

**EXPLORING THE LINK BETWEEN
CHD2 MUTATIONS
AND
DOUBLE STRAND BREAK REPAIR
IN DEVELOPING NEURONS**

**Dr Ian James Tully (MBBCh MRCPCH)
Student Number: 1700901
School of Biosciences
Cardiff University**

**PhD Thesis
2020**

Acknowledgements

First, I would like to thank Professor Adrian Harwood, my lead supervisor, for his guidance, throughout this project – it has been a personally turbulent three years and completing this project would not have been possible without his understanding and support.

I owe a great deal of gratitude to those who supported me through the application process for the Welsh Clinical Academic Track (WCAT) fellowship – in particular Dr Andrew Fry, Professor Julian Sampson, Professor Michael Owen and Dr Alex Murray.

There are too many people within NMHRI who have supported me in various ways to thank each of them individually. I will however signal out Dr William Plumbly, Dr Amy Baldwin and Dr Mouhammed Alsaqati for holding my hand (sometimes quite literally) as I learned how to hold the pipette during the transition from clinical work into academia. Without their patience and willingness to share their expertise, I would be nowhere.

Finally, I would like to offer special thanks to my collaborators, Felix Dobbs and Professor Simon Reed. Some of the most interesting results in this thesis are the fruit of this collaboration and I look forward to continuing our work together in the future.

Dedication

I dedicate this work to my family, who are everything.

DECLARATION

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed (candidate)

STATEMENT 1

Date

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD

Signed (candidate)

STATEMENT 2

Date

This thesis is the result of my own independent work/investigation, except where otherwise stated. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed(candidate)

STATEMENT 3

Date

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed (candidate).....

STATEMENT 4: PREVIOUSLY APPROVED BAR ON ACCESS

Date

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loans after expiry of a bar on access previously approved by the Academic Standards & Quality Committee.

Signed (candidate).....

WORD COUNT: 65,945

SUMMARY OF THESIS

Introduction:

Heterozygous mutations *CHD2* cause intellectual disability and refractory epilepsy. Functional studies demonstrate a deficit in DNA double strand break (DSB) repair via non-homologous end joining (NHEJ) in *CHD2* deficient cells. This project aims to investigate the impact of *CHD2* mutations on the outcomes of DSB repair in developing neurons.

Methods:

A doxycycline inducible CRISPR-Cas9 (iCas9) construct was integrated into the AAVS1 safe harbour. A pipeline for high-throughput analysis of CRISPR experiments was designed and implemented based on nanopore sequencing. This pipeline was used to create a *CHD2*-deficient human induced pluripotent stem cell (hiPSC) line, which was used to investigate the effects of *CHD2* on neurodifferentiation and DSB repair. The pipeline was then adapted to monitor repair outcomes of targeted DSBs in differentiating cells.

Spontaneously occurring DSBs were examined using a novel next-generation sequencing technique, INDUCE-Seq, in which unrepaired DSBs are captured from permeabilised cells in-situ and sequenced in order to provide a snapshot of DSBs existing at the time of extraction in differentiating cells.

Results:

Cas9 induction of DSBs demonstrated changes in the repair, with an increased rate of large deletions in the *CHD2*-deficient cell line. No reproducible change in the pattern of smaller indels was identified.

There was a significant increase of the number of DSBs captured by INDUCE-seq at D19 of differentiation in WT cells, which was not present until D40 in *CHD2*-deficient cells. Differences in the enrichment of DSBs at various histone markers, gene bodies and transcription start sites (TSS) were identified between *CHD2*-deficient cells and WT cells.

Conclusions:

This study demonstrates an impact on the occurrence and repair of DSBs in *CHD2*-deficient cell lines. Integration of RNA-Seq data and analysis of the pattern of spontaneous breakage suggests that altered DSB repair physiology could contribute towards the phenotype exhibited by patients with *CHD2* mutations.

TABLE OF CONTENTS

1	<u>INTRODUCTION</u>	1
	<u>1.1 The CHD2 gene and protein</u>	3
	1.1.1 CHD2 Expression	5
	1.1.2 CHD2 protein structure	7
	1.1.3 Comparison to other CHD2 family proteins	9
	<u>1.2 Clinical relevance of CHD2 mutations in humans</u>	13
	1.2.1 Clinical features of CHD2 related EECO	15
	1.2.2 CHD2 mutations in cancer	17
	1.2.3 Summary of clinical relevance	17
	<u>1.3 Animal models of CHD2 deficiency</u>	18
	1.3.1 Discrepancies between animal and human phenotypes	18
	<u>1.4 Cell models of CHD2 function</u>	21
	1.4.1 Cell models of neurodevelopment	21
	1.4.2 Myogenesis	22
	<u>1.5 Chromatin modification</u>	23
	1.5.1 Chromatin structure and function	23
	1.5.1.1 <i>Introduction to chromatin</i>	23
	1.5.1.2 <i>Histones and chromatin structure</i>	24
	1.5.1.3 <i>Histone tail modifications</i>	25
	1.5.1.4 <i>Nucleosome mobilisation</i>	26
	1.5.1.5 <i>Histone variants</i>	30
	1.5.2 CHD2 mediated chromatin modification	33
	<u>1.6 DNA repair</u>	35
	1.6.1 DNA damage	35
	1.6.2 Double strand breakage in the developing brain	39
	1.6.3 The DNA damage response to DSBs	42
	1.6.3.1 <i>DSB detection and pathway choice</i>	43
	1.6.3.2 <i>Non-homologous end-joining</i>	44
	1.6.3.3 <i>Homologous recombination</i>	47
	1.6.3.4 <i>Alternative end-joining</i>	47
	1.6.4 Summary of DNA damage repair	50
	<u>1.7 Summary and hypothesis</u>	51
	1.7.1 Hypothesis	51
	<u>1.8 Aims and approach</u>	54

2	<u>GENERAL METHODS & MATERIALS</u>	56
2.1	<u>Introduction</u>	56
2.2	<u>Tables of reagents and equipment</u>	57
2.3	<u>Standard cell culture procedures</u>	62
2.3.1	hiPSC culture and maintenance	62
2.3.1.1	<i>Cell line description</i>	62
2.3.1.2	<i>Plate preparation</i>	62
2.3.1.3	<i>Culture media</i>	62
2.3.1.4	<i>Passaging techniques</i>	63
2.3.1.5	<i>Freezing cells</i>	64
2.3.1.6	<i>Thawing cells</i>	64
2.3.1.7	<i>Puromycin kill curve</i>	64
2.3.2	Differentiation of hiPSCs into neurons	65
2.3.2.1	<i>Plating and daily care</i>	65
2.3.2.2	<i>First passage – day 9</i>	65
2.3.2.3	<i>Second passage and subsequent passages</i>	65
2.3.3	Additional culture techniques	66
2.3.4	Nucleofection procedures	67
2.3.4.1	<i>Electroporation</i>	67
2.3.4.2	<i>Lipofection</i>	68
2.3.5	Cell lines and passage numbers for comparisons between wild-type and CHD2 deficient lines	69
2.4	<u>Molecular biology protocols</u>	70
2.4.1	Transformation of bacteria with required plasmid	70
2.4.2	Miniprep	70
2.4.3	Maxiprep	71
2.4.4	Plasmids for inducible Cas9 integration	71
2.4.5	Validation of plasmid integration	73
2.4.5.1	<i>Plasmid integration</i>	73
2.4.5.2	<i>Cas9 protein production</i>	73
2.5	<u>Genetics and Genomics Protocols</u>	74
2.5.1	DNA extraction	74
2.5.2	RNA extraction	74
2.5.3	PCR optimisation	76
2.5.4	DNA product clean-up and quantification	79
2.5.4.1	<i>Spin column purification</i>	79
2.5.4.2	<i>Magnetic bead clean-up</i>	79
2.5.4.3	<i>Qubit quantification</i>	79
2.5.4.4	<i>BioSpectrometer</i>	80
2.5.5	Sequencing	81
2.5.5.1	<i>Oxford Nanopore MinION library preparation</i>	81
2.5.5.2	<i>Whole transcriptome sequencing</i>	85
2.5.5.3	<i>Next-Seq library preparation for INDUCE-Seq</i>	86

2.6	Software packages and programming languages	91
2.6.1	General software packages	91
2.6.2	Custom scripts	91
2.7	Bioinformatics	94
2.7.1	Bioinformatics for nanopore sequencing	94
2.7.1.1	<i>Basecalling</i>	94
2.7.1.2	<i>Quality control and filtering</i>	95
2.7.1.3	<i>Sequence correction</i>	95
2.7.1.4	<i>De-multiplexing</i>	96
2.7.1.5	<i>Sequence alignment</i>	97
2.7.1.6	<i>Alignment, compression and sorting</i>	97
2.7.1.7	<i>Viewing of aligned sequences</i>	97
2.7.1.8	<i>Processing alignments for further analysis</i>	98
2.7.1.9	<i>Variant calling and further analysis</i>	98
2.7.1.10	<i>Sliding window error analysis</i>	99
2.7.2	Transcriptomics	101
2.7.2.1	<i>Trimming and quality control</i>	101
2.7.2.2	<i>Reference indexing and alignment</i>	101
2.7.2.3	<i>Cleaning BAMfiles and marking duplicates</i>	103
2.7.2.4	<i>Generating raw read counts per gene</i>	103
2.7.2.5	<i>RNA-seq data processing in R</i>	103
2.7.3	Data analysis for RNA-Seq	104
2.7.3.1	<i>Sample-to-sample distance measurement</i>	104
2.7.3.2	<i>Gene counts and MA plots</i>	104
2.7.3.3	<i>Differential expression lists and gene ontology</i>	105
2.7.4	INDUCE-seq	106
2.8	Protein analysis	107
2.8.1	Western blotting	107
2.8.2	Immunohistochemistry	108
2.9	Gene editing with inducible Cas9	109
2.9.1	Choice and assembly of gRNA	109
2.9.2	Transfection protocol	110
2.9.2.1	<i>Mature neurons</i>	111
2.9.3	Growth of clonal cell cultures for selection of mutant clones	112
2.9.4	DNA extraction and sequencing library preparation	113
2.9.4.1	<i>Clonal growth of cultures</i>	113

3	SETUP AND TESTING OF AN INDUCIBLE CAS9 GENE EDITING AND NANOPORE SCREENING PIPELINE FOR HIGH THROUGHPUT GENE EDITING EXPERIMENTS	114
3.1	Introduction	114
3.1.1	Creating targeted mutations	114
3.1.1.1	<i>Zinc finger endonucleases and TALENS</i>	115
3.1.1.2	<i>CRISPR-Cas9</i>	116
3.1.1.3	<i>Choice of gene editing approach</i>	121
3.1.2	High throughput analysis of DSB repair outcomes	122
3.1.2.1	<i>Screening overview</i>	122
3.1.2.2	<i>Sanger sequencing</i>	123
3.1.2.3	<i>Next generation sequencing platforms</i>	124
3.1.2.4	<i>Third generation single molecule sequencing platforms</i>	125
3.1.3	Summary and justification for use of nanopore sequencing	129
3.1.4	Aims	130
3.2	Methods	131
3.2.1	Puromycin kill curve	131
3.2.2	Plasmids for inducible Cas9 integration	131
3.2.2.1	<i>GRIN2A mutant library preparation</i>	131
3.2.2.2	<i>Bioinformatic pipeline optimisation</i>	133
3.2.2.3	<i>Running pipeline comparisons</i>	134
3.2.3	Demonstrating the efficacy of inducible Cas9 & nanopore Sequencing pipeline in creating and detecting new mutations	136
3.2.4	Assessment of nanopore error profile	137
3.3	Results	139
3.3.1	Cell line validation	139
3.3.1.1	<i>Demonstration of iCas9 construct insertion at correct Locus</i>	139
3.3.1.2	<i>Demonstration of Cas9 protein production</i>	139
3.3.1.3	<i>Desktop gene editing screen</i>	142
3.3.2	Testing nanopore screening as a testing tool for indels	144
3.3.2.1	<i>Sequencing metrics and barcode decomplexing</i>	144
3.3.2.2	<i>Testing variant screening</i>	146
3.3.3	Subsequent use of systemin creation and description of cell lines	152
3.3.4	Accuracy and output of pipelines using wildtype datasets	155
3.3.5	Sequencing error context	158
3.3.5.1	<i>Sequence composition</i>	158
3.3.5.2	<i>Mononucleotide context of miscalled indels</i>	158
3.3.5.3	<i>Dinucleotide, trinucleotide and tetranucleotide error context</i>	161
3.3.6	Length of errors made in nanopore sequencing	164

3.4	<u>Conclusions and discussion</u>	165
3.4.1	Use of the nanopore for investigation of CRISPR induced mutations	165
3.4.2	Choice of experimental pipeline	167
3.4.3	An idealised Pipeline	169

4	<u>DEVELOPMENT OF A CHD2 MUTANT CELL LINE AND CHARACTERISATION DURING DIFFERENTIATION INTO NEURONS USING RNA-seq</u>	171
4.1	<u>Introduction</u>	171
4.1.1	Criteria for cell line	171
4.1.2	Characterising neurodifferentiation and an introduction to Whole transcriptome sequencing	178
4.1.3	Aims	174
4.2	<u>Methods</u>	175
4.2.1	Set up of <i>CHD2</i> mutant cell line	175
4.2.2	Cell culture and sample collection	176
4.2.3	RNA-Seq library preparation	176
4.2.4	RNA-Seq analyses to be performed	177
	4.2.5.1 <i>Whole dataset analysis</i>	177
	4.2.5.2 <i>Plot counts for neurodifferentiation markers</i>	178
	4.2.5.4 <i>Differential expression lists and gene ontology</i>	178
4.3	<u>Results</u>	179
4.3.1	Confirmation of <i>CHD2</i> mutation	179
	4.3.1.1 <i>Sequencing results</i>	179
	4.3.1.2 <i>Characterisation of mutations</i>	180
4.3.2	RNA-Seq QC	183
4.3.3	RNA-Seq clustering analyses	185
4.3.4	Plot counts for neurodifferentiation markers	188
	4.3.4.1 <i>Summary of transcriptional evidence for differentiation timing</i>	198
4.3.5	Comparisons of transcriptomes at D0, D19 and D40 of differentiation	198
	4.3.5.1 <i>Sample distance, MA plots and PCA plots</i>	204
	4.3.5.2 <i>GO term enrichment</i>	205
	4.3.5.3 <i>Summary of findings from whole transcriptome Analysis</i>	215
4.4	<u>Discussion</u>	
4.4.1	Predicting the impact of our Cas9 induced mutations in <i>CHD2</i>	216
4.4.2	Validating neurodifferentiation with RNA-Seq	218
4.4.3	Insights from whole transcriptome analysis	220
4.4.4	Summary and conclusion	222

5	<u>MODELLING DSB REPAIR IN CHD2 DEFICIENCY USING TARGETED GENOME EDITING</u>	223
5.1	<u>Introduction</u>	223
5.1.1	Overview of experimental approach	223
5.1.2	Aims	226
5.2	<u>Methods</u>	227
5.2.1	Choice of targets and gRNA design	227
5.2.2	Differentiation	227
5.2.3	Gene editing	228
5.2.4	DNA extraction and sequencing	229
5.2.5	Analysis	231
	5.2.5.1 Comparison of DCEs	231
	5.2.5.2 Comparison of Indels	232
5.3	<u>Results</u>	233
5.3.1	Alignment and exploratory analysis of indel counts	233
5.3.2	Double cut excisions	237
	5.3.2.1 <i>A note on the analysis of NRXN1</i>	237
	5.3.2.2 <i>DCE at D0 and D40 of neurodifferentiation</i>	238
5.3.3	Smaller indels at individual gRNA cut sites	242
5.3.4	Transcription at CRISPR targets	246
5.4	<u>Discussion</u>	248
5.4.1	Implications of results	248
	5.4.1.1 <i>DCEs</i>	248
	5.4.1.2 <i>Smaller Indels</i>	248
	5.4.1.3 <i>Relationship between transcriptional activity and mutation detection</i>	249
	5.4.1.4 <i>Potential confounding factors and technical challenges</i>	250
5.4.2	Modelling DCE occurrence	251

6	<u>WHOLE GENOME ASSESMENT OF PHYSIOLOGICAL DOUBLE-STRAND BREAK OCCURRENCE DURING NEURODIFFERENTIATION</u>	256
6.1	<u>Introduction</u>	256
6.1.1	Introduction to INDUCE-Seq	256
6.1.2	Aims	260
6.2	<u>Methods</u>	261
6.2.1	Cell culture and fixation	261
6.2.2	Library preparation	261
6.2.3	Sequencing	261
6.2.4	Analysis	262
6.2.5.1	<i>Genomic context of DSBs</i>	262
6.2.5.2	<i>Enrichment for DSBs captured at histone modification transcription start sites and known fragile sites</i>	262
6.2.5.3	<i>Sequence content at break sites</i>	263
6.2.5.4	<i>Relationship between DSBs and transcription measured by RNA-Seq</i>	265
6.3	<u>Results</u>	266
6.3.1	Quantification and normalisation	266
6.3.1.1	<i>Breaks per run and breaks per cell</i>	266
6.3.1.2	<i>Overlap with protein coding genes</i>	269
6.3.2	Intersection of INDUCE-Seq reads and genomic features	270
6.3.2.1	<i>Relative enrichment for DSBs at PTM histone markers</i>	270
6.3.2.2	<i>Relative enrichment for DSBs at TSS</i>	272
6.3.2.3	<i>Relative enrichment for DSBs at previously described fragile sites</i>	274
6.3.3	Sequence content at break sites	276
6.3.4	Relationship to transcription	278
6.3.4.1	<i>Relationship between FPKM and break count per kb</i>	278
6.3.4.2	<i>Relationship between fold changes in RNA-seq and INDUCE-seq break count</i>	278
6.4	<u>Discussion and conclusions</u>	281
6.4.1	Quantification and localisation of DSBs	281
6.4.2	Enrichment of captured breaks at genomic and epigenomic Features	283
6.4.3	Relationship between DSB count and GC content	287
6.4.4	Integration of RNA-Seq data	288
6.4.5	Summary	290

<u>7</u>	<u>DISCUSSION</u>	292
<u>7.1</u>	<u>Introduction</u>	292
<u>7.2</u>	<u>Summary of investigations and results</u>	293
7.2.1	Setup and testing of an inducible Cas9 gene editing and nanopore sequencing pipeline	293
7.2.2	Creation of a CHD2 mutant and characterisation by RNA-Seq	295
7.2.3	Modelling DSB repair in CHD2 deficiency using targeted genome Editing	296
7.2.4	Whole genome assessment of spontaneous DSB occurrence during differentiation of neurons from induced pluripotent stem cells	298
<u>7.3</u>	<u>Synthesis of results and final conclusions</u>	299
<u>7.4</u>	<u>Impact of findings and suggestions for future research</u>	301
7.4.1	Use of nanopore as a high-throughput screen	301
7.4.2	Understanding of CHD2's role in DSB repair	302
7.4.3	The contribution of DSB to neurodevelopment	303
7.4.4	Patient cell lines	303
<u>7.5</u>	<u>Concluding remarks</u>	304

TABLE OF CONTENTS

APPENDICES	305
APPENDIX I: PYTHON SCRIPTS	305
I CRISPR NANOSCREEN	306
II SLIDING WINDOW ERROR ANALYSIS	313
APPENDIX II: ABBREVIATIONS USED	319
APPENDIX III: SUPPLEMENTARY DATA	323
CHAPTER 3	323
CHAPTER 4	330
CHAPTER 6	336
REFERENCES	339

FULL LIST OF TABLES

1.1	Gene ontology terms of molecular function category directly associated with CHD2	1
1.2	CHD proteins by subclass, including details of identified functions and associated clinical conditions	10
1.3	Selection of histone variants, post translational modifications and associated regulatory functions relevant to this thesis	26
1.4	Types of DNA lesion, type of damage and frequency per cell per day	36
2.1	List of reagents used (3 pages)	57
2.2	List of laboratory consumables used	60
2.3	List of equipment used	61
2.4	Reagent volumes and scaling for plate sizes used in cell culture	62
2.5	Scaling of dilutions for cell culture additives by plate size	67
2.6	Primer sequences used for PCR amplification	78
2.7	Sequences used for first stage of PCR barcoding	83
2.8	Sequences used in two-stage barcoding of amplicons for nanopore sequencing	84
2.9	Buffers required for i-situ library preparation for INDUCE-seq	88
2.10	Software tools used in the analysis of data	91
2.11	Bioinformatic packages used in analysis of genomic data	92
2.12	Python and R packages used in creation of bespoke scripts for data analysis	93
2.13	Primary antibodies used for WB and IF	108
2.14	LiCor secondary antibodies used	108
3.1	Comparison of high-throughput sequencing platforms (2 pages)	127
3.2	Binning of different calls made by the variant screen in optimisation of nanopore sequencing pipeline	135
3.3	P values for comparisons between mean scores for different call types during demultiplexing	150
3.4	New mutations created by iCas9 and detected using nanopore pipeline	152
3.5	total wild-type reads aligned and accuracy of alignments with each tested Pipeline	156
3.6	Sequence composition of each reference amplicon	158
3.7	Mononucleotide context of nanopore sequencing errors by gene	159
3.8	Top ten over-represented dinucleotide sequences upstream of miscalled insertions	163

TABLE OF CONTENTS

4.1	Cut site co-ordinates of gRNA used in <i>CHD2</i> gene editing experiments	175
4.2	List of comparisons made using RNA-Seq data	177
4.3	Transcripts used as markers for different stages of neurodifferentiation	178
4.4	Sample number and condition for RNA extraction	184
4.5	Mean normalised readcounts for expression of neurodifferentiation markers	193
4.6	Number of genes upregulated and downregulated in iCn-CHD2 ^{+/-} cells	204
4.7	Top 20 enriched and depleted GO terms in transcriptomes of iCn-CHD2 ^{+/-} cells at D0 of neurodifferentiation	208
4.8	Top 20 enriched and depleted GO terms in transcriptomes of iCn-CHD2 ^{+/-} cells at D19 of neurodifferentiation	211
4.9	Top 20 enriched and depleted GO terms in transcriptomes of iCn-CHD2 ^{+/-} cells at D40 of neurodifferentiation	214
5.1	gRNAs used to generate DSBs for comparison of repair, and volume of gRNA lipofected	228
5.2	Expected size of DCEs for each gRNA cut site pair	231
5.3	IGV readout and readcount comparison table demonstrating reduction in read depth between cut sites in NRXN1	240
5.4	Comparison of DCEs of predicted sizes found in iCn-WT and iCn-CHD2 ^{+/-} cells with statistical significance values, at D0 of neurodifferentiation	241
5.5	Comparison of DCEs of predicted sizes found in iCn-WT and iCn-CHD2 ^{+/-} cells with statistical significance values, at D40 of neurodifferentiation	241
6.1	Total INDUCE-Seq reads and estimated reads-per-cell for each sample	267
6.2	Relative enrichment for INDUCE-seq reads at histone marker peaks, with between iCn-WT and iCn-CHD2 ^{+/-} cell lines	271
6.3	Relative enrichment for INDUCE-seq reads for distance from TSS, obtained from the Ensembl genome database, with comparison between iCn-WT and iCn-CHD2 ^{+/-} cell lines	273
6.4	Relative enrichment for INDUCE-seq reads for CFS and genes demonstrated to be susceptible to DSB in previous study	275
6.5	Sequence count and GC content of 10bp sequences upstream of DSBs detected by INDUCE-seq for each stage of neurodifferentiation	276
6.6	Pearson coefficients and p-values for relationship between gene transcription levels and DSBs captured per gene by INDUCE-Seq	279
6.7	Regulatory functions of histone markers examined for DSB enrichment	283

FULL LIST OF FIGURES

1.1	Transcripts of the <i>CHD2</i> gene taken from Ensembl genome browser	4
1.2	<i>CHD2</i> protein and <i>CHD2</i> mRNA expression in human tissues	6
1.3	A schematic of human <i>CHD2</i> protein	8
1.4	Schematic representation of histone octamer and nucleosome structure	23
1.5	Flow and filtering of genomic information in line with the central dogma of biology	18
1.6	Choice of DSB pathway and key proteins utilised	49
1.7	Graphical hypothesis	53
2.1	Schematic view of plasmid used to create TET inducible Cas9 insert	72
2.2	Protocol for PCR optimisation and example output	77
2.3	Two stage barcode PCR protocol	83
2.4	Demonstration of sliding-window error analysis	100
2.5	The impact of RNA splicing on reads generated by RNA-Seq	102
3.1	Schematic representation of CRISPR-Cas9 genome editing	119
3.2	Diagram of 96 well plate containing previously sequenced DNA samples	132
3.3	Laboratory and bioinformatic pipeline for nanopore sequencing of Cas9 Induced mutations and variables requiring optimisation	133
3.4	Various permutations of the bioinformatic pipeline used for analysis of wild type sequences	137
3.5	Junction PCR demonstrating insertion of iCas9 construct at AAVS1 locus	139
3.6	Demonstration of Cas9 protein production in response to doxycycline treatment	141
3.7	Cell response to gRNA targeting NANOG	143
3.8	igv readout of data from wells containing <i>GRIN2A</i> deletions, insertions and mixed clones	145
3.9	bar graph demonstrating the average percentage of reads in wells containing Heterozygous mutations stratified by aligner and read depth	148
3.10	Percentage of correctly and incorrectly called reads per well with different Demultiplexing stringencies	149
3.11	Graph demonstrating number of wells available for analysis at different read Depths	151
3.12	Graph comparing the proportion of wells available for analysis by three genetic targets	151
3.13	igv displays of new mutations created using iCas9 and detected using nanopore pipeline	154
3.14	total reads aligned, and accuracy of alignments by each pipeline, aligned with either minimap2 (x) or bwa (o)	157
3.15	Mononucleotide context of nanopore sequencing errors by gene and pipeline	160
3.16	Forward tetranucleotide context of miscalled deletions	162
3.17	Error rate (log scale) by length of miscalled insertion and deletion for nanopore sequencing data for <i>NRXN1</i>	164
3.18	An idealised bioinformatic pipeline for identification of mutant lines from gene editing experiments analysed by nanopore sequencing	170

TABLE OF CONTENTS

4.1	Demonstration of CRISPR-induced CHD2 mutations from first sequencing run	181
4.2	CHD2 heterozygous mutation in sub-cloned cell line	182
4.3	Total reads aligned from each sample number during RNA-seq	184
4.4	PCA and heatmap of sample distance containing all datasets	186
4.5	PCA and heatmap of sample distance, with removal of outlier results	187
4.6.1	Mean and sd of neurodifferentiation markers at D0 of neurodifferentiation	190
4.6.2	Mean and sd of neurodifferentiation markers at D19 of neurodifferentiation	191
4.6.3	Mean and sd of neurodifferentiation markers at D40 of neurodifferentiation	192
4.7.1	Read count comparison for IPS markers at IPS, NPC and neuronal stages of Differentiation	195
4.7.2	Read count comparison for NPC markers at IPS, NPC and neuronal stages of Differentiation	196
4.7.3	Read count comparison for mature neuronal markers at IPS, NPC and neuronal stages of differentiation	197
4.8.1	Sample distance matrix comparing RNA-seq data from iCn-WT and iCn-CHD2+/- cell lines at D0 of neurodifferentiation	199
4.8.2	Sample distance matrix comparing RNA-seq data from iCn-WT and iCn-CHD2+/- cell lines at D19 of neurodifferentiation	200
4.8.3	Sample distance matrix comparing RNA-seq data from iCn-WT and iCn-CHD2+/- cell lines at D40 of neurodifferentiation	201
4.9	MA plots for gene expression comparisons between iCn-WT and iCn-CHD2+/- cell lines at D0, D19 and D40 of neurodifferentiation	202
4.10	Principal component analysis comparing RNA-seq sample distance comparisons between iCn-WT and iCn-CHD2+/-	203
4.11	GO terms enriched in genes downregulated in iCn-CHD2+/- cells at D0 of neurodifferentiation	206
4.12	GO terms enriched in genes upregulated in iCn-CHD2+/- cells at D0 of Neurodifferentiation	207
4.13	GO terms enriched in genes downregulated in iCn-CHD2+/- cells at D19 of neurodifferentiation	209
4.14	GO terms enriched in genes upregulated in iCn-CHD2+/- cells at D19 of neurodifferentiation	210
4.15	GO terms enriched in genes downregulated in iCn-CHD2+/- cells at D40 of neurodifferentiation	212
4.16	GO terms enriched in genes upregulated in iCn-CHD2+/- cells at D40 of Neurodifferentiation	213
4.17	variants found in human population studies which fall within the same domain affected by the in-frame heterozygous variant identified in CHD2 gene-editing experiment	217

TABLE OF CONTENTS

5.1	Schematic representation of DCE (top) and indels (bottom) expected to be found when transfecting multiple gRNA against the same region of DNA.	225
5.2	Results from previous demultiplexing experiments with number of reads successfully demultiplexed as a ratio to the number of reads demultiplexed for all experiments	230
5.3	Workflow for experimental comparison of DSB repair	230
5.4	Scatter charts demonstrating indel counts by size of for four genomic targets in iCn-WT, iCn-CHD2 ^{+/-}) and untreated control cells at D0 of neurodifferentiation	235
5.5	Scatter charts demonstrating indel counts by size of for four genomic targets in iCn-WT, iCn-CHD2 ^{+/-}) and untreated control cells at D40 of neurodifferentiation	236
5.6	Line graphs demonstrating depths of expected DCE +/- 10nt in mutation length at D0 and D40 of neurodifferentiation	239
5.7	IGV readout demonstrating reduction in read depth between cut sites in NRXN1 at D0 and D40 of neurodifferentiation (see also <i>table 5.3</i>)	240
5.8	Comparison of indels within 10bp of a gRNA cut-site in CSMD3 (5.8.1), NRXN1 (5.8.2) and PARK2 (5.8.3) demonstrating depth of indels of different lengths occurring within	243
5.9	Data from RNA-seq experiment demonstrating transcript readcount of each CRISPR target gene at D0 and D40 of neurodifferentiation	247
5.10	Possible versions of equilibrium between breakage, repair and dissociation at DSBs during double-cut experiments	252
5.11	Schematic representation of the equation described in <i>figure 5.10</i>	252
5.12	Schematic description of model used to describe equation in <i>figure 5.10</i>	253
5.13	Computer modelling of maximum time of double cut persistence based on version A of the equilibrium described in <i>figure 5.12</i>	254
6.1	Example stingray plot made from two datasets of 1M randomly generated breakpoints	264
6.2	Estimated INDUCE-Seq reads per cell plotted for iCn-WT and iCn-CHD2 ^{+/-} cells at D0, D19 and D40 of neurodifferentiation	267
6.3	Bedgraphs demonstrating break points determined by INDUCE-seq read start co-ordinates in chromosome 4	268
6.4	Breaks per gene determined by INDUCE-Seq for all protein coding transcripts	269
6.5	Relative enrichment for INDUCE-Seq reads at histone modifications obtained from the INDUCE-Seq database	270
6.6	Relative enrichment for INDUCE-seq reads at 500bp, 1kb and 2kb from transcription start sites, taken from ensembl genome database	272
6.7	Relative enrichment for INDUCE-seq reads at previously identified fragile sites	274
6.8	Stingray plots demonstrating relative representation of and GC content of shared and unique 10bp sequences at break sites	277
6.9	Normalised break count per Kb per gene derived from INDUCE-seq read counts	278
6.10	Log2-fold change for breaks per kb per gene, plotted against Log2fold change per gene from RNA-Seq data.	280
6.11	Schematic of potential for false negative enrichment calls in CHD2 deficient cell line, in the event of an altered histone PTM landscape	284

7.1	Theorised links between CHD2 function, transcription, abnormal DSB repair and profile of DSBs that occur in differentiating neurons	300
-----	---	-----

1: INTRODUCTION

Mutations in the chromodomain Helicase DNA-binding 2 (*CHD2*) gene have been identified as a cause of a severe neurodevelopmental disorder in humans categorised as epileptic encephalopathy of childhood onset (EECO) [1, 2]. Patients with heterozygous mutations in *CHD2* experience seizures, which can be intractable from a young age. The syndrome is associated with learning difficulties, autistic spectrum disorder and an increased risk of neuropsychiatric disorders. How mutations in the *CHD2* gene give rise to this disease phenotype is not currently understood.

CHD2 codes for a chromatin remodelling protein of the same name. Chromatin is the protein which forms the backbone of chromosomes. The eukaryotic genome is wound around this biologically active structure, which is understood to regulate genomic function in a variety of diverse ways. Indeed, the 3D arrangement of chromatin within cells can be thought of as vital for determining the cell type.

The gene ontology database associates *CHD2* with several molecular functions including: transcription, DNA damage repair and muscle differentiation (*Table 1.1*). Further to this, there is published evidence associating *CHD2* with regulation of transcription [3], neurogenesis [4], myogenesis [5], and the maintenance of genome integrity[6].

GO term	GO reference
RNA polymerase II proximal promoter sequence-specific DNA binding	GO_REF:0000024
Chromatin organisation	GO_REF:0000037
Cellular response to DNA damage stimulus	GO_REF:0000107
DNA duplex unwinding	GO_REF:0000108
Muscle organ development	GO_REF:0000024
Histone binding	GO_REF:0000024
Haematopoietic stem cell development	GO_REF:0000107

Table 1.1 Gene ontology (GO) terms of molecular function category directly associated with CHD2

In this thesis, I will explore the link between *CHD2*'s contribution towards genome maintenance via the repair of DSBs and *CHD2*-related EECO, caused by heterozygous mutations the *CHD2* gene.

I will begin, in this introductory chapter, by reviewing the available data available regarding the structure and expression of CHD2 protein, highlighting similarities and differences to other proteins within the CHD family.

I will review the clinical features of *CHD2*-related EECO, based on previously published cases reports, larger case cohort studies and our own unpublished data. I will then also review the phenotypes exhibited in previously published animal models of CHD2 deficiency.

I will outline the structure and function of chromatin, including chromatin remodelling and modification. I will then explore the published data from cell models of CHD2 depletion, with particular focus on its role as a chromatin remodeller with the potential to impact the transcription of genes throughout the genome through modification of histone proteins and the recent work linking CHD2 activity to DSB repair.

I will review the mechanisms by which DSBs occur and the pathways by which they are repaired. I will explore the theoretical pathway by which impaired DSB repair in a developing foetus could lead to a large phenotypic impact on the mature CNS.

This will form the basis of my hypothesis, in which I will detail a prediction that *CHD2* mutations can have an impact on genome repair in the developing central nervous system (CNS), that this can be modelled under laboratory conditions, and that this dysregulation of genome repair could contribute to the neurodevelopmental phenotype exhibited by these patients.

I will conclude with an overview of the approach I used to explore my hypothesis, although a greater depth of technical detail will be provided in the methods and materials section (chapter 2) and the introductory and methods sections of each results chapter (chapters 3 through 7).

1: Introduction

1.1: The CHD2 gene and protein

1.1 THE CHD2 GENE AND PROTEIN

The CHD2 gene is found on chromosome 15q26.1, with genomic coordinates 15:92,900,320-93,028,006 based on the most recent build of the human genome (GRCh38). It is known to be under selective constraint – the gnomAD dataset contains 523 observed missense variants, where 982 would be expected generating an o/e score of 0.53 (95% CI 0.49-0.57, Z score 5.21) indicating a high level of constraint. Similarly, 3 loss of function (LoF) variants are seen, where 111 are predicted, giving an o/e score of 0.01-0.07. These data confirm that missense variants and particularly LoF variants are under evolutionary pressure and are selected against in the general population.

The canonical transcript (RefSeq reference #: NM_001271.4, Ensembl genome browser #: ENSG00000173575) [7] is 9350bp in length and produces a protein of 1828 amino acids. At the time of writing, the Ensembl genome browser lists 29 transcripts, of which 11 are protein coding, 9 contain no open reading-frame, 6 undergo nonsense-mediated decay and 3 retain introns (*figure 1.1*). The majority of the protein coding transcripts overlap with the first half of the canonical transcript, however there are three smaller protein coding transcripts that do not. There is no single protein-coding exon retained in all protein-coding isoforms (*figure 1.1*)[8].

1: Introduction

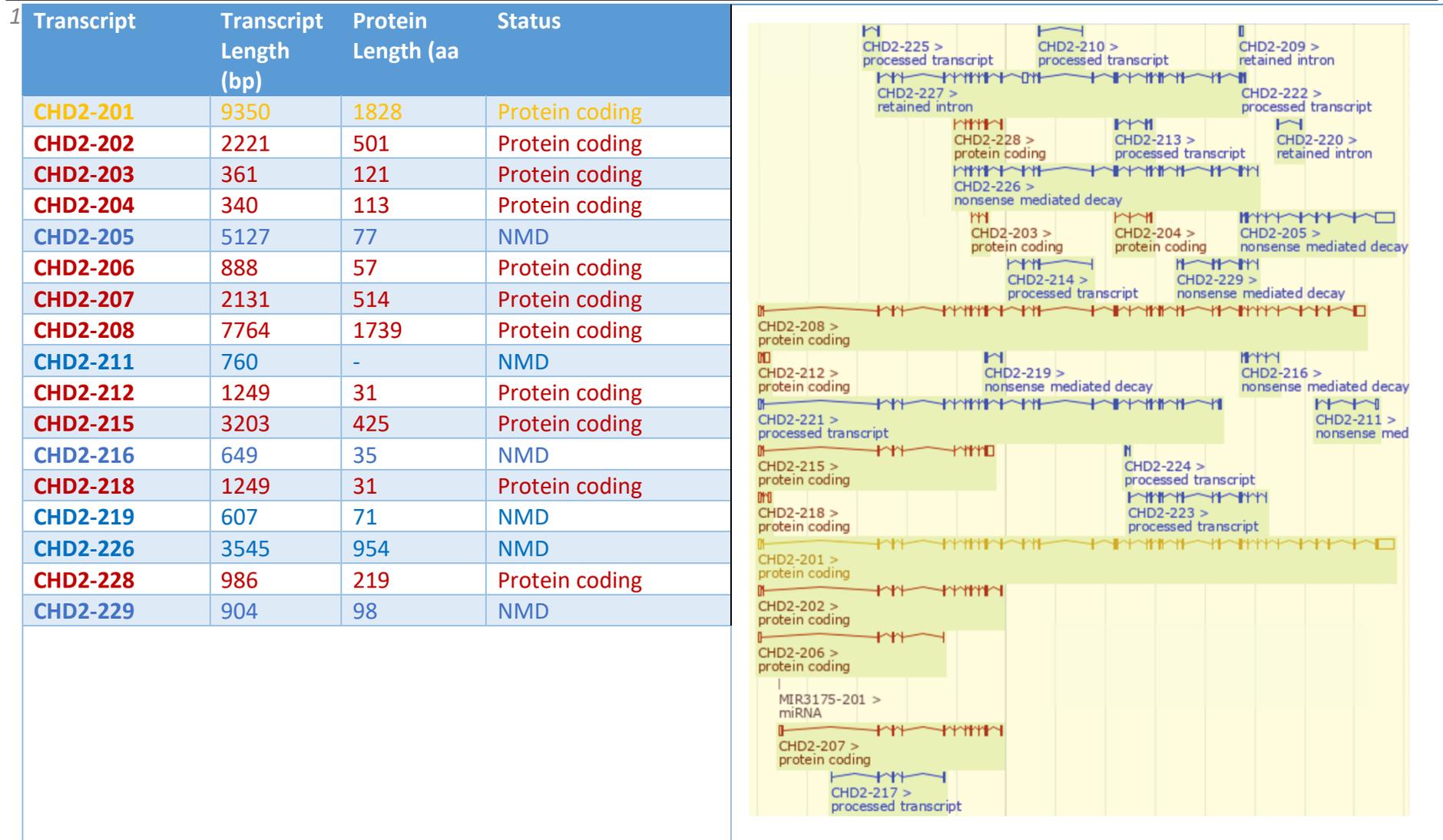


Figure 1.1: Transcripts of the CHD2 gene taken from Ensembl genome browser – Protein Coding transcripts are marked in blue, transcripts with an open reading frame predicted to undergo nonsense mediated decay (NMD) are in red, transcripts with a retained intron (209, 220), or no open reading frame (210, 213, 214, 217, 220-225, 227) which cannot produce a protein are not shown

1: Introduction

1.1: The CHD2 gene and protein

1.1.1 CHD2 Expression

Expression data from the human protein atlas demonstrates expression of mRNA in all tested tissues, and expression of protein [9] in all tissues except skeletal muscle. Protein expression levels are ranked as high in the cerebral cortex, hippocampus, cerebellum, as well as male and female reproductive organs, with all other tissue types ranked as medium or low. Staining for the CHD2 protein is restricted to the nucleus (see *figure 1.2* – image credit human protein atlas)[10].

Tracking of CHD2 expression in mouse embryos demonstrated widespread expression in the early stages of embryogenesis and persistently high levels of expression in the developing brain. Postnatally, expression was preserved in the cerebellum, hippocampus and neocortex, in keeping with the human data from the human protein atlas [11]

Of course, it is not possible to conclude that expression patterns in the developing human brain will be identical to those seen in the mouse, however the fact that the post-natal expression patterns are so similar and the known importance of CHD2 in neurodevelopment suggests that cautious optimism as to the relevance of this data is appropriate.

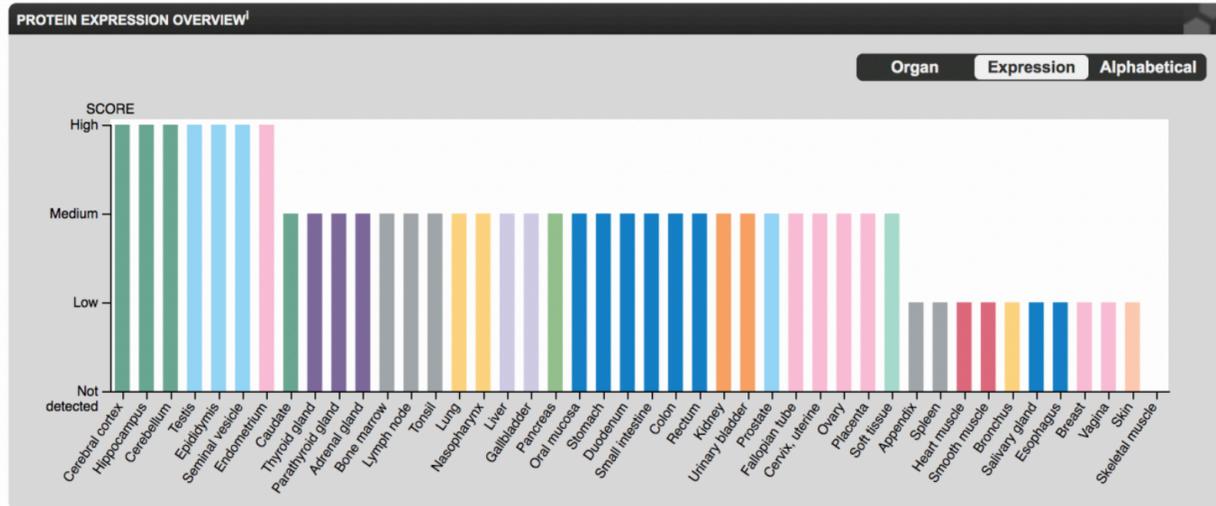
The Allen Brain Atlas [12] dataset indicates that CHD2 messenger RNA (mRNA) expression peaks in the developing human neocortex at 9 weeks post-conception, but that expression is maintained throughout development. Again, high levels of postnatal expression are identified in the cerebellum[12].

Interestingly, despite being widely expressed, the phenotype seen in humans with CHD2 mutations appears to be confined to the central nervous system (CNS). This could indicate a particularly important role in the development and function of the brain, or that its function is redundant or duplicated in other human tissues.

1: Introduction

1.1: The CHD2 gene and protein

A



B

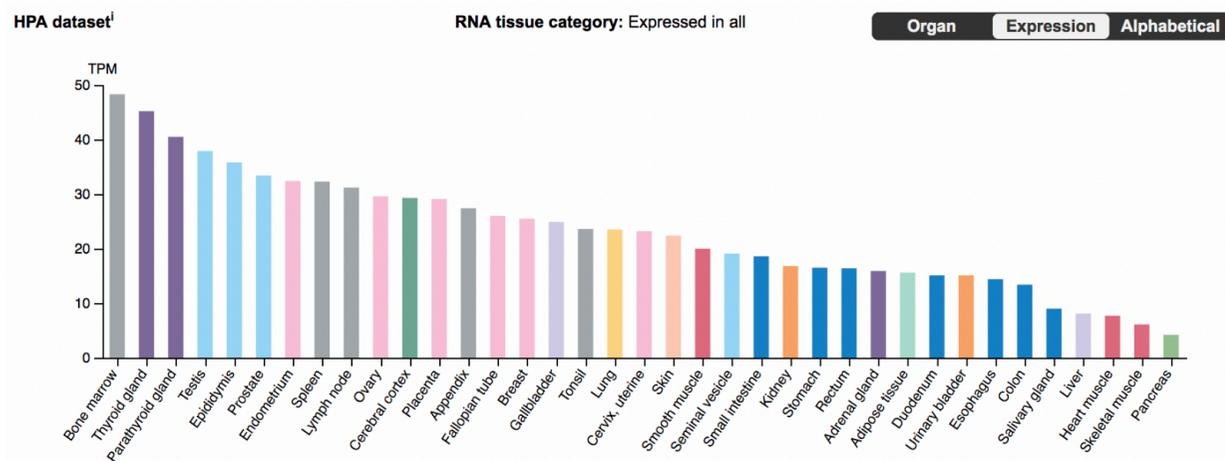


Figure 1.2:
 A – CHD2 protein expression
 B – CHD2 mRNA expression in human tissues, image credit Human Protein Atlas, used under Creative Commons Attribution-ShareAlike 3.0 International License

1: Introduction

1.1: The CHD2 gene and protein

1.1.2 CHD2 Protein Structure

The canonical human CHD2 transcript codes for a 1828 amino acid (aa) length protein [8], with a weight of 211.3 kDa[13]. Putative functional domains include two chromodomains, two helicase domains and a DNA binding region (*figure 1.3*). The C-terminus also includes a poly-A-Ribose (PAR) binding domain.

CHD2 is a member of the SNF2-like family of helicase-related enzymes, which includes all known ATP-dependant chromatin remodelling factors. The CHD subfamily is characterised by the presence of dual chromodomains N-terminal to the SNF2 domain. Liu described the structure of CHD2 as: [14] an ATPase domain at aa 450-1129, chromodomain at aa 1-450 and DNA binding domain at aa 1129-1828[14].

Luijsterburg et al [6] delineate the structure further describing the structure as; chromodomains at aa 1-461, ATPase/helicase domain at 462-951, putative SANT-SLIDE motif at aa 462-951, further describing a PARP1-binding site at 1392-1610 and PAR binding site at 1611-1828.

The SANT-SLIDE domain has been previously demonstrated as part of the DNA binding region in chromatin remodellers, including in the *saccharomyces cerevisiae* homologue of CHD1 and the human isoform of CHD7. Its function is not completely elucidated, but there is evidence that it is required for binding of nucleosomes and plays a role in the remodelling of nucleosomes; proteins with mutations in this domain exhibit normal ATP turnover and nucleosome binding, but defects in the movement of nucleosomes[15].

The function of the PAR and PARP1 domains is explored further in sections 1.5.3 and 1.6.2.

Amongst the missense variants present in gnomAD[16] (*figure 1.3*) there is a bias towards the C-terminus of the protein – it is also worth noting that the helicase domains and DNA-binding domains are depleted for missense variants in this population database, indicating a particular intolerance of variation in these regions of the protein.

1: Introduction

1.1: The CHD2 gene and protein

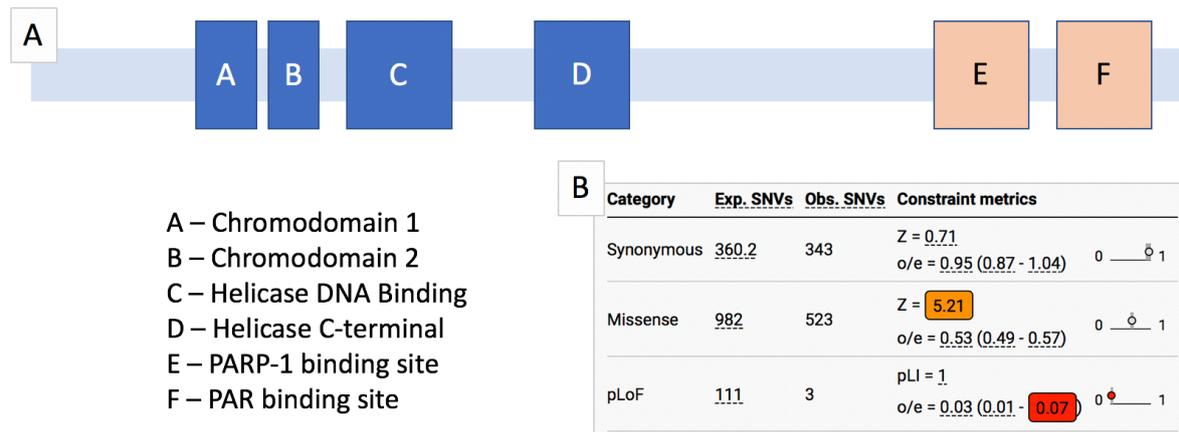


Figure 1.3 – A: schematic of human CHD2 protein, with dark blue domains being well established and described in the UniProt database, and peach domains being suggested by Luijsterburg et al. B: evolutionary constraint metrics for missense variants from the gnomAD database demonstrating intolerance of missense and putative loss of function (pLoF) variation. For missense mutations, a Z-score of above 3.0 indicates that missenses are significantly depleted in the gene represented compared to the average rate of missenses across the protein coding genome. pLoF scores approaching 1 indicate a higher likelihood of intolerance of LoF mutations humans. This data indicates that both missense mutations and loss of function mutations are depleted in CHD2 in healthy controls in the human population, compared to the average rate with which they occur throughout the protein coding genome

1: Introduction

1.1: The CHD2 gene and protein

1.1.3 Comparison to other CHD2 family proteins

Nine Chromodomain Helicase DNA Binding (CHD), and one CHD-like (CHD1L), proteins have been described in *Homo sapiens* to date[1]. They are characterised by tandem chromodomains on the N-terminal side of their ATPase domain[17]. They are divided into subclasses based on shared protein domains and DNA sequence similarities. CHD2 is a member of subclass I of the CHD family. (*table 1.2*).

Thus far, mutations have been linked with disease in CHD1, 2, 3, 4, 7 and 8. Abnormal neurodevelopment is a feature in all these conditions, ranging from mild developmental delay and speech apraxia seen in mutations of CHD3 to potentially profound developmental delays seen in CHD2 mutations.

1: Introduction

1.1: The CHD2 gene and protein

<i>Protein</i>	<i>Sub-class</i>	<i>Features</i>	<i>Identified functions</i>	<i>Clinical conditions associated with germLine mutation (inheritance)</i>	<i>Features of syndrome</i>
CHD1	I	Subfamily specific chromodomains	Maintenance of euchromatin Binds at promoters of actively transcribed regions Remodelling during DSB repair by homologous recombination (HR) [18]	Pilarowski-Bjornsson syndrome[19] (AD)	Speech apraxia, autisti spectrum disorder (ASD)
CHD2	I	Subfamily specific chromodomains	Binds at promoters of actively transcribed regions Remodelling during DSB repair by NHEJ Myogenesis	EECO – heterozygous mutations (AD)	Speech delay, global developmental delay (GDD), Epilepsy ASD Increased risk of mental health disorders
CHD3	II	paired PHD Zn-finger-like domains	Co-regulation of gene expression with sumo-protease SENP1 [20] Chromatin remodelling during HR[205]	Snijders Blok-Campeau syndrome [21] (AD)	Speech apraxia, macrocephaly, Impaired speech, characteristic facial features, joint laxity, undescended testes

Table 1.2 (1/3): CHD proteins by subclass, including details about identified functions and associated clinical conditions

1: Introduction

1.1: The CHD2 gene and protein

<i>Protein</i>	<i>Sub-class</i>	<i>Features</i>	<i>Identified functions</i>	<i>Clinical conditions associated with germLine mutation (inheritance)</i>	<i>Features of syndrome</i>
CHD4	II	paired PHD Zn-finger-like domains	Histone remodelling as component of NuRD DNA damage response [22] via parylation – phosphorylated by ATM Chromatin remodelling during HR[205]	Sifrim-Hitz-Weiss syndrome (AD) [23]	Intellectual disability (ID), hearing loss, distinctive facial features
CHD5	III	Subfamily specific chromodomains	Tumour suppressor – implicated in p53 signalling, apoptosis, senescence[24] Role in neurogenesis, linked to polycomb repression and gene expression[25]	Possible association with peri-ventricular nodular heterotopia Expression altered in cancer – suppression leads to increased H2AX accumulation and poor prognosis. [26]	
CHD6	III	Subfamily specific chromodomains	Oxidative DNA damage repair (not DSB) [27]	None yet identified	n/a

Table 1.2 (2/3): CHD proteins by subclass, including details about identified functions and associated clinical conditions

1: Introduction

1.1: The CHD2 gene and protein

Protein	Sub-class	Features	Identified functions	Clinical conditions associated with germLine mutation (inheritance)	Features of syndrome
CHD7	III	Subfamily specific chromodomains	Evidence that CHD7 supresses p53 activity during development[28] Interacts with SOX2 during neurodevelopment[29]	CHARGE syndrome – heterozygous mutations[30, 31]	Developmental delay, cardiac defects, conductive hearing loss, oesophageal atresia, choanal atresia, coloboma, genitourinary abnormalities, growth restriction
CHD8	III	Subfamily specific chromodomains	Interacts with CHD7 [32] Mutations affect CpG island methylation at CTCF sites [33]	Risk factor for ASD– heterozygous mutations[34, 35]	Autistic spectrum disorder, macrocephaly, gastrointestinal problems
CHD9	III	Subfamily specific chromodomains	Putative role in rRNA biogenesis[36]	None yet identified	n/a
CHD1L	()	No chromodomains – carboxyl domain binds to PARP chains	Required in pre-implantation embryo for blastocyst formation (mice) [37] Binds to PAR chains during first stage of DSB response[205]	Part of critical deleted region in 1q21.1 deletion syndrome [38]	Developmental delay, microcephaly, cardiac abnormalities, cataracts [38]

Table 1.2 (3/3): CHD proteins by subclass, including details about identified functions and associated clinical conditions

1.2 CLINICAL RELEVANCE OF *CHD2* MUTATIONS IN HUMANS

Epilepsy is one of the most common neurological disorders, with lifetime incidence estimated to be close to 3%. In patients with epilepsy, seizures often coexist with a number of associated neurodevelopmental disorders including; intellectual disability (ID), specific learning disabilities (LD), behavioural difficulties, autistic spectrum disorder (ASD) and an increased risk of psychiatric illnesses including depression, mania and schizophrenia[39]. There is considerable overlap between the genes in which rare heterozygous mutations are known to cause epilepsy syndromes and those in which common variants are found to contribute towards risk for schizophrenia on genome wide association studies (GWAS)[40].

Epileptic encephalopathies (EE) are a group of conditions in which developmental delay or regression exists in parallel with persistent epileptic seizures. The International League Against Epilepsy (ILAE) defines EE as a condition in which, “the epileptic activity itself may contribute to cognitive and behavioural impairments beyond what might be expected from the underlying pathology alone”[41]. In many such disorders, the seizures can be refractory to treatment with multiple combinations of anticonvulsants. In others, developmental dyscrasia may be the only outward sign of the relentless electroencephalographic (EEG) changes[42], [43, 44].

Epileptic encephalopathy can be further classified into acquired and idiopathic. Acquired EE develops following a significant neurological insult such as: head injury, hypoxic brain injury, encephalitic infection, or during the progression of an untreatable neurodegenerative disease such as a neuro-metabolic disorder. Idiopathic cases, that is cases where no obvious CNS insult occurred that could explain the diagnosis, are now generally thought to be genetic in origin[45, 46].

It can be challenging to determine the relative contributions of underlying defect, high dose medication and aberrant electrical activity to the developmental phenotype and so ILAE have more recently proposed the new category of Developmental Epileptic Encephalopathies (DEE). As developmental delay in patients with *CHD2* mutations can exist regardless of seizure activities, it is likely that both the seizures and the ID are sequelae of the same underlying phenomenon, rather than the seizures *causing* the ID and therefore DEE may be a better description for this clinical phenotype.

1: Introduction

1.2: Clinical relevance of CHD2 mutations

With the advent of high-throughput sequencing (HTS) techniques (also referred to as next-generation-sequencing (NGS)) as a tool for research into and latterly clinical diagnosis of patients with EE/DEE, finding a genetic diagnosis is becoming possible for a greater number of these patients.

There is often considerable overlap in the clinical presentation of patients with EECO, and therefore two primary approaches to reaching a genetic diagnosis are used; panel testing and whole exome sequencing with gene filtering by bioinformatic panel. For example, the Molecular Genetics Laboratory at Great Ormond Street Hospital currently offers an 82 gene panel (including *CHD2*) for diagnosis of early-infantile epileptic encephalopathy[47]. The 100,000 genomes' panel app project aims to provide up-to-date and peer reviewed panels for filtering the results of whole exome sequencing (WES) in search of specific diagnoses – their bioinformatic panels for syndromic epilepsy and genetic epilepsy syndromes recommend analysis of 423 genes confirmed to be associated with epilepsy syndromes after peer review[48].

1: Introduction

1.2: Clinical relevance of *CHD2* mutations

1.2.1 Clinical features of *CHD2* related EECO

A clinical syndrome associated with *CHD2* haploinsufficiency was first described in patients with larger deletions of 15q26.1-2 [49]. Veredice et al described a deletion involving 56 genes, associated with early onset myoclonic epilepsy. Subsequent case reports narrowed the potential critical region to two genes, *RGMA* and *CHD2*[50, 51] [52]. This allowed for *CHD2* to be identified as a potential target for large sequencing studies that confirmed its association with epileptic encephalopathy.

In one such study targeted next-generation resequencing of 65 genes in 500 patients with epileptic encephalopathy with no identified aetiology was performed. Carvill et al (2013) [42] found *CHD2* mutations in six patients – the highest of any ‘candidate gene’ included in the panel. As a point of comparison, there were eight individuals with Dravet syndrome, which had been well established as linked to *SCN1A* by the time of publication.

At the time of writing, there were 17 published case reports of point mutations in *CHD2* [42, 53] and 11 published cases of patients with deletions [49, 51, 52, 54] including some or all of the gene. The most common clinical features include ID and seizures. The seizure types described include: generalised tonic-clonic seizures (GTCS), atonic seizures, absences, myoclonic seizures, eyelid myoclonus and photosensitivity. A compulsion towards photic self-induction of seizures is an unusual feature [55] which, although not present in all cases, can provide a savvy clinician with a clue towards the diagnosis. One patient described the compulsion; “the sun makes me look at it”.

Mental health is also an emerging concern in this cohort. Two recent case reports describe psychotic features in patients. In one there was a history of brief, transient psychoses[56]. The other describes both negative and positive features of psychosis, although not sufficient for a diagnosis of schizophrenia[57].

Further to this, there were an additional 13 mutations described without significant case detail provided. These include variants associated with epileptic encephalopathy[58-62], identified as a risk factor for photosensitivity in epilepsy, and cases in which autistic spectrum disorder (ASD) without epilepsy was the presenting feature[63-66]. Copy number variants in which *CHD2* is deleted have also been identified in patients with simplex autism[67].

1: Introduction

1.2: Clinical relevance of CHD2 mutations

As is often the case, the first patient reports published described patients at the more severe end of the spectrum of phenotypes, however it is now increasingly understood that there is significant patient to patient variability. The reason for this variability is not known, however variable expressivity and penetrance are frequently features of autosomal dominant genetic disorders. The determinants of severity are likely to be multifactorial, including both cryptic polygenic risk caused by other inherited SNPs, and environmental or lifestyle factors[23].

There are 11 cases with confirmed pathogenic mutations described in the DECIPHER database [68, 69] which includes data from a cohort 13,600 patients enrolled in the Deciphering Developmental Disorders (DDD) study [70], and has expanded to include data about mutations discovered in the setting of Clinical Genetics in the UK. DECIPHER requests various biometrics as part of the phenotype data required for submission, including height weight and occipital-frontal circumference (OFC). The cumulative frequency curves for HC place the range for patients with *CHD2* mutations at -4 to 0 standard deviations from the normal population, whereas height, birthweight and weight at time of assessment roughly follow the expected distribution.

1: Introduction

1.2: Clinical relevance of *CHD2* mutations

1.2.2 *CHD2* mutations in cancer

CHD2 mutations are one of the most common findings on exome sequencing of tumour tissue in chronic lymphocytic leukaemia (CLL) (5.3% of cases) and monoclonal B lymphocytosis (MBL) (7% of cases) patients. The mutations were either frameshift, nonsense or occurred in highly conserved protein domains, and are predicted to be deleterious using CONDEL, a method combining 5 *in silico* analysis pathways to provide a prediction of deleteriousness[71]. Rodrigues et al 2015 suggest that it may be an oncogene and function as a driver mutation in CLL.

There has been a recent report of an association between *CHD2* mutations and bone marrow dysfunction, although not malignancy[72]. The authors described a cohort of ten patients with syndromic neutropenia, defined as a persistent low white cell count associated with an increased risk of infections *and* other clinical features. One patient presented with recurrent peri-anal abscess, intermittent neutropenia and a background of epilepsy and developmental delay. The exome sequencing revealed a frameshift mutation in *CHD2*, c.5094dupC p.(Pro1699Alafs*3).

1.2.3 Summary of clinical relevance

The primary focus of research thus far into the effects of *CHD2* mutations in humans has focussed on its role in neurodevelopment. Mutations were originally described in patients with epileptic encephalopathy, however more recent work has identified *CHD2* mutations as a potential cause of simplex autism, and as a risk factor for photosensitivity in otherwise simple epilepsy.

Recent biometric data also indicates that patients with *CHD2* mutations have a lower-than average head circumference, despite being of average height and weight. MRI data in some patients from our cohort suggests that cerebral atrophy may also be an associated feature.

More recently, *CHD2* has been associated with haematological conditions, emerging as a driver mutation and possible prognostic indicator in CLL. A single patient has also been identified with cyclic neutropenia associated with a *CHD2* mutation – it remains to be seen whether this association is genuine and further work will need to be done reviewing existing cases to discover if this is a recurrent feature.

1.3 ANIMAL MODELS OF CHD2 DEFICIENCY

The gene cluster containing *CHD2* is highly conserved and appears in mice, opossum, and zebrafish. The human CHD2 protein has 96.5% homology to the mouse version[1].

The earliest mouse model, generated using a gene trap approach, demonstrated an increase in perinatal lethality, a proliferation of non-neoplastic lesions throughout the body and an increase in renal abnormalities, but no neurological phenotype[73, 74].

In one study, a gene-trap approach was used to create heterozygous mice, which were then bred to produce homozygous mice. The homozygous mice were shown to have multiple in-utero haemorrhages, attributed by the authors to disordered haematopoiesis[75]. There was also increased perinatal lethality in the heterozygous mice. Those that survived had a shorter life expectancy, attributed to an increased rate of lymphoma. The majority (14/21) had succumbed to lymphoma by 26 weeks of age, compared to 1/6 of the wild type mice used for comparison[75].

A comet assay involves performing an electrophoresis on cells after they have been treated with a damaging agent, and before the DNA breaks have had a chance to heal. The resulting 'tail' structure seen on the gel can be measured to provide a quantitative assessment of DNA damage. In a cell model derived from mouse embryonic fibroblast (MEF) cells, comet assay demonstrated that *CHD2* mutant cells exhibit an increased sensitivity to ultraviolet (UV) and X-ray irradiation[76].

Unlike the mice which have yet to develop an outwardly detectable neurological phenotype despite displaying changes in neuronal migration, the zebrafish model, created by morpholino antisense oligonucleotides, exhibits seizures and photosensitivity. The zebrafish also exhibit absent swim bladder and curvature of the spine[77].

1.3.1 Discrepancies between animal and human phenotypes

Considering the high level of homology between mouse and human proteins, it is surprising that the mice do not exhibit a neurodevelopmental phenotype. It is also perhaps surprising, given its implied role in DSB repair (see section 1.6) that no cancer-risk phenotype has been identified in humans thus far. There are several possible explanations for these discrepancies.

1: Introduction

1.3: Animal models of CHD2 deficiency

First, although it is somewhat obvious to state, humans *are not* mice. There may be redundancies in the human CHD system that do not exist in the mouse and vice-versa, which allow one organism to compensate for certain cellular functions, but not others.

Secondly, the methods used to create the mutations in the mouse and zebrafish models are not 100% specific and without whole exome, or whole transcriptome analysis of these organisms, it cannot be ascertained with certainty that they do not harbour any additional mutations that could explain the phenotypes.

It is also worth noting that the earlier mouse models exhibit congenital abnormalities affecting the heart and kidneys (also not seen in the human phenotype) [73, 74], whereas later models exhibit a phenotype related to impaired genome maintenance[75]. It is difficult to know what causes the differences between these models and therefore both must be treated with a degree of scepticism, although the subsequent cellular studies support an impairment of DSB repair[76].

Finally, it is possible that aspects of the phenotype have been missed in both mouse and human cases. As far as can be ascertained from the published literature, *chd2* mutant mice have never been subjected to photo-stimulation in an attempt to mimic the human photosensitivity phenotype, nor have behavioural studies that may elicit a more subtle neurobehavioral phenotype been conducted in CHD2 mutant mice.

Similarly, from the published evidence regarding human patients with confirmed pathogenic CHD2 mutations, it is not clear that regular Full Blood Count (FBC) monitoring that could detect a dyscrasia of haematopoiesis are being conducted. The evidence available is probably not of the quality to suggest the institution of such a screening program. It is also worth noting that the patients described thus far are young – if a cohort of women with *BRCA1/2* mutations were only assessed up until 25 years of age, it would be difficult to appreciate an increased risk of malignancy, however if the same cohort were followed until the age of 40, the increased risk would become more obvious.

It is also worth noting that other CHD proteins have been implicated in various DNA repair pathways, including CHD1, CHD1L[205], CHD3[205], CHD4[205][22], CHD5[78] and CHD7[79]. There is even evidence that the phenotype in patients with CHARGE syndrome caused by CHD7 mutation is related to dysregulation of the cell cycle checkpoint p53[28] – an essential protein in the management of genome maintenance[80]. Furthermore, there is evidence that CHD8, implicated in an autosomal dominant form of simplex autism interacts

1: Introduction

1.3: Animal models of CHD2 deficiency

with CHD7 [32], identifying a role interacting with DNA repair pathways for at least CHD2, 4, 5, 7 and 8.

In none of these examples are germline mutations associated with an increased cancer risk. It is therefore reasonable to state that a specific cancer-risk phenotype is *not* necessarily a cardinal feature of germline mutations in proteins which contribute to the function of DNA repair pathways.

1.4 CELL MODELS OF CHD2 FUNCTION

1.4.1 Cell models of neurodevelopment

It is helpful to provide a brief summary of embryonic neurodevelopment in order to frame the findings of previous studies into the role of CHD2 in CNS development. Homem et al provide an excellent summary [81], from which the following paragraph is adapted.

In the developing mammalian cortex, neuroepithelial cells expressing glial cell markers form the most apical layer of the developing cortex – the ventricular zone (VZ). The expression of these glial cell markers defines the cells as Radial Glial (RG). RG cells divide asymmetrically to one further RG, to regenerate the pool of progenitor cells, and one cell that becomes either a post-mitotic neuron or intermediate progenitor cell (IP). In turn, IPs typically divide once more to form a pair of cortical neurons. A second population of RG cells, the outer radial glial (oRG) cells also arise from RG divisions and continue to divide and produce IPCs. Further complexity has been identified in the form of multiple basal neural progenitors, arising in the subventricular zone.

In an attempt to explore the link between *CHD2* mutations and the human neuropsychiatric phenotype, another group used short hairpin RNAs to knockdown CHD2 expression in mice and took sections of brain at different stages of development. In their model, *chd2* suppression reduced the proliferating pool of RG neural progenitor cells (NPCs) and led to an increase in the generation of IPs. They suggested that this impact on the renewal of NPCs could affect cortical development and might contribute towards a phenotype of abnormal neurodevelopment[82].

Further to this, they demonstrated a relationship between CHD2 expression and REST expression. REST acts as a transcription factor, suppressing the expression of neuronal genes. The findings suggested that CHD2 functions to suppress REST expression, which was offered as a possible explanation for the impact on the impact on RG NPC pool maintenance; reduced CHD2 levels cause an increase in REST, which dysregulates the transcriptional network responsible for maintaining the pluripotent state[82].

Whereas murine models of neurodevelopment can be developed through direct observation and analysis of foetal brain tissue, there are obvious ethical constraints on such research in developing human embryos. Instead, inferences must be made from

1: Introduction

1.4: Cell models of CHD2 function

differentiation of human Embryonic Stem Cells (hESC) or human Induced Pluripotent Stem Cells (hiPSC). A detailed protocol for this procedure can be found in *section 2.1.2*.

CHD2 has been identified as a target of the transcription factor NKX2-1, which itself has been demonstrated to control cell-type specification in the medial ganglionic eminence (MGE), a transient ventral telencephalic structure. The study in question developed MGE-like progenitors and MGE-derived cortical interneurons (cIN). *CHD2* was enriched in cINs developed in this fashion and cells depleted for *CHD2* using lentivirus transfected siRNA demonstrated impaired cIN specification. cINs are inhibitory GABAergic neurons [83] meaning that in broad terms, a reduction in cIN number or activity could lead to a hyper-excitable state in the developed cortex and lower threshold for triggering seizure activity.

These *CHD2* depleted cells also displayed abnormal electrophysiological function and dysregulated expression in other genes known to be mutated in epilepsy, such as sodium channels[4].

Both disruption of the RG pool and abnormal differentiation and function of GABAergic cell populations could contribute to disordered neurodevelopment and epileptogenesis and these studies provide potential insight into some of the mechanisms underpinning *CHD2* related EECO.

1.4.2 Myogenesis

Although possibly not strictly relevant to its role in neurodevelopment and therefore falling slightly outside the remit of this monograph, further evidence for *CHD2*'s role in cellular differentiation and cell type specification comes from studies into its role in myogenesis[5]. In interaction with the muscle specific transcription factor MyoD, *CHD2* was demonstrated to deposit the histone marker 3.3 (H3.3) at lineage specific sites prior to differentiation. It was also demonstrated that *CHD2* knockdown, achieved with siRNA, caused a loss of H3.3 deposition at myogenic loci, but not at the promoters of housekeeping genes.

Although not specifically relevant to CNS development, this study links the specific chromatin remodelling function of *CHD2* (see section 1.5.2) as being linked to the differentiation of specialised tissues in concordance with a tissue-specific transcription factor.

1.5 CHROMATIN MODIFICATION

1.5.1 Chromatin Structure and function

1.5.1.1 Introduction to Chromatin

First observed in 1878 as a thread-like structure in the nucleus, chromatin is now understood to be a biologically active, interaction-rich protein scaffold, around which DNA is compacted in eukaryotic cells.

Many genes influencing cell function and differentiation exert their action by remodelling chromatin and altering the physiochemical microenvironment of chromatin at specific and diverse sites throughout the whole human genome.

A full recounting of our understanding of chromatin's function would take many pages, and indeed entire textbooks have been written on the subject - with the rapidity of advance in the field of chromatin study, these textbooks are likely already out-of-date. What follows is a precis of the most important aspects of chromatin biology, in order to provide a framework for understanding CHD2's role as a chromatin remodeller and interactor.

1: Introduction

1.5: Chromatin modification

1.5.1.2 Histones and chromatin structure

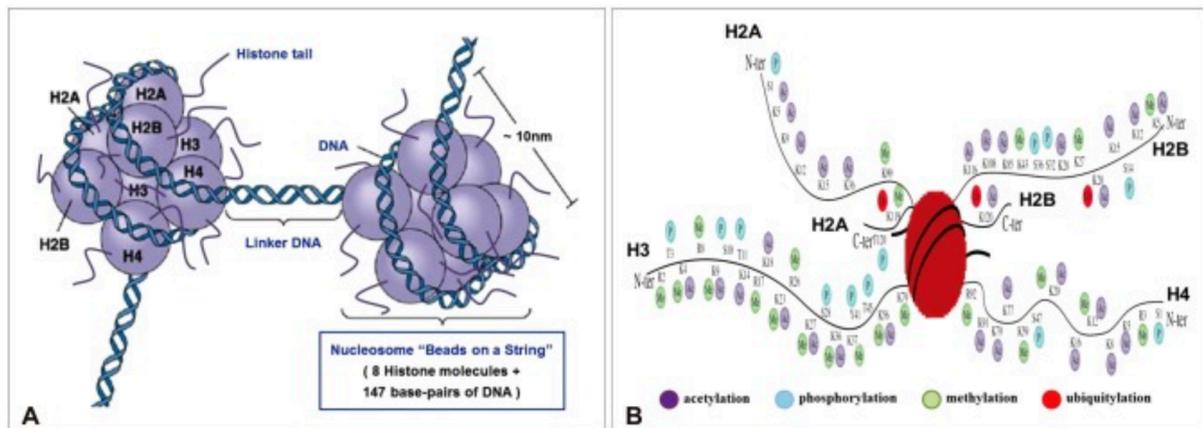


Figure 1.4 – Schematic representation of histone octamer and nucleosome structure. A) demonstrates the coiling of DNA around the histone octamer, made up of eight core histone proteins, B) demonstrates various modifications of the histone molecule ‘tails’ that alter the physiochemical microenvironment and thus regulate the biological activity of chromatin. Image taken from Kim YZ: ‘Altered Histone Modifications in Gliomas’, *Brain Tumor Res Treat.* 2014 Apr; 2(1): 7–21 [84]

The nucleosome is the first level of chromatin organisation. Each nucleosome consists of two copies of histone proteins H2A, H2B, H3 and H4 assembled into an octamer (figure 1.4 – A), around which 146-147bp of DNA is wrapped 1.7 times[85, 86]. The nucleosomes are linked by the H1 linker histone.

Nucleosomes are then compacted into a 30nm fibre, featuring “two-start” higher order organisation; or to put it another way, arrayed in zig-zag formation with disulphide bonds providing stability[87].

Chromatin can be further compacted by the multi-protein polycomb repressor complex[87]. Compacted chromatin is referred to as heterochromatin, and open chromatin is referred to as euchromatin.

The folding of DNA into chromatin fibres hinders accessibility to transcription and replication factors that control which genes are expressed at any one time. Therefore, chromatin must be dynamically modified to unfold and thus regulate nuclear function[88] (figure 1.4 - B).

In general, heterochromatin is thought to be inactive, whereas euchromatin is active. Given that even densely packed chromatin remains biologically potent, it is useful to define “active”. By general convention and for the purposes of this thesis, “active” chromatin and “active” genes refer to those genes which are highly transcribed and translated to proteins. “Inactive”, chromatin “inactive” genes, conversely, refer to genes where transcription is suppressed, or silent.

1: Introduction

1.5: Chromatin modification

A third transcriptional state also exists – some chromatin regions are referred to as “bivalent” or “poised”. In bivalent regions, the genes are primed or poised for transcription, but not yet activated[89]. Bivalent regions are a feature of totipotent and pluripotent cells lines, and interactions with transcription factors and chromatin remodellers can commit them to an active or inactive state as the cell undergoes differentiation.

1.5.1.3 Histone tail modifications

Each histone also has a 20-35 residue segment, rich in basic amino acids, which protrudes from the surface of the nucleosome. This tail is a frequent target of post-translational modification (PTM). Multiple PTMs have been identified affecting the core histones, along with proteins that make and read these modifications. A complete recounting of these is beyond the scope of this thesis, however a table of histone variants and PTMs relevant to the work presented in this thesis can be found in *table 1.3* [90].

These histone modifications alter the physiochemical properties of the histone proteins and imbue them with biological information which can be “read” by histone-interacting proteins and chromatin remodellers[91].

A recent effort has been made to collate a database of histone modifications and their interactors[92]. For human cells, this included 413 site-specific regulator-histone relations and 168 histone regulators affecting the four core histones, highlighting the vast complexity of the system. H3 is by far the most frequently modified histone protein, however a large number of interactions involving H4 were also included, as well as smaller numbers affecting H2A and H2B.

CHIP-Seq is a technique in which DNA is crosslinked to associated proteins using formaldehyde before fractionating and antibody pulldown of target proteins. The DNA is then released from the protein target and sequenced, providing a map of binding sites for each protein. By performing CHIP-Seq targeting multiple, previously well studied histone modifications, a regulatory map identifying 15 different chromatin states has been assembled and validated[90].

This approach yielded states divided into promoters (active, weak and poised), enhancers (strong and weak), insulators, transcriptional transition, transcriptional elongation and weak transcription, polycomb repressed, heterochromatin and two different states associated with repetitive regions[90].

1: Introduction

1.5: Chromatin modification

The bivalent chromatin regions described above are characterised by a mixture of the markers associated with active and inactive regions[89, 90].

Histone marker	Regulatory functions
H3.3	Found at active chromatin and primed enhancers[93-95]
γH2AX	H2AX phosphorylation is a marker of DSB[96, 97]
H2AZ	Found at TSS and regulatory elements. Conflicting evidence exists regarding its effect on transcription and this is likely mediated by PTM[98]
H3K4me1	Found at primed enhancer sites [99] – associated with DNA hypomethylation Fernandez, 2015 #1235} and active chromatin
H3K4me2	Enrichment defines transcription factor binding regions[100]
H3K4me3	Associated with transcription initiation[101]
H3K27ac	Enrichment separates active enhancers from poised enhancers[102]
H3K27me3	Promiscuous repressive marker [103] associated with heterochromatin compaction and polycomb repressor complex binding

Table 1.3: Selection of histone variants post-translational modifications and associated regulatory functions relevant to this thesis

1.5.1.4 Nucleosome Mobilisation

As well as being subject to biologically relevant PTMs, nucleosomes are subject to disassembly, eviction, sliding and spacing by chromatin remodellers. The position of nucleosomes throughout the genome also has the capacity to alter genomic function. Studies of organisms with simple genomes, such as *dictyostelium* demonstrated that nucleosome position was relatively stable, with up to 80% of nucleosomes being present in the same place throughout a cell population. In these organisms, transcription start sites (TSS) are occupancy-free, and occupancy of histones and transcription factors is mutually exclusive[104].

The situation in the mammalian genome is more complex. 98% of the genome is non-protein-coding and histone position is highly mobile. MNase digestion reveals regions that are more accessible to proteins – in human cell lines, these accessible areas are not mutually exclusive to nucleosome occupancy[105]. These MNase sensitive areas have been shown to cluster at TSS. Highly accessible areas, as defined by increased sensitivity to digesting by DNase, are found at regulatory regions, often many kilobases distant from transcription start sites. [104, 105]

A recent study into position of nucleosomes during neuro-differentiation compared nucleosome positioning in hiPSC and NPCs derived from those cells. It demonstrated an

1: Introduction

1.5: Chromatin modification

increase in the number of highly positioned nucleosomes in the NPC state, but also demonstrated that highly positioned nucleosomes were not a marker for increased transcription[104]. They concluded that although small subgroups of genes may exist where nucleosome positioning is important, there was no strong correlation between positioning and activity at a genome-wide level.

The finding that nucleosome position does not correlate directly with gene activity hints at the importance of a further level of organisation; 3D genome architecture. Using new technologies that allow for enrichment of DNA sequences according to their 3D topographical association, the regulatory contribution of higher-order chromatin structures is now being explored.

The methods by which nucleosomes mobilise are not well understood – the ‘beads on a string’ model often used as an analogy for chromatin structure implies that nucleosomes slide up and down the DNA strand, however this may not be the case. There is evidence that conformational changes in nucleosomes translocate the DNA itself[106]. Given the number of ATP-lysing interactions that can occur at the sites of each histone, this model with nucleosomes as a motor is perhaps more likely.

Chromatin Conformation Capture (3C) and high-throughput chromatin conformation capture (HiC) provide this data for local and genome-wide associations respectively[107]. Although HiC is a relatively recent development this work has provided further evidence that the 3-dimensional structure of the nucleus is vital to gene activity regulation in mammals. Folding of the genome into Topographically Associated Domains (TADs) allows distant regulatory elements such as enhancers and transcription factor binding sites into close association with gene bodies.[108] There is also evidence that some transcription factors have specificity for the shape of the genome, rather than a sequence motif[109].

This evidence, beginning to link the 3D topology of the nucleus to nucleosome position and chromatin remodelling provides a potential insight into how chromatin remodelling proteins can impact the function of developing cell lines and organisms[108]. A change in the speed and efficiency at which a particular chromatin modification is made could disrupt this topology and lead to less stringent regulation of a multitude of downstream genes, reflected in a shifting transcriptome and thus a shifting proteome.

From the point of view of the central dogma of molecular biology [110], which states that information passes from genome to protein via the transcriptome but not the other way,

1: Introduction

1.5: Chromatin modification

the 3D architecture of the genome can be considered an information filter. The topographical arrangement of DNA regulates access to the second level of filtering, transcription factors acting directly on the genome in order to regulate gene activity. As chromatin is modified, the shape of the genome and the nature of the filter changes allowing for cell-type specification and dynamic function of differentiated cells in response to endogenous and exogenous stimuli, via changes in the intracellular proteome (*figure 1.5*).

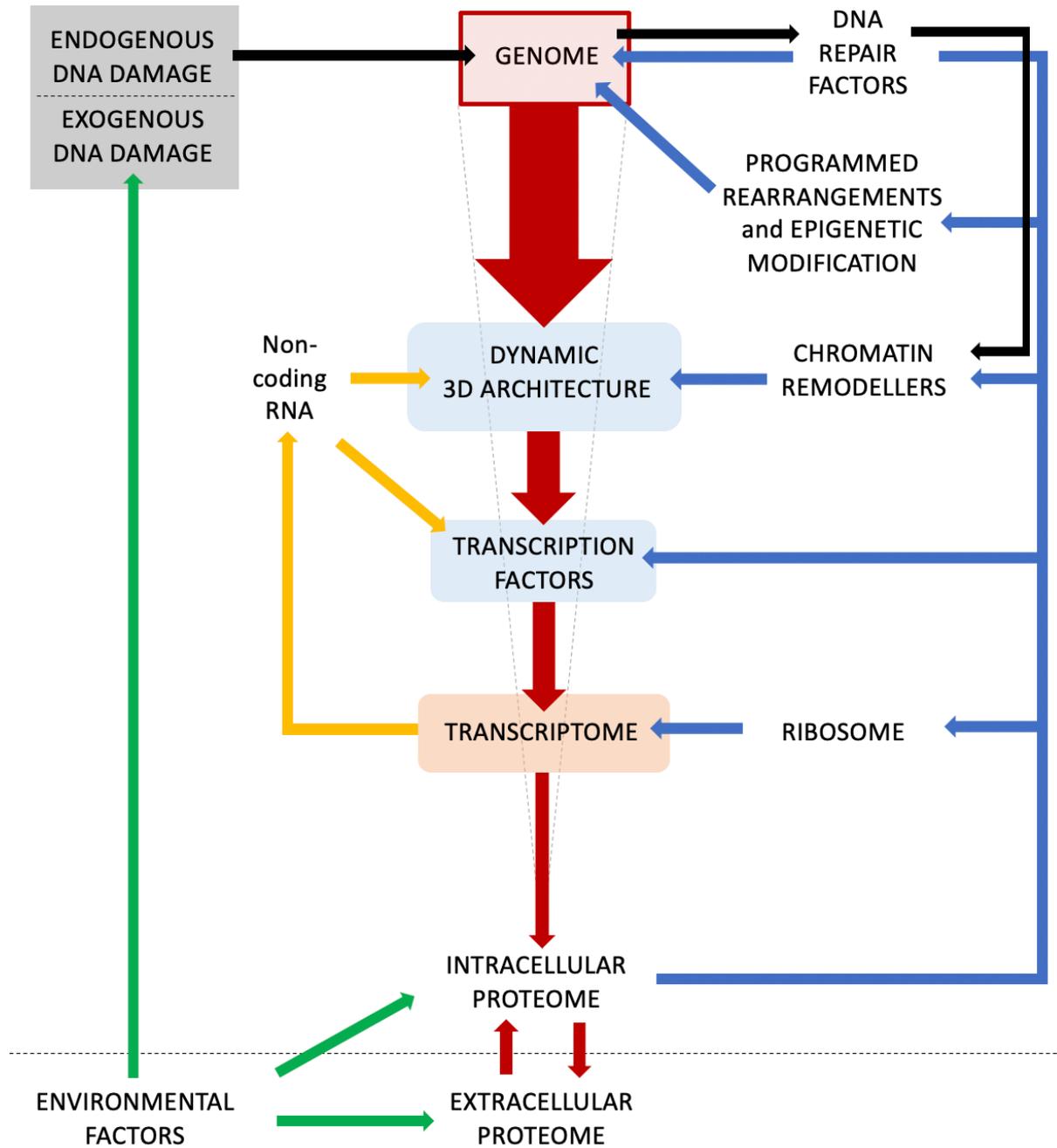


Figure 1.5 – flow and filtering of genomic information, in line with the central dogma of biology. Red arrows represent the flow of information from the genome to the proteome, green arrows environmental influences, orange arrows feedback by RNA mechanisms and blue arrows feedback by protein factors. Black arrows represent DNA damage and its influence.

1: Introduction

1.5: Chromatin modification

1.5.1.5 Histone variants

Adding further to the informational complexity of the histone code, three of the four histone subunits (H2A, H2B and H3) exhibit a diverse range of histone variants.

The canonical histone subunits are deposited in a replication dependant fashion during S phase [111] – H3.1, H2A, H2B and H4 are assembled into a nucleosome and incorporated behind the replication fork by the chaperone protein Chromatin Assembly Factor 1 (CAF1). Conversely, histone variant subunits are incorporated in a replication independent fashion.

H2A and H2B are the canonical variants of histone 2. The H2A.Z variant is the most structurally divergent variant, only having 60% sequence homology with the canonical H2A. Despite this, the structure of nucleosomes containing H2A.Z is similar to those containing H2A, although current thinking is that a reduced number of hydrogen bonds between this form of H2A and the other parts of the nucleosome octamer results in a less stable, more mobile nucleosome[112], the function of H2A.Z remains the cause of some confusion. Studies have identified different roles including: occupancy at the promoters of both active and suppressed genes, stabilisation and destabilisation the nucleosome, and both repression and activation of transcription [112].

The H2AX variant is crucial for genome stability – phosphorylation of H2AX, resulting in γ H2AX is the first stage in initiating the response to DSBs[96, 97]. The role of γ H2AX in NHEJ is explored further in section 1.6.

The vertebrate-specific variants, macroH2A and H2A.B are thought to activate and repress transcription respectively[112].

There is evidence that variants in H3 are related to the maintenance of transcription of active genes. H3.1 and H3.2 are the canonical variants incorporated during DNA replication. The genes encoding these proteins are found in long tandem arrays (Histone gene cluster 1 and Histone gene cluster 2 respectively) and expressed promiscuously during S-phase. Interestingly, the RNA transcripts do not have the usual poly-adenylated tails and are subject to rapid degradation. Five replacement variants have been described; CenH3 is a variant incorporated at centromeres, and indeed can be seen as defining the centromere region by its presence[113], H3t is a testes specific form of the protein, two recently discovered primate-specific variants – H3.X and H3.Y – which may play a role in transduction of exogenous signals affecting cell growth [114], and the ubiquitously expressed H3.3.

1: Introduction

1.5: Chromatin modification

H3.3, encoded by two genes with identical products – H3f3a and H3f3b [111]– is found at sites of transcription, as defined by CHIP-Seq detection of RNA polymerase. It is thought that the canonical nucleosomes are ejected by the activity of RNA polymerases and that H3.3 is assembled to replace them[115]. H3.3 accumulates modifications associated with gene activation, including methylation of K4, K36 and K79 and acetylation at K9 and K14, and is depleted for K9 and K27 methylations associated with gene suppression [94]. It has been implicated as a regulator of role in gene activation[116], epigenetic memory of transcriptional state across cell division[117], and in maintenance of pluripotency in ESCs[37].

As with H2A.Z, there is also some confusion regarding the exact regulatory role of H3.3. There is evidence of H3.3 being deposited both at heterochromatic sites and at the promoters and enhancers of silent genes, suggesting that H3.3 presence alone is not sufficient to activate full length gene transcription [118]. It may be that it primes promoter regions but requires the action of specific transcription factors in order to complete activation of target genes.

Total H3.3 depletion leads to chromosome instability by disruption of heterochromatic regions and cell death by activation of p53 [119], a master cell-cycle checkpoint regulator and a tumour suppressor gene which functions by triggering apoptosis in cells with unrepaired DNA lesions. Given the ubiquitous and essential nature of H3.3 is however difficult to know if much can be inferred from this finding as to H3.3's physiological function.

Interestingly, in post-mitotic and senescent cells such as mature neurons H3.3 gradually replaces H3.1 and H3.2 throughout the entire genome. This activity is also seen in mice. Animals with downregulated H3.3 exhibit a reduction of synapses and impaired long-term memory[120]. In humans, H3.3 expression is upregulated in patients experiencing depression, and downregulated in patients undergoing antidepressant therapy, further demonstrating its importance in the CNS[121].

H3.3 has also been demonstrated to be crucial for genome stability in the developing murine embryo. Mice with knockout of H3f3a and heterozygous mutations in H3f3b suffered multiple developmental abnormalities caused by increased apoptosis mediated by p53 and in embryos rescued by p53 suppression mature cells displayed a wide range of karyotypic abnormalities. Upregulation of H2AX, associated with DNA damage recognition as described above, was also noted[119]. Interestingly despite H3.3's putative role in maintaining

1: Introduction

1.5: Chromatin modification

transcription, whole transcriptome sequencing only demonstrated dysregulation of transcription affecting 5% of genes.

In summary; although histone variants are incorporated in a non-random pattern throughout the genome, in the case of H2A.Z and H3.3 there are significant inconsistencies in the data regarding the link between variant deposition and active transcription. It is likely that the histone variants regulate genome activity in concert with: their PTMs, transcription factors, their target sequences, non-coding RNA (ncRNA), micro RNA (miRNA) and other as yet unidentified elements. That no direct on/off switch cause and effect has yet been identified highlights the complexity of the genome regulation and the chromatin interactome.

1.5.2 CHD2 mediated chromatin modification

So, we come back to CHD2 and its function as a chromatin remodeller. It has been demonstrated that the presence of DNA alone is enough to trigger lysis of ATP by CHD2. ATPase activity is nearly unmeasurable in the presence of only core histones, detectable in the presence of plasmid DNA and high in the presence of plasmid DNA complexed with histone proteins[14]. It has been demonstrated to catalyse nucleosome assembly when naked plasmid DNA is present with core histone proteins, ATP and NAP-1 *in vitro*.

Interestingly, when a truncated CHD2 protein, not containing its chromodomains is incubated with chromatin and DNA, the rate of ATPase activity actually increases. The presence of the chromodomains can therefore be characterised as inhibiting, or perhaps tuning, the ATPase activity of the SNF2-DNA-binding domains. In keeping with observations in other chromatin remodellers, it is thought that the chromodomains “read” the PTMs of the histone tail in order to regulate the protein’s activity. The chromodomains are also sufficient for chromatin remodelling to take place *in vitro*, even when the protein is truncated to remove the DNA binding domain.

To summarise, the SNF2-ATPase domain and chromodomains are essential for the control of remodelling. When the DNA-binding domain is bound to DNA substrate, the rate of ATP use increases, along with, it is assumed the rate of remodelling[14].

Relatively little is known about the exact purpose of CHD2-mediated remodelling activity. A recent high throughput analysis provided further evidence for a role for CHD2 in the regulation of active transcription. By performing a detailed re-analysis of CHIP-Seq data available on the publically accessible Encyclopaedia of DNA Elements (ENCODE) consortium database [122, 123], Siggins et al [3] demonstrated that CHD2 is co-recruited with CHD1 in association with RNA polymerase II (Pol II) machinery, at active chromatin regions. They also demonstrated that CHD2 knockdown leads to an increased occupancy of H3 at these active sites. They also demonstrated a decrease in the relative enrichment of variant H3.3.

As previously mentioned, CHD2 has been demonstrated to deposit histone variant 3.3 in cooperation with the master transcription factor MyoD, during myogenesis [5], leading to activation of the transcription of lineage-specific genes.

CHD2 has also been shown to co-precipitate with Oct3/4, one of the factors necessary for induction and maintenance of pluripotency [124, 125], at bivalent chromatin sites defined by co-localisation of H3K27me and H3K4me in pluripotent mESCs.

1: Introduction

1.5: Chromatin modification

Knockdown of Oct3/4 reduced CHD2 occupancy at these sites, demonstrating a further interaction between a master transcription factor. Depletion of both CHD2 and Oct3/4 showed similar effects; an increase in the enrichment of H3.3 and H3K27me at bivalent genes. [126] These results imply that both CHD2 and Oct3/4 are required for maintenance of the bivalent state – similar to the scenario described by Harada et al [5], where both MyoD and CHD2 were required for proper expression of lineage specific genes during myogenesis.

It is suggested, therefore, that CHD2 functions in pluripotent cells to maintain the poised nature of bivalent regions through regulation of H3.3 deposition, and therefore maintain the poised nature of these regulatory regions for proper transcription during differentiation. The implications are that CHD1 and CHD2 are recruited to transcribed regions of the genome to regulate chromatin architecture, enhancing nucleosome disassembly and access to the genome.

A further interaction between CHD2 and H3.3 was demonstrated in response to DNA damage. Although γ H2AX is known to initiate DNA damage signalling, it is also well established that H3.3 is deposited during DNA damage repair[127]. By expression of a SNAP-tagged H3.3 protein which can be detected fluorescent labelling, it was demonstrated that knockdown of CHD2 reduces this damage-induced H3.3 deposition[6].

In summary, several interactions between CHD2 and H3.3 have been identified. These have been localised to sites of active transcription in co-recruitment with CHD1, and sites of poised bivalent chromatin states in pluripotent stem cells. It is suggested that CHD2 contributes to the maintenance of these bivalent sites. CHD2 has also been demonstrated to be required for H3.3 deposition in response to DNA damage. The DNA damage response is described in more detail below, including the possible consequences of dysregulating the DNA damage response.

1.6 DNA REPAIR

1.6.1 DNA Damage

Every cell in the human body experiences numerous DNA damaging events per day, leading to a wide variety of lesions. Endogenous lesions include removal of bases from the phosphate backbone by depurination or depyrimidation, aberrant methylation, base transition via cytosine deamination, oxidation [139] and more damaging lesions such as single strand breaks (SSB) and DSBs [148] (see *table 1.4*). These lesions include endogenous lesions that occur as a result to normal cellular metabolism, the generation of reactive oxygen species, base misincorporation during replication, modification of bases by alkylation [128] and damage caused by aberrant action of transcriptional and replicative machinery [129, 130]. A large number of lesions also occur as a result of exogenous agents including background radiation, ultraviolet spectrum sunlight and pollutants.

There are several possible consequences of the accumulation of DNA damage in cells. First, if DNA damage is unrepaired then cell cycle checkpoints can be activated. This can lead to cell senescence or programmed cell death by apoptosis [80]. Increasing levels of cell senescence are linked to aging[131] – as cells stop cycling and dividing, tissues lose the ability to regenerate leading to reduced function.

The other potential consequence of DNA damage is mutation. A mutation can be defined as a change in the sequence of the genome. Such changes can include removal of genetic information – a deletion, addition of genetic information – insertions, or a change in genetic information at a single nucleotide – the single nucleotide variant or SNV.

When mutations occur in protein coding exons, splice sites between exons, or conserved non-coding regulatory regions, there can be an impact of the protein translated from the gene effected. These impacts can be difficult to predict, and considerable effort goes into the classification of germline mutations in the field of Clinical Genetics[132].

DNA lesion	Type of DNA damage	Lesions per cell per day
Abasic site	Depurination	10,000 ^a
Abasic site	Depyrimidation	500 ^a
Base transition	Cytosine deamination	100-500 ^a
3-meA	SAM-induced methylation	600 ^a
7-meG		4000 ^a
O6-meG		10-30 ^a
8-oxo-dG	Oxidation	400-1500 ^a
SSB	Ionising radiation	100-500 ^b
DSB	Causes, including replication fork collapse, transcription and exogenous damage	10-50 ^b

Table 1.4 – types of DNA lesion, type of damage and frequency per cell per day. *a* - figures taken from Ciccio and Elledge (2010)[139], *b* – figures taken from Mehta and Haber (2014)[148]

It can, however, usually be assumed that certain types of mutation are likely to result in the loss of the affected allele. If an SNV introduces a stop codon then the transcription of a gene will be prematurely terminated leading to an abnormal RNA that typically undergoes nonsense-mediated decay. If an insertion or deletion (indel) occurs that shifts the reading frame of the triplet code of a protein coding region then all the amino acids coded downstream of the indel – this will also usually introduce a premature stop codon leading to nonsense mediated decay. Intronic and splice-site mutations affecting the splicing of immature mRNA to form mature transcripts can similarly result in effective gene allele loss. Finally, if multiple exons or an entire copy of a gene is deleted as part of a larger deletion or copy number variant (CNV), then it will no longer produce RNA transcript.

Other types of mutation can constitutively activate a gene that is normally under tight regulation. For example, a chromosomal rearrangement that brings a silenced gene into a region downstream of a strongly activated promoter would switch on production of that gene. It is also possible for SNVs to change the amino acid code in a way that confers 'gain of

1: Introduction

1.6: DNA repair

function' to a protein; such a mutation affecting an ion channel could lead to abnormal ion balance within neurons or cardiac myocytes tissue and increase the risk of seizures or arrhythmias respectively.

The most well-understood clinical consequence of a high mutational burden is the development of malignancy[133]. When a combination of tissue-specific tumour suppressor proteins and DNA repair proteins are deactivated, in combination with activating mutations affecting growth factor signal transduction pathways, then uncontrolled growth of cells with unstable genomes occurs, jeopardising the survival of the host organism. It is also now widely accepted that accumulation of DNA damage in benign cell populations is a substantial contributor to the process of aging[133].

DSBs are amongst the most toxic forms of DNA damage [134]. It is estimated that dividing mammalian cells experience 10-50 DSBs per day per cell[135, 136]. They have the potential to cause translocations, loss of genetic material and premature cell death if not repaired[137]. Even when they are repaired, the repair mechanisms are error prone and there is a high chance of mutations developing at the DSB site (see section 1.6.3).

DSBs can occur anywhere in the genome, however their occurrence is not completely stochastic. Common Fragile Sites (CFS) are regions more prone to DSB occurrence in somatic cells[138-141]. They are also preferentially involved in sister chromatid exchange, translocations and deletions and often implicated in the amplification of oncogenes that give rise to various malignancies.

They also occur frequently in late-replicating regions of the genome and AT-rich regions. These AT rich regions are prone to the formation of complex secondary structures that can stall the process of DNA replication, leading to DSBs related to replication fork collapse. They are also enriched for the transcribed regions of long genes, where conflict between transcription apparatus and replication machinery is felt to be the cause[140].

Common CNVs and transcriptionally associated CFS are both known to be tissue specific and there is evidence of a link between the CFS associated with large transcriptional units and common somatic CNVs[130]. Aphidicolin is a commonly used reagent to induce increased replication stress and thus amplify the signal from CFS . By treating fibroblasts with aphidicolin, it was demonstrated that CFS corresponded to commonly seen fibroblast-specific CNVs.

