Characterising Structural Variants in Patients with Craniosynostosis using Short-Read and Long-Range Technologies

Yang Pei

Jesus College, University of Oxford

DPhil in Medical Sciences

Supervisors: Associate Professor Stephen R. F. Twigg, Professor Richard J. Gibbons, Professor Andrew O. M. Wilkie

Word Count: 46,888

Acknowledgement

Completing a research project and crafting a DPhil thesis has been an extraordinary journey, one made possible by the support of numerous individuals. I am deeply grateful to each one of them.

First and foremost, I am immensely grateful to my supervisors, Assoc Prof Stephen Twigg and Prof Andrew Wilkie, whose guidance, expertise, and continual support have been pivotal in shaping this thesis. Their insightful feedback and encouragement have not only directed the course of my research but have also instilled in me a profound sense of purpose – both in helping patients and their families as well in contributing to the field of clinical genetics and science in large. I particularly admired their meticulous and thorough scientific approaches, which have been instrumental in shaping how I would conduct my own research in science. My gratitude also extends to my third supervisor, Prof Richard Gibbons, for his constructive feedback and valuable discussions throughout all our progress meetings.

I am grateful to everyone in the clinical genetics group, both past and present. I'd like to thank Dr Eduardo Calpena, who provided incredible amount of hands-on support both experimentally and bioinformatically. He has been a patient mentor, a valuable colleague, and a dear friend throughout my entire project. I also wish to thank Dr Hana Mlcochova, Dr Ummi Abdullah, and Dr Maria Giebler for their incredible support and friendship during my initial time in the lab. I would like to thank the students in the lab, including Dr J. Heather Vedovato dos Santos, Mr Isaac Walton, and particularly Dr Rebecca Trim (Née Tooze), for the great moments we have shared throughout the

years. I wish to thank Dr Dagmara Korona for her help with several experiments as well as her support at coffee breaks making difficult days easier to get through.

I want to thank Ms Jill Brown and Dr Barbara Xella for their help with experimental parts of my project (FISH and Bionano OGM, respectively). I wish to thank my thesis committee, Dr Alistair Pagnamenta and Prof Jim Hughes, for their excellent advice and encouragement throughout our meetings.

I am thankful to Jesus College, University of Oxford, for their support in enabling me to attend and/or present my work at several conferences, notably the European Society of Human Genetics conference (2023). Moreover, their continuous support throughout my endeavour as a competitive student athlete has been immensely encouraging. Additionally, being honoured as a Jesus College Graduate Scholar was a profound privilege that I deeply appreciated.

I would like to thank all patients and their families for making this project possible. Knowing I have helped multiple families in their journeys looking for diagnoses gave me significant sense of achievement and purpose. Additionally, meeting a family in person and witnessing how my research positively impacted their lives was an incredible honour.

I would like to thank my friends and family. I am grateful for all my friends in Oxford, especially people I have met during my time at the Oxford University Company of

Archers. I extend my gratitude and love to my father, Mr Yuhai Pei (裴玉海), and my mother, Prof Xiangping Zhang（张香平）. Their finical support made this project possible for me in the first place. Their belief in my abilities and their constant encouragement have been the driving force that propelled me forward. Finally, I'd like to thank my partner, Ms Ecaterina Pogorenii, for your patient and support throughout my journey at Oxford and in the years to come in the future!

This research was made possible through access to data in the National Genomic Research Library, which is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The National Genomic Research Library holds data provided by patients and collected by the NHS as part of their care and data collected as part of their participation in research. The National Genomic Research Library is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure.

Statement of Authorship

I declare that the content of this thesis is my own work, with the exceptions of the following collaborations: Whole genome sequencing (WGS) data from the 100,000-genome project (100kGP) [**section 2.4**] were generated by Genomics England. Relevant data include raw alignment data (.bam files) and structural variant (SV) files (.vcf files) from variant callers Manta and Canvas. Fluorescence in situ hybridisation (FISH) for case 10 (**section 2.6**) was carried out under the supervision of Ms Jill Brown (MRC MHU Unit, WIMM, Oxford), who performed the probe precipitation, denaturation, hybridisation, and slides processing as well as image generation. Oxford Nanopore Technology (ONT) WGS data (**section 2.8**) were generated from DNA samples that I personally extracted and prepared for sequencing by the team under the supervisor of Prof Greg Elgar (Genomics England), as part of a 100kGP pilot program. Data processing up to the point of individual SV calling using Sniffles/Sniffles2 was carried out by the Elgar group. Case 19 (**section 5.3**) was a collaboration project where the initial WGS analysis was carried out by Dr Alistair Pagnamenta (Centre for Human Genetics, Oxford). Induced Pluripotent Stem Cell (iPSCs) culturing, RNA extraction, and RNA-seq data generation for case 10 (**section 5.4**) was carried out fully by Dr Dagmara Korona (Clinical Genetics Group, Oxford). Illumina WGS data for case 2 (**section 5.5**) were generated by Novogene and data processing up to alignment (.bam files) was carried out by Dr Simon McGowan (WIMM CCB, Oxford). FISH images for case 2 (**section 5.5**) were generated entirely by Great Ormond Street Hospital Clinical Genetics team.

Publications

Hyder Z[#], Calpena E[#], **Pei Y[#]**, Tooze RS[#], Brittain H, Twigg SRF, Cilliers D, Morton JEV, McCann E, Weber A, Wilson LC, Douglas AGL, McGowan R, Need A, Bond A, Tavares ALT, Thomas ERA; Genomics England Research Consortium; Hill SL, Deans ZC, Boardman-Pretty F, Caulfield M, Scott RH, Wilkie AOM. Evaluating the performance of a clinical genome sequencing program for diagnosis of rare genetic disease, seen through the lens of craniosynostosis. Genet Med. 2021 Dec;23(12):2360-2368. doi: 10.1038/s41436-021-01297-5. Epub 2021 Aug 25. PMID: 34429528; PMCID: PMC8629760. **[#]Contributed equally**

Moore AR, Yu J, **Pei Y**, Cheng EWY, Taylor Tavares AL, Walker WT, Thomas NS, Kamath A, Ibitoye R, Josifova D, Wilsdon A, Ross A, Calder AD, Offiah AC, Wilkie AOM; Genomics England Research Consortium; Taylor JC, Pagnamenta AT. Use of genome sequencing to hunt for cryptic second-hit variants: analysis of 31 cases recruited to the 100 000 Genomes Project. J Med Genet. 2023 Nov 27;60(12):1235-1244. doi: 10.1136/jmg-2023-109362. PMID: 37558402; PMCID: PMC10715503.

Conference Oral Presentation

European Human Genetics Conference (ESHG) 2023

- Characterising clinically relevant complex structural variants in craniosynostosis using long-range technologies

Table of Contents

Acknowledgement ..... *i*
Statement of Authorship ..... *iv*
Publications ..... *v*
Conference Oral Presentation ..... *vi*
Table of Contents ..... *vii*
List of Figures ..... *xi*
List of Tables ..... *xv*
List of Abbreviations ..... *xvi*
Abstract ..... *xx*
Chapter 1 Introduction *1*
1.1 Context of this thesis **2**
1.2 Structural variants (SVs) **3**
1.3 SV mechanisms **4**
1.4 SV detection approaches **8**
1.5 SVs in the human population **14**
1.6 SVs in human diseases **17**
1.7 Craniosynostosis and SVs **21**
1.8 Summary, hypothesis, and aims **27**
Chapter 2 Methods *30*
2.1 Ethics **31**
2.2 Array data analysis **31**
2.3 WES data analysis **31**
2.4 100kGP WGS data analysis **32**
2.4.1 VarCount annotation 34
2.4.2 DGV annotation 35
2.4.3 Gap regions annotation 37
2.4.4 Coding regions annotation 37
2.4.5 CRS panel annotation 38
2.4.6 Lumpy concordant matches 39
2.4.7 Segregation analysis 39
2.4.8 SV filtering and prioritisation 40
2.4.9 SV call verification and further analysis 42
2.4.10 Mosaicism detection 43
2.4.11 100kGP rare disease screening 44
2.4.12 SVRare analysis 45

