chromosome.X.gff3.gz
```

Clone the github repository for *Reform*:

```
git clone https://github.com/gencorefacility/reform
```

Finally, run *Reform* to generate CMTX3 patient-specific ChrX reference FASTA:

```
python3 reform/reform.py \
--chrom="X" \
--upstream_fasta="Upstream_ChrX.fa" \
--downstream_fasta="Downstream_ChrX.fa" \
--in_fasta="TOTAL_CMTX3_REGION.fa" \
--in_gff="CMTX3_annotation_file.gtf" \
--ref_fasta="Homo_sapiens.GRCh38.dna.chromosome.X.fa" \
--ref_gff="Homo_sapiens.GRCh38.97.chromosome.X.gff3"
```

The output of this script is a *'reformed'* FASTA file, **'CMTX3_reformed_chrX.fa'**, which will be used to construct our CMTX3 reference genome (see Appendix E)

Appendix D

Deleting the CMTX3 Insertion Sequence from Chromosome 8

Here, we summarise the process of deleting a 77,846 bp region, corresponding to the CMTX3 Chr8 insertion sequence, from the Chr8 reference FASTA

This protocol assumes that the following steps have been completed (see Appendix C):

- Chr8 reference FASTA has been downloaded from Ensembl
- Character length of the FASTA header line has been determined

Step 1: Cut out the sequence located upstream of the region to be deleted

The final nucleotide before the deleted region is Chr8:144,542,927 (i.e. character 144,542,982 within the FASTA)

Extract the relevant sequence and print output to text file:

```
cat Homo_sapiens.GRCh38.dna.chromosome.8.fa | tr -d '\n' | cut -c 56 - 144542982 > upstream_inse
rtion_8.txt
```

Step 2: Cut out the sequence located downstream of the region to be deleted

The first nucleotide after the deleted region is Chr8:144,620,773 (i.e. character 144,620,829 within the FASTA)

Extract the relevant sequence and print output to text file:

```
cat Homo_sapiens.GRCh38.dna.chromosome.8.fa | tr -d '\n' | cut -c 144620829- > downstream_insert
ion_8.txt
```

Step 3: Join files together to create the custom Chr8 FASTA

String the two files together, in the correct order, to create the custom Chr8 file:

```
cat upstream_insertion_8.txt downstream_insertion_8.txt | tr -d '\n' > chr8_without_insertion_se
q.txt
```

Convert the text file to FASTA format:

```
mv chr8_without_insertion_seq.txt chr8_without_insertion_seq.fa
```

Set line width to 60 characters, to comply with FASTA format:

```
fold -w 60 chr8_without_insertion_seq.fa > wrap60_chr8_without_insertion_seq.fa
```

Add header line to comply with FASTA format, using Nano Text Editor:

```
nano wrap60_chr8_without_insertion_seq.fa
```

The resulting FASTA file, **‘wrap60_chr8_without_insertion_seq.fa’**, will be used to construct our CMTX3 reference genome (see Appendix E)

Appendix E

Generating custom CMTX3 reference genome FASTA

Below is the script that was run to construct the custom CMTX3 reference genome.

This script incorporates the custom-made ChrX and Chr8 files (generated in Appendices C & D, respectively) within the hg38 reference genome, to generate a complete reference genome for the alignment of CMTX3 patient sequencing data.

```
#!/bin/bash

first=(1 2 3 4 5 6 7)
middle=(9 10 11 12 13 14 15 16 17 18 19 20 21 22 MT)


# Download reference FASTA files for Chr 1 - 7

for chromosomes in ${first[@]}; do
  curl "https://ftp.ensembl.org/pub/release-97/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.
  chromosome.${chromosomes}.fa.gz" | zcat >> chr1to7_hg38_Rel97.fa
done


# Download reference FASTA files for Chr 9 - MT

for chromosomes in ${middle[@]}; do
  curl "https://ftp.ensembl.org/pub/release-97/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.
  chromosome.${chromosomes}.fa.gz" | zcat >> chr9toMT_hg38_Rel97.fa
done