Transcribed regions are also prone to formation of hybrid structures between the template strand and the transcribed RNA, which have also been associated with replication fork collapse and DSB occurrence[142].

DSBs also occur in a programmatic fashion in some regions of the genome: V(D)J recombination – the process by which antibody and T-Cell receptor diversity is generated in maturing lymphocytes, and class switch recombination (CSR) which changes the class of immunoglobulin heavy (IgH) chains produced in lymphocytes. These breaks are created by specific proteins, the RAG1 and RAG2 complexes, and are under strict biological control [143]. As they are not of direct relevance to the developing brain and can be seen as an overlapping process rather than a complete homologue of damage repair, they are not considered further in this thesis. There is, however, emerging evidence that DSBs also occur in a programmatic pattern in the developing brain – this is explored in the next section.

1.6.2 Double strand breakage in the developing brain – evidence for a contribution towards cellular physiology

It is now well established that the genome of the adult brain contains a high level of somatic mosaicism, with mosaicism defined as “the existence different genomes within the cells of a monozygotic individual” [144] Somatic mosaicism is known to be the cause of several genetic diseases affecting neurodevelopment[145, 146], and there is evidence to suggest that acquired somatic mutations contribute towards age related neurodegeneration[147]. The differences between the genomes of each cell include: SNVs, indels, larger CNVs (1Mb < in size), translocations and even the loss or gain of an entire chromosome[148]. It is not known whether this mosaicism is important for normal neurodevelopment, however it is thought that deficiencies in DNA repair pathways [149] can affect somatic mosaicism profiles, as can somatic retrotransposition of mobile DNA elements[150].

It is understood that defects in single base repair pathways - such as base excision repair, nucleotide excision repair, and transcription-coupled repair - can cause SNVs and indels, [144] and that deficiencies in NHEJ can cause an increase in the rate of CNVs. It has also been recently demonstrated that a different indel profile can be detected when either NHEJ or A-EJ function at a DSB site. [151]

The role of DNA repair and somatic brain mosaicism is currently an area of significant interest, both in terms of understanding normal CNS function and in terms of its relationship to neurodevelopmental, neuropsychiatric and neurodegenerative disease.

In two concurrent studies [152], two separate classes of DSBs were noted in neuronal stem progenitor cells (NSPCs). DSBs are prone to translocating to distal regions in the genome – using a technique involving bait strands, the rate of DSB translocation was measured. In both cases it should be understood that only the minority of DSBs translocate and most are repaired *in situ*, however the translocation rate was used as a proxy measure of DSB rates.

The first study [152] identified 27 recurrent DSB clusters (RDCs) in murine NSPCs, 24 of which were present in genes with roles in neuronal function including: cell adhesion, neurotransmission and synaptic surface proteins. Nine of the human analogues have been found in recurrently occurring CNVs identifiable in healthy adult brains. They suggest that these RDCs could give rise to the associated CNVs in adult brains and that this may be part of normal CNS development[153].

1: Introduction

1.6: DNA repair

In the second study [154], they report recurrent DSBs occurring within 200bp of the TSS of highly transcribed genes. As these DSBs occur across divergent cell types, and the RDCs identified in the study referenced above did not occur within 200bp of TSS, the authors proposed that these findings represent two different classes of DSBs.

Rather than representing a pathologic process that must be minimised, the generation of somatic brain mosaicism may be essential for the generation of the full diversity of neurons required for normal neurological function[155]. If this is the case, then DSB formation is likely to play a role in the regulation of copy number change in the developing brain.

Several laboratory studies give credence to this hypothesis. DSB formation has been reported in mouse experiments of neuronal stimulation, in which mice undergo a new learning experiment before being euthanized and investigated. In these studies, γ H2AX was used as an analogue of DSB accumulation. γ H2AX accumulation was most notable in the dentate nucleus, which is known to be associated with learning and memory tasks, establishing a link between adaptive CNS function and DSB accumulation. [156] Interestingly, amyloid protein accumulation (of the type seen in Alzheimer's disease) was demonstrated to increase the rate of these stimulation induced DSBs.

It is possible that these stimulation-induced DSBs play a role in synaptic plasticity and long term neuronal maturation. Maturation in response to external stimuli has long been thought to depend upon initiation of transcription programs required for synaptic plasticity. These transcriptional programs rely upon two sets of genes, categorised by the time-course of their upregulation.

Early response genes are enriched for transcription factors and are transcribed independent of changes in the proteome. It has been demonstrated that transcription factors sit pre-bound at important sites ahead of activation in response to the calcium influx that characterises neuronal stimulation[157]. These early response genes govern the expression kinetics of late response genes which in turn govern processes such as: neurite outgrowth, synapse development, and maturation.

The early response genes were recently demonstrated to be upregulated when neurons were subjected to treatment with etoposide, an agent known to cause DSBs[158]. Given that the general response to DSBs is downregulation of transcription, this led to inquiry and demonstration that DSBs are responsible for altering the topological landscape in neurons, acting as a molecular switch in order to rapidly upregulate the transcription of these

1: Introduction

1.6: DNA repair

early response genes. Incubation with cultured neurons with KCl led to an upregulation of the early response genes *Fos* and *Npas4* with a concomitant upregulation in γ H2AX. A murine fear conditioning experiment in which hippocampal lysate was extracted 15 minutes after fear conditioning (during which it is assumed new hippocampal plasticity will occur) also demonstrated similar responses [158].

In summary, there is an increasing body of evidence for the non-random and programmatic occurrence of DSBs in neurons in response to stimuli, as part of the neuroplastic processes involved in learning and memory development. The dynamics of DSB repair are therefore likely to be important not only in maintenance of a healthy genome during CNS development, but in the long-term adaptive plastic function of the brain. I will now review the molecular mechanisms underpinning DSB repair and their relationship to chromatin remodelling and *CHD2* mutation.

1.6.3 The DNA Damage Response to DSBs

Given the extent of the bombardment experienced by cells on a daily basis, it should not be a surprise that various mechanisms have evolved for the repair of DNA damage. There are a multitude of described genetic disorders caused by deficiencies in various factors within DNA damage repair pathways.

There are too many disorders to give each a full and detailed description here however in general the symptoms and signs of DNA repair disorders include: an increase in the rates of specific cancers, premature ageing, neurodegeneration, UV or radiation sensitivity, dysmorphic facial features, abnormal growth, and microcephaly[159-161].

Three DSB repair pathways are known to exist in eukaryotic cells; homologous recombination (HR), non-homologous end joining (NHEJ), and alternative end-joining (A-EJ). Each has different properties and is mediated by separate protein pathways. Inhibition of the core factors of one pathway does not reduce the efficacy of the others [6], and may in fact lead to an increase in the other's utilisation. This has been demonstrated by the use of small molecules which block one pathway, leading to an increased signature of the another [162].

It has long been established that chromatin undergoes modification in response to DNA damage and double strand breaks[163]. In DSB signalling H2AX is phosphorylated promiscuously (forming γ H2AX) in a 2Mb region around the break. Both chromatin relaxation and compaction have been demonstrated in the time course following DNA damage induction. It is often stated that compacted chromatin must be "relaxed" in order to provide space for the core repair factors to ligate the DSB ends however the truth may be more complex than this, with different histone modifications interacting with elements of the repair machinery to stimulate or suppress their activity. It is also possible that chromatin modification functions as part of the cell's attempt to build a molecular "bridge" around the two broken ends of DNA and reduce the risk of translocations or loss of sequences telomeric to the breakpoint. If the DNA ends dissociate from one another then repair will not be possible. Multiple studies have identified roles for different CHD proteins in different repair pathways[7,139,148] (see *table 1.2*). These studies have also demonstrated abrogation of specific repair pathways in absence of the relevant CHD proteins.

As an example of these more complex dynamics CHD2 mediated H3.3 deposition has recently been identified as both associated with and necessary for the repair of DSBs via

1: Introduction

1.6: DNA repair

NHEJ[7]. As well as demonstrating a loss of chromatin expansion when H3.3 was suppressed, co-immunoprecipitation (coIP) demonstrated that H3.3 was an interactor of KU.

In the remainder of this section, I review the mechanisms which influence whether NHEJ, A-EJ or HR are utilised and the current understanding of the mechanisms and protein pathways of each DSB repair pathway. *Figure 1.6* provides an overview of the process including the key proteins involved at each stage.

1.6.3.1 DSB Detection and repair pathway choice

The mechanisms by which pathway choice is made are not fully understood. It is known that NHEJ is the primary DSB repair pathway, and the most commonly utilised at all points in the cell cycle. Even during the G2 phase when the more accurate mechanism of HR is active, it is only (HR) is only utilised in repair of around 15% of DSBs[164].

Several factors influence the choice of repair pathway used and when considering these mechanisms it is useful to think of repair in two categories – repair requiring resection and repair without resection. NHEJ requires no resection, whereas HR and A-EJ both require resection of DNA ends to be utilised.

Key components of each pathway compete with each other at DSB ends [165], and that the cellular concentration of Ku70/80 (KU) - a key complex regulating NHEJ - is at least ten times higher than the concentrations of all other known proteins implicated in DSB repair[166]. Interplay between KU, the MRN protein complex and Ataxia Telangiectasia Mutated (ATM) [167] – a master regulator of the DNA damage response named for the neurodegenerative condition that results from its absence – influences the pathway choice through a number of mechanisms including: DNA resection, phosphorylation of several hundred target proteins, generation of poly-A-ribose (PAR) chains via recruitment of PARP1 and PARP2, and chromatin remodelling.

The differing phosphorylation profiles of these various components in response to cell-cycle linked cyclin-dependant kinase (CDK) provides a mechanistic link between cell-cycle phase and pathway choice[168]. In particular CDK phosphorylates multiple proteins involved in the process of DSB end resection in a cell-cycle dependant fashion. This relationship to cell cycling means that A-EJ and HR are greatly downregulated in favour of NHEJ in non-cycling cells such as post-mitotic neurons. [169]

1: Introduction

1.6: DNA repair

If end resection is utilised, it occurs in two phases. The first phase, called 'end clipping' is carried out by the structure-specific nuclease MRE11 and CtIP. [169] These nucleases create around 20bp of overhang, which allows the DSB to be repaired by A-EJ. The second step, appropriately called 'extensive resection' creates a much larger stretch of ssDNA as is required for HR. [169]

Once resection has been initiated, a number of interacting factors influence whether the break is repaired by HR or A-EJ. These will be explored further in *section 1.6.3.3 – Homologous Recombination*.

1.6.3.2 Non-homologous End Joining

NHEJ (also sometimes referred to as classical NHEJ (cNHEJ)) is the most commonly utilised pathway of DSB repair in mammalian cells. It is highly conserved with analogues of NHEJ being observed in some bacterial species[170]. Minimal end processing is required to repair the break in the genome and it accepts a wide diversity of substrates in terms of end-sequence and architecture[166]. This mechanistic flexibility enables NHEJ to be used for DSBs that occur from a variety of sources and at all points in the cell cycle.

Generally, NHEJ requires the presence of junctional homology of 4bp or less in order to ligated DSBs. Repair of blunt DNA overhangs has been demonstrated to be highly efficient and can be performed with a minimal subset of core cNHEJ proteins. Other more complex overhangs can be resected by non-core NHEJ associated nucleases in order to facilitate ligation. The choice of which sub-pathway is utilised appears to depend on the shape of these overhangs[171].

The first stage of NHEJ is the binding of KU [172]. KU is present at extraordinarily high intracellular concentrations and associates with DSBs within 5 seconds of the lesion occurring [6]. KU binds DNA in a ring conformation regardless of the sequence, providing the mechanistic flexibility required to repair the full spectrum of DSBs that occur in a normal cell. Once it is bound it can translocate down the DNA strand and act as a binding site for other proteins required for repair – it has been described as a protein 'tool-belt' allowing for the nucleases, kinases, and ligases required for end repair to bind. It also antagonises binding of proteins involved in other DSB repair pathways[166].

1: Introduction

1.6: DNA repair

DNA-dependant protein kinase catalytic subunit (DNA-PKcs) has a high affinity for Ku bound to DNA ends and forms the DNA-PK complex, which is responsible for the phosphorylation of several other proteins in the NHEJ pathway[173].

The DNA Ligase IV (LigIV) and X-ray repair cross complimenting protein 4 (XRCC4) complex are the central components of NHEJ. XRCC4 forms a duplex with XRCC4-like factor (XLF) providing a sheath around the paired DSB ends, which are then covalently ligated by LigIV[6, 136].

Perhaps the most well studied nuclease involved in overhang resection is Artemis. Artemis is a DNA nuclease, with a well-established function in processing hairpin structures at DSBs, the likes of which are found during programmatic V(D)J recombination. It has been demonstrated that 20-50% of DSBs caused by ionising radiation require Artemis for repair [174].

Other nucleases that have been implicated in NHEJ include; PNKP-like factor, the MRN protein complex (including the MRE11, RAD50 and NBS1 proteins), WRN [175], FEN1 and EXO1. The extent to which these are essential for NHEJ is uncertain, and each protein may function on more complicated break architectures. It is likely that utilisation of sub-pathways of NHEJ is determined by the break architecture[6].

γ H2AX is known to be involved in the signalling for all DSB repair pathways[97], however more recently H3.3 has been linked to NHEJ. A recent study began by examining the link between PAR chains and DSB repair, demonstrating that PARP1 was required for chromatin expansion. Further to this they co-immunoprecipitated CHD2 in association with PAR chains and demonstrated that knockdown of either was necessary for the expansion of chromatin tracts.

Interestingly, although it had previously been believed that PARP1 was the initiating step in A-EJ [176] a more recent study actually demonstrated that when treated with a PARP inhibitor; KU and XRCC4 failed to aggregate, linking PARP1 function instead to NHEJ[6]. In investigating the function of CHD2 at DSBs it was demonstrated that PARP1 is necessary for the aggregation of H3.3 but not of γ H2AX, which appeared independently at DSB sites.

The process suggested by Luijsterburg et al (2016) states that γ H2AX signalling is an early stage of the DNA damage response, leading to PAR-chain accumulation – these PAR chains accumulate CHD2. CHD2 alters the chromatin microenvironment by H3.3 deposition,

1: Introduction

1.6: DNA repair

causing expansion of local tracts around the DSB sites and interacting with KU to initiate NHEJ[6].

It has long been believed that the lack of a template made NHEJ unreliable and mutagenic. Recently it has suggested that instead of mutagenesis NHEJ creates physiologically controlled changes in the DNA substrate sequence, however this remains a point of controversy[177].

Although it is almost certain that NHEJ can introduce sequence variants at repair sites, there is a growing body of evidence that A-EJ may be more mutagenic[178, 179]. Both mechanisms of repair are known to create indels at the DSB site although the size of the templated insertions is likely larger in A-EJ[180]. The difference in insert size is likely a function of the resections required to generate the requisite homology for repair of asymmetrical ends by each pathway (0-4bp for NHEJ, 2-10bp for A-EJ), and the various actions of the DNA polymerases needed to fill these gaps (Pol μ and Pol λ for NHEJ and Pol θ for A-EJ). Sub-pathways of NHEJ that require resection are therefore more likely to generate insertions than core pathway joining of blunt ended DSBs.

There is also a substantial body of evidence that A-EJ may increase the risk of DSB ends translocating to another chromosomal site although again this has been demonstrated with repair by both pathways[181]. There is also conflicting evidence that translocations demonstrated in lymphoid cell lines carry mutational signatures more associated with NHEJ than A-EJ[182].

1: Introduction

1.6: DNA repair

1.6.3.3 Homologous Recombination

HR can be conceptually divided into three stages: pre-synapsis, synapsis and post synapsis. [183]

During pre-synapsis the DSB is processed to form an extended region of ssDNA by a variety of endonucleases including MRE11, Rad50, Exo1, DNA2 and CtIP[169]. The ssDNA strand is then bound by RPA preventing the formation of secondary structures. Rad51 filaments are assembled by several mediators (Rad51 B/C/D, XRCC2/3 and BRCA2) that regulate the interaction between RPA which blocks access to the strand and Rad51 itself. [169, 183]

During synapsis, Rad51 promotes the formation of 'D-loops' with the homology template and DNA synthesis occurs.[183] [169] Multiple negative regulators of HR have been identified that redirect repair down other pathways such as A-EJ and single strand annealing, including Pol θ , BLM, FANCI, FANCD1 and Rad52 amongst others. [169]

During post-synapsis, three sub-pathways are available depending on whether a second strand is present. If a second strand is present, then synthesis dependant strand annealing (SDSA), which avoids the risk of crossovers and aberrant recombination. If no second strand is available then Break Induced Replication (BIR), in which the D-loop becomes a replication fork, is utilised. This maintains chromosomal integrity at the cost of loss of heterozygosity (LOH) of the repaired region. The third pathway, double holiday junction (dHJ) is used during meiotic recombination rather than repair in somatic cells and therefore will not be described further. [183]

1.6.3.4 Alternative End Joining (A-EJ)

Less is understood about A-EJ, which is also sometimes referred to as micro-homology mediated repair (MMEJ). Although the terminology is not used consistently in the literature, for the purposes of this thesis, A-EJ will be used.

It has been established that although A-EJ can function as a backup to NHEJ, its kinetics are considerably slower. By examining the process of V(D)J recombination, it has been established that when NHEJ is abrogated by homozygous knockout of *LigIV*, the kinetics of V(D)J with Lig I or Lig III are around 10x slower[184, 185].

1: Introduction

1.6: DNA repair

It was previously theorised that A-EJ was not one pathway, but rather a series of backup proteins with overlapping functions to the core factors of NHEJ, each of which could act as a backup mechanism for part of the canonical pathway. This theory is now largely discredited. If it were accurate, then knocking out different proteins in the NHEJ pathway would lead to different outcomes due to the idiosyncrasies of the various replacement proteins and it has been demonstrated that knockdown of XRCC4 and of LigIV produce the same outcomes in terms of change in mutational and translocative profiles. Further evidence that A-EJ is indeed a completely separate pathway comes from the identification of A-EJ proteins in *E. coli* species known to lack NHEJ apparatus[181, 185].

The proteins involved in the A-EJ pathway are also less understood than those of cNHEJ. As described above, there is evidence that PARP1 is important in initiating the pathway, although its recent implication as also being crucial for NHEJ casts its core role in A-EJ into some doubt. There is the possibility that PARP1 is a key initiator of both pathways and that other factors interact with it to determine pathway fate. XRCC1 and DNA ligases I and III have also been identified as essential for A-EJ, as has the DNA polymerase Pol θ [181] which plays a role in the inhibition of HR.

Although CHD1L has been linked to DSB repair generally via interaction with PAR chains, CHD2 has been linked to NHEJ[6], CHD4[186] knockdown has been shown to impair HR, and the p400 remodeller has been shown to *inhibit* A-EJ [187], the chromatin remodelling mechanisms which promote A-EJ remain to be described.

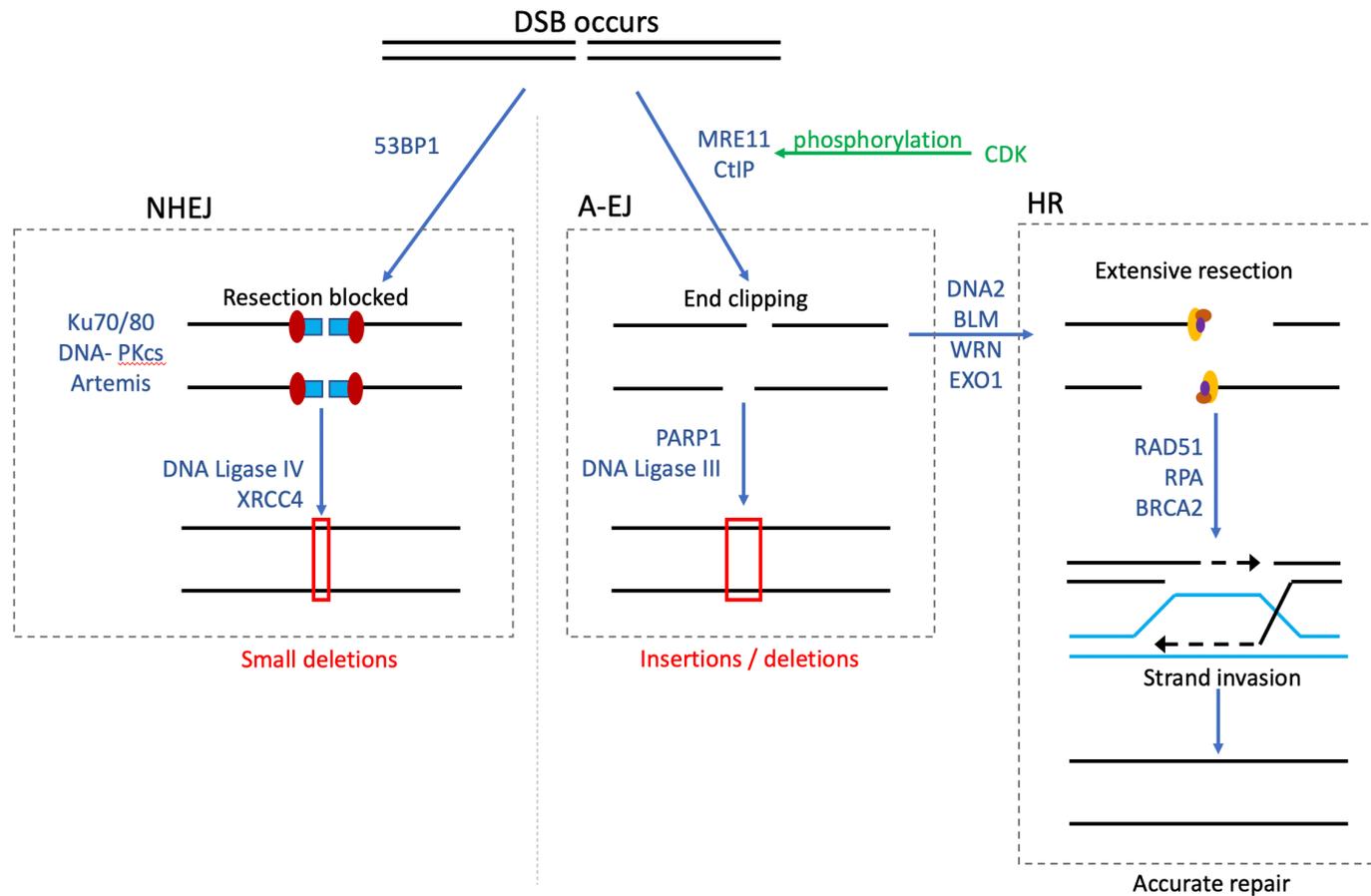


Figure 1.6 : Choice of DSB repair pathway and key proteins utilised. Cell cycle dependant phosphorylation of structure specific endonucleases MRN and CtIP by CDK promotes end resection and utilisation of A-EJ and HR. Adapted from Ceccaldi et al (2016) [169]

1.6.4 Summary of DSB repair

DSB repair is essential for cellular function, with possible consequences of poor repair including: cell death by apoptosis via p53 checkpoint activation, cell senescence, chromosomal translocations, chromosomal non-disjunction at mitosis, and introductions of indels of various sizes at the break sites. Accordingly, multiple mechanisms have evolved to repair DSBs and prevent sequence alterations or CNVs occurring.

The three currently identified pathways are complex and it has been established that chromatin remodelling plays an important regulatory roll in DSB repair. Further, it has been established that CHD2 knockdown can impair NHEJ.

Given that it has also been established that A-EJ exhibits a different and possibly more severe mutational profile than NHEJ and that an impairment of NHEJ should be detectable by examination of DSB repair at known cut sites.

1.7 Summary and Hypothesis

CHD2 is a chromatin remodeller. Heterozygous mutations in *CHD2* have been established as a cause of potentially severe neurodevelopmental disorders, epilepsy and psychiatric diseases in humans. Mice with similar mutations exhibit renal abnormalities and a significantly increased rate of lymphoma. Interestingly, zebrafish with analogous mutations exhibit a neurological phenotype and abnormal swim bladders.

Cell models demonstrate a role for CHD2 in the differentiation of muscle and neurons. It has been demonstrated that CHD2 binds at TSS associated with poised state chromatin, and can interact with master transcription factors to initiate differentiation. It has been associated with increased utilisation of histone variant H3.3 and knockdown increases the relative abundance of canonical H3.1.

Chromatin remodelling has been demonstrated to be an essential stage in the DDR cascade. In particular, CHD2-related chromatin remodelling has been described in association with the repair of DSBs via the most commonly utilised pathway in eukaryotic cell lines - NHEJ.

Where it has been established that DSB repair is important for maintaining the health of the genome, there is also evidence that it has a role in programmatic DNA rearrangements that occur in response to stimulation in circuitry associated with learning and memory. The evidence describes an association between these programmatic rearrangements and the neuroplastic response to external stimulation in learning experiments and fear conditioning in mouse models.

These findings form the basis for the hypothesis stated below, which will be explored in the remainder of this thesis.

1.7.1 Hypothesis

I propose that heterozygous mutations in *CHD2* will have an effect on the dynamics of cNHEJ, with an outcome that can be measured by observing the junctional sequences that form at the DSB repair sites. I further propose that this dysregulation of DSB repair could have an effect on the plastic function of mature neurons and that it could contribute to the neurodevelopmental disorder and epilepsy phenotypes exhibited by human patients with these heterozygous *CHD2* mutations.

Previous studies have demonstrated that in the absence of NHEJ, A-EJ can mediate programmatic DSB repair, albeit in a less efficient manner[188, 189]. By impairing the cNHEJ

1: Introduction

1.7: Summary and hypothesis

pathway, CHD2 mutations could therefore result in increased utilisation of A-EJ and possibly HR. As HR is cell cycle dependent, in mature post-mitotic neurons it is unlikely to contribute significantly towards DSB repair and therefore I predict that an increase in the utilisation of A-EJ will be demonstrable[178] if NHEJ is inhibited (*figure 1.7*).

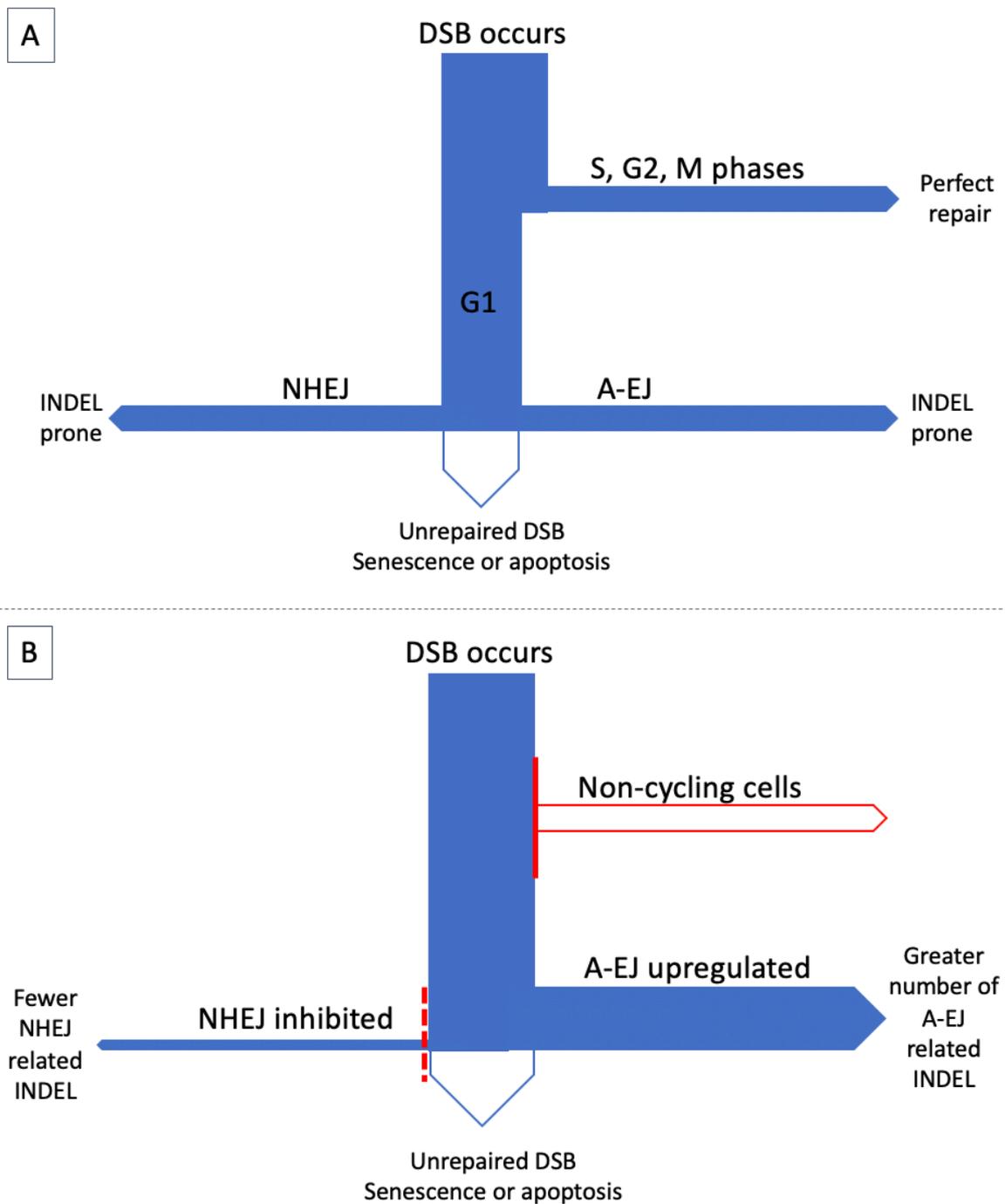


Figure 1.7: Graphical hypothesis. A – the usual process of DSB repair, depending on stage in cell cycle, breaks are repaired by one of three pathways, with unrepaired breaks triggering cell senescence of apoptosis via p53 checkpoint signalling. B – hypothesised result of CHD2 mutation in post-mitotic cells; as HR is unavailable, a greater number of breaks will be repaired by A-EJ with change in pattern of indels indicating this.

1.8 Aims and approach

To investigate the process of DSB repair, it is useful to think of the process from several angles: the characteristics of the repair, the dynamics of repair, the number of DSBs detectable and the genomic locations of the DSBs. When modelling neurodevelopment, it is useful to make observations in totipotent stem cells, pluripotent NPSCs and mature neurons. The process is increasingly well characterised and appears to be suitable for investigation of the cellular mechanisms underpinning human neurodevelopment[190]. There is evidence that neurons derived from hiPSCs more closely resemble foetal neurons than mature adult neurons, making them well suited to this endeavour[191].

The first aim was to design and validate a system that will facilitate sequence level analysis of break sites at high depth and throughput. Next-generation-sequencing (NGS) platforms have the potential to provide far more detail about the range of sequence variants that appear throughout a cell culture than older techniques such as Sanger sequencing. In chapter 3, I will describe the design and experimental validation of such a system.

The next aim was to establish a cell line that allows the creation of targeted DSBs. In chapter 4, I will describe the set-up of a cell line with a doxycycline inducible Cas9 endonuclease and the validation of this cell line through demonstrable editing in a variety of targets. I will demonstrate the use of this cell line to create a heterozygous CHD2 mutation on an isogenic background and characterise the behaviour of this cell line during differentiation via whole transcriptome sequencing (RNA-Seq).

The next aim is to use this inducible CRISPR endonuclease expressing cell line to investigate changes in DSB repair at pre-determined cut-sites. I targeted the endonuclease against a range of genomic targets known to be utilised during either neuronal differentiation, mature neurons, or both. For each target, I investigated the character of junctional sequences at break sites caused by the action of a single endonuclease. I also measured the rate of excisions performed between multiple cut sites within the same 1000 base pair region as a proxy for the investigation of repair dynamics – the reasoning underpinning this approach is explored further in chapter 5.1. In both cases, the experiments were performed in both hiPSC and in mature neurons, in order to explore changes in repair characteristics and dynamics between cycling cells and post-mitotic cells. In chapter 5 I will describe these techniques in detail and the outcome of these experiments.

1: Introduction

1.8: Aims and approach

To assess genome wide DNA stability regarding DSB occurrence, I utilised a novel method of capturing unrepaired DSBs *in situ* from within cell nuclei. This will be fully explored in chapter 6. A brief precis is that by ligating sequencing adapters directly to the DNA ends of fixed and permeabilised cells *in situ*, then using a sequencing flow cell to capture the ligated adapters, a library can be prepared without the requirement for PCR amplification or *in vitro* transcription. This unbiased capture approach allows for mapping of the geography of breaks and will identify fragile sites prone to DSBs – these can be matched to known CFS, loci associated with recurrent pathogenic CNVs, transcription start sites of active and suppressed genes and any other genomic feature of interest. Further, a comparison was made between the abundance and localisation of DSBs occurring in wild-type cells and in *CHD2* mutated cells. In section 1.6.2 I summarised the evidence that DSBs play a programmatic role in neuronal function – it is possible that genome-wide capture of unrepaired DSBs may provide some novel insight into this process and help to describe any perturbation related to aberrant chromosome remodelling. In chapter 6, I will explore the relationship between breaks captured and transcriptional programming in iPSCs at day 0, day 19 and day 40 of differentiation into neurons.

In chapter 7, I will synthesise the results chapters 3 - 6 to provide an overview of DSB repair changes between wild type and *CHD2* mutant cells. I will provide a detailed summary of how the new outputs from this project enhance our understanding of *CHD2* mutations and of DSB repair including repair in post-mitotic cells. Using this information, I will attempt to conclude as to whether dysregulation of DSB repair in *CHD2* mutant cells could be a significant contributor to the phenotype exhibited by patients harbouring germline heterozygous mutations in this gene.

2: GENERAL METHODS & MATERIALS

2.1 Introduction

In this section I will describe the methods common to all further chapters throughout the project. These include cell culture and differentiation techniques, DNA and RNA extraction, PCR optimisations and performance, nucleofection techniques, plasmid operations, molecular techniques, and data analysis tools.

Specific techniques relevant to each chapter are described in the introductory and methods sections of each chapter. The techniques described in these sections are more novel techniques requiring a greater degree of detail in their description; particularly those where the idiosyncrasies of the techniques may influence the results of the relevant experiments.

Similarly, specifics regarding the statistical and data analyses used for each set of experiments is included in the relevant methods sections, and full copies of any novel computation scripts included for perusal in the appendices.

The full list of reagents, consumables and equipment used can be found in *tables 2.1, 2.2 and 2.3* respectively. An attempt has been made to be as exhaustive as possible with regard to recording catalogue numbers to enable reproducibility, however this was not possible for all equipment. Where a catalogue number was not deducible from the companies' online catalogue '-' has been used.

Unless otherwise stated, all procedures described in this section were performed by the author of this thesis.

2: Methods

2.2: Tables of reagents and equipment

2.2 Tables of reagents and equipment

Reagent	Supplier	Catalogue Number
1D sequencing kit (R9.4.1)	Oxford Nanopore	SQK-LSK109(/108)
1D ² sequencing kit (R9.5.1)	Oxford Nanopore	SQK-LSK309(/308)*
2-Mercaptoethanol	Thermo Fisher Scientific	31350010
5-alpha Competent E.Coli	New England Biolabs	C29871
Accutase	Stemcell technologies	07922
Adenosine 5'-Triphosphate (ATP) 10mM	New England Biolabs	P0756S
Agarose	Sigma Aldrich	05066
Agencourt AMPure XP	Beckman Coulter	A63881
B27 supplement –RA	Thermo Fisher Scientific	17504001
B27 supplement +RA	Thermo Fisher Scientific	A3582801
Bolt 4-12% Bis-Tris Gels	Thermo Fisher Scientific	NW04120BOX
Bolt running buffer (x20)	Thermo Fisher Scientific	B000102
Bolt Transfer buffer	Thermo Fisher Scientific	BT00061
BSA, Molecular Biology Grade 20mg/mL	New England Biolabs	B9000S
Chloroform	VWR chemicals	67-66-3
Colorimetric Protein Assay kit	Thermo Fisher Scientific	22662
Corning Matrigel	Corning	356234
CutSmart Buffer	New England Biolabs	B7204S
DMSO	Sigma Aldrich	D2650
DNA QuickExtract	Lucigen	QE09050
Doxycycline	Sigma Aldrich	D9891
DTT (1,4-Dithiothreitol)	Sigma Aldrich	R08861
Dulbecco's Modified Essential Medium / F12	Thermo Fisher Scientific	12634028
E8 Flex medium	Thermo Fisher Scientific	A2858501
Essential 8 (E8) medium	Thermo Fisher Scientific	A1517001
Ethanol	VWR chemicals	64-17-5
Fibronectin	Sigma-Aldrich	F0162
Flow cell wash kit	Oxford Nanopore	EXP-WSH002
Gentle Cell	Stemcell Technologies	07174
Geltrex	Thermo Fisher Scientific	12760021
Gibco Dulbecco's Phosphate Buffered Saline (DPBS)	Thermo Fisher Scientific	14190250

Table 2.1 (page 1/3): List of reagents used. *Nanopore sequencing is a rapidly developing technology and several kits / flow cells became deprecated over the course of the project. The most recent catalogue number is used for references sake, however where possible in the relevant results section, the kit used will be addressed.

2: Methods

2.2: Tables of reagents and equipment

Reagent	Supplier	Catalogue Number
Gibco Neurobasal Medium	Thermo Fisher Scientific	21103049
Gibco OptiMEM Medium	Thermo Fisher Scientific	31985062
GoTaq G2 colourless master mix	Promega	M7832
Invitrogen Proteinase K 20mg/mL	Thermo Fisher Scientific	AM2546
Invitrogen Qubit DNA HS Kit	Thermo Fisher Scientific	Q32854
Invitrogen UltraPure™ BSA (50 mg/mL)	Thermo Fisher Scientific	AM2616
Isopropanol	VWR chemicals	67-63-0
KAPA Library Quantification Kit Platforms	Roche	7960255001
KAPA mRNA HyperPrep Kit	Roche	AA11190110
Laminin	Sigma Aldrich	L2020
LDN193189 (LDN) supplement	Stemcell Technologies	72147
Lipofectamine RNAiMAX	ThermoFisher Scientific	13778075
Methanol	VWR chemicals	67-56-1
MinION Flow cell 106	Oxford Nanopore	R9.4.1(/R9.4)*
MinION Flow cell 107	Oxford Nanopore	R9.5.1(/R9.5)*
miRNAeasy kit	Qiagen	217004
MycoZap	Lonza	LT07-918
N2 supplement	Thermo Fisher	17502001
NEB Quick Blunting Kit (1 reaction = 25µl)	New England Biolabs	E1201L
NEBNext Quick T4 DNA ligase module	New England Biolabs	E6056S
NextSeq 500/550 High Output Kit v2.5	Illumina	20024906
Nuclease Free Water	Promega	P1193
NuPAGE Sample reducing agent	Thermo Fisher Scientific	NP0004
P3 Primary Cell 4D-Nucleofactor X Kit	Lonza	V4XP-3012
Penicillin-Streptomycin-Glutamate (PSG)	Thermo Fisher Scientific	10378016
Penicillin-Streptomycin solution (PenStrep)	Sigma Aldrich	P433
Poly-D-Lysine	Thermo Fisher Scientific	A38890401
Pre-stained Protein Standard	New England Biolabs	P7719
Protease Inhibitor	Abcam	ab65621
PureYield Plasmid Miniprep kit	Promega	A1222
PureYield Plasmid Maxiprep kit	Promega	A2392
Puromycin	Thermo Fisher Scientific	A1113802
Reduced Growth Factor Matrigel	Corning	356231
RevitaCell	Thermo Fisher Scientific	A2644501
RIPA lysis and extraction buffer	Thermo Fisher Scientific	89900
SB431542 (SB) supplement	Stemcell Technologies	72232
SPRI beads: CleanNGS Beads / SPRIselect	GC Biotech/ Beckman Coulter	CNGS-0005/ B23318

Table 2.1 (page 2/3): List of reagents used. *Nanopore sequencing is a rapidly developing technology and several kits / flow cells became deprecated over the course of the project. The most recent catalogue number is used for references sake, however where possible in the relevant results section, the kit used will be addressed.

2: Methods

2.2: Tables of reagents and equipment

Reagent	Supplier	Catalogue Number
Sybr Safe	Thermo Fisher Scientific	S33102
T4 DNA Ligase Reaction Buffer 10x 6mL	New England Biolabs	B0202S
T4 DNA Ligase (2,000,000 units/m)	New England Biolabs	M0202M
Triton X-100	Thermo Fisher Scientific	85111
Tween 20	Thermo Fisher Scientific	85113
Ultra II End-repair / dA tailing module	New England Biolabs	E7546S
Versene	Thermo Fisher Scientific	15040066
Wizard SV Gel and PCR clean-up system	Promega	A9281
Y27632 dihydrochloride – ROCK inhibitor	Tocris	1254
Zymo Research Genomic DNA Clean & Concentrator	Cambridge Bioscience	D4011

Table 2.1 (page 3/3) : List of reagents used. *Nanopore sequencing is a rapidly developing technology and several kits / flow cells became deprecated over the course of the project. The most recent catalogue number is used for references sake, however where possible in the relevant results section, the kit used will be addressed.

2: Methods

2.2: Tables of reagents and equipment

Item	Supplier(s)	Catalogue Number(s)
6 well nunc cell culture plate	ThermoFisher	140675
12 well nunc cell culture plate	ThermoFisher	150628
24 well nunc cell culture plate	ThermoFisher	142475
96 well nunc cell culture plate	ThermoFisher	161093
100mm nunc petri dish	ThermoFisher	5500-0010
2mL Stripettes	Corning	4020
5mL Stripettes	Corning	4050
10mL Stripettes	Corning	4488
25mL Stripettes	Corning	4250
20 µL pipette tips	Clearline	713178B
200 µl pipette tips	Clearline	713179B
1000 µl pipette tips	Clearline	713180B
1.5mL Eppendorf tubes	ThermoFisher	69715
Cryovials	ISS	CRY012
50mL Falcon tubes	Corning	352070
15mL Falcon tubes	Corning	352095
30mL Universal containers	Starlab	E1412-3010
96 well PCR plates	ThermoFisher	AB0600
8 well PCR strips	ThermoFisher	AB0264G
PCR plate seals	ThermoFisher	AB0558

Table 2.2 : List of laboratory consumables used

2: Methods

2.2: Tables of reagents and equipment

Equipment	Manufacturer	Catalogue Number
Strippette gun	ThermoFisher	9501
Research plus 1000µL pipette	Eppendorf	31240000121
Research plus 200µL pipette	Eppendorf	31240000083
Research plus 20µL pipette	Eppendorf	31240000032
Research plus 10µL pipette	Eppendorf	31240000016
Xplorer plus 1200µL multichannel pipette	Eppendorf	4861000830
Xplorer 1200µL electronic pipette	Eppendorf	4861000732
Xplorer 100µL electronic pipette	Eppendorf	4861000716
E4 10µL electronic pipette	Rainin	E4-10XLS+
Xplorer plus 100µL multichannel pipette	Eppendorf	4861000139
Xplorer plus 0.5-10µL multichannel pipette	Eppendorf	4861000112
Galaxy 170 R Incubator	New Brunswick scientific	EPCO170R-120-0000
Sub Aqua Pro heat bath	Grant	15147025
Gel tank for electrophoresis	VWR	700-0136
Mini-Gel tank for western blot	Invitrogen	NW2000
Transfer tank for western blot	BioRad	EPCO170R-120-0000
Power source (300V)A for electrophoresis and WB	VWR	700-0113
Mastercycler x50s	Eppendorf	6311000045
iBright CL1000	Invitrogen	A32749
Odyssey CLx	Li-Cor	-
Nanodrop biospectrometer	ThermoFisher	-
Qubit Biospectrometer	Eppendorf	61350000025
AMAXA 4D nucleofactor	Lonza	AAF-1002B/X
Megafuge 40R	ThermoFisher	75004503
Centrifuge 5810R	Eppendorf	5810000060
Mini Centrifuge 5434	Eppendorf	5828000210
Innova 44 shaker	New Brunswick Scientific	M1282-0002
Grant	Heating block	-
Vortex Genie 2	Scientific Industries	-
AX2202 Scale	Sartorius	-
Magnetic rack	Millipore	20-400
Safe 2020 laminar flow hood	Thermo Scientific	51026637
CKX41 Microscope	Olympus	-
NextSeq 550	Illumina	SY-415-1002
MinION	Oxford Nanopore	-

Table 2.3 : List of equipment used

2: Methods

2.3: Standard cell culture procedures

2.3 Standard cell culture procedures

For general purposes, the volume of media and other reagents used are appropriate for a single well of a 6 well plate. For scaling of these numbers, please refer to *table 2.4*.

Plate size	Culture medium volume per well (mL)	Basement matrix / PBS for wash / dissociation reagent volume per well (mL)
6cm dish	12mL	6mL
6 well plate (6WP)	2mL	1mL
12 well plate (12WP)	1mL	500µl
24 well plate (24WP)	500µl	250µl
96 well plate (96WP)	120µl	60µl

Table 2.4: Reagent volumes and scaling for plate sizes used in cell culture

2.3.1 hiPSC culture and maintenance

2.3.1.1 Cell line description

Our work is conducted using the IBJ4 human induced pluripotent stem cell line. This was a gift from Josh Chenoworth at the Lieber Institute for Brain Development, MD, USA.

The hiPSC line was derived from the BJ fibroblast line (ATCC #CRL-2522) using the non-integrating STEMCCA Cre-Excisable Constitutive Polycistronic Lentivirus Kit (Millipore, #SCR531).

2.3.1.2 Plate preparation

hiPSCs were cultured on Nunc Plates (ThermoFisher). Plates are coated with either Geltrex or Matrigel. In both cases, 60µL of matrix is suspended in 6mL of DMEM/F-12, 1mL of the suspension is used to coat each well. The plates are incubated at 37°C for an hour and washed with PBS before use. Matrigel matrix can be used to prepare a second plate. The choice of basement membrane used by the lab changed over the course of the project, however all cell lines were transitioned to the new matrix simultaneously in order to minimise the variability that this may have engendered.

2.3.1.3 Culture media

- **E8** Used for general maintenance of stem cell cultures. E8 is changed 24 hours after passage of cells, and every 48 hours thereafter until cells are confluent and ready for further passage or use.

2: Methods

2.3: Standard cell culture procedures

- **E8 Flex** E8 with stabilised foetal growth factor – used for high volume cell culture, allowing for longer periods between change of media. Where used, this will be clearly stated
- **N2B27** Neuronal differentiation medium prepared using 100mL DMEM/F12, 50mL Neurobasal, 1mL N2 supplement, 1mL B27 supplement (+ / - RA depending on the stage of differentiation), 1.5mL PSG and 150µl b-mercaptoethanol

2.3.1.4 Passaging techniques

- **Gentle Cell** Used when passaging stem cells. For routine passage (either 1:3 or 1:6) of hPSC, media is aspirated and replaced with Gentle Cell. The cells are incubated at 37°C for 1-2 minutes before Gentle Cell is aspirated and the cells washed with PBS. 1mL media is added and are suspended by gentle scratching with a 5mL strippette. The cells are diluted to 3mL or 6mL and split into destination wells (1mL per well). A further 1mL of media is added to each well for a final volume of 2mL per well. Media is changed 24 hours after passage.
- **Versene** Used when passaging other cell types, such as NPCs or neurons. The media is aspirated and the cells washed with PBS. 1mL of Versene is added to the cells, which are incubated for 2 minutes at 37°C. Versene is aspirated and the cells are washed with 1mL PBS before the cells are suspended in media by gentle scratching with a 5mL strippette and split appropriately between destination wells.
- **Accutase** Used when a single cell suspension is required. Media is aspirated and cells washed once with PBS. 1mL accutase is applied and the cells incubated for 9 minutes at 37°C. Culture media is added to stop the accutase reaction and cells are broken into a single cell suspension by gentle pipetting before being transferred to a 1.5mL Eppendorf tube.

Cells are centrifuged at 200g for 5 minutes. The media and accutase is aspirated, leaving a pellet, which is re-suspended in the required media for use.

2: Methods

2.3: Standard cell culture procedures

2.3.1.5 Freezing cells

Cells are treated with ROCK-inhibitor for 1 hour, then washed with PBS and dissociated using 1mL gentle-cell. After 1 minute, gentle cell is aspirated and cells are washed with 1mL PBS before suspension in E8.

The cell suspension is centrifuged at 400rpm in a 15mL falcon tube for 5 minutes, then re-suspended in 0.5mL E8. 0.5mL E8 containing 20% DMSO is added, making a final cell suspension of 1mL E8 with 10% DMSO. Cells are transferred to a cryovial and frozen overnight in the -80°C freezer unit before transferring to liquid nitrogen for long-term storage.

2.3.1.6 Thawing cells

A 6WP is prepared with Matrigel, washed with PBS and loaded with 1mL media containing 1:50 Revitacell (10µL).

Cells are removed from liquid nitrogen storage and transferred to tissue-culture on dry-ice. The cryovial is then warmed in a 37°C water bath until only a small ice-crystal remains.

The thawed cell suspension is transferred to a 15mL Falcon tube and diluted with a further 9 mL culture media to dilute out the DMSO, which is toxic at room temperature.

The dilute cell suspension is centrifuged at 400rpm for 5 minutes. The media is aspirated, and cells are re-suspended in 1mL fresh E8 media, before transfer to the prepared plate. The final concentration of Revitacell in the receiving well is 1:100.

2.3.1.7 Puromycin kill curve

When testing for integration of constructs, it is useful to know the concentration of certain molecules which will cause cell death in wild type cells.

WT stem cells were plated at a density of 100,000 per well. An ascending dose of puromycin, from 0.1µg/mL to 0.5µg/mL was added to the wells. Light microscopy was used to find the well with the lowest concentration of puromycin in which complete cell death occurred. This demonstrated complete cell death in wild type cells at a concentration of 0.4µg/mL.

2: Methods

2.3: Standard cell culture procedures

2.3.2 Differentiation of hPSC into neurons

2.3.2.1 Plating and daily care

For the first stage of differentiation into neurons, 12WP are prepared with reduced growth-factor Matrigel. Cells are passaged onto these plates using the technique described above and grown in E8 culture medium until confluent.

Once the cells have become confluent, they are washed with PBS and maintained in 1.5mL of N2B27 media. 0.75mL of the media is changed every 48 hours. For the first 4 days, this is supplemented with LDN at 1:10,000 and SB at 1:2000. For days 6-12, only LDN is added. From day 12 to day 26, un-supplemented N2B27 is used.

After day 26, the media is changed to N2B27 with RA supplementation. This is continued until the cells are fully differentiated.

2.3.2.2 First Passage – Day 9

After 9 days, the differentiating cells are passaged onto fibronectin coated 12 well plates prepared with 15µg/mL fibronectin in PBS, incubated for 1 hour and washed x1 with PBS.

1 hour prior to passage the media is changed, and the cells treated with ROCKi. At the time of passage, this conditioned media is collected in a 15 mL falcon tube. The cells are treated with versine for 1-2 minutes at 37°C then washed with PBS.

The wells are scratched and re-suspended in the saved media, then diluted to 2:3 and passaged onto the new plates. After 24 hours, the media is changed to remove the ROCK-i and cell debris.

2.3.2.3 Second Passage and Subsequent Passages – Day 19

Passage is carried out as per day 9, except onto plates coated with Poly-D Lysine-Laminin. The plates are prepared by incubation with 1:100 poly lysine in PBS for 2 hours. After 2 PBS washes, the plates are incubated with 1:100 laminin in PBS. After a further 2 PBS washes to remove any toxic remnants, the cells are passaged as described above.

2.3.3 Additional substances used in cell culture

See *table 2.5* for volumes / concentrations used, scaled to plate size.

- Rock inhibitor y-27632 (various sources)
The apoptosis inhibitor Y-27632 reduces cell death by its action on the Rho-kinase apoptotic pathway. It is used during passage or other treatment of cells that could result in a high degree of cell death.
- RevitaCell Supplement (ThermoFisher #A2644501)
RevitaCell is a proprietary ROCK inhibitor with greater pathway specificity than Y-27632. It is sold in solution with antioxidant and free radical scavengers and offers an improvement in cell recovery from thawing and single cell passaging
- Penicillin-Streptomycin (Pen-Strep)
Used to treat infection in cell culture wells, where cultures were important, for example latter stages of experiments requiring long lead times, or cell lines yet to be expanded. In general, if cells were simply cycling as stock or near the start of easy-to-repeat experiments, the cultures were simply disposed of in order to reduce the impact of an unmatched variable
- Puromycin
Used as a selection marker for cells transfected with puromycin-resistance encoding plasmids. The required concentrations were decided using a kill-curve experiment, detailed in chapter 3
- Doxycycline
Used to induce expression from TET switch containing plasmids, integrated into the host genome. For validation of the concentrations used, please see chapter 3

2: Methods

2.3: Standard cell culture procedures

Plate size	ROCK inhibitor y27632 volume per well: dilution factor (μL , ratio)	Revitacell supplement volume per well (μL)	Pen-Strep volume per well (μL)	Doxycycline volume per well (μL)
6cm dish	24	60	120	Not used
6 well plate (6WP)	4	10	20	4
12 well plate (12WP)	2	5	10	2
24 well plate (24WP)	0.5	2.5	5	1
96 well plate (96WP)	0.125	1.25	2.5	Not used
Dilution factor	1:500	1:200	1:100	1:500

Table 2.5: Scaling of dilutions for cell culture additives by plate size

2.3.4 Nucleofection procedures

Several methods exist for the introduction of proteins and DNA constructs into cell lines. The choice of method depends on the cargo to be transfected and the cell line being transfected. Although variations exist, the methods roughly segregate into: lipofection, in which a cationic lipid formation is associated with the desired cargo and added to the cell culture, electroporation, in which an electrical pulse is used to disrupt the cell membranes and allow diffusion of the required cargo, and viral transfection in which a viral vector is used to infect the cell culture with the required cargo. During the course of this project, lipofection and nucleofection were used.

The method chosen for each transfection and the reasoning behind each decision will be covered in the specific methods section of each results chapter.

2.3.4.1 Electroporation

Electroporation (also referred to as nucleofection) was performed using the Lonza Amaxa 4D nucleofactor unit. The P3 primary cell 4D-Nucleofactor X Kit was used for all transfections.

Before starting, the cells were treated with Y27632 rock inhibitor for a minimum of 1 hour and a destination plate was coated with matrigel, then filled with 2mL pre-warmed culture media. Y27362 was also added to the pre-warmed media. Four destination wells were used for each reaction.

2: Methods

2.3: Standard cell culture procedures

Nucleofection solution is prepared by combining 82 μ L Lonza P3 solution with the 18 μ L of the provided supplement. A maximum of 10 μ L cargo to be nucleofected is added to the solution.

One confluent well from a 6 well plate was used for each nucleofection. A single cell suspension was made using accutase, following the protocol described in section 2.1.1.4, with a small modification; the pellet made after the first centrifugation step is re-suspended in PBS and centrifuged a second time to wash any trace of accutase and medium from the cells as this can interfere with the electroporation.

The pellet is re-suspended in the nucleofection solution and transferred to the cuvette. The cuvette is placed in the nucleofector and pulse program Ca-137 is used – this is the pulse code recommended by Lonza for use with hiPSC stem cells – regrettably it is not possible to determine the settings used from the code provided as this is treated proprietorially by Lonza.

The cells are transferred dropwise into the prepared plate using a Pasteur pipette. 24 hours after nucleofection, the media is changed to remove the Y27632. Cell selection is begun after 48 hours, to give the plasmid enough time to begin expression its selection marker.

2.3.4.2 Lipofection

Lipofection was performed using Lipofectamine RNAiMAX. Cells were passaged into a 24-well plate one day prior to lipofection. The passage-media was aspirated and replaced with 400 μ L E8 – E8 flex is not used as there is some evidence it can inhibit the transfection reaction.

5 μ L RNA is diluted into 45 μ L optimem in a 1.5mL Eppendorf tube. 3 μ L lipofectamine RNAiMAX is diluted in a second tube with 47 μ L optimem. The dilutions are combined to make a 100 μ L solution containing the RNA and the lipofectamine. This is incubated for 5 minutes at room temperature to allow complexes to form.

The 100 μ L is then added drop-wise to the well of the 24-well plate. Media is changed after 24 hours to remove the potentially toxic lipofectamine.

2.3.5 Cell lines and passage numbers for comparisons between wild type and CHD2 deficient lines

Throughout this thesis, a comparison is made between two cell lines; a wild type cell line with inducible Cas9 cassette at AAVS1 locus, and the same cell line harbouring a mutation in *CHD2*. Detailed methods for the creation of these cell lines can be found in chapters 3 and 4. Their characterisation by RNA-Seq is detailed in chapter 4.

At all stages, comparisons are made between cell lines matched for passage number +/- 3 passages, in order to control for age related divergence. The WT cells are of the same isogenic background as the *CHD2* mutant cells[192].

The inducible Cas9 cassette was ligated into AAVS1 at passage 36. The characterisation by RNA-Seq (chapter 4) and the analysis of DSB occurrence in differentiating cell lines (chapter 6) and the analysis of DNA repair in mature neurons (chapter 5), was performed on cells differentiated from the same starter culture. The first stage of the analysis of repair of targeted lesions (chapter 4) was performed on cells at passage number 42.

2.4 Molecular biology protocols

This is the general protocol used to prepare plasmids for nucleofection into our cell lines. Details regarding plasmid choice and manipulation can be found in the methods sections of the relevant results chapters. The majority of this work was undertaken with pre-ligated plasmids, available commercially. Where plasmid assembly took place, this will be described in the relevant section

2.4.1 Transformation of bacteria with required plasmid

5-alpha Competent *E. coli* (High Efficiency) (New England Biolabs #C29871) are defrosted on ice for 10 minutes. 25µL of *E. coli* is mixed with 1-5µL plasmid, and the tube flicked 5 times. The mixture is incubated on ice for 30 minutes, then heat-shocked for exactly 30 seconds at 42°C in a pre-warmed heating block, then returned to ice for a further 5 minutes.

950µL of Super Optimal broth with Catabolite repression (SOC) media is pipetted, and the cultures incubated at 37°C for 60 minutes. The cultures are shaken at 250rpm, inverted to mix cells and diluted 1:6 in LB broth.

50-100µL of each dilution is spread onto a selection plate and incubated for 16 hours at 37°C. The next morning, colonies are picked for maxiprep or miniprep.

2.4.2 Mini-prep

Mini-preps were performed using the PureYield MiniPrep kit (Promega # A1222), using the standard protocol. Starter cultures are grown in LB broth containing the appropriate selection antibiotic, at 37°C, in a shaker at 250rpm.

600µL of culture is transferred to a 1.5mL micro-centrifuge tube and mixed with cell lysis buffer by inverting 6 times. 350µL neutralisation solution is added and mixed by inversion. The column is centrifuged for 3 minutes and the supernatant transferred to the mini-column. Endotoxin and column washes are applied to the column sequentially and centrifuged. 30µL of elution buffer is used to elute the plasmid DNA.

Extracted DNA was stored at -20°C until required.

2: Methods

2.4: Molecular biology protocols

2.4.3 Maxiprep

250 μ L starter culture is incubated in 100mL LB broth containing appropriate selection marker, overnight in a shaker at 37°C / 250rpm. Cells are pelleted by centrifuge at 5000g for 10 minutes.

Cell pellet is re-suspended in 12mL of resuspension solution and 12 mL cell lysis solution is added. The tube is inverted 3-5 times to mix, then incubated at room temperature for 3 minutes.

12mL neutralisation solution is added to stop the reaction. Cell lysate is centrifuged for 20 minutes at 14000g at room temperature in a fixed angle rotor.

DNA is purified by using an assembly of clearing column stacked onto a binding column. The lysate is placed in the clearing column and the assembly placed in a vacuum manifold. The vacuum draws the lysate through the clearing column, then the binding column.

The clearing column is discarded, and the binding column washed sequentially with 5mL endotoxin removal wash and 20mL column wash. The membrane is dried for 5 minutes.

The binding column is assembled to draw flow-through into a 1mL Eppendorf tube in the Eluator Vacuum Elution Device. 1mL nuclease free water (NFW) is added to the membrane and incubated for 1 minute. The vacuum manifold is used to elute the DNA from the membrane.

2.4.4 Plasmids for inducible Cas9 integration

AAVS1-TALEN-L (Addgene # 59025) and AAVS1-TALEN-R plasmids (Addgene # 59026) and the pAAVS1-PDi-CRISPRn plasmid (Addgene # 73500) were prepared from bacterial stabs (*figure 2.1*). These were nucleofected into the cells by electroporation (*see section 2.3.4.1*)

Electroporated cells were seeded onto matrigel and cultured in standard E8 medium for 48 hours to allow plasmid expression to reach functional levels. Cells were then treated with puromycin selection marker for 7 days, with a matched un-transfected control.

Emerging colonies were passaged and expanded. The expanded cultures were frozen down and placed in liquid nitrogen to create a stock of WT-iCas9 cells for further work.

2: Methods

2.4: Molecular biology protocols

A:



B:

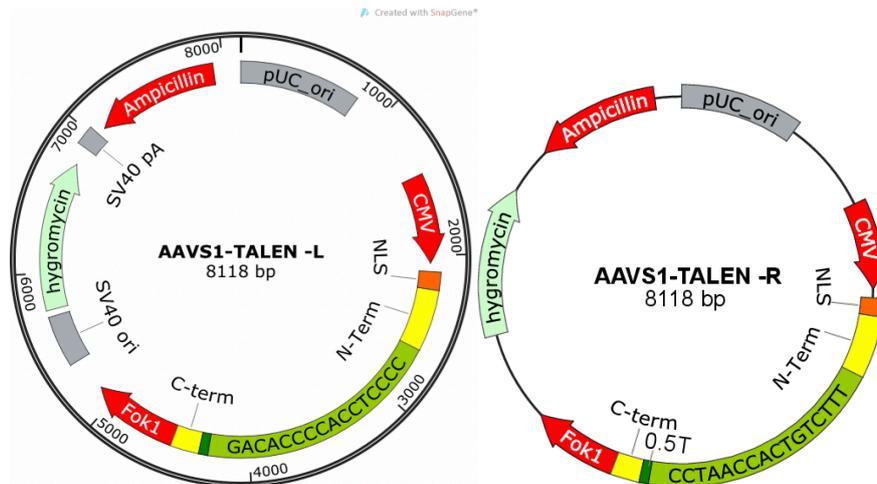


Figure 2.1: schematic view of plasmid used to create TET inducible Cas9 insert. Key domains include: AAV 3'/5'; Adeno-Associated Virus (AAV) integration site 1 homology sequences, TREG3; tetracycline induction switch, 3xFLAG; FLAG protein tag for detecting production of Cas9 with standardised antibody system, Cas9; CRISPR Associated protein 9 nuclease (without gRNA), PuroR; puromycin resistance protein. B: TALEN-L and TALEN-R plasmids with sequences targeting AAVS1 locus (light green)

2: Methods

2.4: Molecular biology protocols

2.4.5 Validation of plasmid integration

2.4.5.1 Plasmid integration

To identify whether the construct was spliced into the correct location in the host genome, PCR primers were selected to create an amplicon crossing the junction between the *AAVS1* locus and the integrated sequence.

2.4.5.2 Cas9 protein production

Western blot and immunohistochemistry were used to examine Cas9 production. The Cas9 protein has a FLAG tag sequence attached, to allow for ease of detection. Putative WT-iCas9 cells were plated in on a 12WP and treated with doxycycline for 48 hours, then washed in PBS and maintained in E8 for a further 72 hours.

Protein extraction was conducted from three wells every 24 hours, yielding three samples at each time point: samples at 0 hours, 24 hours of treatment, 48 hours of treatment, 24 hours post treatment, 48 hours post treatment and 72 hours post treatment. Controls not treated with Doxycycline but plated at the same time as the experimental wells were also extracted. Protein was extracted and quantified, and Western Blots run according to the protocols detailed in *section 2.8*.

Mouse anti-FLAG-M2 primary antibodies were used, with Rabbit anti-GAPDH as a control. Li-Cor secondary Abs at 680nm and 800nm were used to generate the signal. To generate the dose response curve, an average ratio between GAPDH and FLAG was taken across the three samples extracted at each time point. The dose response curve was plotted using Microsoft Excel.

Cells were fixed for immunohistochemistry after 48 hours of treatment with doxycycline. The same primary antibody against FLAG-M2 was used for imaging of Cas9, with rabbit anti-mouse secondary.

2.5 Genetics and Genomics Protocols

Different sequencing platforms are used throughout this project where appropriate. In this section I will provide a summary of the machines and materials used. A detailed introduction to the relevant technologies and justification for their use for each part of the project will be given in the relevant results chapter methods sections.

2.5.1 DNA extraction

DNA due to be used for PCR or amplicon sequencing can be extracted using Lucigen DNA QuickExtract solution. This extraction is rapid and therefore highly scalable, however as there is no clean-up step the DNA extracted is continuously exposed to cellular enzymes and therefore prone to degradation. Where the extracts will be used immediately, the rapid turn-around of this method makes it desirable.

For PCR, DNA is extracted using Lucigen DNA QuickExtract solution. Cells are washed with PBS, then suspended in 500 μ l of the solution and transferred to a 1.5mL Eppendorf tube.

The suspension is vortexed for 15 seconds, then incubated at 65°C for 5 minutes in a heat block. The suspension is vortexed for a further 15 seconds then heated to 98°C and allowed to cool on the desktop before freezing.

2.5.2 RNA extraction

RNA extraction is performed using the QIAzol Lysis reagent (QIAGEN #79306). Confluent cells are washed x2 with 1mL PBS and then scratched to suspend. Cells transferred to a 1.5mL Eppendorf tube and centrifuged at 200g or 5 minutes. The supernatant is aspirated, taking care not to disturb the pellet.

The cell pellet is re-suspended in 500 μ L QIAzol lysis reagent and pipetted. The mixture is incubated at room temperature for 5 minutes. At this stage the extraction is frozen at -80°C until all samples are ready for use.

To complete the extraction, 100 μ L chloroform is added to the defrosted homogenate and the mixture shaken vigorously for 15 seconds, before incubating at room temperature on the desktop for 3 minutes.

2: Methods

2.5: Genetics and genomics protocols

The mixture is centrifuged at 12,000g for 15 minutes at 4°C and the aqueous upper phase transferred to a new tube. 250µL isopropanol is added and the mixture vortexed, before a further 10-minute incubation step at room temperature on the bench top.

The mixture is centrifuged at 12,000g for 10 minutes at 4°C. The supernatant is discarded and the 500µL molecular grade ethanol (75%) is added.

The mixture is centrifuged at 7500g for 5 minutes at 4°C, the supernatant is aspirated, and the RNA pellet is allowed to air-dry to remove any residual ethanol. The pellet is then dissolved in RNase free water for clean-up and further use.

2.5.3 PCR Optimisation

Unless otherwise stated all PCR reactions were performed with GoTaq G2 colourless master mix (Promega #M7832).

Optimisations were performed using two pairs of primers (A forward, A reverse, B forward, B reverse) designed using Benchling's primer wizard [193] and chosen for desired amplicon size and low primer penalty score.

Reactions were plated in the first four columns of a 96 well plate, allowing for four combinations of primer to be tested at a range of temperatures using along a gradient on a thermal cycler. Reactions were run on the MasterCycler X50 (Eppendorf #631000042). Estimated annealing temperature was calculated using the formula $T_m - 5^\circ\text{C}$. The primers were then optimised along a gradient of temperatures $\pm 4^\circ\text{C}$ from this theoretical melting point (*figure 2.2*).

In accordance with the standard GoTaq enzyme protocol, initial T_m was 95°C for 2 minutes, followed by 32 cycles consisting of 30 sec 95°C T_m , 30 sec TA, 1 minute 72°C extension, followed by 5 minutes extension step at 72°C .

Agarose gels for sizing of DNA products and measuring success of PCR reactions are made using 1% high-quality agarose suspended in TAE buffer. This is microwaved until the agarose has melted and is clear of air bubbles. Invitrogen SYBR-safe gel stain (ThermoFisher #S33102) is added at 1:10,000.

NEB 100bp (NEB# N3231S) or 1kb (NEB# N0468S) DNA ladder is used as a marker of amplicon size, and samples are mixed in a ratio of 1:6 with promega PCR loading dye (Promega #G1881) or NEB PCR loading dye (NEB #B7025S) prior to loading onto the gel.

Gels are run in TAE buffer for 45 minutes at 90V. Images are captured using the Invitrogen iBright CL1500 imaging system and transferred to a workstation for analysis via the Thermo Fisher Cloud service. Images are then manipulated in ImageJ [194] in order to provide the best contrast for detection of secondary amplifications.

2: Methods

2.5: Genetics and genomics protocols

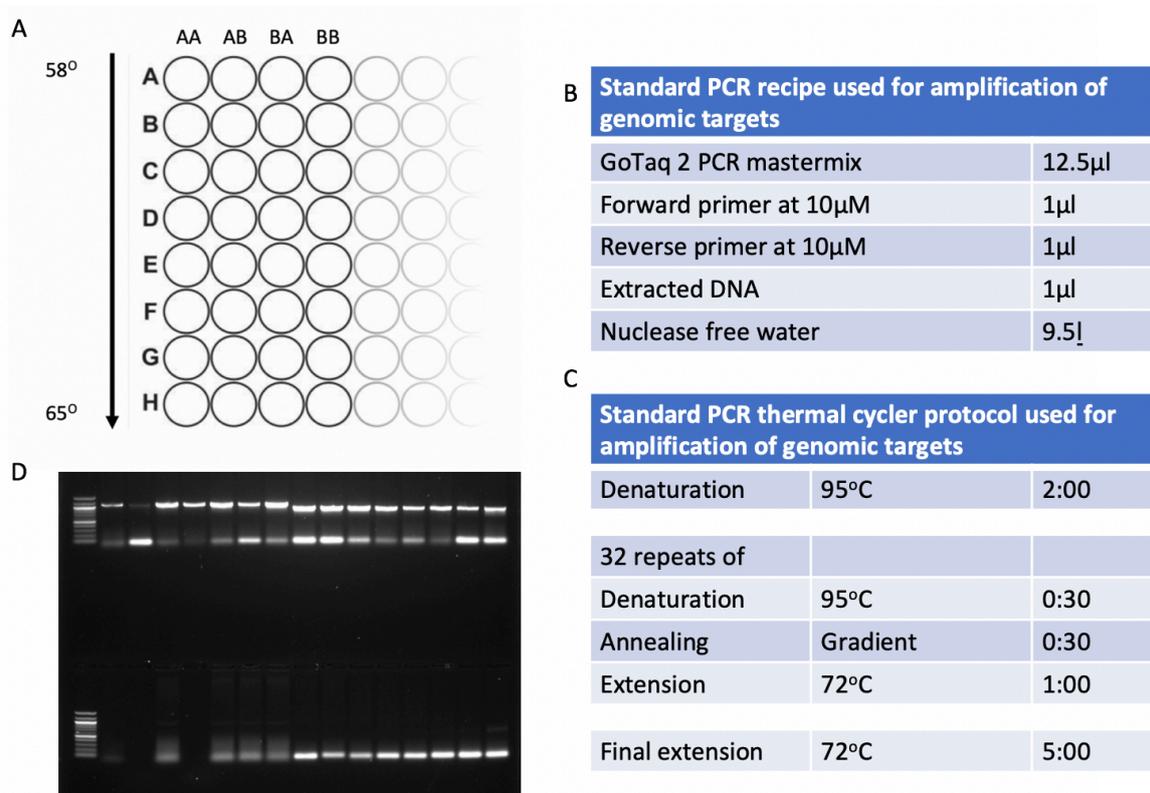


Figure 2.2 –for PCR optimisation and example output. A: schematic representation of PCR plate with primer pairs arranged by row and temperature gradient programmed by column, B: PCR mixture recipe, C: Thermal cycle program, D: output for optimisation of CSMD3 primers

2: Methods

2.5: Genetics and genomics protocols

TARGET	PRIMER	SEQUENCE	OPTIMAL T _A
AUTS2	FWD A	GTGATATCGTCTGAGGATGG	55.9°C
	FWD B	GGTTCATTTTCTAGCCTGAC	
	REV A	CCATGCTGTTCTCATGATAG	
	REV B	GAGGAGTTCCTCTGCACAAG	
CHD2_A	FWD A	TTCAAGCGATCCTTCTGCCTCA	63.1°C
	FWD B	TAGACTGCAGGTGTGTGCCAC	
	REV A	GCTAGCCTAGGACAAGCTAGCA	
	REV B	ACCATCCAATTATAGGGAGGAAAGA	
CHD2_B	FWD A	AAAGGGTGGCAGTGGTTCTCAT	58.7°C
	FWD B	AGCCGCTGTCTTACCTTTCTGT	
	REV A	CCTGAAAGAGGAGGGAGTGCTT	
	REV B	ACTGAACTTCAGGCATGAGGGT	
CSMD3	FWD A	TGGTGGTTGCTATAACGAGCCA	65°C
	FWD B	CCCGTTGGAGAGATGGTTGAGT	
	REV A	ATGTTGATGTGGCAGTGCTTGG	
	REV B	TACCACAGGTCCTAGGGAAGCT	
NRXN1	FWD A	GGGCTGTTTTCTTCTGTGCCTC	63.1°C
	FWD B	CTCACTGCTTCCTTTTGGCCTG	
	REV A	CTGGTGCATGGGGTCTTCCTAA	
	REV B	TGTTTCAGCGGCCTTTTCGTC	
PARK2	FWD A	GTTGGTCAGTTACATGTCAC	58.7°C
	FWD B	GCTTCCCTTCTACCACGGAG	
	REV A	GAAATGGGTGGTAGCATAGAG	
	REV B	GTTAGGCTTCATCTTGCAGG	

Table 2.6: Primer sequences used for PCR amplification – primers in red are the pair chosen – the last column contains the optimal annealing temperature for each reaction. Note annealing temperatures are for sequences above with additional 15bp of M13F/R

2.5.4 DNA product clean-up and quantification

2.5.4.1 Spin column Purification

Spin column purification is a procedure by which DNA samples are adsorbed by a silica membrane and washed to remove PCR primer dimers and reagents from upstream, which may interfere with downstream experimental procedures. Several kits are available, following similar principals – for this thesis, the Wizard SV Gel and PCR clean-up system (Promega #A9281) were used unless stated otherwise.

Samples were mixed with an equal volume of proprietary membrane binding solution and applied to the spin column membranes. The samples were centrifuged to remove all remaining reagents, then washed twice with proprietary wash solution, containing alcohol. Finally, the samples were eluted in a small volume (30-50µl) of nuclease free water (NFW) for quantification and downstream analysis.

2.5.4.2 Magnetic bead clean-up

Solid phase reversible immobilisation beads (SPRI) are polystyrene beads, with a magnetite coating. They reversibly bind DNA in solution, aided in combination with a crowding agent such as glycerol (such agents are often packaged as ‘binding buffers’ or similar).

The beads are paramagnetic and were attracted to a magnet on a rack holding tubes. The beads pellet and can be washed with ethanol to remove impurities. The ratio of beads : solution can also be used to size-select DNA strands of different lengths, making it useful for various DNA library preparation procedures.

Various products exist, and some are included in library preparation kits. For nanopore sequencing, Ampure XP beads (Beckman and Coulter #A63880) are used. For INDUCE-Seq, SPRI beads (ABM #G951) are used, and for RNA-Seq, beads are included in the KAPA library prep kit.

2.5.4.3 Qubit quantification

Qubit fluorimetry uses fluorescent dyes to determine the quantification of DNA or protein in a sample. UV absorbance was measured at 260 and 280nm, in a sample presented in a specialised tube.

2: Methods

2.5: Genetics and genomics protocols

2.5.4.4 BioSpectrometer

The Eppendorf BioSpectrometer works on similar principals. Samples to be tested are loaded into a cuvette between two hydrophobic surfaces, to form a column of fluid. Spectrometry using selectable wavelengths is then used to quantify the contents and purity of the sample.

Purified PCR products tend to register a concentration of 200-1000ng/ μ l. Purity is given in the form of ratio of absorbances (A260/A230 and A260/280).

Although the accuracy of the BioSpectrometer is felt to be lower than that of the Qubit, it is quicker to use and suitable for most applications.

2.5.5 Sequencing

Oxford Nanopore sequencing was used for chapters 3 and 5, and Illumina NextSeq for chapters 4 and 6. A discussion of the relative benefits of nanopore sequencing can be found in the introduction of Chapter 3. Here follows a summary of the technical aspects of library preparation.

The nanopore library preparations were novel in several ways, which will be addressed. The sequencing whole transcriptome sequencing performed for chapter 4 was performed according to well established protocols, which will be referenced but not transcribed. The library preparation for INDUCE-Seq (chapter 6) is novel and is detailed in the chapter-specific methods, however the sequencing run itself is performed according to standard protocols, which again are not transcribed.

2.5.5.1 Oxford Nanopore MinION library preparation

DNA was extracted using DNA Quick Extract (see above), and primary PCR, optimised prior to extraction, is used to amplify the region of interest (800bp – 1200bp) and ligate M13F and M13R tags.

The product from primary PCR was diluted by a factor of 1:100 in NFW (dilutions of 1:1000 are used if this is not successful). Secondary (or barcode) PCR was used to ligate 15bp barcodes to the 3' and 5' ends of each sequence (*tables 2.7 & 2.8 & figure 2.3*) so that up to 96 samples could be run simultaneously. Such approaches are well established for tracking samples in high-throughput sequencing experiments[195]

Samples were then pooled and cleaned with Promega Wizard SV Gel and PCR Clean-Up system, in order to remove primer-dimer artefacts and reagents that may interfere with the library preparation.

The pooled and cleaned samples were quantified on the Eppendorf Nanodrop Spectrometer. 250 femtomoles of DNA was taken forward for library preparation (usually around 500nL).

The library preparation follow was performed using either the SQK 109 (1D) or SQK 309 (1D²) kits. The underlying scientific principles of nanopore sequencing are explored in depth in the introduction to chapter 3. Briefly, nanopore library preparation follows fairly basic principles:

- End repair and A tailing of fragments using the

2: Methods

2.5: Genetics and genomics protocols

- 3 steps (2 for 1D) consisting of ligation of required adapter proteins, followed by wash steps with AMPure XP beads
- Priming of flow cell with running buffer and flow cell tether
- Mixing purified, adapted DNA library with library loading beads and running buffer
- Loading the flow cell and beginning the sequencing run

2: Methods

2.5: Genetics and genomics protocols

Stage	Primer name	Structure
Primary PCR	Forward-M13F	GTAAAACGACGGCCANNNN...
Primary PCR	Reverse-M13R	GGAAACAGCTATGACCATGNNNN...
Barcode PCR	M13F-Barcode	BBBB...GTAAAACGACGGCCA
Barcode PCR	M13R-Barcode	BBBB...GGAAACAGCTATGACCATG

Table 2.7: Sequences used for first stage of PCR barcoding

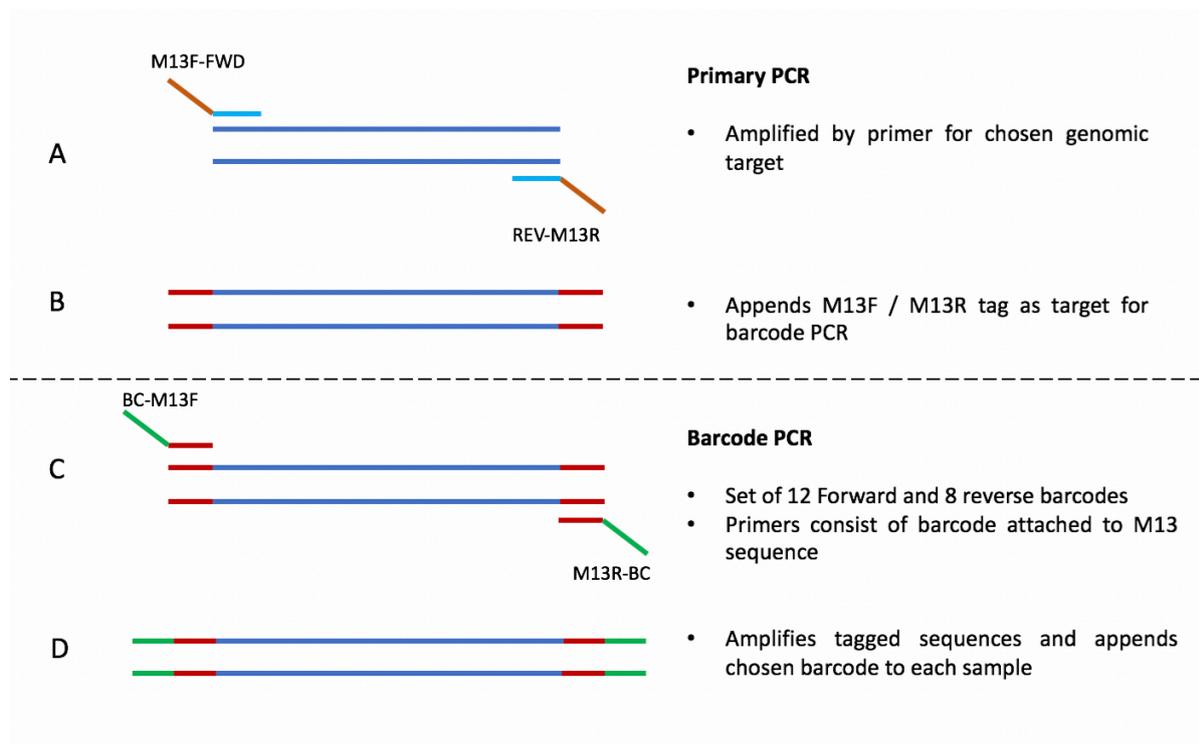


Figure 2.3: Two stage barcode PCR protocol – A&B show primary PCR performed to amplify target sequence and append M13F and M13R tags. C&D show second stage barcode PCR, where unique forward and reverse barcode sequences are appended to primers targeting the M13F/R tags appended in step A&B

2: Methods

2.5: Genetics and genomics protocols

Name	Sequence
M13F	GTAAAACGACGGCCA
M13R	GGAAACAGCTATGACCATG
A	AGAACGACTTCCATACTCGTGTGA
B	AACGAGTCTCTTGGGACCCATAGA
C	AGGTCTACCTCGCTAACACCACTG
D	CGTCAACTGACAGTGGTTCGTA
E	ACCCTCCAGGAAAGTACCTCTGAT
F	CCAAACCAACAACCTAGATAGGC
G	GTTCTCGTGCAGTGTCAAGAGAT
H	TTGCGTCTGTTACGAGAACTCAT
1	AAGAAAGTTGTCGGTGTCTTTGTG
2	TCGATTCCGTTTGTAGTCGTCTGT
3	GAGTCTTGTGTCCAGTTACCAGG
4	TTCGGATTCTATCGTGTTTCCCTA
5	CTTGTCAGGGTTTGTGTAACCTT
6	TTCTCGCAAAGGCAGAAAGTAGTC
7	GTGTTACCGTGGGAATGAATCCTT
8	TTCAGGGAACAAACCAAGTTACGT
9	AACTAGGCACAGCGAGTCTTGGTT
10	AAGCGTTGAAACCTTTGTCCTCTC
11	GTTTCATCTATCGGAGGGAATGGA
12	CAGGTAGAAAGAAGCAGAATCGGA

Table 2.8: Sequences used in two-stage barcoding of amplicons for nanopore sequencing

2: Methods

2.5: Genetics and genomics protocols

2.5.5.2 Whole transcriptome sequencing

Whole transcriptome sequencing (RNA-Seq) libraries were prepared from extracted samples by Joanne Morgan, of the core sequencing team at the Neuroscience and Mental Health Research Institute (NMHRI) of Cardiff University. Library preparation was performed from prepared samples using the KAPA mRNA HyperPrep kit, following the usual principals of RNA-Seq library preparation:

- mRNA is captured from samples using magnetic beads, then eluted using elution buffer
- mRNA is fragmented
- Double stranded cDNA is synthesised from the mRNA, then A-tailed for adapter ligation
- Cleanup is performed to remove un-ligated adapters and other reagents
- Adapter ligated library is amplified by PCR, followed by a further bead cleaning stage to remove primers and reagents
- Library QC is performed by qPCR using the KAPA Library Quantification Kit

2: Methods

2.5: Genetics and genomics protocols

2.5.5.3 Next-Seq Library Preparation for INDUCE Seq

The library preparation for the INDUCE-Seq experiments described in *Chapter 6* differs from the usual Next-Seq illumina library preparation protocol in several important ways. The materials required for this are included in the master reagents table (*table 2.1*).

This protocol was created by Felix Dobbs, Research Fellow at the Wales Gene Park. All remaining stages of the library preparation were performed in collaboration with Felix Dobbs. All buffer and wash recipes can be found in *table 2.9*.

One confluent well of a 12WP (300,000 – 500,000 cells) was treated with accutase for 10 minutes to dissociate the colonies into single cells. The single cell suspension was aspirated and centrifuged at 400g for 5 minutes. The accutase supernatant was aspirated and the cells re-suspended in 75µL PBS.

25µL aliquots of cell suspension were seeded into three wells of the 96 well plate, pre-coated with 1:50 PDL and incubated for 2 hours before seeding. The suspension was incubated at room temperature for 10 minutes, to settle and bind to the PDL coating.

175µl of 4% PFA was then added to the wells, in order to fixate the cells. This was incubated for a further 10 minutes at room temperature.

At the date of library preparation, stored cells are washed, washed and incubated with lysis buffer (LB1) for 1 hour at 4°C, then washed and incubated with LB2 for a further hour at 37°C.

In situ DSB blunting was performed by incubating cells with cut smart buffer at room temperature 3 times, followed by incubation in blunting buffer for 1 hour at room temperature.

Cells were washed 3 times with 1x cutsmart buffer at room temperature and incubated for 2 minutes. This was repeated three times, before incubating the cells with the A-tailing mix for 30 minutes at 37°C.

The cut-smart wash was repeated a further three times, followed by incubation with the DSB ligation mixture overnight at 16°C, to ligate the P5 adapters to the blunted and A-tailed DSBs. As only unrepaired DNA ends were exposed in the culture for ligation, only unrepaired DSBs will have the P5 adapters ligated.

Once adapters are ligated, 10 sequential washes with HSW at room temperature are required to remove the un-ligated adapters, each with a 2-minute incubation at room

2: Methods

2.5: Genetics and genomics protocols

temperature, followed by a 2-minute incubation with PBS, then a 2-minute incubation with NFW to remove the remaining salt buffer.

The cells are incubated with DNA extraction buffer for 5 minutes at room temperature. The plate is then incubated in a thermo-shaker at 65°C for 1 hour, shaking at 800rpm.

Raw DNA sample concentration is measured at this stage, using the qubit following the procedure in general methods. This allows for estimation of the total number of cells extracted, based on the average cellular DNA concentration of 6.6pg.

The remainder of the library preparation follows similar principals to a standard Illumina library preparation. The extracted DNA is sonicated, size selected, the fragment ends are repaired and A-tailed, and P7 adapters are ligated. The end result is that the majority of sequences have P7 adapters ligated, but only those strands which were available for P5 ligation in-situ, that is unrepaired DSBs, have both sets of adapters ligated.

DNA is sonicated in the Bioruptor plus, in high mode, with 30 seconds on alternating with 30 seconds off, for 20 cycles in 100µL. To check fragment size, 2µL of each sample is tested using Tapestation High Sensitivity screentape. If fragmentation is adequate, then size selection is carried out on the remaining sample, using SPRI beads at a 1:1 ratio, to remove DNA strands <150bp in length. P5 dimers and over-fractionated DNA is removed at this stage. A further Tapestation check step is used to determine if the size selection has been successful.

P7 adapters are ligated using a mix consisting of 30µL End Prep Reaction Mixture, 15µL NEB Ultra II Ligation Master Mix, 0.5µL NEBNext ligation Enhancer and 1.25µL 1µM Truncated Illumina P7 adapter per sample. This is incubated for 15 minutes at 20°C. A further SPRI bead purification step is carried out, and the final product checked on the Tapestation for size distribution. 1µL is taken forward for qPCR library quantification.

2: Methods

2.5: Genetics and genomics protocols

Table 2.9.1

Lysis buffer 1 (LB1)					
Reagent	Final conc	Stock conc	Volume (μ l)		
Tris-HCL	10mM	1M	100	200	500
NaCl	10mM	1M	100	200	500
EDTA	1mM	0.5M	20	40	100
Triton X-100	0.2%	10%	200	400	1000
H2O	pH 8 at 4°C		9580	19160	47900
Total			10mL	20mL	50mL

Table 2.9.2

Lysis buffer 2 (LB2)					
Reagent	Final conc	Stock conc	Volume (μ l)		
Tris-HCL	10mM	1M	100	200	500
NaCl	150mM	1M	1500	3000	7500
EDTA	1mM	0.5M	20	40	100
SDS	0.3%	10%	300	600	1500
H2O	pH 8 at 25°C		8080	16160	40400
Total			10mL	20mL	50mL

Table 2.9.3

High-salt wash buffer (HSW)					
Reagent	Final conc	Stock conc	Volume (μ l)		
Tris-HCL	10mM	1M	100	200	500
NaCl	2M	5M	4000	8000	20000
EDTA	2mM	0.5M	40	80	200
Triton X-100	0.5%	10%	500	1000	2500
H2O	pH 8 at 25°C		5360	10720	26800
Total			10mL	20mL	50mL

Table 2.9.4

DNA extraction buffer					
Reagent	Final conc	Stock conc	Volume (μ l)		
Tris-HCL	10mM	1M	100	200	500
NaCl	100mM	1M	1000	2000	5000
EDTA	50mM	0.5M	1000	2000	5000
SDS	1.0%	10%	1000	2000	5000
H2O	pH 8 at 25°C		6900	13800	34500
Total			10mL	20mL	50mL

Table 2.9 –p1/2 - buffers required for in-situ library preparation for INDUCE-Seq

2: Methods

2.5: Genetics and genomics protocols

Table 2.9.5

Blunting Mix				
Reagent	Vol (1x)	Vol (5x) (μ L)	Vol (10x) (μ L)	Vol (100x) (μ L)
NFW	37.5	188.75	377.5	3775
Blunting buffer 10x	5	25	50	500
BSA 20mg/mL	0.25	1.25	2.5	25
dNTPs 1mM	5	25	50	500
Blunting Enzyme mix	2	10	20	200
Total	50	200	500	5000

Table 2.9.6

A-tailing mix				
Reagent	Vol (1x)	Vol (5x) (μ L)	Vol (10x) (μ L)	Vol (100x) (μ L)
NFW	42	210	420	4200
NEBNext dA-Tailing Buffer (10x)	5	25	50	500
Klenow Fragment 4'-5' exo	3	15	30	300
Total	50	250	500	5000

Table 2.9.7

DSB ligation Mix				
Reagent	Vol (1x)	Vol (5x) (μ L)	Vol (10x) (μ L)	Vol (100x) (μ L)
Nuclease-free water	37.5	187.5	375	3750
T4 ligase buffer 10x	5	25	50	500
UltraPure BSA 50 mg/mL	1	5	10	100
P5 adapter 10 μ M	2	10	20	200
ATP 10mM	4	20	40	400
T4 ligase 2,000,000 units/mL	0.5	2.5	5	50
Total	50	250	500	5000

Table 2.9.8

DNA Extraction Buffer				
Component	Vol (1x) (μ l)	Vol (5x) (μ l)	Vol (10x) (μ l)	Vol (100x) (μ l)
DNA extraction buffer	95	475	950	9500
Proteinase K 20mg/mL	5	25	50	500
Total	100	500	1000	10000

Table 2.9 – p2/2 - buffers required for in-situ library preparation for INDUCE-Seq

2: Methods

2.5: Genetics and genomics protocols

The remainder of the library preparation follows similar principals to standard Illumina library preparation. The extracted DNA is sonicated, size selected, the fragment ends are repaired and A-tailed, and P7 adapters are ligated. The end result is that the majority of sequences have P7 adapters ligated, but only those strands which were available for P5 ligation in-situ, that is unrepaired DSBs, have both sets of adapters ligated.

DNA is sonicated in the Bioruptor plus, in high mode, with 30 seconds on alternating with 30 seconds off, for 20 cycles in 100 μ L. To check fragment size, 2 μ L of each sample is tested using Tapestation High Sensitivity screentape. If fragmentation is adequate, then size selection is carried out on the remaining sample, using SPRI beads at a 1:1 ratio, to remove DNA strands <150bp in length. P5 dimers and over-fractionated DNA is removed at this stage. A further Tapestation check step is used to determine if the size selection has been successful.

P7 adapters are ligated using a mix consisting of 30 μ L End Prep Reaction Mixture, 15 μ L NEB Ultra II Ligation Master Mix, 0.5 μ L NEBNext ligation Enhancer and 1.25 μ L 1 μ M Truncated Illumina P7 adapter per sample. This is incubated for 15 minutes at 20°C. A further SPRI bead purification step is carried out, and the final product checked on the Tapestation for size distribution. 1 μ L is taken forward for qPCR library quantification.

The pooled adapter-ligated library (40 μ L) is mixed with 40 μ L of 0.2N NaOH and 40 μ L of 20nM Tris-HCL. 1179 μ L HT1 is added to bring the total volume to 1299 μ L. A PhiX control is prepared for parallel sequencing.

2.6 Software Packages and programming languages

2.6.1 General software packages

A multitude of packages were used to analyse the data arising from the procedures detailed above, and throughout this thesis. These included generic packages such as Microsoft Excel, used for calculations and basic graphs, and proprietary packages linked to various procedures. These will be referenced where appropriate – in addition, *tables 2.10 & 2.11* contain full lists, with references where they are available.

2.6.2 Custom Scripts

This project utilised a number of custom scripts for bespoke data analysis and bioinformatic processing. Descriptive details of the important custom scripts can be found in the relevant results chapters; and the scripts themselves can be found in full in Appendix I. The majority of these scripts are written in either bash or python 3. A full list of the packages used within python and bash, along with brief descriptions, can be found in *table 2.12*.

Software	Description
Atom [196]	Desktop code editor
Python 3.6 [197]	Coding language
Python 3.7 [197]	Update released July 2018
R 3.6.1 [198]	Coding language frequently used for data analysis
R Studio [198]	Running environment for data analyses in R
Bash	Unix shell and command language used in Mac OS terminal
Microsoft Excel	Spreadsheet program used for basic data processing and analysis
ImageJ [199]	Open source image processing software designed by the National Institutes for Health (NIH)
LiCor Image Studio	Proprietary software used for analysis of Western Blot images acquired on Odyssey CLx
Terminal	Unix based command line for Mac OSX
Thermo cloud	Online cloud service for transferring iBright images and analysis files to desktop

Table 2.10 – Software tools used in the analysis of data

Package	Full name and description
Burrows Wheeler Aligner (bwa) [200]	Latest iteration of widely used aligner, capable of handling short and long reads
Minimap2 [201]	New aligner specifically designed for long reads
Nglmr [202]	CoNvex Gap-cost alignMents for Long Reads – badly named aligner for mapping long reads to large reference genomes
STAR aligner [203]	Specialised splice-junction aware aligner for management of RNA-Seq data
Picard [204]	Suite of tools for processing RNA-Seq data
Samtools [205]	Package for handling and manipulating alignment files (.sam, .bam)
Bedtools [206]	Package for performing arithmetic on various sequence files, such as alignment files
igv [207]	Package for viewing genome alignments
FastQC [208]	Package for performing quality control on sequencing libraries
Canu [209]	Error correction and assembly package for nanopore data
Subread [210]	Used for making raw counts of aligned RNA-Seq data
bamtools [211]	Used for QC of aligned RNA-Seq reads
Bam-readcount [212]	Generates arithmetic tables of counts from bam files
Guppy	Base-caller for nanopore data
Albacore	Base-caller for nanopore data, deprecated 2018
Porechop [213]	De-multiplexing tool for nanopore data allowing for use of custom barcodes

Table 2.11 – bioinformatic packages used in analysis of genomic data

Package	Description
PYTHON 3 [197]	
Os [197]	Allows operating system dependant functionality (eg, 'open')
Sys [197]	Allows access to variables supplied from command line
Matplotlib [214]	Library of graphical plotting tools
Plotly [215]	Library of graphical plotting tools
Numpy [216]	Allows use of mathematical matrixes in the form of arrays
Scipy [216]	Contains many tools for scientific and statistical calculations
Seaborn [214]	Adjunct to matplotlib, providing graphical enhancements to output
Pandas	Allows for importing and manipulation of large amounts of data into tables – aka Data Frames
re	Allows use of regular expressions in python syntax
random	Random number generator for statistical modelling
glob	Allows searching for path-names via python
operator	Efficient coding for applying Python's intrinsic functionality to various data functions.
Biopython[249]	Package containing numerous modules used for manipulation of sequence data
R	
Bioconductor	Suite of tools for analysis of high throughput genomic data
BioMaRt	R wrapper to access data from Ensembl's biomaRt database [217, 218]
Pheatmap	Makes Pretty heatmaps
Ggplot2 [219]	Library of graphical plotting tools
DESeq2 [220]	Suite of tools for analysis of changing expression in RNA-Seq data (DE = Differently Expressed)

Table 2.12 – Python and R packages used in the creation of bespoke scripts for data analysis

2.7 Bioinformatics

Bioinformatics is a relatively new scientific discipline, existing at the intersection of molecular biology, biostatistics and computer science. With the advent of whole genome sequencing (WGS), new tools and techniques have been developed to aid in the rapid processing of the ‘big data’ generated by such experiments. The study of bioinformatics and the development of these tools is an entire field of research. For the purposes of this section, bioinformatics refers to the ‘pipeline’ of processing that genomic data undergoes to provide usable data for scientific enquiry.

2.7.1 Bioinformatics for nanopore sequencing

Nanopore sequencing (*chapters 3-5*), RNA-seq (*chapter 4*) and INDUCE-seq (*chapter 6*) have some pipeline idiosyncrasies, which are addressed in detail below.

A complete list of bioinformatic packages can be found in *tables 2.10 and 2.11*.

2.7.1.1 Basecalling

MinKNOW is Oxford Nanopore’s proprietary desktop sequencing platform. It provides real-time sequencing metrics including: available pore count, pores currently sequencing, pores available for sequencing and recovering pores and damaged / unavailable pores. It gives a real-time assessment of the read count, total base count, histograms of the read lengths and a heat map of the quality scores of the bases sequenced. It is important to note that the quality pertains to the electrical signal and does not directly correspond to QC as encoded in FASTQ files.

Basecalling is the process by which electrical signal generated by the sensors of the sequencing platform is transmuted into recognisable genetic code. Genetic code sequences are stored as either FASTQ files – which also contain base quality scores for each base, or .fasta files, which do not.

Whereas Illumina platforms output FASTQ files, ONP outputs a third file type, known as FAST5, containing data in the form of current fluctuations. These .fast5 files were base-called using Oxford Nanopore’s proprietary command-line tools, which are explored further in Chapter 3.

2: Methods

2.7: Bioinformatics

2.7.1.2 Quality Control (QC) and filtering

At this stage, reads with low overall quality scores were filtered from the analysis. The quality score was provided in the FASTQ file as an ASCII character, corresponding to a logarithmic scale of confidence in the base-call (1 error in 10, 100, 1000 etc). Where further QC was necessary, FastQC is used.

2.7.1.3 Sequence correction

The accuracy of nanopore data can be improved through two methods: read correction and consensus polishing. The aim in this experiment was *not* to develop a consensus sequence, but to display the full range of heterogeneity in DNA repair in each sample. Therefore, consensus polishing tools such as nanopolish are *not* considered.

Canu is a polishing pipeline tool which utilises three separate steps: polishing, trimming and assembly. The steps of the pipeline can be accessed together. The correction protocol of canu is considered in this thesis, as a tool for improvement of nanopore read accuracy.