2.5 Breakpoint PCR & Dideoxy-sequencing 47
2.6 Fluorescent in situ hybridisation (FISH) 50
2.6.1 BAC clone culture and DNA extraction. 50
2.6.2 Probe Preparation by Nick Translation 52
2.6.3 Metaphase cell harvesting 53
2.6.4 Slide preparation 53
2.6.5 Slide pre-treatment and chromosomal denaturation 54
2.6.6 Probe precipitation, denaturation, and hybridisation 54
2.6.7 Probe cleaning, blocking, and detection 55
2.7 EBV-transformed lymphoblastoid cell lines generation 56
2.8 Whole genome sequencing using Oxford Nanopore Technologies (ONT) 57
2.8.1 DNA extraction for ONT sequencing 57
2.8.2 ONT data processing 58
2.9 Bionano Optical Genome Mapping (OGM) 59
2.9.1 Bionano high molecular weight (HMW) DNA extraction 59
2.9.2 Bionano direct labelling and staining (DLS) 60
2.9.3 Bionano Saphyr data collection 61
2.9.4 Bionano OGM data analysis 62
2.10 Illumina WGS vs ONT WGS vs Bionano OGM 66
2.10.1 Bionano reference “truth” callset benchmark 67
2.10.2 Bionano SV type normalisation 67
2.10.3 ONT/Illumina sensitivity evaluation 68
2.10.4 ONT *de novo* calling performance evaluation 68
Chapter 3 Results – 100kGP WGS analysis 70
3.1 Introduction 71
3.2 Pathogenic and likely pathogenic SVs 74
3.2.1 Case 24: large chr6 DEL – rediscovery 75
3.2.2 Case 25: DEL affecting *GPC3* - rediscovery 77
3.2.3 Case 29: DUP at *ARX* locus – novel diagnosis 78
3.2.4 Case 21: DUP at *HOXC* gene cluster – novel diagnosis 80
3.3 Structural Variants of Unknown Significance (SVUS) 83
3.3.1 Case 1: chr2 DEL affecting *CERS6* 83
3.3.2 Case 31: small DEL affecting *KIAA0825* 88
3.3.3 Case 27: chr9 DUP affecting *ABL1* locus 91
3.4 Summary 93
Chapter 4 SVs involving the HOXC gene cluster at 12q 97
4.1 Introduction 98
4.2 Evolution and biology of *HOX* gene clusters 98
4.3 Case 21: *HOXC* DUP in craniofacial abnormality 100

4.4 Case 26: *HOXC* DEL in joint and limb malformation 106
4.5 Case 18: *HOXC* CPX split-DUP in congenital heart disease 108
4.6 Interpreting the clinical relevance of *HOXC* SVs 115
4.7 *HOXC* SVs and future direction 127
Chapter 5 Results – Long-range technologies and complex events 130
5.1 Introduction 131
5.2 Case 12: Saethre-Chotzen syndrome 134
5.3 Case 19: complex INS affecting *ENPP1* 140
***5.4* Case 10: CPX INS at *FGF9 locus* 144**
5.4.1 SV identification from WGS 145
5.4.2 SV characterisation with Bionano OGM 152
5.4.3 SV characterisation with FISH 153
5.4.4 Functional analysis: DeepC 157
5.4.5 A competing hypothesis: *FOXP2* frameshifting DEL 158
5.4.6 Functional analysis: RNA-seq 160
5.4.7 Case 10 summary: navigating the dual dilemma of competing hypotheses 164
5.5 Case 2: CPX SV at *KCNJ2* & *16* locus 166
5.5.1 Case 2: syndromic CRS with unusual gum phenotype 167
5.5.2 WGS revealed the complex nature of the event 169
5.5.3 Resolving case 2 complex SV requires both Bionano and FISH 174
5.5.4 Assessing the clinical relevance of the *KCNJ* SVs 178
5.5.5 Unbalanced reciprocal translocation 185
5.5.6 Approaching the detection limit of Bionano OGM 188
5.6 Case 16: CPX SV involving *PLCB4* 189
5.6.1 SV characterisation using WGS, PCR, and dideoxy- sequencing 192
5.6.2 SV characterisation using Bionano OGM 197
5.6.3 Case 16 variant interpretation 203
5.7 Discussion and Summary: challenges and future directions 208
5.7.1 Uncovering clinically relevant SVs 208
5.7.2 Bionano OGM: overcoming technological and computational constraints 209
5.7.3 Addressing the remaining genetic diagnostic gaps in CRS 210
Chapter 6 Result – Comparative analysis of Bionano OGM, Illumina WGS, and ONT WGS 214
6.1 Introduction 215
6.2 Bionano benchmarks 219
6.3 Illumina and ONT WGS performance – Overview 223
6.4 Illumina and ONT WGS performance – SV types 224
6.5 ONT performance – *de novo* calling 227
6.6 Discussion 229

Chapter 7 Closing Remarks 235
7.1 Introduction 236
7.2 Research summary 237
7.3 Technologies and SV detection 240
7.4 Future work 244
7.5 Conclusion 246
References 248
Appendix 276

List of Figures

Figure 1 SVs can be categorised into several subclasses 4
Figure 2 DSBs can be repaired via NAHR and MMEJ/NHEJ....5
Figure 3 The fork stalling and template switching (FoSTeS) mechanism 7
Figure 4 Microhomology mediated break-induced replication (MMBIR) 8
Figure 5 Evolution of SV detection technologies....9
Figure 6 Three-dimensional organisation of chromatids reveals crucial long-range regulatory domains as TADs 19
Figure 7 Chromosome conformation capture (3C) technique 19
Figure 8 DUPs affecting the SOX9-KCNJ2 locus produce different phenotypes 21
Figure 9 Craniosynostosis (CRS) is the premature fusion of one or more skull sutures....22
Figure 10 Mesenchymal cells (MSCs) provide essential niche for suture development 24
Figure 11 Summary of both sequence variants and CNVs in DECIPHER cases with CRS 26
Figure 12 The 100kGP analysis pipeline was designed to identify SV candidates from WGS data in the primary CRS cohort 33
Figure 13 Soft (fuzzy) matching approach for identifying calls representing the same event in different samples 35
Figure 14 Bionano OGM analysis pipeline 63
Figure 15 A large 3.4 Mb de novo DEL was identified on chr6 in the case 24 proband 76
Figure 16 Hemizygous deletion affecting GPC3 in the proband 78
Figure 17 A DUP detected by SVRare at the ARX locus 80
Figure 18 A paternally inherited 286 kb tandem DUP was identified in the case 21 proband 82
Figure 19 Case 1 is a family with three affected individuals and a chr6 de novo deletion in CERS6 is a candidate causative SV....87
Figure 20 Chr5 DEL affecting the last exon of KIAA0825 – a known recessive disease gene for polydactyly 90
Figure 21 Chr9 tandem DUP affecting ABL1, FUBP3, EXOSC2, and PRDM12....92
Figure 22 Four human HOX clusters in the human genome with a total of 39 HOX genes....100
Figure 23 Case 21 family consists of three affected individuals....103
Figure 24 Case 21 HOXC DUP was verified using breakpoint PCR and dideoxy-sequencing 105

Figure 25 Case 26 is a family of four in the 100kGP with three affected individuals 106
Figure 26 A small 64.5 kb DEL was identified in all three affected individuals in case 26 107
Figure 27 Case 18 proband with congenital heart disease without noticeable craniofacial abnormality 108
Figure 28 Case 18 is a family affected by congenital heart disease (CHD) 109
Figure 29 A large CNV gain was detected via WGS in case 18 111
Figure 30 Detailed reads analysis of case 18 SV revealed the complex nature of this event 112
Figure 31 Four possible structures can explain the WGS data with the complex break junctions and the CNV gains 113
Figure 32 Bionano OGM was used to determine the orientations of the two largest purple segments in tandem 114
Figure 33 Long-range PCR and dideoxy-sequencing confirmed the Alt 1 structure as the true structure of this complex SV 115
Figure 34 HOXC cluster and its genomic context 119
Figure 35 DeepC predicted TADs and interpretations for the HOXC cluster for the reference, case 18, and case 21 data 121
Figure 36 Eh mouse model contains a large 46 Mb INV affecting the genomic environment of the Hoxc cluster without directly affecting the coding region of the Hoxcs 123
Figure 37 case 18 SV affects many compelling candidate genes other than the HOXCs 127
Figure 38 Case 12 presented familial Saethre-Chotzen syndrome where both the proband and the mother are affected 135
Figure 39 Bionano OGM analysis identified a large INV ~200 kb telomeric of TWIST1 136
Figure 40 Breakpoint PCR and dideoxy-sequence verification of the INV 138
Figure 41 SVs affecting the TWIST1 critical regulatory region 139
Figure 42 The chr6 SV in Case 19 was identified by WGS and subsequently fully characterised by Bionano OGM 143
Figure 43 Case 10 is a family with sporadic syndromic CRS 145
Figure 44 100kGP WGS analysis detected a de novo interlinked split-DUP on chr13 147
Figure 45 Break point PCR and dideoxy-sequencing verified the two break junctions 148
Figure 46 The origin of the two large, duplicated segments were investigated 149
Figure 47 Three alternative configurations for case 10 CPX SV 152

Figure 48 Bionano OGM detected five molecules successfully spanning the 244 kb informative segments 153
Figure 49 A three coloured FISH was designed to verify the true configuration of the case 10 SV 154
Figure 50 An image from the FISH analysis of the control sample 155
Figure 51 Four example FISH images from the patient cell line 156
Figure 52 DeepC prediction suggests an expansion of the FGF9 TAD due to the SV in Alt 2 compared to the reference 158
Figure 53 Dideoxy- sequencing analysis of the de novo single nucleotide DEL affecting FOXP2 in case 10 159
Figure 54 Bulk RNA-seq data visualised using volcano and box plot 161
Figure 55 Gene ontology enrichment analysis was carried out against the PANTHER pathway database with Fisher's Exact test 164
Figure 56 Case 2 proband has syndromic multi-suture synostosis with hypertrichosis and gum hypertrophy 168
Figure 57 Two CNV gains were confirmed from WGS data 170
Figure 58 The complex nature of case 2 SV was revealed by WGS and verified by breakpoint PCR and dideoxy- sequencing 172
Figure 59 For case 2, three alternative hypotheses can be constructed to explain the WGS data 174
Figure 60 RVA analysis extracted 5 molecules spanning the 17DUP region in the case 2 proband 176
Figure 61 Manual realignment of each of the five supporting molecules detected by RVA 176
Figure 62 FISH was carried out by GOSH diagnostic lab 177
Figure 63 The reciprocal FISH was designed and carried out by GOSH as the initial FISH in Figure 62 cannot fully characterise all three Alts 178
Figure 64 Several cases in the literature have implicated SV/CNVs at the KCNJ locus with the hypertrichosis and gingival hypertrophy phenotype 180
Figure 65 DeepC prediction on the case 2 SV Alt 3 configuration 183
Figure 66 Seemingly balanced translocations may contain complex structures at the break junctions 186
Figure 67 Proposed mechanism of how the case 2 SV might have arisen 187
Figure 68 The case 16 trio was recruited to the 100kGP due to syndromic multisuture CRS 191
Figure 69 WGS analysis revealed the complex nature of the case 16 SV 193
Figure 70 Breakpoint PCR and dideoxy- sequencing verified the four breaks detected by WGS in case 16 194
Figure 71 A total of 12 possible alternative configurations can explain the WGS data of the case 16 SV 196