# Download reference ChrY FASTA file

curl "https://ftp.ensembl.org/pub/release-97/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.
chromosome.Y.fa.gz" | zcat >> chrY_hg38_Rel97.fa

# String together the reference chromosomes in the correct order, incorporating the
custom-built Chr8 and ChrX
# This generates a FASTA file containing the patient-specific custom reference genome

cat chr1to7_hg38_Rel97.fa wrap60_chr8_without_insertion_seq.fa chr9toMT_hg38_Rel97.fa
CMTX3_reformed_chrX.fa chrY_hg38_Rel97.fa > CMTX3_Reference_Genome.fa
```

The output of this script is a FASTA file, **'CMTX3_Reference_Genome.fa'**, which will be used for downstream analysis.

Appendix F

Editing Coordinates of Target Capture Sequences for the Custom CMTX3 Genome

This is an overview of the protocol for editing the location of the target capture bait and probe sequences so they align with the custom-built CMTX3 reference genome.

Step 1: Downloading the control hg38 bait and probe sequences

Technical note: *one 'bait' is comprised of multiple 'probe' sequences targeting the same promoter*

The *.bed* files specifying the coordinates for the capture baits and probes were downloaded:

```
wget https://s3.amazonaws.com/dovetail.pub/capture/human/h_baits_v1.0.bed
wget https://s3.amazonaws.com/dovetail.pub/capture/human/h_probes_v1.0.bed
```

Step 2: Modifying the affected ChrX baits and probes

The CMTX3 SV results in a net gain of 77,854 bp to ChrX

- Therefore, any baits/probes located downstream of the CMTX3 SV must be shifted by +77,854 bp

For any bait located **after** the CMTX3 SV breakpoint (chrX:140,420,783), we added the value of 77,854 to the 'start' and 'end' positions:

```
awk '{if ($1=="chrX" && $2>140420783) ($2+=77854) ($3+=77854); print$0}' \
OFS="\t" h_baits_v1.0.bed > CMTX3_h_baits_v1.0.bed

awk '{if ($1=="chrX" && $2>140420783) ($2+=77854) ($3+=77854); print$0}' \
OFS="\t"h_probes_v1.0.bed > CMTX3_h_probes_v1.0.bed
```

Step 3: Modifying the affected Chr8 baits and probes

Deleting the CMTX3 insertion sequence from Chr8 results in a net loss of 77,846 bp from this chromosome

- First, any baits/probes contained within the deleted sequence must be transferred from Chr8 to the CMTX3 SV coordinates on ChrX
- Then, any baits/probes located downstream of the deleted sequence must be shifted by -77,846 bp

The *.bed* files were manually inspected to identify if any baits/probes were contained within the deleted sequence. One bait sequence was identified:

```
chr8    144605579   144606199   bait_v1.0_28765 0   .
```

Two probe sequences were identified:

```
chr8    144605579   144605699   probe_v1.0_121524   0   .
chr8    144606079   144606199   probe_v1.0_121525   0   .
```

Their coordinates were manually modified to align with the custom CMTX3 chromosome X. Updated bait coordinates:

```
chrX    140483435   140484055   bait_v1.0_28765 0   .
```

Updated probe coordinates:

```
chrX    140483435   140483555   probe_v1.0_121524   0   .
chrX    140483935   140484055   probe_v1.0_121525   0   .
```

Updated coordinates were inserted into the patient-specific *.bed* files using Nano, whilst the original bait/probes on Chr8 were deleted:

```
nano CMTX3_h_baits_v1.0.bed
nano CMTX3_h_probes_v1.0.bed
```

For any bait/probe located **after** the Chr8 deletion (i.e. after chr8:144,542,927), we subtracted the value of 77,846 from the 'start' and 'end' positions:

```
awk '{if ($1=="chr8" && $2>144542927) ($2-=77846) ($3-=77846); print$0}' \
OFS="\t" CMTX3_h_baits_v1.0.bed > CMTX3_h_baits_v1.0.bed
awk '{if ($1=="chr8" && $2>144542927) ($2-=77846) ($3-=77846); print$0}' \
OFS="\t" CMTX3_h_probes_v1.0.bed > CMTX3_h_probes_v1.0.bed
```

Appendix G

Analysis of Promoter Capture Micro-C Data

Analysis of Promoter Capture Micro-C data was based on the Promoter Panel Data Analysis Documentation published by Dovetail Genomics: `https://dovetail-capture.readthedocs.io/en/latest/index.html`

Preparing for analysis

Clone the github repository from Dovetail Genomics:

```
git clone https://github.com/dovetail-genomics/capture.git
```

Make a genome file for each reference genome:

```
cut -f1,2 hg38.fa.fai > hg38.genome

cut -f1,2 CMTX3_Reference_Genome.fa.fai > CMTX3_Reference_Genome.genome
```

The use of variables simplifies switching between different samples and reference genomes. A list of variables is supplied below:

```
ref_fasta= # path to reference fasta (.fa)
read1= # path to read1 (.fastq.gz)
read2= # path to read2 (.fastq.gz)
sample= # sample ID
output= # path to output directory
ref_genome= # path to reference genome file (.genome)
temp= # path to temp directory
probes= # path to directory containing probe files
ref= # ID for the reference genome to be used (i.e. 'CMTX3' or 'hg38')
```

Processing the raw data

This section outlines processing the raw sequencing reads for downtream analysis.

Align reads using BWA MEM algorithm:

```
bwa mem -5SP -T0 -t16 ${ref_fasta} ${read1} ${read2} \
-o  ${output}/${sample}_aligned.sam
```

Process the raw data with Pairtools:

```
# find valid ligation events

pairtools parse --min-mapq 40 --walks-policy 5unique --max-inter-align-gap 30 --nproc-in 8 \
--nproc-out 8 --chroms-path ${ref_genome} ${output}/${sample}_aligned.sam \
> ${output}/${sample}_parsed.pairsam

# sort the pairs

pairtools sort --nproc 16 --tmpdir=${temp} \
${output}/${sample}_parsed.pairsam > ${output}/${sample}_sorted.pairsam

# remove PCR duplicates

pairtools dedup --nproc-in 8 --nproc-out 8 \
--mark-dups --output-stats ${output}/${sample}_stats.txt \
--output ${output}/${sample}_dedup.pairsam ${output}/${sample}_sorted.pairsam

# generate files for downstream analysis

pairtools split --nproc-in 8 --nproc-out 8 \
--output-pairs ${output}/${sample}_mapped.pairs \
--output-sam ${output}/${sample}_unsorted.bam ${output}/${sample}_dedup.pairsam
```

Prepare the BAM file for downstream analysis:

```
# sort BAM file using SAMtools

samtools sort -@16 -T ${temp} \
-o ${output}/${sample}_mapped_PT.bam ${output}/${sample}_unsorted.bam

# index the BAM file using SAMtools

samtools index ${output}/${sample}_mapped_PT.bam

# generate a BAM file that is compatible with the CHiCAGO pipeline

samtools view -@ 16 -Shu -F 2048 \
${output}/${sample}_mapped_PT.bam | samtools sort -n -T ${temp}/temp.bam \
--threads 16 -o ${output}/${sample}_chicago.bam -
```

Quality control assessment

First, assess the Micro-C library statistics generated by Pairtools. Running the 'get_qc.py' script provided by Dovetail Genomics will generate a summary file containing important QC metrics:

```
python3 ./capture/get_qc.py -p ${output}/${sample}_stats.txt
```

Assess performance of the Target Enrichment process by counting the number of reads that map to captured regions ('on-target'reads). Captured regions are specified in the *.bed* file (note probe coordinates have been padded by 200 bp)

```
samtools view ${output}/${sample}_mapped_PT.bam \
-L ${probes}/${ref}_h_probes_200_v1.0.bed \
-@ 16|awk -F "\t" '{print "@"$1}'|sort -u|wc -l \
> ${output}/${sample}_on_target_count.txt
```

Compare the sequencing coverage at targeted regions versus non-targeted regions using Mosdepth, and print the relevant summary information to a text file:

```
mosdepth -t 16 -b ${probes}/${ref}_h_probes_v1.0.bed \
-x ${output}/${sample}_probes \
-n ${output}/${sample}_mapped_PT.bam

head -1 ${sample}_probes.mosdepth.summary.txt > ${sample}_coverage_summary.txt

tail -2 ${sample}_probes.mosdepth.summary.txt >> ${sample}_coverage_summary.txt
```

Detecting chromatin loops using CHiCAGO

This section outlines the detection of chromatin interactions using the CHiCAGO package, specifically designed for Capture Hi-C datasets.

Ensure the CHiCAGO R package and its dependencies have been installed in RStudio

Prepare for analysis by cloning the github repository from Dovetail Genomics:

```
git clone https://github.com/dovetail-genomics/chicago.git
```

Download the 'Design Files' required for running CHiCAGO with the Dovetail Pan Promoter Enrichment Panel Micro-C libraries. These files are supplied by Dovetail, and use a 10 kb binsize for fragmenting the genome.

```
wget https://s3.amazonaws.com/dovetail.pub/capture/human/CHiCAGO_files/h_10kDesignFiles.tar.gz
tar xzvf h_10kDesignFiles.tar.gz
```

- **IMPORTANT NOTE:** the 'Design Files' need to be modified for the custom-built CMTX3 reference genome. This was done according to the instructions outlined in Dovetail Genomics' analysis documentation:
  `https://dovetail-capture.readthedocs.io/en/latest/interactions.html#generate-your-own-chicago-design-directory`

'chinput' files are used as input for CHiCAGO. Generate them using the 'bam2chicago.sh' script supplied by Dovetail Genomics:

```
# to use hg38 design files, run the following command:

./chicago/chicagoTools/bam2chicago.sh ${output}/${sample}_chicago.bam \
./h_10kDesingFiles/pooled_10kb_120bp.baitmap \
./h_10kDesingFiles/digest10kb_pooled120bp.rmap \
10kb_chinput_${sample}

# to use the custom CMTX3 design files, run the following command:

./chicago/chicagoTools/bam2chicago.sh ${output}/${sample}_chicago.bam \
./h_10kb_CMTX3_DesignFiles/CMTX3_pooled120bp.baitmap \
./h_10kb_CMTX3_DesignFiles/CMTX3_digest10kb_pooled120bp.rmap \
10kb_chinput_${sample}
```

Run the CHiCAGO pipeline to detect chromatin interactions. Note that in this exploratory analysis, a lenient CHiCAGO cutoff score of 3 was utilised, to enable investigation of 'sub-threshold' interactions (3 < CHiCAGO score < 5), as well as significant interactions (CHiCAGO score > 5).The default CHiCAGO cutoff score is 5.

```
# to use hg38 design files, run the following command:

Rscript ./chicago/chicagoTools/runChicago.R --design-dir ./h_10kDesingFiles \
--cutoff 3 \
--export-format interBed,washU_text,seqMonk,washU_track \
./10kb_chinput_${sample}/10kb_chinput_${sample}.chinput \
${sample}_chicago_calls_cutoff3

# to use the custom CMTX3 design files, run the following command:

Rscript ./chicago/chicagoTools/runChicago.R --design-dir ./h_10kb_CMTX3_DesignFiles \
--cutoff 3 \
--export-format interBed,washU_text,seqMonk,washU_track \
./10kb_chinput_${sample}/10kb_chinput_${sample}.chinput \
${sample}_chicago_calls_cutoff3
```

Generating contact matrices

Contact matrices were made using Juicer Tools, following the instructions provided in the Dovetail Genomics analysis documentation. This generates *.hic* files that can be visualised using Juicebox software.

Download Juicer Tools:

```
wget https://s3.amazonaws.com/hicfiles.tc4ga.com/public/juicer/juicer_tools_1.22.01.jar
mv juicer_tools_1.22.01.jar ./capture/juicertools.jar
```

Generate the *.hic* file (using the *.pairs* file generated from Pairtools):

```
java -Xmx48000m -Djava.awt.headless=true \
-jar /data/DovetailFiles/capture/juicer_tools_1.22.01.jar \
pre --threads 16 ${output}/${sample}_mapped.pairs ${sample}_contact_map.hic \
${ref_genome}
```