Because of the model canu uses, comparing each k-mer with all other k-mers in the sequencing file to find matched sequences and repair them by identifying the most frequently occurring results, running an entire 100,000 depth sequence output as one process is not computationally feasible. After some experimentation, it was determined that breaking the sequence output into more digestible chunks of 4,000 reads would allow for sequence correction, but without overwhelming the CPU. These outputted chunks are then collated before further processing.

2: Methods

2.7: Bioinformatics

2.7.1.4 De-multiplexing

In order to obtain data in the most cost-effective possible manner, it is not uncommon for multiple experiments to be multiplexed on the same sequencing run. If these experiments contain sequences overlapping the same genomic target, then they must be marked for tracking during the library preparation stage. This usually involves the ligation of barcode DNA tags to the sequences to be analysed.

De-multiplexing can take place at the basecalling stage, or later. For our Illumina experiments, de-multiplexing was performed at the basecalling stage.

Our de-multiplexing approach for nanopore was more idiosyncratic. Porechop [213] is an open source tool for barcode decomplexing of nanopore data. As the barcode sequences are likely to have 1-3 errors per 24 bases, the algorithm uses an accuracy threshold cut-off to place reads into the most likely bins. If the sequence at the start of a read does not reach the percentage accuracy threshold with any of the specified barcodes, it is not binned.

The system is tuneable, so that the minimum accuracy required can be set, along with the minimal differential required between the top two scores for a read. For example, if a read has 91% accuracy for the sequence of barcode 9 and 90% accuracy for the sequence of barcode 4, and the differential is set at 2% ($91 - 90 = <2$), then that sequence will not be binned.

The system is designed to look for sequences with a single barcode appended to one end, or the same barcode appended to both ends. I wrote parent bash-script (porechop_decomplex), which ran the system with only forward barcodes specified, then iterated over each of the 12 forward barcoded bins, but with the reverse barcode specified, so that the 96 combinations could be identified with one command.

2: Methods

2.7: Bioinformatics

2.7.1.5 Sequence Alignment

Where the target organism has a published reference genome available, sequences are aligned to it. This allows for differences between the reference genome and sequenced genome to be identified, and for the sequences to be displayed in a useful graphical manner. The Burrows-Wheeler-Aligner (bwa) has been widely adapted for this task, although other aligners also exist. Aligners usually output into a Sequence Alignment/Map - .sam file.

Multiple aligners are available for matching sequences with a reference genome. With the increasing popularity of nanopore sequencing, several of these programs have specialised algorithms for handling nanopore sequencing data which can be specified when the scripts are launched. In chapter 3 a comparison is made between the burrows-wheeler-aligner (BWA) [200], minimap2 [201], and coNvex Gap-cost alignMents for Long Reads (ngmlr)[202].

In the published literature, two measures are used to describe accuracy – read accuracy and consensus accuracy. Read accuracy compares the sequence identity of reads to a reference genome, whereas consensus accuracy measures the consensus sequence identity. For the purposes of this project we are concerned with variation between reads as a measure of DNA damage repair and so consensus sequences are not sought or used. Therefore, unless explicitly stated otherwise in the text, ‘accuracy’ can be taken to mean ‘read accuracy’.

2.7.1.6 Alignment, compression and sorting

.sam files can be unwieldy and often require compression and sorting for downstream analysis. Samtools is an ubiquitously used package that allows for compression into Binary Alignment/Map (.bam) files as well as sorting and indexing.

2.7.1.7 Viewing of aligned sequences

A genome viewer is a platform that allows visualisation of aligned sequences. Although not always the easiest format in which to analyse data, viewers allow for examination and representation of regions identified through other means. Unless otherwise stated, the integrative genomics viewer (igv) platform was used for all sequence imaging in this project.

2: Methods

2.7: Bioinformatics

2.7.1.8 Processing alignments for further analysis

Samtools stats will be used to generate a list of insertions and deletions of each size. Bam-readcount will be used to generate csv formatted data regarding each reference position in the aligned sequences, including length and depth of deletions, length and depth of insertions and sequence contents of insertions, that can be further analysed to generate statistical comparisons

2.7.1.9 Variant calling and further analyses

Variant calling error-prone sequence reads, such as those arising from nanopore sequencing, is an area of much interest. Chapter 3 explores this problem in considerable depth. The process of variant calling for this data presented a more complex problem. Currently available variant callers for nanopore sequence data require data from processed sequences, in which the semi-stochastic sequencing errors generated by the nanopore are “polished” from an assembled consensus sequence [209]. Unfortunately, the polishing and assembly tools available are written with more standard uses of the nanopore in mind (whole genome assembly) and did not handle the short amplicon sequences as well.

A new variant identifier, named CRISPR_nanoscreen.py, was written in python3 to look for indels in our data. Data is first extracted in tab delimited .txt file format using the bam_readcount package. Column 2 holds the reference position, column 4 the read depth and columns 12 and upwards hold data in a standardised format regarding indels found at this position, including: length, sequence and number of reads containing this deletion.

CRISPR_nanoscreen.py takes the total number of deletions or insertions and stores them in an array that can be viewed in a graphical format. It does so in a way that is agnostic of sequence content for insertions, allowing it to overcome the error rate in sequencing of the insertions.

As the error rate in the nanopore sequence can offset the start/end sites of indels, it also allows similar indels that start near to each-other to be counted together.

For example, if there was a 5bp deletion at reference position 3, with a read depth of ten, then the array would be created; [0, 0, 10, 10, 10, 10, 0, 0]. If a 2bp deletion with a read depth of 3 was then seen at position 4, the array would become [0, 0, 10, 13, 13, 10, 10, 0, 0].

2: Methods

2.7: Bioinformatics

The peak number from the array is called, and the ratio to the average readcount over the area calculated to give a read depth for the indel. So, for our example array, the peak would be 13 – if the average read depth from the bam-readcount file is 75, then this would be called as a deletion present in up to 16% of reads. The wells in which deletions are called can then be visualised in igv to confirm the output from CRISPR_nanoscreen and identify mutations for further investigation. The full code for CRISPR_nanoscreen can be found in the APPENDIX I.

2.7.1.10 *Sliding window error analysis*

As described above the errors in nanopore sequencing are not entirely stochastic; some genomic features are more likely to generate erroneous calls than others; particularly long strings of repeating mono or di-nucleotide sequences.

In order to assess the error profile and assess each pipeline's relative success at handling error-prone sequences, an automated 'sliding window' is used to determine the sequence content in terms of monomer, dimer, trimer, tetramer sequences (*figure 2.4*).

The number of errors occurring in the context of those sequence features is then counted and compiled. By comparing the occurrence of artefacts at each polymer sequence with the occurrence of that sequence within the amplicon, a relative error likelihood can be established. For example, if the trimer AGG represents 3% of the reference sequence but contains 15% of the recorded errors from the sequencing run, then the error ratio is 5:1.

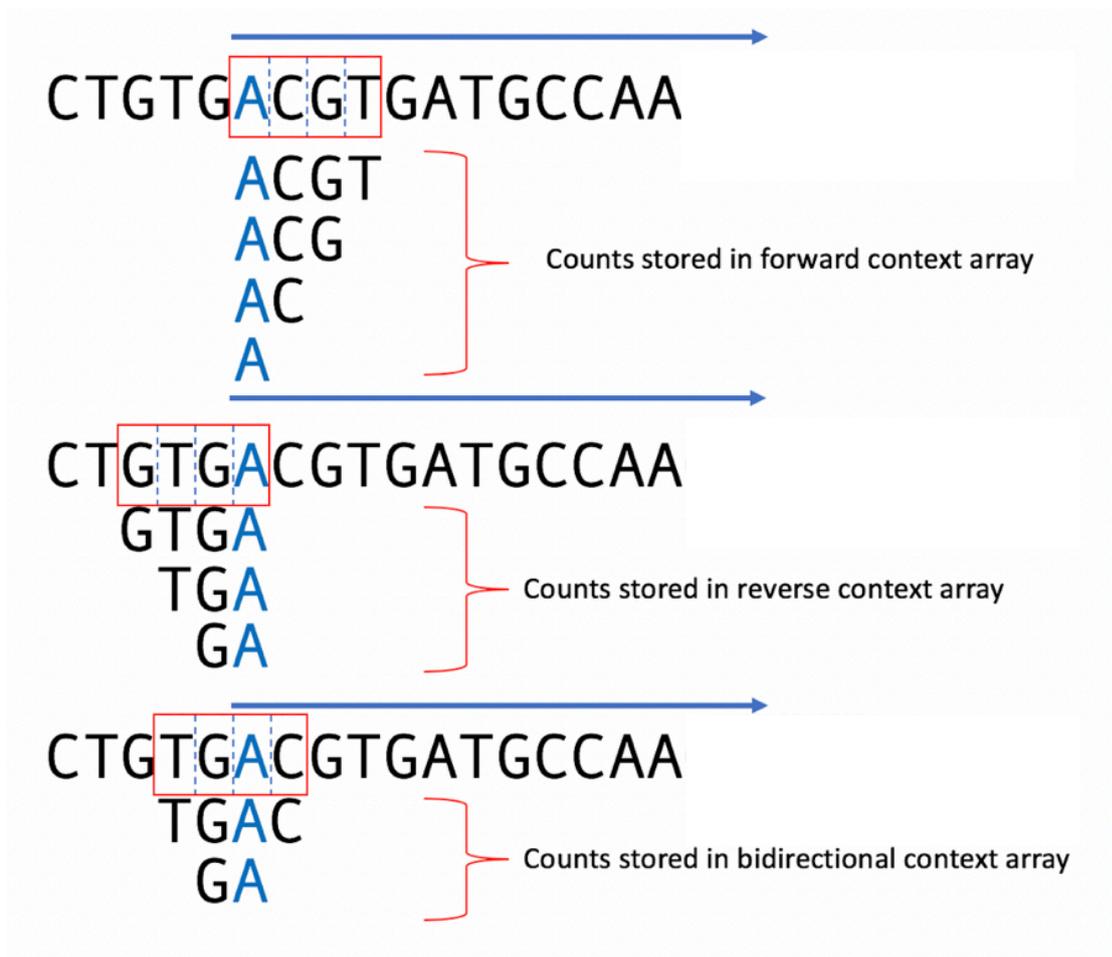


Figure 2.4 – Demonstration of sliding window error analysis.: The first base of each called indel (blue) is used as the starting point to count monomer, dimer, trimer and tetramer in forward, reverse and bidirectional data-structures. The blue arrow indicates the direction of the window as it moves throughout the reference sequence collating a complete dictionary of all forward, reverse or bidirectional sequence tetramers for use in downstream data analysis.

2.7.2 Transcriptomics

Data from RNA-seq was output to the ROCKS storage system of the NMHRI's neurocluster computing node. The steps in this section were performed using the Raven computing cluster provided by Super-Computing Wales (SCW) at Cardiff University. Subsequent steps requiring far less processor horsepower were performed on laptop.

The bioinformatic pipeline for RNA-Seq is similar to the pipeline for second generation WGS, as described in the general methods section. There are however some important differences, which I will highlight below.

2.7.2.1 Trimming and Quality Control

Trimming of RNA-Seq data to remove adapters and bases with low phred scores is performed with trimmomatic[221]. Quality control is performed with FastQC[208].

2.7.2.2 Reference indexing and alignment

As with any genomic assay, reads generated must be mapped to an indexed reference sequence. For RNA-Seq, the genome was indexed with Spliced Transcripts Alignment to Reference (STAR)-aligner, in order to allow its utilisation at the alignment stage[203].

STAR is an aligner specifically written to handle reads from RNA-Seq. It was developed to deal with the vast amount of transcriptome data held by the ENCODE consortium. Compared to WGS, where except for regions of copy number variation or structural defect, reads are expected to cover contiguous stretches of DNA, RNA-seq reads can cross 1 or more exon boundaries[222]. Therefore, an algorithm is needed that can map the first part of a read to one location, then search through exon start sites in order to determine the splicing arrangement of the second half (*figure 2.5*).

The default behaviour is to report one alignment for each read (the alignment with the maximum score). In some instances, if two identical scores are available, the read will be discarded. There is an option, in these cases to select a read at random to be kept, however for this experiment the default behaviour was deemed sufficient.

Reads were mapped to reference genome build GRCh38. This build of the genome was used as it is expected to contain the most up-to date gene list in the associated gtf annotation file. Both the genome and annotation file were downloaded from Ensembl.

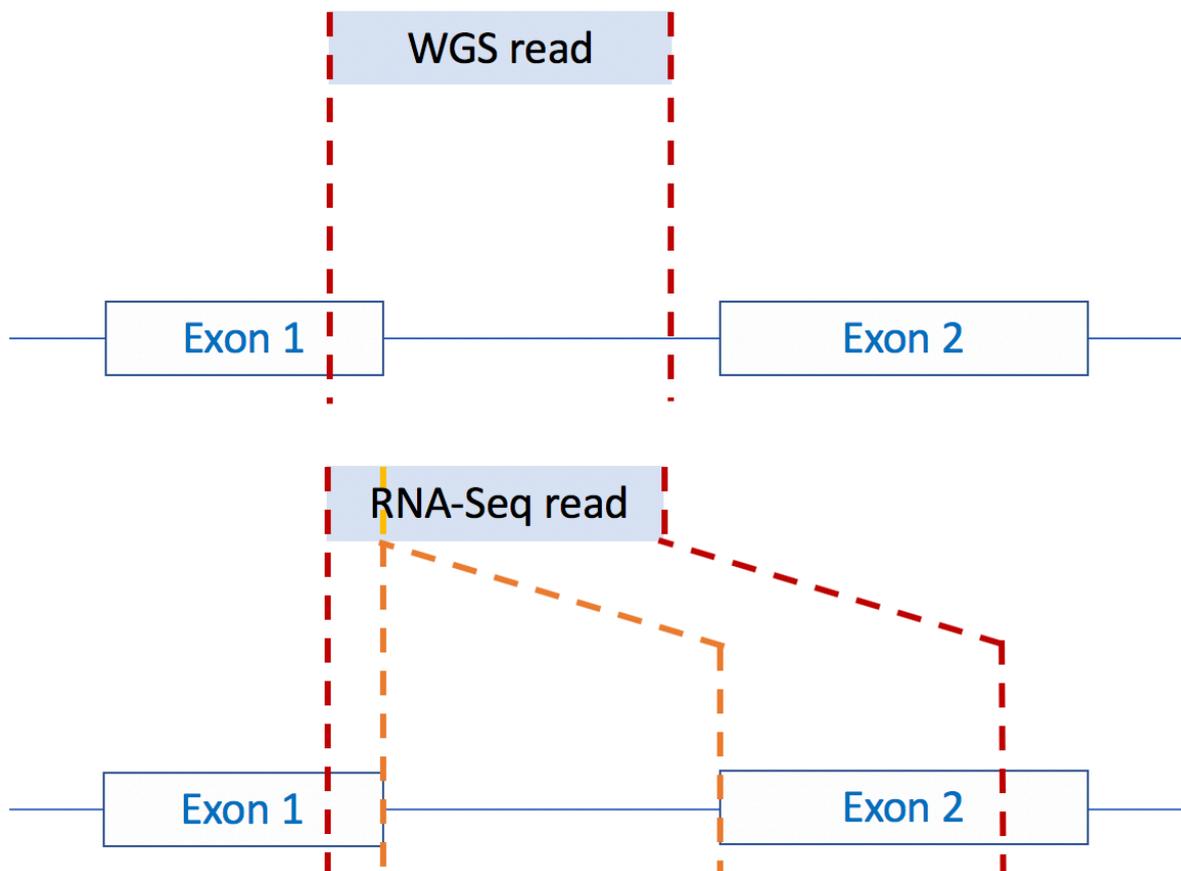


Figure 2.5 – The impact of RNA splicing on reads generated by RNA Seq. Splice aware RNA-seq reads contain only exonic sequences and overlap exonic boundaries. STAR align is a splice aware aligner that can map a read to various isoforms.

2: Methods

2.7: Bioinformatics

2.7.2.3 Cleaning BAMfiles and marking duplicates

The picard package is used to mark and remove duplicate reads – that is reads with identical mapping co-ordinates. High levels of duplication are seen when RNA extraction is of poor quality so that there is greater amplification of the smaller number of extracted molecules[204].

QC can then be performed on the marked bamfiles using bamtools, generating a percentage of duplication[211].

2.7.2.4 Generating raw read counts per gene

The final stage of this process is generating raw counts per gene for each experimental run. This is done using the featureCounts tool[223] from the subreads package. FeatureCounts compares the alignment co-ordinates of the bam file to the gtf file, indexed by STAR in tandem with the genome file. The gtf file contains the genomic co-ordinates for all coding sequences in the genome, linked to the Ensembl gene ID to which they relate.

The output is 18 text files with raw transcript counts for each gene which can be taken forward for further analysis using specialist packages writing in the R programming language.

2.7.2.5 RNA-Seq Data Processing in R

The 18 datasets are compiled into a single table in R and are labelled according to the day of differentiation, their cell-type (WT or CHD2 heterozygote) and their bio-rep number (1-3). The feature counts are collated into a table and the ensembl gene reference numbers are substituted for HGNC gene symbols, using the biomaRt package in R.

A further normalisation step, using the gene length to control for transcript abundance generates a final normalised read-count called Fragments Per Kilobase of transcript per Million mapped reads (FPKM). The FPKM for each gene can then be compared to generate further analyses[224].

2.7.3 Data analysis for RNA-Seq

DESeq2[220] is an R package with extensive tools for processing and visualising RNA-Seq data. One study comparing a plethora of differing RNA-Seq analysis pipelines for up to 48 biological replicates discovered a wide amount of variation in performance. Our chosen bioinformatic toolkit, named Differential Expression Seq2 (DESeq2), was one of the best performing across all replicate counts[225].

Once the data is correctly imported into R and formatted correctly, DESeq2 is called to provide normalised counts for each gene and comparisons between sample groups[224].

2.7.3.1 Sample-to-sample distance measurement

This process performs hierarchical clustering analysis in order to measure the distance between the samples. These are plotted in a heatmap matrix. Related to this, a principal component plot is made to visualise the overall variance between the samples.

This allows us to determine if any samples are outliers from their groups which should be excluded from downstream analysis. It will also allow us to determine if technical replicates for each sample can be collapsed / summed together for further analysis in DESeq2.

2.7.3.2 Gene counts and MA-plots

An MA plot visualises the mean normalised read count against the log-2 fold change (LFC) for each gene. By highlighting the genes in which a statistically significant change has occurred, they provide a useful first visualisation of the spread of the data.

Using the same dataset it is possible to extract and compare the read counts for any individual genes of interest.

2: Methods

2.7: Bioinformatics

2.7.3.3 Differential expression lists and gene ontology

Using DESeq2, the differential expression list can be filtered by p value, to provide a list of genes with statistically significant differential expressions between any two samples of groups. An initial filtering that groups genes with a $p=0.1$ into those with $LFC < 0$ and $LFC > 0$ is the first stage recommended in many guidelines for RNA-Seq, however these thresholds can be changed in order to curate lists of more reasonable lengths for analysis.

This list of genes will be analysed using the online gene ontology annotation toolkit GOrilla [226] to try and determine which functions and pathways are impacted by the change in differential expression between WT and CHD2 mutant cells at each stage of differentiation.

The output of ranked ontology terms from GOrilla is transferred to REVIGO [227] – a web-based suite of tools for visualisation of ontology and pathway terms. From here, datasets are exported for use in Cytoscape [228], providing interactive visualisation of ontology terms in a map format.

2.7.4 INDUCE-seq

The start co-ordinates of each read correspond to the breakpoint from which that read was captured. A bed file is generated from the start co-ordinates of each aligned read using bed intersect. A bedgraph is generated by sorting each bed file into bins of 1kb size length and visualised using the Integrated Genome Browser (IGB).

The co-ordinates for each protein coding gene were obtained from the ensembl browser using biomatr and formatted into a bed-file using Microsoft excel.

Bedtools intersect was used to count the overlaps with each protein coding transcript. The number of breaks per kb for each transcript was calculated and the distributions plotted at each time point in differentiation. Log2fold changes for each gene were calculated and stored for comparison with RNA-Seq data (*chapter 4*).

A bed file containing 1M randomly generated break-sites was generated for comparison with each experimental run. This will be referred to as random break file (RBF).

Bedfiles containing genomic regions of interest such as histone markers from the ENCODE database, transcription start sites (TSS) and protein coding genes or regulatory regions can then be downloaded for comparison with the INDUCE-seq data.

Ratios were calculated between the total number of reads in the RBF and each experimental run. Bedtools intersect was used to count the overlaps between the co-ordinates in each target file and each experimental file. These counts were normalised using the RBF ratios. The enrichment compared to the RBF was calculated between each experimental run and the normalised RBF.

2.8 Protein Analysis

2.8.1 Western blotting

Prior to extraction, cells are cultured on a 12WP. Each well of the plate was washed twice with 500µL PBS before scratching to suspend in 1mL PBS. The cells were pelleted by centrifugation at 900g for 5 minutes at 4°C. The supernatant was aspirated, then the pellet was re-suspended in 150µL extraction buffer (250µL 2x RIPA buffer, 50µl 10x protease inhibitor, 200µl dH₂O makes 500µL).

The suspension was kept on ice and vortexed for 15 seconds every 5 minutes, for 30 minutes. A further centrifuge step was performed – 14,000rpm for 5 minutes at 4°C. The sample was split between 2 Eppendorf tubes; 120µl is combined with 46µl LDS sample buffer and 18µl DTT to make a total volume of 184µL – this sample is heated at 70°C for 10 minutes. The second sample, to be used for protein quantification was stored without treatment. Samples are kept at -80°C until testing.

Protein quantifications were performed on the BMG labtech Clariostar microplate reader using the InesProt pre-set. Protein standards were prepared in triplicate using the ThermoFisher BSA protein assay standards and BioLine Protein Assay Reagent A & B, in a 96 well plate. Protein extractions were prepared with the same reagents in the same plate. After quantification, 10µg of protein for each sample were loaded onto the gel for running.

Proteins were run for 1-2hrs at 100-120V in the Bolt gradient gel, with Novex Bolt running buffer with NEB protein standard for comparison. Novex Bolt transfer buffer was used to transfer the proteins onto nitrocellulose paper, at 4°C, for 1 hour at 120V.

Antigen blocking was performed with 5% milk in TBST. Primary antibody incubation was conducted overnight at 4°C, with antibody solution in 5% milk TBST. Secondary antibody incubation was performed in a light-proof box at room temperature, with secondary antibodies again suspended in 5% milk solution.

Imaging of developed blots was performed on the LiCor Odyssey CLx and analysis performed using the LiCor Image Studio proprietary software.

2: Methods

2.8: Protein Analysis

Antibody target	Supplier	Ref	Species	Ratio IHC	Ratio WB	Band WB
CHD2	abcam	ab167377	Mouse	1:100	1:500	57 kDa
FLAG M2	SIGMA	F1804	Mouse	1:1000	1:1000	55 kDa
GAPDH	abcam	Ab9483	Rabbit	1:200	1:5000	37 kDa
GAPDH	abcam		Mouse		1:5000	

Table 2.13: Primary antibodies used for WB and IF

Antibody target	Ref	Species	Wavelength
Mouse	P/N 925-68020	Goat	680
Mouse	P/N 925-32280	Goat	800
Rabbit	P/N 925-68021	Goat	680
Rabbit	P/N 925-32211	Goat	800

Table 2.14: LiCor secondary antibodies used

2.8.2 Immunohistochemistry

Glass coverslips were prepared for cell growth by placing in a 24 well plate and coating with matrigel according to the above protocol (*section 2.3*). Cells were passaged onto coverslips and grown until stable colonies have formed – usually 24-48 hours for IPS cells.

Cells were fixed by incubating with 4% paraformaldehyde for 10 minutes at room temperature, then washed 3 times with cold PBS.

For intracellular proteins, cells were permeabilised by incubating with 0.1% Triton X-100 for 10 minutes at room temperature and washed 3 times with PBS. Blocking non-specific binding was conducted by incubating with 1% BSA in PBS + 1% tween 20 for 30 minutes.

Primary staining was conducted by diluting primary antibody (*table 2.13*) in 1% PBST and incubating cells at 4°C overnight. Cells were washed 3 times for 5 minutes in PBS and incubated with secondary antibody diluted in 1% BSA for 1 hour, whilst protected from light.

Secondary antibody (*table 2.14*) was decanted and a further 3 5-minute washes with PBS were carried out. Counter staining for nuclear protein was carried out using DAPI at 1µg/mL for 1 minute at room temperature. A further wash of PBS was conducted.

2.9 Gene editing with inducible Cas9

Section 2.4.4 describes the introduction of an inducible Cas9 construct into wild type cell lines. The following protocol describes the use of cells containing this construct at the *AAVS1* genomic safe harbour locus [229] for genome editing.

2.9.1 Choice and assembly of gRNA

GRNA was designed using Benchling's online lab book program. The program provides a procedurally generated list of all possible guides in a region of interest, by scanning the region for PAM sites and providing the sequences 20 base pairs upstream of each PAM.

Benchling scores each gRNA out of 100 for on-target and off-target effects, using an algorithm based upon several previously published studies analysing gRNA sequences [230-232].

For the establishment of cell lines for further study (*chapters 3 & 4*), it is a priority to reduce off-target mutations that could confound the results by introduction of an additional variable. For the investigation of DSB repair (*chapter 5*), gRNA with the greatest cutting fidelity possible were used and so on-target scores were prioritised.

The gRNA identified in the region of interest are ranked by off target effects, and the gRNA with the best on-target effect >50 that does not result in a penalty of >10 from the highest off-target score. There are exceptions and common sense is used to some degree – for example if a gRNA reduces the off-target effect from a maximum of 95 to 84 (penalty of 11) but increases the on-target score from 60 to 80, then this could be considered an acceptable switch. As the scores are only weakly predictive of actual effect, the lack of strict protocol can be justified.

TracrRNA and crRNA are both re-suspended in nuclease free duplex buffer (NFDB) to make 100µM solutions. To create duplexes for transfection, 30µL of each solution is mixed with 40µL NFDB. The mixture is then heated to 95°C for 5 minutes on a heat block and allowed to cool to room temperature on the bench. Assembled duplexes can be used immediately or stored at -20°C for future use.

2.9.2 Transfection protocol

Two versions of the protocol were trialled. The first involved separating cells into a single-cell suspension using accutase, plating and treating with RevitaCell and applying the transfection mixture to the cells in suspension before they settle. Although there is evidence that single-cell suspension increases the chance of efficient transfection [233] this unfortunately also placed a significant strain on the cells and despite the treatment with RevitaCell, often resulted in widespread cell death.

The modified protocol used GentleCell cell releasing reagent to passage the cells the night before transfection, with lipofection of gRNA performed at the start of the next day. This resulted in greater cell survival but still demonstrated high cell editing efficiency and so has been adopted as our default approach.

Cells to be transfected are plated into a 24-well plate, coated with matrigel, in 500 μ L of medium per well containing 2 μ g/mL doxycycline. On the day of transfection, the media volume is reduced to 450 μ L.

5 μ L of the gRNA duplexes are suspended in 45 μ L of optiMEM. In monoplex experiments, where only gRNA is transfected, the entire 5 μ L is used for the same gRNA. In multiplexed experiments[234], where 2 or 3 gRNA are transfected simultaneously, the 5 μ L volume is split between each guide (so, 2.5 μ L each for 2 gRNA and 1.66 μ L each for 3 gRNA). 3 μ L of Lipofectamine RNAiMAX is suspended in 47 μ L optiMEM in a second tube. The two suspensions were combined and left to incubate at room temperature for 10 minutes.

50 μ L of the solution was added dropwise to each of two wells of the 24 well plate, bringing the total volume of medium in the well to 500 μ L.

The cells were cultured with doxycycline at 37°C for a further 48 hours, enough time for all transfected gRNA to degrade. Once cells were confluent, they were passaged for clonal analysis.

2: Methods

2.9: Gene Editing with Inducible Cas9

2.9.2.1 Mature neurons

Primary neurons are challenging to transfect [235], however RNAiMAX has been demonstrated as capable of transfection of siRNAs, albeit in a rat model. Han et al [236] described an experiment in which various conditions are trialled, including; lipofection in DMEM, lipofection in Neurobasal media, starvation of cells prior to transfection and no-starvation prior to transfection. They conclude that transfection is more efficient in cells maintained in neurobasal media.

As B27 supplement used in maintenance of neuronal cultures contains various lipid particles that could interfere with the activity of the lipofectamine complex, on the day of lipofection the media was changed from the differentiation media described in general methods to un-supplemented Neurobasal media. The lipofection was carried out, and after a period of 4 hours, the media was supplemented with B27 and N2 at 1:50. In order to ensure the best possible uptake of gRNA in these difficult-to-transfect cells, the procedure was repeated the following day.

2.9.3 Growth of clonal cell cultures for selection of mutant clones

Prior to passage of lipofected cells, two 6cm dishes were prepared. The dishes were coated with matrigel, washed with PBS and plated with 6mL media containing 1:500 ROCKi. The media was warmed in the plate in an incubator set at 37°C for 30 minutes prior to plating cells.

When cells were confluent, they were pre-treated with ROCKi for 30 minutes, then collected using accutase to form a single-cell suspension. After a 10-minute incubation with accutase, culture media was added to stop the reaction. The resulting suspension was centrifuged at 400rpm for 5 minutes to pellet the cells. The supernatant was aspirated, and the cells re-suspended in 1mL culture media.

One plate was seeded at a concentration of 1:1000 (6µL of cell suspension) and the other at a concentration of 1:500 (12µL of cell suspension). The plates were gently swirled to spread the cells across the matrix.

Plates were incubated with regular media change until colonies had begun to arise from single cells. After 1 day, the medium was changed to remove ROCKi, thereafter medium was changed every 48 hours. Once colonies have begun to emerge, the medium volume was increased to 12mL.

Cells were picked under direct visualisation using video microscope. A 96 well plate was prepared with matrigel basement membrane – each well was filled with 60µL culture medium and the plate was warmed to 37°C. Colonies were scraped from the 6cm dish using a 200µL pipette with P200 tip, set to aspirate 20µL of medium. Once the colony had been detached it was aspirated and moved into a well of the 96 well plate. The aim was to pick 96 colonies for downstream analysis.

2.9.4 DNA Extraction and Sequencing Library Preparation

Clonal colonies in the 96 well plate were grown until confluent. A passage was performed, splitting each colony 1:2 – half the cells went into a new 96 well plate for storage, the other half were placed in 60 μ L DNA Quick extract. DNA was extracted, the region of interest was amplified and barcoded using PCR and the libraries prepared for sequencing using the protocols described above, and in the general methods sections.

2.9.4.1 Clonal growth of cultures

Prior to passage of lipofected cells, two 6cm dishes were prepared. The dishes were coated with matrigel, washed with PBS and plated with 6mL media containing 1:500 ROCKi. The media was warmed in the plate in an incubator set at 37°C for 30 minutes prior to plating cells.

When cells were confluent, they were pre-treated with ROCKi for 30 minutes, then collected using accutase to form a single-cell suspension. After a 10-minute incubation with accutase, culture media was added to stop the reaction. The resulting suspension was centrifuged at 400rpm for 5 minutes to pellet the cells. The supernatant was aspirated, and the cells re-suspended in 1mL culture media.

One plate was seeded at a concentration of 1:1000 (6 μ L of cell suspension) and the other at a concentration of 1:500 (12 μ L of cell suspension). The plates were gently swirled to spread the cells across the matrix.

Plates were incubated with regular media change until colonies had begun to arise from single cells. After 1 day, the medium was changed to remove ROCKi, thereafter medium was changed every 48 hours. Once colonies have begun to emerge, the medium volume was increased to 12mL.

Cells were picked under direct visualisation using video microscope. A 96 well plate was prepared with matrigel basement membrane – each well was filled with 60 μ L culture medium and the plate was warmed to 37°C. Colonies were scraped from the 6cm dish using a 200 μ L pipette with P200 tip, set to aspirate 20 μ L of medium. Once the colony had been detached it was aspirated and moved into a well of the 96 well plate. The aim was to pick 96 colonies for downstream analysis.

3: SETUP AND TESTING OF AN INDUCIBLE CAS9 GENE EDITING AND NANOPORE SCREENING PIPELINE FOR HIGH THROUGHPUT GENE EDITING EXPERIMENTS

3.1 Introduction

In order to explore the aims set out for this thesis (stated in *section 1.8*), two things are necessary. First; a method of reliably creating double strand breaks in targeted regions of the genome in a multitude of cell types, to allow testing at different stages of differentiation. Second, a method for high-depth analysis of the outcomes of DSB repair. In this chapter, I will describe the design and testing of both these systems.

3.1.1 Creating targeted mutations

Several approaches were considered for the investigation of DNA damage repair. These break down into untargeted analyses and targeted analyses. The choice of untargeted analysis is described in detail in chapter 6. In this chapter, we will consider the options available for creating targeted mutations.

The creation of targeted mutations requires either the use of restriction enzymes that create breaks at specific genomic sites, or the use of tuneable endonucleases which create breaks at a chosen point in the genome.

Restriction enzymes can be introduced into mammalian cells via plasmid electroporation[237]. Most enzymes will recognise a motif that repeats throughout the genome. Although this has been used previously for the investigation of DSB repair – the widespread nature of the DNA damage has significant drawbacks. First, as mutations become more abundant, the chance of a mutation causing a significant off-target effect, which could dysregulate the repair pathway being investigated, increases. Second, such widespread damage can cause cell death. The cells that die are likely to be those with the greatest mutational burden – although this can be controlled for, the analysis carried out is unlikely to be reflective of normal physiological function. Third, it is possible that widespread simultaneous damage overwhelms or otherwise changes the DNA damage response cascade, again making such experiments a poor reflection of physiology.

More promising is the introduction of targeted double strand breaks in a fashion that does not create widespread off-target damage, leading to cell survival and well-regulated repair in a manner that mirrors normal physiological function. In recent years, several targeted endonucleases have been identified and made commercially available for use in

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.1: Introduction

biomedical research. These include: zinc-finger (ZF) endonucleases, transcription activator-like nuclease (TALENs) and clustered regularly interspersed short palindromic repeats (CRISPR) with CRISPR associated protein 9 (Cas9).

The three techniques described function by creating targeted DSBs, and either utilising HR to introduce a new sequence - either a mutation, a reporter gene such as green fluorescent protein (GFP) or even a transgene to study, or on cNHEJ and A-EJ to create indels and knock genes out.

3.1.1.1 Zinc Finger Endonucleases and TALENs

TALENs and ZF are similar in terms of action and construction, and so can be considered together. Both endonucleases are guided to their targets by DNA-protein interactions.

ZF endonucleases are amino acid arrays, often found in eukaryotic cells, which recognise DNA triplets. Typically, a 30 amino acid module recognises 3-6 triplets – ZF arrays have been designed that can target all 64 possible triplet combinations[238].

These amino acid arrays are attached to a FokI domain – a non-specific endonuclease, which will cut anywhere in the genome when present as a dimer. The requirement for dimerization means that two ZF probes must be designed, targeting flanking sequences upstream and downstream of the cut-site[238].

Transcription activator-like effector nucleases (TALENs) are protein arrays derived from *xanthomonas* species of bacteria. Like ZFs, they contain a FokI restriction domain which must dimerise to create a DSB. Unlike ZFs, the amino acid arrays recognise individual nucleotides rather than triplets, making TALEN easier to design and tune to different sequences. In addition, the zinc finger motifs can interfere with neighbouring motifs, meaning that design often requires a significant amount of trial and error[239].

Each TALEN targets a 30-40bp length DNA sequence and although some degeneracy has been described in the TALEN code significant off-target mutations have not been described[239].

Editing efficiency is variable with some targets appearing particularly resistant to editing with TALENs. The specificity of the TALEN code that protects from off target mutagenesis is also a drawback; methylation of the target DNA sequence alters its physiochemistry enough to prevent the probe from binding[240].

3.1: Introduction

Another significant consideration is the technical challenge posed by design and construction of new TALEN nucleases, which requires the modular assembly of a novel protein for each target site. Although kit-based systems are available to assist with this the process is still time consuming and relatively expensive.

There is also the issue of protein delivery: lipofection, electroporation or viral vector transduction of TALEN plasmids are required for each desired gene edit. Along with the challenges of protein re-engineering this makes the system less efficient for higher-throughput applications.

3.1.1.2 CRISPR-Cas9

A more recent development offers significant advantages albeit with some slight disadvantages when compared to ZF and TALEN genome editing. Named for the unusual repeating sequences identified in some species of bacteria (the clustered regularly interspersed short palindromic repeats) and the associated proteins, the impact of the CRISPR Cas9 genome editing system on biomedical research is hard to overstate.

Since it was first demonstrated that CRISPR-Cas9 could edit the human genome in 2013[241], there has been an explosion in biological and biomedical research using this approach[242]. It has been widely utilised both in basic science research, and as a potential tool for gene therapy in humans[243].

Understanding the function of Cas9 is useful and a summary of the timeline of its discovery can be helpful in understanding the role of each component. See also *figure 3.1*.

When first identified in 2002[244], the unusual feature of repeating sequences spread throughout the genomes of some species of archaea and bacteria were initially thought to be a unique mobile element. Initially these CRISPR sequences were hypothesised to be mobile elements, or somehow involved in the regulation of transcription.

The CRISPR sequences are arrayed as loci consisting of non-contiguous repeats separated by 'spacer' sequences. The spacer sequences were noted to have sequence homology with foreign elements including bacteriophages[245]. This led to the hypothesis, subsequently proven, that CRISPR sequences conferred adaptive immunity against phages. Seemingly random motifs from the genomes of invading phages are incorporated *de novo* into the spacer sequences in cultures that develop phage resistance – a greater number of incorporated sequences is associated with a greater degree of resistance[246]. Eventually, it

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.1: Introduction

was understood that the CRISPR sequences were part of an adaptive phage-resistance system in certain prokaryotic bacteria[247].

The functions of different Cas proteins were found to range from preparation of the CRISPR-RNA (crRNA) used to guide the immune response by Cas6 [248], to the all-important Cas9 nuclease, which was demonstrated as the only protein required for guided cleavage to occur[249].

The final essential factor for Cas9 mediated target cleavage to be understood was the trans-activating-CRISPR RNA (tracrRNA) [250]. TracrRNA contains a 25nt stretch with almost perfect complementarity to the repeat regions in the CRISPR array, which were subsequently termed CRISPR-RNA (crRNA)[250]. The two RNAs can therefore form a duplex molecule, which guides the Cas9 cleavage.

Full characterisation of the Cas9 mediated cleavage *in vitro*, and demonstration that cleavage was guided by Watson-Crick pairing, and could therefore be programmed by changing the sequence of the crRNA followed [251], leading rapidly to the development of tools capable of intracellular editing.

The use of Watson-Crick base pairing makes CRISPR considerably easier to design than ZF and TALEN. The protein component remains identical, as does all of the RNA complex, save for 20bp of the crRNA [252]. Indeed, multiple services now offer design tools for choosing a guide RNA and several biotech firms provide bespoke RNA nucleotides with various optimisations and alterations to facilitate CRISPR based editing. These RNA oligonucleotides are cheap and easy to assemble.

The range of potential targets in a genome is only limited by the presence of the sequence immediately upstream in the format NGG. This Protospacer Adjacent Motif (PAM) site is present throughout the human genome and must be present for successful action of the nuclease (*figure 3.1*). Based on the frequency of 'GG' dinucleotides in the human genome there are an estimated 161,284,793 PAM sites available[253], allowing for the possibility of editing the vast majority of protein-coding loci in the human genome. Even the absence of this motif may not be the hindrance it was once believed to be. Cleavage, albeit at around 20% efficiency, has also been demonstrated at non-canonical PAM sites of the format NRG (where R = A/G)[254, 255].

Secondly, CRISPR is also known to be highly efficient. Rates of indel formation up to 88% have been reported. It is worth noting however, that this is an upper estimate using well

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.1: Introduction

optimised gRNA in one cell type. The rate of indel formation for each experiment is more unpredictable and is known to be dependent on the method of transfection (see below) [256], the cell type [256], the genomic target chosen [229] and the specific gRNA sequence [229].

This highly efficient cutting, and propensity to bind to less specific sites, is also the main potential disadvantage of CRISPR-Cas9, particularly when considered as a potential mode for gene therapy in human beings. Compared to TALEN and ZF nucleases, Cas9 has more potential to create off-target edits [242].

The 'seed region' of the crRNA – that is the region 10-12bp adjacent to the PAM site, is critical for recognition. The remainder of the sequence can increase target cutting rate, but also risks increasing the propensity for off-target mutations. In total, up to 5bp worth of mismatches can be tolerated by the system, allowing for protein binding and cutting to occur not just at identical sites elsewhere in the genome, but at sites harbouring mismatches [232]. There is some evidence that truncated gRNA of 17-18bp length can reduce off-target effects without having a notable effect on the efficiency at the target effect [257].

Systematic investigations of on target and off target efficiencies have been conducted resulting in algorithms that can be used to estimate the likely on target [232] and off-target [230] effects of CRISPR with a particular gRNA. Subsequent investigation of the scores generated by these algorithms demonstrates that they possess only weak predictive value and that our understanding of the factors influencing the success and precision of genome editing is still incomplete.

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.1: Introduction

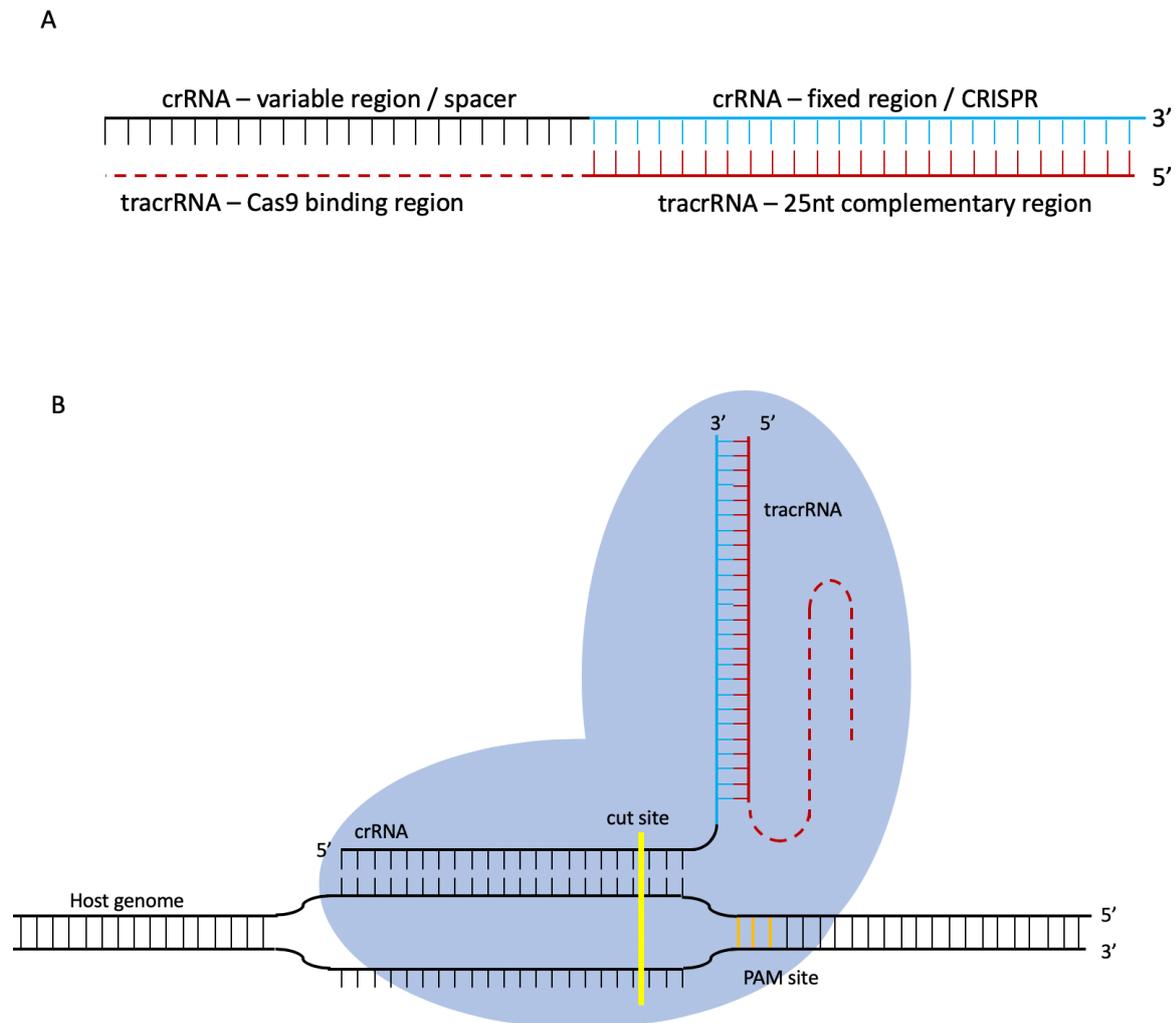


Figure 3.1: schematic representation of CRISPR Cas9 genome editing. A: crRNA and tracrRNA duplex to create gRNA, B: gRNA combines with Cas9 protein to guide gene editing – cut occurs 3nt upstream of the PAM motif site (orange) – the PAM site does not comprise part of the crRNA, but must be of nucleotide composition NGG and occur immediately upstream of the gRNA sequence in the target genome

One recent study by Chakrabati et al [258] demonstrated editing with differing efficiencies in 1248 of 1491 (83%) target sites, meaning that 17% of gRNA was unable to promote gene editing at the target locus. This study also described a significant impact of the nucleotides chosen 2, 3, 4 and 5 in determining the type of mutation and the accuracy of editing. It had previously been suggested that editing was reproducible at genomic targets across cell types and replicates [151] however, the Chakrabati study refined this observation further by demonstrating that *some* gRNA are prone to ‘precise’ editing with reproducible

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.1: Introduction

mutations across replicates, whereas others are prone to imprecise editing where greater variation in the size and characteristics of indels is seen at repair sites.

Although these results have not yet been integrated into the scoring systems used in CRISPR design they represent a further step forward in understanding the dynamics of this system. Another high throughput study published in 2019 provides confirmatory evidence of the importance of target sequence in determining the type of mutation that occurs during NHEJ repair[259].

It has also demonstrated that chromatin states and nucleosome positioning (and therefore by inference, cell type) influence the relative abundance of mutations[260]. This is relatively intuitive; genes that are actively transcribed and maintained in a euchromatic state are more accessible to the Cas9 protein than genes in heterochromatic regions where transcription is suppressed.

There are several options available for the delivery of the gRNA-Cas9 protein complex. These include: nucleofection (by electroporation or lipofection) of the protein complex, nucleofection of a plasmid coding the protein complex [261], nucleofection of an mRNA to be translated by the cell's own ribosomes [256], or viral transfection of the plasmid[262].

As only 20bp of gRNA changes for every Cas9 probe, and Cas9 is known to complex with tracrRNA when both are produced intracellularly, it is possible to set up a cell line that produces Cas9 protein and simply transfect the relevant gRNA[263].

It is also possible to create a cell line that produces Cas9 protein, gRNA, or both in response to administration of a chemical agent such as doxycycline[264, 265]. The advantages of this approach include a higher level of reproducibility – once the cells have been demonstrated to produce Cas9, no optimisation of external factors is necessary. More gentle transfection techniques can also be utilised – getting a short RNA duplex into cells is far easier than transfecting a plasmid of several kb length, or an even larger protein-gRNA complex.

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.1: Introduction

3.1.1.3 Choice of gene editing approach

For the majority of this project editing efficiency is a far greater concern than editing accuracy. Particularly in assessing the characteristics of indels formed (*Chapter 6*) an idealised system would have 100% editing efficiency. For this reason, Cas9 will be utilised throughout this project. In addition, the ease and low cost of designing new guides is extremely attractive from a practical point of view.

The TALEN system will be used in the set-up of our cell line. The AAVS1 genomic safe harbour is an adenovirus integration site. Previous phenotypic studies of cells with integration here demonstrate no detectable change in cell behaviour and so it has become a popular site for the integration of trans-genes into the human genome [266]. TALENs targeting AAVS1 already exist and have been well described. These TALENs will be used to set up a knock-in of doxycycline inducible Cas9 for downstream work[229].

3.1: Introduction

3.1.2 High Throughput Analysis of DSB Repair Outcomes

3.1.2.1 Screening overview

The simplest approach to screening for indels using PCR is simply to run the products on a high percentage agarose gel and look for differences in sample migration. The downside is that smaller indels may be harder to identify using this approach; a 1bp, 2bp or 4bp insertion may be sufficient to frameshift the region and knock out the target gene but is difficult to detect on a gel electrophoresis. It is also difficult to accurately quantify the amount of each sized product by this method using standard PCR.

Mismatch endonuclease digestion can be thought of as a modified version of this approach. Mutant and wild type DNA are hybridised by melting and stepped cooling in a thermal cycler (10°C of cooling every 5 minutes is a common protocol). This leads to a combination of WT-WT hybrids, mutant-mutant hybrids, and mutant-WT hybrids. The WT-mutant hybrids contain mismatches, which are cleaved by the enzyme[267] and can be detected on gel electrophoresis.

Of the widely used commercial kits available, T7E1 can detect mutations present at 5%-95% of a sample, whereas Surveyor can detect mutations at 10-90%. In both cases this means that mutations present at a lower percentage may be lost. [267].

In the case of the surveyor assay[268] a maximum of 60% of hetero-duplexes are cleaved making proper quantification of mutations difficult. As with PCR electrophoresis quantification of the fraction of the strands mutated can be difficult[267].

Both endonuclease assays and gel electrophoresis can struggle to identify smaller lesions. A 1bp insertion or deletion will still have the desired effect in an experiment that aims to create a frameshifted knockout of a gene and these can be missed by both techniques. There is also concern that larger mutations, although easily identifiable by PCR, would be missed by the surveyor endonuclease as they fail to form duplexes with the wild type controls.

Both of these techniques are also hampered by the lack of sequence information, providing only a roughly qualitative description of the DNA sequence. Mutations thought to be genuine therefore need to be described in detail. This can lead to the potentially frustrating scenario of promising looking indels being revealed as in-frame and of dubious impact on the protein being produced.

3.1: Introduction

3.1.2.2 Sanger Sequencing

For decades, Sanger sequencing – named for its inventor Professor Fred Sanger, has been the ‘gold standard’ of sequence level genetic analysis. The original version of the technique depended on random fractionating of nucleic acids and incorporating radiolabelled nucleotides onto strand ends[269]. This first protein coding sequence described using this technique was the coat protein of bacteriophage MS2, in 1972.

A variation developed in 1977 became the foundation of the first commercially available sequencing machines. This involved the incorporation of chemical analogues of nucleotides – dideoxynucleotides (ddNTs) – that terminate the polymerase reaction. Fluorescent labelling or radiolabelling could be used to detect the last base in each reaction[270].

‘First generation’ sequencing platforms used increasing refinements of this technique and were adopted worldwide for both research use and clinical interrogation of DNA sequences. It is understood to be highly accurate but is relatively low throughput; Sanger sequencing is capable of generating a sequence of length roughly 1kb. Investigation of tumours, or cell cultures containing multiple genomes results in difficult-to-interpret outputs.

In terms of practicality for this project, Sanger sequencing offers a high degree of accuracy, but not the depth required to interrogate cell cultures for low level mutations. Quantification of mutations present at low percentages may be possible using the types of ‘shotgun sequencers’ that used multiplexed sanger sequencing for the early stages of the human genome project[271], however these have been superseded by next generation sequencers [272] (also referred to as second generation sequences (SGS)) and more recently third generation sequencing sequencers (TGS). A note on terminology – the term ‘High Throughput Genetic Sequencing’ (HTGS) can refer to either SGS or TGS. Where possible, sequencing techniques will be referred to specifically to avoid the potential for confusion.

3.1: Introduction

3.1.2.3 Next generation sequencing platforms

Since the completion of the human genome project, the fall in the cost of sequencing has been precipitous allowing for its widespread adoption as a tool in research and clinical medicine [273, 274]. As such, the use of NGS has been previously demonstrated in the analysis of multiplexed CRISPR samples[275].

NGS describes a series of technologies capable of performing sequence analysis of millions of DNA strands in a single experimental run. A whole human genome can be sequenced to a depth suitable for clinical diagnoses to be made on the majority of coding sequences in less than 24 hours. This confluence of new laboratory techniques and the continuous improvement in micro-processing power has transformed biomedical research and clinical genetics.

The first NGS technique to be widely adopted was sequencing-by-synthesis, utilising a luminescent reporter to indicate when a nucleotide is incorporated into a strand. As nucleotides are incorporated ATP sulfurylase converts pyrophosphate into ATP, which acts as a substrate for luciferase[276, 277].

Clonal sequences are bound to beads and amplified in an emulsion PCR (emPCR). The beads are washed over a flow cell composed of wells large enough to hold one bead each. As each bead theoretically contains many copies of the same sequence, the signal of each nucleotide incorporation is amplified to a readable level.

This technique can generate read lengths of 400-500bp. One drawback is that homopolymers are difficult to quantify. The amount of light released is proportional to the number of nucleotides incorporated, but this can be challenging to quantify accurately, and so particularly longer strings of homo-polymers can be miscalled. (Roche 454 and 454 GS FLX)

A more recently developed variation of this approach involves the release and measurement of hydrogen ions and fluctuations in pH. This allows for improved sample turn-around time but does not overcome pyrosequencing's difficulties with homopolymer runs[278].

The most widely adopted NGS technique is the 'bridge amplification' technique, originally developed by Solexa but acquired and promoted widely by Illumina. Instead of binding the clonal DNA strands to beads, they are first ligated to adapters which can complex with complementary adapters on the surface of a flow cell. A solid phase PCR produces clusters of identical strands for analysis. The flow cell is washed with a fluorescent reversible-

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.1: Introduction

terminator dNTP. Once the fluorophores are released, a further dNTP can be bound. Again, the clustering of identical strands amplifies the signal from each nucleotide in sequence.

These machines produce shorter sequences (initially as short as 35bp), but allow for paired end sequencing, where the same strand is read from both ends sequentially providing a greater degree of accuracy.

Since the first platforms became available, subsequent designs such as the HiSeq, MiSeq and NextSeq have provided advances in read length, sequencing depth and cost.

3.1.2.4 Third generation single molecule sequencing platforms

Recently, single molecule real-time cell (SMRT-cell) sequencing and nanopore sequencing have emerged. They use different techniques to achieve the same aim: long read sequencing of single DNA molecules.

Nanopore sequencing relies on the translocation of single-stranded DNA through a protein nanopore, embedded in a membrane. Each in-pore nucleotide will create a slightly different amount of electrical resistance, causing a change in the current generated across the pore. This fluctuating current is then read and converted into sequence information[279]. It is thought that up to 5 in-pore nucleotides affect the current level at any one time.

Library preparation involves the ligation of motor proteins to the sequence to be analysed, which can either be derived from any source, including whole genomes or PCR amplicons.

The earliest nanopore flow cells provided utilised the R7 pore protein and had low accuracies – around 70%; arguably of little use to anyone not researching ways to improve nanopore flow cell accuracy[280]. The newer R9.4 flow cells have improved accuracy – up to 90% read identity on alignment[281], although this is still relatively low compared to SGS platforms and certainly not of the fidelity required to pass benchmarking for clinical use.

2D chemistry also utilised the R9.4 pores. Library preparation involved ligation of a hairpin adapter to the DNA molecules, in order to provide a simulacrum of paired end sequencing and improve read accuracy. Read accuracy of up to 97% was reported[282], but with significantly shorter reads than 1D sequencing.

Subsequently, 1D² chemistry utilising a new R9.5 pore, achieves similar improvements in accuracy, but without joining the strands together. A 1D² adapter ligation library preparation involves an extra ligation step to attach the second strand to the membrane

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.1: Introduction

adjacent to the protein pore and hence feed the paired strand through sequentially. 1D² sequencing is reported by Oxford Nanopore (ONP) to provide read accuracy of up to 97% for up to 60% of strands.

At the time of writing, the R9.5 flow cells have been supplanted by R10 flow cells, which use 1D chemistry, but read the sequence passing through the pore twice. Data from ONP suggests that reliable accuracies of up to 94% can be achieved with this platform. It was not available at the time the experiments detailed here were conducted and is included in this introductory section for the sake of completeness.

ONP is the only company with commercially available nanopore sequencing platforms. The most commonly used of these, the minION, fits in the palm of the hand and can be run using a standard desktop PC or laptop. It also has a unique advantage: the experiments can be stopped once enough data is acquired, and the flow cells washed and stored for further use.

As well as relatively low accuracy, the ONP flow cells generate less data than SGS platforms – with a maximum of 3-5Gb per flow cell. They have been widely used for sequencing smaller genomes in a research setting, however whole genome analysis of a human genome is challenging. One recent publication detailed an effort requiring 53 flow cells in order to provide 30x coverage[283].

The other commercially available technology, SMRT cell available from Pacific Biosciences (PacBio), tethers a strand of DNA to the bottom of a well known as a Zero Mode Waveguide (ZMW). Each ZMW is illuminated from below – the wavelength of the light is too wide to pass through a well. The diameter of the ZMW is small enough that the light beam passed through the bottom cannot reach the top of the pore, allowing it to act as a miniscule microscope. Fluorophore linked nucleotides are washed through the cell – as they are incorporated into the strand, the signal from a single fluorophore is read[284].

The PacBio platform shares many of the advantages and disadvantages of the nanopore: long reads, but lower throughput and lower accuracy. It's large footprint and relative expense means that it was not further considered for use.

A summary of the sequencing platforms considered can be found in *table 3.1*.

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.1: Introduction

Sequencing Platform	Amplification technique	Sequencing technique	Read length	Reads per run	Advantages	Disadvantages
454	Bead emPCR	SBS with Pyrosequencing	400-500bp		Long reads compared to illumina	Difficulty resolving homo polymers Largely deprecated by IonTorrent
IonTorrent	Bead emPCR	SBS with semiconductor sequencing	400-500bp		Long reads compared to illumina Rapid run time	Difficulty resolving homo polymers
Illumina HiSeq	Bridge amplification	SBS with fluorescent reporting	150bp paired end	5 billion		
Illumina MiSeq	Bridge amplification	SBS with fluorescent reporting	300bp paired end	25 million	Most accurate sequencing platform currently available Small footprint Short run time Most accurate sequencing technique	
Illumina NextSeq	Bridge amplification	SBS with fluorescent reporting	150bp paired end	400 million	Short run time Most accurate sequencing technique	

Table 3.1 (page 1/2) – Comparison of high throughput sequencing platforms with relative advantages and disadvantages considered when designing experiments

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.1: Introduction

Sequencing Platform	Amplification technique	Sequencing technique	Read length	Reads per run	Advantages	Disadvantages
Oxford Nanopore MinION R9.4	None PCR	Nanopore single molecule sequencing 1D chemistry	Up to 100,000kb	Variable depending on read length	Longer reads allowing for detection of larger indels Possibility of identifying SV Can be analysed in real time Re-usability of flow cells	Lower accuracy Difficulty resolving homo-polymers
Oxford Nanopore MinION R9.5	None PCR	Nanopore single molecule sequencing 1D ² chemistry	Up to 100,000kb	Variable depending on read length	Improved accuracy compared to R9.4	More difficult library preparation Increased accuracy on 10-20% of reads only
PacBIO		SMRT cell single molecule sequencing	Up to 100,000kb		Longer reads allowing for detection of larger indels Improved homo-polymer resolution	Cost Unsuitable for amplicon sequencing

Table 3.1 (page 2/2) – Comparison of high throughput sequencing platforms with relative advantages and disadvantages considered when designing experiments

3.1.3 Summary and justification for use of nanopore sequencing

There were two main considerations in the choice of nanopore long-read sequencing as the appropriate platform for this project.

First, the character of the expected mutations. Aberrant end-joining results in the formation of indels, defined as insertion or deletion of genetic material between 1bp-10kb in length. Short read sequencing characteristically struggles to identify such mutations accurately, indeed it is one of the challenges presented diagnosis using WES or WGS for diagnostic purposes [285].

It was felt therefore that long read sequencing would provide us with useful data that short read sequencing could not. The trade-off; lower accuracy data that makes point mutations more difficult to detect, was felt not to be of significant detriment for the assessment of larger indels.

The second factor is one of expense. Although nanopore sequencing is currently prohibitively expensive for WGS of organisms, like humans, with gigabase-scale genomes [283], the ability to stop a run once enough data has been acquired allows costs for ultra-deep amplicon sequencing to be kept relatively low, by allowing us to run smaller scale experiments.

The most cost-effective approach for finding targeted mutations present in a sub-clonal population is amplicon sequencing. A region of 1000bp can be sequenced at much greater depth for a much lower cost than a whole genome. This approach only works if it is known where the mutations might occur, either from an endonuclease treatment, a known recurrent break site or a targeted cut made with a guided gene editing tool.

The relatively low accuracy of nanopore sequencing and the fact that this represents a previously untested use of the technology were causes for concern and so much of this chapter deals with the testing and optimisation of the nanopore pipeline in order to determine if it is a suitable technique for investigation of DNA repair outcome.

3.1.4 Aims

The remainder of the chapter describes the methods used in fulfilling this design and provides evidence for their efficacy using both test datasets and experimentally generated data.

- 1 Set up a cell line with inducible CRISPR-Cas9 expression capable of creating reproducible genome edits at targeted sites in hIPSCs
- 2 Set up and test a sequencing and bioinformatic pipeline capable of high depth sequencing for investigation and characterisation of sub-clonal mutations, using the ONP MinION platform and a cell line created for a previous project harbouring a known mutation in GRIN2A
- 3 Demonstrate efficacy of cell line and sequencing pipeline in generation of new mutants
- 4 Demonstrate optimisation of various bioinformatic pipelines in order to generate the most useful possible data for use elsewhere in the project

3.2 Methods

3.2.1 Cas9 cell line setup and cell line validation

A full list of PCR primer sequences used in this chapter can be found in *tables 2.6 & 2.7*. All mutations were created using the inducible Cas9 construct spliced into the AAVS1 locus, described in section 2.4.4 (*figure 2.1*).

Western blot and immunohistochemistry with primary antibodies targeting the FLAG tag on the induced Cas9 protein were used to validate Cas9 protein production.

A desktop test of Cas9 function was performed by transfecting gRNA against a known pluripotency factor, NANOG[286]. Knockdown of NANOG in pluripotent cultures leads to rapid loss of pluripotent morphology in effected cells [229]. This has been previously used to demonstrate effective integration and editing using the same iCas9 system. Lipofection was performed using a gRNA previously demonstrated to have high gene editing efficacy in iPSCs targeting the pluripotency factor NANOG[269]. The cells were imaged at 24-hour intervals to examine morphology, with loss of morphology being taken as confirmation of NANOG knockout.

3.2.2 Screening for known mutations

3.2.2.1 GRIN2A mutant library preparation

GRIN2A mutations were created by Dr William Plumbly in the same IBJ4 hiPSC cell line used throughout this project, whilst working at the NMHRI as part of his doctorate research project. The DNA extractions used in this experiment were a gift from him. The mutations were previously identified by PCR of the targeted area, followed by Sanger sequencing of the clones determined to harbour amplicons of different lengths.

The four samples chosen for this experiment were; *GRIN2A* wild type (*GRIN2A*-WT), an extract from a pure clone containing a heterozygous 16bp deletion (*GRIN2A*-DEL), an extract from a pure clone containing a heterozygous 41bp insertion (*GRIN2A*-INS), and an extract from a well containing a mixture of cells with the heterozygous 16bp deletion and cells with the heterozygous 41bp insertion (*GRIN2A*-MIX). All mutations arose from the same DSB site in the same CRISPR experiment. The samples were plated randomly on a 96 well plate and the location of each sample type recorded (*figure3.2*).

PCR primers to amplify a region of 832bp (914bp including M13 tags and barcode sequences) were optimised according to the protocol described in *section 2.5.3*.

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.2: Methods

DNA from these cell lines had been stored in TE at -20°C for a period of 2 years prior to use. 24 1µL samples from each of the four cell lines was plated, at random, in a 96 well plate.

At the time this work was conducted, SQK-LSK108 was the most advanced sequencing kit available for ONP sequencing. Library preparation was performed using the appropriate protocol. Sequencing was conducted on a 1D flow cell (MIN-106, R9.4), for 4 hours with the aim of generating 400,000 reads.

	1	2	3	4	5	6	7	8	9	10	11	12
A	Mix	Del	Ins		Del	Mix			Mix	Del	Mix	
B		Ins	Del	Ins		Ins		Del	Ins		Ins	Del
C	Ins	Mix		Del		Mix	Ins	Del	Del	Mix		
D	Del	Ins		Mix	Mix	Del		Ins		Ins	Del	Mix
E		Del	Ins		Del	Mix	Ins		Mix	Del	Ins	Ins
F	Ins	Mix	Del	Mix		Mix	Del	Mix		Ins		Del
G	Mix	Del	Mix		Ins	Del	Mix		Ins	Mix	Del	Ins
H	Del	Mix	Del	Mix	Ins	Ins	Del	Mix		Ins		Mix

Figure 3.2; diagram of 96WP containing DEL, INS, MIX and WT (blank) DNA samples

3: Setup and testing of an inducible Cas9 gene editing and nanopore sequencing pipeline

3.2: Methods

3.2.2.2 Bioinformatic Pipeline Optimisation

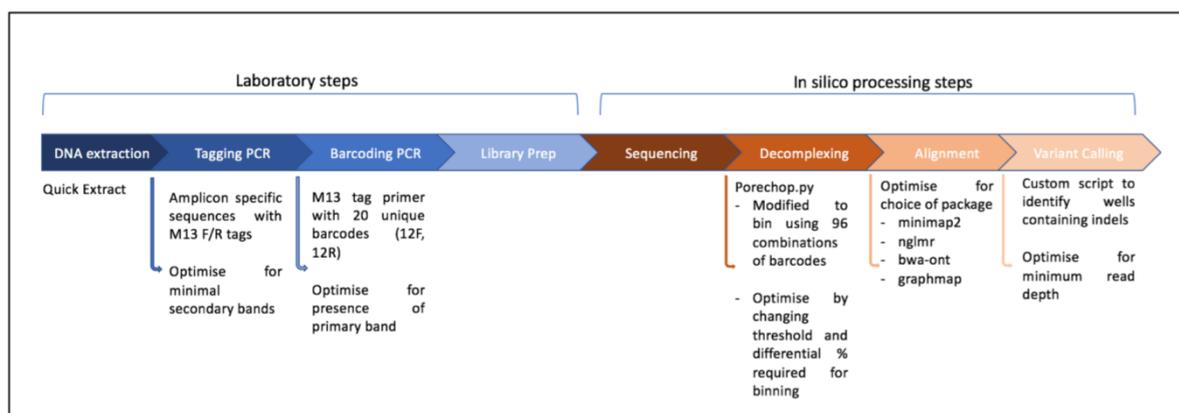


Figure 3.3: laboratory and bioinformatic pipeline for nanopore sequencing of Cas9 induced mutations and variables requiring optimisation

In order to determine the best bioinformatic pipeline for use throughout the remainder of the project, data is analysed using a number of different variables within the data analysis pipeline. There are two potentially tuneable variables in the nanopore pipeline: stringency settings of porechop and choice of aligner. (Figure 3.3)

The tuneable settings on porechop are accuracy and threshold. Accuracy refers to the required sequence homology between the barcode DNA sequenced and expected sequence of the barcode DNA for a read to be marked for binning. The threshold is the minimum difference for between the top two homology scores required for a single read to be binned. The default settings are an accuracy of 75% and a threshold of 3% (75_3) - so if a read has 82% homology to barcode A, and the next highest homology score is 65%, it will be binned in barcode A. If, however, the next highest homology score is 80%, then the 3% threshold score is not reached, and the read will not be binned. Accuracy_Threshold scores of 70_1, 75_3 and 80_5 were chosen to correspond to loose, moderate and stringent settings, based on preliminary data (not shown).

The choice of aligners is also important for nanopore sequencing. As discussed in the introduction to this chapter, nanopore data is different from SGS data and therefore differently designed models are required. As we also discussed, our data consists of neatly stacked amplicons rather than long range reads with overlapping segments and therefore there was doubt as to whether the aligners designed for nanopore data would be usable. Based on preliminary data, three – bwa, minimap2 and nglmr – were chosen for comparison. LAST and Graphmap[287] were rejected for comparison as the alignments generated were

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.2: Methods

very poor – possibly reflecting the difference between the data they were designed for and our amplicon sequence data.

I also wanted to determine if the average read depth in the sequence data from well had an impact on usability of the results for identification of new mutations. In order to determine this, a minimum alignment depth requirement of 1 to 25 was used.

3.2.2.3 Running pipeline comparisons

For each variable, insertion score and deletion score for each well was calculated and binned according to the cell line. So, if a well was known to contain GRIN2A-DEL, then the deletion score was added to the GRIN2A-DEL_deletion_score array, and the insertion score was added to the GRIN2A-DEL_insertion_score array. If the system is working effectively, then the GRIN2A-DEL wells should have a high average deletion score and a low average insertion score. See *table 3.2* for a full reference of the expected scores for each well-type.

A bash script was used to iterate over the FASTQ files and apply the 225 different variable combinations to the analyses pipeline and combine the results for analysis.

As the data was not normally distributed, the Mann-Whitney U test was used to determine whether there was a significant difference between the means in each category. Graphs to visualise the data were compiled using plotly.py

TYPE OF CALL	Name of call type	Expected result
Percentage of reads containing an insertion in a well containing a known insertion (GRIN2-INS well)	True_INS score	50%
Percentage of reads containing an insertion in a well containing a known deletion (GRIN2A-DEL well)	Miscall_INS score	0%
Percentage of reads containing an insertion in a well containing a mixture of known insertion, known deletion and wild type DNA (GRIN2A-MIX well)	Mix_INS score	~10%
Percentage of reads containing an insertion found in a well containing wild type DNA (GRIN2A-WT well)	False_INS score	0%
<hr/>		
Percentage of reads containing a deletion in a GRIN2A-DEL well	True_DEL score	50%
Percentage of reads containing a deletion in a GRIN2A-INS well	Miscall_DEL score	0%
Percentage of reads containing a deletion in a GRIN2A-MIX well	Mix_DEL score	~25-30%
Percentage of reads containing a deletion in a GRIN2A-WT well	False_DEL score	0%

Table 3.2: Binning of different calls made by the variant screener in optimisation of nanopore pipeline. The expected result column reflects the results of Sanger sequencing of these samples – the GRIN2A-INS and GRIN2A_DEL wells are assumed to be truly heterozygous, with no extracts from WT cells. Based on the sanger sequencing results, the GRIN2A_MIX well contained more deletions than insertions – this is reflected in the relative weighting of the expected scores. The expected scores are based on the results of previous sanger sequencing of these mutations (data not shown) from each sample used

3.2.3 Demonstrating the efficiency of the inducible Cas9 & nanopore sequencing pipeline in creating and detecting new mutations

Since the design and testing of this pipeline, it has been used to generate mutations in a variety of neuro-developmentally relevant genes, including: *CACNA1C*, *CHD2*, *CHD8*, *FADS2*, *GJAJ5*, *GJAJ8*, *SETD1A* and *TSC1*. These mutations were created by a number of colleagues working within the Harwood laboratory, including: Dr Bridget-Ann Kenny, Irene Serpa, Dr Shane Wainwright, Iker Martinez, and Gemma Wilkinson. The mutations were created following the protocol described in *section 2.9*.

For sequencing, in most cases the 1D² (R9.5) flow cell was used, along with the appropriate library preparation kit. For earlier experiments, this was the SQK-LSK308 kit, in later experiments this was replaced this with the SQK-LSK309 kit.

In all cases, mutations were created using a protocol described in *section 2.8* by and analysed using the pipeline described above in *section 2.6.2.1*, with slight modifications where new developments had deprecated the tools described above. Any additional analysis of sequence data will be described below, and was performed by the author.

3.2.4 Assessment of nanopore error profiles

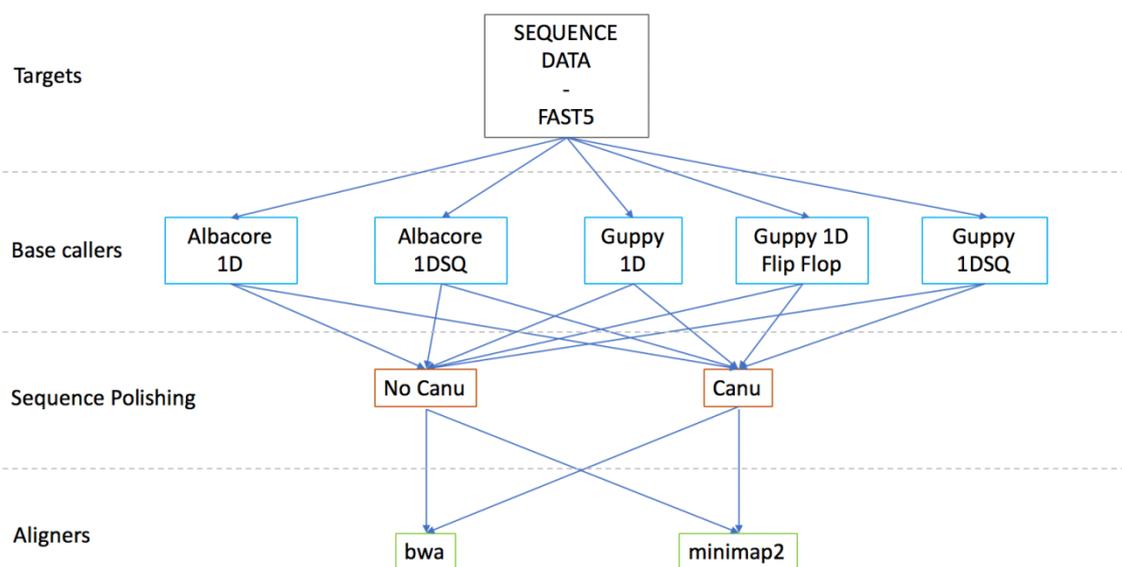


Figure 3.4: Various permutations of the bioinformatic pipeline compared for analysis of wild type sequences. The pipeline will be run with each of the five base callers, with / without sequence polishing and aligned with *minimap2* and *bwa*. This gives a total of twenty permutations, the results of which will be compared

The generation of such large volume sequence data over a diverse array of targets afforded the opportunity to perform an assessment of the various tools used for nanopore analysis.

After reviewing data pertaining to the most well-sequenced targets used throughout this thesis, three were selected for testing various permutations of the pipeline. Criteria were that the libraries sequenced contained the lowest number of PCR artefacts aligning to other regions of the genome. The targets selected were: the amplicons within *CHD2* used in generation of our knockouts in *chapter 4*, *FADS2* – used by Dr Bridget-Ann Kenny in her knockout experiment, and *NRXN1* – used as one of the targets to test the impact of *CHD2* mutations on DSB repair (see *chapter 6*).

The FAST5 data from these runs were processed with 20 permutations of the pipeline, with changes between base-callers, polishing (yes or no) and choice of aligner (*figure 3.4*, see *section 2.7* for description of the bioinformatic steps). A comparison is made of the utility of each pipeline permutation, including measurements of:

- Number of reads alignable to reference sequence
- Accuracy of alignment to reference sequence as determined by samtools stats

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.2: Methods

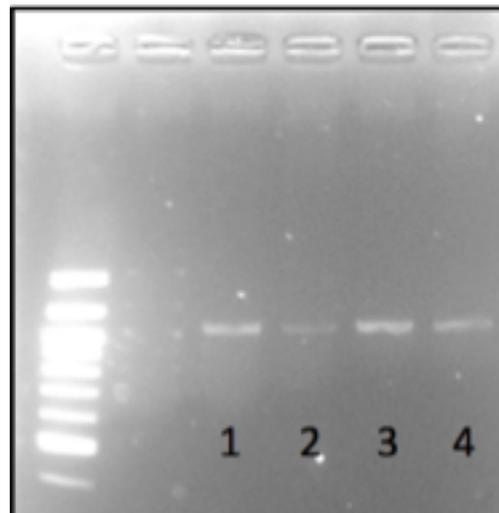
- Patterns of errors, determined by a sliding window analysis (*see chapter 2.6.2.1 – 9*)
and figure 2.4

3.3 Results

3.3.1 Cell line Validation

3.3.1.1 Demonstration of *iCas9* construct insertion at correct locus

Junction PCR demonstrated insertion of the plasmid into the AAVS1 cell line in two separate clones isolated from Puromycin selection (*figure 3.5*).



*Figure 3.5, Junction PCR demonstrating insertion of *iCas9* construct at AAVS1 locus – lane 1 & 2 DNA extract from clone A, lane 3 & 4 DNA extract from clone B*

3.3.1.2 Demonstration of Cas9 protein production

A puromycin kill curve experiment (data not shown) demonstrated complete death of WT-hIPCS cells not containing the construct at a concentration of 4 μ g/mL puromycin. Two days after nucleofection with the TALEN plasmids and *iCas9* plasmid as described above, the medium was changed to include this concentration of puromycin. After 1 week of treatment, multiple puromycin-resistant colonies had emerged.

Cells treated with doxycycline demonstrated binding with the FLAG-M2 antibody, (*figure 3.6*). The signal was not well localised to the nucleus, however untreated control cells did not show a comparable antibody response (not shown), leading to the conclusion that the response seen was indeed genuine.

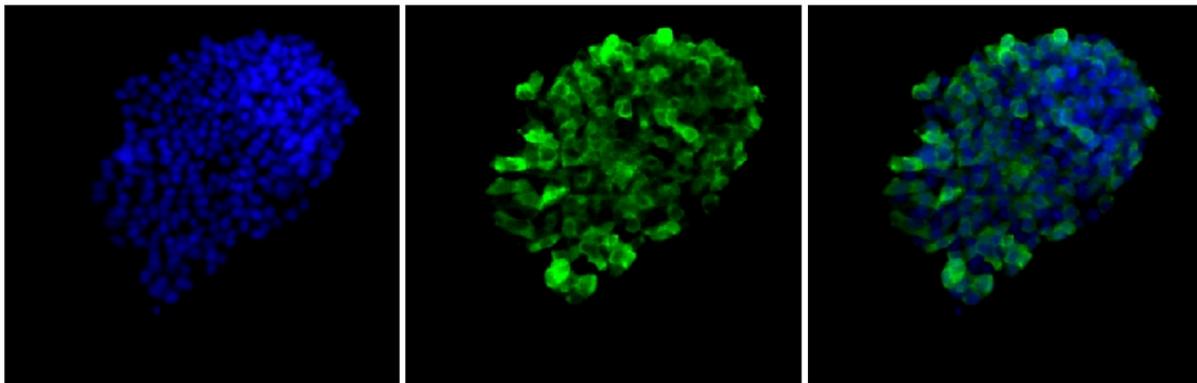
The average of the three experiments conducted showed a clear dose-response curve, showing an increased ratio of FLAG-M2 signal to GAPDH signal at 24 hours (7.31, SEM 3.0), peak at 48 hours (14.76, SEM 4.3), with a decline in the ratio over 72 hours after the doxycycline was removed (24 hours off: 11.3, 3.79, 48 hours off: 3.56, SEM 0.29, 72 hours off:

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.3: Results

0.22, SEM 0.12). There was no detectable signal in the control wells, confirming that Cas9 expression in this cell line is dependent on the presence of doxycycline and by 72 hours, the FLAG-M2 signal had returned to near baseline.

A:



B:

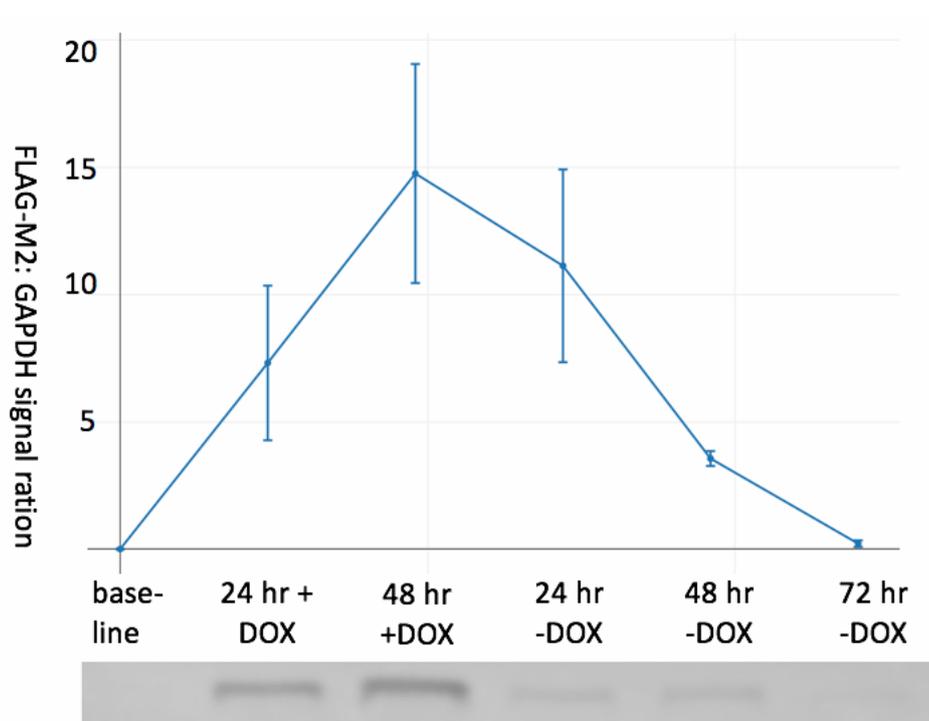


Figure 3.6 Demonstration of Cas9 protein production in response to doxycycline treatment
A: demonstrating DAPI nuclear staining (blue), FLAG-M2 staining (green) and combined images in CnIPS after treatment with $2\mu\text{g}/\mu\text{l}$ doxycycline for 48 hours

B: dose-response curve showing ratio of FLAG-Tag antibody signal to GAPDH antibody signal for 48 hours treatment with doxycycline at $2\mu\text{g}/\mu\text{l}$ and 72 hours culture in E8 C: Western Blot images of signal from FLAG-Tag antibody

3.3.1.3 Desktop Gene Editing Screen

As a rapid screen for editing efficiency, gRNA targeting NANOG was transfected. This screen has been previously described as a good benchtop check of editing efficacy [269]; NANOG is a pluripotency factor and effective knockout leads to rapid loss of pluripotent cell morphology. As the cell colonies rapidly deteriorate, the DNA is not extracted for sequencing. Rather, this check provided confidence in advancing to the next stage of investigation.

All of the cell lines transfected with NANOG_ gRNA demonstrated rapid changes in morphology within 48 hours of treatment, followed by collapse of culture and widespread cell death within 120 hours. The change in morphology happened regardless of the plating density of the cells, demonstrating a high level of efficacy even at higher cell densities.

The control wells did not demonstrate similar effects, demonstrating that successful gene editing of NANOG was the most likely cause for the loss of morphology. Neither gRNA, or lipofectamine alone was sufficient to cause the differentiation, in presence or absence of doxycycline (*figure 3.7*).

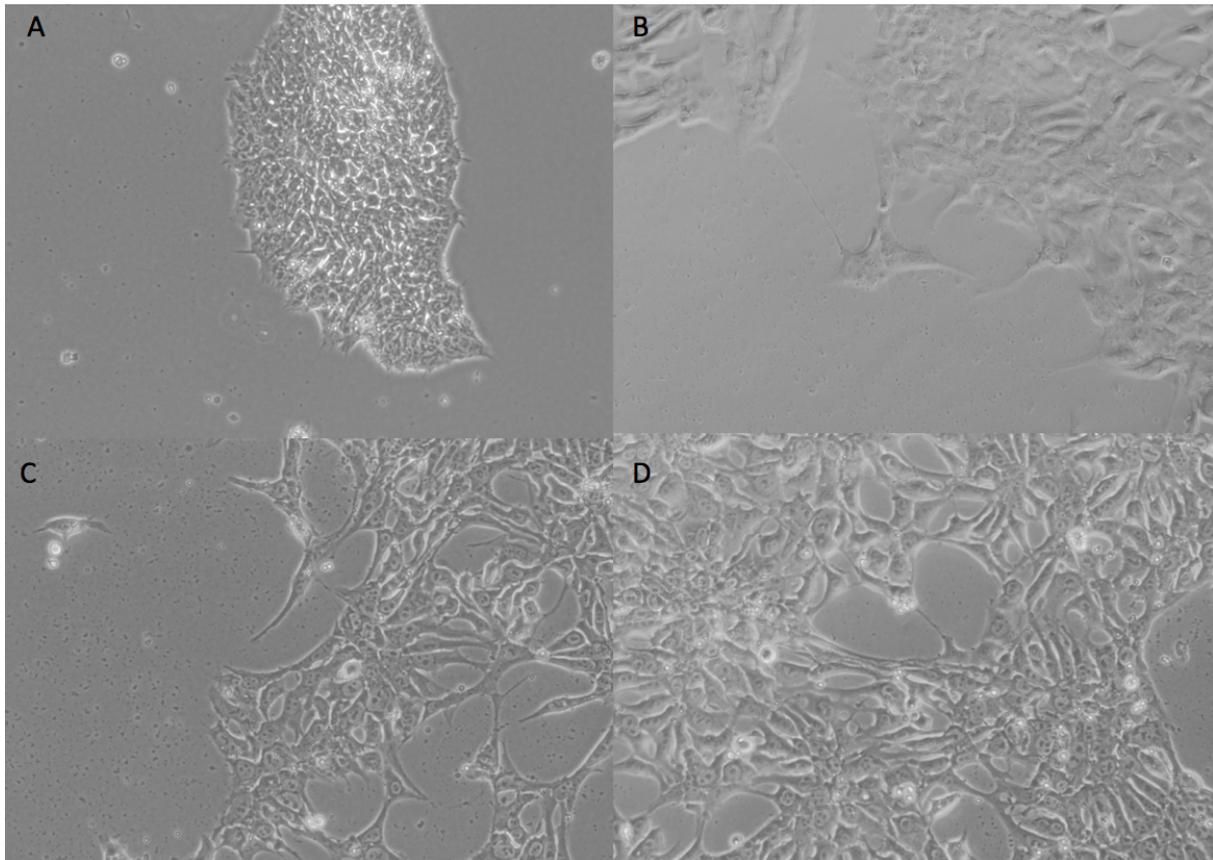


Figure 3.7: Cell response to gRNA targeting NANOG. The figure demonstrates iCas9 cells 48 hours post transfection with gRNA NANOG_A.

A: control well treated with doxycycline and lipofectamine RNAiMAX but without gRNA exhibiting standard stem cell morphology – small, tightly packed cells with a clearly defined margin. B, C and D; cells plated at 0.2×10^6 cells/mL, 0.4×10^6 cells /mL, and 0.6×10^6 cells/mL respectively. Cells demonstrated loss of stem cell morphology at 48 hours post-treatment as the maintenance of pluripotency was disrupted, increasing in size, losing tight packing arrangement and developing characteristics of fibroblasts and neuronal progenitor cells

3.3.2 Testing nanopore sequencing as a screening tool for indels

The second part of the pipeline to be set up was the nanopore screening pipeline. The data in this section demonstrates its utility using a pair of previously described CRISPR-generated mutations within the same amplicon of the *GRIN2A* gene. This gene was mutated by Dr William Plumbly as part of his PhD project – it was chosen for this project because of the availability of high-quality DNA extracts and validated sanger data describing the mutations.

3.3.2.1 Sequencing metrics and barcode decomplexing

The MinION flow cell began the run with 550 available pores and was run overnight for a total runtime of 2 hours, 17 minutes generating 274,796 reads.

Data was successfully demultiplexed and aligned to the reference sequence, converted to binary format and sorted. Mutations could be clearly visualised at the cut site in igv. The visual calling of mutations is a fairly subjective process, however the DEL wells were clearly mutated, as were the INS wells. Interestingly, the deletions aligned to the reference genome more accurately than the insertions, which were scattered around the cut site on the alignment visualisation. (*Figure 3.8*)

This demonstrated that insertions and deletions could be identified from the nanopore data, but that the amount of noise arising from stochastic sequencing errors created a potential obstacle that needed to be overcome with further data processing.

3.3: Results

3.3.2.2 Testing Variant Screening

The best results for the pileup engine varied by aligner, porechop parameters and minimum read depth at cut site (MRD). It is not reasonable to present all 225 data points in text format, however there are important comparisons to be made.

In determining which aligner was best suited to identify insertions and deletions, the assumption was made that in wells containing heterozygous deletions (DEL wells) (see *figure 3.2*) 50% of the reads should contain a deletion and 0% contain an insertion, with the inverse would be true for INS wells. The absolute ratio of reads containing insertions, deletions and WT sequence in the MIX wells are not known.

Alignment with bwa-ont was found to consistently generate the highest insertion scores in INS wells. With barcode settings at their most stringent (T80, D5) the mean insertion score was 45.97% (Standard error of the mean (SEM) 1.83%) at MRD of 12 and 45.97% (SEM 1.83%) at a MRD of 24, compared to minimap2 (38.8% (SEM 2.52%) at MRD 12, $p = 0.013$ and 40.92% (SEM 1.52%) at MRD 24, $p = 0.02$), and compared to ngmlr (39.91% (SEM 2.54%) at MRD 12, $p = 0.018$ and 42.14% (SEM 1.35%) at MRD 24 $p = 0.032$) (*figure 3.9*).

Changing the alignment package made no statistically significant difference in the scores for deletion calls in DEL wells as measured by student's T-test, although minimap2 produced higher scores at all read depths – 43.94%(SEM 2.64%) at MRD12 and 45.98% (SEM 1.86%) at MRD of 24 compared to bwa-ont (41.26% (SEM 2.01%) at MRD12 $p = 0.11$ and 42.45%(SEM 1.78%) at MRD 12, $p = 0.07$) and ngmlr (41.44% (SEM 2.60%) at MRD 12, $p = 0.11$ and 41.46% (SEM 2.88%) at MRD 24, $p = 0.07$) (*figure 3.12*).

Changing the minimum threshold for barcode accuracy had the biggest impact on the scores for deletion calls in DEL wells and for insertion calls in INS wells. Using minimap2 for deletion calls in DEL wells and a barcode differential of 5%, with a MRD of 24; a barcode accuracy threshold of 70% produced a mean deletion score of 33.32% (SEM 3.70%) for deletions compared with threshold 75% (deletion score 43.96% (SEM 2.011), $p = 0.02$) and a threshold of 80% (deletion score 45.98% (SEM 1.87%), $p = 0.005$). There was no significant difference between deletion scores at a threshold of 75% and 80% ($p = 0.33$).

Using bwa-ont for insertion scores in INS wells with a barcode differential of 5% and MRD of 24, an accuracy threshold of 70% produced a mean insertion score of 31.53% (SEM 3.51%) compared to a threshold 75% (41.91% (SEM 2.51%), $p = 0.01$) and threshold 80%

3.3: Results

(45.97% (SEM 1.83%), $p= 0.001$). Again, there was no statistically significant difference in between thresholds of 75% and 80% ($p= 0.07$).

All aligners, threshold scores and differential scores generated a statistically significant difference in the deletion scores and the insertion scores for DEL wells, INS wells, WT wells and MIX wells at read depth of 12 and of 24 (*figure 3.10, table 3.3*).

Increasing the minimum depth mutation site required for a well to be included in the analysis impacted the number of wells available for analysis. In several wells where either the barcode PCR stage had failed, or the porechop decomplexing had been unable to bin sequences, the read depth was <10 (*figure 3.11*).

Increasing the minimum depth required at the mutation site impacted the number of wells available for analysis. In several wells where either the barcode PCR stage had failed, or the porechop decomplexing had been unable to bin sequences, the read depth was <10 (*figure 3.11*). This can be seen with reads aligned from different targets, however not always to the same extent. There was a greater fall-off in the number of wells available for analysis in the GRIN2A experiment than in experiments targeting CHD2 and CANCA1C (*figure 3.12*)

There is a trade-off between the minimum read depth set and the accuracy of the binning of reads. At a higher minimum read depth, the results become more reliable, however there are an increasing number of wells where the not enough data is demultiplexed. Similarly, increasing the stringency of the threshold and differential settings increases the accuracy of the data, but at the cost of reducing the number of reads per well.

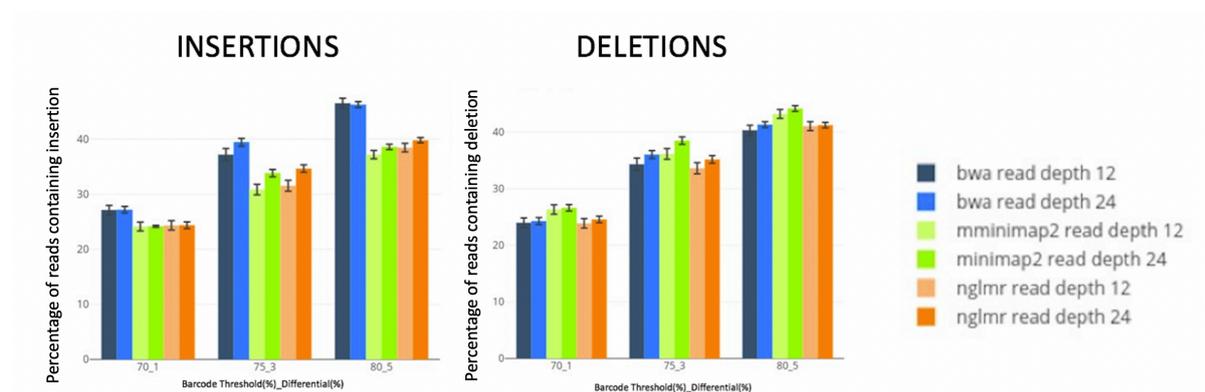
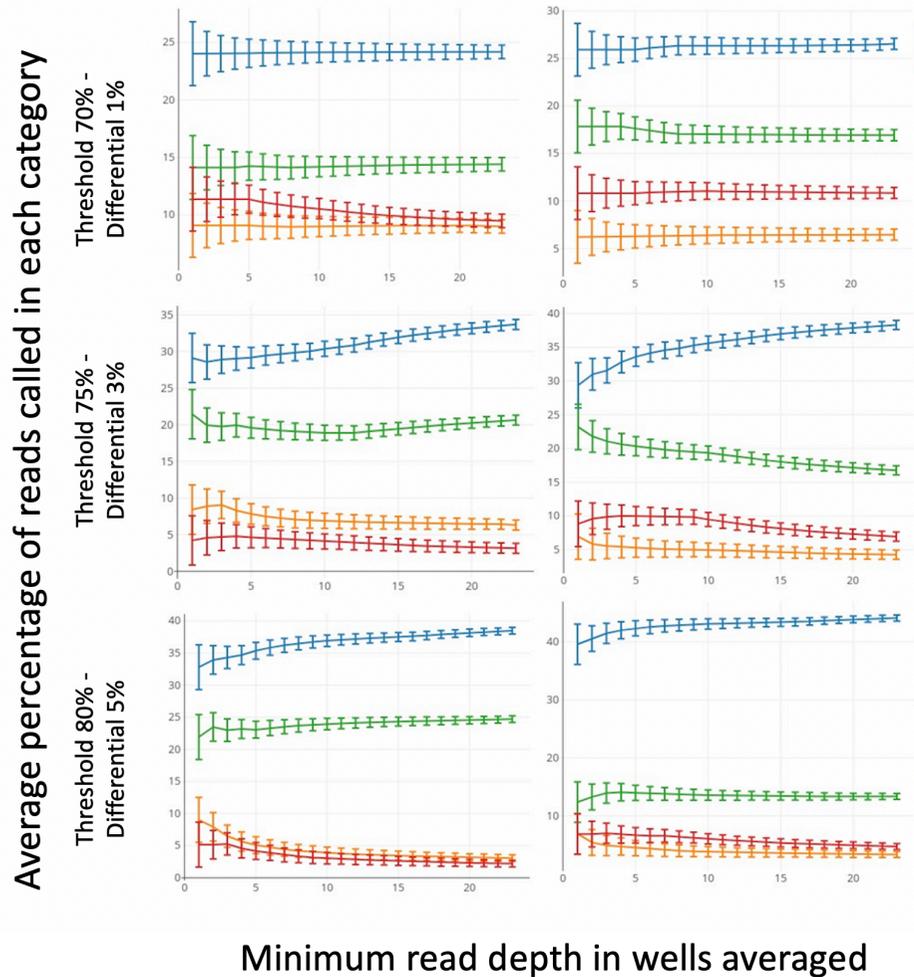


Figure 3.9: bar graph demonstrating the average percentage of reads (y axis) in wells containing known heterozygous mutations (left - heterozygous 41bp insertion, right – heterozygous 16bp deletion) in which that mutation is identified, stratified by aligner used (colour) minimum read depth of wells taken into account (shade). For each graph, data is displayed at three different stringency settings on porechop (x axis). In both cases, for a perfect sequencing run of a pure heterozygous clone, one would expect exactly 50% of the reads to contain the mutation. It is therefore assumed that bwa (dark blue) was more accurate in identifying insertions, where minimap 2 was more accurate in identifying deletions.