Figure 72 Bionano OGM data from the unaffected mother in case 16 198
Figure 73 Bionano OGM data for the case 16 proband 200
Figure 74 The case 16 maternal grandmother was recruited later to investigate the origin of the CPX SV using PCR and Bionano OGM 202
Figure 75 DeepC predictions for the case 16 SV 205
Figure 76 One hypothesis can potentially explain the pathogenic mechanism in the case 16 family 206
Figure 77 A total of 35 individuals from 13 CRS families were analysed on the ONT PromethION as part of the 100kGP pilot program 217
Figure 78 Example of an INS call made by the Bionano Access software.....221
Figure 79 Confusion matrix and metrics for the INS and DUP events called by the Bionano Access pipeline 222
Figure 80 Performance evaluation for Illumina and ONT compared to the truth set of the normalised Bionano SV calls.....223
Figure 81 SV detection evaluation for ONT and Illumina compared to Bionano OGM SV calls, stratified into the three detected SV types.....225
Figure 82 Example INS captured by ONT but not detected by Illumina WGS 226
Figure 83 De novo calling performance improved in batch 3 using Sniffles2 compared to the first two batches using Sniffles 227
Figure 84 Nabsys EGM reads the electrical signal caused by tags on DNA molecules to generate a consensus map to compare to the reference map.....243

List of Tables

Table 1 Common NGS and long-range technologies for SV detection 12
Table 2 DGV to Manta/Canvas call conversions 36
Table 3 SV-coding region intersection criteria 38
Table 4 SV segregation, filtering, and prioritisation strategies 41
Table 5 Mosaicism detection method based on the $Rmos$ value 44
Table 6 List of regions screened in the 100kGP rare disease cohort 44
Table 7 SV segregation, filtering, and prioritisation strategies for the SVRare analysis 46
Table 8 Primers used for breakpoint PCR including the expected product length 48
Table 9 Standard PCR setup using Taq DNA polymerase 49
Table 10 PCR program with Taq DNA polymerase 49
Table 11 LongAmp PCR setup using LongAmp Taq DNA polymerase 49
Table 12 PCR program with LongAmp Taq DNA polymerase 50
Table 13 BAC clones for FISH of case 10 52
Table 14 Run metrics from Bionano recommendations to determine data quality 65
Table 15 Bionano OGM benchmarking error matrix 67
Table 16 ONT/Illumina WGS sensitivity matrix against TP Bionano OGM calls 68
Table 17 ONT de novo calling performance evaluation 69
Table 18 100kGP cases where a candidate SV was identified through WGS analysis 72
Table 19 Genes directly affected by case 18 complex SV 126
Table 20 Summary of the 20 cases analysed using Bionano OGM 133
Table 21 Major differences between the three batches in ONT analysis 216
Table 22 ONT data quality for all 35 individuals, including coverage and N50 218
Table 23 OGM SV types were normalised into DEL, tandem DUP, and INS. 219
Table 24 five coding, potentially de novo candidates from the ONT-Sniffles2 analysis 228
Table 25 Seven likely/pathogenic SVs were identified in the CRS cohort of 144 cases 238

List of Abbreviations

Abbreviation	Meaning
100kGP	100,000 Genomes Project
3C	Chromosome conformation capture
3GS	Third generation sequencing
aCGH	Microarray-based comparative genomic hybridisation
ADHD	Attention deficit hyperactivity disorder
AMC	arthrogryposis multiplex congenita
ARCND2	Auriculocondylar syndrome 2
ASD	Autism spectrum disorder
ATS	Andersen-Tawil syndrome
BAC	Baterial artificial chromosomes
BAM	Binary alignment map
BIR	Break-induced replication
BND	Break end (variant/call/read)
BR	Broad range
BSA	Bovine serum albumin
CCB	Centre for Computational Biology
CFS	Common fragile site
CGH	Comparative genomic hybridisation
CHD	Conginetal heart disease
CI	Confidence interval
CML	Chronic myeloid leukaemia
CMT1A	Charcot–Marie–Tooth Type 1A
CN	Copy number
CNV	Copy number variant
CON	Conversion (variant)
COXPD31	Combined oxidative phosphorylation deficiency 31
CPX	Complex SV/events
CRS	Craniosynostosis
CV	Coefficient of variation
DDG2P	Developmental Disorders panel in the Gene2Phenotype database
DECIPHER	DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources
DEL	Deletion
DGAP	Developmental Genome Anatomy Project
DGS	DiGeorge syndrome
DGV	Database of Genomic Variants
DIG	Digoxigenin

DLS	Direct Label and Stain (Bionano)
DLS DNA	Labelled and stained DNA
DSB	Double-stranded break
DTT	Dithiothreitol
DUP	Tandem duplication
EBV	Epstein–Barr virus
EGM	Electronic genome mapping
FDR	False discovery rate
FGF	Fibroblast growth factor
FGFR	Fibroblast growth factor receptor
FISH	Fluorescence in situ hybridisation
FITC	Fluorescein isothiocyanate
FN	False negative
FoSTeS	Fork stalling and template switching
FP	False positive
GACI	Generalised arterial calcification of infancy
GBoCM	Genetic Basis of Craniofacial Malformation
GE	Genomics England
GeCIP	Genomics England Clinical Interpretation Partnership
gnomAD	Genome Aggregation Database
GOSH	Great Ormond Street Hospital
GTEx	Genotype-Tissue Expression
HMW	High molecular weight
HPO	Human Phenotype Ontology
HR	Homologous recombination
HS	High sensitive
HTC	Generalised hypertrichosis
ID	Intellectual disability
IGV	Integrative Genomics Viewer
INS	Insertion
INV	Inversion
iPSC	Induced pluripotent stem cells
IQ	Intelligence quotient
LB	Luria-Bertani
LCR	Low copy repeat
LOEUF	Loss-of-function observed/expected upper bound fraction
MANE	Matched Annotation from National Center for Biotechnology Information and European Bioinformatics Institute
MAPH	Multiplex amplification and probe hybridisation
MEPS	Minimal efficient processing segment
MHU	Molecular Haematology Unit

MLPA	Multiplex ligation-dependent probe amplification
MMBIR	Microhomology-mediated break-induced replication
MMEJ	Microhomology-mediated end joining
MRC	Medical Research Council
MSC	Mesenchymal cells
MT	Mitochondria/mitochondrially-encoded
NAHR	Non-allelic homologous recombination
NEB	New England Biolabs
NGM	Next generation mapping
NHEJ	Non-homologous end joining
NHS	National Health Service
NLV	Negative label variance
OGM	Optical Genome Mapping
OMIM	Online Mendelian Inheritance in Man
ONT	Oxford Nanopore Technologies
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PDA	Patent ductus arteriosus
PFO	Patent foramen ovale
pHaplo	Predicted probability of haploinsufficiency
pLI	Probability of loss-of-function intolerance
PLV	Positive label variance
PMSF	Phenylmethylsulfonyl fluoride
pTriplo	Predicted probability of triplosensitivity
QC	Quality control
qPCR	Quantitative polymerase chain reaction
RBC	Red blood cell
RCF	Relative centrifugal force
REC	Research ethics committee
RNA-seq	RNA sequencing
RT	Room temperature
RT-qPCR	Quantitative reverse transcription polymerase chain reaction
RVA	Rare variant analysis
SDS	Sodium dodecyl sulphate
SGB	Simpson-Golabi-Behmel syndrome
sHET	Selection coefficient of heterozygous loss-of-function variants
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SSC	Saline-sodium citrate buffer
SSCT	Saline-sodium citrate buffer with Tween
SV	Structural variant

SVUS	Structural Variant of Unknown Significance
SYNS3	Multiple synostoses syndrome 3
T2T	Telomere-to-Telomere
TAD	Topologically associated domain
TN	True negative
TP	True positive
TRA	Translocation
UCSC	University of California, Santa Cruz
VCF	Variant call format
VSD	Ventricular septal defect
VUS	Variant of uncertain significance
WBC	White blood cell
WCHG	Wellcome Center of Human Genetics
WES	Whole exome sequencing
WGS	Whole genome sequencing
WIMM	Weatherall Institute of Molecular Medicine
WSI	Wellcome Sanger Institute
ZLS	Zimmerman-Laband syndrome

Abstract

Background: Structural variants (SVs) are large genomic alterations that can span from 50, to millions of base pairs, with potential clinical implications depending on their location and consequence. Characterisation and interpretation of SVs through conventional methods remain challenging due to both experimental and computational limitations. In contrast, emerging long-range technologies, such as Bionano Optical Genome Mapping (OGM) and Oxford Nanopore Technology (ONT), offer promise in significantly improving SV detection and interpretation.

Craniosynostosis (CRS) is characterised by the premature fusion of one or more cranial sutures, affecting 1 in 2000 births. Genetically, CRS is a highly heterogeneous condition, with to date 66 associated “Green” genes in the Genomics England PanelApp. This genetic complexity makes CRS an ideal candidate for genetic studies, and yet many CRS cases still lack a clear underlying molecular cause following appropriate targeted investigations.

Aim: My project set out to identify clinically relevant SVs in patients with CRS lacking a genetic diagnosis, with the hypothesis that elusive pathogenic SVs may contribute to the diagnostic gap in CRS.

Method and data: Three genomic approaches were employed to investigate clinically relevant SVs:

1. 114 CRS families were analysed by Illumina WGS as part of the 100,000 Genomes Project (100kGP);
2. I employed Bionano OGM to investigate 20 families to both further characterise complex SV candidates and to identify pathogenic SVs that may have been overlooked by conventional technologies;
3. 8 trios were analysed by ONT as part of a 100kGP pilot sequencing programme.

Results: The 100kGP analysis identified several candidate SVs, with the most compelling SVs affecting the *HOXC* cluster – a group of genes previously unexplored in relation to CRS. Bionano OGM successfully identified and characterised several SVs, including a large elusive inversion affecting *TWIST1*, an ambiguous paired duplication on chr13, a chromothripsis-like event spanning over a ~3 Mb region on chr20, and an unbalanced reciprocal translocation affecting a novel disease locus near *KCNJ2*. In contrast, ONT data faced initial challenges in SV calling, while subsequent optimisation efforts yielded significant improvements, demonstrating ONT's unique advantages in characterising insertions.

Conclusion: Investigating clinically relevant SVs in CRS patients revealed compelling SVs that could explain some of the diagnostic gaps. Bionano OGM demonstrated excellent clinical utility in SV analysis, while stringent sample requirements and cost are major challenges for routine diagnostic implementation. ONT showed promise in SV analysis, but computational calling of variants needs further improvement for wider application.

Chapter 1 Introduction

1.1 Context of this thesis

The past 60 years have witnessed extraordinary progress in our ability to diagnose the causes of constitutional genetic disease in individual patients. Amongst the notable milestones, crucial discoveries include the identification of trisomy 21 as the cause of Down syndrome (1959)[1,2], the first demonstration of the cause of a genetic disease using analysis of DNA with α-globin gene deletion in Haemoglobin Bart's hydrops fetalis (1974)[3], and the draft sequence of the human genome (2001).[4] Recent years have seen a pivotal shift for genetic diseases diagnosis, highlighted by a landmark publication in 2010. This work introduced the first use of exome sequencing to identify the basis of a previously unsolved genetic disease, the craniofacial disorder Miller syndrome.[5] This breakthrough laid the groundwork for the past decade's efforts in genetic research.

Subsequently, genomic technologies have evolved to a degree that, at the start of my thesis research in 2019, whole genome sequencing (WGS) was being used to diagnose rare diseases in individual patients, for example in the Genomics England (GE)/National Health Service (NHS) 100,000 Genomes Project (100kGP). However, much of the initial focus on the use of short-read (typically 150 bp) paired-end WGS, primarily aiming to improve the diagnosis of small variants, such as single nucleotide polymorphisms (SNPs) or small insertion-deletion. Although WGS in theory holds the potential to identify larger structural variants (SVs), the practical success and the barriers to achieving this were largely uncharted.

Therefore, I set out to address this gap, focusing on the identification of SVs in a specific disorder, craniosynostosis (CRS), the premature fusion of cranial suture(s). In this Introduction, I will first define the landscape of SVs, review the challenges associated with their identification, and outline why CRS represents a valid patient cohort to investigate the optimal detection methods for SVs and their potential contribution to undiagnosed genetic diseases.

1.2 Structural variants (SVs)

Originally, the term structural variants (SVs) referred to a subset of large DNA alterations known as sub-microscopic SVs ranging from 1 kb - 3 Mb.[6] With advances in sequencing/microscopic technologies and computational power, SVs now more commonly refer to any DNA alterations larger than 50 bp.[7] SVs are broadly categorised into several subclasses, including deletions (DELs), tandem duplications/duplications (DUPs), insertion (INSs), inversions (INVs), translocations (TRAs), and complex (CPX) events, as illustrated in **Figure 1**. Copy number variant (CNV) is an umbrella term describing several types of SVs. For example, DEL events often manifest as CNV loss, while DUPs and INSs represent the majority of the CNV gain events. On the other hand, INVs and translocations are typically copy neutral. CPX SVs are single events with multiple breakpoints that cannot be classified under any other subclass of SVs.

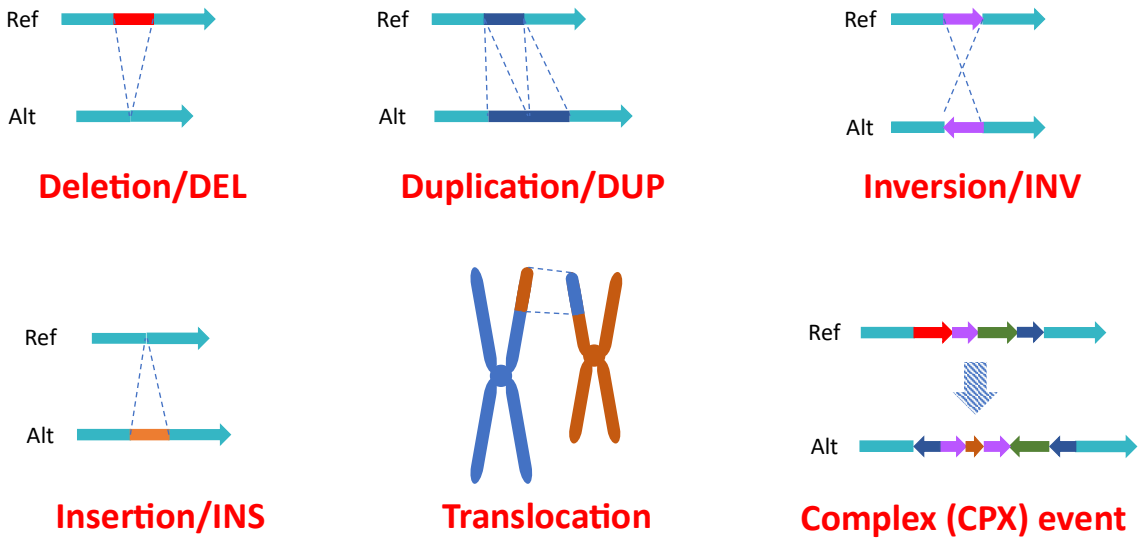


***Figure 1 SVs can be categorised into several subclasses**. Ref represents the reference allele, while Alt illustrates the allele with a SV.*

1.3 SV mechanisms

SVs can arise from various DNA repair mechanisms, such as non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), microhomology mediated end joining (MMEJ), as well as DNA replication-based mechanisms, including microhomology-mediated break-induced replication (MMBIR) and fork stalling and template switching (FoSTeS).[8] In human constitutional genetic conditions, the process of meiosis is the most relevant stage for SV generation. Meiosis involves one round of DNA replication, followed by homologous recombination (HR), and two rounds of cell division. Notably, meiosis deliberately creates double-stranded DNA breaks (DSBs) to enable the subsequent HR.[9] These programmed DSBs serve as the precursor for SV generation via break-induce repair mechanisms. Alternatively, SV may arise from the initial round of DNA replication via replicative based mechanisms. These processes are crucial to understand SV mechanisms in human genetic conditions.

HR, NHEJ and MMEJ are cellular mechanisms responsible for repairing DSBs. NAHR is mediated by highly homologous sequences, often common repetitive elements such as low copy repeats (LCRs), Alu elements, and pseudogenes. These homologous sequences involved in NAHR must also reach a certain length, referred to as minimal efficient processing segments (MEPS), estimated to be between 300-500 bp[10]. An example of an NAHR mechanism resulting in an SV is illustrated in **Figure 2a**, where the misaligned homologous sequences led to a DEL during the repair of a DSB via NAHR. Clinically, a well-characterised NARH-mediated human disease is DiGeorge/velocardiofacial syndrome (DGS). The DGS critical region at 22q11.2 is flanked by four highly homologous LCRs with a sequence similarity exceeding 95%.[11] These LCRs act as substrates for NAHR, giving rise to various sizes of recurrent pathogenic DELs at the DGS locus.