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.3: Results



Read called with correct mutation in well containing heterozygous mutation	True call
Read called as wrong type of mutation in well containing heterozygous mutation	Miscall
Read called containing mutation in well containing both mutations (INS left, DEL right)	MIX call
Read called as containing mutation in well containing wild type DNA (INS left, DEL right)	False positive

Figure 3.10 – Percentage of correctly and incorrectly called reads per well, with different porechop demultiplexing stringencies. The percentage of reads correctly called increases and the percentage of wells containing miscalled mutations or false positive calls falls, as wells with lower read depth are excluded (x axis)

This demonstrates that the pipeline is able to successfully distinguish between wells containing insertions, deletions, mixed clones and wild type DNA extractions

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.3: Results

Aligner	Porechop parameters		Well type	Well type	p.value
	Threshold (%)	Differential (%)	A	B	
BWA For insertions	70	1	INS	WT	0.0001
			INS	DEL	4.82 x10 ⁻⁵
			INS	MIX	0.015
			MIX	WT	0.008
			MIX	DEL	0.001
	75	3	INS	WT	4.36x10 ⁻⁵
			INS	DEL	1.1x10 ⁻⁵
			INS	MIX	0.002
			MIX	WT	0.007
			MIX	DEL	0.009
	80	5	INS	WT	6.17x10 ⁻⁵
			INS	DEL	6.17x10 ⁻⁵
			INS	MIX	0.007
			MIX	WT	0.0001
			MIX	DEL	0.0001
Minimap2 for deletions	70	1	DEL	WT	2.5x10 ⁻⁵
			DEL	INS	2.05x10 ⁻⁵
			DEL	MIX	0.057
			MIX	WT	0.016
			MIX	INS	0.0001
	75	3	DEL	WT	5.6x10 ⁻⁶
			DEL	INS	8.19x10 ⁻⁵
			DEL	MIX	0.00056
			MIX	WT	0.0002
			MIX	INS	8.19x10 ⁻⁵
	80	5	DEL	WT	0.0001
			DEL	INS	4.05x10 ⁻⁵
			DEL	MIX	0.00022
			MIX	WT	0.0002
			MIX	INS	0.0001

Table 3.3: p values for comparisons between mean scores for different call types during demultiplexing, by threshold and differential setting used in the pipeline. The p values describe the significance of the difference between the number of reads called as insertions or deletions in each sample type (DEL = well containing pure deletion, INS = well containing pure insertion, MIX = well containing mixed clone of INS, DEL and WT, WT = well containing no mutations – see also figure 3.10)

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.3: Results

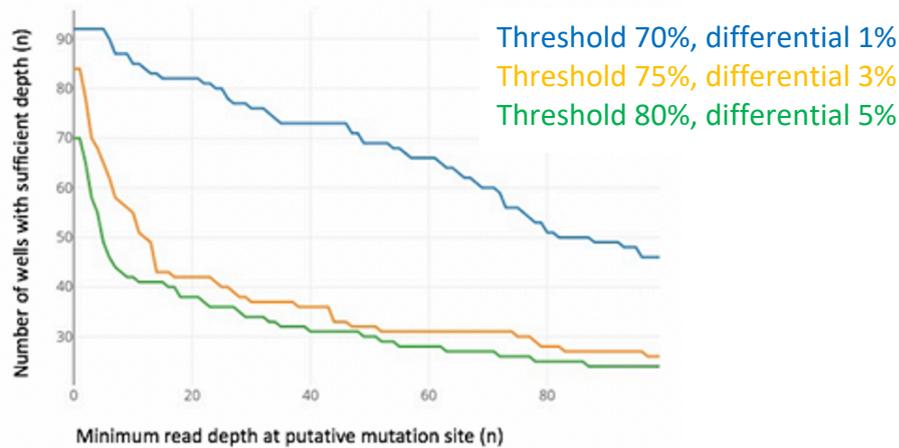


Figure 3.11 – graph demonstrating number of wells available for analysis at MRD 0 – 80 reads in the GRIN2A sequencing run. Note the rapid drop-off in the number of wells available at higher demultiplexing stringencies (orange, green)

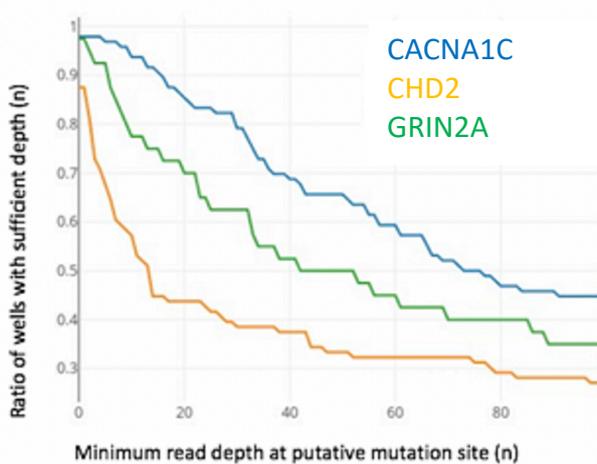


Figure 3.12 – graph comparing the proportion of wells sequenced available for analysis at minimum read depth 0-80 in three sequencing runs of different genomic targets, with the same demultiplexing settings (demultiplexing threshold 75%, demultiplexing differential 3%, aligned with minimap2): GRIN2A, CHD2 and CACNA1C sequencing run. This demonstrates that the fall off in wells is variable between targets and may be more significant for the GRIN2A run on which the system was tested than other sequencing runs

3.3.3 Subsequent use of system in creation and description of cell lines

This system has subsequently been used for the creation of multiple cell lines. The detected success rates for each editing experiment and the spectrum of mutations described can be found in *table 3.4*. Igv demonstration of the mutations can be found in *figure 3.13*. These genes were chosen by colleagues (pre and post doctoral) for use in their research, rather than for any inherent properties of the target sequence.

Gene	Reference sequence length (nt)	gRNA locus as nt of reference sequence	Wells identified n_mutations / n_wells sequenced (%)	Notes
CACNA1C	756	303 344 521	8/96 (8.33%)	Triple gRNA transfection Previous failure with single gRNA
CHD1L	1028	651 704 752	15/42(35.7%)	Triple gRNA transfection
CHD2	811	556	0/60 (0%)	Single gRNA used, no successful edits
CHD2	912	670	4/40 (15%)	Single gRNA used – required subcloning, all wells mixed
CHD8	905	544 598 685	42/50 (84%)	Triple gRNA used Very high editing rates 4 wells chosen
FADS2	709	329	3/96 (3.13%)	Single gRNA
GJAJ5	1423	538 858	40/60 (66.6%)	Triple gRNA transfection
SETD1A	1050	434 659 675	4/96 (4.16%)	Triple gRNA transfection
SLC6AC	1258	363 453 1044	72/72 (100%)	Triple gRNA transfection

Table 3.4 (page 1/2) – new mutations created and detected using iCas9-nanopore pipeline

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.3: Results

Gene	Reference sequence length (nt)	gRNA locus as nt of reference sequence	Wells identified n_mutations / n_wells sequenced (%)	Notes
CACNA1C	756	303 344 521	8/96 (8.33%)	Triple gRNA transfection Previous failure with single gRNA
CHD1L	1028	651 704 752	15/42(35.7%)	Triple gRNA transfection
CHD2	811	556	0/60 (0%)	Single gRNA used, no successful edits
CHD2	912	670	4/40 (15%)	Single gRNA used – required subcloning, all wells mixed
CHD8	905	544 598 685	42/50 (84%)	Triple gRNA used Very high editing rates 4 wells chosen
FADS2	709	329	3/96 (3.13%)	Single gRNA
GJAJ5	1423	538 858	40/60 (66.6%)	Triple gRNA transfection
SETD1A	1050	434 659 675	4/96 (4.16%)	Triple gRNA transfection
SLC6A	1258	363 453 1044	72/72 (100%)	Triple gRNA transfection

Table 3.4 (page 2/2) – new mutations created and detected using iCas9-nanopore pipeline

3.3.4 Accuracy and output of pipelines using wild-type datasets

In order to further describe the accuracy of the nanopore pipeline and determine the best set of parameters for use in further experiments, a run was performed using three well-sequenced targets from elsewhere in this project. The amplicons chosen were in CHD2, NRXN1 and FADS2.

The average accuracy, given by the number of mismatches present in the aligned sequences, and total reads aligned for each pipeline are shown in *table 3.5* and *figure 3.14*. Note; as there is no way to determine the raw number of sequences of each target present in the sequence library if CHD2 only made 15% of the library compared to 40% NRXN1 and 45% FADS2, then the means should be weighted accordingly. It was felt that the total number of reads mapped to *all* targets was a more meaningful measure of pipeline performance. The accuracy measurements given are, however, means.

In general, all pipeline add-ons to normal 1D sequencing led to a significant decrease in the number of reads available for alignment. For 1D basecalling, Guppy outperformed Albacore in terms of accuracy and output, regardless of the aligner used.

BWA produced a greater number of reads, of statistically significant lower alignment accuracy than minimap2 when used for alignment of 1D data($p=0.0082$). This difference disappeared for 1DSQ reads base-called by Albacore, and for any reads that had undergone correction with Canu ($p=0.948$)

The Guppy model for 1DSQ base calling produced the lowest accuracy and fewest number of reads of any pipeline when aligned with BWA. Even when aligned with minimap2, the accuracy was comparable to the accuracy of 1D base-calling but with a far smaller fraction of reads aligned.

The 1D Flip-Flop model of basecalling produced very few useable reads, of accuracy comparable with standard 1D basecalling.

Canu correction, in general, decreased the number of aligned reads but significantly increased their accuracy ($p=0.002$ for bwa, $p=0.0007$ for minimap2). The exception to this general rule is the sequences called by the 1DSQ and 1D_Flip_Flop models. In these cases, accuracy *and* readcount were notably increased.

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.3: Results

Aligner	Correction with Canu	PIPELINE	Total reads aligned	Alignment accuracy (mean, %)	Standard error of accuracy (%)			
BWA	Canu	Albacore 1DSQ	4469	0.987	0.005			
		Guppy 1DSQ	5981	0.963	0.005			
		Guppy Flip Flop	52469	0.972	0.003			
		Albacore 1D	55382	0.948	0.006			
		Guppy 1D	65519	0.955	0.004			
		No Canu	Albacore 1DSQ	8174	0.938	0.005		
			Guppy 1DSQ	34650	0.765	0.018		
			Guppy Flip Flop	142	0.837	0.009		
			Albacore 1D	183432	0.823	0.002		
			Guppy 1D	186658	0.841	0.004		
			mmap2	Canu	Albacore 1DSQ	4449	0.985	0.004
					Guppy 1DSQ	5938	0.967	0.005
Guppy Flip Flop	51886	0.975			0.002			
Albacore 1D	54441	0.957			0.007			
Guppy 1D	64935	0.961			0.004			
No Canu	Albacore 1DSQ	7984			0.94	0.002		
		Guppy 1DSQ		21588	0.825	0.005		
		Guppy Flip Flop		117	0.864	0.004		
		Albacore 1D		159448	0.844	0.005		
		Guppy 1D		163491	0.86	0.005		

Table 3.5 – total wild-type reads aligned and accuracy of alignments with each tested pipeline

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.3: Results

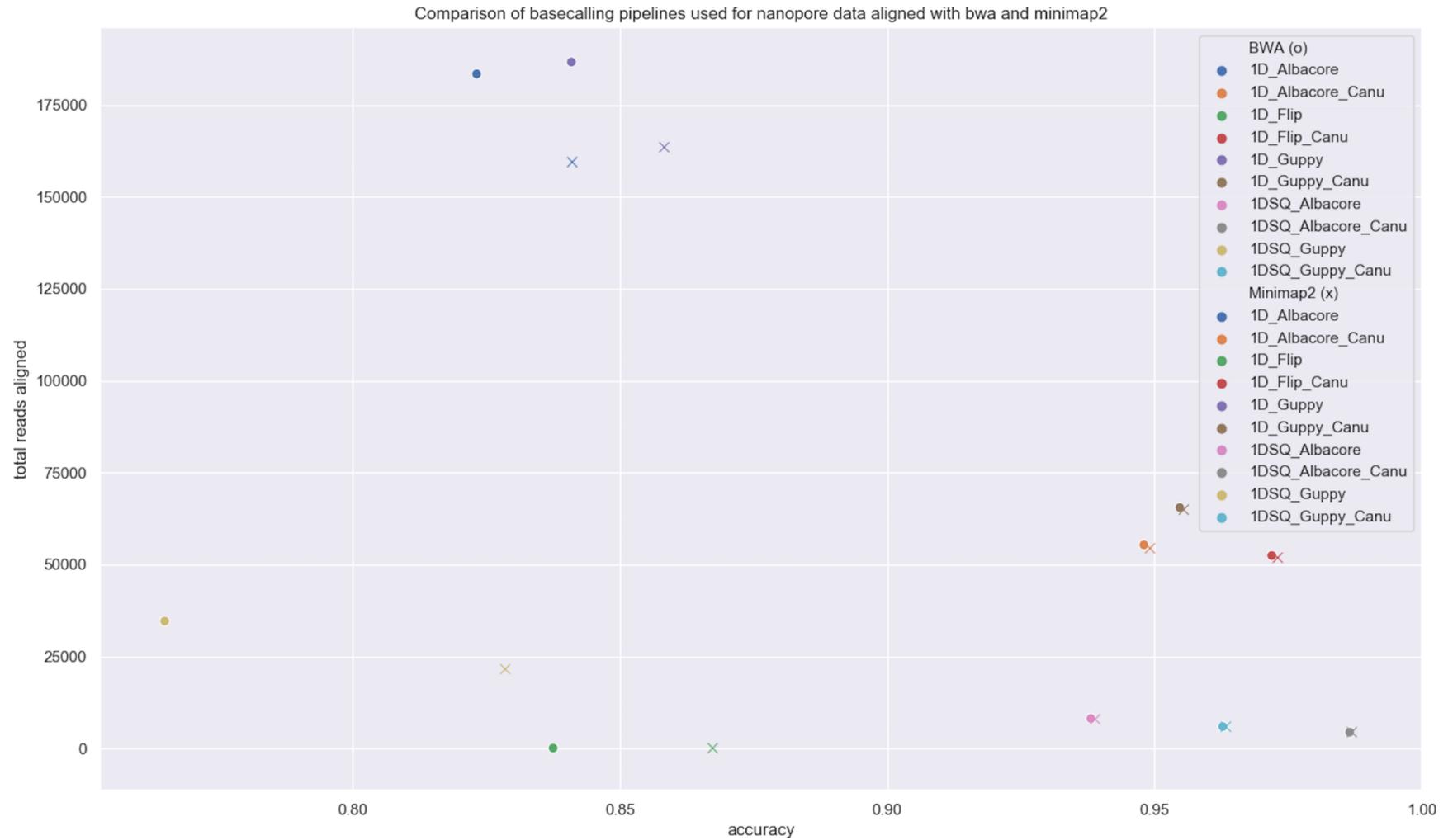


Figure 3.14 – total reads aligned, and accuracy of alignments by each pipeline, aligned with either minimap2 (x) or bwa (o). Colours on legend correspond to different pipeline choice (flow cell 1D or 1DSQ, base-caller Albacore, Flip-flop, Guppy, Canu correction or none)

3.3.5 Sequencing error context

The following data focuses on the pattern rather than the abundance of errors in each sequence structure. All data presented are as ratios, rather than raw numbers; the raw numbers of errors in each sequence context themselves relate to the accuracy counts described in *section 3.3.4*. Therefore in all cases, it can be assumed that the more accurate pipelines have a lower raw number of errors in each context than their less accurate counterparts.

3.3.5.1 Sequence composition

The sliding window generated sequence composition of each reference sequence can be found in *table 3.6*. Full tabulation of the makeup of each reference sequence in dinucleotides, trinucleotides and tetranucleotides would take several thousand lines for little additional insight.

	CHD2	FADS2	NRXN1
Length in nucleotides (nt)	821	711	834
A (nt, %)	220 (26.8)	136 (19.1)	122 (14.6)
C (nt, %)	159 (19.4)	204 (28.7)	260 (31.2)
G (nt, %)	192 (23.4)	189 (26.6)	315 (37.8)
T (nt, %)	249 (30.3)	182 (25.6)	137 (16.4)
GC content (%)	42.8	55.3	68.9

Table 3.6 – sequence composition of each reference amplicon

3.3.5.2 Mononucleotide context of miscalled indels

The ratios of starting mononucleotide for deletions to the number of times that mononucleotide is represented in the reference sequence can be found in *table 3.7* and *figure. 3.15*

No clear pattern emerged – to highlight a few examples; in 1D_Albacore base called data, errors in CHD2 alignments were over-represented in Adenine, in FADS2 and NRXN1 errors in Adenine was under-represented. In NRXN1, errors beginning in thymine were significantly over-represented.

There was also inconsistency in the pattern between base calling pipelines for the same reference sequence – Adenine was over-represented in CHD2 reads aligned from 5 / 9 base callers, but thymine and guanine were over-represented in others.

Minimap2			A	C	G	T
Albacore	1D	CHD2	2.21	0.40	0.52	0.68
		FADS2	0.68	1.02	1.05	1.16
		NRXN1	0.30	0.27	0.22	4.81
	1DSQ	CHD2	1.12	0.95	1.03	0.91
		FADS2	0.99	1.08	1.05	0.85
		NRXN1	1.35	0.98	0.87	1.00
	1D Canu	CHD2	1.44	0.82	0.87	0.83
		FADS2	0.83	1.02	1.18	0.93
		NRXN1	0.52	0.40	1.67	1.05
	1DSQ Canu	CHD2	0.59	0.28	0.43	2.25
		FADS2	0.26	0.95	1.75	0.82
		NRXN1	1.10	0.55	0.76	2.32
Guppy	1D	CHD2	1.80	0.61	1.08	0.48
		FADS2	0.85	0.95	0.88	1.28
		NRXN1	0.36	0.36	0.37	4.24
	1DSQ	CHD2				
		FADS2	0.74	0.93	1.02	1.25
		NRXN1	0.74	0.53	1.36	1.31
	FlipFlop	CHD2	0.82	1.32	1.01	0.95
		FADS2	0.82	1.12	1.13	0.87
		NRXN1	1.04	1.17	0.97	0.71
	1D Canu	CHD2	1.28	0.78	1.22	0.72
		FADS2	0.64	0.86	1.60	0.82
		NRXN1	1.86	0.50	0.80	1.64
	1DSQ Canu	CHD2	0.82	0.76	0.87	1.41
		FADS2	1.10	0.79	0.92	1.24
		NRXN1	0.77	0.84	0.85	1.86
	FlipFlop Canu	CHD2	0.98	0.51	1.83	0.68
		FADS2	0.93	1.09	.095	1.02
		NRXN1	0.57	0.64	1.45	1.05

Table 3.7 – Mononucleotide context of nanopore sequencing errors by gene. Each number represents the ratio of number of errors recorded starting at a specific nucleotide to how often that nucleotide appears in the reference sequence

3: Setup and testing of an inducible Cas9 gene editing and nanopore sequencing pipeline

3.3: Results

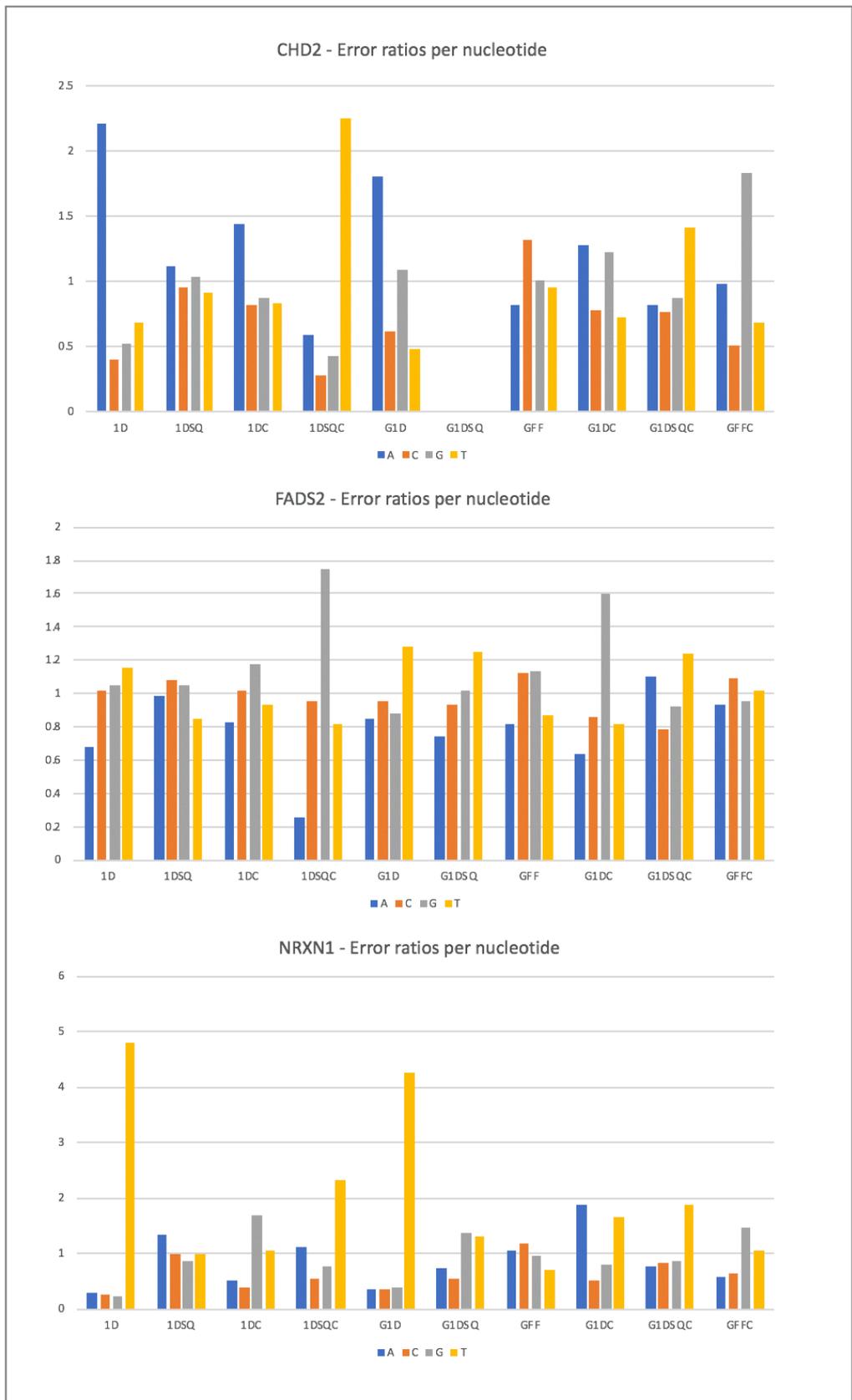


Figure 3.15: mononucleotide error context of nanopore sequencing by gene and pipeline, for all three targets, based on alignment with minimap2. , grouped by basecalling pipeline. 1D = Albacore 1D, 1DSQ = Albacore 1DSQ, 1DC = Albacore 1D Canu, 1DSQC = albacore 1DSQ Canu, G1D = Guppy 1D, G1DSQ = Guppy 1DSQ, GFF = Guppy Flip-Flop, G1DC = Guppy 1D Canu, G1DSQC = Guppy 1DSQ Canu, GFFC = Guppy Flip Flop Canu

This demonstrates that the mononucleotides with the most frequent errors vary significantly from pipeline to pipeline and from target to target within the same pipeline

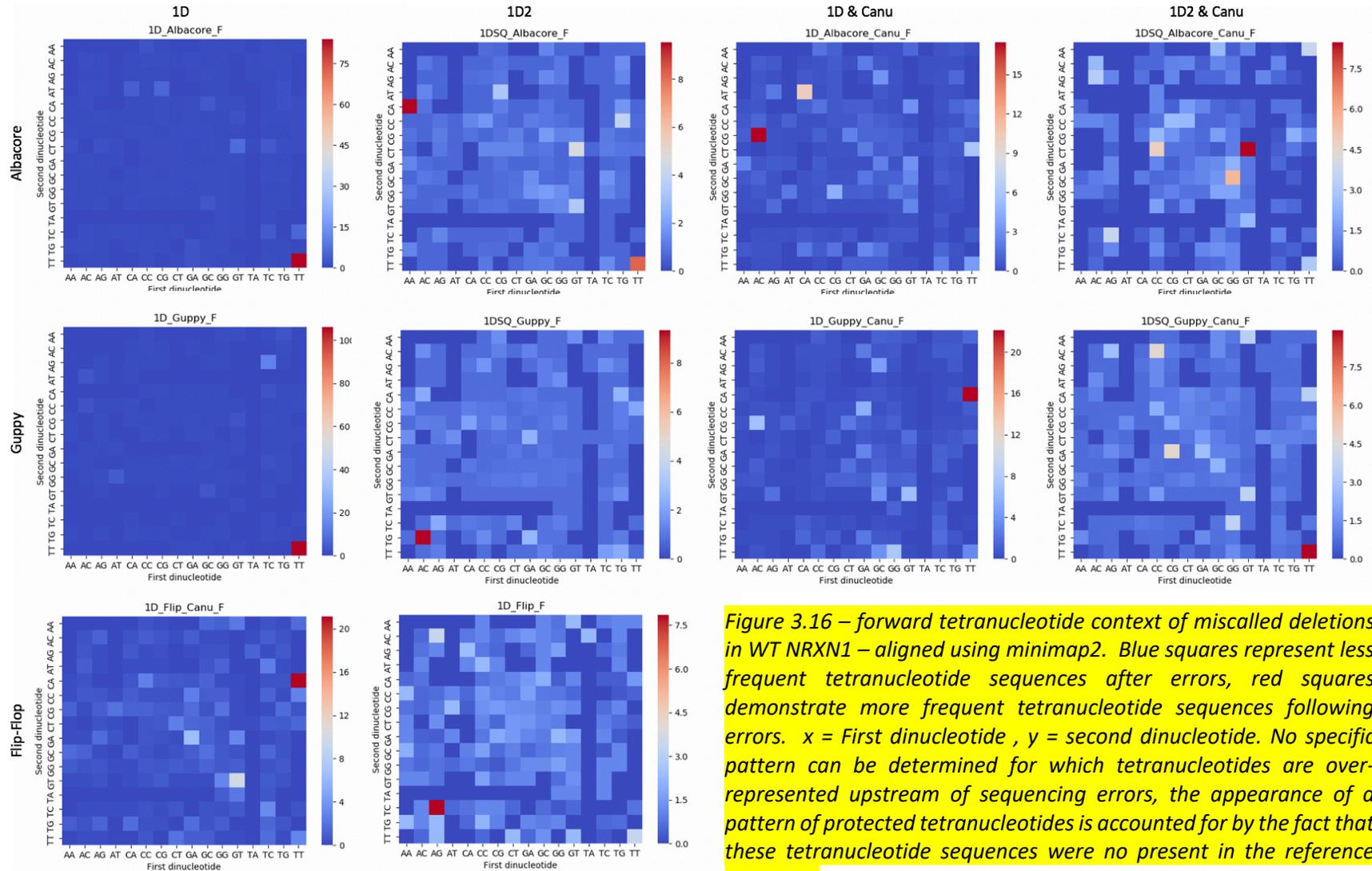
3.3.5.3 Dinucleotide, Trinucleotide and Tetranucleotide Error Contexts

When viewed in isolation, the dinucleotides 3' and 5' of the first base of the sequencing error appear to have a similarly unpredictable pattern to the mononucleotides, with certain dinucleotides being over-represented, depending on the reference sequence, but a shifting pattern between pipelines and between genes.

The *figure 3.16* demonstrate the relatively random context of the errors when the tetranucleotide sequence upstream of the errors is considered either in terms of the sequence upstream. Similar data for sequence downstream and bilateral tetranucleotide context, or downstream of the first base of the error and can be found in the supplementary material.

3: Setup and testing of an inducible Cas9 gene editing and nanopore screening pipeline

3.3: Results



1D Albacore			1D Guppy			
	CHD2	FADS2	NRXN1	CHD2	FADS2	NRXN1
	Dinucleotide, ratio					
1	AG 2.027904	TC 1.194310	TT 8.249940	AG 1.364937	TC 1.404247	TT 9.271541
2	GA 1.272490	GC 1.180625	TC 1.056092	CA 1.352060	AT 1.290035	TC 1.541147
3	CT 1.142644	CT 1.128224	GC 0.812825	CT 1.334760	GG 1.261269	AT 1.438404
4	GG 1.023378	TG 1.024718	CT 0.812530	AC 1.186902	GA 1.217014	AG 0.766622
5	CC 0.978250	TA 1.015163	GA 0.798725	TA 1.044864	TT 1.170206	TG 0.738357
6	AA 0.974488	CG 1.009192	CC 0.775597	TG 0.985788	CT 0.989838	AC 0.733086
7	TC 0.945675	CA 1.002134	AG 0.774149	TC 0.978951	GT 0.964706	GT 0.730618
8	AT 0.910633	GG 0.988020	GT 0.759282	AT 0.973578	CA 0.929356	GA 0.711863
9	TG 0.857752	GT 0.965643	CG 0.739829	GC 0.947873	CG 0.912760	CA 0.708928
10	GT 0.848326	AT 0.962614	GG 0.73660	GT 0.924176	TG 0.884363	GG 0.686177

Table 3.8 - top ten overrepresented dinucleotide sequences, upstream of miscalled insertions representing dinucleotide and normalised ratio of errors to frequency in the target sequence – for example if AG dinucleotides form 2% of the target sequence, but contain 4% of the errors then the ratio will be 2.0.

Table 3.8 demonstrates the top ten over-represented dinucleotides immediately upstream of miscalled deletions. Unlike the tetranucleotide sequence contexts, there is some consistency between different pipelines, with the most common dinucleotide contexts remaining identical for the same target sequence between different pipelines, however it is difficult there are few similarities in the other nine most over-represented dinucleotides. Similar data for trinucleotide and tetranucleotide sequences and for miscalled insertions shows a similar variation between platforms and can be found in the supplementary data section (appendix III).

3.3.6 Length of errors made in nanopore sequencing

The vast majority of the errors were mis-calls of 1 and 2bp length. *Figure 3.17* demonstrates the rapid fall in error rates as the length of error is increased from 1 to 200. The data displayed is for NRXN1, however the pattern is near identical for CHD2 and FADS2 (data not shown).

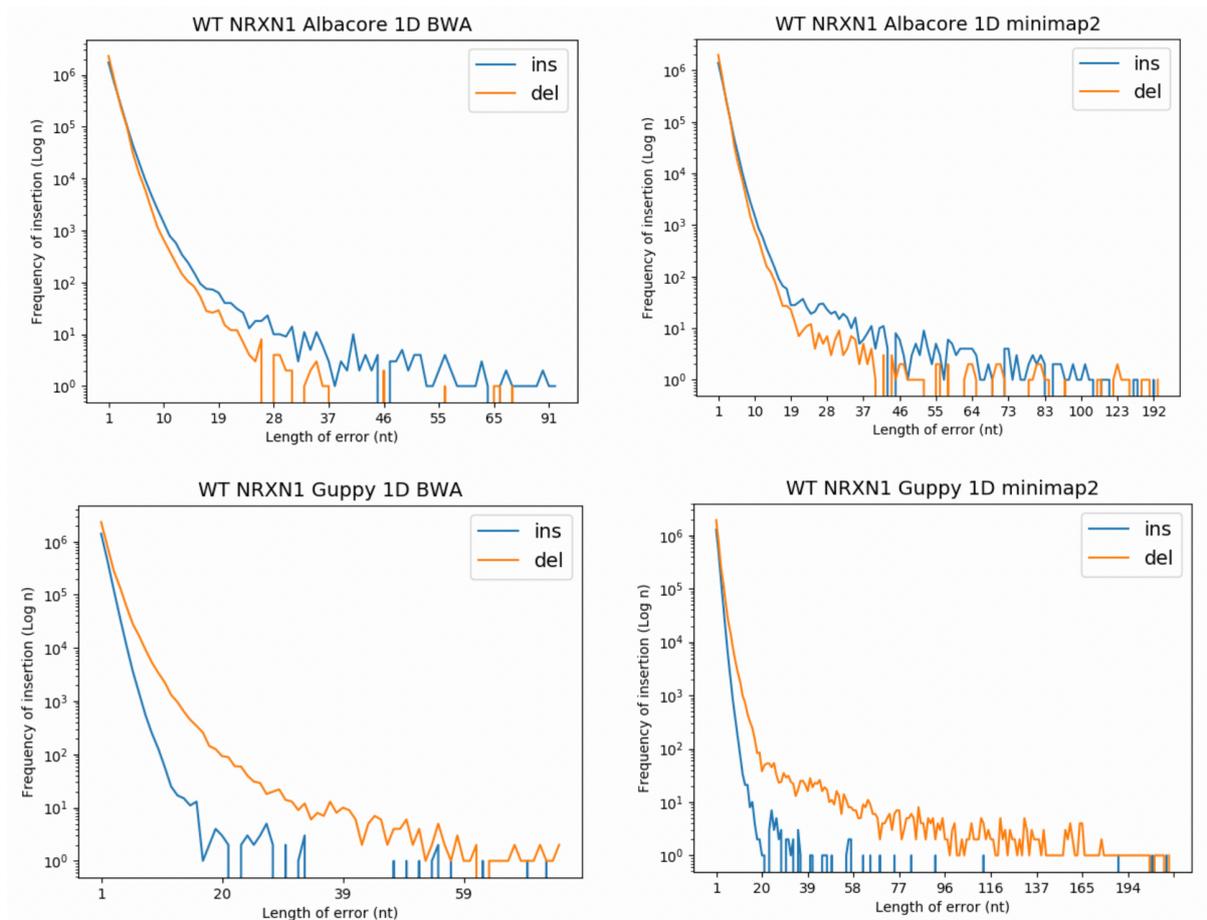


Figure 3.17 – Error rate (log scale) by length of miscalled insertion and deletion for nanopore sequencing data for NRXN1

3.4 Conclusions and Discussion

3.4.1 Use of the nanopore for investigation of CRISPR induced mutations

The data in this chapter supports the use of multiplexed nanopore sequencing as a suitable high-throughput screening tool to find CRISPR induced mutations. The cost and turn-around time, as well as the generation of detailed sequencing data, make it an attractive alternative to established screening mechanisms, reliant on either restriction endonuclease assays or sanger sequencing.

A similar approach has been described using second generation sequencing on the LifeTechnologies IonPGM platform[275]. Although this approach provides more accurate reads, it is less well suited to the description of larger deletions created by co-transfection of multiple gRNA[285]. The cost and workload for library preparation is also higher than that for the system described.

The single cell sub-cloning step ensures that mutations present in as low as 1.04% (1/96) of cells can be detected. Theoretically with a longer sequencing run a 384 well plate could be multiplexed and analysed, allowing for detection of mutations occurring at 0.02%.

The accuracy of minION generated reads is an obvious concern for any use. In particular, it provides a challenge in the interpretation of mixed wells, containing either >1 mutation or a combination of mutation and wild type. hPSCs can be tricky to break into a single cell suspension; they have a low threshold for triggering apoptosis or differentiating when stressed and it is difficult to be certain that all colonies on a plate grown from a single-cell dilution genuinely arise from a single cell, instead of 1-5 cells that stuck together during the resuspension or otherwise coalesced after plating.

Our accuracy rates for 1D sequencing with Albacore and Guppy are 85-85%, slightly lower than reported in previous publications[281]. This could be an artefact of amplicon sequencing featuring intronic regions. Intronic regions are highly variable and our targets may well contain genuine mutations that have sequenced correctly, but do not align to the reference sequence and therefore are included in these figures. Without running a proven DNA standard containing such intronic mutations as a comparison, it is not possible to be certain in this regard.

We were not able to identify any particular reproducible pattern for clustering of sequencing errors between different targets. There are, however, both caveats potential solutions for these issues.

To begin with the caveat – in each target there were certain sequences at which read errors were vastly over-represented; in CHD2 insertions were over-represented by a factor of 39.7 at AGAA tetranucleotide sites, in NRXN both insertions and deletions were over-represented at TT and TTT, in FADS2 insertions were over-represented at AGGT. The possible reasons for the over-representation of miscalls at these sites include; genuine intronic indel being detected by the sequencing platform and wider genomic context around these sites making them more difficult for the R9 pores to read.

It is likely that sequences such as this will be unique to each region amplified. Whatever the cause, the increased number of errors at these sites is sufficient that they could bias the overall accuracy.

Another caveat is that the sequencing artefacts look different to the mutations one may see with Cas9 gene editing[258, 288]. As per the graphs in *figure 3.17*, 1bp deletions and insertions are miscalled an order of magnitude more frequently than 2bp errors and 4 orders of magnitude more frequently than errors of 20bp.

This means that, although we may expect around 15% of sequences at any one locus to harbour miscalls of 1bp in length, the expected rate of miscalls for longer mutations is significantly lower. The variant screening algorithm has the option to set the minimum length of indel to include – and the minimum call count to set. In our experience, setting the minimum call to 3bp and the minimum read depth to report at 15% prevents errors from being called.

The other concern arising from the data is that there seems to be a relative difficulty in resolving the location of insertions using 1D nanopore data. From our test set of known GRIN2A mutations, the CRISPR_nanoscreen algorithm was able to count the reads containing insertions accurately, so the wells did not escape identification, however as can be seen in *figure 3.10*, the exact locus of the insertion determined by the aligner varied considerably compared to the locus of the deletion, which was relatively well resolved.

It could also be considered unusual that our other runs have thus far only identified deletions, however this could be an artefact of how the majority of the runs were conducted. For CACNA1C, FADS2, CHD8, GJAJ5, GJAJ8 and SETD1A, multiple gRNAs were used in order to

create an excision between two cut sites. There is evidence that single cut sites are more likely develop insertions than deletions[258]. In order to determine whether our pipeline is capable of detecting these, further experiments such as the CHD2 run, where only one gRNA was used, will be necessary.

3.4.2 Choice of experimental pipeline

All pipelines except for 1D called by Albacore and Guppy resulted in a significant loss of total number of reads aligned. Increased accuracy could be obtained by using Canu to error-correct reads (*figure 3.14*).

It was anticipated that as Guppy is the newest base-calling package, that either Guppy's Flip-Flop model, or Guppy 1DSQ basecalling would result in the greatest read accuracy. The data suggests that this is not the case.

This finding is in conflict with the published data about the utility of these base-callers. In particular, Ryan Wick (author of porechop, used in all pipelines) of Monash University in Melbourne keeps an up-to-date analysis of tools used for nanopore data available on his GitHub page[289]. This suggests that, based on their team's highly benchmarked standard dataset, Flip-Flop has an increased read-accuracy as compared to albacore and the standard guppy model.

A potential reason for these differences is the difference between the libraries used in our experiments and the benchmarking dataset used by Wick et al. Our libraries consist of amplicons amplified from *Homo sapiens* genomes, of 700-900bp length created by PCR (and therefore expected to be identical). The dataset used for benchmarking is a 5.5Mb *Klebsiella pneumoniae* chromosome, with reads ranging from 22-134kb in length.

Given that our data suggests differing accuracies for different mononucleotides, it is tempting to suggest that simply determining the nucleotide composition of the sequencing target may provide a potential estimate of expected sequencing fidelity. *Klebsiella* has a GC fraction of 57.5% as compared to our amplicons: CHD2 42.8%, NRXN1 68.7%, and FADS2 55.3%. However, a close reading of our results contradicts this supposition. FADS2 has a GC fraction close to that of *Klebsiella* but the accuracy of these alignments was no higher than the alignments for the other targets.

The published dataset was collected using the 9.4 (1D) nanopore, whereas ours was collected using the 9.5 (1DSQ) nanopore. As the capture process for 1D reads is identical in

both pores, it is difficult to imagine how this could have had an impact on our results but remains worth highlighting. Whatever the cause of the discrepancy, given the lack of consensus between our data and this published benchmark set, we must be explicitly aware of the limitations of our dataset.

It is also worth noting that our data do not and cannot answer questions about consensus assembly accuracy, as they only target a very narrow region of the genome. A broader experiment, targeting a larger genomic region (or even a whole genome) would be necessary to assess the relative utility of these pipelines for assembly accuracy.

As stated above, there was no repeating pattern of error-prone sequences demonstrated by our experiments. It is therefore likely that the structure of the strand to be amplified, rather than the nucleotides themselves may be responsible for the biases seen in the errors called by nanopore sequencing.

For example, a sequence with a 10bp dimer run containing five adenine dinucleotides (AAAAAAAAAA) would exhibit a greater number of miscalls than another sequence that had five adenine dinucleotides separated by varied sequences. It is therefore not the AA dinucleotide, nor the total adenine content of the target, but the wider context of the sequence that determines whether AA dinucleotides are more likely to be poorly sequenced.

It is worth noting that the target sequences used to generate our data were chosen for expediency, rather than for their representative nature of the human genome. It is fortunate coincidence that they have a range of GC contents at least somewhat representative of the genome as a whole. Repeating a similar analysis with a much wider range of more carefully curated amplicons would help to validate the results generated here.

Because of the short lengths of the reads, understanding the error bias of more complicated arrangements is not possible, for example, if runs of dinucleotides are flanked by dinucleotides of different composition, would this effect the error rate?

With slight modifications the analysis pipeline could be geared towards a genome-wide analysis – in theory there is no upper limit to the nt length of the sliding window used for results binning and properly trained neural network package such as tensorflow could be used to screen for patterns beyond simple strand composition. Although this work is beyond the scope of this thesis it is a potentially fertile field for further work.

3.4.3 An idealised Pipeline

For the common lab procedure of identifying cell lines containing mutants from Cas9 and other genome-editing experiments, the 1D pipeline has proven sufficient (see section 3.4.1). Based on our deeper analysis in this section, there is a significant possibility that the lower output of more accurate pipelines would result in a greater well-fallout and therefore a higher risk of missing mutations.

There is also the concern that running an error correction program such as *canu* on the pooled sequence data would render the library useless. First, it may attempt to correct the barcode sequences; second, if mutations are present at low levels (<15% of reads, for example), it may regard these as sequencing inaccuracies and correct them. This risk is likely to be particularly significant for shorter indels generated by NHEJ repair of a single cut-site.

An idealised approach based on our data, would be to use the pipeline described in *figure 3.18* to generate 1D reads from Guppy *and* 1DSQ reads from Albacore. These base-called reads can then be pooled and demultiplexed. The variant screen can be used to identify wells with possible mutations. As every read pertaining to this well should, in theory, contain the mutation, the demultiplexed reads for wells of interest can be safely corrected with *canu* in order to obtain a more reliable description of the mutation present in that well. This approach may not be as useful for describing mutations in wells with multiple genotype as again, there is a risk of *Canu* overcorrecting mutations present at lower frequencies.

Running two separate basecalling pipelines is, however, time consuming and a more realistic approach may be simply to use 1D reads from Guppy, then perform the post-demultiplexing correction. Given that 1D library preparation is around 45 minutes shorter than 1DSQ library preparation this also saves on bench time.

These results also inform the interpretation of the data generated in chapter 5, in which the system described here is used to investigate abnormalities in DNA repair. The ways in which these findings inform our understanding of that data is better addressed in context of that data, and so will be considered in the discussion section of chapter 5.

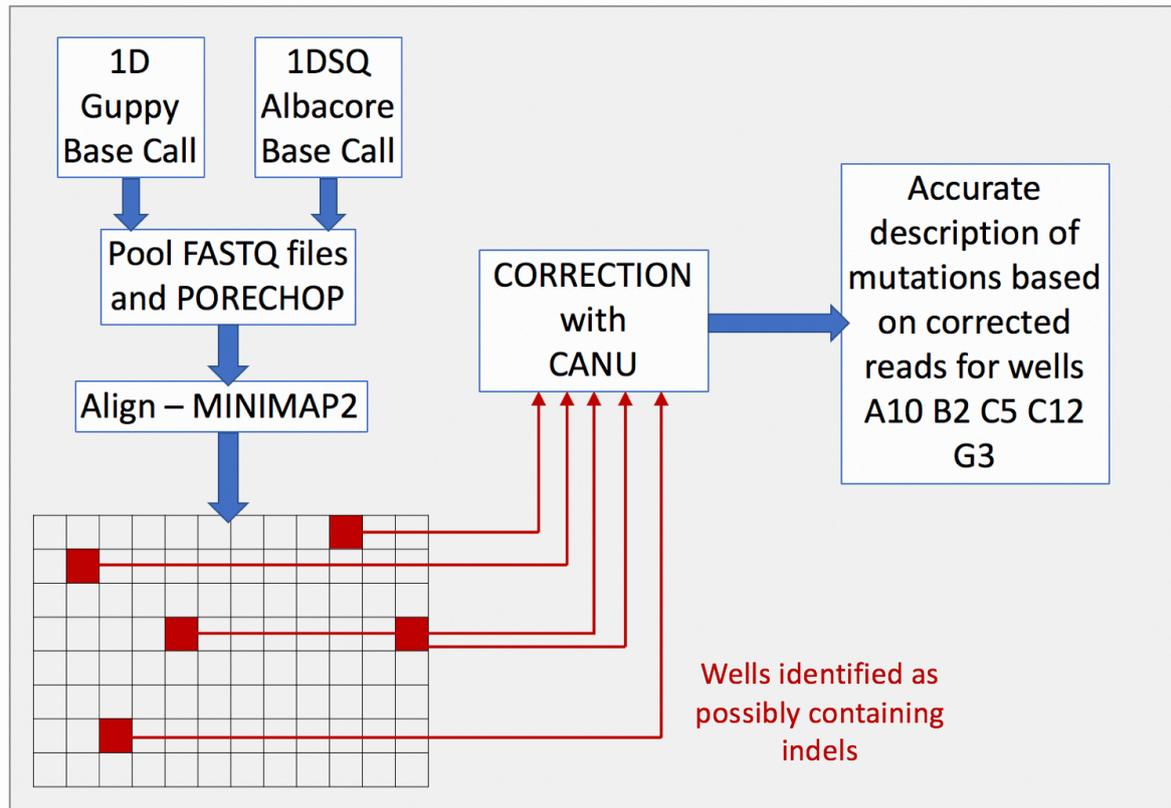


figure 3.18 – An idealised bioinformatic pipeline for identification of mutant lines from gene editing experiments analysed by nanopore sequencing, based on analysis of multiple pipelines for amplicons <1kb length

4: DEVELOPMENT OF A CHD2 MUTANT CELL LINE AND CHARACTERISATION DURING DIFFERENTIATION INTO NEURONS USING RNA-SEQ

4.1 INTRODUCTION

In the previous chapter, I described the set-up and testing of a cell line capable of creating targeted double strand breaks using a doxycycline inducible Cas9 expression and lipofection of gRNAs against the desired target. In this chapter, I describe the creation of a CHD2 mutant line using the i-Cas9 system. This cell line is necessary to investigate DSB repair in CHD2 mutant lines – the stated aim of this thesis.

I also characterise this cell line by whole transcriptome sequencing (RNA-Seq) at different stages of induced neurodifferentiation in culture.

4.1.1 Criteria for cell line

The aim of this project is to create a cell model of CHD2-related EECO. To date all known cases of CHD2 related neurodevelopmental disorders have occurred in patients with heterozygous mutations. A significant number of these patients harbour null mutations such as: nonsense mutations, copy number loss, or indels that shift the reading frame of the gene (see also section 1.2). There are also patients who harbour missense mutations in CHD2 that change the coding sequence but still produce a protein. As described in our introductory section these patients can be affected with an indistinguishable phenotype.

This spectrum of mutations in patients suggests that one likely disease mechanism of CHD2-related EECO is haploinsufficiency. Furthermore, population data analyses found in the exome aggregation consortium (ExAC) database suggest that CHD2 is intolerant of loss-of-function variation. Therefore creating a non-specific knockout (frameshift or premature stop codon) is a sufficient approach for modelling this disease. This is a more straightforward procedure than trying to directly replicate a missense variant with proven pathogenicity. The decision was taken to aim for a frameshift indel mutation, due to the relative ease of achieving this with the iCas9 cell line described in *chapter 3*.

Our evidence from chapter 3 suggests that transfecting multiple gRNA simultaneously increases the potential rate of mutations (section 3.3.3), however as this cell line was intended for further study, it was decided to use single gRNAs for each transfection, thereby reducing the number of potential off-target hits.

4.1.2 Characterising neurodifferentiation and an introduction to whole transcriptome sequencing

This project aims to investigate the biological process of DSB repair at different time-points during induced neurodifferentiation. In order to determine if these results are valid and aid in their interpretation it is necessary to determine that the mutant cells are differentiating in a similar manner to the WT cell line.

Several approaches to cell characterisation are available: simple morphological assessment with light microscopy, electrophysiological assessment of function for mature neurons, protein-based assays such as immunohistochemistry and western blot, and transcription analysis using RNA-Seq.

Light microscopy provides a useful benchtop diagnostic of the progress of cell cultures – certainly cultures that look like they are not behaving in the expected manner can be spotted by experienced investigators, however it does not provide the detail necessary for a full characterisation of differentiation.

Electrophysiological experiments are desirable as they give a true picture of culture activity[290]. There is evidence that electrical activity is not just a feature of neuronal maturation, but that it is necessary for maturation to take place [291]. Electrical stimulation may play a significant part in triggering plasticity cascades that cause morphological changes, such as neurite outgrowth[292] in fully committed cells of various neurological lineages. The decision not to use these experiments in this project reflects the focus of the project on intracellular signalling pathways for damage repair, rather than changes in whole culture activity and cell migration.

Protein based assays are perhaps the most well-established method of determining cell lineage and are used throughout biological research and clinical practice. They are of particular clinical utility in the field of oncology, where the expression of certain cell surface markers can determine not only prognosis, but help to decide the most appropriate course of chemotherapeutic treatment[293].

The widespread use of protein-based assays has led to the development of a well characterised set of markers, against which antibodies are commercially available. Multiple targets are used in these experiments. The ones explored in more detail in this thesis are: NANOG and OCT4 which are used to confirm pluripotency, PAX6 and SOX2 which are used to identify NPCs, and MAP2 and PSD95 which are used to identify mature neurons.

4.1: Introduction

OCT4 (Octamer-binding transcription factor 4) has been demonstrated to form part of the cocktail of 'Yamanaka' factors (named for their discoverer) necessary for the induction of induced pluripotent stem cells from fibroblasts[294]. NANOG - named for 'Tir Na Nog', the mythologic Celtic land of the ever-young - has been demonstrated as a downstream target of the Yamanaka factors[295] and is necessary for the maintenance of pluripotency [229].

PAX6 (paired box gene 6) is one of the earliest detectable proteins known to promote neurogenesis[296]. Sox2 (SRY-box 2) is present in hiPSCs and NPCs with levels expected to fall rapidly as neurons mature[297].

MAP2 (microtubule assembly protein 2) is crucial for neuronal cytoskeleton function and is precociously expressed in mature neurons[298]. PSD95 (post-synaptic density protein 95, more recently renamed to DLG4) has been demonstrated as crucial for the maturation of synapses in mature neurons[299].

The mRNA coding for these proteins will also be upregulated and downregulated accordingly at the relevant stages of neurodifferentiation and so in more recent years, with the advent of affordable high-throughput sequencing, RNA-Seq has been used for characterisation of cell lines[300].

RNA-Seq provides a direct and unbiased assessment of all expressed loci within a cell or cell population. The relative abundance of each transcript in the human transcriptome can be described as a comparison between two cell populations in order to assess the transcriptional impact of a mutation or environmental change such as: treatment with drug, small molecule, or a change of culture medium [301].

An advantage to RNA-Seq is that if irregularities are identified in comparison with the expected pattern of transcription, the dataset for the remaining transcriptome is already available for analysis. To put it another way; with RNA-Seq the same dataset can be used to determine if the mutant cell lines are differentiating correctly *and* to begin investigating changes in the phenotype relating to the mutation.

One of the myriad ways in which CHD2 mutations could affect their phenotype is by the dysregulation of transcription[1]. The binding of CHD2 at sites of active transcription suggests that it has a role in maintaining chromatin state at these sites and it is known that chromatin state correlates with transcriptional activity. It is also possible that dysregulation of CHD2's involvement in specific biological pathways (such as DSB repair and myogenesis) could impact the transcriptome in unpredictable ways. Within the arcane chemical soup that

4.1: Introduction

forms the chromatin microenvironment there are likely to be feedback mechanisms that respond to the interruption, prolongation and completion of any biological process.

As described above, RNA-Seq provides an abundance of data and not all of it will be directly relevant to the questions being asked. Within the vast datasets generated it is possible to query changes in specific genes which will be used as an assessment of neurodifferentiation.

In order to understand wider perturbations in the transcriptome related to change in culture conditions or gene knockout, a wider lens is required. Gene ontology (GO) is the formalised systematic representation of our current knowledge regarding the functions of each protein-coding and non-coding RNA transcript in the human genome[302]. The ontology relates to three aspects; molecular function, cellular component and biological process. The GO terms themselves are an area of active research and are maintained by the Gene Ontology Consortium[303, 304].

In our analysis we will compare WT and mutant cells at time points in differentiation to determine if any deregulated genes cluster with GOs that may be of interest regarding the pathogenesis of EECO in patients with *CHD2* mutations.

4.1.3 Aims

4.1.3.1 Primary aims

- a) Develop a iPSC-iCas9 line containing a loss of function mutation in *CHD2*
- b) Assess the differentiation markers at D0, D19 and D40 of differentiation into neurons in WT-iCas9 cells and *CHD2*-mutated-iCas9 cells

4.1.3.2 Secondary aims

- a) Compare the transcriptomes of WT-iCas9 cells and *CHD2*-mutated iCas9 cells to investigate patterns of gene expression which could explain the neurodevelopmental abnormalities exhibited in patients with heterozygous *CHD2* mutations

4.2 METHODS

4.2.1 Set up of CHD2 mutant cell line

In order to generate a cell line best suited for downstream analysis, prevention of off-target events was prioritised. Although use of multiplexed gRNA targeting the same exon can increase the chances of generating the desired knock-out mutation, an increase in the number of guide sequences also increases the risk of off-target effects. Three gRNAs were designed for use in non-multiplexed experiments, targeting exons 2, 3 and 4 of the *CHD2* gene (table 4.1).

gRNA name	exon	start	end	Predicted cut site	Strand
CHD2_g556	2	556	576	560	-
CHD2_g653	3	653	673	657	-
CHD2_g908	4	908	928	912	+

Table 4.1: cut site co-ordinates for gRNA used in CHD2 editing experiments. Co-ordinates are based on the canonical transcript NM_001271.4

The gRNA was transfected into WT-iCas9 cells using the protocol described in chapter 3. Cells were picked with the aim of generating 96 colonies per gRNA. DNA was extracted, amplified by PCR and barcoded using the protocol in section 2.5..

Each target was analysed on separate minION runs, with an aim of 500,000 reads per target. Data was analysed using CRISPR_nanoscreen (described in chapter 3) and wells likely to have mutation inspected visually in igv.

Where necessary, sub-cloning of cells with a mixture of different mutations was performed using the same protocol used for single cell plating in chapter 3. This process was repeated to aim for a clone of >90% purity for further investigation.

A short nanopore run on a 1DSQ flow cell, generating 100,000 reads, with canu correction of reads prior to alignment to the reference sequence (see chapter 3) to generate was used in order to confirm the mutation.

The effects of the mutation on the protein structure were investigated using mutalyzer [305], an online tool used to check mutation co-ordinates and generate human genetic variant society (HGVS) standardised nomenclature for the effect of the mutation on gene and protein. For any in-frame deletions, in-silico analysis to predict the impact on protein function was performed using PROVEAN [306] and SIFT-indel [307].

4.2.2 Cell culture and sample collection

Cells were differentiated into neurons using the protocol described in *chapter 2*. At D0, D19 and D40, protein and RNA extractions were performed in triplicate. Cells were also fixed in a 96 well plate for analysis with INDUCE-seq, as described in *chapter 6*. For all experiments, three separate wells were extracted to generate three biological replicates.

Extractions used for RNA-Seq, western blot and INDUCE-Seq were all taken at the same time-points from differentiations using the same starter cultures. This allows for cross-referencing of the results of these experiments during analysis.

4.2.3 RNA-Seq Library Preparation

Library preparation was conducted by Joanne Morgan of the core sequencing team in line with the protocol outlined in section 2.5.5.2. Three biological replicates were tested for each sample. It is understood that the higher the number of biological replicates, the lower the likely false discovery rate (FDR) of differently expressed genes (DEG) will be.

4.2.5 RNA-Seq Analyses to be performed

For all analyses detailed below, the comparisons to be made are found in *table 4.2*. The first three experiments will be the main focus of this chapter. The changes between each stage of development are relevant to the data presented in *chapter 6* where relevant the results of these comparisons and will be addressed there.

	Experiment A	Experiment B
hIPSC cells	D0 WT	D0 CHD2
Neuronal progenitor cells (NPC)	D19 WT	D19 CHD2
Mature neurons (MN)	D40 WT	D40 CHD2
WT hIPSC>NPC	D0 WT	D19 WT
WT hIPSC>MN	D0 WT	D40 WT
WT NPC>MN	D19 WT	D40 WT
CHD2 hIPSC>NPC	D0 CHD2	D19 CHD2
CHD2 hIPSC>MN	D0 CHD2	D40 CHD2
CHD2 NPC>MN	D19 CHD2	D40 CHD2

Table 4.2 – list of comparisons made using RNA-Seq data

4.2.5.1 Whole dataset analysis

Sample to sample distance will be used to provide the first estimate of similarity of difference between WT and CHD2^{+/-} cells at each stage in differentiation.

In determining significant read count differences between each transcript, our exploratory analysis (data not shown) determined that for these experiments LFC >0.5 and < 0.5, with p-value of < 0.05 would provide the most useful analysis. MA plots are used to visualise the spread of the data. Lists of significantly differently expressed genes are extracted using the same criteria.

Gene lists where there is a significant difference in the LFC are analysed using GOrilla and REVIGO [346] gene ontology toolkits in order to consider which developmental programs are differently regulated between the CHD2 mutant and WT cell lines. Visualisation of these differences will be provided by Cytoscape [347].

4.2.5.2 Plot counts for neurodifferentiation markers

In order to determine whether our cell lines are at the appropriate stage of differentiation, transcript levels of genes that code for proteins markers of each stage of neurodifferentiation were compared.

As well as the markers for hPSC, NPC and mature neuron classification described in the introduction[124, 125, 295-299, 308, 309], markers for intermediate progenitors and immature neurons[296, 310, 311] were also compared. A full list of the genes to be compared can be found in *table 4.3*.

IPS markers	NPC markers	Intermediate Progenitors	Immature Neurons	Mature Neurons
NANOG	PAX6	TBR2	TBR1	MAP2
OCT4	SOX2	ASCL1	TUBB3	PSD95
SOX2				

Table 4.3 Transcripts used as markers for different stages of neuronal differentiation

4.2.5.3 Differential expression lists and gene ontology

Using DESeq2, the differential expression list can be filtered by p value, to provide a list of genes with statistically significant differential expressions between any two samples of groups. An initial filtering that groups genes with a $p=0.1$ into those with $LFC < 0$ and $LFC > 0$ is the first stage recommended in many guidelines for RNA-Seq, however these thresholds can be changed in order to curate lists of more reasonable lengths for analysis. Our exploratory analysis (data not shown) determined that for these experiments $LFC > 0.5$ and < -0.5 , with p-value of < 0.05 provided a more useful analysis.

This list of genes will be analysed using the online gene ontology annotation toolkit GOrilla [226] to try and determine which functions and pathways are impacted by the change in differential expression between WT and CHD2 mutant cells at each stage of differentiation.

The output of ranked ontology terms from GOrilla is transferred to REVIGO [227] – a web-based suite of tools for visualisation of ontology and pathway terms. From here, datasets are exported for use in Cytoscape [228], providing interactive visualisation of ontology terms in a map format.

4.3 RESULTS

4.3.1 Confirmation of CHD2 mutation

4.3.1.1 Sequencing results

40 viable colonies were chosen for DNA extraction using PCR primers to generate an amplicon of 816bp in length, including M13F/R tags and barcodes. The minION flow cell had 440 pores available at the start of sequencing. The minION was run for 15 hours overnight due to the low number of pores and generated 318,280 reads.

Reads were decomplexed successfully with a minimum threshold of 75% and differential of 3% and aligned to the 816bp reference sequence. CRISPR_nanoscreen identified potential deletions in over 20% of reads in 10 (25%) of the wells. On visualisation of the data, the mutations were a mixture of an 11bp deletion and a 6bp deletion, with smaller numbers of other mutations represented.

After two rounds of sub-cloning, a colony 50% of the reads contained an in-frame 6bp deletion and 50% were the 11bp frameshift was identified, leading to the conclusion that this cell line was compound heterozygous(*figure 4.1*).

A final sequencing run of 60,542 reads was performed using 1DSQ workflow. After basecalling with albacore and sequence correction with canu, 41,424 reads were aligned to the reference genome. 20,997 (50.6%) reads contained the 6bp deletion, 20,427 reads (49.3%) contained the 11bp deletion (*figure 4.2*).

4.3: Results

4.3.1.2 Characterisation of mutations

The experiment generated a compound heterozygous mutation in exon 3 of CHD2. The full HGVS description of these mutations are

1) 11bp deletion at position 649:

n.649_659del relative to transcript

c.76_86del relative to translation start site

p.(Glu26Valfs*64)

This frame-shift mutation is predicted to introduce a premature stop codon at residue 64 and is therefore expected to function as a null mutation.

2) 6bp deletion at position 649:

n.649_654del relative to transcript

c.76_81del relative to translation start site

p.(Glu26_Glu27del)

This is an in-frame deletion of two glutamine residues. *In silico* analysis with Provean predicts that the variant will be deleterious to protein function (score -4.02, cutoff -2.5). *In silico* analysis with SIFT-indel predicts the variant is neutral, with a confidence score of 0.91.

As pathogenicity of the second mutation cannot be confirmed, for the purposes of this thesis, this cell line will be referred to as iCn-CHD2^{+/-}. The wild type cell line will be referred to as iCn-WT.

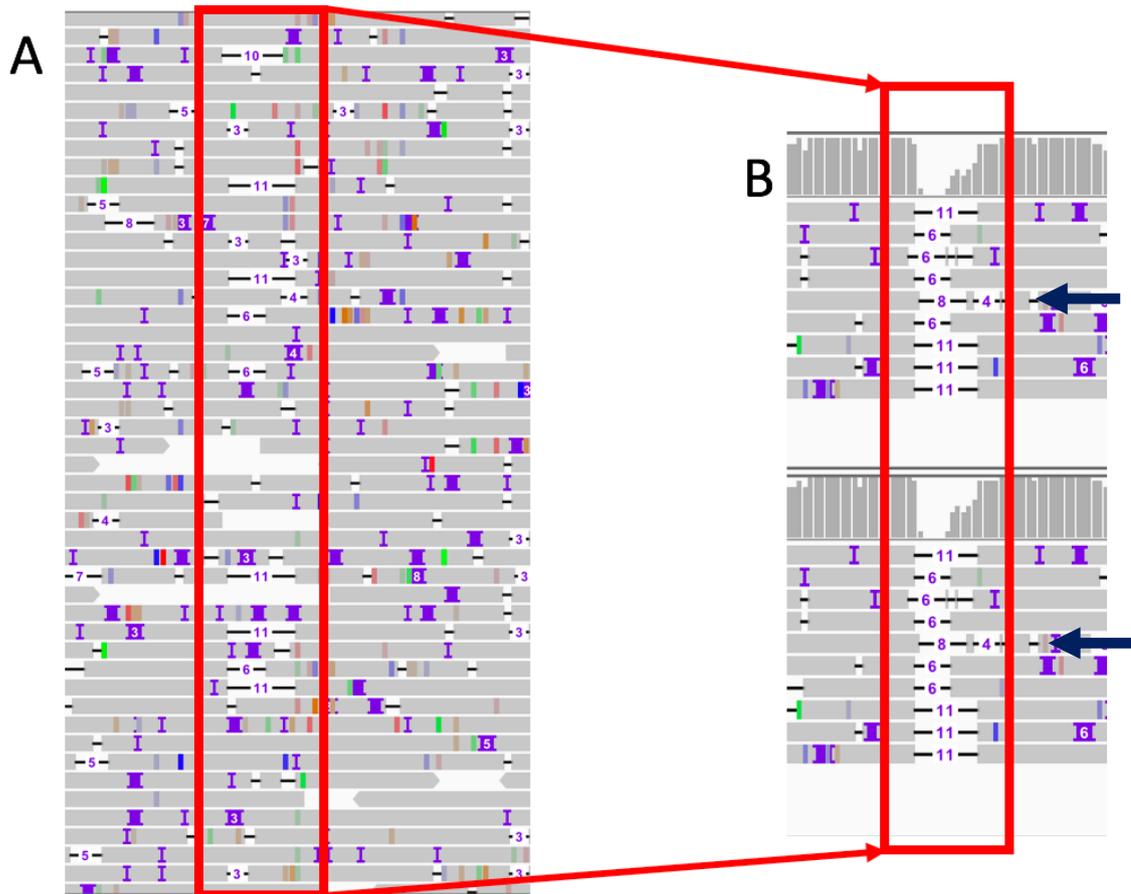


Figure 4.1; Demonstration of CRISPR-induced CHD2 mutations from first sequencing run
A – well D5 taken from first run of CHD2 CRISPR experiment, demonstrating mixture of iCn-WT, 6bp del and 11bp del mutations, B – successfully sub-cloned cell line displaying wells C3 and D1 of the third sequencing run. These wells demonstrate the compound heterozygous mutation described in 4.3.1.1

The arrow highlights a read containing an 8bp and a 4bp deletion. This read disappeared on canu correction and is felt to be an artefact of nanopore sequencing

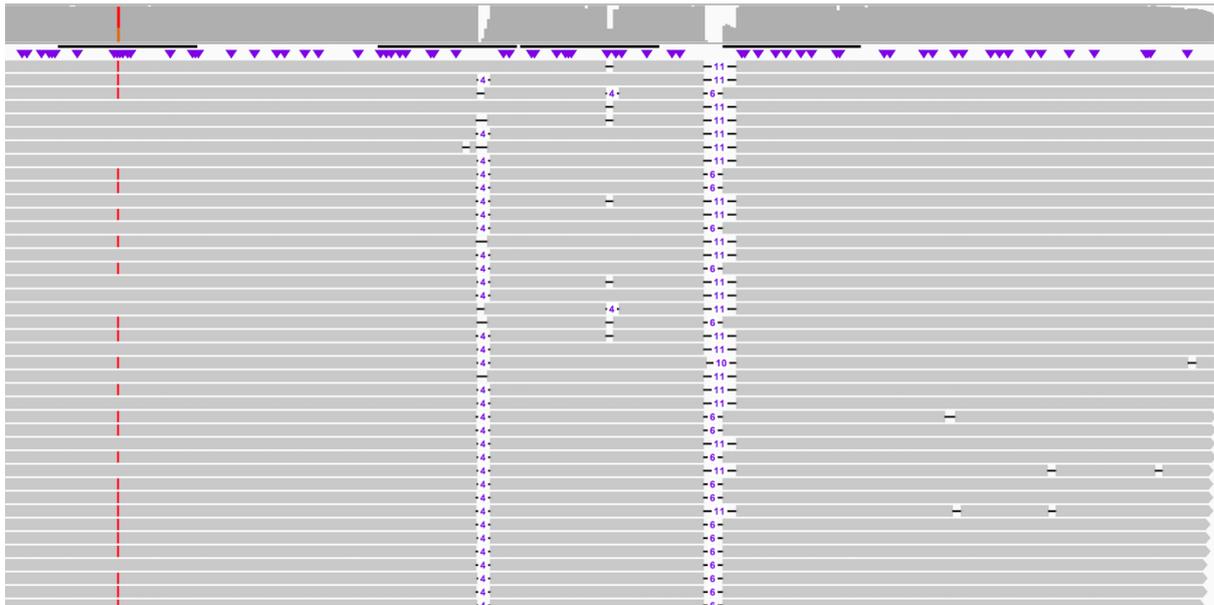


Figure 4.2 – CHD2 heterozygous mutation in sub-cloned cell line. Data displayed is from 1DSQ sequencing after canu correction, displayed in igv. Note resolution of 11bp and 6pb mutation in 1:1 ratio. Other calls are in intronic region and either represent intronic variation or miscalling in long homopolymer region

A homozygous 4bp mutation is seen downstream on the cutsite – this mutation is present in intronic material and does not affect the coding sequence, nor is it predicted to interfere with splicing. It is likely that this represents a benign intronic variant present in our cell line.

4.3.2 RNASeq QC

QC performed both before and after trimming was satisfactory, except for extraction number 7. The number of reads for each was between 19,000,000 and 27,000,000, except for extraction 7 – technical replicates 1 and 2 had 30,000 reads and 26,000 reads respectively.

The mean base Q score was within acceptable limits for all samples, except for the first technical replicate of sample 7. Mean GC content was 43-45% across all samples – again sample 7 was the exception with a mean GC of 36% and 33% for the first and second replicates.

Each sample had <2% total over-represented sequences. In every case they were long poly-A, poly-T or poly-C sequences. Yet again, sample 7 was the exception, with >10% reads being composed of poly-sequences.

The bamtools outputs counted between 38,000,000 and 48,000,000 mapped reads for all of the samples which passed initial QC (*figure 4.3, table 4.4*). The duplicates in each dataset were between 14 and 24% of the mapped reads, which was deemed to be acceptable.

The decision was taken to take forward all datasets except for extraction 7 for further analysis.

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

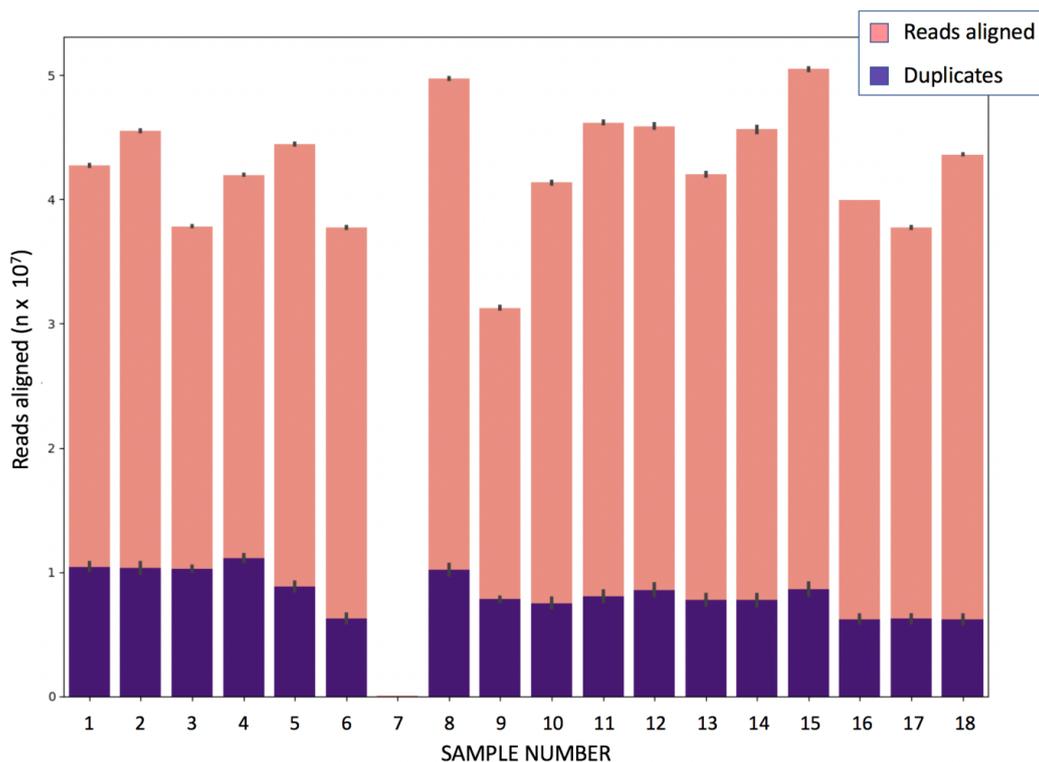


Figure 4.3: total reads aligned from each sample number during RNA-Seq and total duplicate reads detected by bamtools whilst marking duplicates

Sample number	Condition	Abbreviation
1	iCn-WT hiPSC biological replicate 1	WT-IPS-1
2	iCn-WT hiPSC biological replicate 2	WT-IPS-2
3	iCn-WT hiPSC biological replicate 3	WT-IPS-3
3	iCn-CHD2 ^{+/-} hiPSC biological replicate 1	CHD2-IPS-1
5	iCn-CHD2 ^{+/-} hiPSC biological replicate 2	CHD2-IPS-2
6	iCn-CHD2 ^{+/-} hiPSC biological replicate 3	CHD2-IPS-3
7	iCn-WT NPC biological replicate 1	WT-NPC-1
8	iCn-WT NPC biological replicate 2	WT-NPC-2
9	iCn-WT NPC biological replicate 3	WT-NPC-3
10	iCn-CHD2 ^{+/-} NPC biological replicate 1	CHD2-NPC-1
11	iCn-CHD2 ^{+/-} NPC biological replicate 2	CHD2-NPC-2
12	iCn-CHD2 ^{+/-} NPC biological replicate 3	CHD2-NPC-3
13	iCn-WT MN biological replicate 1	WT-N-1
14	iCn-WT MN biological replicate 2	WT-N-2
15	iCn-WT MN biological replicate 3	WT-N-3
16	iCn-CHD2 ^{+/-} MN biological replicate 1	CHD2-N-1
17	iCn-CHD2 ^{+/-} MN biological replicate 2	CHD2-N-2
18	iCn-CHD2 ^{+/-} MN biological replicate 3	CHD2-N-3

Table 4.4 – sample number and condition for RNA extractions

4.3.3 RNA-Seq sample clustering analyses

The transcriptomes at each stage of differentiation clustered together, regardless of whether the samples were taken from the iCn-CHD2^{+/-} line or the iCn-WT line.

Within these differentiation-stage clusters there was appreciable distance between the iCn-CHD2^{+/-} cell lines at iCn-WT lines in D19 (NPCs) and at D40 (neurons) with a smaller distance between samples at D0 (IPS).

The first analysis included all 36 samples, including all biological replicates and technical replicates. From this visual analysis, it was determined that extraction 6 (iCn-CHD2^{+/-} cells at D0, biological replicate 3) and extraction 7 (iCn-WT D19, biological replicate 1) were unusual.

Sample 7 had a significantly lower read-count than the other samples and QC had flagged potential difficulties with an unexpectedly high representation of long A and T repeats sequenced. Sample 6 had passed QC but, despite being an extract from D0 cells, bore more overall resemblance to the D19 and D40 cohorts, leading to concerns about possible sample contamination.

Based on these concerns, the decision was taken to remove the datasets derived from samples 6 and 7 datasets from further analyses – the relative impact of reducing the number of biological replicates will be addressed in the discussion section. The sample distance heatmaps and principal component analyses (PCA) conducted both before and after this sample was removed can be found in *figures 4.4 and 4.5* respectively.

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

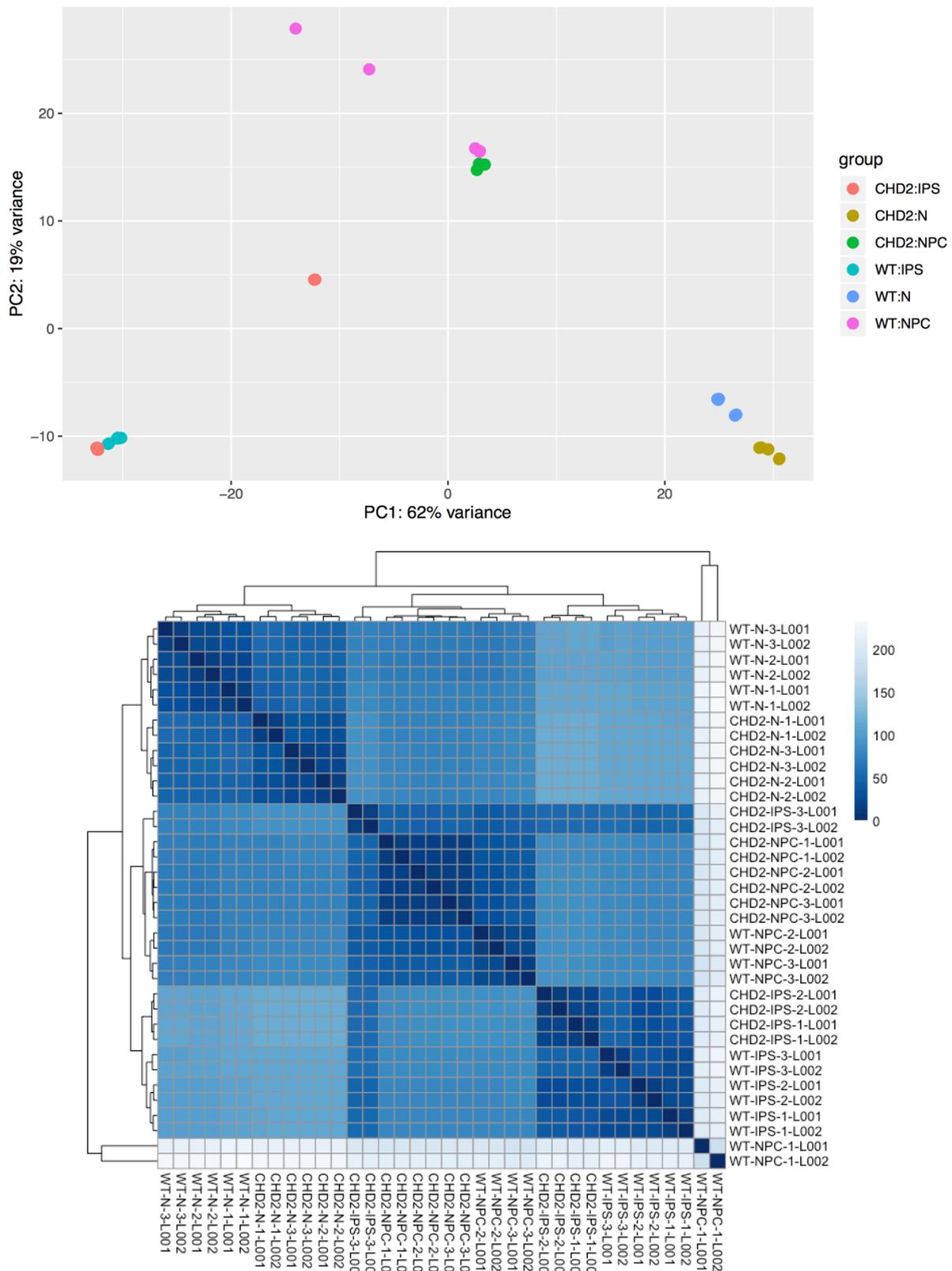


Figure 4.4 – PCA and heatmap of sample distance containing all 36 RNA-seq datasets with both technical replicates of each run (L001 and L002 represent replicates of the same sample)– note the abnormal results for iCn-WT-NPC-1 and CHD2-IPS-3 samples

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

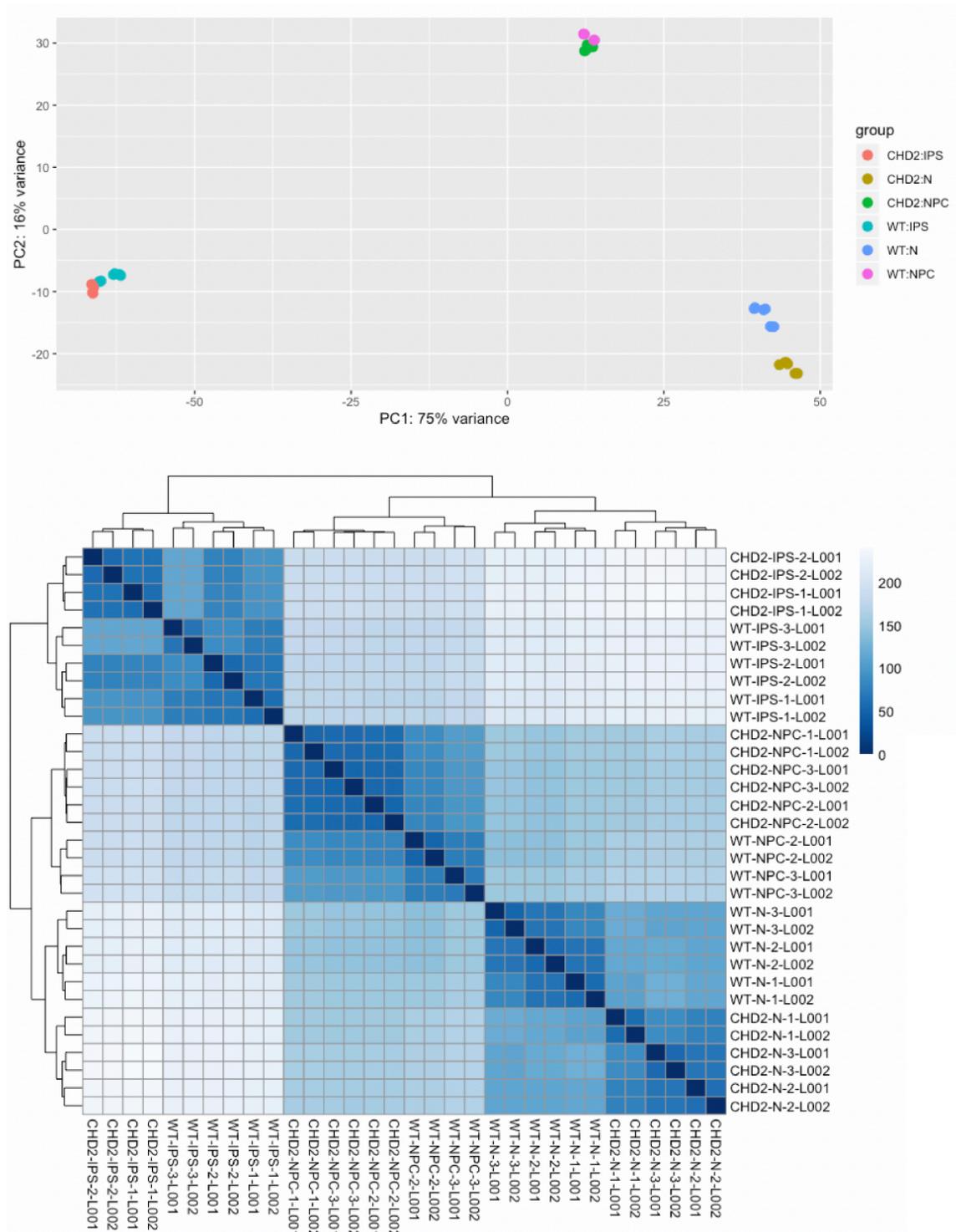


Figure 4.5 – PCA and heatmap of sample distance containing datasets with both technical replicates of each run (L001 and L002 represent replicates of the same sample

Data presented with removal of outlier (iCn-WT-NPC-1 and CHD2-IPS-3) results (see Figure 4.5)

4.3.4 Plot counts for neurodifferentiation markers

At D0, the IPS markers SOX2 and OCT4 are raised in both cell lines, however NANOG is not. There is a statistically significant difference between the iCn-WT cell line and the iCn-CHD2^{+/-} cell line for OCT4 read-counts ($p=0.003$) and for SOX2 ($p=0.012$). In both cases, the normalised readcount was higher in the iCn-WT cells. (*table 4.5 , figures 4.6 & 4.7*)

OCT4 and NANOG transcription both fall between D0 and D19, however SOX2 remains highly transcribed – this is the expected pattern: indeed, SOX2 antibodies are often used as a marker for NPC immunohistochemistry and western blotting.

At D19, the NPC markers PAX6 and SOX2 were elevated in both iCn-WT and iCn-CHD2^{+/-} cell lines. The PAX6 readcount was significantly higher in iCn-WT cells ($p=0.00012$), however there was no significant difference in the SOX2 count ($p=0.68$).

MAP2, a marker of neuronal maturity, was upregulated in both cell lines at D19. Again, there was a statistically significant difference, with a greater number of read counts aligned from the iCn-WT cell line ($p=0.002$) (*table 4.5 , figures 4.6 & 4.7*).

At day 40, the neuronal marker MAP2 was highly transcribed in both cultures. The expression was significantly higher in the iCn-CHD2^{+/-} cell line ($p=0.00036$). PSD95, another marker of neuronal maturity, was upregulated in both cells, to a lower extent than MAP2. Again, there was a statistically significant difference with greater transcription in iCn-CHD2^{+/-} cells ($p=0.005$)

PAX6 remained highly transcribed in the iCn-WT cells at D40, with minimal transcription evident in the CHD2 cells. The difference in PAX6 transcription between cell lines was highly significant ($p=1.42 \times 10^{-9}$).

SOX2 transcription was higher at D40 in the CHD2 cells than at D19. Although SOX2 transcription was still evident in the iCn-WT cells, the levels had begun to fall. Again, there was a high level of statistical significance between transcription levels ($p=5.3 \times 10^{-8}$) (*table 4.5*)

The intermediate progenitor markers TBR2 and ASCL1 were modestly raised at D19 and D40 in both cell lines. ASCL1 was higher at D40 than D19, with the reverse true for TBR2. At D40, there was a significant difference in the transcription levels of both with ASCL1 higher in CHD2 ($p=4.85 \times 10^{-8}$) and TBR2 higher in iCn-WT ($p=2.57 \times 10^{-12}$).

4.3: Results

The immature neuronal markers, TBR1 and TUBB3 were raised at D19 and D40. TBR1 was higher in D40 than D19 in iCn-WT cells, with the opposite pattern seen in iCn-CHD2^{+/-} cells. There was a statistically significant difference at D40 ($p=2.94 \times 10^{-7}$). TUBB3 was higher in D40 than D19 in both cell lines, with no statistically significant difference at D40 ($p=0.07$).

In the case of intermediate progenitor and immature neuronal markers, the transcription levels were considerably lower than the transcription levels for NPCs at D19 and mature neuronal markers at D40.

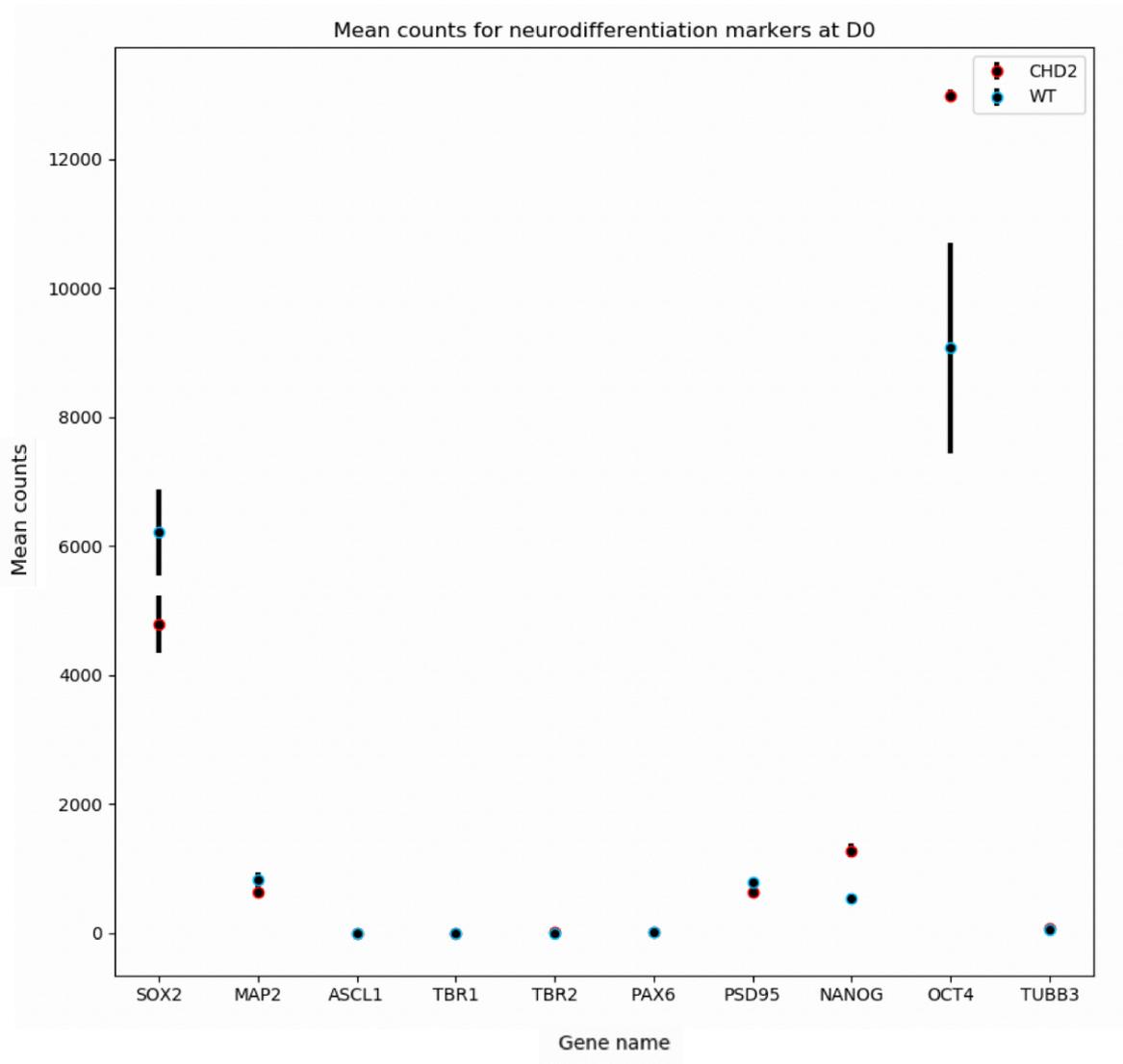


Figure 4.6.1: mean and sd for normalised readcounts of differentiation markers at D0 of neurodifferentiation demonstrating high levels of transcription of OCT4 and SOX2

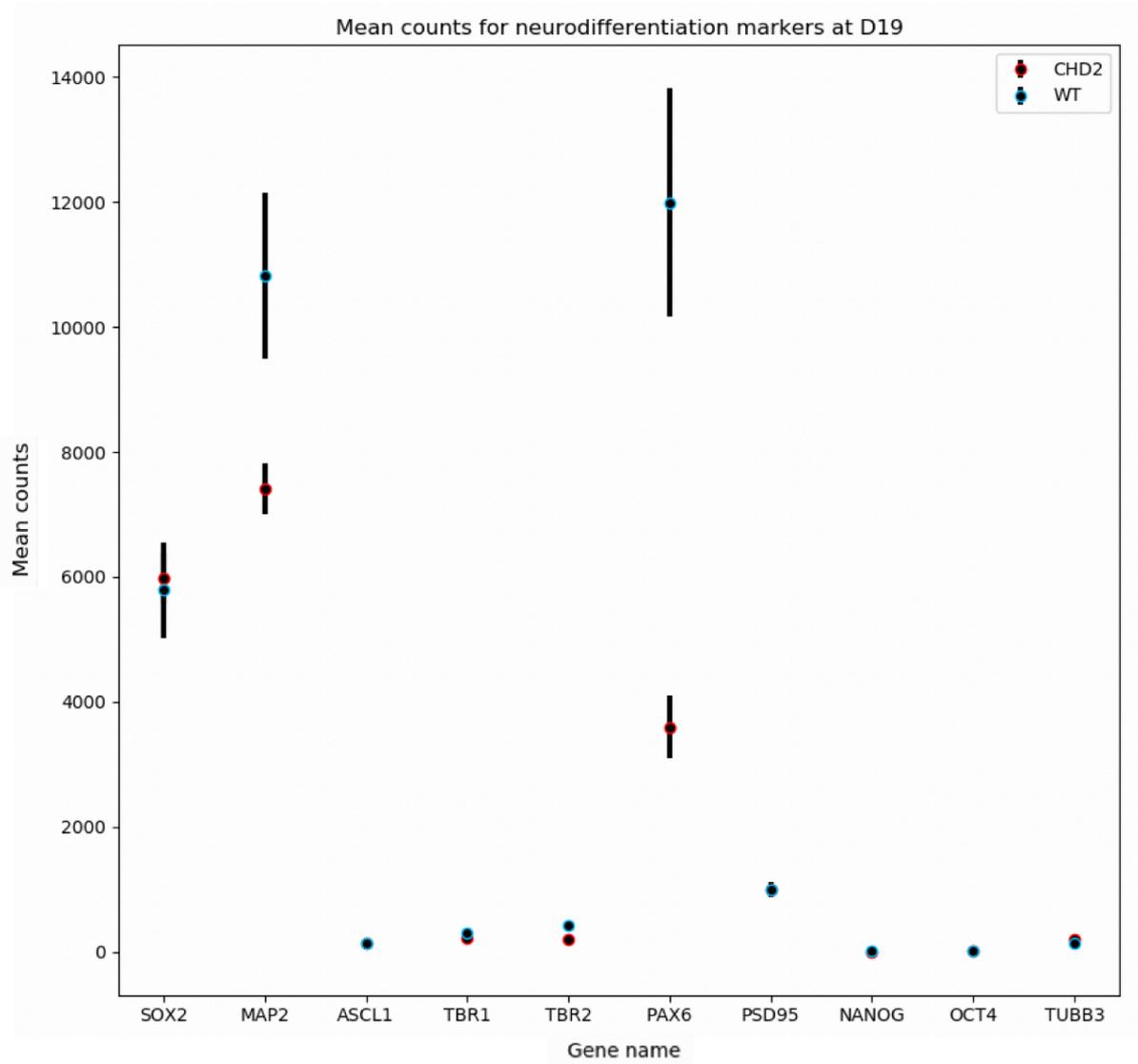


Figure 4.6.2: mean and sd for normalised readcounts of differentiation markers at D19 of neurodifferentiation demonstrating high levels of transcription of PAX6, MAP2 and SOX2

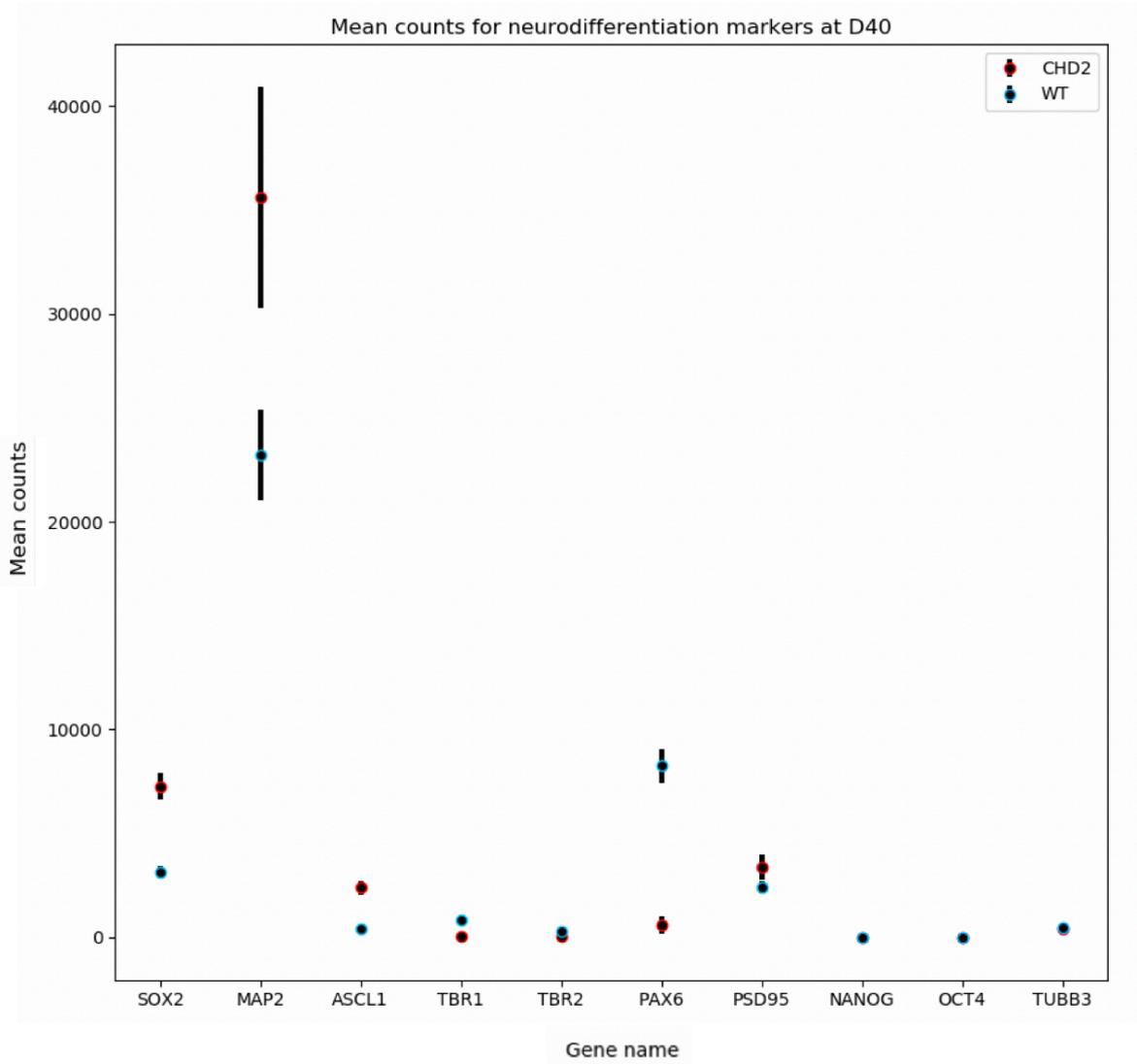


Figure 4.6.3: mean and sd for normalised readcounts of differentiation markers at D40 of neurodifferentiation demonstrating high levels of transcription of MAP2 and increased levels of PSD95

A – D0 of neurodifferentiation			
Gene	Mean normalised readcount in iCn-WT cell: mean (sd)	Normalised readcount in iCn-CHD2 ^{+/-} cells (mean, sd)	P value for comparison between readcounts
NANOG	545.3(68.9)	1278.8 (101.4)	2.06 x10 ⁻⁵
OCT4	9071.5(1633.6)	12976.4(99.4)	0.0031
SOX2	6207.6 (664)	4781.9(444.7)	0.011
PAX6	12.6(7.3)	13.8(5.2)	0.80
TBR2	4.5(2)	6.8(2.1)	0.17
ASCL1	5.5(2)	2(1.8)	0.04
TBR1	1.5(0.9)	1.2(0.5)	0.60
TUBB3	46(10.7)	64.7(19.4)	0.14
MAP2	828.9(117.7)	624.9(50.9)	0.019
PSD95	785.9(32.5)	630.5(21.3)	0.0002

Table 4.5 (page 1/2): mean normalised readcounts for expression of neurodifferentiation markers at: A-D0, B-D19 and C-D40 of neurodifferentiation

B – D19 of neurodifferentiation			
Gene	Normalised readcount in iCn-WT cells (mean, sd)	Normalised readcount in iCn-CHD2 ^{+/-} cells (mean, sd)	P value for comparison between readcounts
NANOG	5(6.9)	1.6(0.4)	0.36
OCT4	9.1(2.1)	6.5(2.4)	0.16
SOX2	10819.4(1334.1)	7403.8(410.8)	0.68
PAX6	11991.4(1833.2)	3597.1(506.6)	0.00012
TBR2	420.3(56)	192.5(20)	0.0002
ASCL1	5785.5(773)	5974.7(413.2)	0.45
TBR1	307.7(47.6)	213.7(19.5)	0.02
TUBB3	140.1(31.9)	196.5(17.7)	0.021
MAP2	130.2(11.3)	141.6(26)	0.01
PSD95	988.7(120.6)	997.1(32.5)	0.002

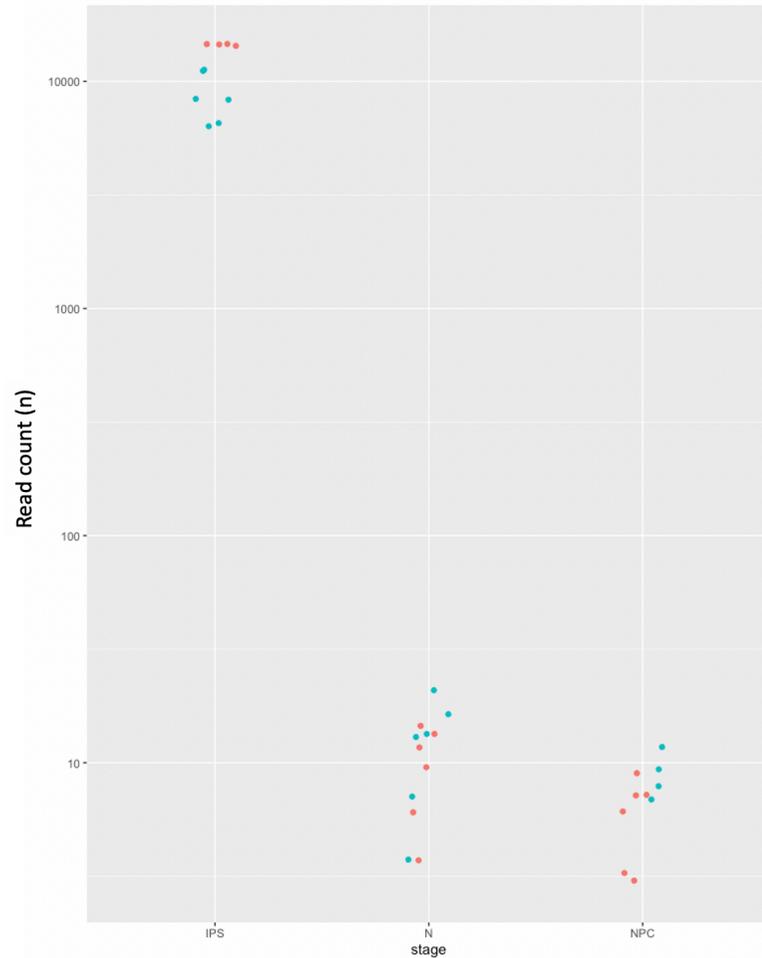
C – D40 of neurodifferentiation			
Gene	Normalised readcount in iCn-WT cells (mean, sd)	Normalised readcount in iCn-CHD2 ^{+/-} cells (mean, sd)	P value for comparison between readcounts
NANOG	1.7(0.7)	2.1(1.7)	0.62
OCT4	13.6(6.8)	10.7(4.7)	0.40
SOX2	23197.8(2166.7)	35609.9(5329.6)	5.30x10 ⁻⁸
PAX6	8245.5(800)	598.6(407.7)	1.49x10 ⁻⁹
TBR2	314.4(13.2)	29.9(11.6)	2.57x10 ⁻¹²
ASCL1	3133.2(295.3)	7271.8(641.1)	4.85x10 ⁻⁸
TBR1	847.8(144)	74.6(64.8)	2.93x10 ⁻⁷
TUBB3	486.2(32.5)	418.3(76.5)	0.07
MAP2	413.4(47.3)	2391.3(330.9)	0.0003
PSD95	2385.4(303.9)	3362.2(593.3)	0.004

Table 4.5 (page 2/2): mean normalised readcounts for expression of neurodifferentiation markers at: A-D0, B-D19 and C-D40 of neurodifferentiation

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

Comparison of RNA-seq readcount for OCT4 transcripts between WT cells and CHD2 mutant cells at IPS cells, NPC and mature neuron stage of differentiation



Comparison of RNA-seq readcount for NANOG transcripts between WT cells and CHD2 mutant cells at IPS cells, NPC and mature neuron stage of differentiation

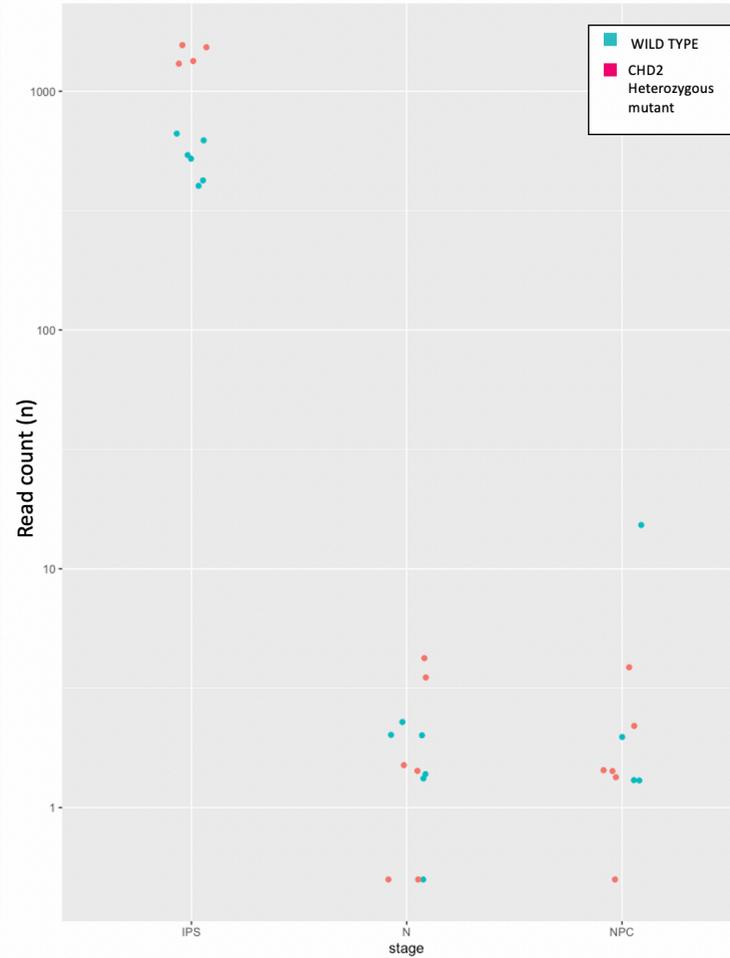


Figure 4.7.1 comparison of levels of pluripotency markers OCT4 and NANOG at three stages of neurodifferentiation, IPS = D0, NPC = D19, N = D40

Both markers are higher at D0 than D19 or D40, which is the expected pattern

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

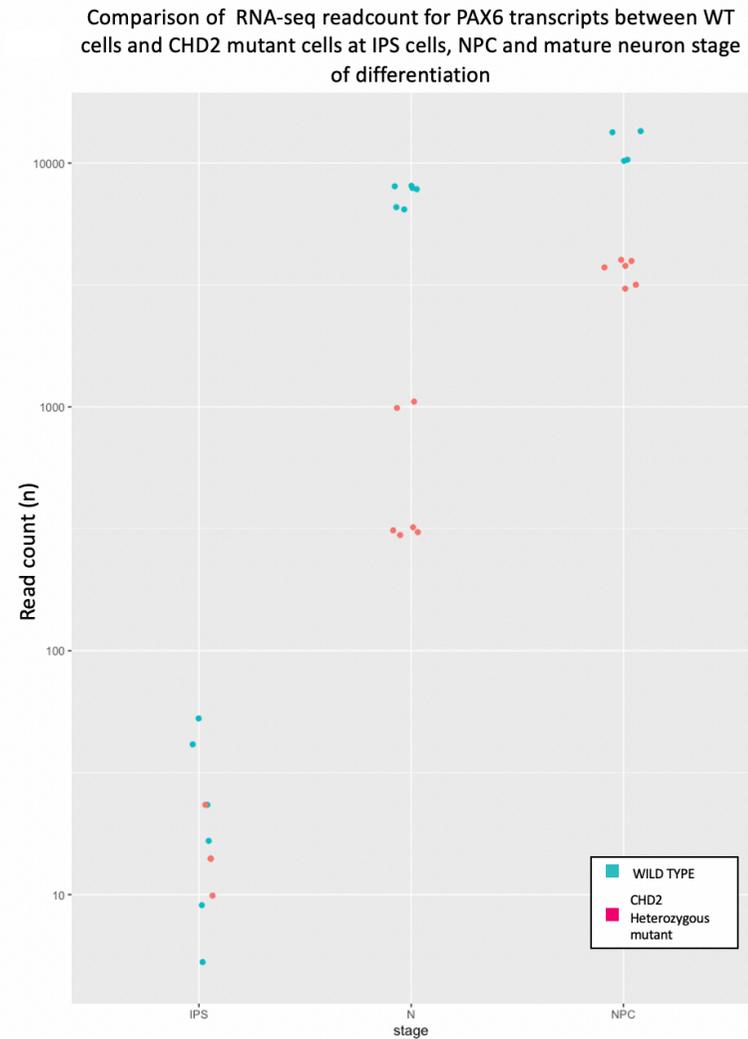
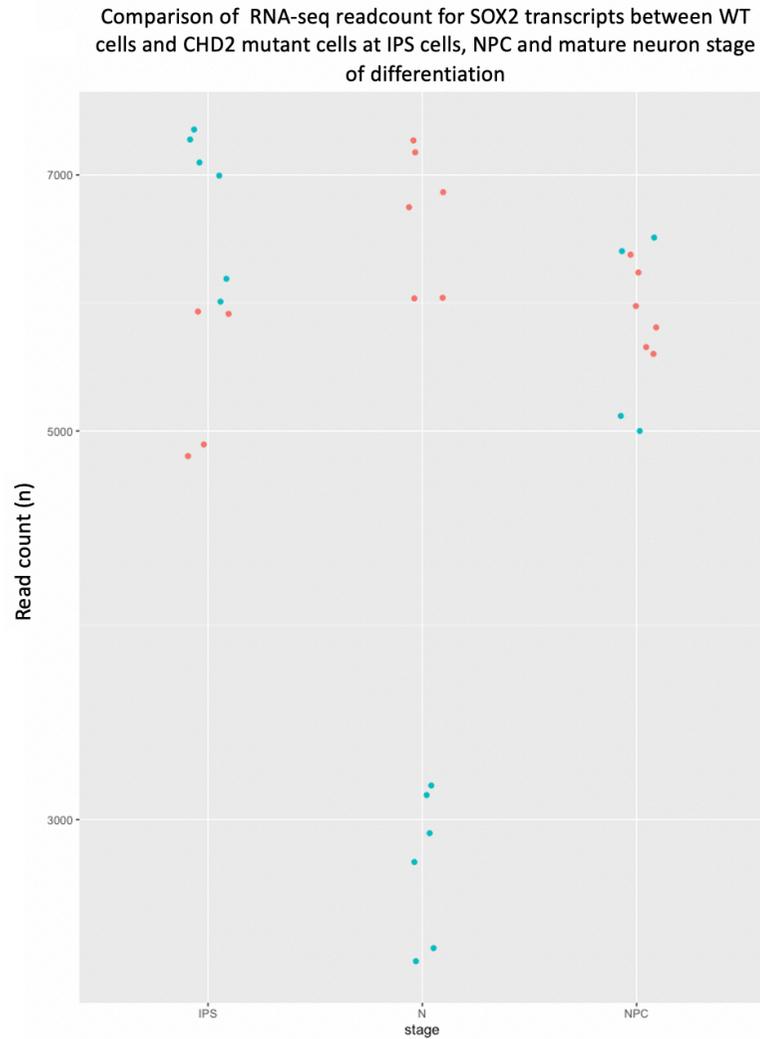


Figure 4.7.2 comparison of levels of NPC markers SOX2 and PAX6 at three stages of neurodifferentiation, IPS = D0, NPC = D19, N = D40

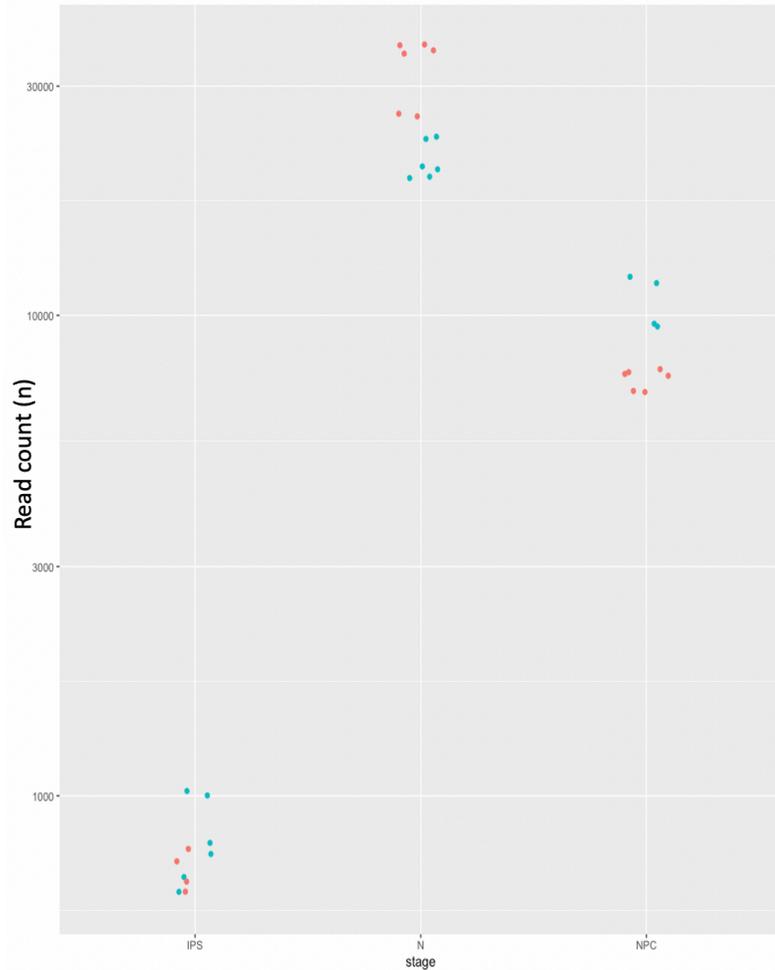
SOX2 is elevated in IPS cells and NPC cells, as expected in both cell lines, however expression continues at D40 in CHD2 mutant line

PAX6 expression is low at D0 and increased by D19

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

Comparison of RNA-seq readcounts for MAP2 transcripts between WT cells and CHD2 mutant cells at IPS cells, NPC and mature neuron stage of differentiation



Comparison of RNA-seq readcounts for PSD95 transcripts between WT cells and CHD2 mutant cells at IPS cells, NPC and mature neuron stage of differentiation

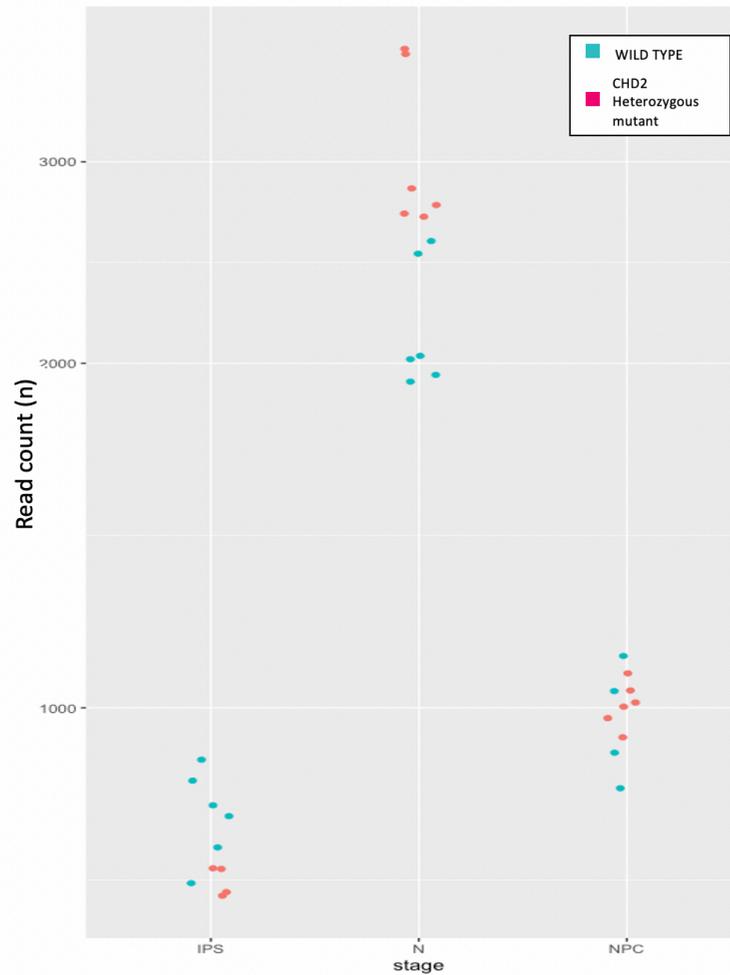


Figure 4.7.3 comparison of levels of neuronal markers MAP2 and PSD95 at three stages of neurodifferentiation, IPS = D0, NPC = D19, N = D40

Both markers start to increase by D19 and are highest at D40

4.3: Results

4.3.4.1 Summary of transcriptional evidence for neurodifferentiation timing

In both the WT-iCn and the CHD2^{+/-} cell lines, the most highly transcribed markers at D0, D19 and D40 were the most appropriate markers for IPS, NPC and mature neurons respectively. It is therefore reasonable to treat these time points as representative of IPS, NPC and mature neuron for further experimentation.