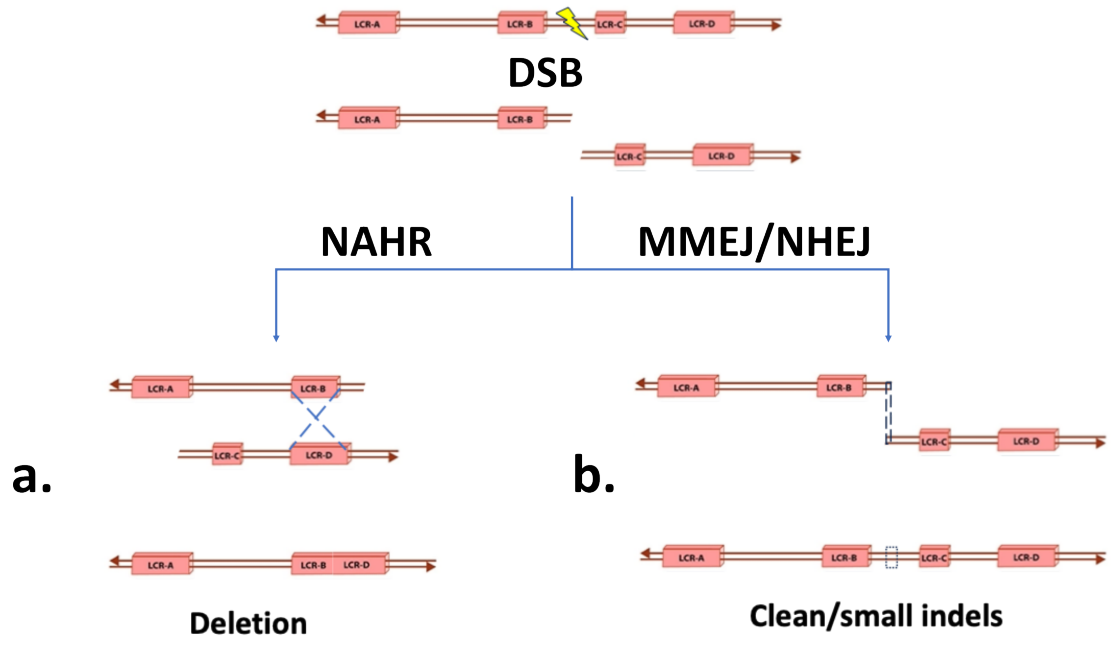


Figure 2 DSBs can be repaired via NAHR and MMEJ/NHEJ. a. *In an attempt to repair the DSB, misaligned NAHR occurs due to highly homologous LCR elements, LCR-B and LCR-D, resulting in a DEL of the LCR-C locus.* ***b.*** *Alternatively, DSB may be repaired via an end joining mechanism with or without presence of microhomology. As end joining is inherently an imprecise mechanism, the process often leaves small indels at the break junction. Figure adapted from Burssed et al (2022).*[8]

NEHJ, another DSB repair mechanism, is characterised by two distinct features. Firstly, unlike NAHR, NHEJ break junctions have little to no homology, with a maximum of 1-4 bp microhomology. Secondly, NHEJ may leave a small indel of random nucleotides at the break as an “information scar” due to the DSB repair process.[12] MMEJ is a similar mechanism to NHEJ, but it uses a slightly longer stretch of homologous nucleotides from 5-25 bp. In human disease, end-joining mechanisms are seen in many types of cancers. For example, in chronic myeloid leukaemia (CML), DSBs occur in both chr9 and chr22 due to various factors.[13] The broken ends of chr9 and chr22 are then repaired and joined to generate a frequently observed 9-22 translocation, known as the Philadelphia chromosome, likely mediated by NHEJ/MMEJ, as evidenced by the microhomology at the break junctions[14].

Replicative mechanisms have also been suggested as models of generating SVs, particularly complex SVs in the human genome. During DNA replication, the replication fork may collapse or stall due to various reasons, such as endogenous DNA lesions and repetitive barriers. Via FoSTeS, as shown in **Figure 3**, the stalled fork may switch and jump from one replication template to another, causing a DEL when switched to a downstream template (forward invasion), or a DUP when switched to an upstream template (backward invasion). Translocations can be formed when FoSTeS involves invasion of templates on different chromosomes.

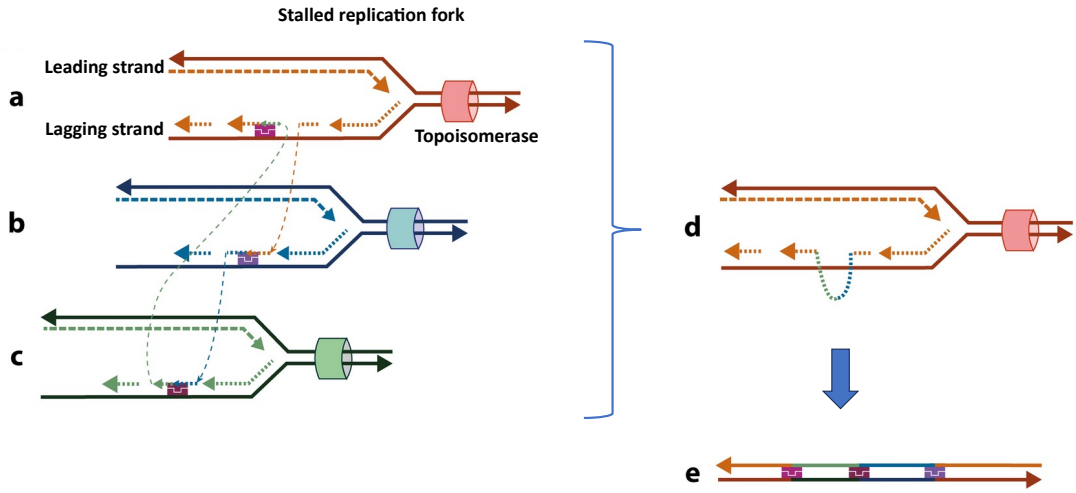


Figure 3 The fork stalling and template switching (FoSTeS) mechanism. *In the event of a stalled replication fork* ***a****, the lagging strand may disengage and invade other replication forks, such as fork* ***b*** *and* ***c****, resulting in the creation of a multi-break junction SV in* ***d****. The end product* ***e*** *contains genetic material from all three loci in forks* ***a****,* ***b****, and* ***c****. Figure adapted from Burssed et al (2022).*[8]

A stalled or collapsed replication fork can further be converted into single-ended DSBs when the fork passes through a nick in the DNA or due to endonuclease activity (**Figure 4a**).[15] Consequently, two repair mechanisms can follow, break-induced replication (BIR) via NAHR requiring a long stretch of homology, or MMBIR, as illustrated in **Figure 4**. With the FoSTeS/MMBIR mechanisms, long-range complex SVs can be generated in the human genome. Some of the highly complex, chromothripsis or chromothripsis-like SVs partially mediated via FoSTeS/MMBIR, are often seen in cancer such as melanoma[16].