There were some differences between the levels of transcription of each marker, at each stage, that may indicate that sub-populations of cells within each culture were maturing at different rates. Without repeating this experiment with a greater number of time points, it is not possible to describe the full time-course of transcription of each marker.

4.3.5 Comparisons of transcriptomes between iCn-WT cells and iCn-CHD2^{+/-} cells at D0, D19 and D40 of neurodifferentiation

Having demonstrated that D0, D19 and D40 exhibit a pattern of transcription of established marker genes consistent with IPS, NPC and mature neurons, in this section we examine the whole transcriptome dataset for larger differences in transcriptional programming at each time point.

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

RNA-SEQ DATA FROM WT AND CHD2 HETEROZYGOUS MUTANT CELL LINES AT D0 OF NEURODIFFERENTIATION

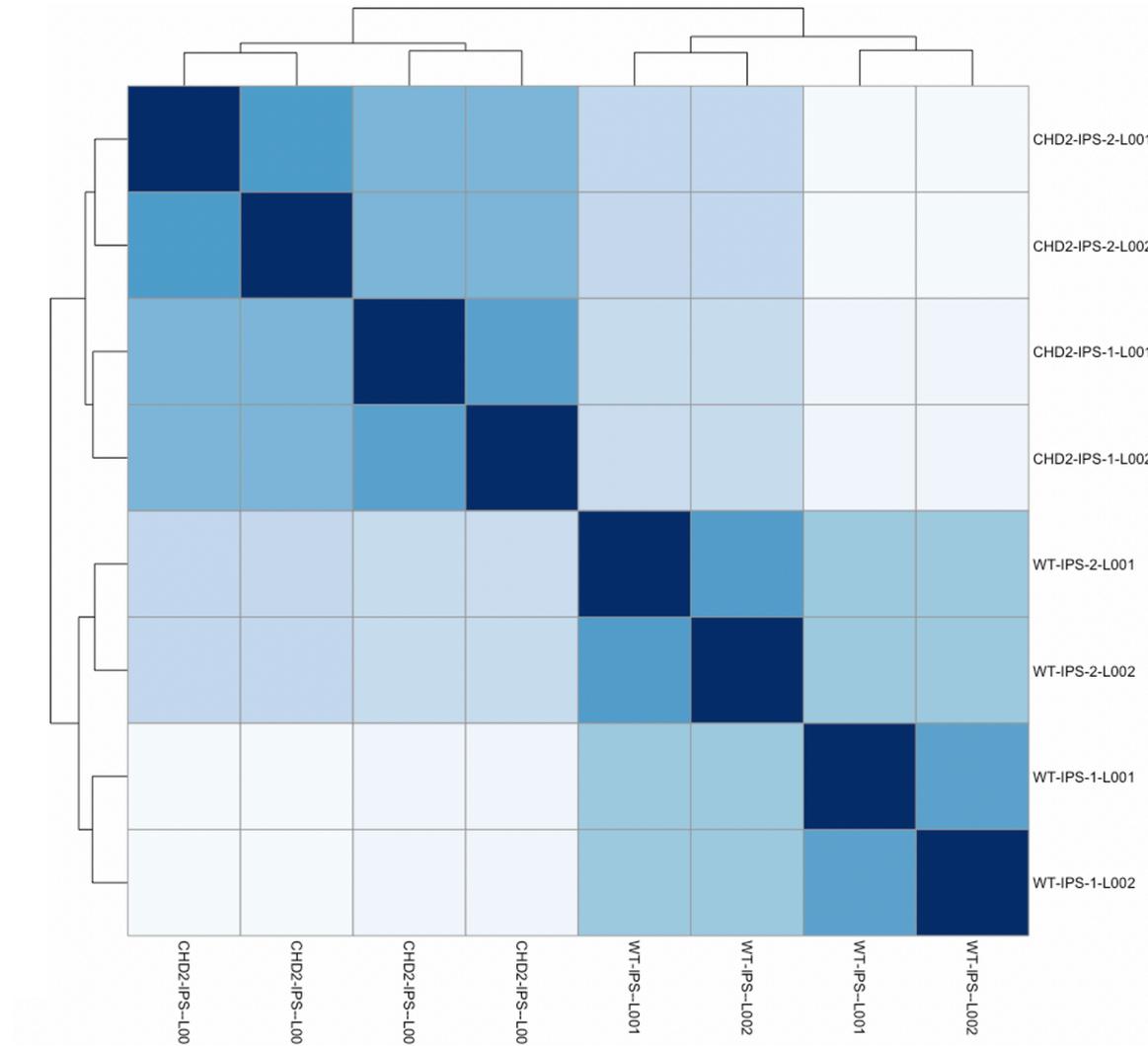


Figure 4.8.1: sample distance matrix comparing RNA-seq data from iCn-WT and iCn-CHD2^{+/-} cell lines at D0 neurodifferentiation – darker squares indicate that samples are more similar to each other

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

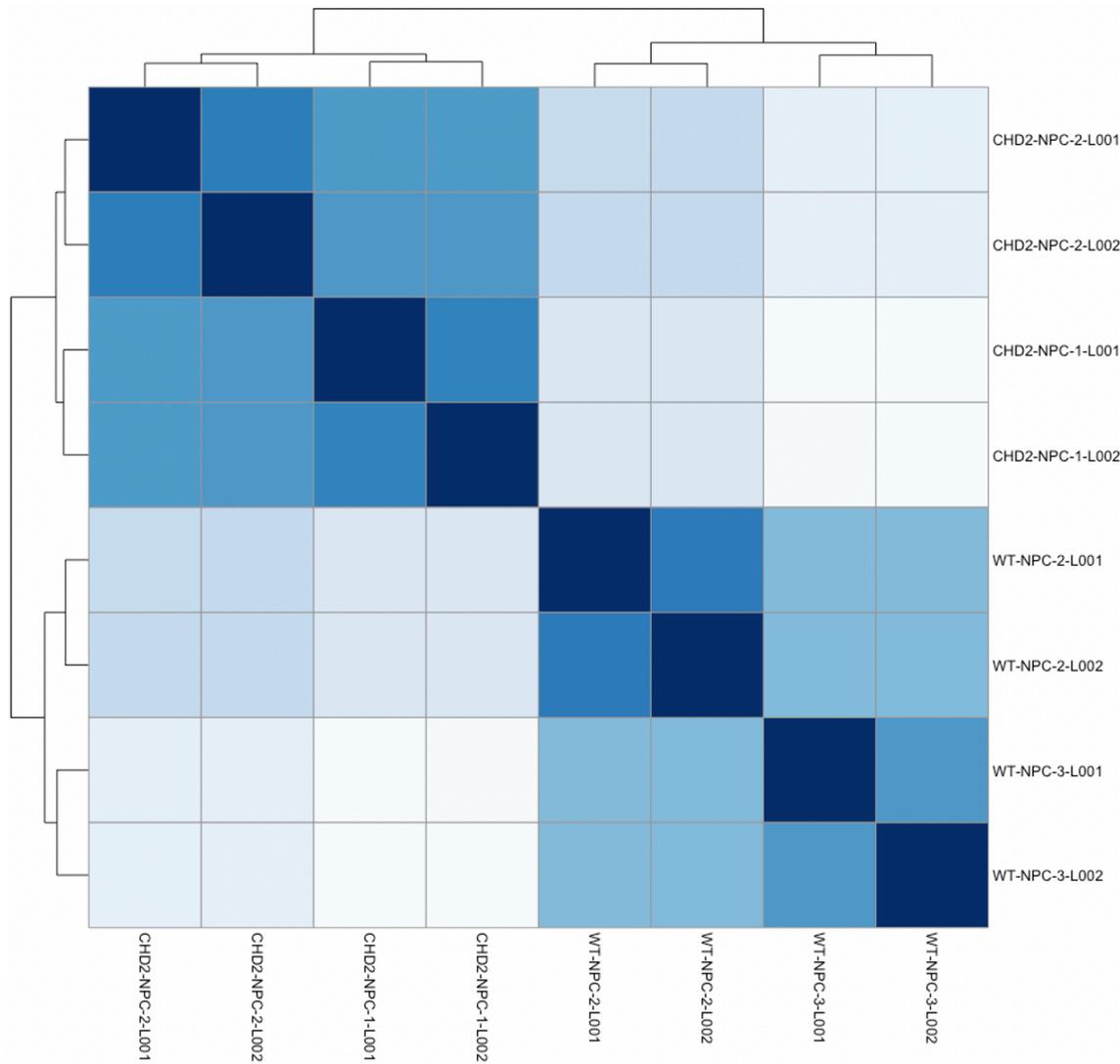


Figure 4.8.2: sample distance matrix comparing RNA-seq data from iCn-WT and iCn-CHD2^{+/-} cell lines at D19 of neurodifferentiation – darker squares indicate that samples are more similar to each other

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

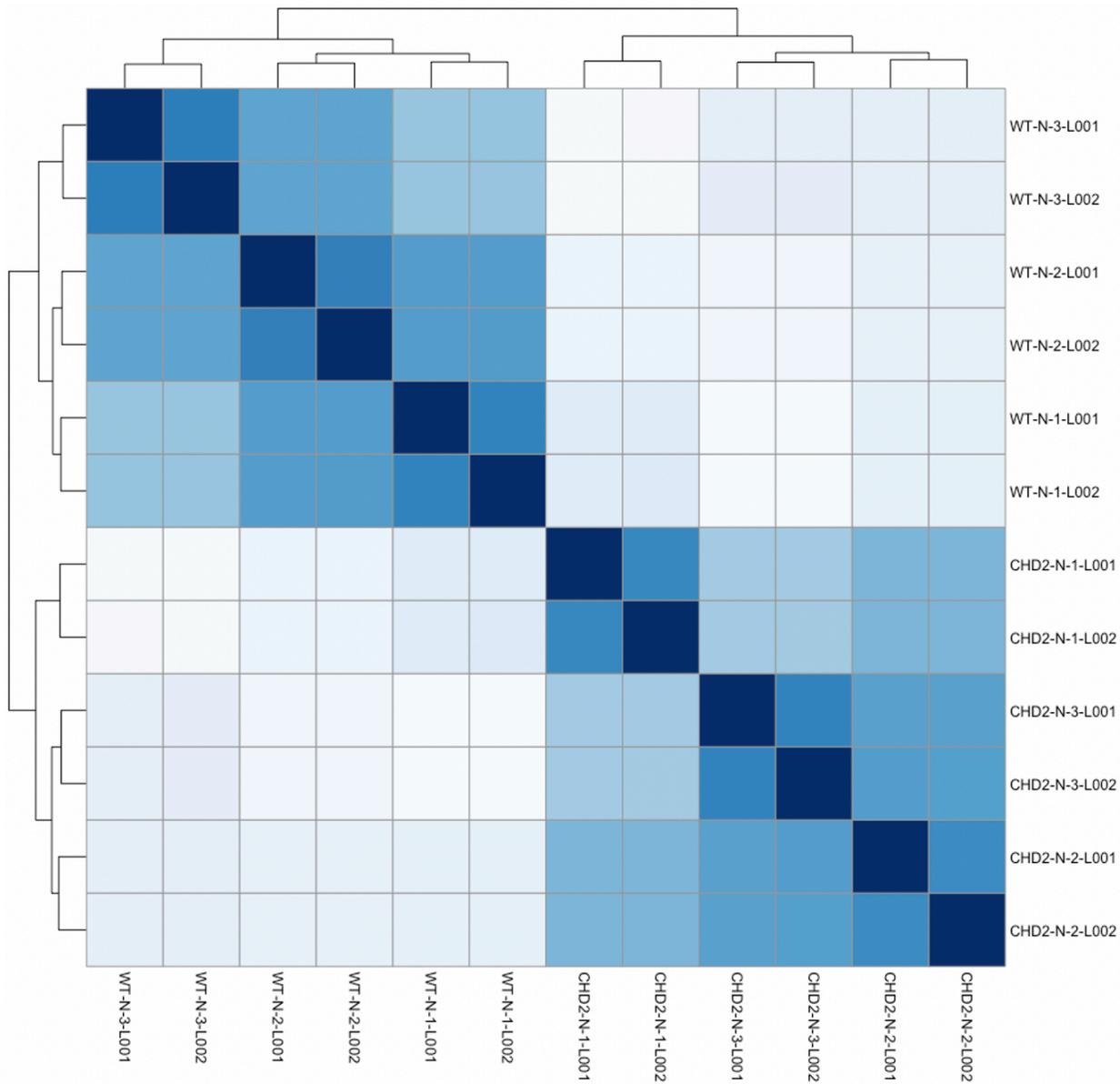


Figure 4.8.3: sample distance matrix comparing RNA-seq data from iCn-WT and iCn-CHD2^{+/-} cell lines at D40 of neurodifferentiation – darker squares indicate that samples are more similar to each other

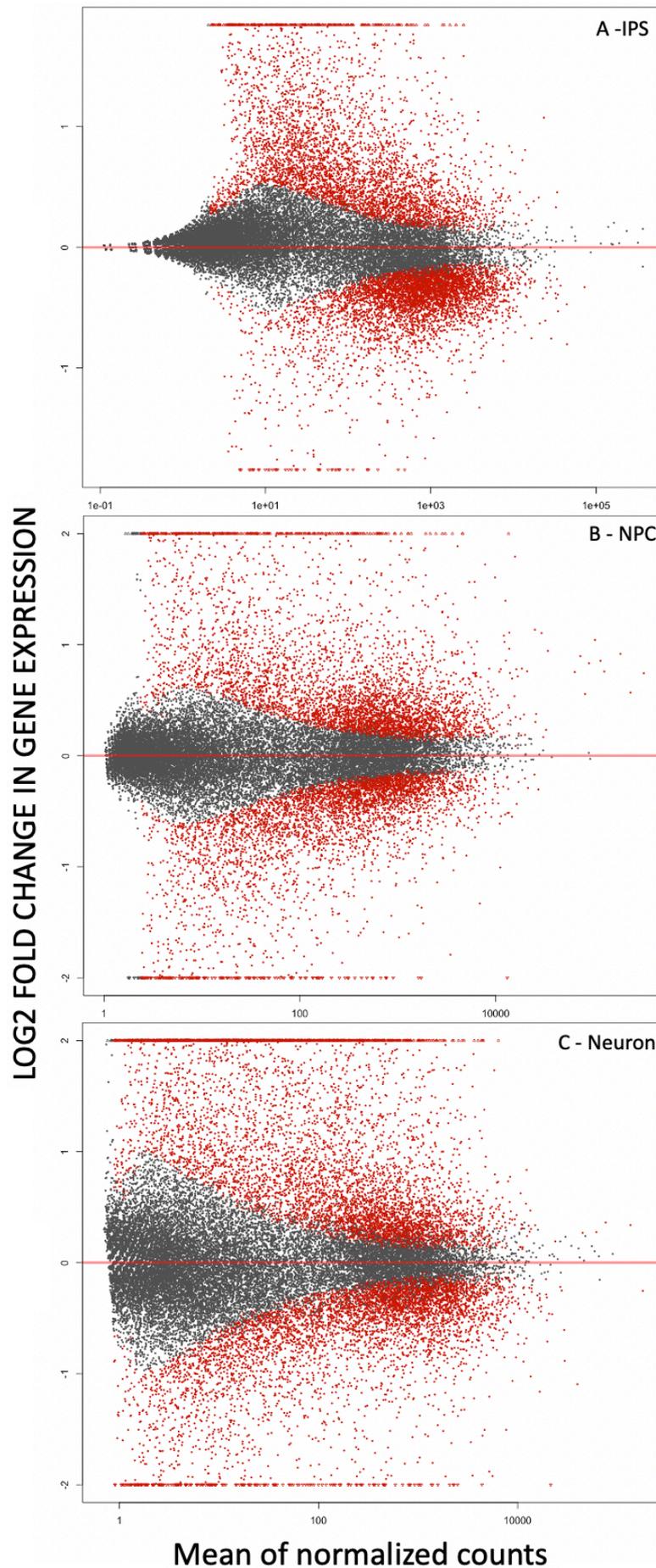


Figure 4.9: MA plots for gene expression comparisons between iCn-WT and iCn-CHD2+/- cell lines at D0 (IPS), D19 (NPC) and D40 (neuron) of neurodifferentiation. Red dots indicate transcripts where there is a statistically significant change in gene expression

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

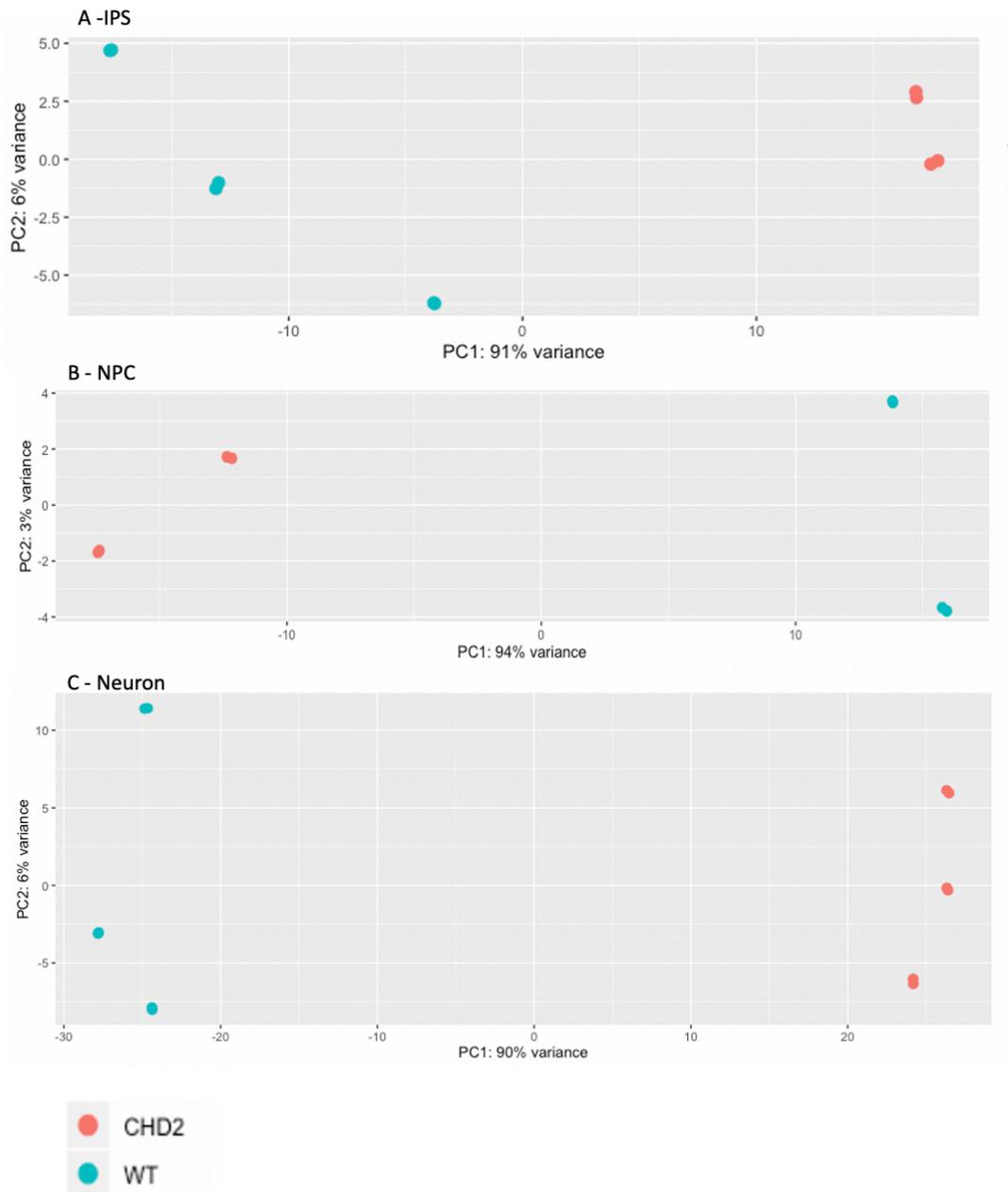


Figure 4.10 – PCA comparing RNA-seq sample distance at D0 (IPS), D19 (NPC) and D40 (neuron) of neurodifferentiation between WT cells and CHD2 heterozygous mutant cells

4.3.5.1 Sample distance, MA plots and PCA plots

Sample distance matrices and PCAs were constructed for *only* the samples to be compared (D0 compared with D0, D19 with D19 and D40 with D40) (*figure 4.8*). At D0, there was a greater difference between samples from different cell lines than there was between samples from the same cell line (*figure 4.8.1*). This pattern was similar at D19 (*figure 4.8.2*) and at D40 (*figure 4.8.3*). At all three time-points there was a greater distance between biological replicates than technical replicates. Hierarchical clustering (visible on the heatmaps) confirms this, with samples clustering first as technical replicates, then as biological replicates from the same cell line.

The MA plots demonstrated a roughly even spread of statistically significant transcription changes (red) at D19 and D40, however at D0 iCn-CHD2^{+/-} cells were more depleted for highly transcribed genes (*figure 4.9*).

Principal component analysis (PCA) (*figure 4.10*) indicated a wide separation for PC1 and for PC2 in all three samples, again confirming that there was a greater difference between runs of the different cell lines than there was within biological replications. To further consider the differences between cell types at each developmental time-point, genes with a LFC > 0.5 and p-value below 0.05 were collated for ontology analysis (*table 4.6*)

Time point	Number of genes upregulated in iCn-CHD2 ^{+/-} (n)	Number of genes downregulated in iCn-CHD2 ^{+/-} (n)
D0	2511	1550
D19	2209	1866
D40	2489	3913

Table 4.6 – number of genes upregulated (LFC >0.5 p=0.05) and downregulated (LFC <-0.5, p=0.05) in iCn-CHD2^{+/-} cell lines compared with iCn-WT cell lines at D0, D19 and D40 of neurodifferentiation

4.3: Results

4.3.5.2 GO term enrichment

iCn-CHD2^{+/-} transcriptomes were depleted for genes related to: regulation of cell migration, anatomical structure development and system processes. Transcriptomes at this stage were enriched for: regulation of metabolic processes, cellular developmental processes, and response to chemical stimulus.

At D19 the iCn-CHD2^{+/-} transcriptomes were depleted for genes related to regulation of cell differentiation, regulation of developmental processes and metabolic processes. They were enriched for genes related to cell adhesion, locomotion

At D40, the iCn-CHD2^{+/-}, the transcriptomes were depleted for genes related to cell-fate commitment and cell differentiation, as well as multiple metabolic processes particularly pertaining to gene expression, transcription and RNA metabolism.

Of particular note, was a depletion of transcription in genes related to forebrain neuron differentiation. iCn-CHD2^{+/-} cell transcriptomes at D40, there was enrichment for genes related to developmental processes, nervous system processes, regulation of nervous system development, and regulation of neurogenesis. There was also enrichment for terms related to transcription and RNA metabolism.

The data relevant to these findings is displayed in graphical format (tree map and enrichment map) for D0 of neurodifferentiation in *figures 4.11 & 4.12*, for day 19 of neurodifferentiation in *figures 4.13 & 4.14* and for D40 of neurodifferentiation in *figures 4.15 & 4.16*. The top 20 enriched GO terms by p-value for each time-point can be found in *tables 4.7, 4.8 and 4.9*.

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

GO TERMS DEPLETED TRANSCRIPTION IN *iCn-CHD2^{+/-}* CELLS AT D0 OF NEURODIFFERENTIATION

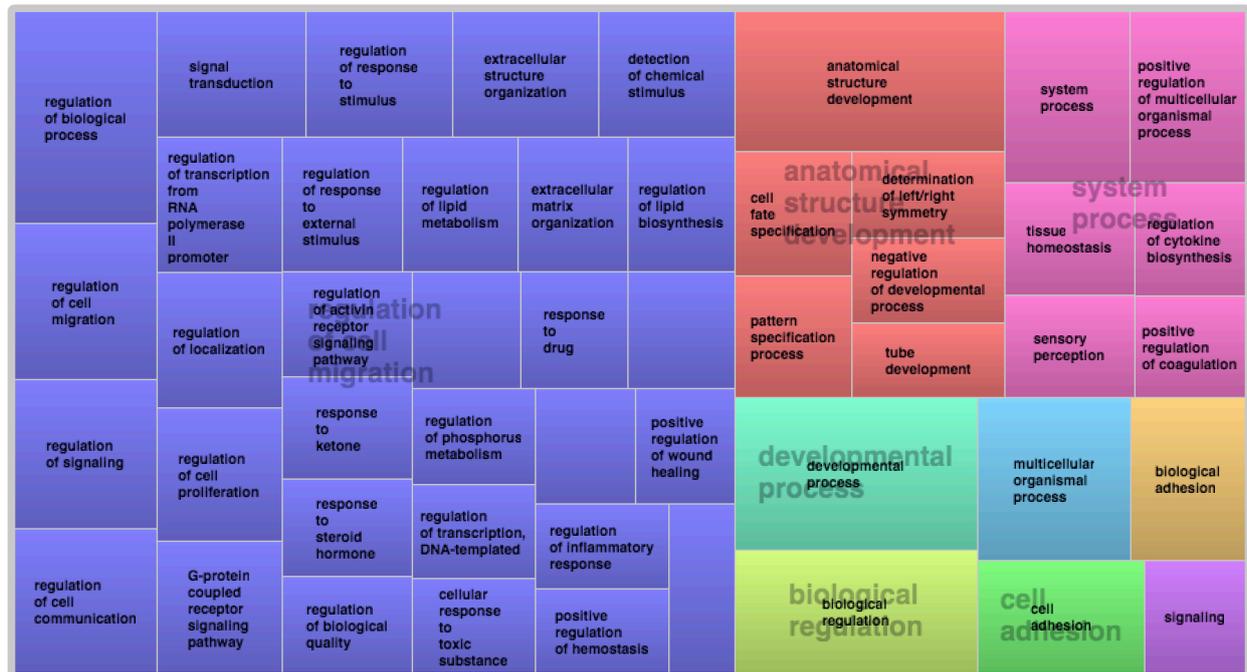
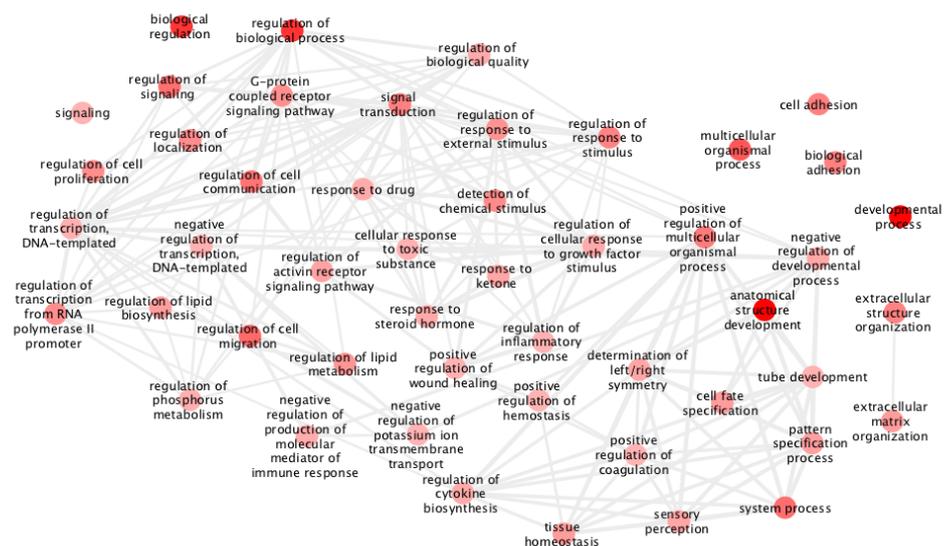


Figure 4.11:

TOP: Tree map of GO terms enriched in genes downregulated in *iCn-CHD2^{+/-}* cells at D0 of differentiation. Same coloured regions denote linked GO terms and size of rectangle denotes p value, with larger spaces being of higher statistical significance.



BOTTOM: Enrichment map of same data, with darker circles representing higher statistical significance and thicker grey lines joining circles denoting greater degree of overlap between lists of genes under each term heading.

Transcription levels of genes clustering under cell migration regulation, anatomical structure development, system processes, biological regulation and developmental process were downgraded in pluripotent stem cells harbouring heterozygous *CHD2* mutations (see also table 4.7).

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

GO TERMS ENRICHED TRANSCRIPTION IN iCn-CHD2^{+/-} CELLS AT D0 OF NEURODIFFERENTIATION

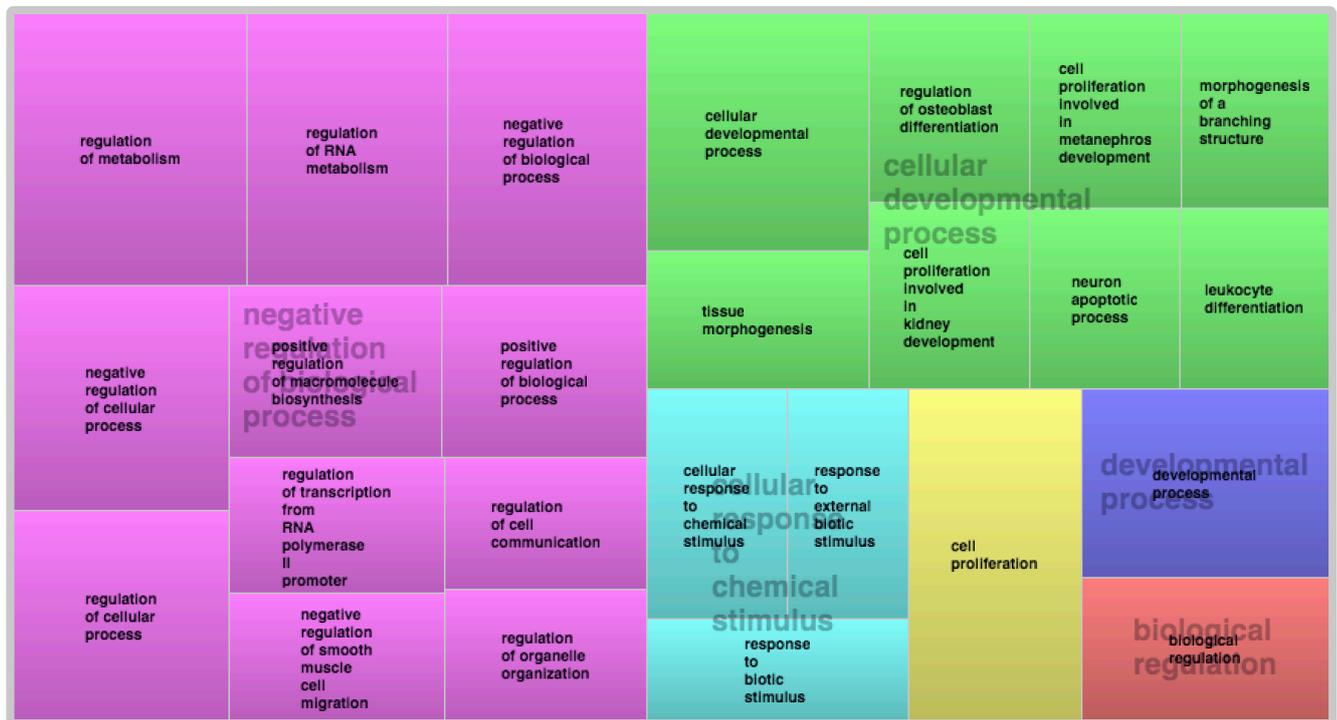
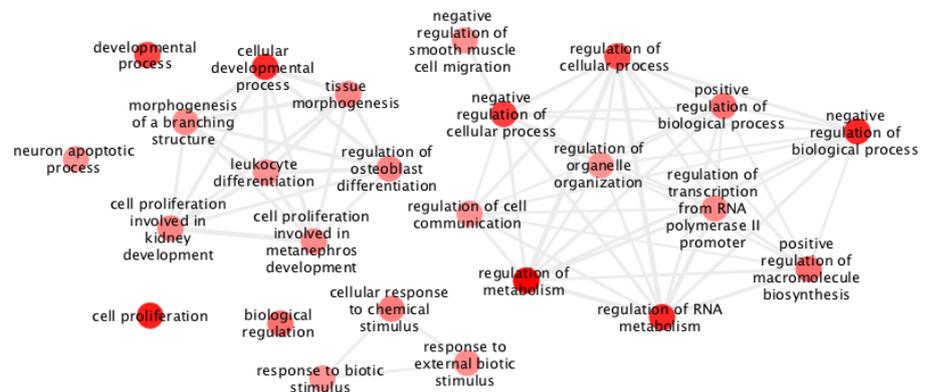


Figure 4.12:

TOP: Tree map of GO terms enriched in genes enriched in iCn-CHD2^{+/-} cells at D0 of differentiation. Same coloured regions denote linked GO terms and size of rectangle denotes p value, with larger spaces being of higher statistical significance.



BOTTOM: Enrichment map of same data, with darker circles representing higher statistical significance and thicker grey lines joining circles denoting greater degree of overlap between lists of genes under each term heading.

Transcription levels of genes clustering under negative regulation of biological process, cellular developmental process and cellular response to chemical were upregulated in pluripotent stem cells harbouring heterozygous CHD2 mutations (see also table 4.7)

ENRICHED IN iCn-CHD2 ^{+/-} CELLS AT D0 OF NEURODIFFERENTIATION		DEPLETED IN iCn-CHD2 ^{+/-} CELLS AT D0 OF NEURODIFFERENTIATION	
Gene ontology group	P value	Gene ontology group	P value
regulation of metabolic process	4.75E-08	anatomical structure formation involved in morphogenesis	1.45E-06
regulation of cellular metabolic process	1.24E-07	developmental process	2.37E-06
regulation of nucleobase-containing compound metabolic process	2.17E-07	regulation of anatomical structure morphogenesis	8.48E-06
cell proliferation	2.61E-07	regulation of cellular amide metabolic process	9.46E-06
regulation of macromolecule biosynthetic process	2.95E-07	cell development	1.11E-05
regulation of macromolecule metabolic process	3.71E-07	negative regulation of cellular process	1.50E-05
regulation of primary metabolic process	4.32E-07	biological regulation	1.61E-05
regulation of nitrogen compound metabolic process	4.50E-07	anatomical structure development	1.87E-05
regulation of RNA metabolic process	5.18E-07	regulation of multicellular organismal process	2.08E-05
regulation of cellular biosynthetic process	5.21E-07	regulation of cellular process	2.48E-05
regulation of biosynthetic process	5.70E-07	actin filament bundle organization	2.59E-05
negative regulation of biological process	5.71E-07	regulation of angiogenesis	3.68E-05
regulation of cellular macromolecule biosynthetic process	7.53E-07	regulation of translation	4.68E-05
cellular developmental process	8.13E-07	negative regulation of multicellular organismal process	5.03E-05
regulation of gene expression	1.78E-06	regulation of biological process	5.41E-05
regulation of RNA biosynthetic process	2.10E-06	angiogenesis	6.09E-05
negative regulation of cellular process	2.59E-06	response to chemical	7.08E-05
developmental process	4.02E-06	regulation of vasculature development	7.49E-05
regulation of nucleic acid-templated transcription	4.10E-06	actin filament bundle assembly	7.56E-05
regulation of transcription, DNA-templated	5.46E-06	regulation of cell adhesion	8.23E-05

Table 4.7: Top 20 enriched and depleted GO terms in transcriptomes of iCn-CHD2^{+/-} cells at D0 of neurodifferentiation

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

GO TERMS DEPLETED FOR TRANSCRIPTION IN iCn-CHD2^{+/-} CELLS AT D19 OF NEURODIFFERENTIATION

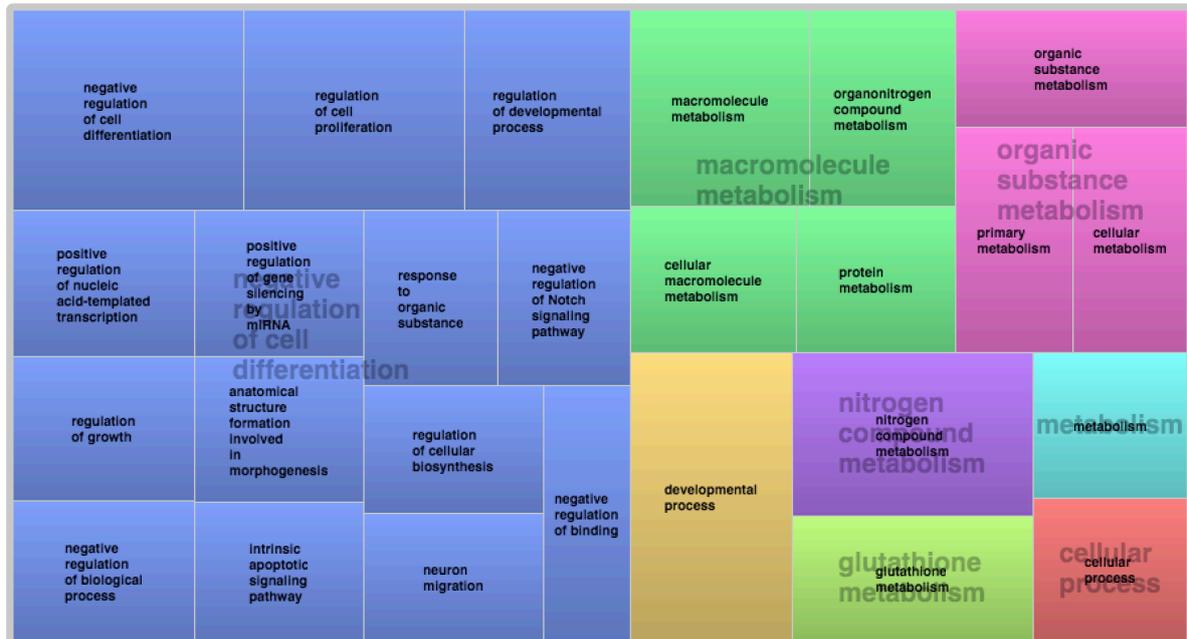
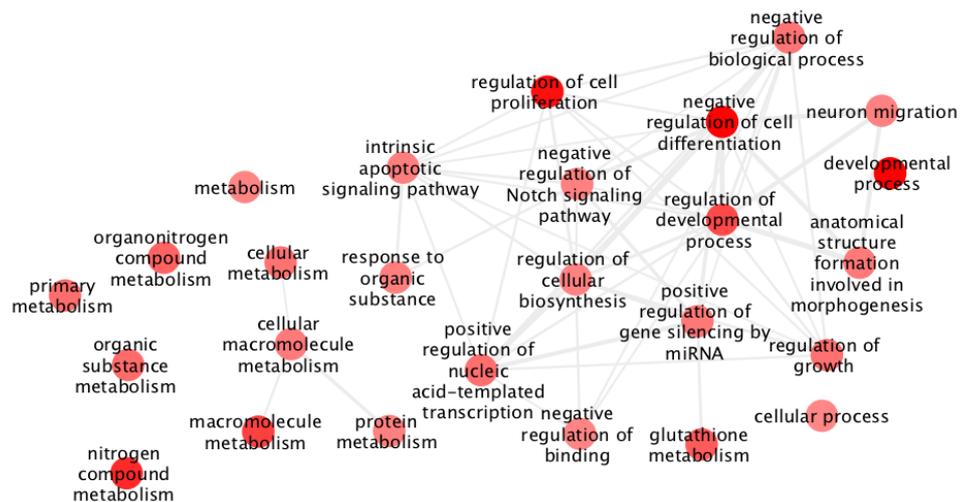


Figure 4.13:

TOP: Tree map of GO terms enriched in genes downregulated in iCn-CHD2^{+/-} cells at D0 of differentiation. Same coloured regions denote linked GO terms and size of rectangle denotes p value, with larger spaces being of higher statistical significance.



BOTTOM: Enrichment map of same data, with darker circles representing higher statistical significance and thicker grey lines joining circles denoting greater degree of overlap between lists of genes under each term heading.

Transcription of genes clustering under negative regulation of cell differentiation, macromolecule metabolism and organic substance metabolism were downregulated at D19 of neurodifferentiation in cells harbouring heterozygous CHD2 mutations (see also table 4.8).

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

GO TERMS ENRICHED FOR TRANSCRIPTION IN iCn-CHD2^{+/-} CELLS AT D19 OF NEURODIFFERENTIATION

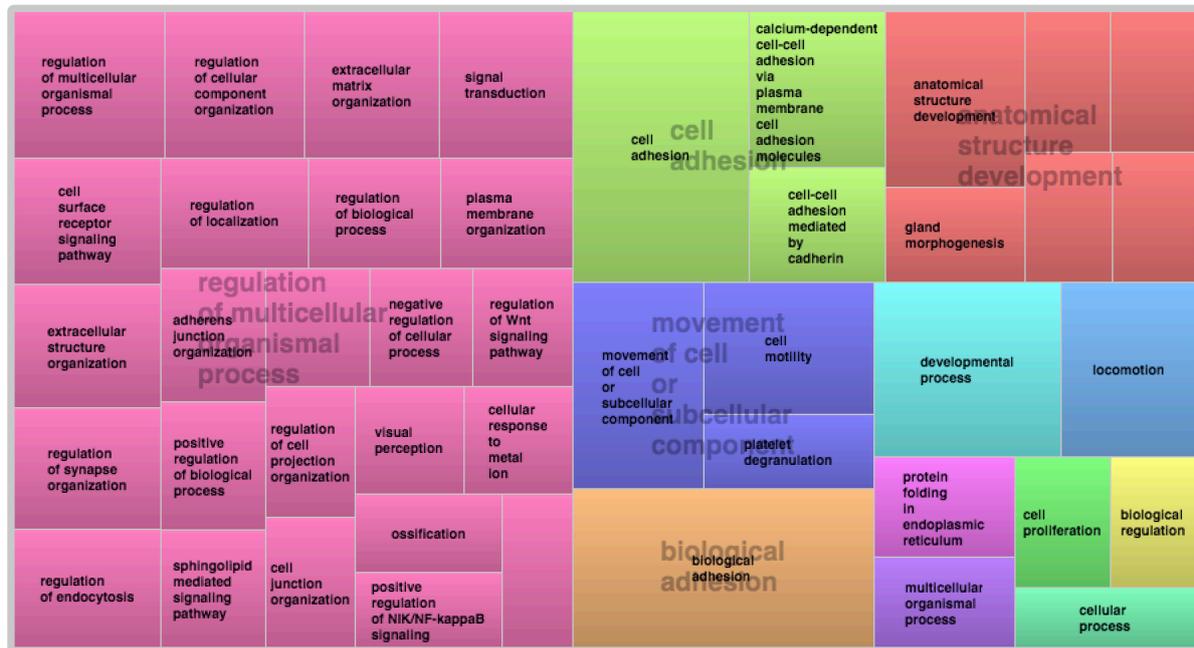
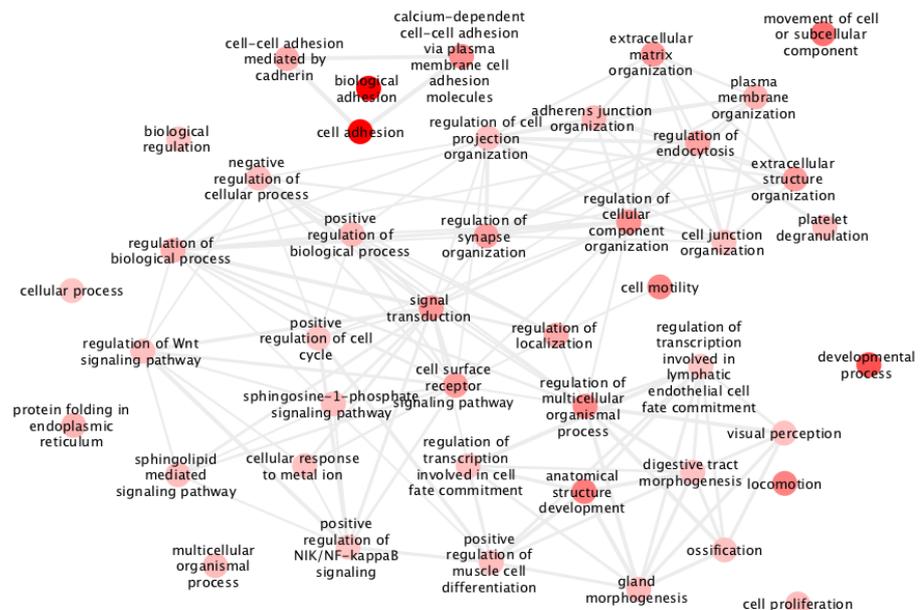


Figure 4.14:

Tree map of GO terms enriched in genes downregulated in iCn-CHD2^{+/-} cells at D0 of differentiation. Same coloured regions denote linked GO terms and size of rectangle denotes *p* value, with larger spaces being of higher statistical significance.



Enrichment map of same data, with darker circles representing higher statistical significance and thicker grey lines joining circles denoting greater degree of overlap between lists of genes under each term heading.

Transcription levels of genes clustering under cell adhesion, regulation of organismal processes, anatomical structure development, and movement of cells were upregulated at D19 of neurodifferentiation in cells harbouring heterozygous CHD2 mutation (see also table 4.8).

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

ENRICHED IN iCn-CHD2 ^{+/-} CELLS AT D19 OF NEURODIFFERENTIATION		DEPLETED IN iCn-CHD2 ^{+/-} CELLS AT D19 OF NEURODIFFERENTIATION	
Gene ontology group	P value	Gene ontology group	P value
biological adhesion	5.87E-14	developmental process	3.71E-07
cell adhesion	5.87E-14	negative regulation of cell differentiation	4.31E-07
cell-cell adhesion	1.53E-10	regulation of cell proliferation	7.93E-07
developmental process	8.99E-10	nitrogen compound metabolic process	3.81E-06
movement of cell or subcellular component	3.11E-08	regulation of cell differentiation	9.19E-06
cell-cell adhesion via plasma-membrane adhesion molecules	4.70E-08	macromolecule metabolic process	1.45E-05
anatomical structure development	1.66E-07	regulation of developmental process	2.54E-05
locomotion	3.05E-07	negative regulation of cellular process	3.07E-05
cell motility	6.00E-07	anatomical structure development	3.18E-05
anatomical structure morphogenesis	6.95E-07	positive regulation of developmental process	4.48E-05
regulation of multicellular organismal process	7.48E-07	regulation of stem cell proliferation	5.11E-05
regulation of multicellular organismal development	7.93E-07	glutathione metabolic process	6.97E-05
cell migration	1.04E-06	negative regulation of developmental process	9.57E-05
calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules	1.31E-06	cell differentiation	9.75E-05
regulation of developmental process	1.33E-06	organonitrogen compound metabolic process	1.14E-04
regulation of cellular component organization	2.22E-06	negative regulation of cell development	1.30E-04
regulation of neuron differentiation	2.33E-06	organic substance metabolic process	1.82E-04
regulation of cell migration	2.53E-06	primary metabolic process	2.19E-04
extracellular matrix organization	3.45E-06	positive regulation of RNA biosynthetic process	2.22E-04
signal transduction	3.93E-06	positive regulation of nucleic acid-templated transcription	2.22E-04

Table 4.8: Top 20 enriched and depleted GO terms in transcriptomes of iCn-CHD2^{+/-} cells at D19 of neurodifferentiation

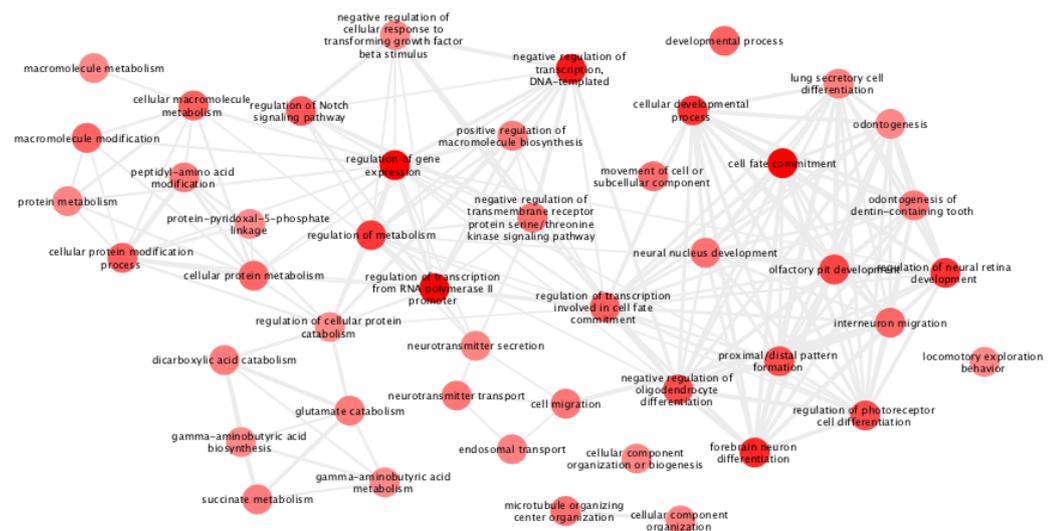
4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

GO TERMS DEPLETED FOR TRANSCRIPTION IN iCn-CHD2^{+/-} CELLS AT D19 OF NEURODIFFERENTIATION



Figure 4.15: TOP: Tree map of GO terms enriched in genes downregulated in iCn-CHD2^{+/-} cells at D0 of differentiation. Same coloured regions denote linked GO terms and size of rectangle denotes p value, with larger spaces being of higher statistical significance.



BOTTOM: Enrichment map of same data, with darker circles representing higher statistical significance and thicker grey lines joining circles denoting greater degree of overlap between lists of genes under each term heading.

Transcription levels of genes clustering under negative regulation of cell differentiation, macromolecule metabolism and organic substance metabolism were downregulated at D40 of neurodifferentiation in cells harbouring heterozygous CHD2 mutations (see also table 4.9).

4: Development of a CHD2 mutant cell line and characterisation using RNA-Seq

4.3: Results

GO TERMS ENRICHED FOR TRANSCRIPTION IN iCn-CHD2^{+/-} CELLS AT D19 OF NEURODIFFERENTIATION

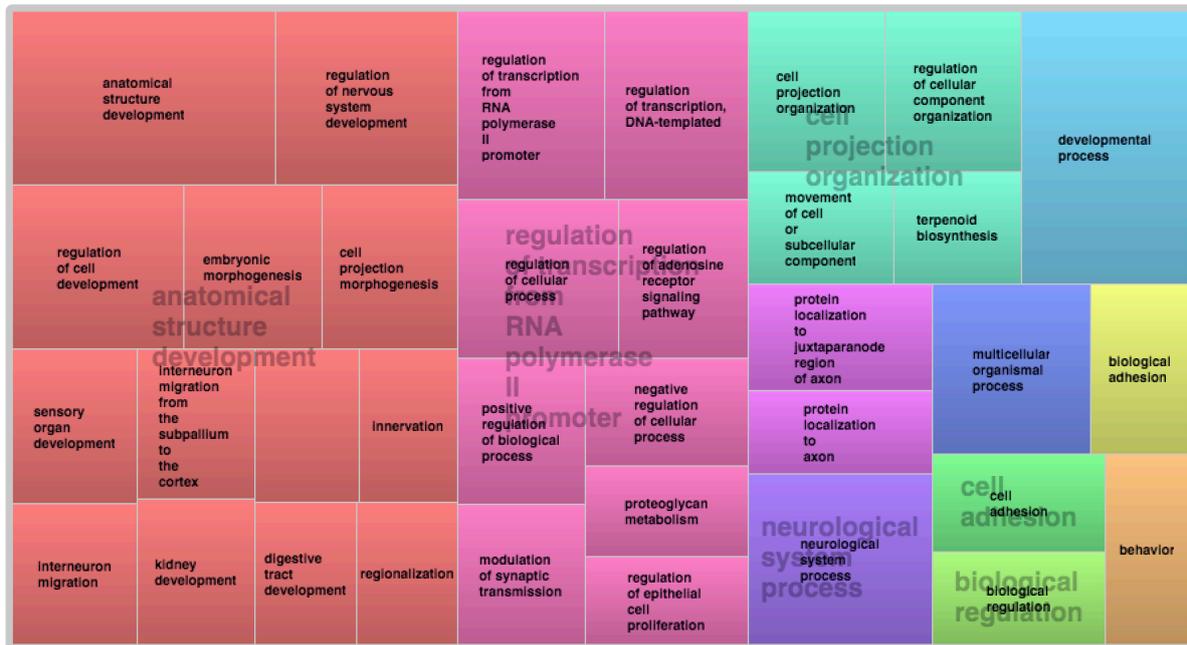
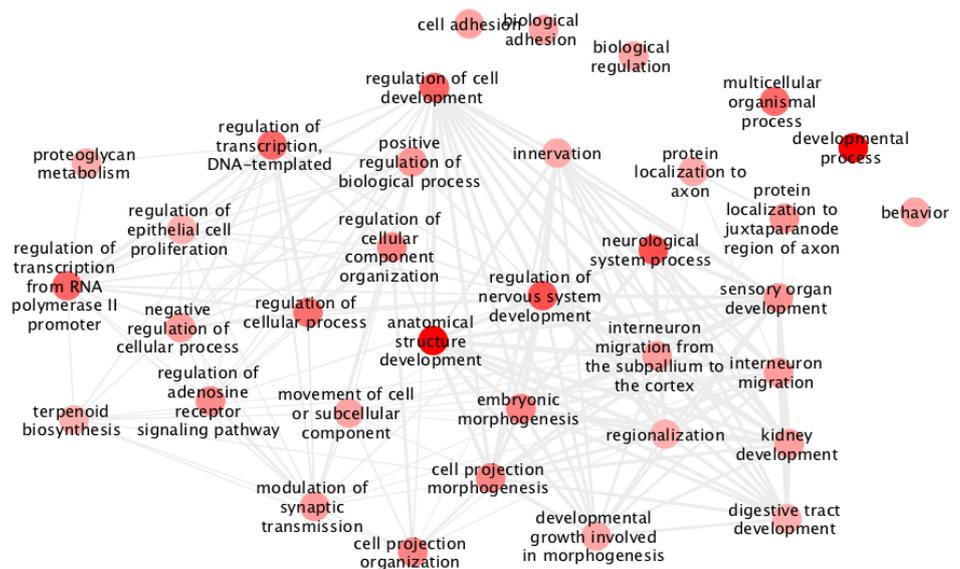


Figure 4.16:

TOP: Tree map of GO terms enriched in genes downregulated in iCn-CHD2^{+/-} cells at D0 of differentiation. Same coloured regions denote linked GO terms and size of rectangle denotes p value, with larger spaces being of higher statistical significance.



BOTTOM: Enrichment map of same data, with darker circles representing higher statistical significance and thicker grey lines joining circles denoting greater degree of overlap between lists of genes under each term heading.

Transcription levels of genes clustering under cell migration regulation, anatomical structure development, system processes, biological regulation and developmental process were upregulated at D40 of neurodifferentiation in cells harbouring heterozygous CHD2 mutation (see table 4.9).

ENRICHED IN iCn-CHD2 ^{+/-} CELLS AT D40 OF NEURODIFFERENTIATION		DEPLETED IN iCn-CHD2 ^{+/-} CELLS AT D40 OF NEURODIFFERENTIATION	
Gene ontology group	P value	Gene ontology group	P value
developmental process	1.18E-10	regulation of macromolecule biosynthetic process	8.87E-08
anatomical structure development	1.18E-10	cell fate commitment	2.71E-07
anatomical structure morphogenesis	1.41E-09	regulation of gene expression	2.79E-07
animal organ development	8.97E-09	regulation of transcription by RNA polymerase II	2.89E-07
regulation of anatomical structure morphogenesis	9.67E-08	regulation of cellular macromolecule biosynthetic process	2.98E-07
nervous system process	1.42E-07	regulation of RNA metabolic process	3.54E-07
regulation of nervous system development	1.43E-07	negative regulation of cellular macromolecule biosynthetic process	4.58E-07
positive regulation of cell development	1.68E-07	regulation of nitrogen compound metabolic process	5.54E-07
positive regulation of nervous system development	1.68E-07	regulation of cellular metabolic process	7.72E-07
positive regulation of neurogenesis	2.82E-07	negative regulation of transcription, DNA-templated	9.48E-07
regulation of cell development	7.73E-07	regulation of macromolecule metabolic process	9.86E-07
regulation of nucleic acid-templated transcription	8.89E-07	negative regulation of macromolecule biosynthetic process	1.15E-06
regulation of transcription by RNA polymerase II	9.48E-07	regulation of RNA biosynthetic process	1.34E-06
regulation of RNA biosynthetic process	1.05E-06	cell differentiation	1.43E-06
regulation of macromolecule biosynthetic process	1.26E-06	negative regulation of RNA metabolic process	1.62E-06
regulation of transcription, DNA-templated	1.34E-06	regulation of nucleic acid-templated transcription	1.62E-06
positive regulation of neuron differentiation	1.37E-06	regulation of cellular biosynthetic process	1.62E-06
multicellular organismal process	1.45E-06	regulation of biosynthetic process	2.26E-06
regulation of biosynthetic process	1.53E-06	forebrain neuron differentiation	2.47E-06
regulation of neuron differentiation	1.53E-06	negative regulation of RNA biosynthetic process	2.68E-06

Table 4.9: Top 20 enriched and depleted GO terms in transcriptomes of iCn-CHD2^{+/-} cells at D40 of neurodifferentiation

4.3.5.3 Summary of findings from whole transcriptome analysis

The RNA-Seq datasets cluster together by cell type, then by cell line (*figure 4.6*), demonstrating that although there are differences between the transcriptomes of each cell line at any given stage of differentiation, the transcriptomes are more similar to one another than they are to cells from the same line at different stages of differentiation.

This hierarchical clustering provides evidence of two important factors. Firstly, the relative close clustering at each stage of differentiation tells us that it is reasonable to make comparisons between the cell lines at these time points, and that the differences between each cell line when controlled for stage of differentiation indicates that heterozygous knock out of CHD2 causes a deviation from the expected transcriptional program during neurodifferentiation.

Although there are differences in the amplitude of expression of each cell marker at each stage of differentiation, they still follow the expected pattern. In other words, at D0 iPSC markers are expressed in both cell lines, at D19 NPC markers are expressed in both cell lines and at D40 neuronal markers are expressed in both cell lines. This confirms that these time-points can be used as proxies for iPSC, NPC and MN for the investigation of DSB repair carried out in *chapters 5 and 6*

Analysis of GO terms enriched and depleted in iCn-CHD2^{+/-} at each time point provides clues as to a possible pathophysiology for the epileptic encephalopathy syndrome exhibited in these patients, with GO terms and transcriptional modules pertaining to regulation of cell development and specialisation of neuronal structures exhibiting differential expression compared to our wild type control cell line.

4.4 DISCUSSION

4.4.1 Predicting the impact of our Cas9 induced mutations in CHD2

Section 4.3.1 demonstrates the successful engineering of an inducible Cas9 cell line, containing a compound heterozygous mutation in CHD2; n.649_659del, p.(Glu26Valfs*64) and n.649_652del, p.(Glu26Glu27del).

In determining whether these mutations are likely to have a deleterious effect on the protein, it is useful to consider them in context of the American College of Medical Genetics (ACMG) guidelines for interpretation of sequence variants[312, 313]. These guidelines provide the framework by which Clinical Geneticists determine whether or not a variant found in a patient using NGS testing is the cause of their clinical condition. As such, they are not *directly* applicable to laboratory research, however they provide a list of criteria which are helpful in deciding how to regard our variants.

It is also worth considering our variants in reference to the previously published sequence variants that have been associated with EECO in human patients, and in context of what is known about the important functional domains of the protein

The ACMG guidelines recommend analysis a variant in several different ways. They recommend searching through the growing body of variants recorded in population databases, including dbSNP[314], ClinVar[315], exAC[316] and GnomAD[16], to determine if the variant of interest has been detected previously in either healthy individuals or those with disease.

They recommend use of computational and predictive data to determine whether there is an impact on the protein. For nonsense and frameshift mutations, this is usually straightforward, however in-silico tools for predicting the effects of in-frame and missense mutations are less robust. The remaining fields are more relevant to patient family history and therefore not used in this context.

To begin with our frameshift mutation; this mutation is not present in any of the listed population databases and predicted by mutalyzer[305] to be a “null variant in a gene where LoF is a known mechanism of disease”. Even without clinical data, these factors would be enough to classify the variant as pathogenic.

The in-frame deletion is more difficult to categorise. No variants affecting the deleted amino acid residues are found in ExAC or gnomAD, however missense mutations in the same

4.4: Discussion

region are identified (figure 4.17). The affected region sits outside of any of the well-described functional domains found on UniProt, however it is very possible that these residues may still play an important role in protein folding.

The in-silico predictions are also unhelpful; provean [306] suggests that the deletion is damaging, whereas SIFT-indel describes it as having a neutral (i.e. benign) effect. Without any further evidence, this variant must be classified as being of uncertain significance (a VUS in medical terms). It is not uncommon for in-silico tools to provide differing predictions and the ACMG guidelines suggest that they can only be used to provide evidence for the deleterious nature of a deletion when multiple lines of in-silico investigation provide the same result.

It is also worth considering that a variant that may be tolerated in the context of a second working copy of the gene may also become more significant if the second copy harbours a null variant. Without going to the extent of setting up two further cell lines, each containing the specific mutation this is impossible determine for sure.

The decision was taken to use this cell line for study; with the frameshift variant present, it is almost certain that CHD2 function will be perturbed. It is possible that the second variant may render this perturbation more severe – this would have the advantage of making our results easier to interpret, but the disadvantage that it is a less accurate model of the heterozygous mutations seen in patients with CHD2 related EECO. In either case, findings reported in the PhD thesis will need to be verified in tissue samples from patients with this condition.

15	93467560	p.Ala24Ala	PASS	synonymous	1	119054	0	0.000008400	
15	93467574	p.Ser29Leu	PASS	missense	2	119734	0	0.00001670	
15	93467575	p.Ser29Ser	PASS	synonymous	6	119812	0	0.00005008	
15	93467586	p.Ser33Leu	PASS	missense	1	120340	0	0.000008310	

Figure 4.17: variants found in human population studies which fall within the same domain affected by the in-frame heterozygous variant identified in CHD2 gene-editing experiment

4.4.2 Validating neurodifferentiation with RNA-Seq

Two questions need to be answered when considering the data presented here; are we confident that the differentiation of our iCn-WT and our iCn-CHD2^{+/-} cell lines follows the expected timing – changing from pluripotent stem cells at D0, to NPCs at D19 and mature neurons at D40? The second is whether we can find any difference in the transcriptomes between iCn-WT and iCn-CHD2^{+/-} cells at D0, D19 and D40 that might provide clues as to the pathophysiology underlying the epilepsy phenotype seen in patients with heterozygous mutations in CHD2?

The first question is perhaps more important for this thesis; I will present data from two further experiments across the next two chapters, with data collected at D0, D19 and D40 of neurodifferentiation. It is important therefore to establish exactly what cell populations are being investigated.

At D0 the cells are expected to be hiPSCs in a pluripotent state. We have good transcriptional evidence that this is the case; in both cell lines OCT4 and SOX2 transcription is elevated. It is worth noting also that although NANOG was transcribed at a significantly lower level than the other pluripotency markers (*figure 4.7, table 4.5*), the transcription levels at D0 for NANOG were considerably higher than those at D19 or D40.

At D19, the cell cultures are expected to have formed NPCs. SOX2 remained elevated in both cell lines as expected, however there was some discrepancy in the transcription of PAX6, which was significantly higher in the iCn-WT cells than in the CHD2 cells. The co-transcription of PAX6 and SOX2 in concordance with a fall in OCT4 and NANOG confirms that the iCn-CHD2^{+/-} cells are behaving as NPCs, however it is possible that there is some attenuation of this signalling cascade in the CHD2 deficient cells.

The presence of cells transcribing MAP2 at this stage suggests that some final differentiation into neurons had already begun at D19. The MAP2 level transcription levels were, however, considerably lower than at D40.

At D40, the cells are expected to be neurons approaching maturity. This is confirmed by the massive upregulation of MAP2 seen in both cell lines. The upregulation was almost twofold higher in CHD2 than in the iCn-WT cells. The iCn-WT cells were still expressing PAX6 at a comparable level to the D19 cultures and the CHD2 cells were still expressing SOX2. Both lines had upregulated PSD95 production to greater levels than in NPCs.

4.4: Discussion

The combination of upregulated MAP2 and PSD95 confirms that there are mature neurons in this cell culture, however the continued transcription of the NPC markers suggests that there were still a population of cells behaving as NPCs. Whereas at D19, NPC markers were predominant, this pattern had reversed at D40. This suggests that D19, although mature neurons may have been present, NPCs were the main constituent of the cultures and that by D40 mature neurons were the dominant cell.

The higher levels of transcription of neuronal markers MAP2 and PSD95 in iCn-CHD2^{+/-} cells compared to iCn-WT cells could indicate that the iCn-CHD2^{+/-} cells are reaching maturity and exiting the cell cycle earlier than the iCn-WT cells. This finding must be interpreted with caution, however; CHD2 mutations are expected to disrupt the transcriptome on a genome wide scale and rather than indicating a greater number of mature cells in the culture, the differing levels of transcription could simply be a feature of the underlying distortion of the transcriptional profile – to put it another way; the same ratio of mature neurons but producing a higher concentration of marker transcripts. In either event, these findings hint at a difference in the way transcription of neurospecific genes proceeds during neurodevelopment in the context of CHD2 mutations.

To summarise; the RNA-seq data provides adequate evidence that D0, D19 and D40 cell cultures are behaving in the transcriptional manner expected for IPS, NPC and neuronal cells respectively. Based on the results described here, it is likely that the D19 cultures contain some mature neurons and that the D40 cultures contain some NPCs. The change in the ratio of these transcription factors is strongly suggestive that NPCs are the majority cell type at D19 and neurons at D40.

With this knowledge, we can confirm that D19 and D40 cultures are reasonable tools for investigating DSB repair in NPCs and mature neurons respectively.

It is tempting to draw conclusions as to the rate of maturation based on the significantly higher enrichment for *MAP2* and *PSD95* transcription in the *CHD2* mutant cell lines. Future work, requiring detailed protein analysis is necessary to confirm such a change in the timing of differentiation as the data presented here is inconclusive.

4.4.3 Insights from whole transcriptome analysis

At all three timepoints, there were multiple GO terms both enriched and depleted in iCn-CHD2^{+/-} when compared to iCn-WT. In this section I will address consistencies and inconsistencies across the entire differentiation and highlight the most interesting differently expressed terms regarding our understanding the clinical phenotype seen in heterozygous CHD2 mutations.

At all three timepoints, there were GO terms that occur in both the enriched and depleted groups. For example, at D0, GO:0044767 (developmental process, defined as "A biological process whose specific outcome is the progression of an integrated living unit: an anatomical structure (which may be a subcellular structure, cell, tissue, or organ), or organism over time from an initial condition to a later condition.") was both enriched ($p=4.02 \times 10^{-6}$) and depleted ($p=2.37 \times 10^{-6}$).

At first glance, this may appear paradoxical, however it can be easily explained when you consider that it functions as a parent term, encapsulating multiple other GO terms that include both 'positive regulation of developmental process' and 'negative regulation of developmental process'. It is also worth keeping in mind that this term applies to 6346 genes. It is perfectly reasonable to assume that some are upregulated and some downregulated to an extent that the parent term scores highly in both enriched and depleted lists.

It is also worth interpreting these lists with the approach that both enriched and depleted genes could have a detrimental effect on end organism development – in this case CNS development. The biochemical cascades governing in-utero development are complex and finely tuned – a mutation that creates any detectable disruption in the transcription of these genes is likely to have significant impacts on outcome.

GO:0044767 was in fact, in the top 20 most significantly upregulated and downregulated terms at D0 and at D19, however does not appear in the 20 most downregulated terms at D40.

Given the large number of genes and daughter terms included under the umbrellas of these parent terms, it is more useful to look for patterns in more specific terms that fall under these broader headings. At D19 GO:0072091 (regulation of stem cell proliferation, 88 genes) was depleted compared to WT cells (9.57×10^{-5}). Also of potential interest is GO:0006749 (glutathione metabolic process, 58 genes), which was depleted (6.97×10^{-5}). Glutathione metabolism has been implicated in the CNS response to toxins and in autistic spectrum

4.4: Discussion

disorder pathogenesis [317] – given the increased rates of ASD in CHD2 mutations this depletion is worthy of further exploration.

At D40, forebrain neuron differentiation GO:0021879 (forebrain neuron differentiation, 74 genes) was depleted in iCn-CHD2^{+/-} cells ($p=2.47 \times 10^{-6}$), as was GO: 0001709 (cell fate commitment, 287 genes) ($p=2.71 \times 10^{-7}$).

Taking a general view, the representation of GO terms related to stem cell development, neurodevelopment and transcriptional regulation, particularly at D19 and D40 of differentiation provide many further avenues for potential exploration within this dataset.

Regarding DSB repair and DNA repair in general, terms relating to these were not enriched or depleted at any point in the neurodifferentiation. This does not rule out a role for these pathways in the pathogenesis of CHD2 related EECO, but it does indicate that further transcriptome analysis is unlikely to elucidate changes in these pathways in this context.

Comparing our list of upregulated and downregulated GO terms to those previously published in a paper using siRNA knockdown of CHD2 cells [4], there does not appear to be any direct overlap. This paper identified upregulated GO terms relating to immune-system function (interferon response, MHC class switching etc), and depletion in developmental process related GO terms, including nervous system development ($p=1.6 \times 10^{-47}$).

In one point of overlap this paper demonstrated changes in interneuron development related to CHD2 knockdown and genes related to GO:1904936 (interneuron migration, 6 genes) were depleted in our D40 cell iCn-CHD2^{+/-} culture.

It is uncertain whether these datasets are directly comparable. The paper in question focused specifically on interneuron development and used a different protocol for neurodifferentiation. It is also worth noting that, given our cell line potentially has up-to 50% CHD2 function remaining (depending on the deleteriousness of our in-frame deletion), we may be attempting to compare a homozygous model with a heterozygous model.

4.4.4 Summary and conclusion

The data presented in this chapter demonstrates the creation of a mutant CHD2 line, in which one allele is a confirmed knockout via frameshift deletion of 11bp and one allele is a 6bp knockout of uncertain impact.

It has been demonstrated that the iCn-CHD2^{+/-} cells transcribe cell lineage markers in the expected pattern for hIPSCs at D0, NPCs at D19 and neurons at D40. There were some differences in the pattern of marker transcription between the WT and mutant cells, however both lines expressed the expected pattern. It is not clear whether these discrepancies are the result of a change in the rate of maturation of the iCn-CHD2^{+/-} cells, or the result of a wider transcriptional dysregulation. In either event developmental timing does not appear to have changed.

What we do have evidence for is a change in the read depth of some transcriptional modules that might influence CNS development. Gene ontology clustering reveals both enrichment and depletion in a wide range of terms and genes related to normal cell development, metabolism and neurodifferentiation. It is possible that these imbalances could contribute towards disorders of neurodevelopment in human beings and hence contribute towards the human phenotype in heterozygous CHD2 mutations.

The focus of this project remains on the contribution of CHD2 towards DSB repair, however this RNA-seq dataset merits further investigation in the future. In particular, analysis for genes enriched in epilepsy, mood disorder, schizophrenia and autistic spectrum disorder may reveal important clues as to the pathogenesis of CHD2-related EECO.

As well as confirming that our D19 and D40 cells are sufficient analogues for NPCs and mature neurons respectively, this dataset has potential relevance to the interpretation of our results in *chapter 6*. This chapter aims to take a whole genome approach to identifying DSBs in culture and it is theorised that an increased rate of transcription is likely to lead to an increased rate of DNA damage, through mechanisms that will be explored in *chapter 6*. One of the aims for *chapter 6* will be to relate the transcription rates found in our RNA-Seq data to the occurrence of DSBs at highly transcribed sites.

5: MODELLING DSB REPAIR IN CHD2 DEFICIENCY USING TARGETED GENOME EDITING

5.1 Introduction

In *chapter 3*, we discussed the basics of Cas9 genome editing and demonstrated the set-up of a cell line which produces Cas9 protein in response to doxycycline stimulation. We demonstrated the creation of new mutant cell lines using single gRNAs and gRNAs translocated against 2-3 targets within the same exonic region. We also demonstrated the use of nanopore sequencing using 1D and 1D² chemistry to detect and quantify these new mutations.

In *chapter 4*, this cell line was used to create a hiPSC line containing a compound heterozygous mutation in CHD2; one allele containing an 11bp frameshift mutation, the other containing a 6bp in-frame of uncertain but likely neutral effect. This cell line, termed iCn-CHD2^{+/-} was characterised by RNA-Seq at D0, D19 and D40 of differentiation according to the neurodifferentiation protocol described in *section 2.3.2*.

We can now proceed to use this cell line and the inducible Cas9 and ONP sequencing pipeline to investigate the impact of CHD2 mutations on DSB repair; the primary aim of this project.

5.1.1 Overview of experimental approach

In this chapter, we will examine the rate of mis-repair and the character of mutations that occur when making targeted DSBs using inducible Cas9. To do this, we will lipofect each cell line with 2-3 gRNA targeting the same 1kb genomic region and analyse the results of these transfections, similarly to the approach described in *chapter 3*.

The choice to use multiple gRNA rather than single gRNA is important for several reasons. First; the use of multiple gRNA increases the chance that at least one of the gRNA will make a cut that results in a mutation[241]. Secondly; as well as allowing us to examine the repair characteristics at a single cut site, using multiple guides allows us to consider repair dynamics.

If a perfect repair occurs then provided the gRNA-Cas9 complex is still present, the strand can be broken a second time. Once an imperfect repair is made the gRNA is no-longer able to recognise the target strand and no further cutting is possible. To put it another way both the DSB created by Cas9 and its perfect repair are reversible in the continued presence

5.1: Introduction

of the Cas9 complex but the creation of an indel is not. Although the dynamics of this pathway are not fully understood there is evidence that they are influenced by the repair pathway used to heal the DSB.

Breaks at the same target in the same cell line have been demonstrated to have reproducible results, Chakrabarti et al[258] studied CRISPR Cas9 editing at 1248 target sites across 450 genes, 649 of which were successfully edited. They noted that some targets were prone to more reproducible (or 'precise') editing than others, with the same indel occurring in a high proportion of reads. They also demonstrated that these more 'precise' edits are prone to a higher editing efficiency – that is a higher success rate for the induction of mutations. They also noted that the mutations in precise targets were more likely to be insertions (up to 80%) whereas the mutations at imprecise target sites were more likely to be deletions (up to 64%).

At the time these experiments were conducted and at the time of writing, these findings had not yet translated into reliable tools to predict the effect of mutations on a specific targets and so a range of targets was picked within our chosen genes, based on the predicted efficacy of the gRNA[232].

The same gRNAs will be used in both the iCn-WT and iCn-CHD2^{+/-} providing an effective control for the potential for variation in repair 'precision' as described in this paper.

The hypothesis states that an inhibition of NHEJ in the iCn-CHD2^{+/-} cells is expected, resulting in an increased utilisation of A-EJ. There is evidence that NHEJ is a faster repair pathway than A-EJ, meaning that each DSB should persist unrepaired for longer if NHEJ is inhibited.

In the experiment described in this chapter 2-3 gRNA were transfected simultaneously to target the same 150-500bp region of a gene. If each DSB made persists for longer, then the chance of a second DSB occurring before the first is repaired is increased. *Chapter 3* describes the outcome of such double cuts – the intervening DNA is lost and an indel is created. In order to distinguish these mutations from smaller indels created by mis-repair of single cuts, they are referred to in this chapter as Double Cut Excisions (DCEs).

The DCEs are predictable in size. They are expected to be the length between the cut-site of the two gRNA-Cas9 complexes that create the DSBs at each end (*figure 5.1*).

Inhibition of NHEJ is known to lead to slower repair of DSB by A-EJ or HR. Delayed repair of DSBs increases the chance of two DSBs existing simultaneously in this model system,

5.1: Introduction

and therefore is theorised to increase the rate of DCEs. Therefore, if heterozygous CHD2 mutations cause an inhibition of NHEJ, we expect to detect this as an increased rate of DCE when multiple gRNA are transfected simultaneously against the same region.

The other approach taken in this chapter is to consider mutations that occur as a result of mis-repair of a single DSB; identified as insertions and deletions localised to a single cut-site that do not span the distance between two cuts (*figure 5.1*). For the avoidance of confusion, these will be referred to as indels, as opposed to the DCEs created by multiple gRNA cutting simultaneously.

If the hypothesis proposed in *section 1.7* is correct, we would expect to see either a change in the size, depth or position relative to the cut-site of the indels between iCn-WT cells and iCn-CHD2^{+/-} cells, or a change in the relative rate of DCEs.

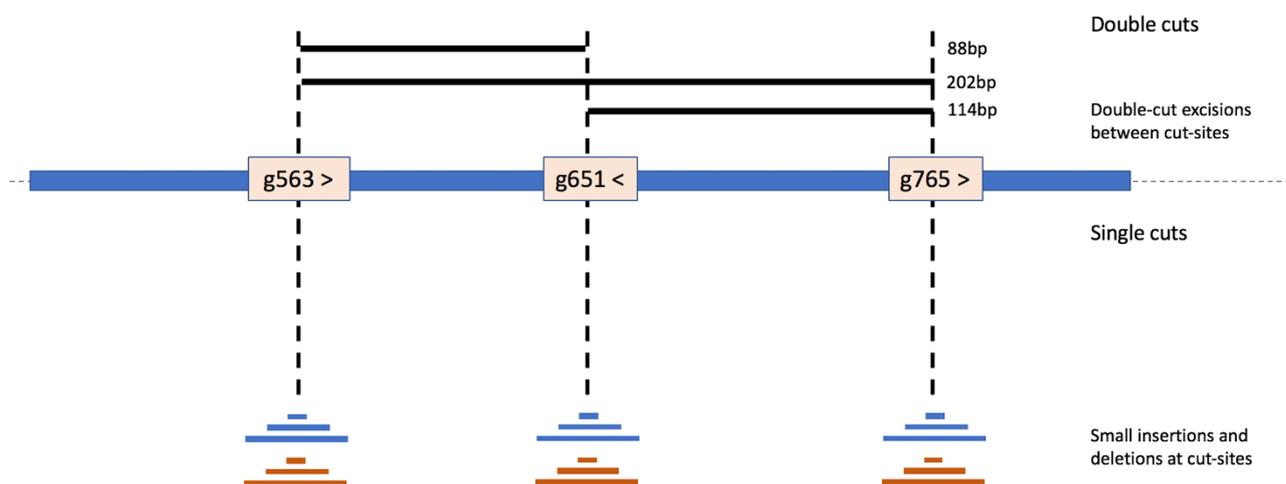


Figure 5.1: Schematic representation of DCE (top) and indels (bottom) expected to be found when transfecting multiple gRNA against the same region of DNA. The numbers of the cut-sites denote the locus of the cut as measured from the start of the amplicon to be analysed.

> denotes a gRNA with PAM site upstream of the gRNA and < denotes a PAM site downstream of the gRNA.

5.1.2 Aims

- 1) Compare the depth of large indels created by transfection of multiple gRNA against the same target between WT cells and cells with a mutation in CHD2
- 2) Compare the depth and length of smaller indels caused by a single gRNA between WT cells and cells with a mutation in CHD2

5.2 Methods

5.2.1 Choice of targets and gRNA design

Long, late replicating genes are known to be disproportionately susceptible to endogenous strand breakage[152]. It is also thought that CRISPR efficacy is in part related to the transcriptional status of a gene, with actively transcribed genes being more susceptible to editing due to euchromatic status conferring a favourable microenvironment for crRNA binding and Cas9 activity.

Gene lengths were ranked using a custom script to access the national centre for biotechnology information (NCBI)'s application programming interface (API). From the list of the longest genes, candidates were selected that are known to be expressed in mature neurons according to the data found in the human protein atlas.

A list of six target genes was chosen; AUTS2, CSMD1, CSMD3, GRID, NRXN1 and PARK2. An initial sequencing run was performed on WT DNA from these targets to ensure that different amplicons could be successfully demultiplexed. Of these genes, GRID and CSMD1 had poor alignment accuracies and depths, despite appropriate results on gel electrophoresis for PCR barcode optimisation. AUTS2, CSMD3, NRXN1 and PARK2 were therefore chosen as targets, with the intention of targeting multiple gRNAs against one target per culture well.

As these cells were not intended to survive for use in further work, when designing gRNAs the on-target scores were prioritised, regardless of how poor the off-target score was[232].

5.2.2 Differentiation

iCn-WT and iCn-CHD2^{+/-} cells were grown on separate 6 well plates, coated with matrigel. Once the cells were 80% confluent, they were passaged onto separate 6 well plates, coated with reduced growth factor matrigel. At the same time, a 24 well plate coated with matrigel was seeded from the same passage.

The cells seeded to the 24 well plate were used as D0 cells. The cells seeded to the 6 well plate were differentiated into mature neurons using the protocol described in *section 2.3.2*.

5.2.3 Gene editing

The strategy for genome editing described in *section 2.9* was used for genome editing. Multiple guides were used for each target region (see figure 5.1) and so the total gRNA volume was split between two or three gRNAs against the same genomic target for each experiment (see *table 5.1*).

Target	Guide RNA name	Volume in Lipofection mix
AUTS2	AUTS2_363	3.3µL
	AUTS2_451	3.3µL
	AUTS2_565	3.3µL
CSMD3	CSMD3_265	3.3µL
	CSMD3_326	3.3µL
	CSMD3_392	3.3µL
NRXN1	NRXN1_1278	5µL
	NRXN1_1759	5µL
PARK2	PARK2_377	3.3µL
	PARK2_438	3.3µL
	PARK2_488	3.3µL

Table 5.1: gRNAs used to generate DSBs for comparison of repair, and volume lipofected

5.2.4 DNA extraction and sequencing

DNA was extracted for sequencing 48 hours after the lipofection as per protocol (see *section 2.5.1*).

Separate PCR reactions are performed against each target amplicon. The three most successfully de-multiplexed barcode combinations were identified by combining the read counts of all previously de-multiplexed FASTQ files and generating a heatmap. These were assigned at random to the Wild-Type Treated (iCn-WT⁺), Wild-Type Negative control (iCn-WT) and CHD2 Treated (iCn-CHD2^{+/-}) cell lines.

Two stage PCR was conducted as described in section 2.5. Analysis of previous runs determined that barcodes C3, C5 and C7 had a high number of reads successfully demultiplexed across previous runs, with relatively even number of reads per barcode. (*figure 5.2*).

1D² library preparation was carried out. The samples were sequenced on the Oxford Nanopore Minion 1DSQ (MIN107) flow cell. Reads were base called using Albacore 1D and 1DSQ base calling.

Base-called reads were demultiplexed using porechop, with an accuracy / threshold setting of 75%/3% (see *chapter 3*). The demultiplexed reads were aligned to the reference sequences with minimap2. For overview of the bioinformatic pipeline please see *figure 5.3*.

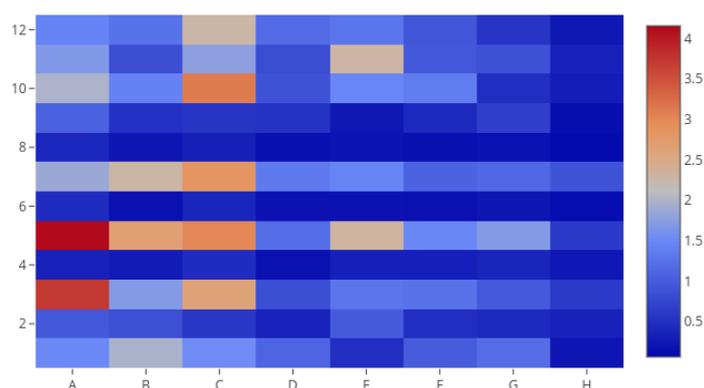


Figure 5.2: 2. Results from previous demultiplexing experiments, with number of reads successfully demultiplexed presented as a ratio to the average number of reads demultiplexed for all experiments. Dark red squares indicate the barcode combinations with the highest number of reads demultiplexed. C3, C5 and C7 were chosen for use in iCn-CHD2^{+/-}, positive control iCn-WT and negative control iCn-WT cell lines respectively

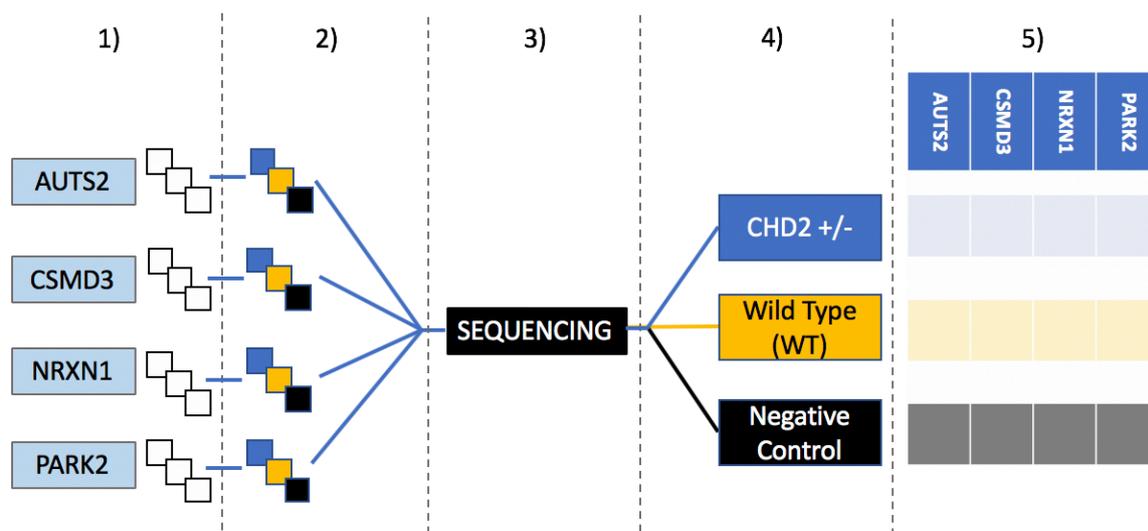


Figure 5.3 – Workflow for experimental comparison of DSB repair. All experiments are pooled and sequenced on the same sequencing run with bioinformatic demultiplexing performed afterwards.