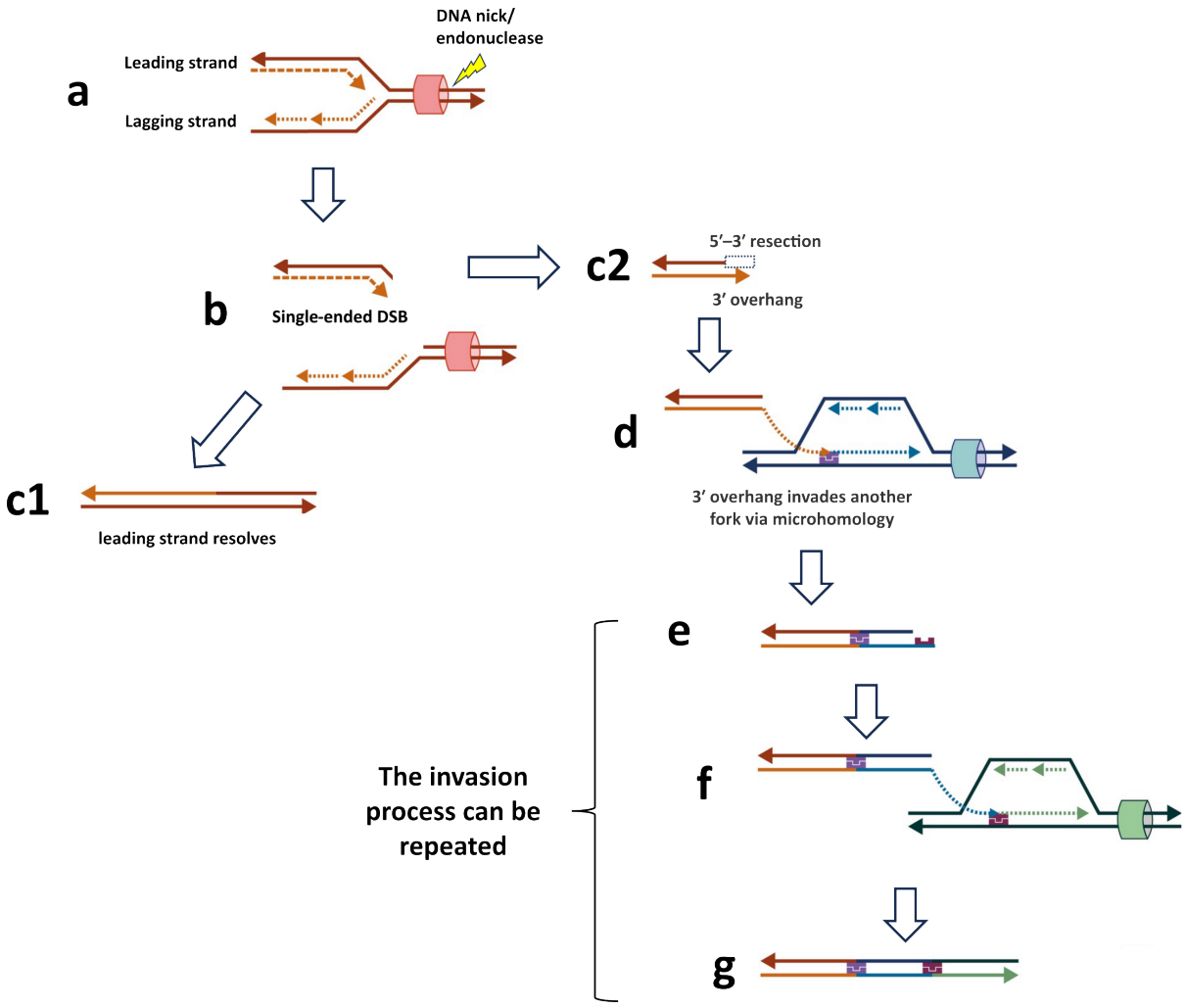


Figure 4 Microhomology mediated break-induced replication (MMBIR). *Single-ended DSBs **b** are created when the replication fork passes through a nick in the DNA or when the fork encounters endonuclease activity shown in **a**. The fork collapses into **c1**, which resolves using the unaffected lagging strand template, and **c2**, where 5'-3' recession occurs, leaving a 3' overhang. The 3' overhang invades another fork via microhomology (**d**). Due to the low processivity of the DNA polymerase during the initial stage of the replication, this invasion process may repeat (**e, f,** and **g**), until DNA polymerase becomes more processive until the end of the chromosome. Figure adapted from Burssed et al (2022).*[8]

1.4 SV detection approaches

Significant effort has been directed towards identifying clinically relevant SVs using diverse methods based on their clinical utility. As summarised in **Figure 5**, SV detection technologies have progressed from physical genome mapping to array/probe based, and finally to sequencing based approaches. This evolution has

seen an increase in the resolution of technologies, allowing detection of a greater variety of SVs/CNVs by delving into individual bases of the DNA molecules. Recent developments in long-range technologies have shifted the primary focus of SV detection to overcoming the limitations of DNA fragmentation in short-read technologies, aiming to achieve a more comprehensive view of complex genomic structures. Nevertheless, each of these technologies has its unique advantages and applications in SV detection in clinical settings.

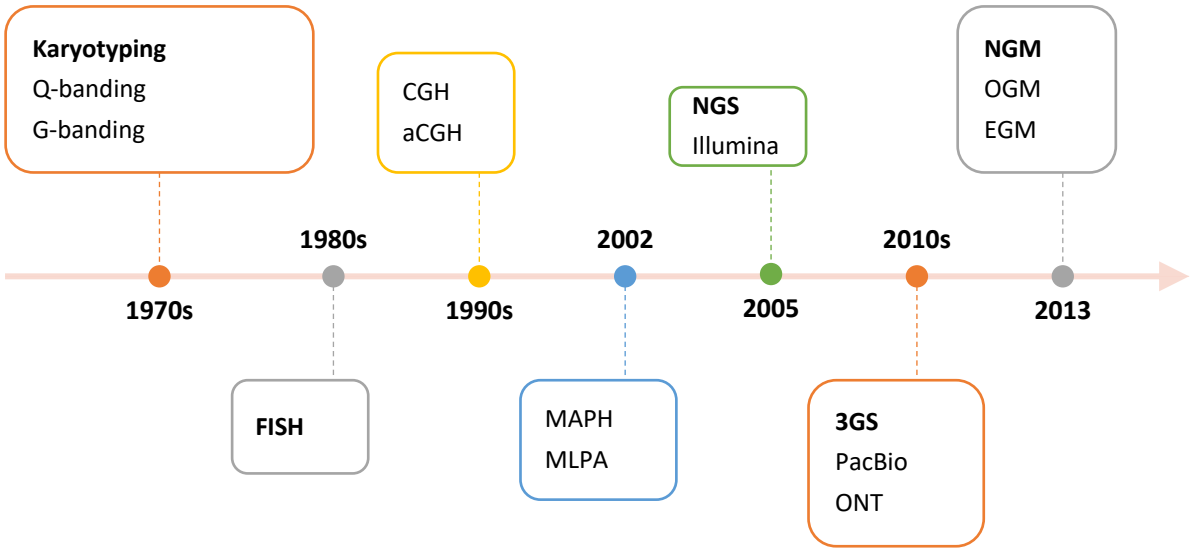


Figure 5 Evolution of SV detection technologies. *The timeline here includes Q/G-banding[17,18], fluorescence in situ hybridisation (FISH)[19], Comparative Genomic Hybridisation (CGH)[20]/Microarray-based CGH (aCGH)[21], multiplex amplification and probe hybridisation (MAPH)[22], multiplex ligation-dependent probe amplification (MLPA)[23], next generation sequencing (NGS)[24], third generation sequencing (3GS)[25], and next generation mapping (NGM)[26]. ONT = Oxford Nanopore Technologies; EGM = Electronic Genome Mapping. Data extracted from Pös et al (2021)[27] and Chen et al (2023)[28].*

Array-based approaches, such as microarray-based comparative genomic hybridisation (aCGH) and SNP arrays, are common methods in routine clinical CNV detection. aCGH compares fluorescently labelled test DNA against a reference DNA through a competitive hybridisation process, with DNA probes ranging from 25 bp –

300 kbp in size derived from oligonucleotides to bacterial artificial chromosomes (BACs), immobilised on a microarray. The difference in emission intensity can therefore determine the copy number (CN) changes in the test DNA compared to the reference DNA.[29] SNP arrays, in comparison, allow more accurate and higher-resolution CNV detection and can have over 900,000 allele-specific probes.[30] This quantitative nature of SNP array further enables the detection of mosaic CNVs.[31] These array approaches are fast and cost-efficient methods for SV detection, particularly in the clinical setting where rapid diagnosis is imperative. However, a major drawback of the array approaches is that they cannot detect balanced events such as INVs and novel insertions (sequences not in the reference genome), as these events require breakpoint information in addition to the CN change to be fully characterised. Furthermore, despite recent improvement, the resolution of array technologies remains at a sizeable level. For example, until recently, clinical aCGH was often recommended to report only CNVs ≥400 kb (>200 kb in the Oxford clinical lab) for the optimal balance of clinical precision and sensitivity.[32] This level of resolution is efficient at detecting large events, but SVs under the detection limit are overlooked completely.

For balanced events, karyotyping and fluorescence in situ hybridisation (FISH) are the other common cytogenetic methods for diagnostic SV detection. Karyotypes are created by staining chromosomes of the cells arrested in the metaphase or prometaphase to generate a recognisable banding pattern, where abnormal banding patterns can indicate the presence of chromosomal abnormalities. Unlike array-based techniques, karyotyping can detect balanced events like translocations and INVs. Additionally, karyotyping can discriminate CN gain events, such as DUPs, inverted DUPs, and duplicated INS, which are indistinguishable using array-based approaches.

The major drawbacks of karyotyping are its significantly lower resolution, typically between 5 – 10 Mb, depending on the chromosome banding resolution, and the requirement of a highly skilled workforce to interpret the banding patterns. FISH, which uses fluorescently labelled probes, offers much better resolution compared to G – banding karyotyping, especially when using multi-coloured FISH. However, FISH is a targeted approach, as the probes must be designed specifically to a target and cannot be used for novel diagnostic discoveries.

In contrast, next generation sequencing (NGS) technologies, such as Illumina WGS, offers base-pair resolution for most regions of the human genome. When combined with dideoxy- ("Sanger") sequencing, NGS enables accurate detection of SV breakpoints, hence allowing for the identification of more complex SVs compared to simple CNVs from array-based methods. However, for SV discovery NGS requires *post hoc* computational processing, ie variant calling. This has proven to be a highly complex and heterogeneous process, yielding diverse sets of SVs depending on the algorithm used by the SV callers. To date, a myriad of SV callers have emerged for NGS data, each with its own advantages. Fast and simple callers rely on changes in read depth to call CNVs, such as CNVnator[33] and Canvas[34]. Combining read depth with split/pair read information, callers such as Lumpy[35] can detect more complex SVs than simple CNVs. Other callers, such as MantaSV[36], identify candidate SVs by firstly using read information, and then performing re-alignment/assembly of abnormal reads locally to validate called SV candidates.[36]

A major limitation of NGS approaches for SV discovery is the need for *post hoc* reassembling of the genome using short-read data derived from fragmented DNA molecules. Prior to SV calling, the short reads (~300 bp paired reads for Illumina WGS, **Table 1**) from NGS must be assembled and aligned to the reference genome. This process is particularly challenging, if not impossible, when dealing with large INSs and certain types of CPX SVs. For INSs larger than the read length, reads fully flanked by the INS cannot be mapped accurately and/or uniquely to the correct loci. Similarly, for certain CPX SVs, short reads cannot span the additional breakpoints all at once, which is necessary to fully resolve the structure of some complex rearrangements. This limitation of short-read NGS may have significant research and clinical implications, which will be thoroughly discussed in detail for specific cases in later chapters.