- 1) Primary PCR and M13 tagging, 2) Barcode PCR to tag WT+, WT- and CHD2^{+/-} with barcodes C3, C5, C7 respectively, 3) samples pooled and sequenced, 4) demultiplexing with porechop, 5) demultiplexing by alignment to separate reference sequences

5.2.5 Analysis

5.2.5.1 Comparison of DCE

Using the locus of each PAM site, it is possible to predict how long in nt any DCEs will be for each target (*table 5.2*). Note that as Cas9 cuts 3nt upstream of the start of the PAM site, if gRNA face in opposite directions, the DCE will be 6nt longer than if they target the same strand.

The number of DCEs of each predicted size will be compared. As the sequencing depth is expected to be different for each condition of each target, the counts will be normalised using the mean of the read counts across all conditions.

To determine if any of differences in the rate of occurrence of DCE are statistically significant, proportion testing will be used.

GENE	gRNA 1	Direction	Cut site (refpos)	gRNA 2	Direction	Cut site (refpos)	Predicted DCE size (nt)
AUTS2	AUTS2_363	FWD	360	AUTS2_451	REV	454	94
	AUTS2_363	FWD	360	AUTS2_565	FWD	562	202
	AUTS2_451	REV	454	AUTS2_565	FWD	562	120
CSMD3	CSMD3_265	FWD	262	CSMD3_326	FWD	323	61
	CSMD3_265	FWD	262	CSMD3_392	REV	395	133
	CSMD3_326	FWD	323	CSMD3_392	REV	395	72
NRXN1	NRXN1_224	REV	227	NRXN1_705	FWD	701	474
PARK2	PARK2_377	FWD	374	PARK2_438	REV	441	67
	PARK2_377	FWD	374	PARK2_488	FWD	485	111
	PARK2_438	REV	441	PARK2_488	FWD	485	56

Table 5.2 : expected size of DCEs for each gRNA cut site pair. Note that Cas9 creates the DSB between the 3rd and 4th nucleotide upstream from the PAM site

5.2.5.2 Comparison of Indels

Using the bam_readcount file two python dictionaries will be generated, one each for deletions and insertions. These will contain data regarding the length and read-depth of every indel which occurs 10bp upstream and downstream of each cut-site, in the format {locus {length of insertion/deletion: depth of insertion / deletion}}.

The depth of the each indel will be normalised according to the ratio of total read counts between iCn-WT and iCn-CHD2^{+/-} cell lines for each alignment, taken from the samtools stats output. Heatmaps will be generated from this normalised data with matplotlib's pyplot module and the seaborn graphical data package for python.

To determine if significant differences exist between the depth of indels of each size at each position, proportion testing will be used, with matrices containing the number of aligned reads which do not contain the mutation, and the un-normalised number of reads containing the mutation. The scipy stats package's Chi-square contingency function will be used to generate p values and a list of mutations with statistically significant changes between iCn-WT and iCn-CHD2^{+/-} mutant lines will be output.

5.3 Results

5.3.1 Alignment and exploratory analysis of indel counts

The number of reads successfully aligned for iCn-WT and iCn-CHD2^{+/-} cells can be found in *tables 5.4 and 5.5* for day 0 and day 40 of neurodifferentiation respectively.

Despite two attempts at PCR amplification and sequencing, no reads from the D40 sample could be successfully aligned to the AUTS2 reference sequence at D40. For the other targets, data from both sequencing runs was combined. The first library produced very few useable reads for any target except CSMD3, which is therefore represented at much greater depth in the D40 dataset.

Although there were clear differences between the number of reads aligned to each reference sequence for both differentiation timepoint and cell type, except for analysis of AUTS2 at D40, this did not affect the planned analysis. As described in the methods, differences in the number of reads for each target are expected and factored into analyses that either use normalisation to a mean based on readcount or statistical proportion testing which can account for comparisons between two groups of different sizes

The normalised counts of all indels of various lengths recorded in each sample can be seen in *figure 5.4* for D0 of neurodifferentiation and *figure 5.5* for D40. It is worth noting that these scatter charts include all sequencing artefacts as well as genuine indels and DCEs created by our experimental pipeline.

All samples at D0 and D40 had a large number of indels recorded of small size (1-8bp length) resulting in a sharp fall in the recorded numbers as the indel size increases. At D0 there were increases in the number of deletions visible in AUTS2, CSMD3 and PARK2, but not NRXN1. There were no immediately obvious increases in insertions of any lengths in any of the targets based on the scatter plotted data.

There were fewer obvious changes in the numbers of deletions of specific lengths in any of the three targets aligned in D40, except for CSMD3. Again, at D40 there are no immediately obvious increases of insertions of specific lengths on scatter plotted data, however there appears to be a greater spread of smaller insertions at D40 than in D0.

The significance of all the increases seen in the scatter plotted data as relating to predicted DCEs will be discussed in *section 5.3.2.1*. The reason for the lack of increases in

5.3: Results

NRXN1 will be addressed in *section 5.3.2.1*. The pattern of smaller indels and any significant changes between iCn-WT and iCn-CHD2^{+/-} cell lines will be addressed in *section 5.3.3*.

5: Modelling DSB repair in CHD2 deficiency using targeted genome editing

5.3: Results

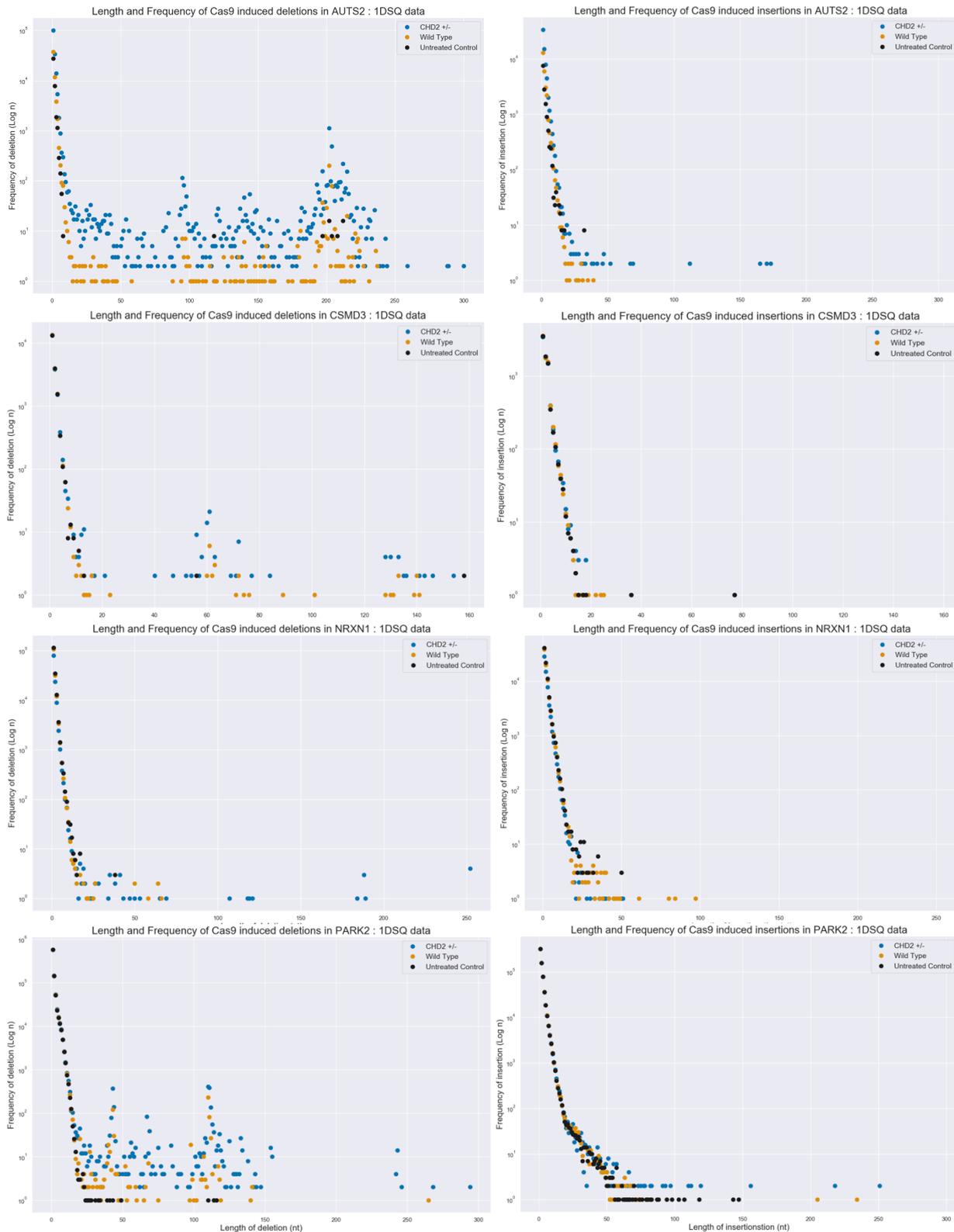


Figure 5.4. Scatter charts demonstrating indel counts by size of for four genomic targets in *iCn-WT* (orange) and *iCn-CHD2^{+/-}* (blue) cells with untreated controls (black) – from 1DSQ sequencing data at D0 of neurodifferentiation

5: Modelling DSB repair in CHD2 deficiency using targeted genome editing

5.3: Results

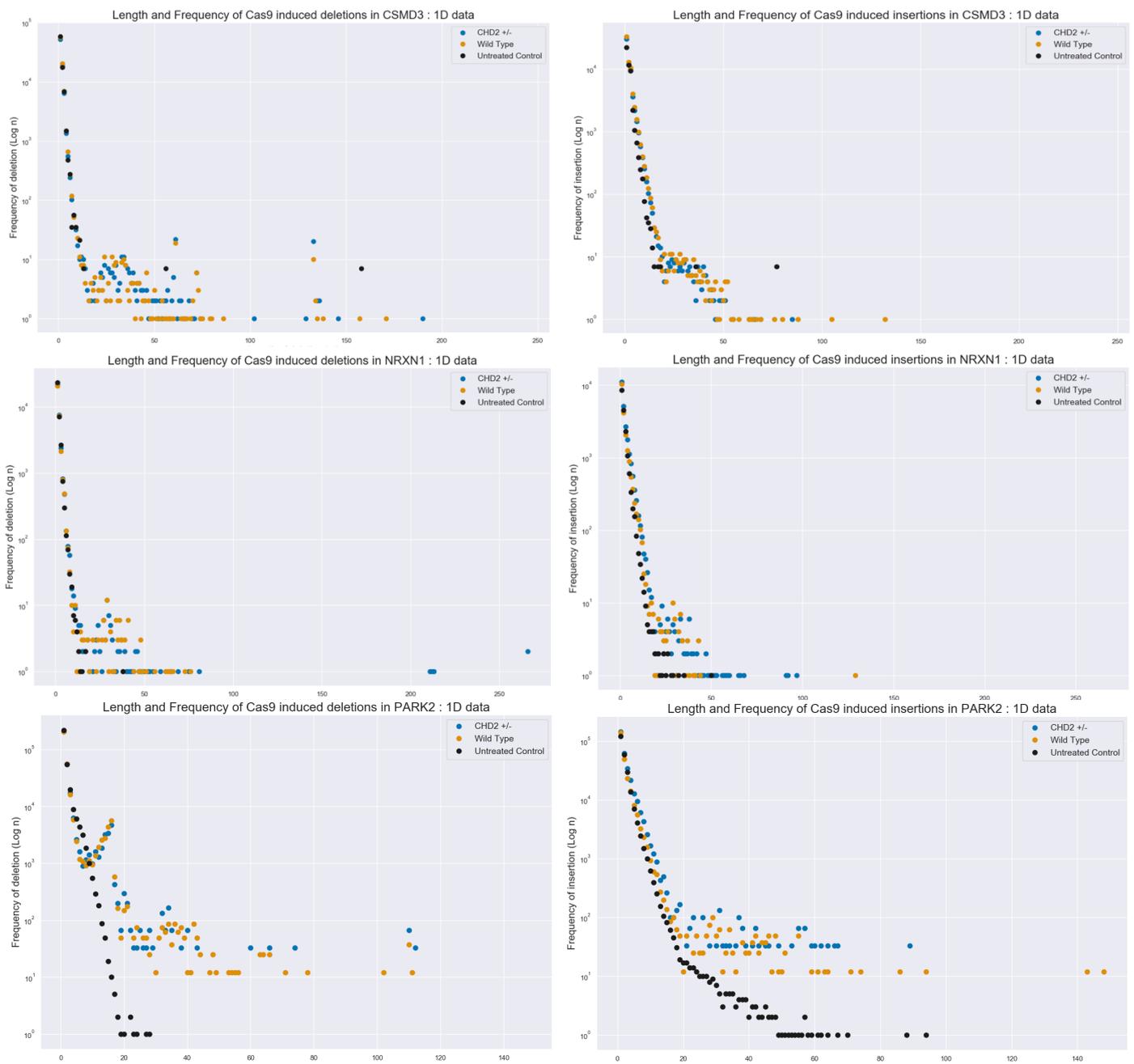


Figure 5.5. Scatter charts demonstrating indel counts by size of for three genomic targets in *iCn-WT* (orange) and *iCn-CHD2^{+/-}* (blue) cells with untreated controls (black) – from 1DSQ sequencing data, taken at D40 of neurodifferentiation

5.3.2 Double cut excisions

5.3.2.1 A note on the analysis of NRXN1

Before detailing the results of DCE counts, attention must be drawn to the special case of NRXN1, which proved resistant to analysis with the informatics pipeline used for the other targets. Initial exploration of the data from the D0 experiment (*figure 5.4*) indicated that there were no cuts of the expected size that would indicate DCE had taken place. Despite this, review of the read graph in igv demonstrated a clear loss of material between the two cut-sites (*figure 5.7*) in the experimental runs, but not in the untreated iCn-WT control.

It is suspected that the aligner, minimap2, did not record the deletion in a way that could be counted by samtools stats or bam_readcount, due to its size relative to the reference allele. The DCE size was 481bp; half the size of the reference sequence – it is suspected that this prevented the read being aligned and therefore prevented the deletion being recorded in the concise idiosyncratic gapped alignment report (CIGAR) string of the samfile. What was recorded, was the depth of the alignment between the two cut-sites. Using a modified version of the script for CRISPR_nanoscreen, described in chapter 3, the % change in the readcount could be calculated against the regions outside of the expected DCE region.

This method proved successful in determining the depth of the DCE (*table 5.3*), however it should be understood that a different method was used when analysing data from this part of the experiment.

5.3: Results

5.3.2.2 DCE at D0 and D40 of neurodifferentiation

At D0, in all four targets the CHD2 mutant line had a statistically significant increase in the number of DCEs of all predicted sizes (*table 5.4*) in the iCn-CHD2^{+/-} cell line compared to the iCn-WT cell line. In fact, the only predicted DCEs that did not achieve statistical significance were deletions of 133nt and 72nt size in CSMD3, both including a second cut with the same gRNA, CSMD3_gRNA_394.

As well as deletions of specific length used for chi-squared proportion testing to investigate significance, there were deletions of +/- 2-3bp length, particularly notable at some of the higher depth cut-sites. These were not incorporated into statistical analysis but can be visualised in *figure 5.5*.

At D40, the pattern was less clear. Although there were a higher normalised number of each predicted DCE in the iCn-CHD2^{+/-} cell line compared to the iCn-WT cell line, this increase only achieved statistical significance in one case; 133bp deletions in CSMD3 (*table 5.5*). Even in this case, the p value was orders of magnitude lower than the majority of those found in the D0 experiment.

Again, a visualisation of the normalised DCE counts can be found in *figure 5.5*.

5: Modelling DSB repair in CHD2 deficiency using targeted genome editing

5.3: Results

D0 of neurodifferentiation

D40 of neurodifferentiation

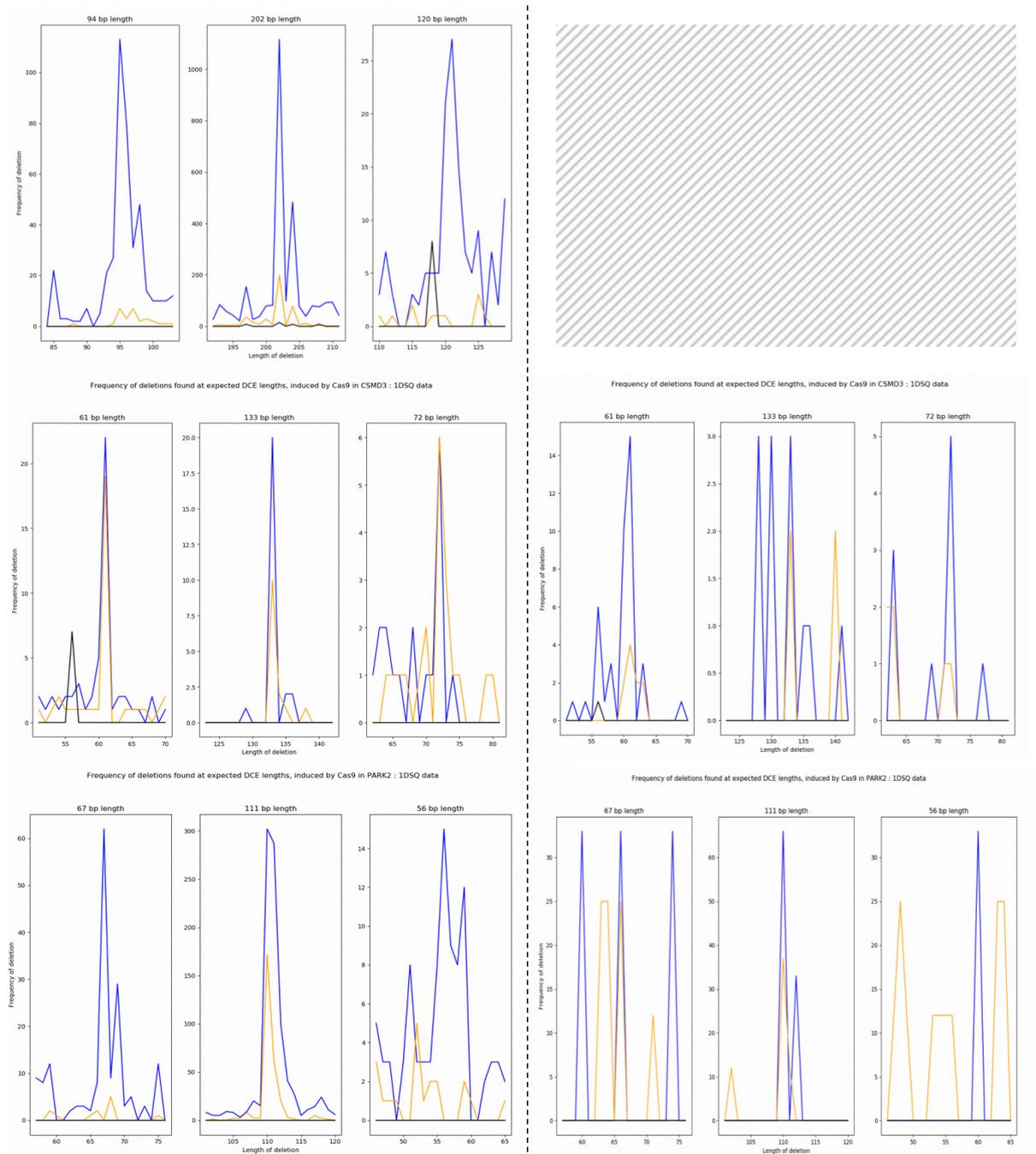
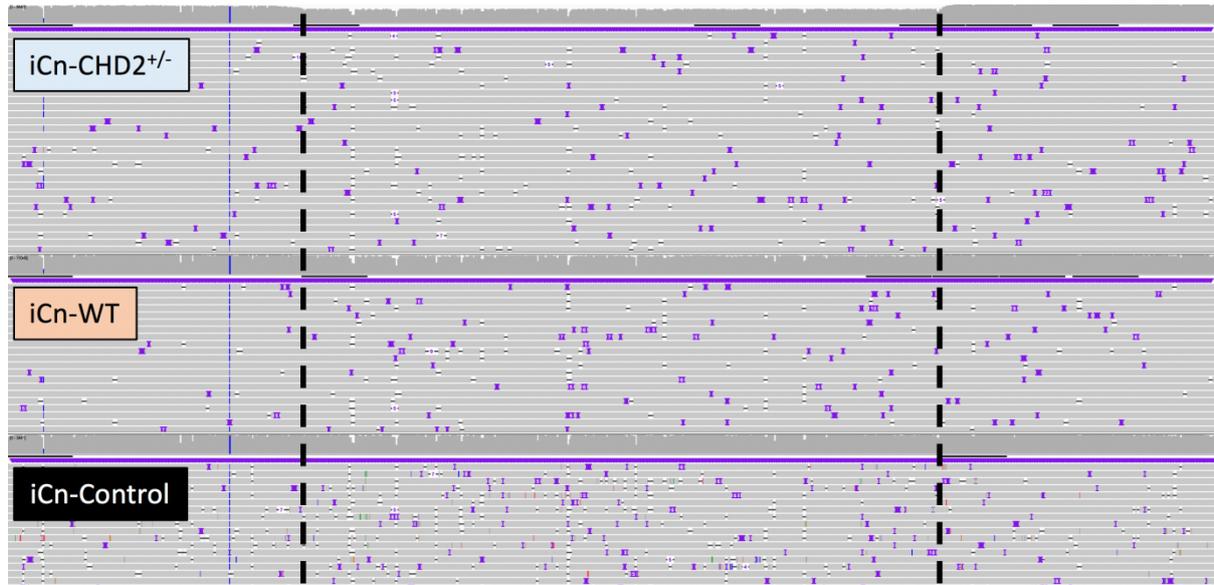


Figure 5.6 :Line graphs demonstrating depths of expected DCE +/- 10nt in mutation length at D0 and D40 of neurodifferentiation. Blue = CHD2^{+/-}, Orange = WT, Black = negative control. Note. AUTS2 sequences did not align from D40 samples and NRXN1 deletions were too long for minimap2 to identify as deletions – see figure 5.7 got NRXN1.

5: Modelling DSB repair in CHD2 deficiency using targeted genome editing

5.3: Results



GENE / cut	Day of differentiation	Cell line at D0	Mean read depth between cuts (n)	Mean read depth outside of cuts (n)	% of reads containing excision
NRXN1 / 22	D0	iCn-WT	10261.58	10815.85	5.12%
		iCn-CHD2 ^{+/-}	5914.81	7361.94	19.66%
		iCn-WT-Control	3381.73	3423.46	1.2%
	D40	iCn-WT	1306.71	1349.67	3.18%
		iCn-CHD2 ^{+/-}	3077.49	3202.14	3.90%
		iCn-WT-Control	1670.42	1679.14	0.52%

Figure 5.7 & Table 5.3: IGV readout and readcount comparison table demonstrating reduction in read depth between cut sites in NRXN1 at D0 and D40 of neurodifferentiation as proxy measure of deletions not identified by minimap2

5: Modelling DSB repair in CHD2 deficiency using targeted genome editing

5.3: Results

Day 0

GENE	Predicted DCE size (nt)	Read depth (observed DCE count) in iCn-WT	Read depth (observed DCE count) in iCn-CHD2 ^{+/-}	P value for difference between normalised DCE counts
AUTS2	94	20028(2)	3545(16)	3.784x10 ⁻¹⁷
	202	20023(7)	3504(57)	4.340x10 ⁻²⁸⁷
	120	20027(3)	3549(12)	2.678x10 ⁻¹¹
CSMD3	61	1981(4)	665(12)	0.003
	133	1981(4)	675(2)	0.981
	72	1982(3)	673(4)	0.135
PARK2	67	62,892(0)	311131(41)	4.22x10 ⁻¹⁹
	111	62,811(81)	30983(189)	1.24x10 ⁻³⁷
	56	62,892(0)	31161(10)	3.24x10 ⁻⁵
NRXN1	474	10,815.85(554)	7361.94(1447.13)	5.495x10 ⁻²⁰⁷

Table 5.4: Comparison of DCEs of predicted sizes found in iCn-WT and iCn-CHD2^{+/-} cells, with statistical significance, at D0 of neurodifferentiation

Day 40

GENE	Predicted DCE size (nt)	Read depth (observed DCE count) in iCn-WT	Read depth (observed DCE count) in iCn-CHD2 ^{+/-}	P value for differences between normalised DCE counts
CSMD3	61	7545(27)	13009(54)	0.608
	133	7558(14)	13015(48)	0.029
	72	7563(9)	13048(15)	0.897
PARK2	67	1974(0)	739(0)	n/a
	111	1974(0)	739(0)	n/a
	56	1977(0)	737(0)	n/a
NRXN1	474	1349(42.96)	3202.14(124.65)	0.283

Table 5.5 Comparison of DCEs of predicted sizes found in iCn-WT and iCn-CHD2^{+/-} cells, with statistical significance, at D40 of neurodifferentiation

5.3.3 Smaller indels at individual gRNA cut sites

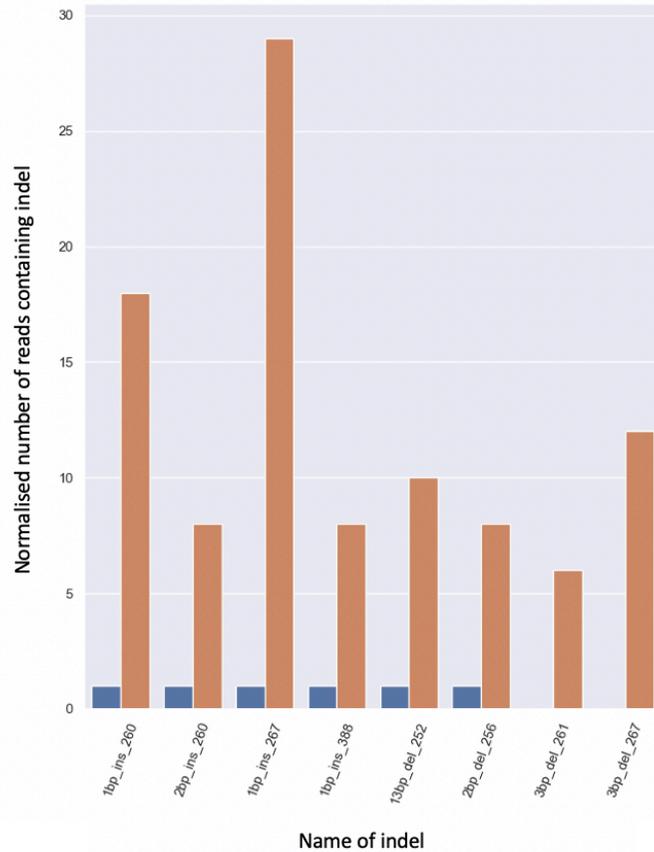
The range of indels of different lengths present within 10bp upstream and downstream of each gRNA cut-site were compared. The raw data is displayed in heatmap format in the supplementary materials section (APPENDIX III). The difference between the depths of reached statistical significance for a large number of these smaller indels and the normalised counts of the indels where statistical significance are displayed in graphical format in *figure 5.8*. Here, we will discuss general trends.

At D0 of differentiation, in CSMD3 and PARK2, the majority of these small indels were more common in the iCn-CHD2^{+/-} cell line than the iCn-WT line. The only exceptions are 1nt deletions and insertions at the same locus (264) in CSMD3, where the iCn-WT line was more commonly affected. In NRXN1, both deletions and insertions were higher in iCn-CHD2^{+/-} from locus 214-223, then higher for the remainder of calls in the WT cells.

The data for AUTS2 is not displayed graphically, as there were far too many significant counts to display in a readable graphical fashion without significant loss of data. The vast majority of the indels were more common in the iCn-CHD2^{+/-} line than the iCn-WT line. The pattern seen will be discussed in detail in section 5.4.

At D40, there were more insertions and deletions in the WT cell line than the iCn-CHD2^{+/-} in CSMD3 and NRXN1 for all indel sizes. The opposite was true for PARK2, with the depth of indels being far greater in the iCn-CHD2^{+/-} line.

Graph displaying indels in CSMD3 where there was a statistically significant difference in normalized count at D0 of neurodifferentiation



Graph displaying indels in CSMD3 where there was a statistically significant difference in normalized count at D40 of neurodifferentiation

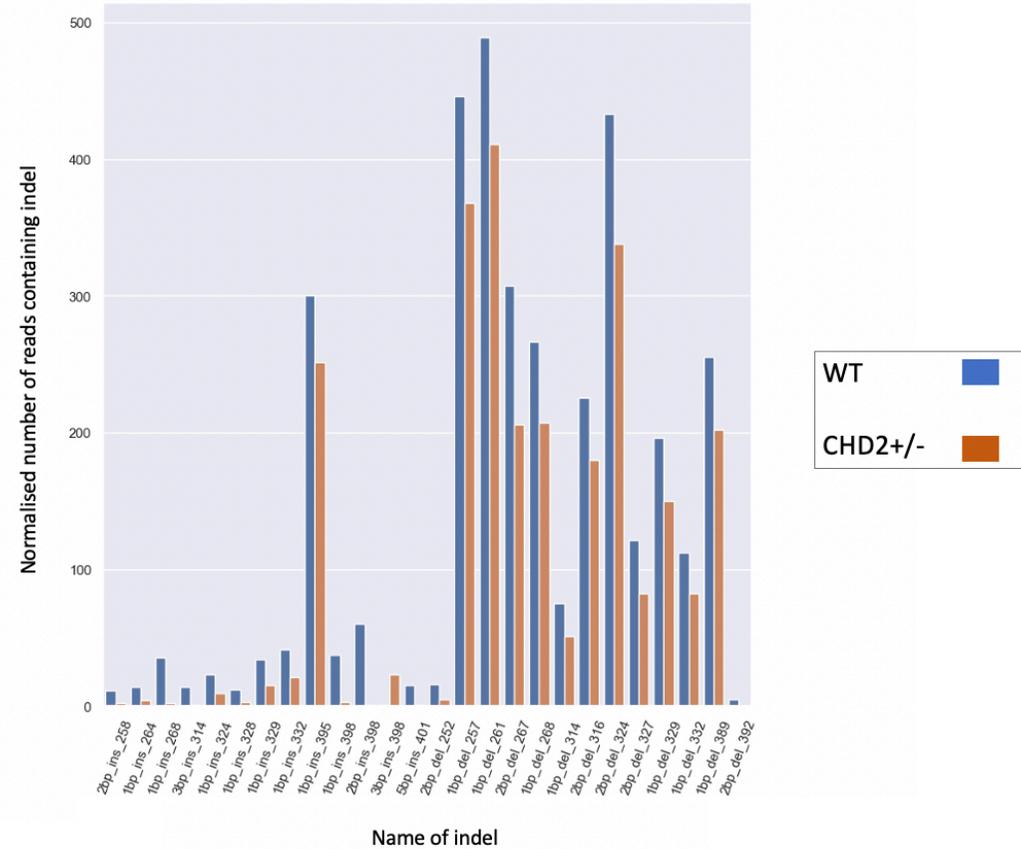


Figure 5.8.1: comparison of indels within 10bp of a gRNA cut-site in CSMD3 where a statistically significant difference existed in the normalised number of reads containing that indel in CSMD3 at D0 and D40. The name of each indel is the length of the indel(bp)_type of indel (ins = insertion, del = deletion)_ the locus of the indel as compared to the amplicon start site. At D0 all indels in CSMD3 were more common in CHD2 heterozygous mutants, however at day 40 of neurodifferentiation the pattern had reversed for all except one 3bp insertion at base 398

5: Modelling DSB repair in CHD2 deficiency using targeted genome editing

5.3: Results

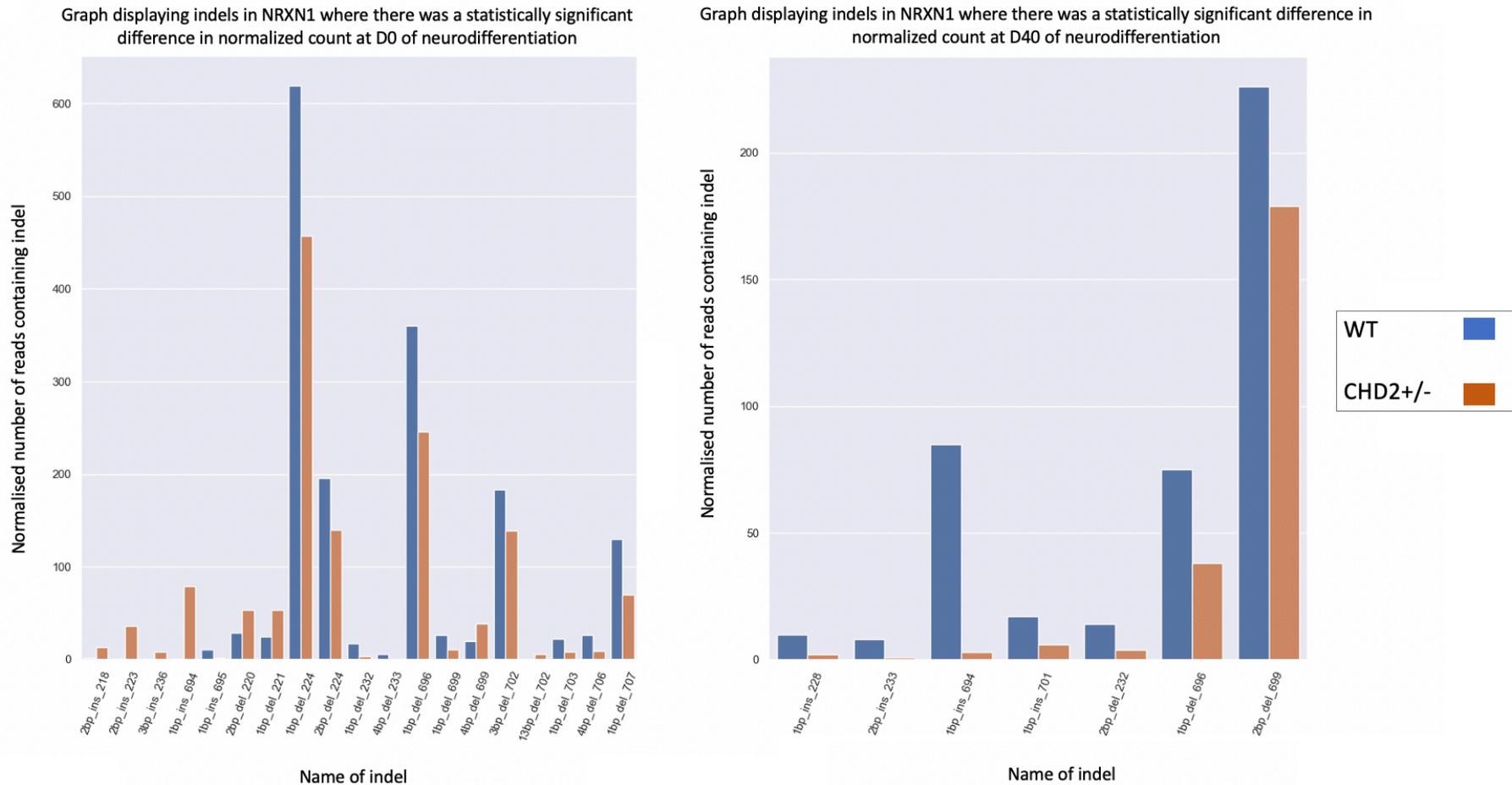
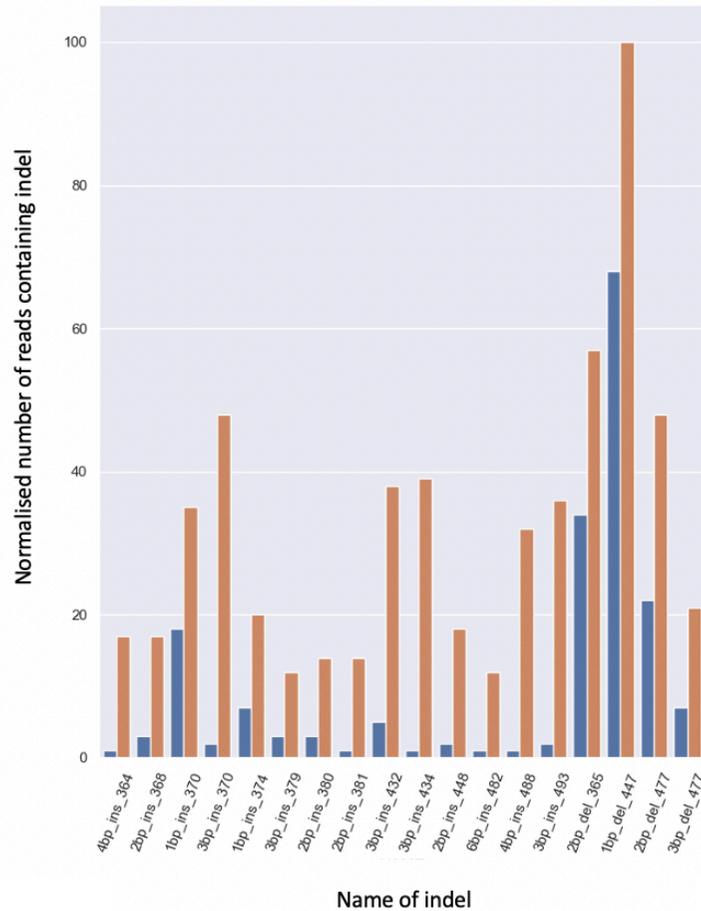


Figure 5.8.2: comparison of indels within 10bp of a gRNA cutsite in NRXN1 where a statistically significant difference existed in the normalized number of reads containing that indel in NRXN1 at D0 and D40. The name of each indel is the length of the indel(bp)_type of indel (ins = insertion, del = deletion)_ the locus of the indel as compared to the amplicon start site. At D0 some indels were more frequent in WT cells and some in CHD2 heterozygous mutants, whereas at of neurodifferentiation all indels where a statistically significant difference existed in frequency were more common in WT cells

5: Modelling DSB repair in CHD2 deficiency using targeted genome editing

5.3: Results

Graph displaying indels in PARK2 where there was a statistically significant difference in normalized count at D0 of neurodifferentiation



Graph displaying indels in PARK2 where there was a statistically significant difference in normalized count at D40 of neurodifferentiation

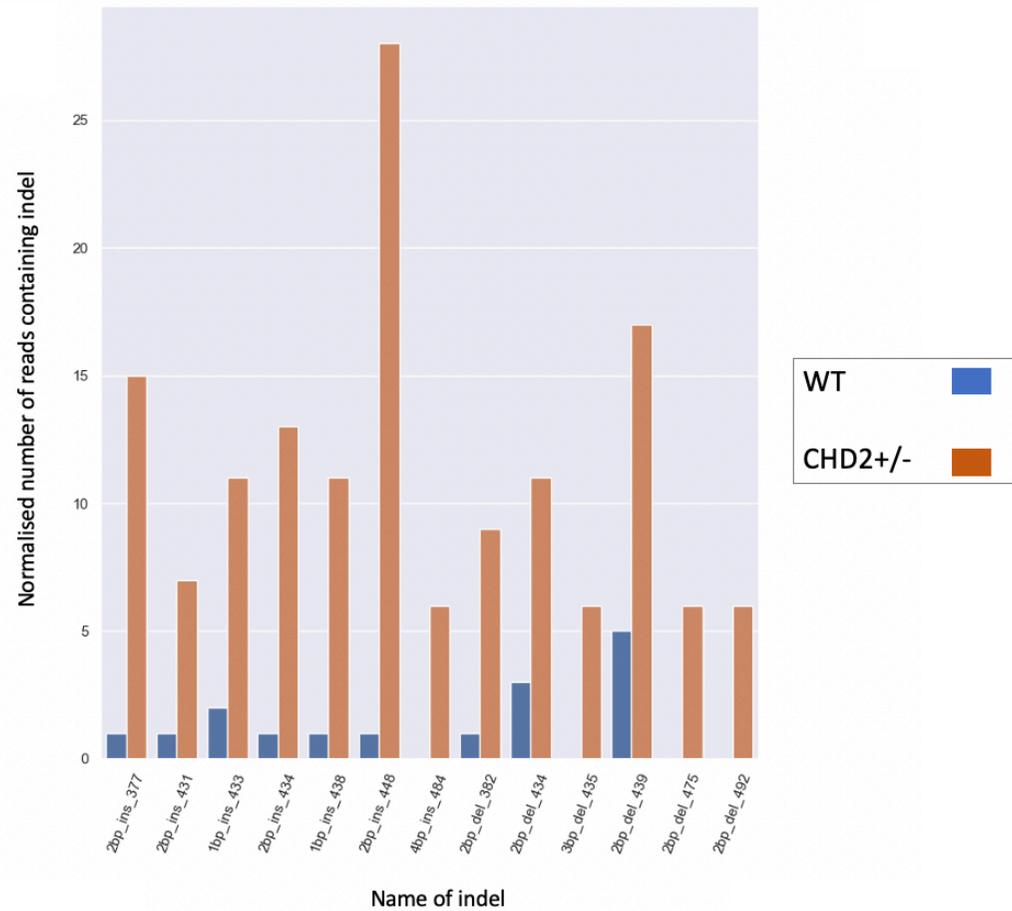


Figure 5.8.2: comparison of indels within 10bp of a gRNA cutsite in PARK2 where a statistically significant difference existed in the normalised number of reads containing that indel in PARK2 at D0 and D40. The name of each indel is the length of the indel(bp)_type of indel (ins = insertion, del = deletion)_ the locus of the indel as compared to the amplicon start site. At D0 and at D40, all indels for which a statistical difference existed in normalised frequency were more frequent in cells containing a heterozygous CHD2 mutation

5.3.4 Transcription at gRNA targets

Figure 5.9 demonstrates the normalised readcounts for AUTS2, CSMD3, NRXN1 and PARK2 at D0 and D40 of differentiation in iCn-WT and iCn-CHD2^{+/-} cell lines. These counts will be discussed in detail in relation to the previously described results in the discussion section (section 5.4):

- AUTS2 was expressed at D0 and D40, with significantly higher expression in iCn-WT cells at D0, and no significant difference in expression levels at D40.
- CSMD3 was expressed at a very low level at D0, with significantly higher expression at D40 – there was no significant difference between expression levels in each cell line at either time point.
- NRXN1 was more highly expressed in the iCn-WT cell line than it was in the iCn-CHD2^{+/-} line at D0, with equal expression at D40.
- PARK2 was expressed at a lower level at D0 than at D40 in both cell lines; at both these time points, there was no significant difference in expression levels. Even at D40, the expression levels of PARK2 were significantly lower than all other genes investigated.

It has been proposed that open chromatin structure leads to a greater accessibility to Cas9 protein and a greater editing rate. If the expression levels had been higher in CHD2 heterozygous mutant cells, it could be theorized that a more open chromatin conformation had predisposed to an increased editing rate, however this is not the case. We can therefore be reasonably certain that expression levels were not a significant influence on the results of this experiment.

Indeed, in the case of AUTS2 and NRXN1, the expression was significantly lower in the iCn-CHD2^{+/-} cells at D0. Despite this, the rate of DCE was higher in the mutant cell line than in the controls for both target

5: Modelling DSB repair in CHD2 deficiency using targeted genome editing

5.3: Results

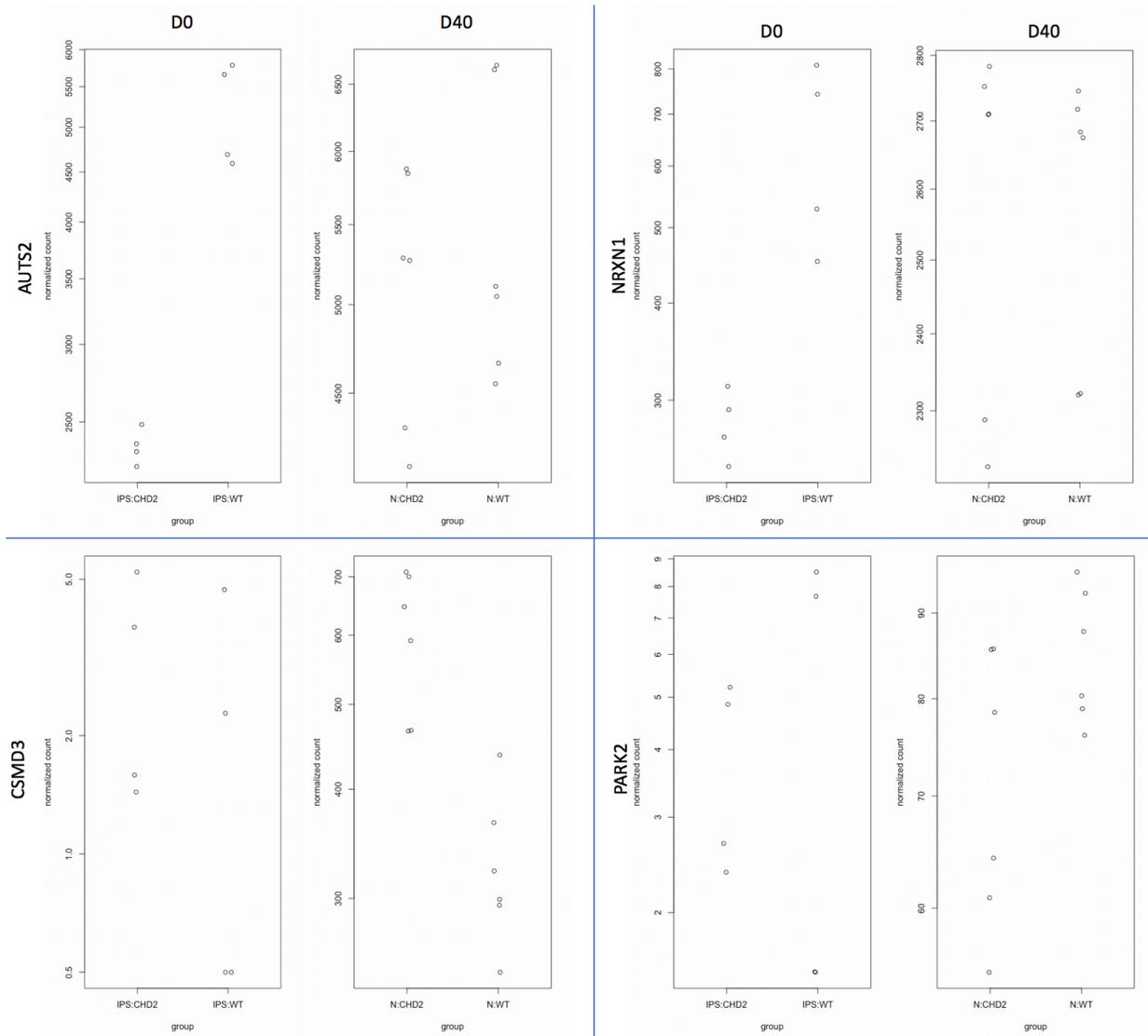


Figure 5.9 – Data from RNA-seq experiment described in chapter 4 demonstrating the RNA readcount of each CRISPR target gene at D0 (IPS) and D40 (N) of neurodifferentiation

Although differences exist in the level of transcription between the WT and CHD2 heterozygous mutant cells, there does not appear to be any relationship to the number of indels or DCE recorded.

For example, AUTS2 was less transcribed at D0 in the CHD2 mutant line, however a statistically significant increase in the number of DCEs was recorded there. Conversely, CSMD3 was transcribed at a higher level at D40 in the CHD2 mutant cells, but the DCE count was still higher than in the wild type cells.

This demonstrates that the increase in DCE seen in CHD2 mutant cells was not related to changes in the level of transcription of the target genes

5.4 Discussion

5.4.1 Implications of results

5.4.1.1 DCEs

The data presented in this chapter provides evidence that the repair of DSB induced by CRISPR-Cas9 is altered in cells with heterozygous CHD2 mutations. The results from hIPSC at D0 of neurodifferentiation are particularly convincing with statistically significant evidence of an increased rate of DCE in all targets analysed.

In all four of the selected targets at D0 there was a statistically significant increase in at least one of the predicted DCEs (pDCE). The effect was greatest in AUTS2 and NRXN1, but also highly significant in PARK2.

At D40, only one pDCE (61bp in CSMD3) met a statistical significance threshold of $p=0.05$, however differences in the 72bp DCE in the same gene approached a significant difference. The pattern for the 133bp pDCE in CSMD3 was entirely equivocal ($p = 0.981$).

The data from cells at D40 of neurodifferentiation is less striking, however the one statistically significant result conforms to the pattern seen at D0 and other results that do not achieve statistical significance nonetheless trend towards the same conclusion.

Taken together, the results from D40 of neurodifferentiation are less convincing than those at D0, however the only changes in pDCE rate were an increased in CHD2 deficient cells. Although the majority of changes in pDCE did not achieve statistical significance, taken with the D0 results they can be seen as supporting evidence of a change in DSB repair.

It is likely that the reduced rate of pDCEs observed in D40 is due to the inherent challenges of transfecting mature neurons. There is evidence that our gRNA did reach its target (in that some editing did occur), but it is probable that this occurred at a much lower level than in the IPSC experiment. Without being able to ensure identical transfection rates, a comparison between DSB repair at D0 and D40 is not possible.

5.4.1.2 Smaller indels

There was no such consistent pattern between targets for the smaller indels found at each cut-site (*figure 5.8*). Although there were statistically significant differences in the number of smaller indels of each size found at each locus between each cell line, in some cases there was an increase seen in the iCn-WT cells, whereas in others there was an increase in the iCn-CHD2^{+/-} cell line.

5.5: Discussion

At D0, in AUTS2 and PARK2, there was a statistically significant increase in small indels found in the CHD2 mutant line. Most indels were higher in CHD2^{+/-} cells in CSMD3, however one 1bp deletion at 264bp (2bp upstream of the cut site – not shown on graph) that was extremely prominent in both cell lines.

The data for indels from NRXN1 are a mixed picture – some that occur less frequently but with a statistically significant difference are more common in the CHD2 deficient line, whereas the more commonly occurring indels are over-represented in WT cells.

The pattern of smaller indels may be dependent on factors other than CHD2 deficiency. For example, if certain sequence breaks are more amenable to repair by A-EJ or HR, then inhibition of NHEJ by knockdown of CHD2 levels would have less impact than it would on sequences better repaired by NHEJ. There is evidence that the outcome of single DSB repair is largely influenced by the sequence surrounding the cut sites and therefore a full analysis of repair at such lesions would require a different experimental design; one that chose a large number of cut sites that could then be stratified for analysis according to the sequence context of the cut site.

5.4.1.3 Relationship between transcriptional activity and mutation detection

Interestingly the order of statistical significance at D0 followed the same pattern as the read counts from RNA-Seq – AUTS2 was most highly transcribed and had the most significant difference in the level of DCEs between iCn-CHD2^{+/-} and iCn-WT cell lines. This was followed by NRXN1, then PARK2. There was minimal transcription at D0 in CSMD3 and no significant difference in the rate of DCE in this gene at D0. It is worth noting also that PARK2 was minimally expressed and that the DCE count per read was an order of magnitude lower in PARK2 than in AUTS2. Conversely, NRXN1 had the highest DCE count per read despite having a lower transcript count in RNA-Seq data than AUTS2.

Although it is tempting to draw conclusions regarding the relationship between transcription and the efficacy of CRISPR-mediated gene editing based on these results, to establish whether the pattern is a genuine result of biological factors, a much larger experiment with genes chosen specifically for stratified transcriptional rates would be required. This pipeline is a possible tool for such investigations, however further work in this regard is beyond the remit of this project.

5.4.1.4 Potential confounding factors and technical challenges

The reduced efficacy of sequencing with nanopore using our pipeline at D40 is a major confounding factor in the interpretation of these results. The DNA extraction, PCR amplification, barcoding and sequencing protocols followed were identical in both cases, as was the bioinformatic pipeline used to analyse the data output. One possibility is that DNA extraction with QuickExtract at D40 is less efficient or more prone to generating a degraded product than at D0.

Neurons are larger by volume than IPS cells and the additional cytoplasmic contaminants in the sample may have impacted on the PCR reactions – although the products were checked by gel electrophoresis before proceeding.

It is also possible that if DSBs are more prone to translocation during repair in mature neurons, or to the generation of larger indels, that this could have disrupted the sequences and prevented alignment. Again though, changes this significant should have presented themselves on gel electrophoresis either as a significant extra band, or disappearance of the expected product in the case of translocation.

It is possible that disruption in CHD2's global chromatin remodelling role could have made the sequences analysed more prone to CRISPR editing, however the RNA-Seq data suggests that transcription of our target genes was either normal or depleted in our mutant cell line. A more open chromatin structure would, in theory, lead to increased transcription and therefore by this indirect measure we can suggest that this did not influence our results.

5.4.2 Modelling DCE occurrence

Our data is sufficient to conclude that at D0 there is an increased rate of DCE in CHD2 deficient cells. It is hypothesised that this represents a reduction in the rate of repair as opposed to an increase in the rate of DSBs occurring. The previously published evidence suggests a role for CHD2 in the *repair of DSBs*, rather than *protection from damage*.

For a DCE to occur the second cut must occur before the first is healed – it is not known if this is the only factor influencing DCE or if there are other influences affecting whether the DNA between the cut-sites dissociates. For the factors we can understand, the equilibrium represented in this experiment can be considered using the equation in *figure 5.10*.

tB_1 represents the time until the first break occurs. From this point tR_1 represents the time until the first break is either repaired correctly to WT DNA or incorrectly resulting in a single indel. Once an indel has occurred the gRNA can no longer bind and no further DSBs are possible at this site and so we exit the equilibrium. If the DSB is repaired correctly then a further DSB can still occur, provided that the Cas9 complex has not dissociated or degraded.

Alternatively, from the first DSB a second break could occur before the first is repaired. The time taken for this to occur is represented as tB_2 . What is less certain is whether this second break in and of itself is sufficient for a DCE to occur (version A), or if there is a chance for the second DSB to be repaired (tR_2), in which case the time until the double-cut dissociates is represented by time to dissociation (tD) (*figure 5.10 & 5.11*).

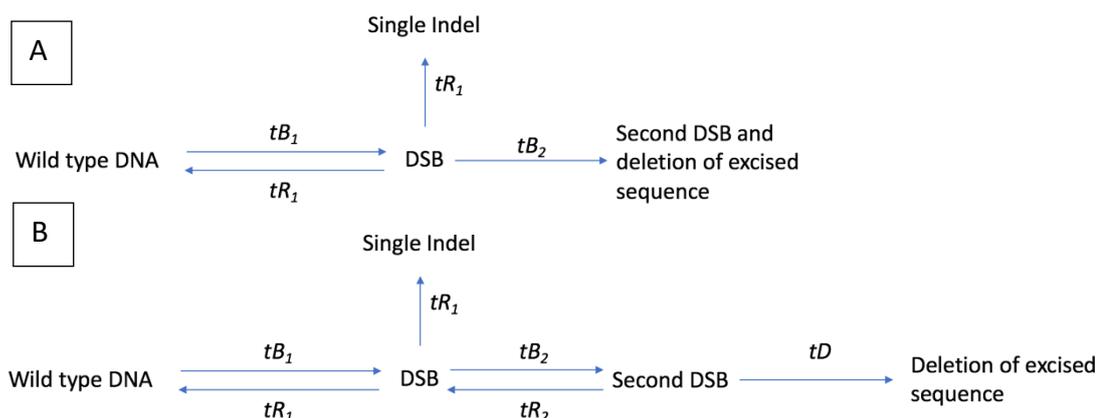


Figure 5.10: possible versions of equilibrium generated at DSBs during double-cut or triple-cut experiments. See body section 5.4.2 text for full explanation of terms.

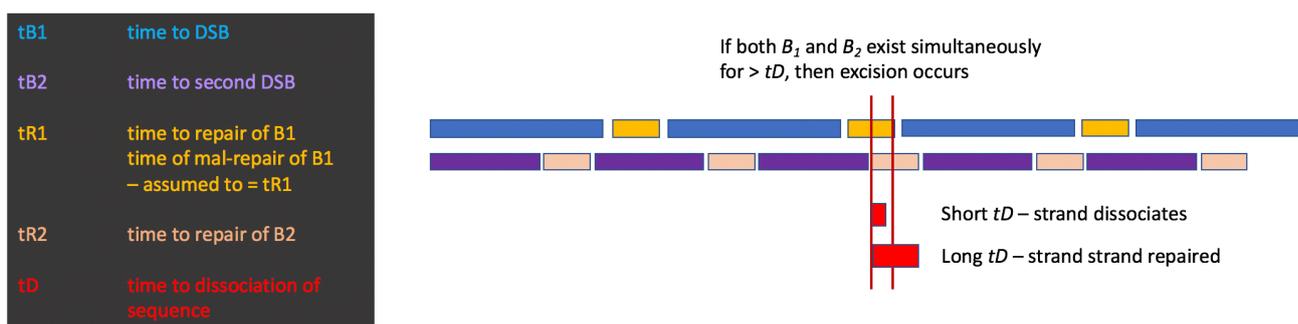


Figure 5.11: schematic representation of equilibrium described in figure 5.10

The process was modelled by setting up a system where a break is recorded if a break constant representing t_{B1} falls below a randomly generated floating point between 0 – 1, and repair occurs if the random floating-point falls below a repair constant RC, representing t_{R1} . By modelling two systems in parallel, different BC and RC could be set for each system, reflecting the fact that different sequences are more prone to both breakage and repair (figure 5.12).

With each time point defined as a new pair of randomly generated floating points, an array was set up to capture the maximum length in time points of unrepaired double-cuts over x number of iterations.

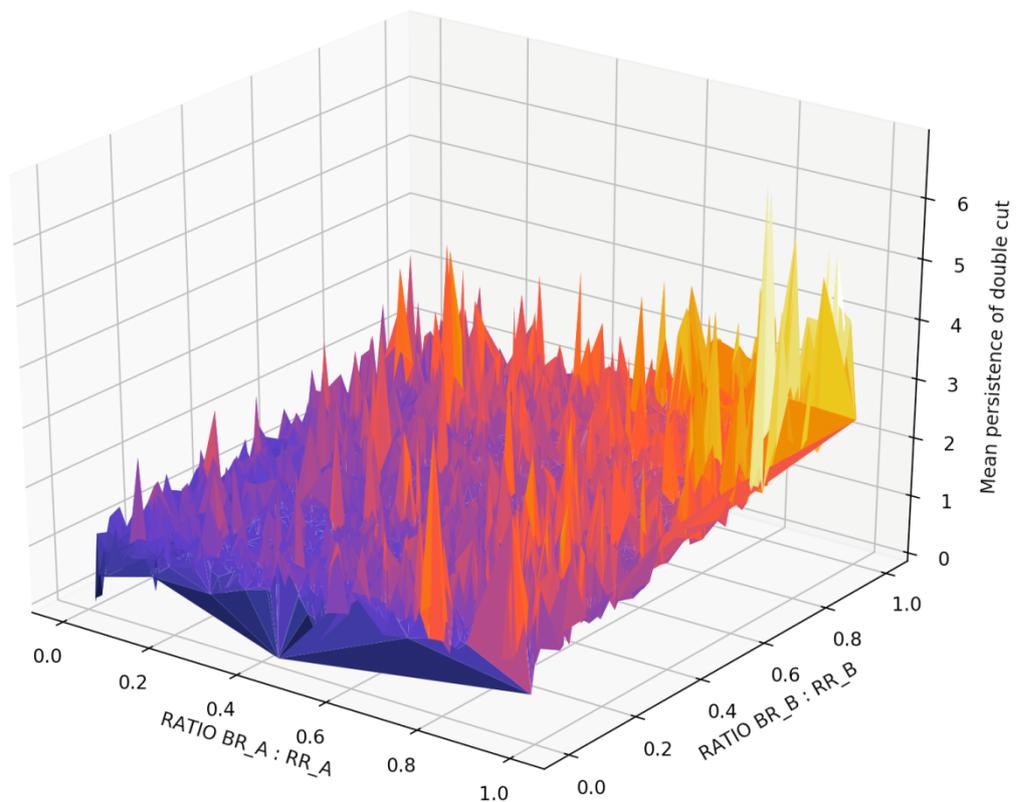


Figure 5.13: Computer modelling of maximum time of double cut persistence based on version A of the equilibrium described in figure 5.12

As these ratios increase, the max persistence of double cut for the iterations increases. As stated above, it is not certain whether constant tD , time to dissociation exists. If we set the tD at an arbitrary value then it would be possible to count the number of iterations that where the maximum double counts exceed it. If it does not, and any momentary existence of a double-cut is sufficient for DCE to occur, then it would be appropriate to consider the same model with a far lower maximum BC:RC ratios, and accept any figure >0 for mean maximum double-count persistence as representing a DCE.

In either situation, we can use this theoretical model to consider our findings. The increase in DCEs seen in CHD2 deficient cells can be interpreted as an increase in the break to repair ratios $tR1:tB1$ and $tR2:tB2$. This could come about either by an increase in the rate of double strand breakage, or by a delay in DSB repair.

As stated in the introduction to this chapter, and detailed in the introduction of this thesis, the evidence suggests that NHEJ is a more rapid process than A-EJ[202, 203]. If NHEJ is inhibited, then we expect DSBs to persist for longer, increasing $tR1$ and $tR2$. Under this

5.5: Discussion

model, this will lead to an increased rate of DCE occurring and indeed this is what our experiment has demonstrated.

From this we can suggest the conclusion that heterozygous CHD2 mutation increases the length of time required for DSBs to be repaired in our cell model. From the previously published evidence regarding the speed at which NHEJ and A-EJ occur, we can also suggest that this increased time to repair represents an inhibition of NHEJ and increased dependence on A-EJ.

6: WHOLE GENOME ASSESMENT OF PHYSIOLOGICAL DOUBLE STRAND BREAK OCCURRENCE DURING NEURODIFFERENTIATION

6.1 Introduction

6.1.1 Introduction to INDUCE-Seq

In *chapter 5* I described the potential impact of CHD2 mutations on the dynamics and outcomes of DSB repair. I inferred from the increase in mutations caused by double-cut excision (DCE) that CHD2 mutations increase the time-to-repair each DSB. In this chapter I approached the assessment of DSB repair by looking at the outcome of repair. By convention, gene editing efficacy has been measured using such approaches due to the difficulty of measuring cleavage directly[258].

Such approaches have the limitation of describing the response to specific damage created by specific reagents that may or may not be generalizable across a whole genome scale or relevant to normal biological function. It is unlikely that anyone will transfect a patient harbouring a CHD2 mutation with two gRNA targeting a specific genomic region. In trying to understand the effect of CHD2 mediated disruption of NHEJ on a patient it is therefore important to also attempt to model normal physiological function.

One possible approach is to search for and analyse recurrent breakpoint clusters. These have been previously identified [152] however they do not localise within a region short enough to be amplified by PCR and therefore investigating these regions with the approach described in *chapter 5* is not possible. It is also possible that these clusters could vary from cell-type to cell-type and therefore reproducibility of the 27 clusters described by Wei et al [168] in our IBJ hiPSC cell lines may be challenging. There is also the non-trivial risk that, given the role of chromatin remodelling in DSB repair, a cell line haplo-insufficient for an important chromatin remodeller such as CHD2 could have a different pattern of recurrent DSB clusters to a wild type cell line. This could lead to a risk of misinterpretation of results if our CHD2 mutant line had *fewer* mutations at a break-point cluster validated in a iCn-WT cell line for example.

A whole-genome approach to the analysis of breaks is therefore desirable. The most straightforward approach would be to use extremely high-depth (100x or higher) sequencing of an entire culture after a period of growth. This would allow for the identification of low level sub-clonal mutations in culture throughout the genome and comparison between iCn-

6.1: Introduction

WT and iCn-CHD2^{+/-} cell lines. Although technically feasible, it would be both an expensive and inelegant approach in terms of both sequencing reagents, and bioinformatic analysis.

Several other approaches using a wider scope to assess the occurrence and repair of DSBs do exist. Protein based assays allow for the assessment of break formation and healing on a nucleus-wide scale; immuno-histochemistry targeting H2AX can generate cell-specific information regarding the abundance and even intra-nuclear localisation of DSBs [6]. As this histone modification is known to form part of the signalling pathway for DSB repair, its accumulation is felt to be an adequate proxy for measurement of DSBs. This approach allows for signal detection from breaks that are correctly repaired, as well as breaks in which mutations occur, however they do not offer single strand resolution in the same way as our sequencing-based approach.

Less precise quantification of DSB occurrence is also possible using an electrophoresis-based technique known as the comet assay. When cells are incubated in an alkaline buffer, broken DNA ends form coiled structures that can be separated by [318] electrophoresis on gel. The name of the assay is taken from the comet shaped tails of DNA fragments that are measured. This assay has been widely used in the assessment of radiation damage [319], and demonstrating the differences caused by mutations in canonical proteins of various repair pathways.

Although both these methods are highly sensitive in the assessment of DNA damage, they do not offer the single strand resolution provided by DNA sequencing.

With the huge global interest in Cas9 as a potential gene therapy and the concern about the potential for harmful off-target editing in human subjects, there has been much recent development in the field of global analysis of DSB occurrence. Indirect methods that require either translocation of the DSB to a known site, or measuring some intermediary of repair are recently being supplanted by direct techniques in which unrepaired DSBs are captured directly from cell culture using various methods[320].

CHIP using an antibody targeting H2AX, or P53 has been previously demonstrated. As these proteins aggregate at DSBs the DNA sequences preferentially cross-linked to them during fixation are thought to correspond to areas of DNA damage. This approach does not directly label the DSBs, rather the protein response to them and as repair proteins do not necessarily bind directly to the break-site itself, this approach still does not allow nucleotide-specific localisation of DSBs. CHIPSeq is also highly dependent on the target chosen; a CHIP-

6: Whole genome assessment of physiological DSB occurrence during neurodifferentiation

6.1: Introduction

Seq experiment targeting one of the core factors of NHEJ (DNA Ligase IV, for example), would not identify DSBs repaired by HR or by A-EJ[321].

Translocation capture sequencing (TC-Seq) is a process by which bait strands of DNA are generated to capture the regions of interest[322]. This process depends on translocation mediated by NHEJ in order to create the hybrid strands, which can then be captured and amplified for sequencing. The 27 neurodevelopmentally genes identified as harbouring recurrent breakpoint clusters mentioned above were identified using this approach[152]. It has also been used to assess: translocations in B lymphocytes, and Cas9 specificity[323].

GUIDE-Seq [257] (Genome-wide, Unbiased, Identification of DSBs Enabled by Sequencing) is an approach that introduces a blunt double-stranded oligodeoxynucleotide (dsODN) at DSBs. dsODNs specific amplification then enriches for the sites of incorporation. A sequencing library is prepared from this dsODN enriched sample, allowing identification of DSBs at a nucleotide resolution.

Although GUIDE-Seq has been demonstrated as a robust method for detection of CRISPR mediated off-target mutations and would therefore be potentially applicable to our inquiry, it depends on NHEJ to integrate the dsODN into a break site and is therefore not suited to investigating a potential disruption of this pathway. It also relies on the incorporation of exogenous genomic material into breaks and despite the author's insistence that this occurs reliably, evidence from experiments that use Cas9 to create specific mutations via introduction of homologous material at DSBs suggest that this is not the case and that such incorporations do not usually occur at all available break sites [324]. It is therefore possible that break-sites are missed.

BLESS is a direct capture technique in which unrepaired DSBs are captured in-situ by end blunting and ligation of biotinylated adapters followed by capture on streptavidin and PCR amplification. It requires at least 10^6 cells per sample and library preparation takes 7-10 days to perform[325]. It has been used in the investigation of replication stress-induced DSBs and to test Cas9 specificity[320].

Breaks Labelling In-Situ and Sequencing (BLISS-Seq) is a direct DSB labelling method that captures DSBs in-situ by ligating oligonucleotides with unique sequences to act barcodes for library amplification via in-vitro transcription[321]. This method overcomes the necessity for functional NHEJ to capture the target sequences but is still prone to potential amplification bias. It has the advantage over BLESS-seq of requiring a far lower number of cells (10^{3-4}) and

6.1: Introduction

being considerably less labour intensive. Again, this approach has been demonstrated to identify CRSIPSR off-target sites.

END-seq is a highly sensitive method designed to operate without the restriction of cell culture and be used for in-vivo samples without fixation. Live cells to be studied are embedded in an agarose plug. Embedded cells are treated with proteinase and RNase and unrepaired ends A-tailed so that they can be captured using an adapter containing Illumina's p5 sequence. The DNA is sonicated, and new ends ligated to the p7 adapter required for sequencing. PCR amplification of the library is performed before sequencing. This technique allows detection of DSBs occurring at very low levels (1 per 10,000 cells) and has been demonstrated to effectively capture physiologically occurring breaks during V(D)J recombination, as well as those induced with a wide variety of break-inducing agents [326]. The high-sensitivity 'snap-shot' approach has also proven effective at describing DNA overhangs at unrepaired ends.

A new assay known as Identification and quantification of DSBs by unbiased flow-cell enrichment and sequencing (INDUCE-seq) has been developed by local collaborator Felix Dobbs and tested extensively in HEK293 cells. It takes a similar approach to BLISS-Seq, but removes the amplification step, meaning that every read sequenced comes from a single DSB site. Fixated cells are permeabilised and adapters ligated to unrepaired DSBs in-situ prior to sonication and size selection. The resulting library is captured on an Illumina flow cell and the only reads capable of being captured are those with the P5 adapters ligated; that is the unrepaired DSBs at the time of library preparation. The technical details are further discussed in the methods chapter (*section 2.5.5.3*).

The advantages of INDUCE-seq are that as an amplification-free assay, each read captured on the flow-cell represents a specific DSB that has been captured by the library preparation process from the cell population being investigated allowing for single cell and single read resolution. The reads can be mapped to the genome, allowing for single nucleotide resolution and providing a resource to query specific genomic regions, motifs and structures to identify patterns in DSB occurrence.

This approach obtains a snapshot of a culture at the time of fixation and by performing bio-replicated experiments consistent patterns can be identified. As previously stated DSBs are a common lesion with around 50 breaks per cell per day previously reported[327] and they occur in a non-random fashion throughout the genome [154].

6.1: Introduction

At the risk of stating the obvious, this assay is unbiased between the identification of DSBs created by exogenous agents and those that occur naturally. It could be used to describe the pattern of DSBs occurring after exposure to: ionising radiation, genotoxic chemical, CRISPR-Cas9 gene editing, any other damaging agent, or breaks that occur naturally within culture conditions. In-fact, like BLISS-seq, it was developed as part of a project to assess off-target effects of Cas9 genome editing.

In this chapter, I will describe an experiment using this technique to compare the frequency and pattern of naturally occurring DSBs between a population of cells with the heterozygous CHD2 mutation and wild-type cells.

Based on the findings discussed in *chapter 5*, which suggest an increased time-to-repair for each cut made in CHD2 deficient cells, we expect to see an increased number of unrepaired DSBs in our iCn-CHD2^{+/-} cell line compared to the iCn-WT cell line at each investigated time-point.

It is also possible that we may identify a change in the localisation of cuts seen. Certain genomic motifs appear more amenable to repair by NHEJ whereas others are more amenable to repair by A-EJ [328] Inhibiting one of these pathways may lead to DSBs being available for capture in the iCn-CHD2^{+/-} cells that are not detectable in the iCn-WT cell line.

Furthermore, it is possible that impairing a chromatin remodeller may alter the three-dimensional structure of the genome in each cell type, and that this change in conformation may impact the number and locus of the DSBs identified.

6.1.2 Aims

- 6.1.2.1 Generate a profile of naturally occurring breaks in cell culture for cells at D0, D19 and D40 of neurodifferentiation
- 6.1.2.2 Compare WT and CHD2 deficient cell lines to determine if there is any significant change in the abundance and pattern of breaks
- 6.1.2.3 Integrate data regarding chromatin modifications during neurodevelopment available from the ENCODE database to consider the relationship between DSB occurrence and chromatin state
- 6.1.2.4 Integrate data from the RNA-Seq experiments described in chapter 4 to assess the relationship between transcription and DSB occurrence

6.2 METHODS

6.2.1 Cell culture and fixation

iCn-WT and iCn-CHD2^{+/-} cells were differentiated into mature neurons, following the protocol described in chapter 2. Cells were collected for fixation and INDUCE-seq library preparation at three time points, D0, D19 and D40 corresponding to IPSC, NPC and Mature Neurons respectively (see also *chapter 4*).

6.2.2 Library Preparation

Library preparation follows the protocol described in section 2.5.5.3. Raw DNA sample concentration is measured at this stage, using the qubit following the procedure in general methods. This allows for estimation of the total number of cells extracted, based on the average cellular DNA concentration of 6.6pg

6.2.3 Sequencing

The library is run on the Illumina Next-Seq using the NextSeq 500/550 High Output Kit v2.5 (75 Cycles) sequencing kit.

Unlike the ONP data, the Illumina data leaves the sequencer already base called. QC was performed with FASTQC. Alignment to the indexed human genome (hg37) was performed using bwa.

Using the estimated weight of DNA extracted from each culture, the number of cells in culture is estimated, by an approximate calculation of 6.6 pg per cell. The ratios between the total number of cells from which DNA was successfully extracted is calculated.

The total number of reads = the total number of DSBs for each culture. This number is normalised to estimate the total number of breaks per cell from culture.

6.2.4 Analysis

6.2.5.1 Genomic context of DSBs

A bed file is generated from the start co-ordinates of each aligned read using bed intersect. A bedgraph is generated by sorting each bed file into bins of 1kb size length and visualised using the Integrated Genome Browser (IGB).

The co-ordinates for each protein coding were obtained from the ensembl browser using biomaRt and formatted into a bed-file using Microsoft excel.

Bedtools intersect was used to count the overlaps with each protein coding transcript. The number of breaks per kb for each transcript was calculated and the distributions plotted at each time point in differentiation. Log2fold changes for each gene were calculated and stored for comparison with RNA-Seq data (*chapter 4*).

6.2.5.2 Enrichment for DSBs captured at histone modification, transcription start sites and known fragile sites

Transcription start sites were downloaded from ensembl using biomaRt. Bedfiles were generated featuring regions 500bp, 1kb and 2kb upstream and downstream from each TSS.

A list of common fragile sites (CFS) was obtained from the HumCFS database[329]. The genomic co-ordinates were formatted into a bed file.

The genomic co-ordinates for the 27 genes found to have increased susceptibility to DSB formation when under replicative stress were obtained from the UCSC genome browser. A bed file containing these co-ordinates was created.

Independent T-testing was performed using the scipy stats python package, was used to compare the enrichment between iCn-WT and iCn-CHD2^{+/-} cells at each time point, for each genomic feature analysed.

6.2.5.3 Sequence content at break sites

Using a similar approach to the sliding window described in *section 2.7.1.10 (figure 2.5)* with the breakpoint of each DSB sequenced as a seed, the sequence 10bp in the direction of the read is captured, resulting in a 10bp window. These are plotted according to frequency of the sequence in the genome and coloured for GC content.

The number of times each 10bp sequence appears in each experiment is counted. The sequences are categorised by whether they are represented uniquely in one experiment or appear in both.

These are plotted in a 'stingray' chart, with log₂-fold change of each 10bp sequence on the x axis, and total number of times the sequence occurs in the genome on the y axis.

$$Score = \log_2\left(\frac{A + 1}{B + 1}\right)$$

Figure 6.1 demonstrates this with two randomly generated datasets from the human genome. 1 million reads of 10bp length are included. The top image displays all 10bp sequences identified, with each data point coloured by GC content of sequence. The bottom image on the right-hand side highlights the sequences that are represented in only one sample.

The average GC content of the shared sequences and the sequences that appear exclusively in each sample are calculated. Chi² contingency testing is used to compare the proportions of each group between both samples.

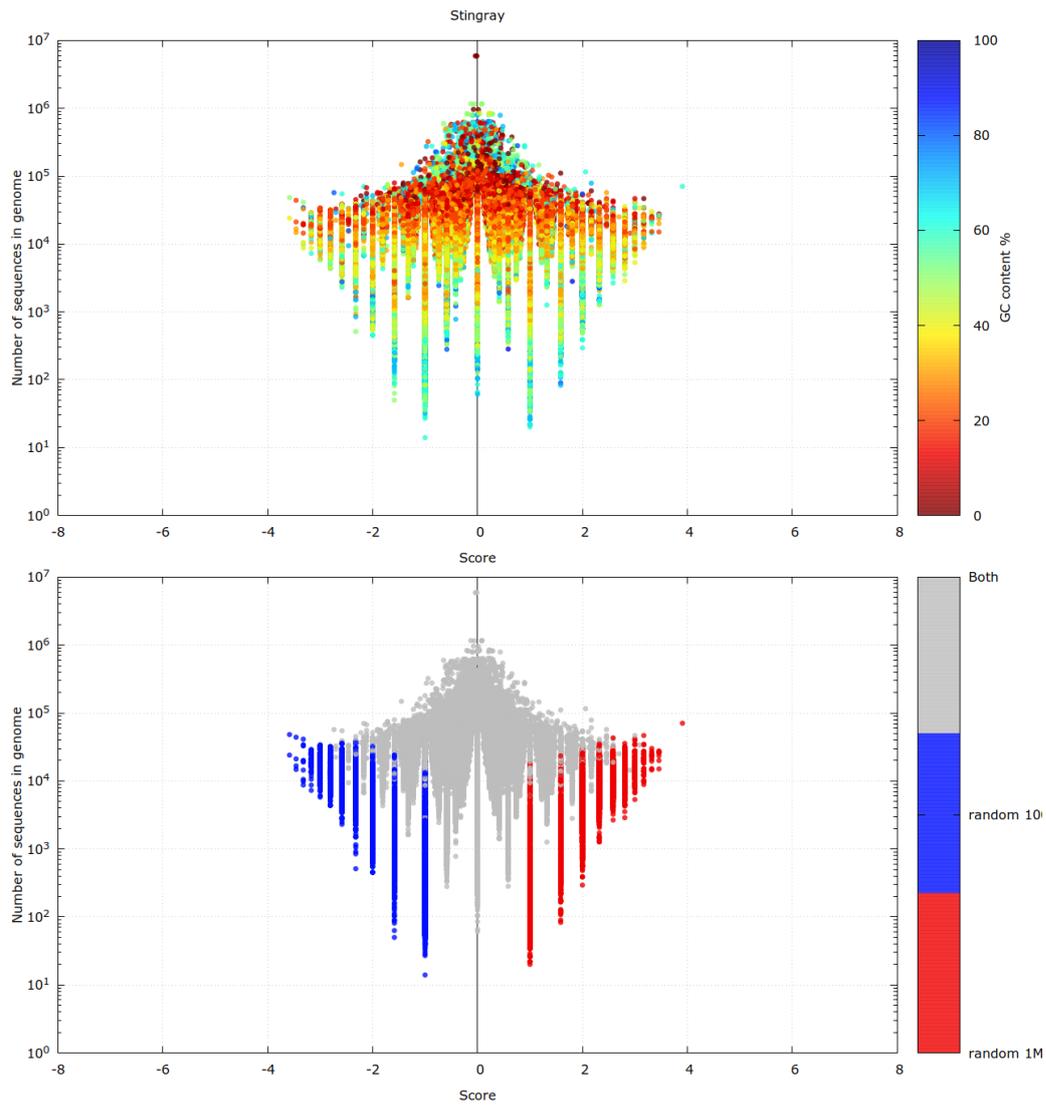


Figure 6.1: example stingray plot made from two datasets of 1M randomly generated breakpoints. Full description in body text of section 6.2.5.3

6.2.5.4 Relationship between DSBs and transcription measured by RNA-Seq

The cells used for these analyses were cultured at the same time, from the same stock cultures, in the same plates as those used in the RNA-Seq experiments described in chapter 4.

To assess the relationship between transcription and DSB occurrence, the FPKM from the RNA-seq data will be plotted against the mean normalised read count per kb per gene from the INDUCE-seq experiments for each cell at each timepoint. Pearson's coefficient and linear regression will be used to assess whether any relationship exists between the rate of transcription and the rate of DSB capture by INDUCE-Seq.

The fold change for RNA-Seq data and INDUCE-seq read counts per kb per gene will also be plotted, comparing iCn-WT and iCn-CHD2^{+/-} cells at each time point. Again, Pearson coefficient and linear regression will be used to assess whether a relationship exists between the fold change in transcription this will be based on the transcript abundance for each gene. In the DSB data, it will be based on the breakpoints per kb analysis described above.

6.3 RESULTS

6.3.1 Quantification and normalisation

6.3.1.1 Breaks per run and breaks per cell

Using the calculation of 6.6pg DNA per cell, the total input DNA was used to estimate the number of cells from which DNA was extracted for each culture. This was used to estimate the total number of reads per cell – corresponding to the number of breaks per cell.

The data is presented in *table 6.1* and *figure 6.2*. There was no significant difference in the number of reads per cell between iCn-WT cells and iCn-CHD2^{+/-} cells at day 0 of differentiation. At day 19, there were significantly more reads per 6.6pg from the iCn-WT cells ($p = 0.002$). At day 40, there were more reads per cell from the CHD2^{+/-} cells, approaching statistical significance ($p = 0.07$).

In the iCn-WT cell line, there was a peak in the number of reads at D19 ($p = 0.001$ vs D0 and $p = 0.149$ vs D40). The read count at D40 was higher than D0, approaching statistical significance ($p = 0.056$). In the CHD2^{+/-} cell line, there was a small but not statistically significant increase between D0 and D19 ($p=0.29$), with a significant spike at D40 ($p=0.016$ vs D0, $p=0.018$ vs D19).

For NPC and Neurons, observation of bedgraphs derived from the endpoints of the INDUCE-seq reads revealed clear changes in the patterns of DSBs observed between timepoints. There were also differences noted between the iCn-WT cells and iCn-CHD2^{+/-} cell lines, however these appeared less obvious than the changes between time-points for the same cell lines (*figure 6.3*).

	Input DNA – ng mean, (SD)	Reads recorded - n mean (SD)	Reads per 6.6pg - n mean, SD	P value for time-point comparison (p)
iCn-WT_D0	49.2(18.5)	1,093,717.7 (111,796)	164.5 9 (48.2)	0.684
CHD2_D0	44.9 (3.72)	1,307,085 (589,276)	198.2 (97.6)	
iCn-WT_D19	95.1 (14.0)	23,152,740 (6,186,557)	1577.9 (240.8)	0.002
CHD2_D19	393.3 (201.54)	16,347,398 (5,614,496)	303.3 (77.1)	
iCn-WT_D40	137.3 (1.25)	20,029,100 (8,658,143)	963.7 (421.0)	0.078
CHD2_D40	65.8 (4.2)	23,999,777 (6,347,784)	2449.6 (791.3)	

Table 6.1 Total INDUCE-Seq reads and estimated reads per cell for each sample, averaged across three technical replicates, with p value calculated using independent T-test.

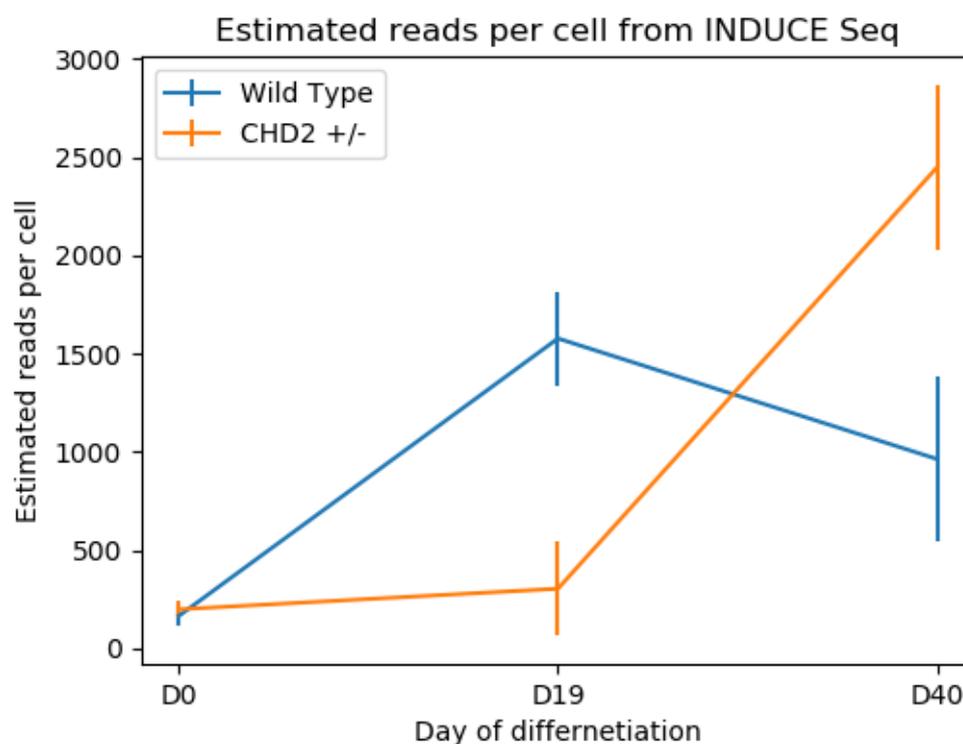


Figure 6.2 – estimated INDUCE-Seq reads per cell plotted for iCn-WT and iCn-CHD2^{+/-} cells, with standard deviation, at D0, D19 and D40 of neurodifferentiation

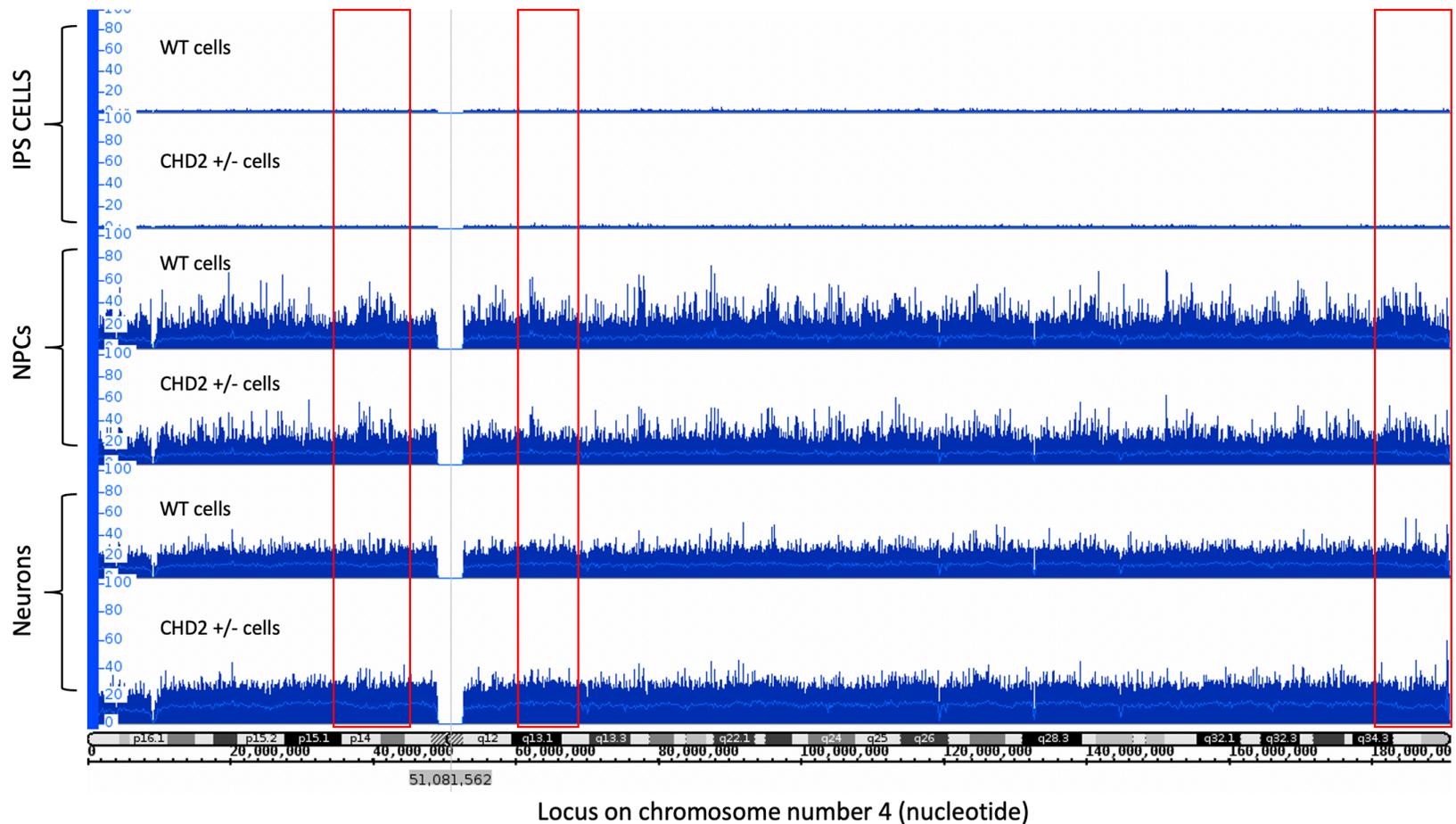


Figure 6.3 – Bedgraphs demonstrating break points determined by INDUCE-seq read start co-ordinates in chromosome 4. Row 1- *iCn-WT_D0*, Row 2- *CHD2_D0*, Row 3- *iCn-WT_D19*, Row 4- *CHD2_D19*, Row 5- *iCn-WT_D40*, Row 6- *CHD2_D40*. Breakpoints are sorted into bins of 1000bp width to demonstrate areas of higher and lower susceptibility. Red boxes highlight areas to demonstrate similarity between *iCn-WT* and *CHD2* cell lines at stages of differentiation, but changes across day of differentiation. Chr4 shown to best demonstrate these features – the whole dataset can be found in the supplementary material

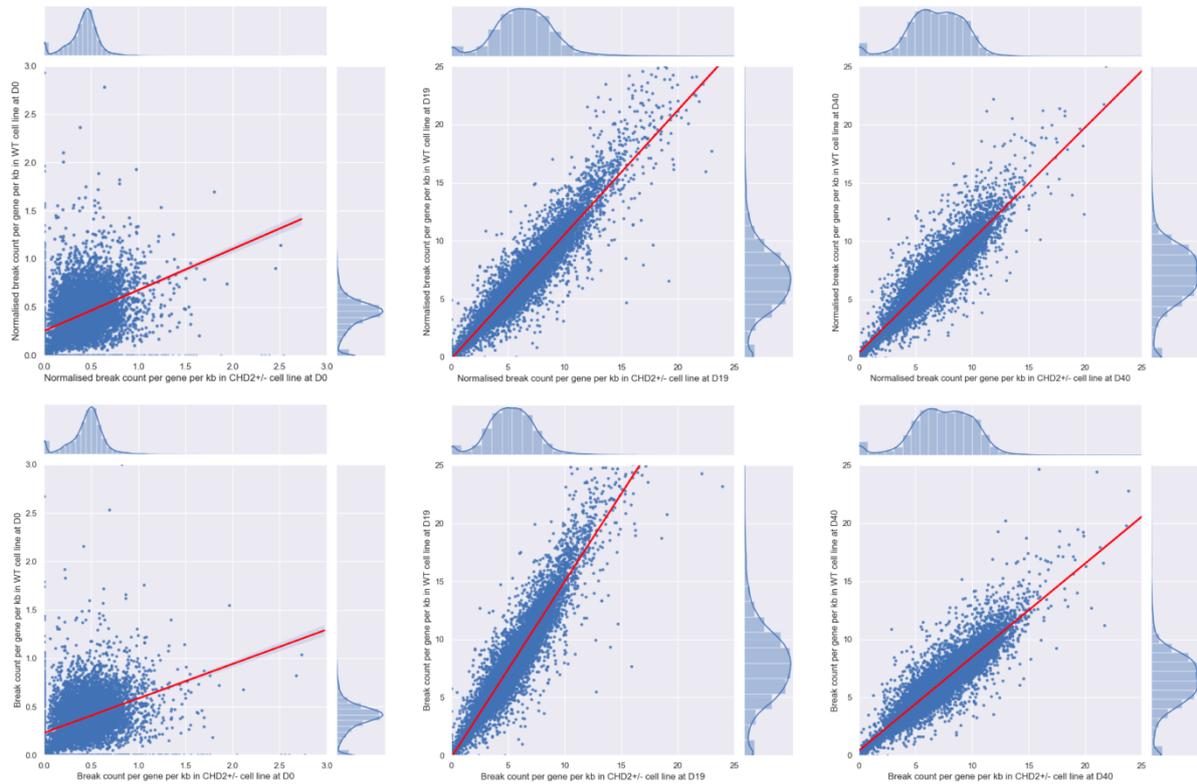


Figure 6.4 – Breaks per gene determined by INDUCE-Seq for all protein coding transcripts. Y axes = read counts per gene for iCn-WT cells, X axes = read counts per gene for iCn-CHD2^{+/-}. Left D0, Middle D19, Right D40, top row normalised, bottom row un-normalised

6.3.1.2 Overlap with protein coding genes

Figure 6.4 demonstrates the normalised and un-normalised break-counts per kb per gene derived from read end loci of INDUCE-seq data. At D0, there was a clear preponderance for breaks inside gene bodies in iCn-CHD2^{+/-} cells ($p=4.826 \times 10^{-9}$).

At D19 there was a greater preponderance for breaks to occur in the gene bodies of iCn-WT cells ($p=1.894 \times 10^{-20}$)

At D40, when normalised for readcount, there was a preponderance for breaks inside gene bodies in iCn-CHD2^{+/-} cells ($p=7.52 \times 10^{-17}$).

6.3.2 Intersection of INDUCE-seq reads and genomic features

6.3.2.1 Relative enrichment for DSBs at PTM histone markers

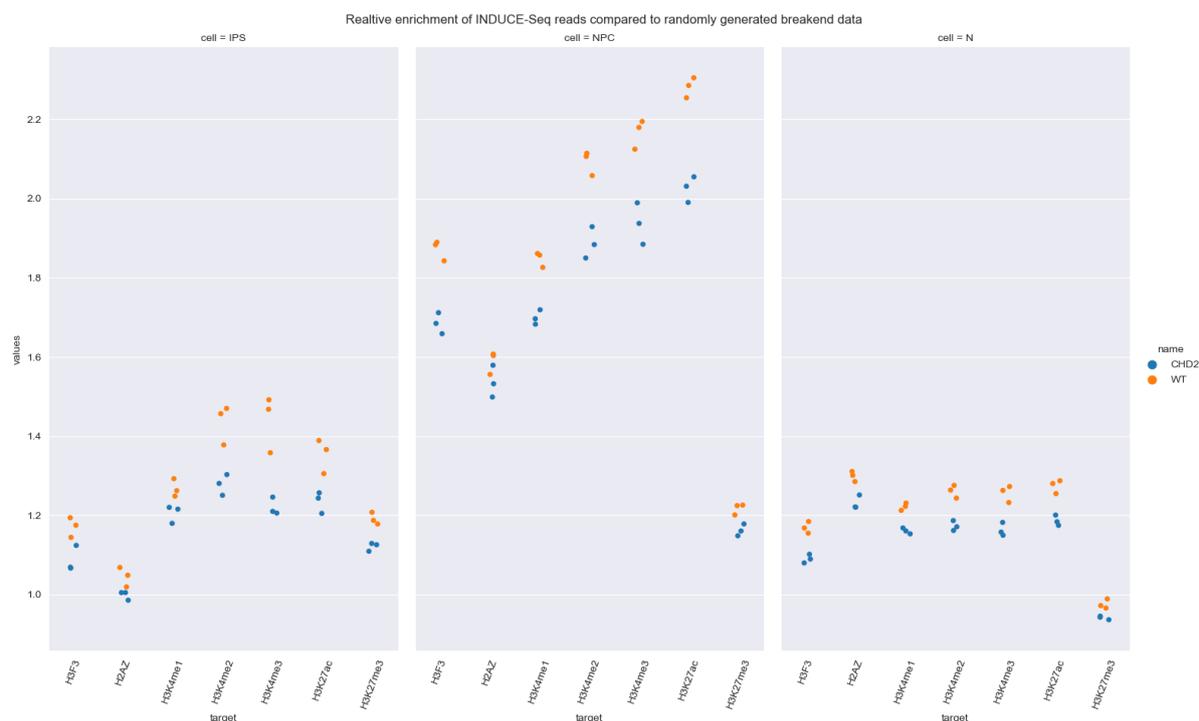


Figure 6.5 – Relative enrichment for INDUCE-seq reads at histone modifications obtained from the ENCODE database

The relative enrichment for break ends derived from INDUCE-seq data expressed as a ratio to the number of breaks in a randomly generated dummy file for each histone marker analysed can be seen in *figure 6.5*.

At D0, H2AZ was the least enriched for DSBs, with no clear difference between cell lines ($p=0.04$). There was no difference for enrichment at D19 ($p=0.141$), however at D40 the marker was the *most* enriched for DSBs in both cell lines, with a statistically significant increase in enrichment in iCn-WT cells ($p=0.006$).

For all other markers, there was an increase in the relative enrichment in iCn-WT cells compared to iCn-CHD2 cells (see *table 6.2*).

H3K27me3 was histone marker with the lowest level of relative enrichment in both cell lines at all stages of neurodifferentiation. H3.3 also had a relatively low enrichment for DSBs in D0 and D40, but not at D19 where it was comparable to other markers.

H3K4me1, H3K4me2, H3K4me3 and H3K27ac had comparable levels of relative enrichment at D0, D19 and D40, however H3K27ac was significantly more enriched than the other markers in WT cells at D19.

6: Whole genome assessment of physiological DSB occurrence during neurodifferentiation

6.3: Results

Overall, there was a clear shift in the relative enrichment patterns for each histone marker in both cell lines between D0 and D40. The pattern of relative enrichment of DSBs at histone markers was the same in iCn-WT and iCn-CHD2^{+/-} cells, however the relative enrichment was higher in iCn-WT cells for each marker at each stage of differentiation, except for H2AZ at D0 and D19.