Table 1 Common NGS and long-range technologies for SV detection

	Illumina	**PacBio**	**ONT**	**OGM**
N50 (reflective of read length)	< 250 bp	30-60 kb	10-60 kb	~300 kb
Maximum read/molecule lengths	~ 300 bp (paired)	> 200 kb	~ 4 Mb	> 2 Mb
Resolution	1 bp	1 bp	1 bp	500 bp

Resolution indicates the minimum detectable size of variants. The Illumina data are based on NovaSeq 6000, which was used for the 100kGP; PacBio is based on Sequel II; ONT is based on PromethION, which is used in the 100kGP long-read pilot program. Illumina, PacBio, and ONT metrics are obtained from Verges et al (2020).[37] OGM N50 is extracted from Bionano "Data Collection Guidelines, Document Number: 30173, Revision: E"; resolution is extracted from the Bionano OGM official website (https://bionano.com/saphyr-systems/, last accessed 26.10.2023); the maximum molecule length is estimated based on the observed molecule length in local data. Max unrestricted read length for ONT has been reported ~ 4 Mb (https://nanoporetech.com/sites/default/files/s3/white-papers/human-genetics-research-white-paper.pdf, last accessed 22.02.2024).

Over the last decade, long-read technologies, commonly referred to as third-generation sequencing (3GS) technologies, have been developed specifically to address the short-read challenge. Traditional NGS methods typically involve fragmenting the DNA molecules to a certain length to achieve the required high throughput. Long-range technologies, in contrast, tend to preserve the long DNA molecule to achieve a longer read length (**Table 1**). Common genome-wide long-read approaches include PacBio and Oxford Nanopore Technologies (ONT). Based on conventional sequencing-by-synthesis approaches, PacBio achieves longer reads by circularising long fragments of DNA molecules.[38] In contrast, instead of sequencing-by-synthesis, ONT passes long DNA molecules through a protein nanopore, measuring the detectable current density change caused by different DNA bases. This way, ONT can produce read lengths of 10-30 kbp, depending on the required throughput[39].

More recently, orthogonal next-generation genome mapping (NGM) technologies have provided a new perspective on genome wide SV detection. Without the amplification/fragmentation challenges and the need for bp level accuracy, NGM technologies examine ultra-long, minimally fragmented DNA molecules (**Table 1**) in nanochannels using sequence-specific tags and labels. Bionano OGM visualises the structure of long DNA molecules with added fluorescent labels[40], offering the possibility of examining the most intact molecules. NGM technologies likely offer the most unbiased and comprehensive SV detection approach without sacrificing the overall throughput. Nevertheless, complex library preparation, lower throughput, and higher run cost are likely major challenges for OGM. EGM technology is further discussed in **section 7.3**.

The continual evolution of genomic technologies holds the promise of better clinical SV detection approaches. However, the choice of the appropriate detection methods remains highly dependent on the relevant clinical context, balancing factors such as resolution, throughput, read length, and cost efficiency. Overall, in combination, these genomic technologies provide a crucial platform to understand the role of SVs in human diseases and, more broadly, human populations in general.

1.5 SVs in the human population

Numerous endeavours have been made to characterise common SVs in the general population. One of the largest aggregation of SVs is the Database of Genomic Variants (DGV), which is a curated collection of human SVs from (originally) 55 published studies, containing over 2.5 million entries from more than 22,300 genomes.[41] The latest version, DGV v107, now contains 75 published studies, with 18,366,594 entries from 54,980 genomes (http://dgv.tcag.ca/dgv/app/, last accessed 28.10.2023). This latest version collated more than double the original number of genomes, and included highly quality short-read studies, such as the Genome Aggregation Database (gnomAD) SV[42], as well as some long-read studies, such as the project described in Audano et al (2019).[43]

DGV's repository predominantly consists of genome data from healthy, clinically unaffected individuals, serving as an excellent resource for filtering against likely clinically benign SVs in the general population. In the context of clinical CNV interpretation, the recommended guidelines typically suggest disregarding CNVs

found in unaffected individuals, such as from DGV. However, there are several considerations. One is that many DGV entries are from technologies with low or imprecise resolution, such as BAC aCGH and targeted MLPA (http://dgv.tcag.ca/dgv/app/search?ref=#tabs-view_all_info_study, under method section, last accessed 21.11.2023). Combined with a conservative merging strategy, the recorded CNV sizes in DGV may overestimate or underestimate the true events. This discrepancy could potentially lead to the misinterpretation of some pathogenic CNVs as benign.[44] To address these issues, the latest versions of DGV have introduced a "Gold-standard" track, consisting of manually curated, high-quality consensus SVs derived from at least two accomplished studies meeting defined quality criteria. Using this Gold-standard track can reduce false positive pathogenic regions in the control population, although this stringent criterion will also underestimate the frequencies of some benign SVs.

Additionally, aggregating and merging SVs from studies with different methods and resolutions inevitably results in information loss. This is particularly problematic for complex SVs. For example, the gnomAD SV project characterised multiple subclasses of complex SVs based on their alternative allele structures and this information was reduced when gnomAD SV was integrated into DGV. While this step was necessary during the process of SV merging, the resulting loss of structural information may lead to downstream misinterpretation of certain clinical findings. This type of oversimplification underscores the critical need for more precise and comprehensive data integration strategies, especially when dealing with complex SVs in the clinical setting.

In general, SV interpretation against a control population is essential to assess clinical relevance, but one must exercise caution when employing databases such as DGV. The querying strategy must be adapted based on the clinical context, such as the prevalence of the condition in question, the abundance of cases affected by SVs, and likely mode of inheritance of the condition. Opting for specific subsets of DGV can often be beneficial to reflect the clinical context in variant interpretation.

On the other hand, there have also been multiple attempts to characterise SVs in case populations. The DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources (DECIPHER)[45], in the latest version v11.22 (16.08.2023), has aggregated 47,222 open-access patient cases, containing 48,013 CNVs in total. In contrast to DGV, DECIPHER provides a great resource to characterise likely clinically relevant SVs in the case population, although a significant portion of the DECIPHER SVs remain as variants of uncertain significance (VUSs).

Overall, the continuous initiatives to characterise SVs in human populations, such as DGV and DECIPHER, provided crucial resources for variant interpretation in the clinical setting. SVs are known to play a critical role in various human diseases, and detecting and interpreting likely pathogenic SV has been challenging. With the advancement of both SV detection technologies and SV databases, we are now better equipped to understand SV-related human diseases and the underlying molecular mechanisms.

1.6 SVs in human diseases

SVs can cause human diseases generally through two mechanisms: dosage sensitivity or misregulation of critical genes. Dosage sensitivity mechanisms include haploinsufficiency and triplosensitivity. Haploinsufficiency occurs when a single copy of the gene is insufficient to maintain its normal function. This is often the consequence of DEL and CN loss events, from either whole gene DELs, or a partial gene DELs resulting in truncated, non-functioning copies. Occasionally, haploinsufficiency can also arise from INVs, translocations, INSs, and CPX events, when the break points occur within a critical gene. To quantify gene haploinsufficiency, the Probability of Loss-of-function Intolerance (pLI) has been developed as a scale to measure the likelihood of a gene being haploinsufficient. pLI is derived by comparing the observed and expected (given an estimated mutation rate) number of protein truncating variants in a gene in large-scale population data.[46] Genes that exhibit high pLI scores (pLI ≥ 0.9) are considered to be extremely intolerant to loss-of-function mutations, ie extremely haploinsufficent, whereas genes with low pLI scores (pLI ≤ 0.1) are deemed haplosufficient[47]. Another metric, the "loss-of-function observed/expected upper bound fraction (LOEUF)", provides a similarly evaluation to loss-of-function genes but is more affected by gene length compared to pLI.

Triplosensitivity is another mechanism contributing to SV pathogenicity, often the consequence of DUPs and CN gain events. One of the most well-documented conditions caused by triplosensitivity is Charcot–Marie–Tooth Type 1A (CMT1A, OMIM 118220), which has been linked to DUPs of *PMP22.*[48,49] Positioned between two large homologous LCRs, the *PMP22* gene is highly predisposed to CN changes.[50] The precise mechanism by which *PMP22* DUP causes CMT1A remains to be fully

understood, but research has suggested that the DUPs and subsequent over-production of PMP22 overwhelms the protein degradation system, leading to the aggregation of incorrectly processed, non-functioning protein.[51]

Lastly, SVs may cause human diseases without directly altering the CN of critical genes, but by affecting the surrounding regulatory elements, and/or the topologically associated domains (TADs). A short-range local effect to a candidate gene can result from an SV removing or duplicating its regulatory elements, such as an enhancer or promoter. A long-range effect can be produced by altering the 3D structure of the genome and subsequently affecting TADs. As illustrated in **Figure 6**, TADs are large regions of the genome that form distinct 3D structures characterised by frequent intra-domain interactions[52]. In human, the identification of TADs have been achieved through Hi-C technique, which is the high-throughput variation of the pioneer capture technique - chromosome conformation capture (3C).[53] As illustrated in **Figure 7**, 3C is a technique that detects and quantifies the spatial proximity between specific DNA sequences by fixing cells, digesting chromatin, re-ligating crosslinked DNA fragments, and then using polymerase chain reaction (PCR) to determine the frequency of interactions.[54] Hi-C further processes ligated DNA fragments for NGS to provide genome wide view of 3D DNA interactions.