Cell	Target	WT mean (enrichment ratio of breaks per kb)	WT standard deviation	CHD2 ^{+/-} mean (enrichment ratio of breaks per kb)	CHD2 ^{+/-} standard deviation	p-value
IPS	H3F3	1.171	0.02	1.087	0.027	0.023
IPS	H2AZ	1.046	0.02	0.999	0.009	0.04
IPS	H3K4me1	1.268	0.018	1.205	0.018	0.027
IPS	H3K4me2	1.435	0.041	1.278	0.021	0.008
IPS	H3K4me3	1.439	0.058	1.221	0.018	0.007
IPS	H3K27ac	1.353	0.035	1.235	0.022	0.016
IPS	H3K27me3	1.191	0.012	1.122	0.009	0.003
NPC	H3F3	1.872	0.021	1.685	0.022	0.001
NPC	H2AZ	1.589	0.023	1.537	0.033	0.141
NPC	H3K4me1	1.848	0.016	1.699	0.015	0.001
NPC	H3K4me2	2.092	0.025	1.887	0.032	0.002
NPC	H3K4me3	2.166	0.03	1.937	0.043	0.003
NPC	H3K27ac	2.281	0.021	2.025	0.027	0.0004
NPC	H3K27me3	1.217	0.011	1.163	0.012	0.01
N	H3F3	1.169	0.012	1.091	0.009	0.002
N	H2AZ	1.299	0.011	1.231	0.014	0.006
N	H3K4me1	1.222	0.007	1.161	0.006	0.001
N	H3K4me2	1.261	0.013	1.173	0.01	0.002
N	H3K4me3	1.256	0.017	1.163	0.014	0.004
N	H3K27ac	1.274	0.014	1.186	0.011	0.002
N	H3K27me3	0.976	0.01	0.942	0.004	0.01

Table 6.2 Relative enrichment for INDUCE-seq reads at histone marker peaks, with comparison between iCn-WT and iCn-CHD2^{+/-} cell lines

6.3.2.2 Relative enrichment for DSBs at TSS

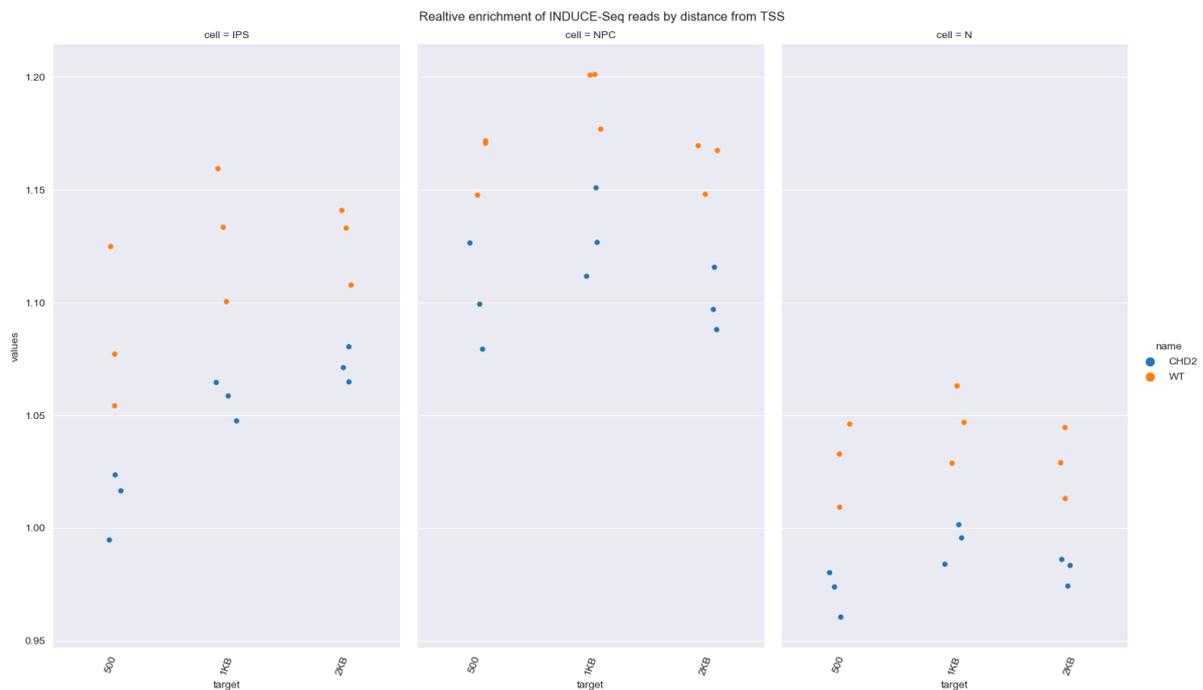


Figure 6.6 - Relative enrichment for INDUCE-seq reads at 500bp, 1kb and 2kb from transcription start sites, taken from ensembl genome database

At D0 and D19, there was relative enrichment of DSBs 1Kb and 2Kb from the TSS. At D0, there was relative enrichment in the iCn-WT cells, but not the iCn-CHD2^{+/-} cells.

At D40, there was relative enrichment for 500bp, 1kb and 2kb from the TSS, but relative depletion for 500bp and 2Kb in iCn-CHD2^{+/-} cells, with no enrichment or depletion at 1Kb.

At all time-points, for all distances from the TSS, there was a statistically significant increase in the relative enrichment of DSBs in iCn-WT cells compared to the iCn-CHD2^{+/-} cells (figure 6.6, table 6.3).

Although the enrichments were at a lower level than for the majority of histone markers (figure 6.5), there was still a reproducible pattern in the enrichments – being highest at D19 and lowest at D40 of neurodifferentiation. As with the histone markers, this pattern was identical for both cell lines, with only the degree of enrichment or depletion changing.

Cell	Distance from TSS	WT mean (enrichment ratio of breaks per kb)	WT standard deviation	CHD2 ^{+/-} mean (enrichment ratio of breaks per kb)	CHD2 ^{+/-} standard deviation	p-value
IPS	500bp	1.0854	0.0294	1.0116	0.0123	0.0306
NPC	500bp	1.1633	0.0111	1.1017	0.0193	0.0173
N	500bp	1.0294	0.0152	0.9716	0.0082	0.0091
IPS	1KB	1.1311	0.0241	1.0569	0.0071	0.014
NPC	1KB	1.1929	0.0114	1.1297	0.0161	0.0106
N	1KB	1.0462	0.014	0.9937	0.0073	0.0093
IPS	2KB	1.1272	0.0141	1.0721	0.0064	0.0074
NPC	2KB	1.1617	0.0097	1.1002	0.0115	0.0045
N	2KB	1.0289	0.0129	0.9813	0.0051	0.0082

Table 6.3 Relative enrichment for INDUCE-seq reads for distance from TSS, obtained from the Ensembl genome database, with comparison between iCn-WT and iCn-CHD2^{+/-} cell lines

6.3.2.3 Relative enrichment for DSBs at previously described fragile sites

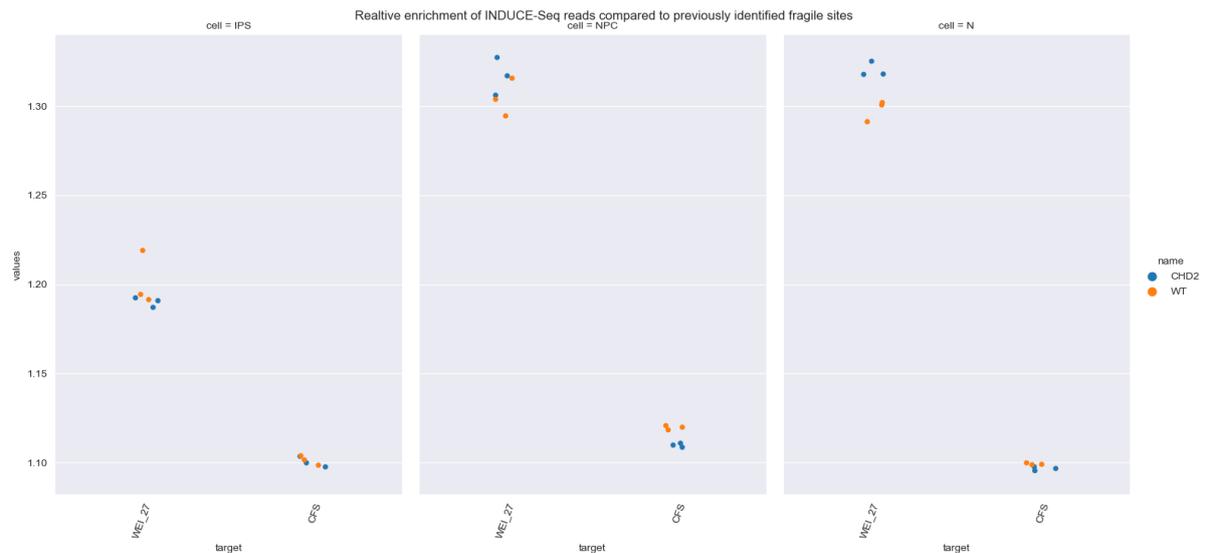


Figure 6.7: Relative enrichment for INDUCE-seq reads at previously identified fragile sites – see body text of section 6.3.2.3 for full explanation

The relative enrichments for previously described CFS and for the long late-replicating genes demonstrated to harbour recurrent breakpoint clusters by Wei et al (referred to as the Wei 27) can be seen in *figure 6.7* and *table 6.4*.

There was a very slight increase in relative enrichment for DSBs at CFS in iCn-WT cells compared to iCn-CHD2 at D19 ($p=0.00048$) and D40 ($p=0.0207$), however not at D0 ($p=0.70$). It should be noted that the change in the enrichment ratios was extremely small in these cases (0.009 at D19 and 0.003 at D40), but the standard deviations were miniscule. Whether this statistical difference has any biological relevance is doubtful.

There was no difference in relative enrichment for breaks within the gene bodies of the Wei 27 at D0 ($p=0.266$) or D19 ($p=0.233$), however enrichment was higher in the iCn-CHD2^{+/-} cells at D40 ($p=0.00581$).

Cell	Target	WT mean (enrichment ratio of breaks per kb)	WT standard deviation	CHD2 ^{+/-} mean (enrichment ratio of breaks per kb)	CHD2 ^{+/-} standard deviation	p-value
IPS	WEI_27	1.20162	0.01238	1.19012	0.00225	0.26603
NPC	WEI_27	1.30469	0.00865	1.31682	0.00865	0.23342
N	WEI_27	1.29799	0.00478	1.32034	0.00343	0.00581
IPS	CFS	1.10127	0.00216	1.10032	0.00245	0.70087
NPC	CFS	1.11963	0.00098	1.10975	0.00092	0.00048
N	CFS	1.09922	0.00046	1.0966	0.00089	0.02072

Table 6.4: Relative enrichment for INDUCE-seq reads for CFS and genes demonstrated to be susceptible to DSB in previous study (Wei_27) [152], obtained from the Ensembl genome database, with comparison between iCn-WT and iCn-CHD2^{+/-} cell lines

6.3.3 Sequence content at break sites

	Sequence Type	WT sequence count	GC % of sequences , WT	CHD2 ^{+/-} sequence count	GC % of sequences, CHD2 ^{+/-}	p-value for comparison of GC% between WT and CHD2
IPS	Shared	368086	39.1	395300	39.7309	1.73E-08
IPS	Exclusive	106259	45.8	82309	47.1758	1.54E-09
NPC	Shared	398625	37.8	401984	37.9598	0.077891718
NPC	Exclusive	89404	45.9	87791	46.4633	0.015988625
Neuron	Shared	373247	38.3	382514	38.4736	0.094071173
Neuron	Exclusive	91460	46.7	84704	46.7136	0.992746041

Table 6.5: sequence count and GC content of 10bp sequences upstream of DSBs detected by INDUCE-seq for each stage of neurodifferentiation. Comparison made of %GC content between sequences in iCn-WT and iCn-CHD2^{+/-} cells, for sequences shared between both cell lines and sequences only identified in one cell line

At D0 of neurodifferentiation, the GC content was higher in sequences 10bp upstream of DSBs in iCn-CHD2^{+/-} cells than in iCn-WT cells (table 6.5), for sequences shared between both cells ($p=1.73 \times 10^{-8}$) and for sequences exclusive to one or another cell line ($p=1.54 \times 10^{-9}$).

At D19, the GC content was significantly higher for reads exclusive to iCn-CHD2^{+/-} cells than reads exclusive to iCn-WT cells ($p=0.0159$), but not for sequences which appeared in both datasets ($p=0.077$).

There was no statistically significant change in the GC content of either the shared sequences ($p=0.094$) or the sequences exclusive to each cell line ($p=0.993$) in neurons.

Most notably, at all stages there was a statistically significant difference between the %GC content of shared sequences and exclusive sequences. The difference was so significant, in fact, that the chi2 contingency test returns a p value of 0.0 at D0, D19 and D40 of neurodifferentiation.

Graphical representation of the skewing of GC content and unique sequences can be found in figure 6.8. The reduced GC content of the shared sequences is largely represented by the red and yellow clouds, whereas the exclusive sequences (represented as blue or red lines) appear as green-yellow vertical lines, grouped by their log2fold ratio.

6: Whole genome assessment of physiological DSB occurrence during neurodifferentiation

6.3: Results

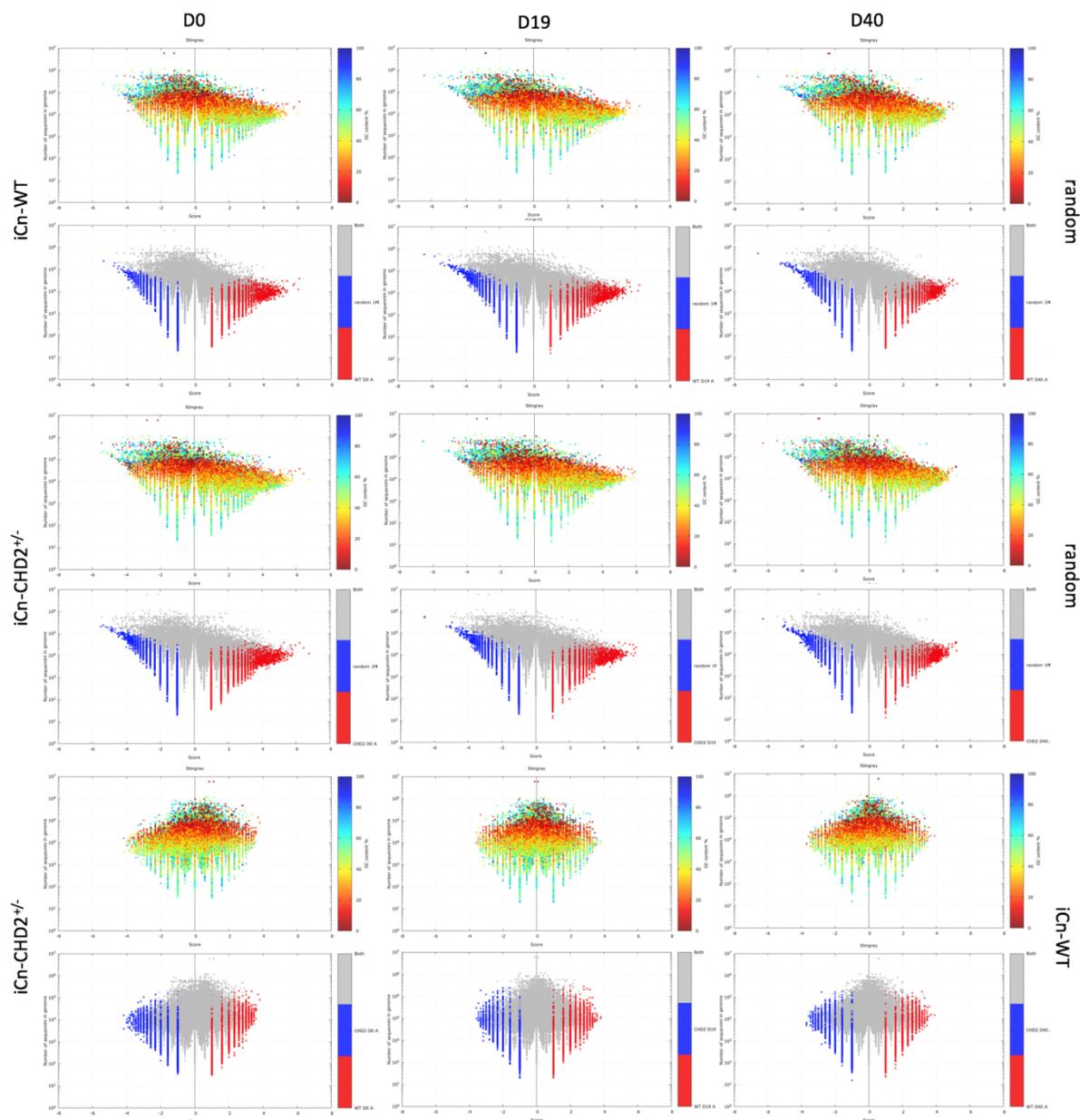


Figure 6.8: Stingray plots demonstrating relative representation of and GC content of shared and unique 10bp sequences at break sites, between iCn-WT and iCn-CHD2^{+/-} at D0, D19 and D40 of neurodifferentiation, plotted against frequency of the sequence in the human genome (log 10 scale)

6.3.4 Relationship to transcription

6.3.4.1 Relationship between FPKM and break count per Kb

There was a clear, if small relationship between FPKM reads from the RNA-Seq data described in chapter 4, and break count per Kb for the same genes in both cell lines, at all stages of differentiation, as demonstrated in *figure 6.9*. The Pearson coefficient and p values can be found in *table 6.6*. At each stage of cell differentiation, genes that were found to be more highly transcribed were more likely to harbour unrepaired DSBs.

This positive correlation between transcription and DSB rate was strongest at D19 of neurodifferentiation for both iCn-WT (Pearson coefficient = 0.213) and iCn-CHD2^{+/-} (Pearson coefficient=0.198).

The positive correlation was weaker, but still statistically significant at D0 0.0858 for iCn-WT and 0.0801 for iCn-CHD2^{+/-}.

6.3.4.2 Relationship between fold changes in RNA-seq and INDUCE-seq break counts

There was no statistical relationship between the log₂fold change for gene transcription and the log₂fold change for DSB occurrence at D0 (p=0.434) or D40 (p=1.09), however a small effect existed at D19 (Pearson coefficient 0.068, p=3.43x10⁻⁷) (*figure 6.10*). This means that at D19, genes where transcription was upregulated in iCn-CHD2^{+/-} cell lines had an increased enrichment for DSBs as inferred from INDUCE-seq data.

6: Whole genome assessment of physiological DSB occurrence during neurodifferentiation

6.3: Results

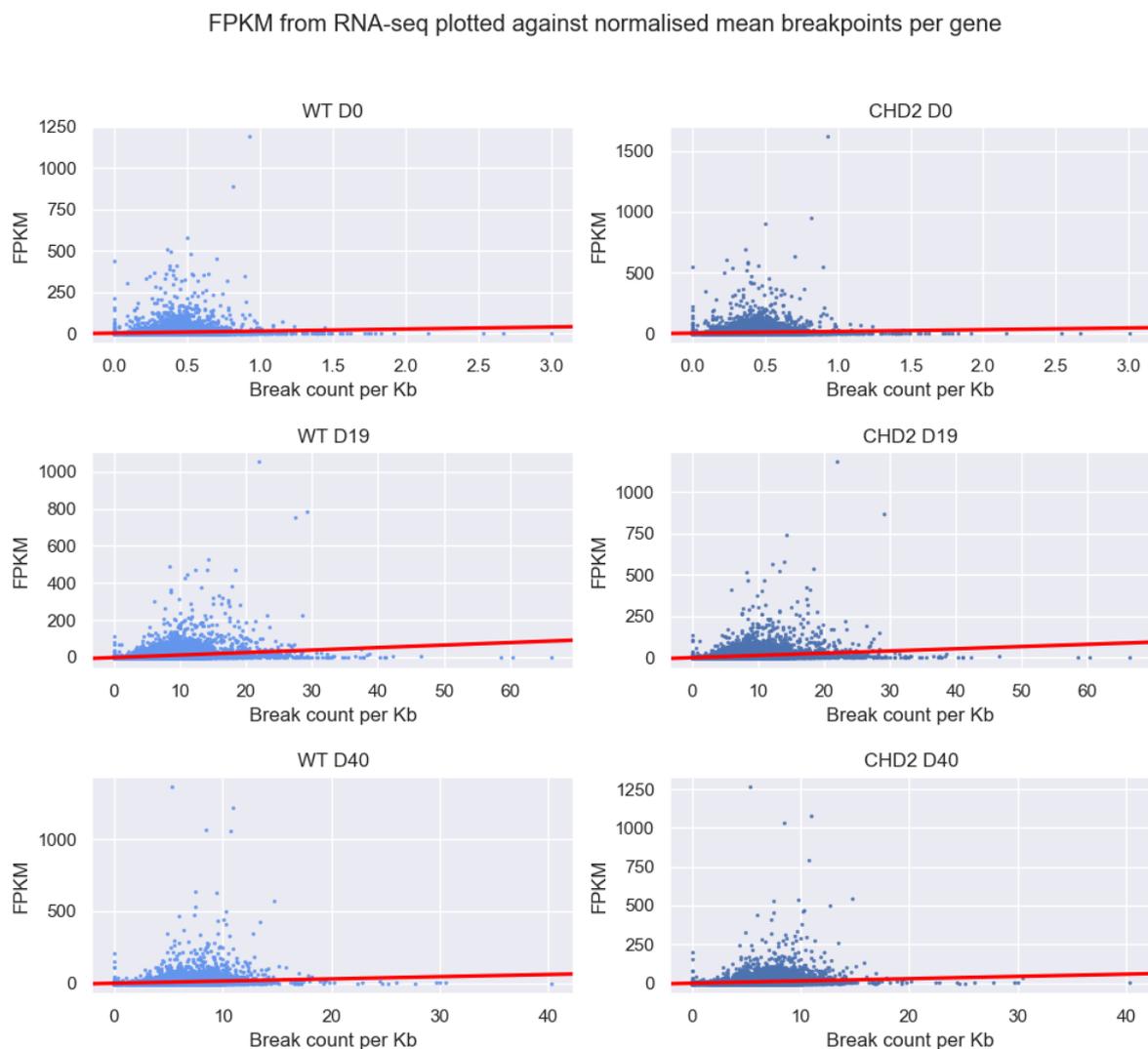


Figure 6.9: Normalised break count per Kb per gene derived from INDUCE-seq read counts, plotted against FPKM from RNA-seq data for all protein-coding genes (blue), with linear regression (red) demonstrating positive association between transcription rate and break rate

Stage of neurodifferentiation	Cell line	Pearson coefficient	p-value
IPS	iCn-WT	0.0858	1.225489×10^{-30}
IPS	iCn-CHD2 ^{+/-}	0.0801	6.445018×10^{-27}
NPC	iCn-WT	0.2129	$1.372829 \times 10^{-182}$
NPC	iCn-CHD2 ^{+/-}	0.1977	$2.918851 \times 10^{-157}$
Neuron	iCn-WT	0.1259	$3.2629736 \times 10^{-64}$
Neuron	iCn-CHD2 ^{+/-}	0.1319	$2.6451583 \times 10^{-70}$

Table 6.6: Pearson coefficients and p-values for relationship between gene transcription levels and DSBs captured per gene by INDUCE-Seq

6: Whole genome assessment of physiological DSB occurrence during neurodifferentiation

6.3: Results

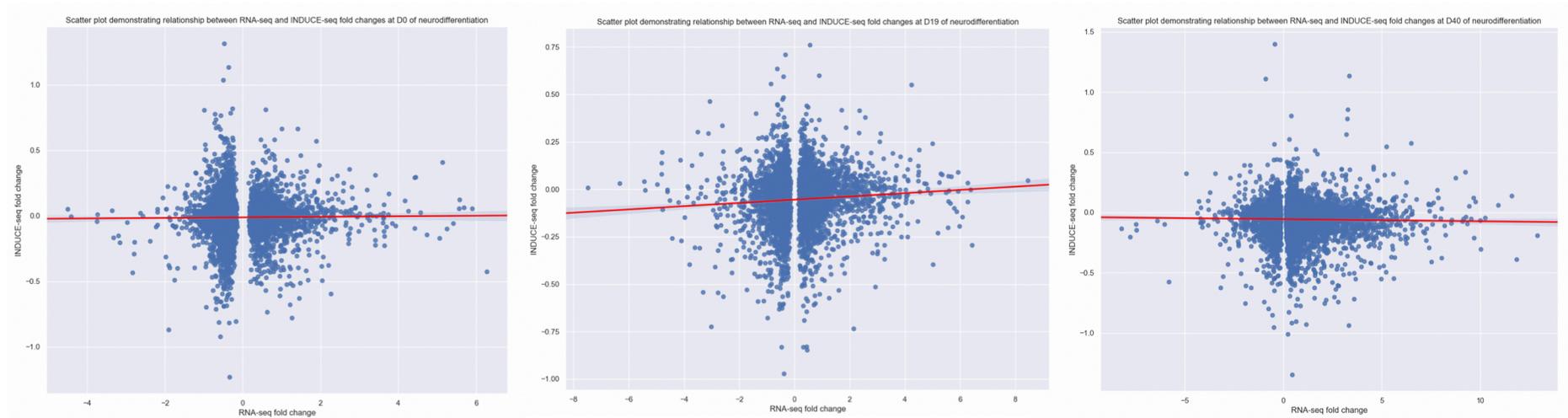


Figure 6.10: Log₂-fold change for breaks per kb per gene, plotted against Log₂fold change per gene from RNA-Seq data. Positive values indicate increased breaks per kb / increased transcription in *iCn-CHD2^{+/-}* cells compared to *iCn-WT* cells. Only the positive correlation between log₂-fold change and breaks per kb at D19 achieved statistical significance.

6.4 Discussion and Conclusions

6.4.1 Quantification and localisation of DSBs

The number of breaks captured for each experiment run gives a snapshot of the genomic landscape of the sample at that specific time-point only. It represents the number of currently unrepaired breaks, rather than a number of DCEs, or indels as described in *chapter 5*. This data makes no assumptions about whether these breaks are repaired correctly or incorrectly, simply that they were available for adapter ligation at the time of permeabilization.

Perhaps the most remarkable finding from all the experiments presented in this thesis is the high number of DSBs observed in cells undergoing neurodifferentiation using our protocol. Compared to previous studies describing the number of DSBs per cell per day, the number of DSBs per cell calculated as breaks per 6.6pg of input DNA provides a massively higher estimate of breaks per-cell at a single snapshot of culture. A commonly quoted figure is that each cell in the body experiences ~50 DSBs per day [327].

BLISS-seq data put the DSB_{max} per cell in a snapshot higher than this number, at 94 breaks[321]. At D0, the estimated mean breaks per cell are not wildly elevated compared to this estimate and would be within the expected margin for increased detection by a potentially more sensitive assay. Certainly though, mean estimates of 1577.9 breaks per cell for WT cells at D19 of differentiation and 2449.6 breaks per cell in CHD2 deficient cells at D40 are considerably higher than one might expect.

It is tempting when faced with such unexpected results to assume that the data is in some way inaccurate, however there are several findings that provide evidence that this is not the case. The INDUCE-seq workflow was designed and engineered using HEK293 cells – a technical control run concurrently on the same flow cell consisting extracts from this cell line (data not shown) demonstrated data consistent with previous runs, demonstrating that there were no technical abnormalities with the sequencing or flow cell incubation.

All technical aspects of cell culture and sample preparation were controlled for; the cells were fixed using the same technique and from a differentiation beginning with the same paired starter cultures. The storage conditions (sealed in lab tape at 4°C) were identical, as was the procedure used for library preparation – the libraries were prepared using the same batches of reagents. No exposure to DNA damaging agents or environments was allowed.

6.4: Discussion

The relationship between break enrichment and previously described biological features (discussed in *section 6.4.2*) suggests that the results are genuine. If the results were artefactual, then one would expect them not to exhibit patterns that concord to known biological markers.

Whether or not describing the ratio of input weight to reads counted in terms of breaks-per cell is necessarily an accurate estimation of the number of cells included in each extraction, quantifying the number of breaks in some sort of ratio to the total DNA input makes sense. With this in mind, we will go forward with our interpretation of these results with the understanding that at D19 there were a higher number of breaks captured in WT cells, whereas at D40 there were a higher number in CHD2 deficient cells (albeit only approaching the threshold of statistical significance, $p=0.078$).

At D0 there was no statistically significant difference between the normalised break count in WT and CHD2^{+/-} cells across the whole genome. Despite this, the break count per kb of protein coding gene was higher in CHD2 deficient cells than WT cells. Put simply, at D0, protein coding genes were more enriched for DSBs in the iCn-CHD2^{+/-} cell line.

This implies that at D0 protein coding genes in WT cells are relatively protected from breakage, either by more rapid repair or less promiscuous breakage compared to protein coding genes in CHD2 deficient cells.

The opposite pattern was seen at D19, with relative enrichment of breaks per Kb of protein coding gene in WT cells compared to CHD2 deficient cells. This implies that at D19 of neurodifferentiation, protein coding genes are relatively protected from breakage in CHD2 deficient cells compared to WT cells.

At D40, as with D0, there was relative enrichment for breaks in gene bodies in CHD2 depleted cells, implying that protein coding genes are relatively protected in WT cells.

It is also possible that NPCs and neurons genuinely experience a far higher rate of DSB than previously appreciated. They are known to be highly transcriptionally active and undergo significant shifts in their chromatin architecture as a developing cell line and as described in the *Chapter 1*: a very high degree of somatic brain mosaicism is present in samples taken from adult patients[144] and programmatic DSBs have been described in neurons [156]

6.4.2 Enrichment of captured breaks at genomic and epigenomic features

Histone marker	Regulatory functions
H3.3	Found at active chromatin and primed enhancers[93-95]
H2AZ	Found at TSS and regulatory elements. Conflicting evidence exists regarding its effect on transcription and this is likely mediated by PTM [98, 330]
H3K4me1	Found at primed enhancer sites [99] – associated with DNA hypomethylation [331] and active chromatin
H3K4me2	Enrichment defines transcription factor binding regions[100]
H3K4me3	Associated with transcription initiation[101]
H3K27ac	Enrichment separates active enhancers from poised enhancers[102]
H3K27me3	Promiscuous repressive marker [103] associated with heterochromatin compaction and polycomb repressor complex binding

Table 6.7: regulatory functions of histone markers examined for DSB enrichment

For ease of reference the regulatory functions of each chromatin marker investigated can be found in *table 6.7*. Two clear patterns emerge from the histone marker enrichment for DSBs, as well as several patterns that are less clear but still worthy of exploration.

Firstly, across all three stages of neurodifferentiation markers associated with active chromatin and transcription have relative enrichment for DSBs detected by INDUCE-seq, whereas there is a relative depletion of DSBs at the heterochromatic marker H3K27me3, which is associated with transcriptional repression.

Secondly, for the majority of chromatin markers representing all stages of the cell cycle there is an increased relative enrichment for DSBs in WT cells compared to CHD2 depleted cells, with the only exception being H2AZ at D19 of neurodifferentiation ($p=0.141$). The separation in enrichment for DSBs at H2AZ only just meets the threshold for statistical significance at D0 ($p=0.04$). At D40, the separation is more significant ($p=0.006$).

The enrichment for DSBs at all examined sites was higher at D19 for both cell lines than D0 or D40, with the exception of the heterochromatic marker H3K27me3, which was enriched at a similar level at D0.

The relationship between enrichment of the most ubiquitous markers, H3.3 and H2AZ, changed with neurodifferentiation. At D0 and D19, H3.3 had a higher relative enrichment for DSBs than H2AZ. At D40, H2AZ had a higher degree of relative enrichment.

This data provides significant evidence that the DSBs captured during neurodifferentiation in these experiments are related to chromatin state, and therefore likely to be influenced by the activity of either transcriptional apparatus or regulatory elements. As

well as being interesting and worthy of consideration in and of themselves, these results recapitulate the previously described pattern of heterochromatin being relatively protective against DSB compared to highly active chromatin states. This provides supporting evidence that our dataset is a genuine reflection of biological activity and that the readcount; DNA input ratio described in *section 6.4.1* may indeed be a genuine reflection of reality.

When considering the consistency of the higher relative enrichment at all chromatin markers in the WT cell line, it is worth considering that the histone map of CHD2 mutant cells may well be different than that of the WT cells. The CHIP-Seq data used was not obtained experimentally as part of this project and is generated from experiments using wild-type cell lines. If the peaks described in these samples describe a lower fraction of the enrichment of each chromatin marker in CHD2 than WT cells, then this could lead to a ‘false negative’ effect in the analysis of the CHD2 data (*figure 6.11*).

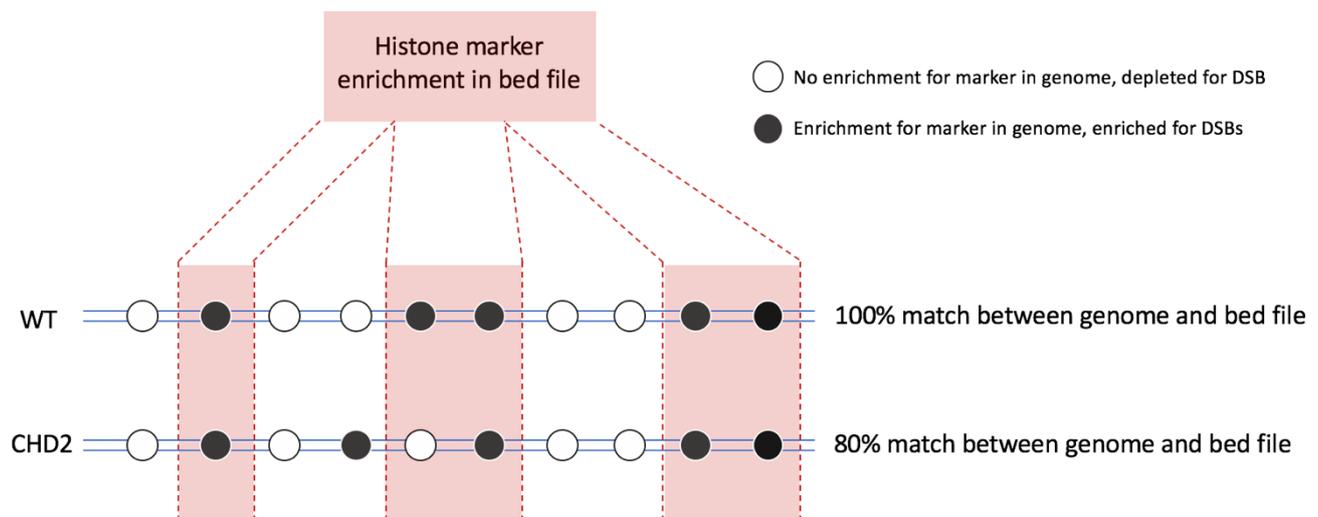


Figure 6.11: schematic of potential for false negative enrichment calls in CHD2 deficient cell line, in the event of an altered histone PTM landscape

6.4: Discussion

From the data available, it is not possible to determine whether the lower enrichment at active chromatin markers in CHD2 depleted cells is mediated by a change in the histone PTM landscape, or by a direct effect on the ratio of DSB:DSR. To investigate this further, a repeat of the experiment with paired CHIP-seq for each marker at each time-point would need to be performed. Although technically feasible, this approach was outside the scope of this project.

What will not change between the cell lines is the TSS map. The data pertaining to relative enrichment at TSS exhibited a changing pattern as the cells differentiated. In all cases, the relative enrichment at TSS was comparable to that seen in the heterochromatic histone marker. The enrichment was highest for regions within 1Kb of transcription start sites, and lowest within 500bp of TSS. Indeed, in the CHD2 deficient cells, DSBs were depleted 500bp from TSS. At all distances measured from TSS, at all stages of neurodifferentiation, enrichment for DSBs at TSS was higher in WT cells than CHD2 depleted cells.

This data implies that the regions immediately local to TSS are relatively protected from DSBs and that regions 1Kb and 2Kb from TSS are no more enriched than those associated with heterochromatic histone markers; certainly DSBs are under-represented in regions flanking TSS compared to regions defined as enhancers and TF binding sites by their chromatin marker enrichments.

This data is surprising when taken in context of previously published data demonstrating that regions close to TSS are highly susceptible to DSBs[154]. It is also surprising that, given the statistically significant over-representation of DSBs per kb of protein coding gene described in *section 6.4.1*, that the relative enrichment at these sites in CHD2 deficient cells was lower than in wild type cells. This could indicate that the BR:RR ratio described in *section 5.4.2* is higher in CHD2 cells for gene bodies, but only at loci within the gene that are distal to the start site.

As with TSS previously described CFS were no more enriched than heterochromatin markers at D0, D19 or D40 of neurodifferentiation. CFS sites are defined by their increased translocation during replicative stress and so an increased occurrence enrichment may not be evident during normal cellular function[138-141, 332, 333]. CHD2 deficiency can be effectively ruled out as a cause of replication stress in and of itself, however It would be interesting to repeat this experiment with aphidicolin treatment or another exogenous

6.4: Discussion

inducement of replication stress and investigate whether the *response* to replication stress is changed in CHD2 deficiency.

The Wei 27 genes previously identified to harbour recurrent breakpoint clusters was relatively enriched for DSBs at D0, D19 and D40, with greater enrichment in NPCs and neurons compared to IPS cells. Interestingly, the relative enrichment was higher in CHD2 than WT cells at D40 of neurodifferentiation ($p=0.00048$). At D40, the enrichment in the Wei 27 was also higher than for all markers associated with active chromatin in the CHD2 deficient cell line, and with comparable enrichment to these markers in WT cells. This increased enrichment in both cell lines is in keeping with the association between transcription and DSB enrichment demonstrated in *section 6.3.4*.

6.4.3 Relationship between DSB count and GC content

The breaks that were identified in both cell lines occurred at a higher frequency, and had a lower GC content than those that occurred uniquely in only one cell line; the unique breaks were identified at a lower rate, in regions of higher GC content. The shared breaks identified had a GC content below the average of the human genome (40.87%, based on the GRCh38 build of the human genome) [334], whereas the breaks that occurred uniquely had an average GC content higher than the genome-wide average at all time points.

It can therefore be inferred that GC content is protective against DSB formation – or that it increases the speed at which these breaks are repaired. As previously stated, this technique takes a snapshot of the unrepaired breaks in the sample and it is not possible to know whether a greater occurrence or change in the break-repair time is responsible for the under-representation of GC across the samples as a whole.

The GC content was higher in the sequences uniquely associated with breaks in CHD2 deficient cells at D0 ($p=1.54 \times 10^{-9}$) and D19 ($p=0.015$) of neurodifferentiation. The GC content at unique sites in WT and CHD2 deficient cells was almost identical ($p=0.9927$) at D40.

These data are difficult to interpret in any great detail; however, they do suggest that the GC content is less protective in CHD2 deficient cells at D0 and D19 than in WT cells. The fact that this pattern is not repeated at D40 is confounding, and these results should be interpreted with caution until they have been repeated.

6.4.4 Integration of RNA-Seq data

Given the data demonstrating enrichment at sites associated with active chromatin, it is perhaps unsurprising that there is an association between transcription and DSB enrichment in the bodies of protein coding genes. The increase in DSB susceptibility at sites of active transcription is well described [320]. A number of mechanisms are been postulated to explain this relationship.

Firstly, a conflict between the process of DNA replication and transcription. Head on collisions in particular are likely to cause DSBs[335]. Secondary structures known as R loops, in which RNA hybridises back to its template DNA are thought to form during these events obstructing DNA replication and creating foci for DNA damage, including DSB[336, 337].

Backtracking of RNA polymerase complexes is a necessary process in the control of transcriptional elongation[338]. As with R loops, stalled transcription forks triggered by this process can create an impediment to replication and cause DSB to occur, when the replication machinery translocates onto the RNA strand; this leads to SSB, which then develops into DSB during the next round of replication.

Several studies have also linked the occurrence of DSB to the initiation of transcription; cases where increased transcription is effect rather than cause of the DSB. This has been demonstrated in topoisomerase mediated DSB production in neurons[158] and in the response to sex hormones in both breast cancer and prostate cancer[339].

The relationship between transcription and DSB enrichment was stronger in D19 and D40 of neurodifferentiation. Again, this finding correlates with previous studies demonstrating an association between neuronal activity and DSB occurrence. In IPS and NPCs, the association was strongest in WT cells, with CHD2 deficient cells showing the strongest correlation at D40. There difference between the cell lines was minimal at each stage and should be interpreted with caution.

There was a weak correlation between fold enrichment in RNA-Seq data and INDUCE-seq data at D19, but no correlation at D0 or D40. The Pearson coefficient for this correlation was low (0.068) but statistically significant ($p=3.43 \times 10^{-7}$). It is interesting that this correlation occurred at the point in neurodifferentiation where the correlation between transcription itself was at its highest. We can speculate that this may indicate DSBs occurring as a means of genomic regulation during cell differentiation and that a disruption of the genes involved

6: Whole genome assessment of physiological DSB occurrence during neurodifferentiation

6.4: Discussion

in the process could have significant downstream consequences, however further work will be needed to replicate this finding and investigate further.

6.4.5 Summary

This dataset represents the first time that INDUCE-Seq has been used to track the change in a developmental time course. It has been used previously to compare two cell lines cultured under different conditions, however these were conditions with a more predictable outcome – incubation with a known endonuclease – and was performed as a test of the system rather than an avenue of new scientific exploration.

This is also the first time that INDUCE-Seq has been used in a comparison where the outcome was not already expected. Furthermore, the system has only previously been used with HEK cells – this is the first time that any of these cell types have been assessed with this library preparation workflow.

This dataset demonstrates clear correlations between DSB occurrence and chromatin activity and transcription rates. There were also clear changes in the pattern of DSB occurrence at protein coding genes, TSS, and chromatin markers between the WT cells and CHD2 deficient cells. Additionally, a cohort of genes previously demonstrated to be susceptible to DSB under replication stress in neurons was found to be relatively enriched for DSBs in the CHD2 deficient line compared to the WT line, by a small but statistically significant margin.

The most striking feature is the massive increase in breaks which formed in WT cells at D19, and in CHD2 deficient cells at D40. Given the body of evidence described above that suggest a relationship between DSBs and neuronal activity, it is possible that these increases are a crucial aspect of transcriptional regulation during neurodevelopment.

The fact that the increases happened later in CHD2 deficient cells could be significant – although the RNA-seq data suggests no delay in maturation in the CHD2 deficient cell lines – compared to the WT, the late activation of such pathways, or their persistence beyond the required time could be part of a wider transcriptional deregulation that could account for some of the abnormal neurodifferentiation previously demonstrated in CHD2 knockdown [5].

It is well established that chromatin state affects transcription, with certain states associated with highly active regions, and other states associated with repression of transcription. The chromatin markers associated with increased transcription are associated with a higher rate of DSBs in this dataset.

6.4: Discussion

This correlates with the widely accepted link between increased transcription and increased DSB occurrence. Several mechanisms for this, including R loop formation and collision between the apparatus of transcription and replication, have been postulated and are described in detail in the introductory chapter.

There is also an emerging body of evidence, described in detail in the introduction to this thesis, that certain DSBs occur in a programmatic fashion in order to upregulate transcription in response to certain stimuli. In particular they appear to play a role in neuroplasticity cascades. It is not clear if chromatin state affects the rate at which these DSBs occur. It is also not known how widespread such programmatic lesions might be in the developing brain.

Simply put; changes in chromatin architecture can increase transcription, which can lead to an increase in DSB frequency. Programmatic DSBs that upregulate transcription of specific genes have also been described – the relationship between these breaks and chromatin architecture is yet to be explored (*figure 6.12*).

The Pearson coefficients described in *section 6.3.4.2* describe a highly significant relationship between transcription and DSB occurrence, however the relationship only accounts for 8-21.2% of the observed distribution. Some of the remaining variation is likely to be due to stochastic variation in DSB occurrence, however it is not possible to estimate exactly how much of an influence this may be. It is also not possible to absolutely state what fraction of the transcription-related DSBs are an *effect* of increased transcription, and which are the *cause* of increased transcription[158].

Further investigation of the pattern of DSBs that occurs during differentiation of neurons will require more detailed experiments. Before further investigations of pathological mutations are attempted, a more detailed description of the DSB map in wild-type cells would be useful.

A suggested protocol would involve differentiating WT cells into neurons and fixing a set of samples every second day in order to provide a higher resolution mapping of DSB enrichment by developmental time-course. This dataset could then be used as a standard with which to compare analyses from other cell lines (myocyte differentiation for example) or neuronal differentiation in the context of other mutations affecting chromatin remodelling.

7: DISCUSSION

7.1 Introduction

This project aimed to determine whether disrupted DSB repair by NHEJ contributes towards the neurodevelopmental phenotype exhibited in patients with heterozygous mutations in CHD2, known as EECO.

In this thesis I have described an increase in the occurrence of excisions created in an experimental Cas9 system in cells with heterozygous CHD2 mutations. I suggest a model in which the CHD2 mutations slow the repair of DSBs. Based on previously published evidence about the relative rates at which NHEJ and A-EJ occur, I suggest that this prolongation of repair represents an inhibition of the NHEJ pathway of DSB repair.

I also describe the pattern of DSBs that occur in cells as they develop from pluripotent stem cells into mature neurons in laboratory conditions. This evidence noted a striking increase in the number of DSBs that occur as the cells differentiate, with a different pattern of DSBs occurring in the cell line with a heterozygous CHD2 mutation.

In this section I will provide a brief summary of the investigations carried out towards their aim and assess how well those aims have been achieved. A detailed discussion of the statistical significance of the findings can be found in the discussion sections of each results chapters; this will not be repeated here and instead an overview will be offered. I will conclude by discussing the wider implications and impacts of the data presented here, and explore further research suggested by our conclusions.

7.2 Summary of investigations and results

7.2.1 Setup and testing of an inducible Cas9 gene editing and nanopore sequencing pipeline

In *chapter 3*, I described the use of nanopore sequencing for high-throughput screening of CRISPR-Cas9 genome editing experiments involving a doxycycline inducible Cas9 hiPSC cell line. This pipeline was demonstrated to be an effective screening tool, able to discriminate between wells containing heterozygous deletions, heterozygous insertions, wild type samples and wells in which there was a mixture of the samples. Provided there was a read depth of >12 in each well, the results were statistically significant for each sample type.

Effective screening of Cas9 experiments was further demonstrated by presentation of summary data from eight further experiments in which the nanopore sequencing pipeline had been successfully used to identify new mutations in a variety of different targets. The results of these experiments suggested that in keeping with previously published data, transfection of multiple gRNA against the same target resulted in a higher rate of successful editing than transfection of a single gRNA.

The error profile of nanopore sequencing was explored. Based on the data acquired there were not clear patterns in the error profile between different targets, however regardless of the pipeline assessed, similar sequences were over-represented at error sites in the same target. This standardisation of error sites across algorithms, independent of actual sequence makeup is in keeping with previously published data[340].

This evidence suggested that intrinsic features of the amplicons being sequenced, rather than predictable patterns are responsible for the error profile seen in each sequencing run. As the electrical signal generated by each pore of the flow cell can be influenced by a variable number of nucleotides (3nt – 5nt, depending on sequence) and the DNA translocates through the pore at variable speeds, resolution of repeat sequences longer than the length influencing the electrical signal is difficult[281]. It is therefore felt to be likely that the specific repeat sequences within each amplicon, rather than the exact makeup of those sequences is likely to determine which regions of each read are most error prone.

It is suggested that a larger assessment of more diverse genomic regions would yield reproducible errors, however as <3Kb was used in total it was not possible to extrapolate genome wide predictions for error profiles from this experiment.

One weakness of the nanopore screening pipeline is its difficulty in localising insertions, although it was able to correctly identify that an insertion existed in the test samples it was not able to localise the start point. Again, this may be related to the idiosyncrasies of nanopore sequencing chemistry. Insertions at DSB sites are often formed of perfect templated repeats of the flanking area[341] and sequences of di or tri-nucleotide repeats are a challenge for the nanopore platform.

There was no such difficulty with testing for known deletions as except in specific circumstances these do not create tandem repeats. All of the new mutations identified were deletions. This is not in keeping with published data which suggests that small insertions are the most common consequence of NHEJ-repaired Cas9 induced DSBs. This could either be a consequence of the nanopore pipeline's difficulty in localising these insertional lesions,, or of the fact that the majority of these experiments were conducted with multiple gRNAs, making double cut excisions (DCE) more likely.

Given the evidence that repair is reproducible at the same site [258, 259], another possibility is that across the relatively limited number of loci that we investigated, we chose loci more prone to deletion during DSB repair insertion, skewing the dataset towards small deletions.

It was concluded that with the higher fidelity pipeline utilising 1DSQ library preparation resulting in a reduced error rate, the platform would be sufficient for targeted analysis of Cas9 experiments aimed at investigating the fidelity of DSB repair in CHD2 deficient cells.

7.2.2 Creation of a CHD2 mutant and characterisation by RNA-Seq

In *chapter 4*, I described the creation of a CHD2 mutant using the pipeline described in *chapter 3*, and characterised the neurodifferentiation of that cell line using RNA-Seq. The cell line was found to be compound heterozygote, with one frameshift mutation predicted to knockout the gene, and one 6bp in-frame deletion of doubtful effect[307]. It was decided that as human cells are known to be susceptible to haploinsufficiency for CHD2 [16], that regardless of the effect of the in-frame deletion these cells would be deficient in CHD2 and therefore suitable for further investigation.

The RNA-seq demonstrated that both the WT and CHD2 cell lines exhibited the expected transcription profile for genes used as markers of neurodifferentiation at D0 (IPS), D19 (NPC) and D40 (neuron). [125, 295-299, 308-311, 342]

It was also demonstrated that there was greater similarity between biological replicates of the same cell line than the other cell line at each stage of differentiation, indicating that CHD2 deficiency had a detectable impact on the transcriptional profile of cells during neurodifferentiation.

Ontology mapping demonstrated statistically significant fold changes in the transcription of genes pertaining to multiple pathways relevant to neurodevelopment at D19 and D40 of neurodifferentiation, demonstrating that the change in transcription profile seen in CHD2 mutations had the potential to affect the process of neurodifferentiation and neurodevelopment.

It is not possible, from the data we have, to determine if this change in profile would be similar if the 6bp deletion was not present. In this context it is perhaps relevant that the only other study containing published data pertaining to RNA-seq in CHD2 knockdown demonstrated results of greater statistical significance than our study[4], providing further evidence to support the suggestion based on in-silico studies of the in-frame deletion that our cell line may have retained some CHD2 function.

7.2.3 Modelling DSB repair in CHD2 deficiency using targeted genome editing

In *chapter 5* I described the use of inducible Cas9 editing and the transfection of multiple guides targeting genes expressed during neurodevelopment to assess the fidelity of DSB repair at D0 and D40 of neurodifferentiation.

At D0 there was a clear increase in the number of DCEs present in CHD2 deficient cells compared to WT cells. At D40 the pattern was far less clear, failing to meet statistical significance in the majority of targets, however in all cases where DCEs were detected there was an increase in the CHD2 deficient cells. At both D0 and D40, the pattern of smaller indels at each gRNA cut site was dependant on the target sequence, rather than the CHD2 status of the cell line.

It was concluded that the increase in DCEs seen in the CHD2 deficient line was likely due to an increase in ratio of break rate to repair rate – that is either an increase in the break rate, or a prolongation of the break repair time. Given that NHEJ is known to be the most rapid DSB repair pathway[178] and that CHD2 deficiency has been previously demonstrated to inhibit NHEJ [6], it can be inferred that there was a demonstrable inhibition of NHEJ in CHD2 deficient cells at D0 with supporting evidence at D40.

Both the number of reads aligned to the target sequences and the break rates in those sequences were lower in D40 than at D0, which may have obscured more clear results (or indeed more confounding results). It is likely that the combination of difficulty lipofecting mature neurons[235] and the choice of DNA extraction method for these cells may have been responsible for the relatively poor data quality of this run.

The nanopore sequencing platform is also an excellent choice for the investigation of larger DCEs – the ability to sequence single reads of 1Kb and longer allows for detection of low-level mutations of several hundred bp length that would not be resolvable by amplicon sequencing on second generation platforms.

The profile of smaller indel mutations created by cuts with single gRNAs was unhelpful in the reckoning of whether NHEJ had been inhibited, in the CHD2 deficient cell line. 1bp insertions and deletions are the most common indel seen in Cas9 gene editing[258, 262] and also the most common type of sequencing error from our pipeline (*chapter 3*). For the detection of mutations in subcloned cell populations our pipeline works well, however it

proved less robust for identification of low readcount mutations in samples derived from a heterogeneous pooled population.

If this experiment or similar is to be repeated using single gRNAs it would be worth considering using a second-generation sequencing platform to more accurately resolve these smaller lesions.

7.2.4 Whole genome assessment of spontaneous DSB occurrence during differentiation of neurons from induced pluripotent stem cells

In *chapter 6*, I presented data from direct DSB capture during neurodevelopment to explore whether there was any change in the pattern of naturally occurring breaks in CHD2 cell lines.

The most striking finding was that in both cell lines there was a significant increase in the number of DSBs detected over the course neurodifferentiation. In WT cells, this increase occurred at D19 whereas in CHD2 deficient cells, the increase occurred at D40. In both cases, the DSB count at D0 was significantly lower and closer to previously published break-counts per-cell than the break counts at D19 or D40[327].

At all three time-points there was a significant relationship between chromatin state as determined by PTM profile and DSB occurrence. In keeping with previous studies, breaks were enriched at markers for active chromatin and relatively depleted at markers associated with heterochromatin. The enrichment was more significant in the WT cells, however this may reflect a change in the chromatin PTM landscape associated with CHD2 deficiency [3, 118] rather than a genuine change in the level of association with each PTM.

In keeping with many previous observations [154], a significant relationship to transcription, based on paired RNA-Seq data (presented in *chapter 4*) was demonstrated at all three time-points. More significantly so at D19 and D40 than at D0.

It was concluded that there were demonstrable differences in the pattern of breaks regarding association with previously described chromatin PTMs in the CHD2 deficient cells. Although both cell lines demonstrated increases in the break count during neurodifferentiation, the CHD2 DSB spike occurred later than the WT DSB spike. It was postulated that this increase could be related to the recently described physiological function of DSBs relating to transcriptional regulation in stimulated neurons[343].

7.3 Synthesis of results and final conclusions

These studies cannot confirm a direct link between NHEJ, CHD2 deficiency and abnormal neurodevelopment, however we have identified several substantial pieces of evidence of a *possible* link.

First, the evidence of prolongation of repair as indicated by the increased rate of DCEs, seen in CHD2 deficient cells. From this it can be inferred that CHD2 mutations affect DSB repair and that in light of previous evidence demonstrating a link to NHEJ the slower rate of DSB repair is caused by a deficiency in this pathway.

Secondly, it provides evidence, supported by a large body of previously published literature, that DSBs are increased at sites of active transcription. There is evidence that this increase is both pathological (in cases related to stalled replication forks) [320] *and* occurs as part of normal physiological transcriptional regulation[343].

Third, the RNA-seq data provides evidence that the transcriptome during neurodevelopment is perturbed by CHD2 deficiency, with both upregulation and downregulation of genes related to neurodevelopmental processes such as forebrain neuron differentiation, neural nucleus development and proximal pattern formation.

Two possible mechanisms by which CHD2 mutations could impact neurodevelopment via abnormal DSB repair therefore exist:

- 1) Given these changes in transcriptional profile, it can be concluded that a different profile of genes will be susceptible to DSBs in CHD2 deficient cells and that a significant subset of these genes will be genes that contribute towards neurodevelopment.
- 2) Given the likely importance of DSB regulation to neurodifferentiation and the shift in pattern of DSBs seen during neurodifferentiation in CHD2 deficient cells, it is possible that abnormalities in DSB repair could actually affect a change in transcriptional profile[343]

The delay in repair of both of these classes of DSB described in *chapter 5* could also contribute towards abnormal genomic regulation and subsequent cellular dysfunction.

It is also possible that the changes in the transcriptional profiles during neurodevelopment indicated by the results of the RNA-seq data are sufficient to explain the neurodevelopmental delays exhibited by patients with CHD2 mutation however, given the

apparently complex relationship between DSB, NHEJ and chromatin organisation, this would likely be a massive oversimplification of a deeply complex biological system.

Similarly, given the evidence regarding the changing pattern of DSBs detected during normal neurodifferentiation, to suggest that a simple increase in mutational burden caused by over-reliance on A-EJ - as suggested by the hypothesis (*section 1.7*) - is responsible for the phenotype would be reductionist at best.

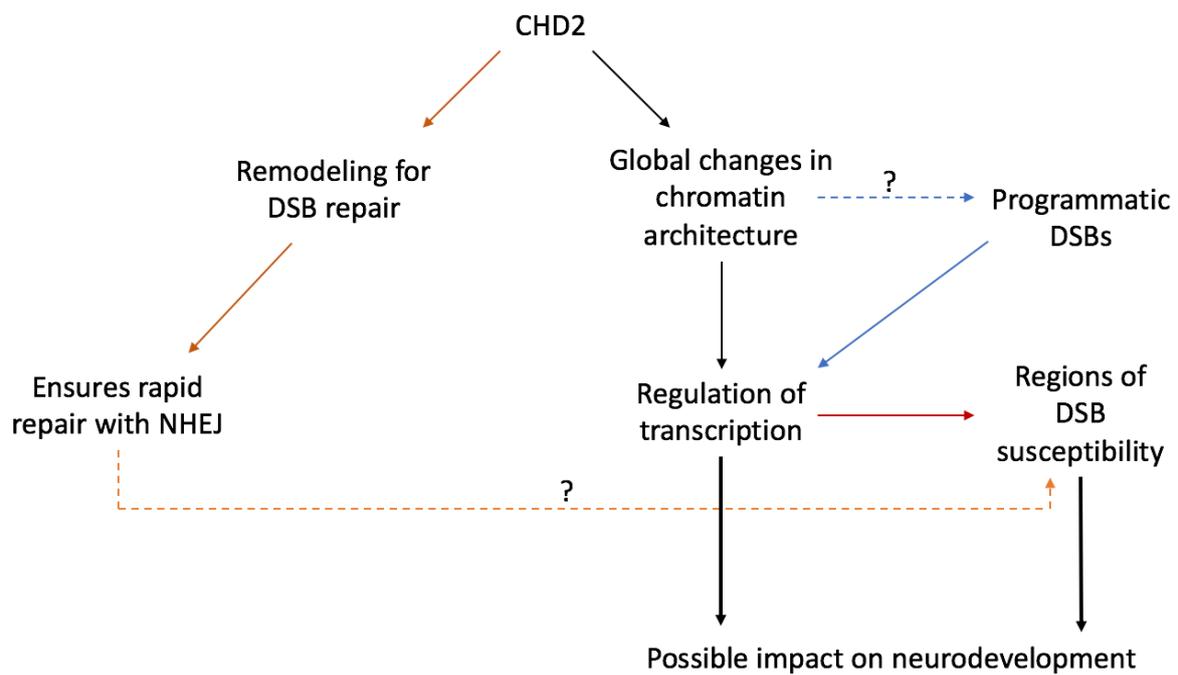


Figure 7.1: Theorised links between CHD2 function, transcription, abnormal DSB repair and profile of DSBs that occur in differentiating neurons

7.4 Impact of findings and suggestions for further research

7.4.1 Use of Nanopore as a high-throughput screen

CRISPR Cas9 is perhaps the most important development in biological research since to polymerase chain reaction. In the few years since successful intracellular editing was first demonstrated, it has revolutionised the investigation of the mechanisms underpinning of genetic disease in human cell lines and animal models. If technical challenges regarding the risk of off target mutations and effective protein delivery can be solved, it has the potential to revolutionise the treatment of both congenital and acquired genetic diseases.

As such, establishing high throughput pipelines for the analysis of Cas9 mediated genome editing experiments is highly desirable. The Oxford Nanopore Minion is an affordable sequencing platform and despite the problems highlighted with resolution of smaller indels, we have successfully used this sequencing pipeline to generate mutations in multiple genomic targets. The pipeline is simple enough that new staff can be trained in its use in a matter of months.

At the time of writing, it is the fastest pipeline currently described for the creation and validation of Cas9 mediated mutations. The low cost of the sequencer and library preparation reagents mean that it is affordable on modest laboratory budgets. Publication and presentation of this pipeline and the supporting data has the potential for significant impact in the scientific community.

As discussed above the sequencing experiment proved less robust for analysing the repair at DSB sites. For identification and quantification of DCEs it was sufficient, however the inherent error rate of nanopore sequencing made it more difficult to interpret the data regarding indels, and so this part of the research was deemed inconclusive.

It will be useful to follow the development of nanopore sequencing; if the R10 pore delivers the promised improvement in read accuracy and if this improvement is present in library preparations involving smaller genomic regions (our 1Kb amplicons, rather than the usual 10-100Kb), then the technical drawbacks may diminish significantly.

7.4.2 Understanding of CHD2's role in DSB repair

Taken together, *chapters 5 and 6* demonstrate clear differences in DSB repair in CHD2 deficient cells. Further work is needed in order to describe this in more detail, particularly as regards the occurrence of smaller indels at single gRNA cut sites. Single cut sites are likely a closer analogue of DSBs that occur under physiological circumstances in human cells than the DCEs.

The modelling of DCEs is important data and will be prepared for publication. With further development, the model presented in the discussion section could provide a reproducible framework for analysis of DNA repair timing that could be used in the analysis of a wide variety of cell lines and culture conditions.

As suggested in *section 7.2.3*, a repeat of the targeted genome editing experiment using technology able to resolve smaller lesions could provide further evidence of the impact of NHEJ inhibition in CHD2 deficient cells. The protocol need not be particularly radical; studies oft quoted in this thesis have done similar experiments with lentiviral gRNA libraries[151, 258]. The only difference would be to provide a comparison between WT cells and CHD2 cells. An argument could even be made that a standardised lentiviral gRNA library would be a useful resource for investigating DSB repair not only CHD2 mutations, but the plethora of genes that mediate and interact with DNA repair pathways.

Another way to increase the power of the study would be to compare to a cell line with a knockout of a protein known to completely abrogate NHEJ, such as DNA ligase IV[159, 160, 324, 344-346]. This would allow for description of a scale of NHEJ function, ranging from normal in the WT cells to complete abrogation in the LigIV^{-/-} line. Based on the results described here, it is suggested that CHD2 mutants would fall somewhere in the middle of such a spectrum.

7.4.3 The contribution of DSB to neurodevelopment

The finding of 10-fold to 20-fold enrichment for DSBs during the process of neurodevelopment is perhaps the most noteworthy output from this project in terms of potential impact on our understanding of genomics and neurology.

Multiple lines of evidence regarding the relationship of these breaks to transcription and their association with active chromatin states suggest that they are indeed genuine, rather than an artefact of the sample preparation. Taken together with other evidence that DSBs have physiological roles in transcriptional regulation and cellular differentiation, along with the evidence of a high rate of somatic mosaicism in adult cells, these findings may prove a fertile ground for further research endeavour.

As a next stage in investigation, repeating the WT neurodifferentiation, with samples taken at regular time points, from D0 to D50 would allow for a better understanding of the how the enrichment and profile of DSBs changes during neurodifferentiation.

Once this has been fully established, it will form a valuable resource for investigating the impact of mutations in a wide variety of DNA-damage-repair genes and chromatin remodellers on the stability of the developing genome in the brain. Understanding the mechanisms by which transcription is regulated in concordance with DSBs in developing neurons also has the potential to significantly advance our understanding of global CNS development.

7.4.4 Patient cell lines

Regarding the impact of pathogenic heterozygous CHD2 mutations on DSB repair in humans and whether this contributes towards the phenotype seen in patients, once the finer details of the cellular model have been fleshed out with the further experiments described above, recapitulating the findings with patient cell lines will be important in understanding the relevance of these findings to the EECO syndrome.

In addition, it would be useful to compare the somatic genomes of older patients harbouring CHD2 mutations with those of healthy age-matched individuals. If a significant additional burden of acquired somatic mutations can be demonstrated, then there are direct implications for patient management as increased somatic mutations could lead to an increased risk of malignancy.

7.5 Concluding remarks

This work provides significant evidence that CHD2 deficiency can cause a change in the dynamics of DSB repair and the profile of physiological DSB occurrence in a cell model of human development. It also provides evidence of a detectable change in the transcriptomes of CHD2 deficient cells during neurodifferentiation. The shift in the transcriptome and the change in the profile of physiological DSBs may be caused by overlapping mechanisms and may influence each other in various ways (*figure 7.1*). Both are felt to be likely contributors to the syndrome of EECO exhibited in human patients with heterozygous CHD2 mutations.

In order to further investigate the role of CHD2 in neurodevelopment, additional sequencing techniques including CHIP-Seq targeting CHD2 and 3C to investigate the impact of CHD2 on the 3D genome could be utilised.

The techniques and findings described in this thesis regarding DSBs in normal neurodevelopment have the potential for significant impact on our understanding of CNS development. A strong case can therefore be made for further investigation in this rapidly developing field. In particular, assembly of a reproducible map of DSBs that occur during neurodevelopment could provide a foundation for further work.

APPENDICES

APPENDIX I: PYTHON SCRIPTS

This section is not an exhaustive compendium of the script used to generate every table and figure in this project – such a compendium would double the length of the thesis for little added value. Where something novel or unique has been attempted however, the code is included here for reference.

I: CRISPR NANOSCREEN

CRISPR nanoscreen iterates over files running bam_readcount, modifying the headers, then screening with two scripts; deletion caller and insertion caller, to generate two reports for each bamfile.

Included here are the parent script, and deletion_caller.sh. The insertion caller script is essentially identical to the deletion caller and therefore is not included here

II: SLIDING WINDOW ERROR ANALYSIS

Several scripts were used, depending on the desired data-output format. All scripts followed the same basic principal – illustrated in the script on the next page, which outputs % for each mononucleotide, dinucleotide and trinucleotide within the reference sequence and the errors identified at each sequence.

CRISPR NANOSCREEN

#CRISPR_nanoscreen.py

#!/bin/bash

#\$1 = gpath (path to workspace)

#\$2 = gene

#\$3 = cutsite

#\$4 = sens

low=\$(((\$3 - 100))

high=\$(((\$3 + 100))

avlow=\$(((\$low+30))

avhigh=\$(((\$high-30))

```
#####  
#####DELETIONS#####  
#####
```

cd \$1

touch \$1/\$2_Deletion_report.txt

echo \$2 Deletion Report >> \$1/\$2_Deletion_report.txt

echo >> \$1/\$2_Deletion_report.txt

echo -e "WELL\tR_D\tD_D\tD_% " >> \$1/\$2_Deletion_report.txt

mkdir \$1/\$2_Count_del

mkdir \$1/\$2_Head_del

for file in \$1/\$2_Sort_del/*.bam; do

file=\${file%*}

bam-readcount -f \$1/\$2_Ref/\$2_Ref.fasta \$file -i FADS2:\$low-\$high >> \$file.txt

mv \$file.txt \$1/\$2_Count_del

done

for file in \$1/\$2_Count_del/*.txt; do

file=\${file%*}

header.sh \$file

mv \$file.out.txt \$1/\$2_Head_del

done

for file in \$1/\$2_Head_del/*; do

file=\${file%*}

echo \$file

deletion_caller.sh \$file \$avlow \$avhigh \$4 4 >> \$1/\$2_Deletion_Report.txt

```
done

echo >> $1/$2_Deletion_report.txt
echo >> $1/$2_Deletion_report.txt
echo >> $1/$2_Deletion_report.txt
echo >> $1/$2_Deletion_report.txt
echo R_D = Read depth - average read depth \in this well \for 200bp around \cut site >>
$1/$2_Deletion_report.txt
echo D_D = Deletion depth - the maximum number of reads \in which a deletion is recorded,
at any point within 200bp of the \cut site >> $1/$2_Deletion_report.txt
echo D_% = Maximum % of reads \in which a deletion is recorded, at any point within 200bp
of the \cut site >> $1/$2_Deletion_report.txt
echo >> $1/$2_Deletion_report.txt
echo >> $1/$2_Deletion_report.txt
echo In calibration data, wells with a read depth \>12 were reliably able to distinguish
between heterozygous deletions, mixed mutations \dirty clones\ and \false positives >>
$1/$2_Deletion_report.txt

cp $1/$2_Deletion_report.txt $1/$2_Deletion_report.xls

#####
#####INSERTIONS#####
#####

touch $1/$2_Insertion_report.txt
echo $2 Insertion Report >> $1/$2_Insertion_report.txt
echo >> $1/$2_Insertion_report.txt
echo -e "WELL\tR_D\tI_D\tI_% " >> $1/$2_Insertion_report.txt

mkdir $1/$2_Count_ins
mkdir $1/$2_Head_ins

for file in $1/$2_Sort_ins/*.bam; do
    file=${file%*}

    bam-readcount -f $1/$2_Ref/$2_Ref.fasta $file -i FADS2:$low-$high >> $file.txt
    mv $file.txt $1/$2_Count_ins
done

for file in $1/$2_Count_ins/*.txt; do
    file=${file%*}
    header.sh $file
    mv $file.out.txt $1/$2_Head_ins
done

for file in $1/$2_Head_ins/*; do
    file=${file%*}
```

```
insertion_caller.sh $file $avlow $avhigh $4 4 >> $1/$2_Insertion_Report.txt  
done
```

```
echo >> $1/$2_Insertion_report.txt  
echo >> $1/$2_Insertion_report.txt  
echo >> $1/$2_Insertion_report.txt  
echo >> $1/$2_Insertion_report.txt  
echo R_D = Read depth - average read depth \in this well \for 200bp around \cut site >>  
$1/$2_Insertion_report.txt  
echo I_D = Insertion depth - the maximum number of reads \in which an Insertion is recorded,  
at any point within 200bp of the \cut site >> $1/$2_Insertion_report.txt  
echo I_% = Maximum % of reads \in which an Insertion is recorded, at any point within 200bp  
of the \cut site >> $1/$2_Insertion_report.txt  
echo >> $1/$2_Insertion_report.txt  
echo >> $1/$2_Insertion_report.txt  
echo In calibration data, wells with a read depth \>12 were reliably able to distinguish  
between heterozygous Insertions, mixed mutations \(\dirty clones\) and \false positives >>  
$1/$2_Insertion_report.txt  
  
cp $1/$2_Insertion_report.txt $1/$2_Insertion_report.xls
```

DELETION CALLER

```
#deletion_caller.sh
#!/usr/bin/env python3

import sys
import os
import numpy
import re
import pandas as pd
import plotly.plotly as py
import plotly.graph_objs as go

#identifies file from parent loop script
myFile = sys.argv[1]
FirstBase = sys.argv[2]
LastBase = sys.argv[3]
Sensitivity = sys.argv[4]
MinLength = sys.argv[5]

Sens = int(Sensitivity)
MinLen = int(MinLength)
NameFile = sys.argv[1]
Name = (NameFile.split(".")[0])
UpstreamAvrg = (int(sys.argv[2]) - 20)
DownstreamAvrg = (int(sys.argv[3]) + 20)

#pulls tld into dataframe
df = pd.read_csv(myFile, sep='\t')
df2 = df.set_index("RefPos ", drop = False)

#gets average readcounts for comparison
df3 = (df2.loc[DownstreamAvrg:DownstreamAvrg + 20,"ReadCount "])
DS_Av = (int(round(df3.mean()))))

df4 = (df2.loc[UpstreamAvrg:UpstreamAvrg + 20,"ReadCount "])
US_Av = (int(round(df4.mean()))))

Av = (int(round((DS_Av + US_Av)/2)))

ref_range = (range(int(sys.argv[2]),(int(sys.argv[3])+150)))
ins_dict = {}.fromkeys((ref_range), 0) #sets up the dictionary for insertion calling
del_dict = {}.fromkeys((ref_range), 0) #sets up the dictionary for deletion calling

#set up variables for RefPos, RefSeq, A, C, T, G, Call1, Call2, Call3, Call4
for x in range (int(sys.argv[2]),int(sys.argv[3])):
    RefPos = (x)
    RefSeq = (df2.loc[int(x),"RefBase "])
```

```
ReadCount = (df2.loc[int(x),"ReadCount "])
Call1 = (df2.loc[int(x), "Call1 "])
Call2 = (df2.loc[int(x), "Call2 "])
Call3 = (df2.loc[int(x), "Call3 "])
Call4 = (df2.loc[int(x), "Call4 "])
Call5 = (df2.loc[int(x), "Call5 "])
Call6 = (df2.loc[int(x), "Call6 "])
Call7 = (df2.loc[int(x), "Call7 "])
Call8 = (df2.loc[int(x), "Call8 "])
Call9 = (df2.loc[int(x), "Call9 "])
Call10 = (df2.loc[int(x), "Call10 "])
Call11 = (df2.loc[int(x), "Call11 "])
Call12 = (df2.loc[int(x), "Call12 "])
Call13 = (df2.loc[int(x), "Call13 "])
Call14 = (df2.loc[int(x), "Call14 "])
Call15 = (df2.loc[int(x), "Call15 "])
Call16 = (df2.loc[int(x), "Call16 "])
Call17 = (df2.loc[int(x), "Call17 "])
Call18 = (df2.loc[int(x), "Call18 "])
Call19 = (df2.loc[int(x), "Call19 "])
Call20 = (df2.loc[int(x), "Call20 "])
Call21 = (df2.loc[int(x), "Call21 "])
Call22 = (df2.loc[int(x), "Call22 "])
Call23 = (df2.loc[int(x), "Call23 "])
Call24 = (df2.loc[int(x), "Call24 "])
Call25 = (df2.loc[int(x), "Call25 "])
Call26 = (df2.loc[int(x), "Call26 "])
Call27 = (df2.loc[int(x), "Call27 "])
Call28 = (df2.loc[int(x), "Call28 "])
Call29 = (df2.loc[int(x), "Call29 "])
Call30 = (df2.loc[int(x), "Call30 "])
Call31 = (df2.loc[int(x), "Call31 "])
Call32 = (df2.loc[int(x), "Call32 "])
Call33 = (df2.loc[int(x), "Call33 "])
Call34 = (df2.loc[int(x), "Call34 "])
Call35 = (df2.loc[int(x), "Call35 "])
Call36 = (df2.loc[int(x), "Call36 "])
Call37 = (df2.loc[int(x), "Call37 "])
Call38 = (df2.loc[int(x), "Call38 "])
Call39 = (df2.loc[int(x), "Call39 "])
Call40 = (df2.loc[int(x), "Call40 "])
Call41 = (df2.loc[int(x), "Call41 "])
Call42 = (df2.loc[int(x), "Call42 "])
Call43 = (df2.loc[int(x), "Call43 "])
Call44 = (df2.loc[int(x), "Call44 "])
Call45 = (df2.loc[int(x), "Call45 "])
Call46 = (df2.loc[int(x), "Call46 "])
```

```
Call47 = (df2.loc[int(x), "Call47 "])
Call48 = (df2.loc[int(x), "Call48 "])
Call49 = (df2.loc[int(x), "Call49 "])
Call50 = (df2.loc[int(x), "Call50 "])
Call51 = (df2.loc[int(x), "Call51 "])
Call52 = (df2.loc[int(x), "Call52 "])
Call53 = (df2.loc[int(x), "Call53 "])
Call54 = (df2.loc[int(x), "Call54 "])
Call55 = (df2.loc[int(x), "Call55 "])
Call56 = (df2.loc[int(x), "Call56 "])
Call57 = (df2.loc[int(x), "Call57 "])
Call58 = (df2.loc[int(x), "Call58 "])
Call59 = (df2.loc[int(x), "Call59 "])
Call60 = (df2.loc[int(x), "Call60 "])
Call61 = (df2.loc[int(x), "Call61 "])
Call62 = (df2.loc[int(x), "Call62 "])
Call63 = (df2.loc[int(x), "Call63 "])
Call64 = (df2.loc[int(x), "Call64 "])
```

```
Call_list = [Call1, Call2, Call3, Call4, Call5, Call6, Call7, Call8, Call9, Call10, Call11, Call12,
Call13, Call14, Call15, Call16, Call17, Call18, Call19, Call20, Call21, Call22, Call23, Call24, Call25,
Call26, Call27, Call28, Call29, Call30, Call31, Call32, Call33, Call34, Call35, Call36, Call37, Call38,
Call39, Call40, Call41, Call42, Call43, Call44, Call45, Call46, Call47, Call48, Call49, Call50, Call51,
Call52, Call53, Call54, Call55, Call56, Call57, Call58, Call59, Call60, Call61, Call62, Call63,
Call64]
```

```
for x in Call_list:
```

```
    if isinstance(x, float):
```

```
        y = 1
```

```
    elif (x[0]) is "-":
```

```
        Del = (int(re.search(r'\d+', x).group())) #number of reads containing an insertion or
deletion
```

```
        DelName = (x.split(":")[0])
```

```
        DelLen = ((len(DelName)) -1)
```

```
        if DelLen >= MinLen:
```

```
            del_range = range(RefPos, (RefPos + DelLen))
```

```
            for x in del_range:
```

```
                del_dict[x] = (del_dict[x] + Del)
```

```
xde = list(del_dict.keys())
```

```
yde = list(del_dict.values())
```

```
dMAX = max(yde)
```

```
max_dperc = ((dMAX/Av) * 100)
min_dfloat = float(sys.argv[4])
max_dround = (round(max_dperc, 2))

Wellsplit = Name.split("_")
Well = Wellsplit[-2]+Wellsplit[-1]

if max_dperc > min_dfloat:
    print (Well, "\t", Av, "\t", dMAX, "\t", max_dround)
```

SLIDING WINDOW ANALYSIS

```
#!/usr/bin/env python3
#sliding_window_analysis_perc.py

import sys
import numpy as np
import pandas as pd
import re
from matplotlib import pyplot as plt
#geno_sameness.py():

bamrc = sys.argv[1]
seq_file = sys.argv[2]

#reads reference sequence
with open (seq_file) as file:
    refseq = list((file.readlines()[1]).upper())

#seq_one = list(seq_one)
#length = len(seq_one)

Nuc = ["A", "C", "G", "T"]

#####
#### DICT SETUPS####
#####
#Create dictionaries for each
mono = {"A":0, "C":0, "G":0, "T":0}
dinuc = {}
trinuc = {}
tetranuc = {}
pentanuc = {}
hexanuc = {}

dinuc_list = []
trinuc_list = []
tetranuc_list = []
pentanuc_list = []
hexanuc_list = []

for i in Nuc:
    for n in Nuc:
        x = i + n
        dinuc_list.append(x)
        dinuc[x] = 0
```

```
###
for i in Nuc:
    for n in dinuc_list:
        x = i + n
        trinuc_list.append(x)
        trinuc[x] = 0

###
for i in Nuc:
    for n in trinuc_list:
        x = i + n
        tetranuc_list.append(x)
        tetranuc[x] = 0

""""###
for i in Nuc:
    for n in tetranuc_list:
        x = i + n
        pentanuc_list.append(x)
        pentanuc[x] = 0

###
for i in Nuc:
    for n in pentanuc_list:
        x = i + n
        hexanuc_list.append(x)
        hexanuc[x] = 0""""

#####
num_lines = sum(1 for line in open(bamrc))
ref_range = range(num_lines+30)
DelDict = {}.fromkeys((ref_range), 0)

with open (bamrc) as file:

    for line in file:
        data = line.split()
        refpos = int(data[1])

        for call in data:
            if call[0] is '-':
                DelName = (call.split(":")[0])
                DelLen = ((len(DelName)) -1)
```

```
        DelRange = range(refpos,refpos+DelLen)
        DelDep = int((call.split(":")[1]))
        for x in DelRange:
            DelDict[x] = DelDict[x] + DelDep
#print (DelDict)
#print (len(refseq))

for x in DelDict:
    if x < len(refseq)-1:
        count = DelDict[x]
        seq = refseq[x]
        mono[seq] += count

for x in DelDict:
    if x < len(refseq)-2:
        count = DelDict[x]
        seq = refseq[x:x+2]
        seq = ".join(seq)
        dinuc[seq] += count

for x in DelDict:
    if x < len(refseq)-3:
        count = DelDict[x]
        seq = refseq[x:x+3]
        seq = ".join(seq)
        trinuc[seq] += count

SeqRange = range(5, num_lines)
#print (SeqRange)

mon_all = sum (mono.values())

def percs (dic):
    dic_all = sum (dic.values())
    for i in dic:
        print (i, dic[i], (dic[i] / dic_all))

print()
percs (mono)
print ()
percs (dinuc)
print ()
percs (trinuc)
```

III: DOUBLE CUT EXCISION MODELLING

The images were generated using the following script:

```
#!/usr/bin/env python3
#MODEL_TWO.py

from random import *
from mpl_toolkits.mplot3d import Axes3D
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns

dataset_a = range(0,1000)
dataset_b = range(0,1000)

def simulate (dataset, BR, RR):
    set = []
    for x in dataset:
        r = random()
        if r < BR: #break occurs
            n = 1
        else:
            n = 0
        set.append(n)

    count = 0

    for n in set:
        count += 1

        if n == 1:
            r = random()
            if r > RR: #then repair fails
                if count < len(set):
                    set[count] = 1

    return (set)

fig = plt.figure()
ax = fig.add_subplot(111, projection='3d')

THEX = []
THEY = []
THEZ = []
counter = 0
```

```
the_count = 0

while the_count < 1000:
    print ("while count", the_count)
    BR_A = random()
    BR_B = random()
    RR_A = random()
    RR_B = random()

    ratio_A = BR_A / RR_A #HIGH = lots of cut
    ratio_B = BR_B / RR_B #HIGH = lots of cut
    set_a = simulate(dataset_a, BR_A, RR_A)
    set_b = simulate(dataset_b, BR_B, RR_B)
    double_cut = []

    counter = 0
    print ("ratio_A", ratio_A)
    print ("ratio_B", ratio_B)
    if ratio_A < 1.0 and ratio_B < 1.0:
        print ("BOTH")
        double_cut = []
        print ("set_a", set_a)
        print ("set_b", set_b)
        for i in set_a:
            if i == 1:
                if counter < len(set_b):
                    if set_b[counter] == 1:
                        double_cut.append(1)
                    else:
                        double_cut.append(0)
                else:
                    double_cut.append(0)
            counter += 1
        print ("double_cut", double_cut)

    graph = []
    tots = []
    sumit = 0
    for x in double_cut:
        if x == 1:
            sumit += 1
        if x == 0:
            tots.append(sumit)
            sumit = 0
        print (x, sumit)

    print("Totals", tots)
```

```
for i in tots:
    if i > 0:
        graph.append(i)
    graph.append(0)
    print (graph)

    x = ratio_A
    z = np.mean(graph)
    y = ratio_B

    THEX.append(x)
    THEY.append(y)
    THEZ.append(z)
else:
    THEX.append(x)
    THEZ.append(0)
    THEY.append(y)
the_count += 1

ax.plot_trisurf(np.array(THEX),np.array(THEY),np.array(THEZ), cmap=plt.cm.CMRmap)
ax.set_xlabel('RATIO BR_A : RR_A')
ax.set_ylabel('RATIO BR_B : RR_B')
ax.set_zlabel('Mean persistence of double cut')
plt.show()
```

APPENDIX II: GLOSSARY OF ABBREVIATIONS

3C	Chromatin conformation capture
6WP	6 well plate
12WP	12 well plate
24WP	24 well plate
96WP	96 well plate
aa	Amino Acid
AAVS	adeno-associated-virus integration site
ACMG	American college of medical genetics
AD	Autosomal dominant
A-EJ	Alternative end joining
aNHEJ	Alternative non-homologous end-joining
ASD	Autistic spectrum disorder
ATM	Ataxia telangiectasia mutated
Bam	Binary alignment map
BLISS	Breaks labelling in-situ and sequencing
Bp	Base pair
Bwa	Burrows-Wheeler aligner
CDK	Cyclin dependant kinase
CAF1	Chromatin assembly factor 1
Cas9	CRISPR associated protein 9
CDK	Cyclin dependant kinase
CFS	Common fragile sites
CHD2	Chromodomain Helicase DNA-binding 2
CHIP-Seq	Chromatin Immuno-precipitation sequencing
CIGAR	Concise idiosyncratic gapped alignment report
cIN	Cortical interneurons
iCn	Inducible CRISPR nuclease
iCn-CHD2 ^{+/-}	Inducible CRISPR cell line with heterozygous mutation in CHD2
iCn-WT	Inducible CRISPR cell line with wild type CHD2 sequence
CLL	Chronic Lymphocytic Leukaemia
cNHEJ	Classical non-homologous end-joining
CNS	Central Nervous System
CNV	Copy number variant
coIP	co-immunoprecipitation
crRNA	CRISPR-RNA
CRISPR	Clustered regularly interspersed short palindromic repeats
CSR	Class switch recombination
DCE	Double-cut excisions
DDD	Deciphering developmental disorders study
ddODN	Double-strand deoxyoligonucleotide
ddNT	dideoxynucleotides
DEE	Developmental Epileptic Encephalopathy
DEG	Differently expressed genes
DESeq2	Differential expression seq-2
DNA-PKcs	DNA dependant protein catalytic subunit
DO	Day 0 of neurodifferentiation

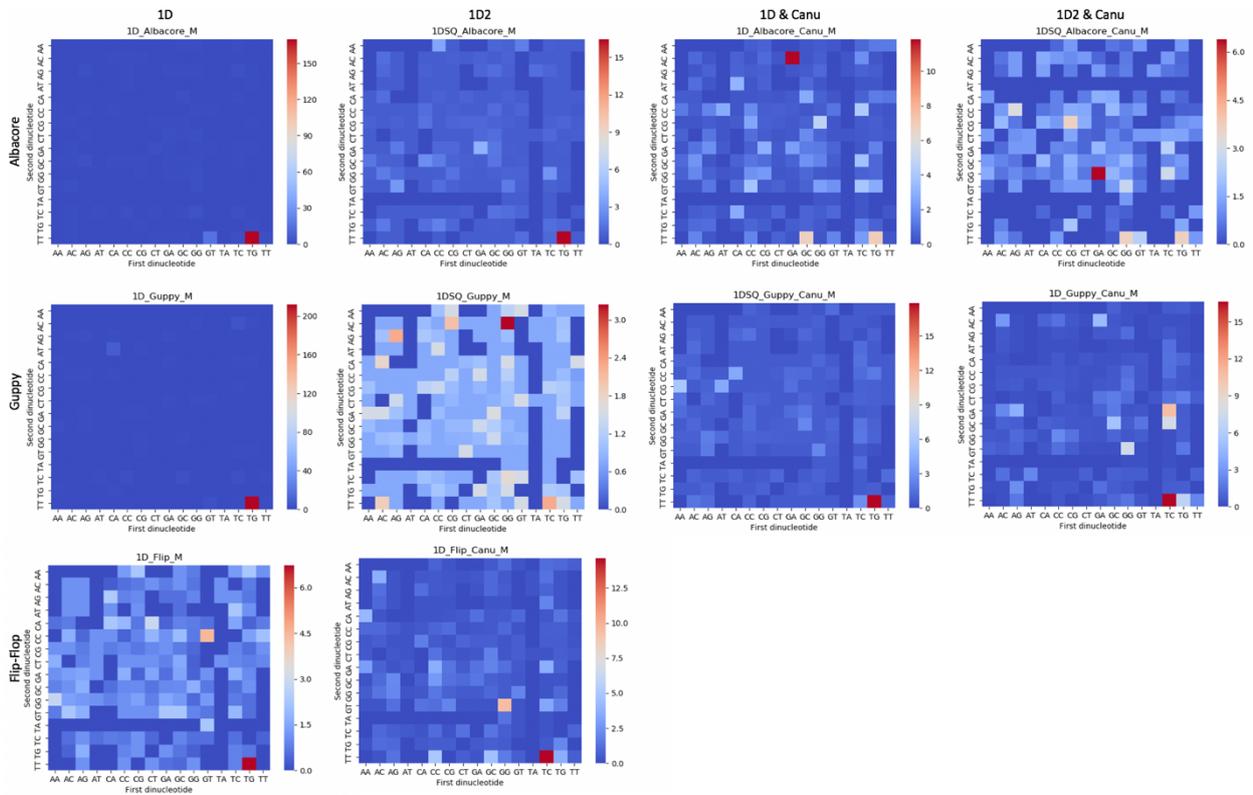
D19	Day 19 of neurodifferentiation
D40	Day 40 of neurodifferentiation
DSB	Double Strand Break
E8	Essential 8 medium
EE	Epileptic Encephalopathy
EEOC	Epileptic Encephalopathy of Childhood Onset
EEG	Electroencephalogram
emPCR	Emulsion PCR
ENCODE	Encyclopaedia of DNA elements
ExAC	Exome aggregation consortium
FBC	Full blood count
FDR	False discovery rate
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
Gb	Gigabase
GDD	Global developmental delay
GFP	Green fluorescent protein
GO	Gene ontology
gRNA	guide RNA
GUIDE-Seq	Genome-wide unbiased identification of DSBs enabled by sequencing
GTCS	Generalised tonic-clonic seizure
GWAS	Genome wide association studies
HiC	High-throughput chromatin conformation capture
hESC	Human embryonic stem cells
HGVS	Human genetic variation society
hiPSC	human Induced Pluripotent Stem Cell
H3.3	Histone marker 3.3
H3.1	Histone marker 3.1
H2AX	Histone 2A-X
H2A	Histone 2A
H2B	Histone 2B
HiC	High throughput chromatin conformation capture
HR	Homologous recombination
HSW	High-salt wash
HTS	High throughput sequencing
ID	Intellectual disability
IGB	Integrated genome browser
IgH	Immunoglobulin heavy chain
Igv	Integrated genomics viewer
ILAE	the International league against epilepsy
INDUCE-seq	Identification and quantification of DSBs by unbiased flow cell enrichment
Indel	DNA insertion or deletion
IP cell	Intermediate progenitor cell
Kb	Kilobase
KU	Ku70/80 complex
LB1	Lysis buffer 1
LB2	Lysis buffer 2
LD	Learning difficulties (eg dyslexia, dyscalculia)

LFC	Log2 fold change
LigIV	DNA Ligase IV
LigIII	DNA Ligase III
Mb	Megabase
MBL	Monoclonal B-lymphocytosis
MEF	Mouse Embryonic Fibroblasts
mESC	Mouse embryonic stem cells
MGE	Medial ganglionic eminence
miRNA	micro RNA
MMEJ	Microhomology mediated end-joining
MN	Mature neurons
MRD	Minimum read depth
mRNA	messenger RNA
NCBI	National center for biotechnology information
ncRNA	non-coding RNA
NEB	New England Biosciences
NFDB	Nuclease-free duplex buffer
NFW	Nuclease free water
NgImr	convex gap-cost alignments for long reads
NGS	Next generation sequencing
NHEJ	Non-Homologous End Joining
NIH	National institutes for health
NMHRI	Neuroscience and mental health research institute
NPC	Neuronal progenitor cells
NSPC	Neuronal stem progenitor cells
Nt	Nucleotide
OFC	Occipito-frontal circumference
ONP	Oxford Nanopore
oRG cell	Outer radial glial cell
PacBIO	Pacific biosciences
PAM	Protospacer adjacent motif
PAR	Poly-A-Ribose
PBS	Phosphate buffered saline
PCA	Principal component analysis
pDCE	predicted Double cut excision
PDL	Poly-D-Lysine
PFA	Paraformaldehyde
PTM	Post-translational modification
QC	Quality control – used pertaining to sequencing data
RBF	Random break file
RDC	Recurrent DSB cluster
RFS	Rare fragile sites
RG cell	Radial glia cell
Sam	Sequence alignment map
SCW	Super-computing Wales
Sd	Standard deviation
SEM	Standard error of the mean

SGS	Second generation sequences
siRNA	Short interfering RNA
SMRT	Single molecule real-time
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SOC	Super-optimal broth with catabolite repression
SPRI	Solid phase reversible immobilisation beads
ssODN	Single-strand deoxyoligonucleotide
STAR-aligner	Spliced transcripts alignment to reference
TAD	Topographically associated domain
TALENs	Transcription activator-like nucleases
tracrRNA	trans-activating-CRISPR RNA
tD	Dissociation constant for double-cut excisions
tR	Repair time constant for CRISPR induced DSB
tB	Break time constant for CRISPR induced DSB
TC-Seq	Translocation-capture sequencing
TGS	Third-generation sequencers
TSS	Transcription start site
UV	Ultraviolet
VUS	Variant of uncertain significance
VZ	Ventricular zone
WES	Whole exome sequencing
WGS	Whole genome sequencing
WT	Wild Type
XRCC4	X-ray cross complimenting 4
ZF	Zinc-finger
ZMW	Zero mode waveguide

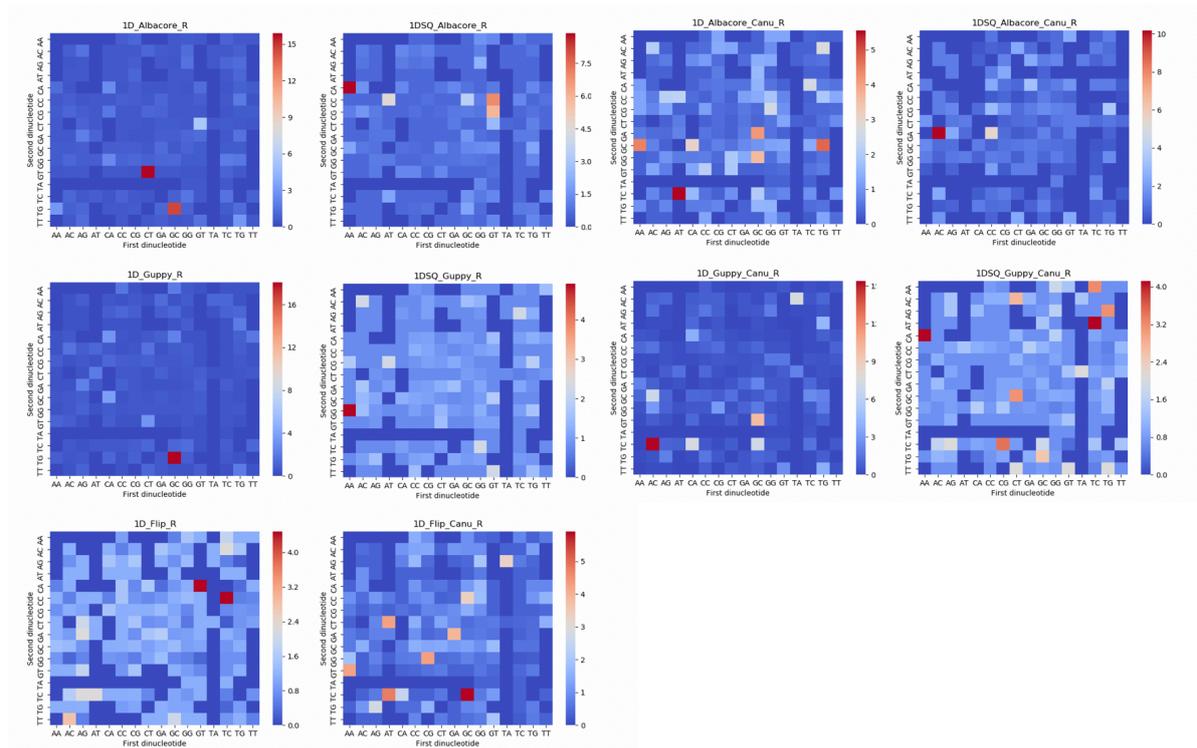
APPENDIX III: SUPPLEMENTARY DATA

CHAPTER 3



Bilateral tetranucleotide context of miscalled deletions in WT NRXN1 – aligned using minimap2. Higher numbers show overrepresented sequences. x = downstream dinucleotide, y = upstream dinucleotide (see section 3.3.5.3)

APPENDIX III: SUPPLEMENTARY DATA



Reverse tetranucleotide context of miscalled deletions in WT NRXN1 – aligned using minimap2. Higher numbers show overrepresented sequences. x = first dinucleotide, y = second dinucleotide (see section 3.3.5.3)

1D Albacore			1D Guppy			
	CHD2	FADS2	NRXN1	CHD2	FADS2	NRXN1
	Trinucleotide, ratio					
1	AGC 4.362821	GCC 1.663503	TTT 36.975633	CTT 2.559257	TCT 2.398233	TTT 43.316509
2	GAC 1.922873	CTA 1.663503	GAT 1.987937	CAT 2.533406	CGA 2.129774	TCA 3.205586
3	TGC 1.583543	TAC 1.397342	GTC 1.905107	AGC 1.895746	GGG 1.977648	ATG 2.383641
4	TCG 1.555265	TCT 1.324279	TCC 1.424688	TAC 1.868664	TTC 1.778713	GTA 1.643890
5	GAG 1.539555	TGT 1.280897	TCT 1.219871	ACA 1.868664	GAC 1.521267	TCG 1.232918
6	AAT 1.450761	AGC 1.242082	CCA 1.146887	TCG 1.492900	ATC 1.460417	TCC 1.150723
7	AGG 1.413877	CAA 1.242082	TTC 1.098597	ACC 1.421809	ATG 1.399566	ACC 1.056787
8	CTG 1.357322	TCC 1.207580	GCC 1.026032	TGG 1.356187	CTG 1.379282	GTC 1.027431
9	ATG 1.319619	CTC 1.181328	GGT 1.025030	AGG 1.327022	GGC 1.369141	AGA 0.958936
10	TAG 1.319619	GGA 1.164452	CTG 1.011106	CTG 1.175362	GAG 1.293077	CAC 0.958936

Top ten overrepresented trinucleotide sequences, upstream of miscalled deletions (see section 3.3.5.3)

1D Albacore			1D Guppy			
	CHD2	FADS2	NRXN1	CHD2	FADS2	NRXN1
	Tetranucleotide	Tetranucleotide	Tetranucleotide	Tetranucleotide	Tetranucleotide	Tetranucleotide
1	GAAG 9.897141	GAGT 4.657807	TGTT 169.968637	AAAC 4.549790	CAAT 7.302083	TGTT 212.061845
2	GAAA 5.466992	CGGC 4.269657	GTTT 15.903498	AGCA 3.696704	AGAT 4.868056	CAAT 8.630424
3	TGAC 3.958856	ACTG 3.105205	GGTC 2.981906	ATTA 3.601917	CATT 4.259549	TCAC 4.520698
4	CGGA 3.110530	GGTT 3.105205	AGGC 2.609168	GCGA 3.127980	GCTC 3.194661	GTTT 3.287781
5	GCGA 2.544979	TTTC 2.328904	TCAG 2.236429	TCAC 3.127980	TCTC 3.143953	TCTC 3.082294
6	GCAG 2.504583	GATC 2.328904	GACT 1.739445	AGCT 3.127980	AGAA 3.042535	GCTC 2.951530
7	GAGA 2.403591	GGCG 2.328904	GGGT 1.689747	GCTT 2.843619	AAGT 2.738281	CCTC 2.301446
8	AACC 2.262204	AGGG 1.863123	CTGA 1.656614	GAAG 2.488166	CTCG 2.434028	AAGC 2.260349
9	GAGG 2.262204	GCTC 1.746678	ACGT 1.490953	ATCT 2.356141	GAGA 2.231192	AGAG 2.054863
10	GACT 2.262204	GCAG 1.707863	TCTT 1.490953	AGAG 2.345985	GTGA 2.129774	AGCC 1.64389

Top ten overrepresented and tetranucleotide sequences, upstream of miscalled deletions (see section 3.3.5.3)

1D Albacore			1D Guppy		
CHD2	FADS2	NRXN1	CHD2	FADS2	NRXN1
Dinucleotide	Dinucleotide	Dinucleotide	Dinucleotide	Dinucleotide	Dinucleotide
AA 4.139283	GT 2.013419	TC 12.104995	AG 3.017075	TA 2.579653	TT 8.083172
AG 2.103398	TA 1.700277	TT 2.217112	GA 2.113610	TC 2.111571	TC 7.388746
TT 0.982692	CA 1.410694	AT 1.202773	AA 1.959809	CA 1.258810	AT 1.192546
AT 0.949299	TG 1.323209	AA 0.453426	AT 0.747385	GT 1.213716	GT 0.916645
GA 0.855924	TC 1.145085	CT 0.316603	GT 0.695327	CG 1.189969	CT 0.596402
TA 0.662041	CC 0.996772	CC 0.302212	CC 0.683006	AT 0.995247	AC 0.432690
TG 0.646593	GC 0.845801	CA 0.298188	CA 0.647107	CT 0.908704	CC 0.351477
CA 0.509647	CT 0.805955	GG 0.266686	GG 0.644674	AC 0.878954	GC 0.314511
CC 0.458061	GG 0.795566	GT 0.247794	GC 0.616022	GG 0.855598	CA 0.310389
GC 0.448717	AC 0.755582	TG 0.212354	CT 0.563103	TG 0.829372	GG 0.287343

Top ten over-represented dinucleotide sequences upstream of miscalled insertions (see section 3.3.5.3)

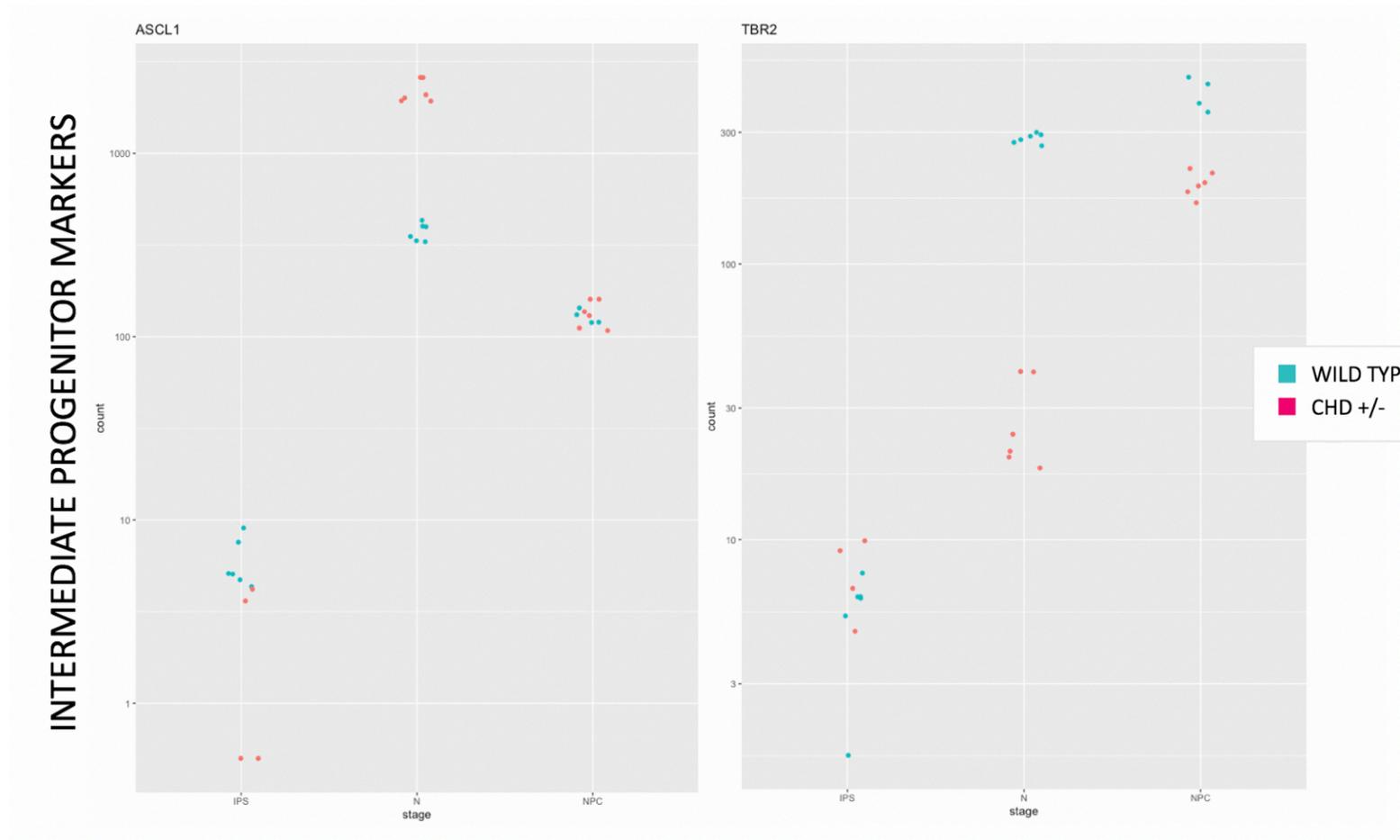
1D Albacore			1D Guppy		
CHD2	FADS2	NRXN1	CHD2	FADS2	NRXN1
Trinucleotide	Trinucleotide	Trinucleotide	Trinucleotide	Trinucleotide	Trinucleotide
AAG 14.751827	GTA 24.810167	TCT 50.917407	AAG 6.537754	TAG 5.679398	TCT 30.058684
AGA 3.488955	TAT 3.270431	TTC 2.468851	AGA 5.221860	TCC 3.816315	TTC 10.137067
AGG 3.343306	CAA 2.571236	ATG 2.285269	GAA 4.482020	TAT 3.786265	GTC 4.192969
TTT 1.742213	TGT 2.165251	TTT 1.663837	AGG 3.631664	CGG 2.343879	ATG 2.221729
ATT 1.481316	TCT 2.016655	CAT 0.841941	AGC 2.640094	CAT 2.163580	TTT 1.829659
AGC 1.298861	CAT 1.908474	TCG 0.835259	ATT 1.563335	GTG 1.651153	CTT 1.256634
GAA 1.197979	CCG 1.804376	CTT 0.686198	GTG 1.259011	ATA 1.622685	CAT 1.110864
ATA 1.041611	TAG 1.804376	AAG 0.584682	TCC 1.172135	TCT 1.527233	ACA 1.034629
TAA 0.965801	GGG 1.691602	GTC 0.517861	CCA 1.079829	CTC 1.458065	TCG 0.898493
ATG 0.953339	GTC 1.435299	TCC 0.494474	CAC 1.031479	AAA 1.442387	TCC 0.773249

Top ten over-represented trinucleotide sequences upstream of miscalled insertions (see section 3.3.5.3)

1D Albacore			1D Guppy		
CHD2	FADS2	NRXN1	CHD2	FADS2	NRXN1
Tetranucleotide	Tetranucleotide	Tetranucleotide	Tetranucleotide	Tetranucleotide	Tetranucleotide
AGAA 39.722449	AGGT 16.690476	TTTC 185.060049	AAAG 19.391797	TGTA 10.817901	TTTC 101.720328
AAAG 9.824686	ATTC 6.202542	TTTT 8.469529	AGAA 17.621092	TCTC 9.285365	TTTT 75.800159
TGTT 7.100388	ATCA 5.112398	TGTT 7.216640	GAGA 11.539666	ATCA 6.851337	GTCT 10.618557
GCAG 5.390904	TAGG 4.962033	TCTT 7.041236	TGCA 9.283307	GACG 5.949846	TCTT 9.720063
TAAT 4.943238	TCTG 4.736486	GCAT 5.512711	TAAT 7.478219	GCTA 5.048354	TGTT 6.044409
GATG 4.369469	GCTA 4.510940	GTCT 4.510400	GAAG 6.221105	ATTC 4.462384	GGTC 5.390960
ACTA 4.192925	TGCC 3.909481	GACC 1.353120	GCAG 6.188871	TTCT 3.966564	GCAT 5.309278
TCAT 3.277102	TGTA 3.157658	TGCA 1.202773	GGTC 3.223370	GAAC 3.786265	GGGT 4.998890
TGCA 3.045388	GGGC 3.007293	GCTC 1.075207	CCGT 3.094436	GTGT 3.245370	GTTT 2.940523
AAGA 2.939461	GACA 2.988497	GTTT 1.002311	AGGT 2.664653	GAAA 3.245370	GACC 2.001190

Top ten over-represented tetranucleotide sequences upstream of miscalled insertions (see section 3.3.5.3)

CHAPTER 4



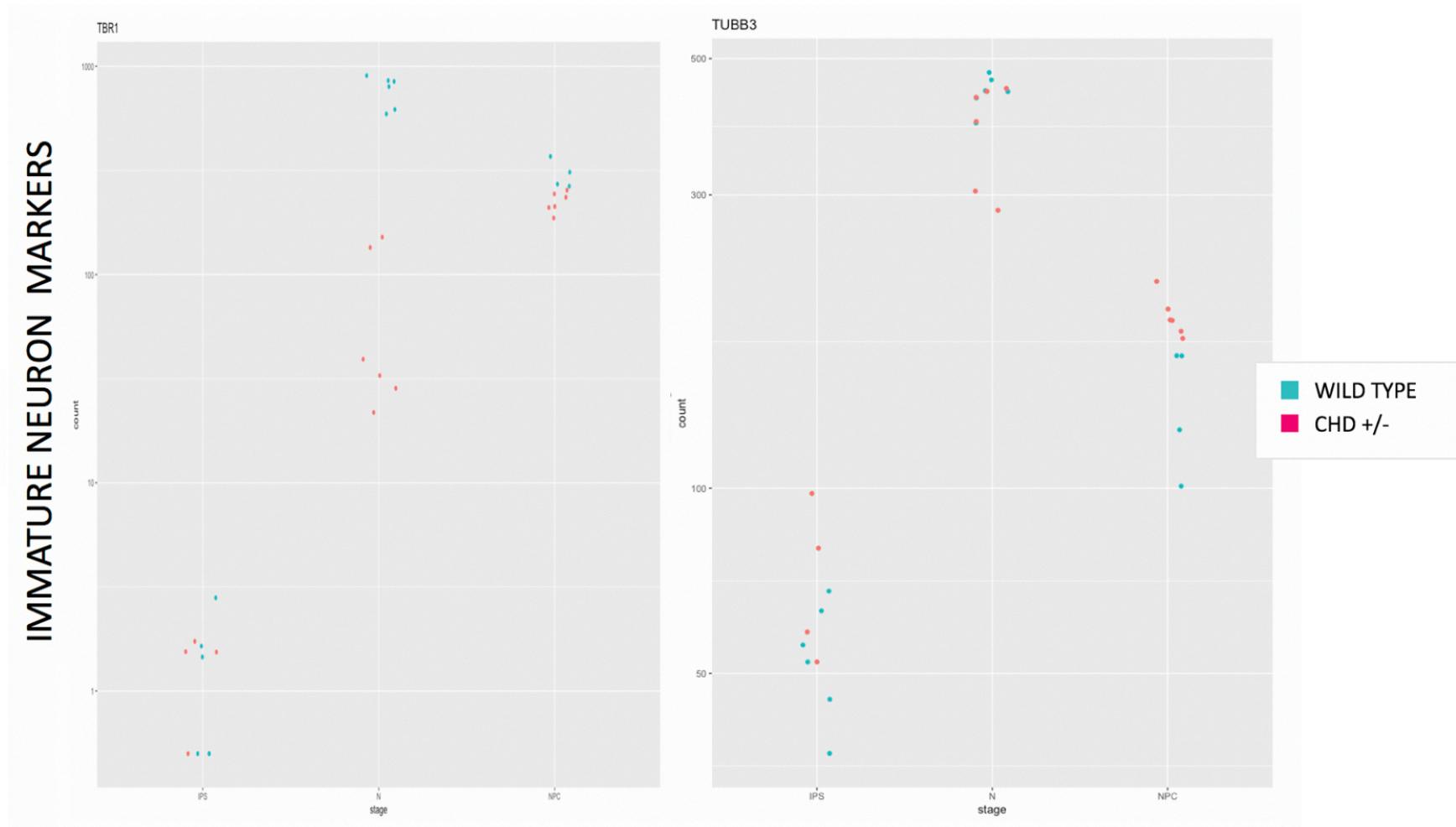
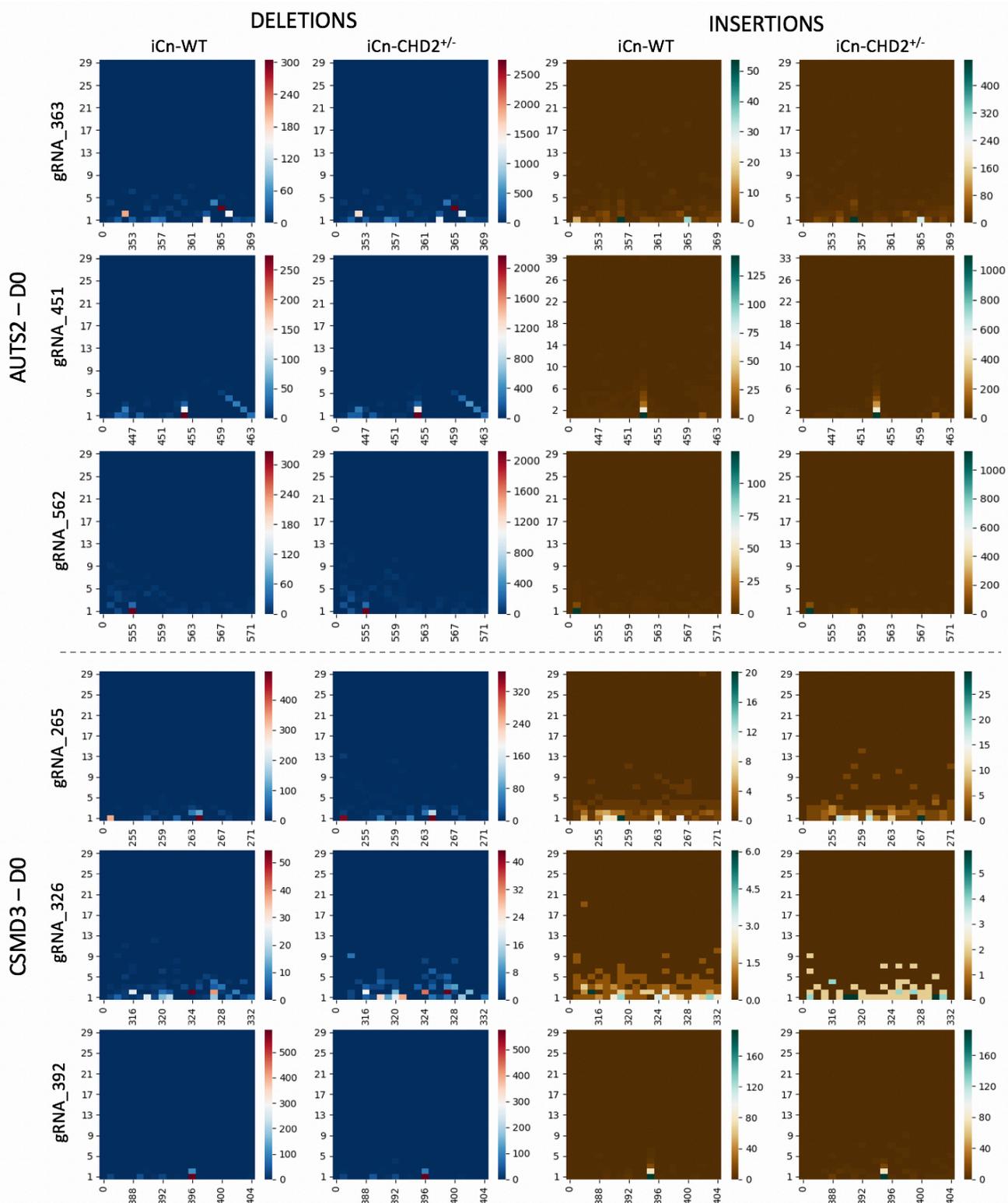
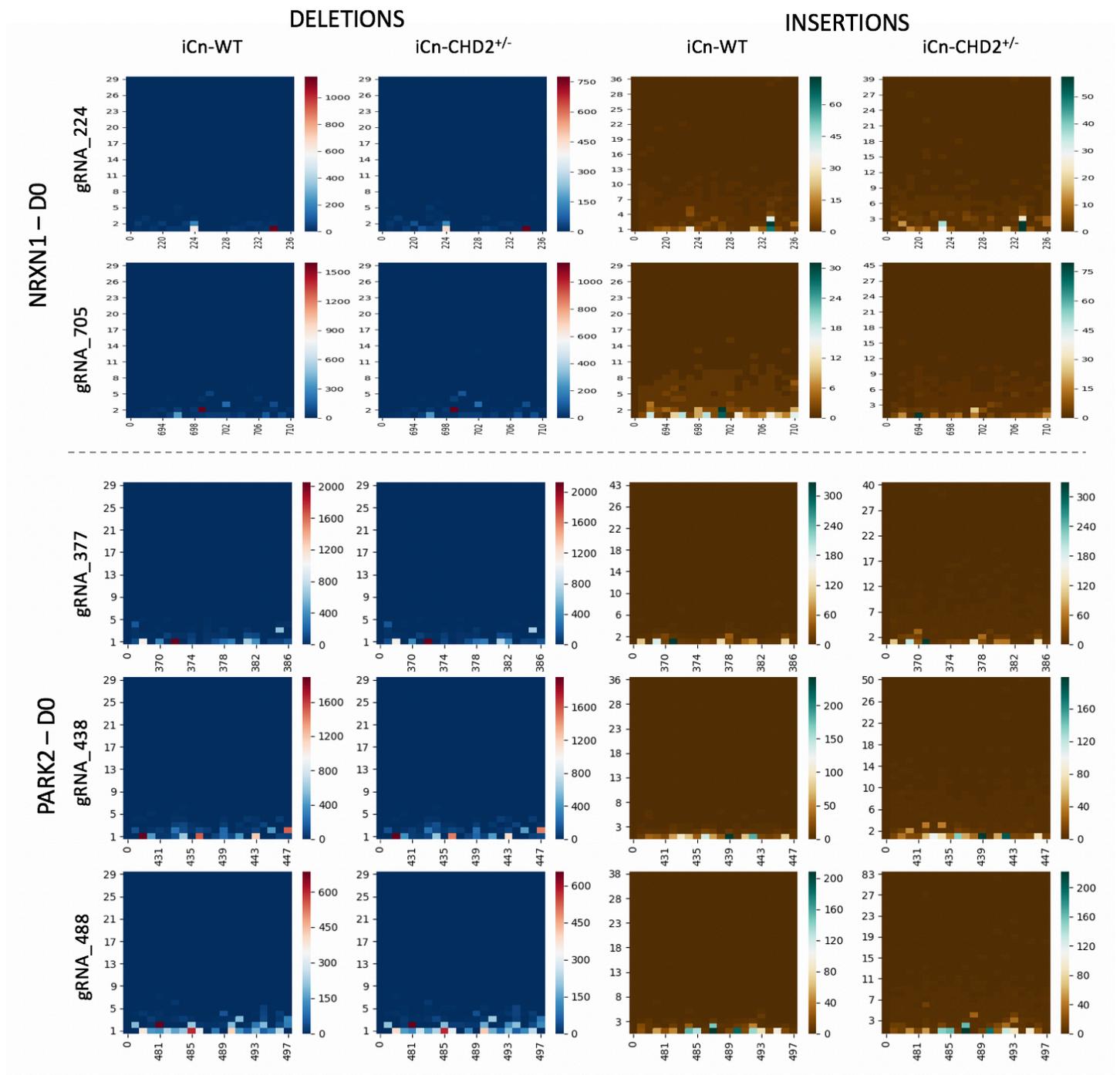


Figure 4.8.4: read count comparison for immature neuron markers at IPS, Neuron (N) and NPC stages of differentiation

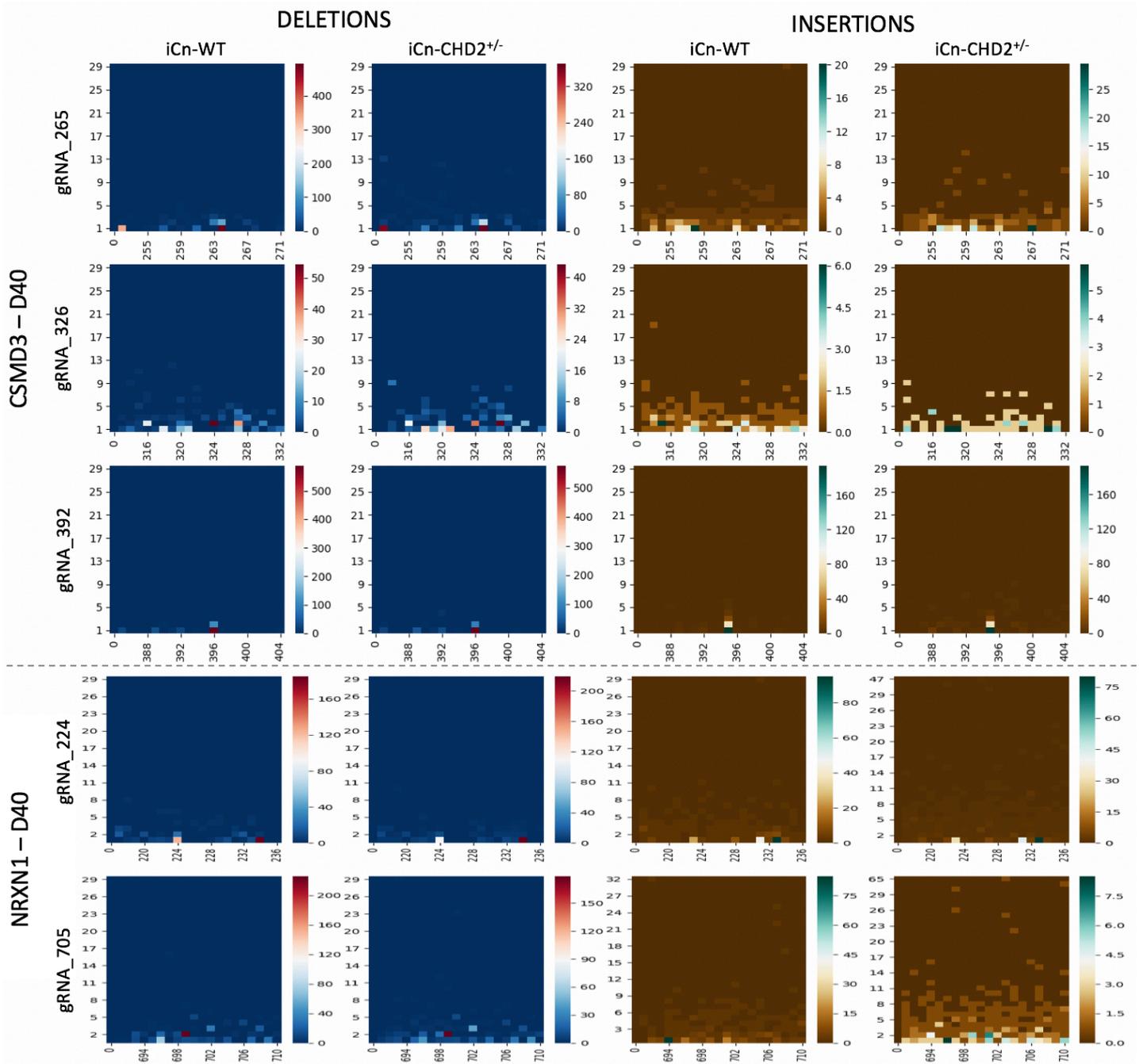
CHAPTER 5



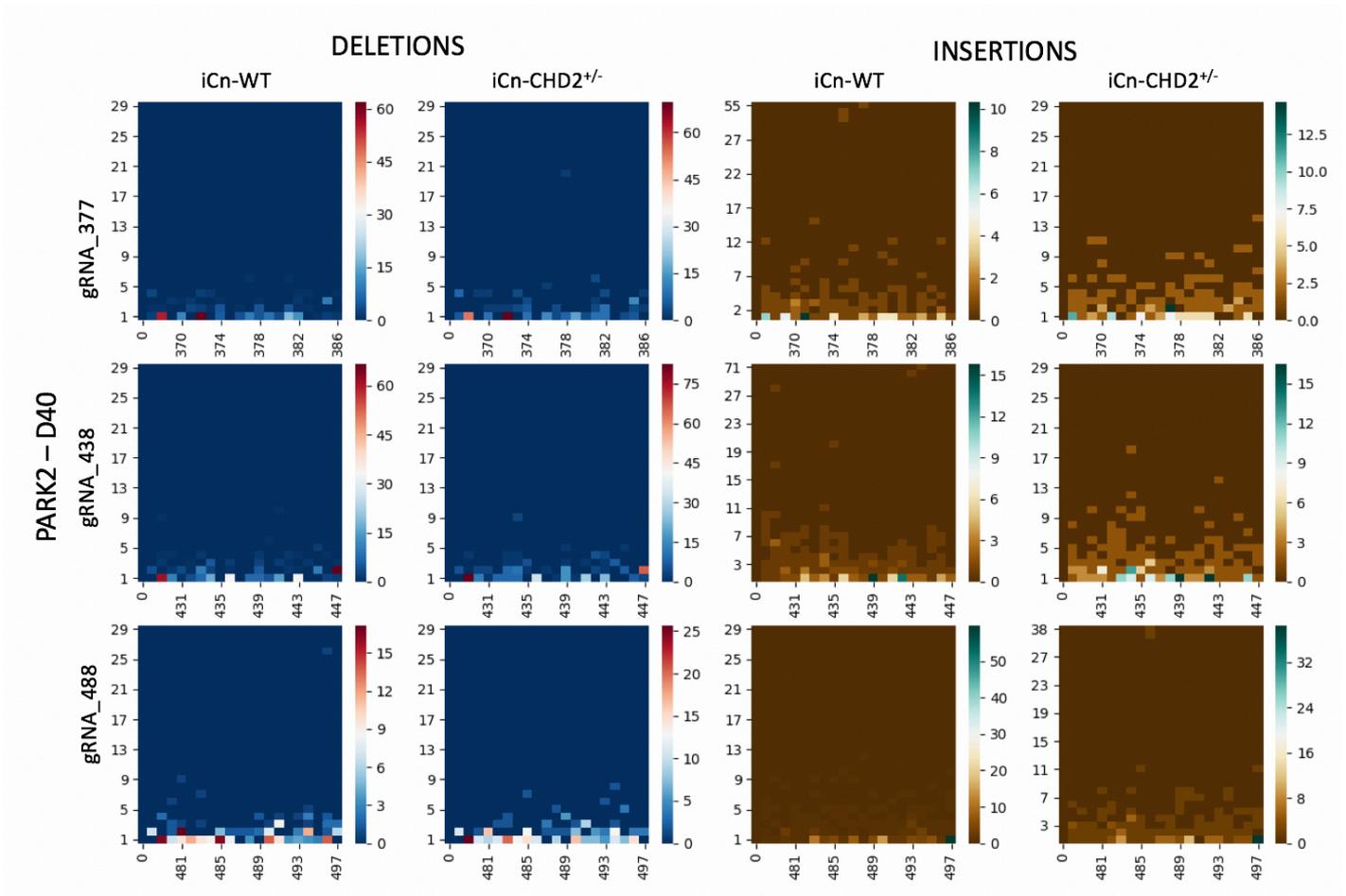
Heatmaps demonstrating depth of indels of different lengths occurring within 10bp upstream or downstream of each gRNA cut site at DO. Y axis = length of indel, x axis = position on reference sequence, with cutsite at centre, colourmap = number of reads containing deletion of length x at position y



Heatmaps demonstrating depth of indels of different lengths occurring within 10bp upstream or downstream of each gRNA cut site at DO. Y axis = length of indel, x axis = position on reference sequence, with cutsite at centre, colourmap = number of reads containing deletion of length x at position y

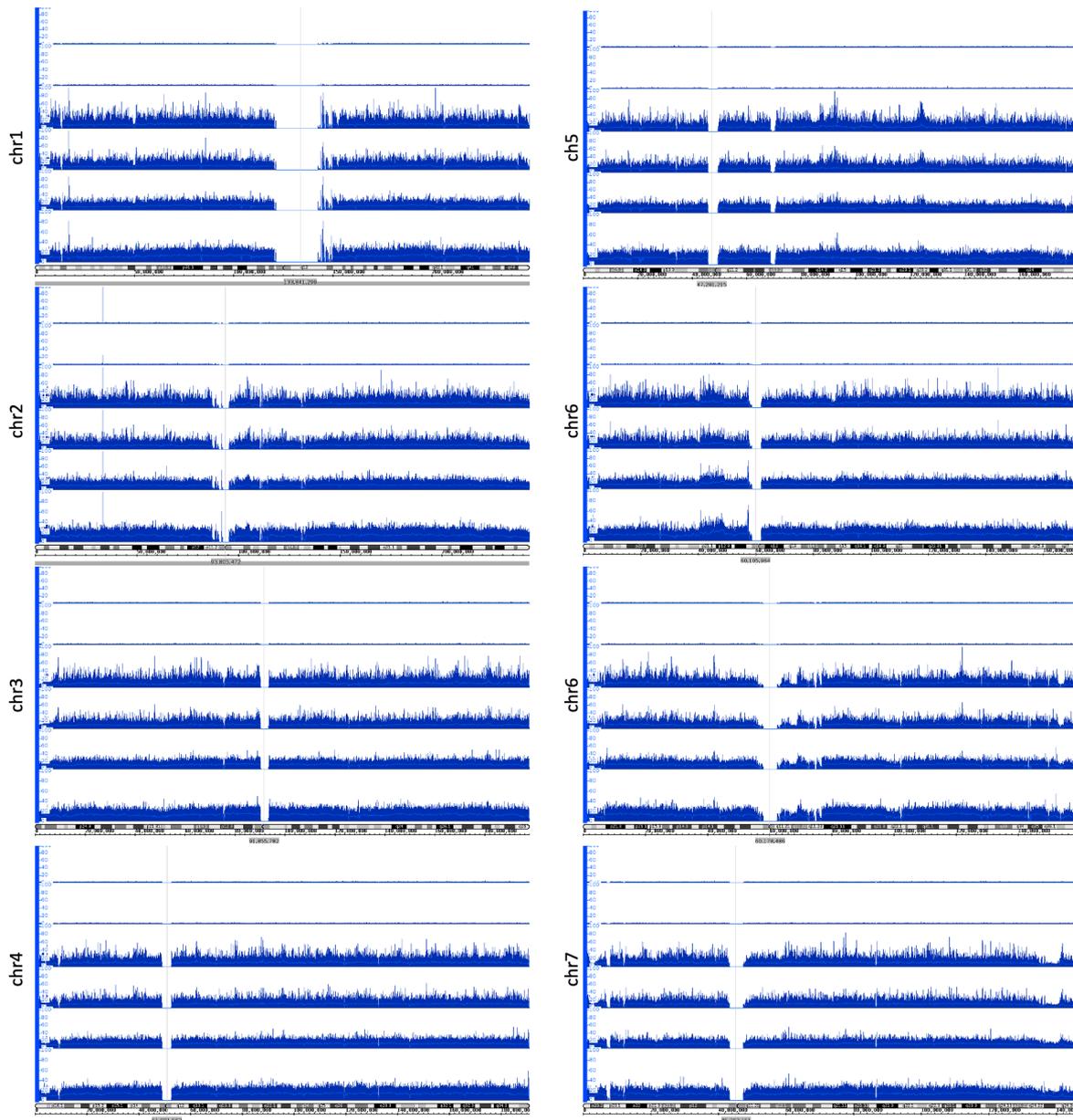


Heatmaps demonstrating depth of indels of different lengths occurring within 10bp upstream or downstream of each gRNA cut site at D40. Y axis = length of indel, x axis = position on reference sequence, with cutsite at centre, colourmap = number of reads containing deletion of length x at position y

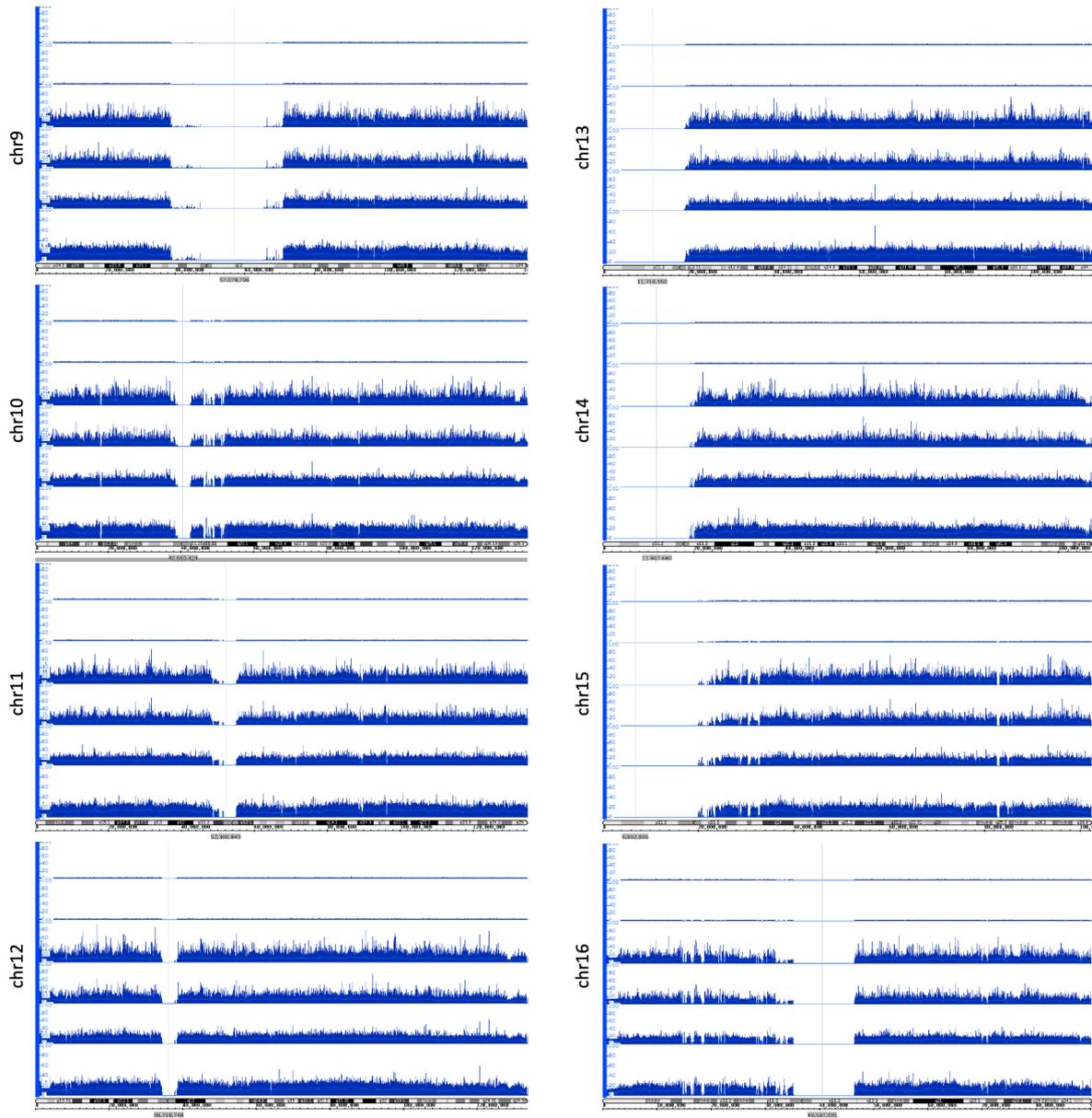


Heatmaps demonstrating depth of indels of different lengths occurring within 10bp upstream or downstream of each gRNA cut site at D40. Y axis = length of indel, x axis = position on reference sequence, with cutsite at centre, colourmap = number of reads containing deletion of length x at position y

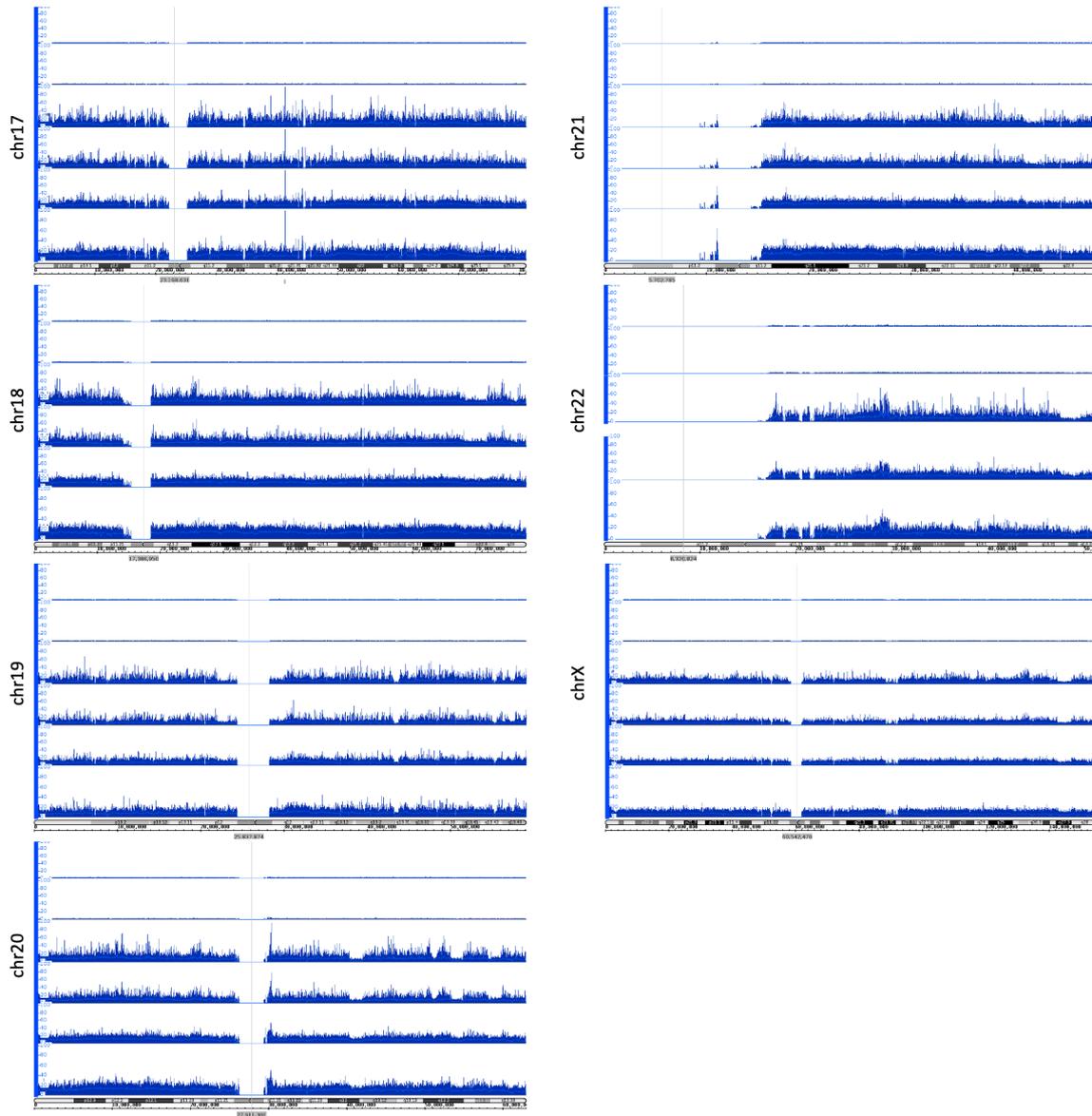
CHAPTER 6



Bedgraphs demonstrating break points determined by INDUCE-seq read start co-ordinates in chromosome 4. Row 1 - iCn-WT_D0, Row 2 - CHD2_D0, Row 3 - iCn-WT_D19, Row 4 - CHD2_D19, Row 5 - iCn-WT_D40, Row 6 - CHD2_D40. Breakpoints are sorted into bins of 1000bp width to demonstrate areas of higher and lower susceptibility.



Bedgraphs demonstrating break points determined by INDUCE-seq read start co-ordinates in chromosome 4. Row 1 - iCn-WT_D0, Row 2 - CHD2_D0, Row 3 - iCn-WT_D19, Row 4 - CHD2_D19, Row 5 - iCn-WT_D40, Row 6 - CHD2_D40. Breakpoints are sorted into bins of 1000bp width to demonstrate areas of higher and lower susceptibility.



Bedgraphs demonstrating break points determined by INDUCE-seq read start co-ordinates in chromosome 4. Row 1- iCn-WT_D0, Row 2 -CHD2_D0, Row 3 -iCn-WT_D19, Row 4 - CHD2_D19, Row 5 - iCn-WT_D40, Row 6 - CHD2_D40. Breakpoints are sorted into bins of 1000bp width to demonstrate areas of higher and lower susceptibility.

REFERENCES

1. Lamar, K.J. and G.L. Carvill, *Chromatin Remodeling Proteins in Epilepsy: Lessons From CHD2-Associated Epilepsy*. Front Mol Neurosci, 2018. **11**: p. 208.
2. OMIM. *Epileptic Encephalopathy, Childhood-onset; EEOC*. OMIM 2019 [cited 2019 20/8/2019]; Available from: <https://www.omim.org/entry/615369?search=epileptic%20encephalopathy%2C%20childhood&highlight=%28epilepsy%7Cepileptic%29>.
3. Siggens, L., et al., *Transcription-coupled recruitment of human CHD1 and CHD2 influences chromatin accessibility and histone H3 and H3.3 occupancy at active chromatin regions*. Epigenetics Chromatin, 2015. **8**(1): p. 4.
4. Meganathan, K., et al., *Regulatory networks specifying cortical interneurons from human embryonic stem cells reveal roles for CHD2 in interneuron development*. Proc Natl Acad Sci U S A, 2017. **114**(52): p. E11180-E11189.
5. Harada, A., et al., *Chd2 interacts with H3.3 to determine myogenic cell fate*. EMBO J, 2012. **31**(13): p. 2994-3007.
6. Luijsterburg, M.S., et al., *PARP1 Links CHD2-Mediated Chromatin Expansion and H3.3 Deposition to DNA Repair by Non-homologous End-Joining*. Mol Cell, 2016. **61**(4): p. 547-562.
7. Sayers, E.W., et al., *Database resources of the National Center for Biotechnology Information*. Nucleic Acids Res, 2019. **47**(D1): p. D23-D28.
8. Zerbino, D.R., et al., *Ensembl 2018*. Nucleic Acids Res, 2018. **46**(D1): p. D754-D761.
9. Uhlen, M., et al., *Proteomics. Tissue-based map of the human proteome*. Science, 2015. **347**(6220): p. 1260419.
10. Thul, P.J., et al., *A subcellular map of the human proteome*. Science, 2017. **356**(6340).
11. Kasah, S., C. Oddy, and M.A. Basson, *Autism-linked CHD gene expression patterns during development predict multi-organ disease phenotypes*. J Anat, 2018. **233**(6): p. 755-769.
12. Shen, E.H., C.C. Overly, and A.R. Jones, *The Allen Human Brain Atlas: comprehensive gene expression mapping of the human brain*. Trends Neurosci, 2012. **35**(12): p. 711-4.
13. UniProt, C., *UniProt: a worldwide hub of protein knowledge*. Nucleic Acids Res, 2019. **47**(D1): p. D506-D515.
14. Liu, J.C., C.G. Ferreira, and T. Yusufzai, *Human CHD2 is a chromatin assembly ATPase regulated by its chromo- and DNA-binding domains*. J Biol Chem, 2015. **290**(1): p. 25-34.
15. Ryan, D.P., et al., *The DNA-binding domain of the Chd1 chromatin-remodelling enzyme contains SANT and SLIDE domains*. EMBO J, 2011. **30**(13): p. 2596-609.
16. Konrad J Karczewski, L.C.F., Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, Laura D Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A Watts, Daniel Rhodes, Moriel Singer-Berk, Eleanor G Seaby, Jack A Kosmicki, Raymond K Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X Chong, Kaitlin E Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S Ware, Christopher Vittal, Irina M Armean, Louis Bergelson, Kristian Cibulskis, Kristen M Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam

- Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E Talkowski, The Genome Aggregation Database Consortium, Benjamin M Neale, Mark J Daly, Daniel G MacArthur, *Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes*. bioRxiv 2019. **531210**(doi: <https://doi.org/10.1101/531210>).
17. Woodage, T., et al., *Characterization of the CHD family of proteins*. Proc Natl Acad Sci U S A, 1997. **94**(21): p. 11472-7.
 18. Zhou, J., et al., *Human CHD1 is required for early DNA-damage signaling and is uniquely regulated by its N terminus*. Nucleic Acids Res, 2018. **46**(8): p. 3891-3905.
 19. Pilarowski, G.O., et al., *Missense variants in the chromatin remodeler CHD1 are associated with neurodevelopmental disability*. J Med Genet, 2018. **55**(8): p. 561-566.
 20. Rodriguez-Castaneda, F., et al., *The SUMO protease SENP1 and the chromatin remodeler CHD3 interact and jointly affect chromatin accessibility and gene expression*. J Biol Chem, 2018. **293**(40): p. 15439-15454.
 21. Snijders Blok, L., et al., *CHD3 helicase domain mutations cause a neurodevelopmental syndrome with macrocephaly and impaired speech and language*. Nat Commun, 2018. **9**(1): p. 4619.
 22. Polo, S.E., et al., *Regulation of DNA-damage responses and cell-cycle progression by the chromatin remodelling factor CHD4*. EMBO J, 2010. **29**(18): p. 3130-9.
 23. Sifrim, A., et al., *Distinct genetic architectures for syndromic and nonsyndromic congenital heart defects identified by exome sequencing*. Nat Genet, 2016. **48**(9): p. 1060-5.
 24. Baykara, O., et al., *CHD5 is a potential tumor suppressor in non small cell lung cancer (NSCLC)*. Gene, 2017. **618**: p. 65-68.
 25. Egan, C.M., et al., *CHD5 is required for neurogenesis and has a dual role in facilitating gene expression and polycomb gene repression*. Dev Cell, 2013. **26**(3): p. 223-36.
 26. Hall, W.A., et al., *Low CHD5 expression activates the DNA damage response and predicts poor outcome in patients undergoing adjuvant therapy for resected pancreatic cancer*. Oncogene, 2014. **33**(47): p. 5450-6.
 27. Moore, S., et al., *The CHD6 chromatin remodeler is an oxidative DNA damage response factor*. Nat Commun, 2019. **10**(1): p. 241.
 28. Van Nostrand, J.L., et al., *Inappropriate p53 activation during development induces features of CHARGE syndrome*. Nature, 2014. **514**(7521): p. 228-32.
 29. Engelen, E., et al., *Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes*. Nat Genet, 2011. **43**(6): p. 607-11.
 30. Jongmans, M.C., et al., *CHARGE syndrome: the phenotypic spectrum of mutations in the CHD7 gene*. J Med Genet, 2006. **43**(4): p. 306-14.
 31. Aramaki, M., et al., *Phenotypic spectrum of CHARGE syndrome with CHD7 mutations*. J Pediatr, 2006. **148**(3): p. 410-4.
 32. Batsukh, T., et al., *CHD8 interacts with CHD7, a protein which is mutated in CHARGE syndrome*. Hum Mol Genet, 2010. **19**(14): p. 2858-66.
 33. Ishihara, K., M. Oshimura, and M. Nakao, *CTCF-dependent chromatin insulator is linked to epigenetic remodeling*. Mol Cell, 2006. **23**(5): p. 733-42.
 34. Bernier, R., et al., *Disruptive CHD8 mutations define a subtype of autism early in development*. Cell, 2014. **158**(2): p. 263-276.
 35. Barnard, R.A., M.B. Pomaville, and B.J. O'Roak, *Mutations and Modeling of the Chromatin Remodeler CHD8 Define an Emerging Autism Etiology*. Front Neurosci, 2015. **9**: p. 477.

36. Salomon-Kent, R., et al., *New Face for Chromatin-Related Mesenchymal Modulator: n-CHD9 Localizes to Nucleoli and Interacts With Ribosomal Genes*. J Cell Physiol, 2015. **230**(9): p. 2270-80.
37. Snider, A.C., et al., *The chromatin remodeling factor Chd11 is required in the preimplantation embryo*. Biol Open, 2013. **2**(2): p. 121-31.
38. Mefford, H.C., et al., *Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes*. N Engl J Med, 2008. **359**(16): p. 1685-99.
39. Tellez-Zenteno, J.F., et al., *Psychiatric comorbidity in epilepsy: a population-based analysis*. Epilepsia, 2007. **48**(12): p. 2336-44.
40. Owen, M.J., *New approaches to psychiatric diagnostic classification*. Neuron, 2014. **84**(3): p. 564-71.
41. Scheffer, I.E., et al., *Classification of the epilepsies: New concepts for discussion and debate-Special report of the ILAE Classification Task Force of the Commission for Classification and Terminology*. Epilepsia Open, 2016. **1**(1-2): p. 37-44.
42. Carvill, G.L., et al., *Targeted resequencing in epileptic encephalopathies identifies de novo mutations in CHD2 and SYNGAP1*. Nat Genet, 2013. **45**(7): p. 825-30.
43. Engel, J., Jr. and E. International League Against, *A proposed diagnostic scheme for people with epileptic seizures and with epilepsy: report of the ILAE Task Force on Classification and Terminology*. Epilepsia, 2001. **42**(6): p. 796-803.
44. Blume, W.T., et al., *Glossary of descriptive terminology for ictal semiology: report of the ILAE task force on classification and terminology*. Epilepsia, 2001. **42**(9): p. 1212-8.
45. Tully, I., et al., *Admissions to paediatric intensive care units (PICU) with refractory convulsive status epilepticus (RCSE): A two-year multi-centre study*. Seizure, 2015. **29**: p. 153-61.
46. Holland, K.D. and B.E. Hallinan, *What causes epileptic encephalopathy in infancy?: the answer may lie in our genes*. Neurology, 2010. **75**(13): p. 1132-3.
47. Street, G.O. *Clinical Genetics Department Test Directory*. 2019; Available from: <http://www.labs.gosh.nhs.uk/laboratory-services/genetics/molecular-genetics-service>.
48. Project, G. *Panel App*. 2019; Available from: <https://panelapp.genomicsengland.co.uk>.
49. Veredice, C., et al., *Early onset myoclonic epilepsy and 15q26 microdeletion: observation of the first case*. Epilepsia, 2009. **50**(7): p. 1810-5.
50. Li, M.M., et al., *Characterization of a cryptic 3.3 Mb deletion in a patient with a "balanced t(15;22) translocation" using high density oligo array CGH and gene expression arrays*. Am J Med Genet A, 2008. **146A**(3): p. 368-75.
51. Dhamija, R., et al., *Microdeletion of chromosome 15q26.1 in a child with intractable generalized epilepsy*. Pediatr Neurol, 2011. **45**(1): p. 60-2.
52. Capelli, L.P., et al., *Deletion of the RMGA and CHD2 genes in a child with epilepsy and mental deficiency*. Eur J Med Genet, 2012. **55**(2): p. 132-4.
53. Lund, C., et al., *CHD2 mutations in Lennox-Gastaut syndrome*. Epilepsy Behav, 2014. **33**: p. 18-21.
54. Courage, C., et al., *15q26.1 microdeletion encompassing only CHD2 and RGMA in two adults with moderate intellectual disability, epilepsy and truncal obesity*. Eur J Med Genet, 2014. **57**(9): p. 520-3.
55. Thomas, R.H., et al., *CHD2 myoclonic encephalopathy is frequently associated with self-induced seizures*. Neurology, 2015. **84**(9): p. 951-8.
56. Verhoeven, W.M., et al., *Absence epilepsy and the CHD2 gene: an adolescent male with moderate intellectual disability, short-lasting psychoses, and an interstitial deletion in 15q26.1-q26.2*. Neuropsychiatr Dis Treat, 2016. **12**: p. 1135-9.

57. Bernardo, P., et al., *CHD2 mutations: Only epilepsy? Description of cognitive and behavioral profile in a case with a new mutation*. *Seizure*, 2017. **51**: p. 186-189.
58. Zhou, P., et al., *Novel mutations and phenotypes of epilepsy-associated genes in epileptic encephalopathies*. *Genes Brain Behav*, 2018. **17**(8): p. e12456.
59. Rim, J.H., et al., *Efficient strategy for the molecular diagnosis of intractable early-onset epilepsy using targeted gene sequencing*. *BMC Med Genomics*, 2018. **11**(1): p. 6.
60. Chen, X., et al., *Phf8 histone demethylase deficiency causes cognitive impairments through the mTOR pathway*. *Nat Commun*, 2018. **9**(1): p. 114.
61. Wang, Y., et al., *Genetic Variants Identified from Epilepsy of Unknown Etiology in Chinese Children by Targeted Exome Sequencing*. *Sci Rep*, 2017. **7**: p. 40319.
62. Helbig, K.L., et al., *Diagnostic exome sequencing provides a molecular diagnosis for a significant proportion of patients with epilepsy*. *Genet Med*, 2016. **18**(9): p. 898-905.
63. Hamdan, F.F., et al., *De novo mutations in moderate or severe intellectual disability*. *PLoS Genet*, 2014. **10**(10): p. e1004772.
64. O'Roak, B.J., et al., *Recurrent de novo mutations implicate novel genes underlying simplex autism risk*. *Nat Commun*, 2014. **5**: p. 5595.
65. Neale, B.M., et al., *Patterns and rates of exonic de novo mutations in autism spectrum disorders*. *Nature*, 2012. **485**(7397): p. 242-5.
66. Lebrun, N., et al., *Autism spectrum disorder recurrence, resulting of germline mosaicism for a CHD2 gene missense variant*. *Clin Genet*, 2017. **92**(6): p. 669-670.
67. Turner, T.N., et al., *Genomic Patterns of De Novo Mutation in Simplex Autism*. *Cell*, 2017. **171**(3): p. 710-722 e12.
68. Bragin, E., et al., *DECIPHER: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation*. *Nucleic Acids Res*, 2014. **42**(Database issue): p. D993-D1000.
69. Firth, H.V., et al., *DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources*. *Am J Hum Genet*, 2009. **84**(4): p. 524-33.
70. Deciphering Developmental Disorders, S., *Prevalence and architecture of de novo mutations in developmental disorders*. *Nature*, 2017. **542**(7642): p. 433-438.
71. Rodriguez, D., et al., *Mutations in CHD2 cause defective association with active chromatin in chronic lymphocytic leukemia*. *Blood*, 2015. **126**(2): p. 195-202.
72. Gauthier-Vasserot, A., et al., *Application of whole-exome sequencing to unravel the molecular basis of undiagnosed syndromic congenital neutropenia with intellectual disability*. *Am J Med Genet A*, 2017. **173**(1): p. 62-71.
73. Marfella, C.G., et al., *A mutation in the mouse Chd2 chromatin remodeling enzyme results in a complex renal phenotype*. *Kidney Blood Press Res*, 2008. **31**(6): p. 421-32.
74. Marfella, C.G., et al., *Mutation of the SNF2 family member Chd2 affects mouse development and survival*. *J Cell Physiol*, 2006. **209**(1): p. 162-71.
75. Nagarajan, P., et al., *Role of chromodomain helicase DNA-binding protein 2 in DNA damage response signaling and tumorigenesis*. *Oncogene*, 2009. **28**(8): p. 1053-62.
76. Rajagopalan, S., J. Nepa, and S. Venkatachalam, *Chromodomain helicase DNA-binding protein 2 affects the repair of X-ray and UV-induced DNA damage*. *Environ Mol Mutagen*, 2012. **53**(1): p. 44-50.
77. Suls, A., et al., *De novo loss-of-function mutations in CHD2 cause a fever-sensitive myoclonic epileptic encephalopathy sharing features with Dravet syndrome*. *Am J Hum Genet*, 2013. **93**(5): p. 967-75.

78. Wu, M.C., et al., [*Methylation of CHD5 Gene Promoter Regulates p19(Arf)/p53/p21(Cip1) Pathway to Facilitate Pathogenesis of Acute Myeloid Leukemia*]. *Zhongguo Shi Yan Xue Ye Xue Za Zhi*, 2019. **27**(4): p. 1001-1007.
79. Chu, X., et al., *Genotranscriptomic meta-analysis of the CHD family chromatin remodelers in human cancers - initial evidence of an oncogenic role for CHD7*. *Mol Oncol*, 2017. **11**(10): p. 1348-1360.
80. Villa, M., et al., *Coupling end resection with the checkpoint response at DNA double-strand breaks*. *Cell Mol Life Sci*, 2016. **73**(19): p. 3655-63.
81. Homem, C.C., M. Repic, and J.A. Knoblich, *Proliferation control in neural stem and progenitor cells*. *Nat Rev Neurosci*, 2015. **16**(11): p. 647-59.
82. Shen, T., et al., *CHD2 is Required for Embryonic Neurogenesis in the Developing Cerebral Cortex*. *Stem Cells*, 2015. **33**(6): p. 1794-806.
83. Hu, J.S., et al., *Cortical interneuron development: a tale of time and space*. *Development*, 2017. **144**(21): p. 3867-3878.
84. Kim, Y.Z., *Altered histone modifications in gliomas*. *Brain Tumor Res Treat*, 2014. **2**(1): p. 7-21.
85. Luger, K., et al., *Crystal structure of the nucleosome core particle at 2.8 Å resolution*. *Nature*, 1997. **389**(6648): p. 251-60.
86. Marino-Ramirez, L., et al., *Histone structure and nucleosome stability*. *Expert Rev Proteomics*, 2005. **2**(5): p. 719-29.
87. Dorigo, B., et al., *Nucleosome arrays reveal the two-start organization of the chromatin fiber*. *Science*, 2004. **306**(5701): p. 1571-3.
88. Peterson, C.L. and M.A. Laniel, *Histones and histone modifications*. *Curr Biol*, 2004. **14**(14): p. R546-51.
89. Bernstein, B.E., et al., *A bivalent chromatin structure marks key developmental genes in embryonic stem cells*. *Cell*, 2006. **125**(2): p. 315-26.
90. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types*. *Nature*, 2011. **473**(7345): p. 43-9.
91. van Eijk, P., et al., *Nucleosome remodeling at origins of global genome-nucleotide excision repair occurs at the boundaries of higher-order chromatin structure*. *Genome Res*, 2019. **29**(1): p. 74-84.
92. Xu, Y., et al., *WERAM: a database of writers, erasers and readers of histone acetylation and methylation in eukaryotes*. *Nucleic Acids Res*, 2017. **45**(D1): p. D264-D270.
93. Chen, P., et al., *H3.3 actively marks enhancers and primes gene transcription via opening higher-ordered chromatin*. *Genes Dev*, 2013. **27**(19): p. 2109-24.
94. McKittrick, E., et al., *Histone H3.3 is enriched in covalent modifications associated with active chromatin*. *Proc Natl Acad Sci U S A*, 2004. **101**(6): p. 1525-30.
95. Lin, C.J., M. Conti, and M. Ramalho-Santos, *Histone variant H3.3 maintains a decondensed chromatin state essential for mouse preimplantation development*. *Development*, 2013. **140**(17): p. 3624-34.
96. Bassing, C.H., et al., *Increased ionizing radiation sensitivity and genomic instability in the absence of histone H2AX*. *Proc Natl Acad Sci U S A*, 2002. **99**(12): p. 8173-8.
97. Fernandez-Capetillo, O., et al., *H2AX: the histone guardian of the genome*. *DNA Repair (Amst)*, 2004. **3**(8-9): p. 959-67.
98. Giaimo, B.D., et al., *The histone variant H2A.Z in gene regulation*. *Epigenetics Chromatin*, 2019. **12**(1): p. 37.
99. Rada-Iglesias, A., *Is H3K4me1 at enhancers correlative or causative?* *Nat Genet*, 2018. **50**(1): p. 4-5.

100. Wang, Y., X. Li, and H. Hu, *H3K4me2 reliably defines transcription factor binding regions in different cells*. Genomics, 2014. **103**(2-3): p. 222-8.
101. Huang, X., et al., *Stable H3K4me3 is associated with transcription initiation during early embryo development*. Bioinformatics, 2019.
102. Creyghton, M.P., et al., *Histone H3K27ac separates active from poised enhancers and predicts developmental state*. Proc Natl Acad Sci U S A, 2010. **107**(50): p. 21931-6.
103. Wiles, E.T. and E.U. Selker, *H3K27 methylation: a promiscuous repressive chromatin mark*. Curr Opin Genet Dev, 2017. **43**: p. 31-37.
104. Harwood, J.C., et al., *Nucleosome dynamics of human iPSC during neural differentiation*. EMBO Rep, 2019. **20**(6).
105. Mueller, B., et al., *Widespread changes in nucleosome accessibility without changes in nucleosome occupancy during a rapid transcriptional induction*. Genes Dev, 2017. **31**(5): p. 451-462.
106. Bilokapic, S., M. Strauss, and M. Halic, *Structural rearrangements of the histone octamer translocate DNA*. Nat Commun, 2018. **9**(1): p. 1330.
107. Belton, J.M., et al., *Hi-C: a comprehensive technique to capture the conformation of genomes*. Methods, 2012. **58**(3): p. 268-76.
108. Lee, S.H., et al., *Mapping the spectrum of 3D communities in human chromosome conformation capture data*. Sci Rep, 2019. **9**(1): p. 6859.
109. Zhou, T., et al., *Quantitative modeling of transcription factor binding specificities using DNA shape*. Proc Natl Acad Sci U S A, 2015. **112**(15): p. 4654-9.
110. Crick, F., *Central dogma of molecular biology*. Nature, 1970. **227**(5258): p. 561-3.
111. Marzluff, W.F., et al., *The human and mouse replication-dependent histone genes*. Genomics, 2002. **80**(5): p. 487-98.
112. Rudnizky, S., et al., *H2A.Z controls the stability and mobility of nucleosomes to regulate expression of the LH genes*. Nat Commun, 2016. **7**: p. 12958.
113. Shelby, R.D., K. Monier, and K.F. Sullivan, *Chromatin assembly at kinetochores is uncoupled from DNA replication*. J Cell Biol, 2000. **151**(5): p. 1113-8.
114. Wiedemann, S.M., et al., *Identification and characterization of two novel primate-specific histone H3 variants, H3.X and H3.Y*. J Cell Biol, 2010. **190**(5): p. 777-91.
115. Schwartz, B.E. and K. Ahmad, *Transcriptional activation triggers deposition and removal of the histone variant H3.3*. Genes Dev, 2005. **19**(7): p. 804-14.
116. Tamura, T., et al., *Inducible deposition of the histone variant H3.3 in interferon-stimulated genes*. J Biol Chem, 2009. **284**(18): p. 12217-25.
117. Ng, R.K. and J.B. Gurdon, *Epigenetic memory of an active gene state depends on histone H3.3 incorporation into chromatin in the absence of transcription*. Nat Cell Biol, 2008. **10**(1): p. 102-9.
118. Szenker, E., D. Ray-Gallet, and G. Almouzni, *The double face of the histone variant H3.3*. Cell Res, 2011. **21**(3): p. 421-34.
119. Jang, C.W., et al., *Histone H3.3 maintains genome integrity during mammalian development*. Genes Dev, 2015. **29**(13): p. 1377-92.
120. Maze, I., et al., *Critical Role of Histone Turnover in Neuronal Transcription and Plasticity*. Neuron, 2015. **87**(1): p. 77-94.
121. Lepack, A.E., et al., *Aberrant H3.3 dynamics in NAc promote vulnerability to depressive-like behavior*. Proc Natl Acad Sci U S A, 2016. **113**(44): p. 12562-12567.
122. Davis, C.A., et al., *The Encyclopedia of DNA elements (ENCODE): data portal update*. Nucleic Acids Res, 2018. **46**(D1): p. D794-D801.
123. Consortium, E.P., *The ENCODE (ENCyclopedia Of DNA Elements) Project*. Science, 2004. **306**(5696): p. 636-40.

124. Takahashi, K. and S. Yamanaka, *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors*. Cell, 2006. **126**(4): p. 663-76.
125. Liu, X., et al., *Yamanaka factors critically regulate the developmental signaling network in mouse embryonic stem cells*. Cell Res, 2008. **18**(12): p. 1177-89.
126. Semba, Y., et al., *Chd2 regulates chromatin for proper gene expression toward differentiation in mouse embryonic stem cells*. Nucleic Acids Res, 2017. **45**(15): p. 8758-8772.
127. Adam, S., S.E. Polo, and G. Almouzni, *Transcription recovery after DNA damage requires chromatin priming by the H3.3 histone chaperone HIRA*. Cell, 2013. **155**(1): p. 94-106.
128. Ciccia, A. and S.J. Elledge, *The DNA damage response: making it safe to play with knives*. Mol Cell, 2010. **40**(2): p. 179-204.
129. Lee, J.A., C.M. Carvalho, and J.R. Lupski, *A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders*. Cell, 2007. **131**(7): p. 1235-47.
130. Wilson, T.E., et al., *Large transcription units unify copy number variants and common fragile sites arising under replication stress*. Genome Res, 2015. **25**(2): p. 189-200.
131. McHugh, D. and J. Gil, *Senescence and aging: Causes, consequences, and therapeutic avenues*. J Cell Biol, 2018. **217**(1): p. 65-77.
132. Nykamp, K., et al., *Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria*. Genet Med, 2017. **19**(10): p. 1105-1117.
133. Hoeijmakers, J.H., *DNA damage, aging, and cancer*. N Engl J Med, 2009. **361**(15): p. 1475-85.
134. Wilson, M.D. and D. Durocher, *Reading chromatin signatures after DNA double-strand breaks*. Philos Trans R Soc Lond B Biol Sci, 2017. **372**(1731).
135. Lieber, M.R., et al., *Mechanism and regulation of human non-homologous DNA end-joining*. Nat Rev Mol Cell Biol, 2003. **4**(9): p. 712-20.
136. Lieber, M.R., *The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway*. Annu Rev Biochem, 2010. **79**: p. 181-211.
137. Seluanov, A., Z. Mao, and V. Gorbunova, *Analysis of DNA double-strand break (DSB) repair in mammalian cells*. J Vis Exp, 2010(43).
138. Debacker, K. and R.F. Kooy, *Fragile sites and human disease*. Hum Mol Genet, 2007. **16 Spec No. 2**: p. R150-8.
139. Fungtammasan, A., et al., *A genome-wide analysis of common fragile sites: what features determine chromosomal instability in the human genome?* Genome Res, 2012. **22**(6): p. 993-1005.
140. Ma, K., et al., *Common fragile sites: genomic hotspots of DNA damage and carcinogenesis*. Int J Mol Sci, 2012. **13**(9): p. 11974-99.
141. Mefford, H.C. and E.E. Eichler, *Duplication hotspots, rare genomic disorders, and common disease*. Curr Opin Genet Dev, 2009. **19**(3): p. 196-204.
142. Gan, W., et al., *R-loop-mediated genomic instability is caused by impairment of replication fork progression*. Genes Dev, 2011. **25**(19): p. 2041-56.
143. Sekiguchi, J.M., et al., *Nonhomologous end-joining proteins are required for V(D)J recombination, normal growth, and neurogenesis*. Cold Spring Harb Symp Quant Biol, 1999. **64**: p. 169-81.

144. McConnell, M.J., et al., *Intersection of diverse neuronal genomes and neuropsychiatric disease: The Brain Somatic Mosaicism Network*. Science, 2017. **356**(6336).
145. Poduri, A., et al., *Somatic activation of AKT3 causes hemispheric developmental brain malformations*. Neuron, 2012. **74**(1): p. 41-8.
146. Lee, J.H., *Somatic mutations in disorders with disrupted brain connectivity*. Exp Mol Med, 2016. **48**: p. e239.
147. Lodato, M.A., et al., *Aging and neurodegeneration are associated with increased mutations in single human neurons*. Science, 2018. **359**(6375): p. 555-559.
148. McConnell, M.J., et al., *Mosaic copy number variation in human neurons*. Science, 2013. **342**(6158): p. 632-7.
149. Freed, D., E.L. Stevens, and J. Pevsner, *Somatic mosaicism in the human genome*. Genes (Basel), 2014. **5**(4): p. 1064-94.
150. Baillie, J.K., et al., *Somatic retrotransposition alters the genetic landscape of the human brain*. Nature, 2011. **479**(7374): p. 534-7.
151. van Overbeek, M., et al., *DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks*. Mol Cell, 2016. **63**(4): p. 633-646.
152. Wei, P.C., et al., *Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells*. Cell, 2016. **164**(4): p. 644-55.
153. Weissman, I.L. and F.H. Gage, *A Mechanism for Somatic Brain Mosaicism*. Cell, 2016. **164**(4): p. 593-5.
154. Schwer, B., et al., *Transcription-associated processes cause DNA double-strand breaks and translocations in neural stem/progenitor cells*. Proc Natl Acad Sci U S A, 2016. **113**(8): p. 2258-63.
155. Muotri, A.R. and F.H. Gage, *Generation of neuronal variability and complexity*. Nature, 2006. **441**(7097): p. 1087-93.
156. Suberbielle, E., et al., *Physiologic brain activity causes DNA double-strand breaks in neurons, with exacerbation by amyloid-beta*. Nat Neurosci, 2013. **16**(5): p. 613-21.
157. West, A.E. and M.E. Greenberg, *Neuronal activity-regulated gene transcription in synapse development and cognitive function*. Cold Spring Harb Perspect Biol, 2011. **3**(6).
158. Madabhushi, R., et al., *Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes*. Cell, 2015. **161**(7): p. 1592-605.
159. Altmann, T. and A.R. Gennery, *DNA ligase IV syndrome; a review*. Orphanet J Rare Dis, 2016. **11**(1): p. 137.
160. Grunebaum, E., A. Bates, and C.M. Roifman, *Omenn syndrome is associated with mutations in DNA ligase IV*. J Allergy Clin Immunol, 2008. **122**(6): p. 1219-20.
161. Shen, J., et al., *Mutations in PNKP cause microcephaly, seizures and defects in DNA repair*. Nat Genet, 2010. **42**(3): p. 245-9.
162. Li, G., et al., *Small molecules enhance CRISPR/Cas9-mediated homology-directed genome editing in primary cells*. Sci Rep, 2017. **7**(1): p. 8943.
163. Luijsterburg, M.S. and H. van Attikum, *Chromatin and the DNA damage response: the cancer connection*. Mol Oncol, 2011. **5**(4): p. 349-67.
164. Beucher, A., et al., *ATM and Artemis promote homologous recombination of radiation-induced DNA double-strand breaks in G2*. EMBO J, 2009. **28**(21): p. 3413-27.
165. Chanut, P., et al., *Coordinated nuclease activities counteract Ku at single-ended DNA double-strand breaks*. Nat Commun, 2016. **7**: p. 12889.
166. Chang, H.H.Y., et al., *Non-homologous DNA end joining and alternative pathways to double-strand break repair*. Nat Rev Mol Cell Biol, 2017. **18**(8): p. 495-506.

167. Zhou, Y., et al., *Regulation of the DNA Damage Response by DNA-PKcs Inhibitory Phosphorylation of ATM*. *Mol Cell*, 2017. **65**(1): p. 91-104.
168. Escribano-Diaz, C., et al., *A cell cycle-dependent regulatory circuit composed of 53BP1-RIF1 and BRCA1-CtIP controls DNA repair pathway choice*. *Mol Cell*, 2013. **49**(5): p. 872-83.
169. Ceccaldi, R., B. Rondinelli, and A.D. D'Andrea, *Repair Pathway Choices and Consequences at the Double-Strand Break*. *Trends Cell Biol*, 2016. **26**(1): p. 52-64.
170. Shuman, S. and M.S. Glickman, *Bacterial DNA repair by non-homologous end joining*. *Nat Rev Microbiol*, 2007. **5**(11): p. 852-61.
171. Chang, H.H., et al., *Different DNA End Configurations Dictate Which NHEJ Components Are Most Important for Joining Efficiency*. *J Biol Chem*, 2016. **291**(47): p. 24377-24389.
172. Difilippantonio, M.J., et al., *DNA repair protein Ku80 suppresses chromosomal aberrations and malignant transformation*. *Nature*, 2000. **404**(6777): p. 510-4.
173. Davis, A.J., B.P. Chen, and D.J. Chen, *DNA-PK: a dynamic enzyme in a versatile DSB repair pathway*. *DNA Repair (Amst)*, 2014. **17**: p. 21-9.
174. Kurosawa, A., et al., *The requirement of Artemis in double-strand break repair depends on the type of DNA damage*. *DNA Cell Biol*, 2008. **27**(1): p. 55-61.
175. Kusumoto, R., et al., *Werner protein cooperates with the XRCC4-DNA ligase IV complex in end-processing*. *Biochemistry*, 2008. **47**(28): p. 7548-56.
176. Robert, I., F. Dantzer, and B. Reina-San-Martin, *Parp1 facilitates alternative NHEJ, whereas Parp2 suppresses IgH/c-myc translocations during immunoglobulin class switch recombination*. *J Exp Med*, 2009. **206**(5): p. 1047-56.
177. Betermier, M., P. Bertrand, and B.S. Lopez, *Is non-homologous end-joining really an inherently error-prone process?* *PLoS Genet*, 2014. **10**(1): p. e1004086.
178. Iliakis, G., T. Murmann, and A. Soni, *Alternative end-joining repair pathways are the ultimate backup for abrogated classical non-homologous end-joining and homologous recombination repair: Implications for the formation of chromosome translocations*. *Mutat Res Genet Toxicol Environ Mutagen*, 2015. **793**: p. 166-75.
179. Deriano, L. and D.B. Roth, *Modernizing the nonhomologous end-joining repertoire: alternative and classical NHEJ share the stage*. *Annu Rev Genet*, 2013. **47**: p. 433-55.
180. Yu, A.M. and M. McVey, *Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions*. *Nucleic Acids Res*, 2010. **38**(17): p. 5706-17.
181. Simsek, D. and M. Jasin, *Alternative end-joining is suppressed by the canonical NHEJ component Xrcc4-ligase IV during chromosomal translocation formation*. *Nat Struct Mol Biol*, 2010. **17**(4): p. 410-6.
182. Lieber, M.R., *Mechanisms of human lymphoid chromosomal translocations*. *Nat Rev Cancer*, 2016. **16**(6): p. 387-98.
183. Heyer, W.D., K.T. Ehmsen, and J. Liu, *Regulation of homologous recombination in eukaryotes*. *Annu Rev Genet*, 2010. **44**: p. 113-39.
184. Masani, S., et al., *Redundant function of DNA ligase 1 and 3 in alternative end-joining during immunoglobulin class switch recombination*. *Proc Natl Acad Sci U S A*, 2016. **113**(5): p. 1261-6.
185. Han, L. and K. Yu, *Altered kinetics of nonhomologous end joining and class switch recombination in ligase IV-deficient B cells*. *J Exp Med*, 2008. **205**(12): p. 2745-53.
186. Smith, R., et al., *CHD3 and CHD4 recruitment and chromatin remodeling activity at DNA breaks is promoted by early poly(ADP-ribose)-dependent chromatin relaxation*. *Nucleic Acids Res*, 2018. **46**(12): p. 6087-6098.

187. Taty-Taty, G.C., et al., *Control of alternative end joining by the chromatin remodeler p400 ATPase*. Nucleic Acids Res, 2016. **44**(4): p. 1657-68.
188. Boboila, C., et al., *Alternative end-joining catalyzes robust IgH locus deletions and translocations in the combined absence of ligase 4 and Ku70*. Proc Natl Acad Sci U S A, 2010. **107**(7): p. 3034-9.
189. Tsuji, H., et al., *Involvement of illegitimate V(D)J recombination or microhomology-mediated nonhomologous end-joining in the formation of intragenic deletions of the Notch1 gene in mouse thymic lymphomas*. Cancer Res, 2004. **64**(24): p. 8882-90.
190. Lee, H.J., et al., *Alteration of Genomic Imprinting Status of Human Parthenogenetic Induced Pluripotent Stem Cells during Neural Lineage Differentiation*. Int J Stem Cells, 2019.
191. Topol A, T.N., Brennand KJ, *A Guide to Generating and Using hiPSC Derived NPCs for the Study of Neurological Disease*. Journal of Visualised Experiments, 2015(96): p. 52495.
192. Baghbaderani, B.A., et al., *Detailed Characterization of Human Induced Pluripotent Stem Cells Manufactured for Therapeutic Applications*. Stem Cell Rev, 2016. **12**(4): p. 394-420.
193. Software, B. *Benchling*. 2019; Available from: <https://benchling.com>.
194. Schneider, C.A.R., W.S & Eliceiri, K.W., *NIH Image to ImageJ: 25 years of image analysis*. Nature Methods, 2012. **9**(7): p. 671-675.
195. Stahlberg, A., et al., *Simple multiplexed PCR-based barcoding of DNA for ultrasensitive mutation detection by next-generation sequencing*. Nat Protoc, 2017. **12**(4): p. 664-682.
196. Atom. *Atom*. 2019.
197. -, *Python manual online*. 2018.
198. Team, R.C. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing 2019.
199. Schneider, C.A., W.S. Rasband, and K.W. Eliceiri, *NIH Image to ImageJ: 25 years of image analysis*. Nat Methods, 2012. **9**(7): p. 671-5.
200. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform*. Bioinformatics, 2009. **25**(14): p. 1754-60.
201. Li, H., *Minimap2: fast pairwise alignment for long nucleotide sequences*. arXiv, 2017. **1708.01492**.
202. Fritz J Sedlazeck, P.R., Moritz Smolka, Han Fang, Maria Nattestad, Arndt von Haeseler, Michael Schatz, *Accurate detection of complex structural variations using single molecule sequencing*. BioRxiv, 2017.
203. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner*. Bioinformatics, 2013. **29**(1): p. 15-21.
204. Picard, *Picard*. 2019.
205. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
206. Quinlan, A.R., *BEDTools: The Swiss-Army Tool for Genome Feature Analysis*. Curr Protoc Bioinformatics, 2014. **47**: p. 11 12 1-34.
207. Robinson, J.T., et al., *Integrative genomics viewer*. Nat Biotechnol, 2011. **29**(1): p. 24-6.
208. S, A., *FASTQC: a quality control tool for high throughput sequence data*. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>. 2010.
209. Koren, S., et al., *Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation*. Genome Res, 2017. **27**(5): p. 722-736.

210. Liao, Y., G.K. Smyth, and W. Shi, *The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads*. *Nucleic Acids Res*, 2019. **47**(8): p. e47.
211. Bamtools, *Bamtools*. 2019.
212. Griffith, O. *Bam-readcount*. 2019.
213. Wick, R., *Porechop*. Online, 2017.
214. IEEE, *Matplotlib: a 2D graphics environment*. 2007. **9**(3): p. 90-95.
215. Plotly Technologies Inc. Title: Collaborative data science Publisher: Plotly Technologies Inc. Place of publication: Montréal, Q., *plotly*. 2015.
216. Jones E, O.E., Peterson P, et al. *SciPy: Open Source Scientific Tools for Python*. 2001 [cited 2019 2019].
217. Guberman, J.M., et al., *BioMart Central Portal: an open database network for the biological community*. Database (Oxford), 2011. **2011**: p. bar041.
218. Smedley, D., et al., *The BioMart community portal: an innovative alternative to large, centralized data repositories*. *Nucleic Acids Res*, 2015. **43**(W1): p. W589-98.
219. Wickham, H., *ggplot2: Elegant Graphics for Data Analysis*. 2016: Springer-Verlag New York.
220. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. *Genome Biol*, 2014. **15**(12): p. 550.
221. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data*. *Bioinformatics*, 2014. **30**(15): p. 2114-20.
222. Baruzzo, G., et al., *Simulation-based comprehensive benchmarking of RNA-seq aligners*. *Nat Methods*, 2017. **14**(2): p. 135-139.
223. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features*. *Bioinformatics*, 2014. **30**(7): p. 923-30.
224. Chatterjee, A., Ahn A, Rodger EJ, Stockwell PA, Eccles MR, *A Guide for Designing and Analyzing RNA-Seq Data*, in *Gene Expression Analysis*. 2018. p. 35-80.
225. 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. 839-51.
226. Eden, E., et al., *GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists*. *BMC Bioinformatics*, 2009. **10**: p. 48.
227. Supek, F., et al., *REVIGO summarizes and visualizes long lists of gene ontology terms*. *PLoS One*, 2011. **6**(7): p. e21800.
228. Shannon, P., et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks*. *Genome Res*, 2003. **13**(11): p. 2498-504.
229. Mandegar, M.A., et al., *CRISPR Interference Efficiently Induces Specific and Reversible Gene Silencing in Human iPSCs*. *Cell Stem Cell*, 2016. **18**(4): p. 541-53.
230. Doench, J.G., et al., *Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9*. *Nat Biotechnol*, 2016. **34**(2): p. 184-191.
231. Doench, J.G., et al., *Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation*. *Nat Biotechnol*, 2014. **32**(12): p. 1262-7.
232. Hsu, P.D., et al., *DNA targeting specificity of RNA-guided Cas9 nucleases*. *Nat Biotechnol*, 2013. **31**(9): p. 827-32.
233. Takenobu N, K.H., Marumoto T, Sakuma T, Yamamoto T, Tani K, *Single-Cell-State Culture of Human Pluripotent Stem Cells Increases Transfection Efficiency*. *BioResearch Open Access*, 2016. **5**(1): p. 127-136.
234. Zheng, Q., et al., *Precise gene deletion and replacement using the CRISPR/Cas9 system in human cells*. *Biotechniques*, 2014. **57**(3): p. 115-24.

235. Sariyer, I.K., *Transfection of neuronal cultures*. Methods Mol Biol, 2013. **1078**: p. 133-9.
236. Han, Z., et al., *Establishment of Lipofection Protocol for Efficient miR-21 Transfection into Cortical Neurons In Vitro*. DNA Cell Biol, 2015. **34**(12): p. 703-9.
237. Manivasakam, P., et al., *Restriction enzymes increase efficiencies of illegitimate DNA integration but decrease homologous integration in mammalian cells*. Nucleic Acids Res, 2001. **29**(23): p. 4826-33.
238. Kim, H.J., et al., *Targeted genome editing in human cells with zinc finger nucleases constructed via modular assembly*. Genome Res, 2009. **19**(7): p. 1279-88.
239. Joung, J.K. and J.D. Sander, *TALNs: a widely applicable technology for targeted genome editing*. Nat Rev Mol Cell Biol, 2013. **14**(1): p. 49-55.
240. Juillerat, A., et al., *Comprehensive analysis of the specificity of transcription activator-like effector nucleases*. Nucleic Acids Res, 2014. **42**(8): p. 5390-402.
241. Cong, L., et al., *Multiplex genome engineering using CRISPR/Cas systems*. Science, 2013. **339**(6121): p. 819-23.
242. You, L., et al., *Advancements and Obstacles of CRISPR-Cas9 Technology in Translational Research*. Mol Ther Methods Clin Dev, 2019. **13**: p. 359-370.
243. Moutal, A., et al., *CRISPR/Cas9 editing of Nf1 gene identifies CRMP2 as a therapeutic target in neurofibromatosis type 1 (NF1)-related pain that is reversed by (S)-Lacosamide*. Pain, 2017.
244. Jansen, R., et al., *Identification of genes that are associated with DNA repeats in prokaryotes*. Mol Microbiol, 2002. **43**(6): p. 1565-75.
245. Bolotin, A., et al., *Complete sequence and comparative genome analysis of the dairy bacterium Streptococcus thermophilus*. Nat Biotechnol, 2004. **22**(12): p. 1554-8.
246. Barrangou, R., et al., *CRISPR provides acquired resistance against viruses in prokaryotes*. Science, 2007. **315**(5819): p. 1709-12.
247. Gasiunas, G., et al., *Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria*. Proc Natl Acad Sci U S A, 2012. **109**(39): p. E2579-86.
248. Carte, J., et al., *Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes*. Genes Dev, 2008. **22**(24): p. 3489-96.
249. Garneau, J.E., et al., *The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA*. Nature, 2010. **468**(7320): p. 67-71.
250. Deltcheva, E., et al., *CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III*. Nature, 2011. **471**(7340): p. 602-7.
251. Jinek, M., et al., *A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity*. Science, 2012. **337**(6096): p. 816-21.
252. Hsu, P.D., E.S. Lander, and F. Zhang, *Development and applications of CRISPR-Cas9 for genome engineering*. Cell, 2014. **157**(6): p. 1262-78.
253. IDTDNA. *Estimation of PAM sites in the human genome*. online 2019 [cited 2019 15th May 2019].
254. Zhang, X.H., et al., *Off-target Effects in CRISPR/Cas9-mediated Genome Engineering*. Mol Ther Nucleic Acids, 2015. **4**: p. e264.
255. Zhang, Y., et al., *Comparison of non-canonical PAMs for CRISPR/Cas9-mediated DNA cleavage in human cells*. Sci Rep, 2014. **4**: p. 5405.
256. Yu, X., et al., *Improved delivery of Cas9 protein/gRNA complexes using lipofectamine CRISPRMAX*. Biotechnol Lett, 2016. **38**(6): p. 919-29.
257. Tsai, S.Q., et al., *GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases*. Nat Biotechnol, 2015. **33**(2): p. 187-197.

258. Chakrabarti, A.M., et al., *Target-Specific Precision of CRISPR-Mediated Genome Editing*. Mol Cell, 2019. **73**(4): p. 699-713 e6.
259. Chen, W., et al., *Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair*. Nucleic Acids Res, 2019. **47**(15): p. 7989-8003.
260. Horlbeck, M.A., et al., *Nucleosomes impede Cas9 access to DNA in vivo and in vitro*. Elife, 2016. **5**.
261. Kime, C., et al., *Efficient CRISPR/Cas9-Based Genome Engineering in Human Pluripotent Stem Cells*. Curr Protoc Hum Genet, 2016. **88**: p. Unit 21 4.
262. Henser-Brownhill, T., J. Monserrat, and P. Scaffidi, *Generation of an arrayed CRISPR-Cas9 library targeting epigenetic regulators: from high-content screens to in vivo assays*. Epigenetics, 2017. **12**(12): p. 1065-1075.
263. Bell, S., et al., *A Rapid Pipeline to Model Rare Neurodevelopmental Disorders with Simultaneous CRISPR/Cas9 Gene Editing*. Stem Cells Transl Med, 2017. **6**(3): p. 886-896.
264. Zhu, Z., F. Gonzalez, and D. Huangfu, *The iCRISPR platform for rapid genome editing in human pluripotent stem cells*. Methods Enzymol, 2014. **546**: p. 215-50.
265. Gonzalez, F., et al., *An iCRISPR platform for rapid, multiplexable, and inducible genome editing in human pluripotent stem cells*. Cell Stem Cell, 2014. **15**(2): p. 215-226.
266. Papapetrou, E.P. and A. Schambach, *Gene Insertion Into Genomic Safe Harbors for Human Gene Therapy*. Mol Ther, 2016. **24**(4): p. 678-84.
267. Vouillot, L., A. Thelie, and N. Pollet, *Comparison of T7E1 and surveyor mismatch cleavage assays to detect mutations triggered by engineered nucleases*. G3 (Bethesda), 2015. **5**(3): p. 407-15.
268. Qiu, P., et al., *Mutation detection using Surveyor nuclease*. Biotechniques, 2004. **36**(4): p. 702-7.
269. Sanger, F., G.G. Brownlee, and B.G. Barrell, *A two-dimensional fractionation procedure for radioactive nucleotides*. J Mol Biol, 1965. **13**(2): p. 373-98.
270. Sanger, F., S. Nicklen, and A.R. Coulson, *DNA sequencing with chain-terminating inhibitors*. Proc Natl Acad Sci U S A, 1977. **74**(12): p. 5463-7.
271. Waterston, R.H., E.S. Lander, and J.E. Sulston, *On the sequencing of the human genome*. Proc Natl Acad Sci U S A, 2002. **99**(6): p. 3712-6.
272. Heather, J.M. and B. Chain, *The sequence of sequencers: The history of sequencing DNA*. Genomics, 2016. **107**(1): p. 1-8.
273. Genomes Project, C., et al., *A global reference for human genetic variation*. Nature, 2015. **526**(7571): p. 68-74.
274. Muir, P., et al., *The real cost of sequencing: scaling computation to keep pace with data generation*. Genome Biol, 2016. **17**: p. 53.
275. Bell, C.C., et al., *A high-throughput screening strategy for detecting CRISPR-Cas9 induced mutations using next-generation sequencing*. BMC Genomics, 2014. **15**: p. 1002.
276. Ambardar, S., et al., *High Throughput Sequencing: An Overview of Sequencing Chemistry*. Indian J Microbiol, 2016. **56**(4): p. 394-404.
277. Ronaghi, M., *Pyrosequencing sheds light on DNA sequencing*. Genome Res, 2001. **11**(1): p. 3-11.
278. Malapelle, U., et al., *Ion Torrent next-generation sequencing for routine identification of clinically relevant mutations in colorectal cancer patients*. J Clin Pathol, 2015. **68**(1): p. 64-8.

279. Jain, M., et al., *The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community*. Genome Biol, 2016. **17**(1): p. 239.
280. Laver, T., et al., *Assessing the performance of the Oxford Nanopore Technologies MinION*. Biomol Detect Quantif, 2015. **3**: p. 1-8.
281. Rang, F.J., W.P. Kloosterman, and J. de Ridder, *From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy*. Genome Biol, 2018. **19**(1): p. 90.
282. Tyler, A.D., et al., *Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications*. Sci Rep, 2018. **8**(1): p. 10931.
283. Jain, M., et al., *Nanopore sequencing and assembly of a human genome with ultra-long reads*. Nat Biotechnol, 2018. **36**(4): p. 338-345.
284. Rhoads, A. and K.F. Au, *PacBio Sequencing and Its Applications*. Genomics Proteomics Bioinformatics, 2015. **13**(5): p. 278-89.
285. Kim, B.Y., et al., *Optimized detection of insertions/deletions (INDELs) in whole-exome sequencing data*. PLoS One, 2017. **12**(8): p. e0182272.
286. Hayashi, Y., et al., *Structure-based discovery of NANOG variant with enhanced properties to promote self-renewal and reprogramming of pluripotent stem cells*. Proc Natl Acad Sci U S A, 2015. **112**(15): p. 4666-71.
287. Sovic, I., et al., *Fast and sensitive mapping of nanopore sequencing reads with GraphMap*. Nat Commun, 2016. **7**: p. 11307.
288. Janssen, J.M., et al., *The Chromatin Structure of CRISPR-Cas9 Target DNA Controls the Balance between Mutagenic and Homology-Directed Gene-Editing Events*. Mol Ther Nucleic Acids, 2019. **16**: p. 141-154.
289. Ryan R. Wick, L.M.J., Kathryn E. Holt, *Performance of neural network basecalling tools for Oxford Nanopore sequencing*. bioRxiv, 2019(<http://dx.doi.org/10.1101/543439>).
290. Fink, J.J. and E.S. Levine, *Uncovering True Cellular Phenotypes: Using Induced Pluripotent Stem Cell-Derived Neurons to Study Early Insults in Neurodevelopmental Disorders*. Front Neurol, 2018. **9**: p. 237.
291. Kemp, P.J., et al., *Improving and accelerating the differentiation and functional maturation of human stem cell-derived neurons: role of extracellular calcium and GABA*. J Physiol, 2016. **594**(22): p. 6583-6594.
292. Zhang, Q., et al., *Electrical Stimulation with a Conductive Polymer Promotes Neurite Outgrowth and Synaptogenesis in Primary Cortical Neurons in 3D*. Sci Rep, 2018. **8**(1): p. 9855.
293. Iqbal, N. and N. Iqbal, *Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications*. Mol Biol Int, 2014. **2014**: p. 852748.
294. Goh, P.A., et al., *A systematic evaluation of integration free reprogramming methods for deriving clinically relevant patient specific induced pluripotent stem (iPS) cells*. PLoS One, 2013. **8**(11): p. e81622.
295. Pan, G. and J.A. Thomson, *Nanog and transcriptional networks in embryonic stem cell pluripotency*. Cell Res, 2007. **17**(1): p. 42-9.
296. Englund, C., et al., *Pax6, Tbr2, and Tbr1 are expressed sequentially by radial glia, intermediate progenitor cells, and postmitotic neurons in developing neocortex*. J Neurosci, 2005. **25**(1): p. 247-51.
297. Papanayotou, C., et al., *A mechanism regulating the onset of Sox2 expression in the embryonic neural plate*. PLoS Biol, 2008. **6**(1): p. e2.

298. Sanchez, C., J. Diaz-Nido, and J. Avila, *Phosphorylation of microtubule-associated protein 2 (MAP2) and its relevance for the regulation of the neuronal cytoskeleton function*. Prog Neurobiol, 2000. **61**(2): p. 133-68.
299. El-Husseini, A.E., et al., *PSD-95 involvement in maturation of excitatory synapses*. Science, 2000. **290**(5495): p. 1364-8.
300. Wu, J.Q., et al., *Dynamic transcriptomes during neural differentiation of human embryonic stem cells revealed by short, long, and paired-end sequencing*. Proc Natl Acad Sci U S A, 2010. **107**(11): p. 5254-9.
301. Koch, C.M., et al., *A Beginner's Guide to Analysis of RNA Sequencing Data*. Am J Respir Cell Mol Biol, 2018. **59**(2): p. 145-157.
302. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles*. Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
303. The Gene Ontology, C., *The Gene Ontology Resource: 20 years and still GOing strong*. Nucleic Acids Res, 2019. **47**(D1): p. D330-D338.
304. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium*. Nat Genet, 2000. **25**(1): p. 25-9.
305. Wildeman, M., et al., *Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker*. Hum Mutat, 2008. **29**(1): p. 6-13.
306. Choi, Y. and A.P. Chan, *PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels*. Bioinformatics, 2015. **31**(16): p. 2745-7.
307. Hu, J. and P.C. Ng, *SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins*. PLoS One, 2013. **8**(10): p. e77940.
308. Yuan, A., et al., *Neurofilaments at a glance*. J Cell Sci, 2012. **125**(Pt 14): p. 3257-63.
309. Shi, G. and Y. Jin, *Role of Oct4 in maintaining and regaining stem cell pluripotency*. Stem Cell Res Ther, 2010. **1**(5): p. 39.
310. Kim, E.J., et al., *Ascl1 (Mash1) defines cells with long-term neurogenic potential in subgranular and subventricular zones in adult mouse brain*. PLoS One, 2011. **6**(3): p. e18472.
311. Menezes, J.R. and M.B. Luskin, *Expression of neuron-specific tubulin defines a novel population in the proliferative layers of the developing telencephalon*. J Neurosci, 1994. **14**(9): p. 5399-416.
312. Green, R.C., et al., *ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing*. Genet Med, 2013. **15**(7): p. 565-74.
313. Richards, S., et al., *Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology*. Genet Med, 2015. **17**(5): p. 405-24.
314. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation*. Nucleic Acids Res, 2001. **29**(1): p. 308-11.
315. Landrum, M.J., et al., *ClinVar: improving access to variant interpretations and supporting evidence*. Nucleic Acids Res, 2018. **46**(D1): p. D1062-D1067.
316. Karczewski, K.J., et al., *The ExAC browser: displaying reference data information from over 60 000 exomes*. Nucleic Acids Res, 2017. **45**(D1): p. D840-D845.
317. Durieux, A.M.S., et al., *Cortical and subcortical glutathione levels in adults with autism spectrum disorder*. Autism Res, 2016. **9**(4): p. 429-435.
318. Collins, A.R., *The comet assay for DNA damage and repair: principles, applications, and limitations*. Mol Biotechnol, 2004. **26**(3): p. 249-61.

319. Wang, Y., et al., *Evaluation of the comet assay for assessing the dose-response relationship of DNA damage induced by ionizing radiation*. *Int J Mol Sci*, 2013. **14**(11): p. 22449-61.
320. Bouwman, B.A.M. and N. Crosetto, *Endogenous DNA Double-Strand Breaks during DNA Transactions: Emerging Insights and Methods for Genome-Wide Profiling*. *Genes (Basel)*, 2018. **9**(12).
321. Yan, W.X., et al., *BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks*. *Nat Commun*, 2017. **8**: p. 15058.
322. Klein, I.A., et al., *Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes*. *Cell*, 2011. **147**(1): p. 95-106.
323. Frock, R.L., et al., *Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases*. *Nat Biotechnol*, 2015. **33**(2): p. 179-86.
324. Hu, Z., et al., *Ligase IV inhibitor SCR7 enhances gene editing directed by CRISPR-Cas9 and ssODN in human cancer cells*. *Cell Biosci*, 2018. **8**: p. 12.
325. Crosetto, N., et al., *Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing*. *Nat Methods*, 2013. **10**(4): p. 361-5.
326. Canela, A., et al., *DNA Breaks and End Resection Measured Genome-wide by End Sequencing*. *Mol Cell*, 2016. **63**(5): p. 898-911.
327. Mehta, A. and J.E. Haber, *Sources of DNA double-strand breaks and models of recombinational DNA repair*. *Cold Spring Harb Perspect Biol*, 2014. **6**(9): p. a016428.
328. Symington, L.S. and J. Gautier, *Double-strand break end resection and repair pathway choice*. *Annu Rev Genet*, 2011. **45**: p. 247-71.
329. Kumar, R., et al., *HumCFS: a database of fragile sites in human chromosomes*. *BMC Genomics*, 2019. **19**(Suppl 9): p. 985.
330. Marques, M., et al., *Reconciling the positive and negative roles of histone H2A.Z in gene transcription*. *Epigenetics*, 2010. **5**(4): p. 267-72.
331. Fernandez, A.F., et al., *H3K4me1 marks DNA regions hypomethylated during aging in human stem and differentiated cells*. *Genome Res*, 2015. **25**(1): p. 27-40.
332. Thys, R.G., et al., *DNA secondary structure at chromosomal fragile sites in human disease*. *Curr Genomics*, 2015. **16**(1): p. 60-70.
333. Kim, J., et al., *Replication Stress Shapes a Protective Chromatin Environment across Fragile Genomic Regions*. *Mol Cell*, 2018. **69**(1): p. 36-47 e7.
334. Piovesan, A., et al., *On the length, weight and GC content of the human genome*. *BMC Res Notes*, 2019. **12**(1): p. 106.
335. Kim, N. and S. Jinks-Robertson, *Transcription as a source of genome instability*. *Nat Rev Genet*, 2012. **13**(3): p. 204-14.
336. Lin, Y.L. and P. Pasero, *Transcription-Replication Conflicts: Orientation Matters*. *Cell*, 2017. **170**(4): p. 603-604.
337. Garcia-Muse, T. and A. Aguilera, *Transcription-replication conflicts: how they occur and how they are resolved*. *Nat Rev Mol Cell Biol*, 2016. **17**(9): p. 553-63.
338. Core, L.J., J.J. Waterfall, and J.T. Lis, *Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters*. *Science*, 2008. **322**(5909): p. 1845-8.
339. Schaefer-Klein, J.L., et al., *Topoisomerase 2 Alpha Cooperates with Androgen Receptor to Contribute to Prostate Cancer Progression*. *PLoS One*, 2015. **10**(11): p. e0142327.
340. Ip, C.L.C., et al., *MinION Analysis and Reference Consortium: Phase I data release and analysis*. *F1000Res*, 2015. **4**: p. 1075.

REFERENCES

341. Rodgers, K. and M. McVey, *Error-Prone Repair of DNA Double-Strand Breaks*. J Cell Physiol, 2016. **231**(1): p. 15-24.
342. Li, Y., et al., *Transcriptome analysis reveals determinant stages controlling human embryonic stem cell commitment to neuronal cells*. J Biol Chem, 2017. **292**(48): p. 19590-19604.
343. Khan, F.A. and S.O. Ali, *Physiological Roles of DNA Double-Strand Breaks*. J Nucleic Acids, 2017. **2017**: p. 6439169.
344. Pan-Hammarstrom, Q., et al., *Impact of DNA ligase IV on nonhomologous end joining pathways during class switch recombination in human cells*. J Exp Med, 2005. **201**(2): p. 189-94.
345. Jiang, J., et al., *Molecular and immunological characterization of DNA ligase IV deficiency*. Clin Immunol, 2016. **163**: p. 75-83.
346. Gatz, S.A., et al., *Requirement for DNA ligase IV during embryonic neuronal development*. J Neurosci, 2011. **31**(27): p. 10088-100.