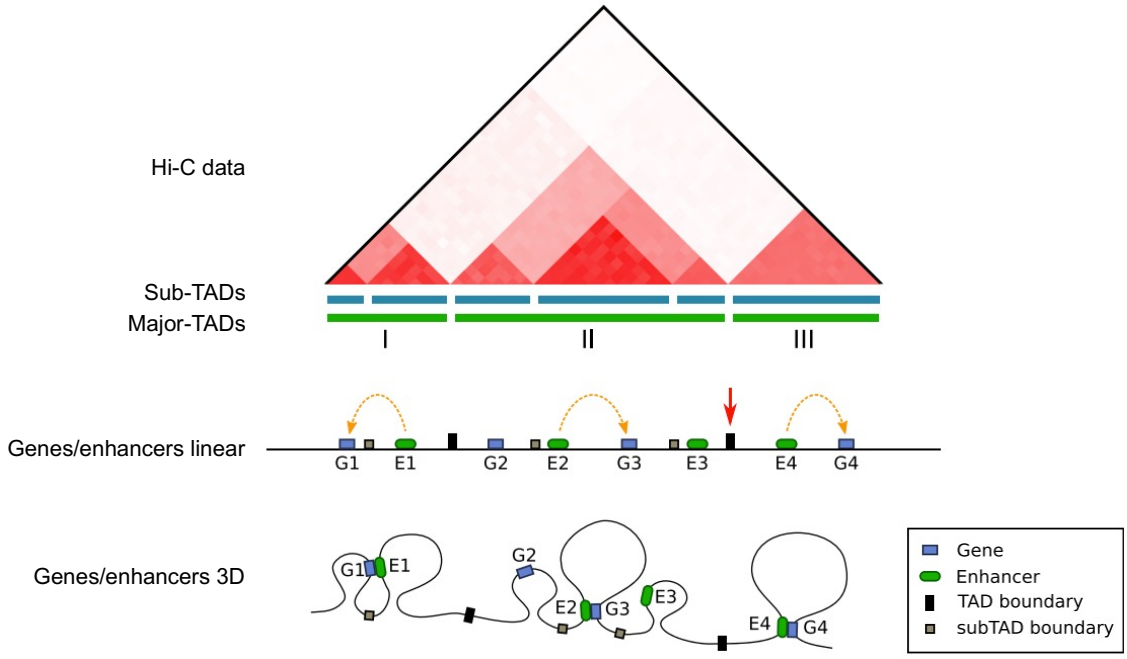


Figure 6 Three-dimensional organisation of chromatids reveals crucial long-range regulatory domains as TADs. *Linear representation of chromosomes cannot appreciate the true spatial distance between genes and enhancers. Using Hi-C technique, the 3D interactions between genomic regions can be characterised and categorised into TADs, separated by TAD boundaries. The true 3D proximity between genes and regulatory elements can therefore be illustrated more clearly. Figure adapted from Szalaj & Plewczynski (2018).*[55]

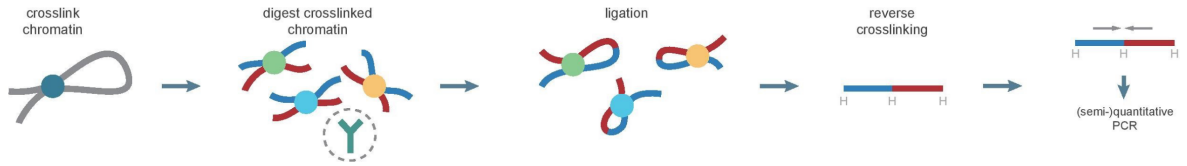


Figure 7 Chromosome conformation capture (3C) technique. *Firstly, cells are treated with formaldehyde to preserve their 3D structures of chromatins. The crosslinked chromatins are digested with restriction enzymes. Next, DNA ligation is performed, favouring the joining of DNA fragments in close spatial proximity. Finally, the crosslinking is reversed, and PCR is used to quantify the frequency of interactions between specific DNA regions. Figure adapted from Li et al (2014).*[56]

TAD boundaries, such as illustrated in **Figure 6**, tend to have significant enrichment of binding sites for the transcription factor CTCFs, which have been proposed to define TAD boundaries.[57] TADs can have long-range effect as they typically span hundreds

of kb to a Mb, with the larger TADs spanning a few Mb.[58] Genes within a TAD may be subjected to a similar regulatory context and are often co-expressed, while genes in a TAD are usually well insulated from the effects of neighbouring TADs. Large SVs can completely remove or duplicate TADs, while smaller SVs at the TAD boundary or critical loci, such as CTCF sites, can cause TAD fusion and/or create neo-TADs.[59] A well characterised example of alterations of TAD function in development is provided by the *SOX9-KCNJ2* locus. SOX9 is a key developmental gene that plays a crucial role in skeletal development.[60] Genetic abnormalities at the SOX9 locus have been associated with a range of human diseases, including Campomelic dysplasia (OMIM 114290), Pierre Robin sequence (OMIM 261800), Cooks syndrome, and Sex Reversal. These disorders (except Sex Reversal) affect skeletal development and can present clinically as craniofacial and limb abnormalities. The heterogeneous phenotype is thought to be caused by disturbance of the complex TADs at the *SOX9-KCNJ2* locus, as illustrated in **Figure 8**. Using mouse models and 3C methods, it was shown that pathogenic *SOX9/KCNJ2* misexpression is a result of specific types of TAD disruption, including inter-TAD DUPs containing *KCNJ2*, intra-TAD DUPs within the *SOX9* TAD, and inter-TAD INVs.[61,62] In contrast, other TAD disturbances at the same loci, such as inter-TAD DUPs without gene involvement or simple TAD fusion by removing CTCF sites, do not cause *SOX9/KCNJ2* misexpression or produce abnormal phenotypes.[61,62]

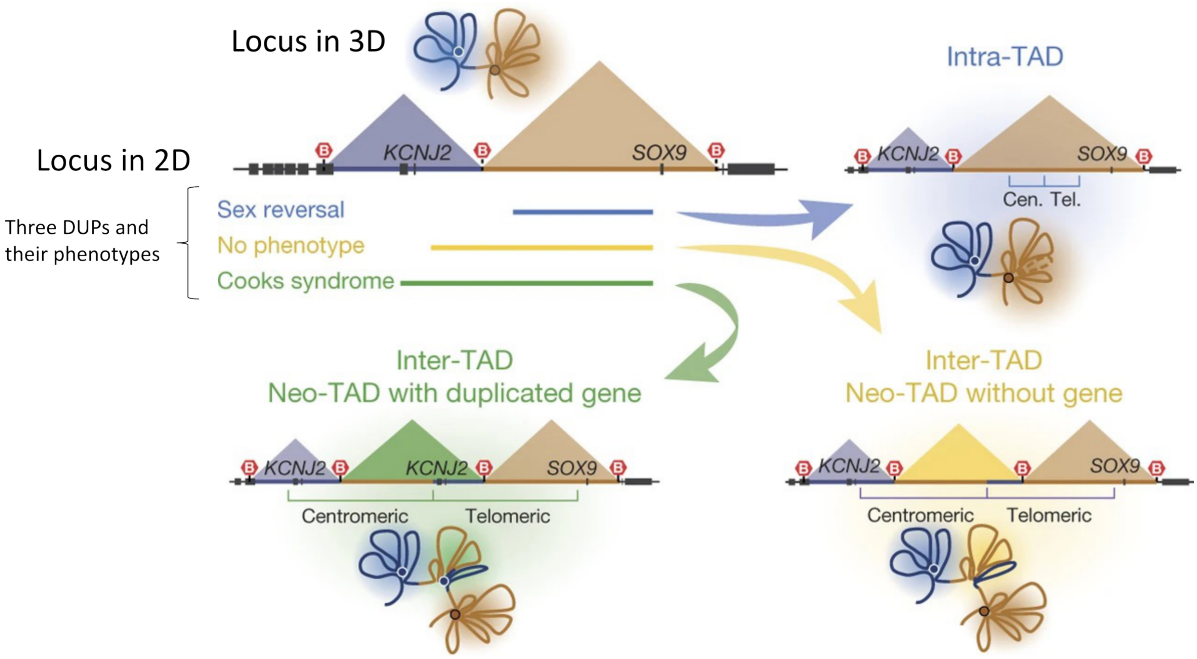


Figure 8 DUPs affecting the SOX9-KCNJ2 locus produce different phenotypes. *Two well-defined TADs can be seen in both the linear/2D and spatial/3D representation of the SOX9-KCNJ2 locus. Three DUPs, depending on the affected genes and TAD boundaries (B with red hexagon), have been implicated in three different phenotypes. Figure adapted from Franke et al (2016).*[61]

1.7 Craniosynostosis and SVs

Craniosynostosis (CRS) is a clinically and genetically heterogeneous condition characterised by the premature fusion of one or more skull sutures, as shown in **Figure 9**, affecting ~1/2000 births[63]. CRS can manifest as either non-syndromic, where suture fusion is an isolated finding, or syndromic with additional clinical features. In syndromic CRS such as Apert syndrome, Crouzon syndrome, Muenke syndrome, Pfeiffer syndrome, and Saethre-Chotzen syndrome, cases can often be diagnosed through specific clinical features related to each syndrome.

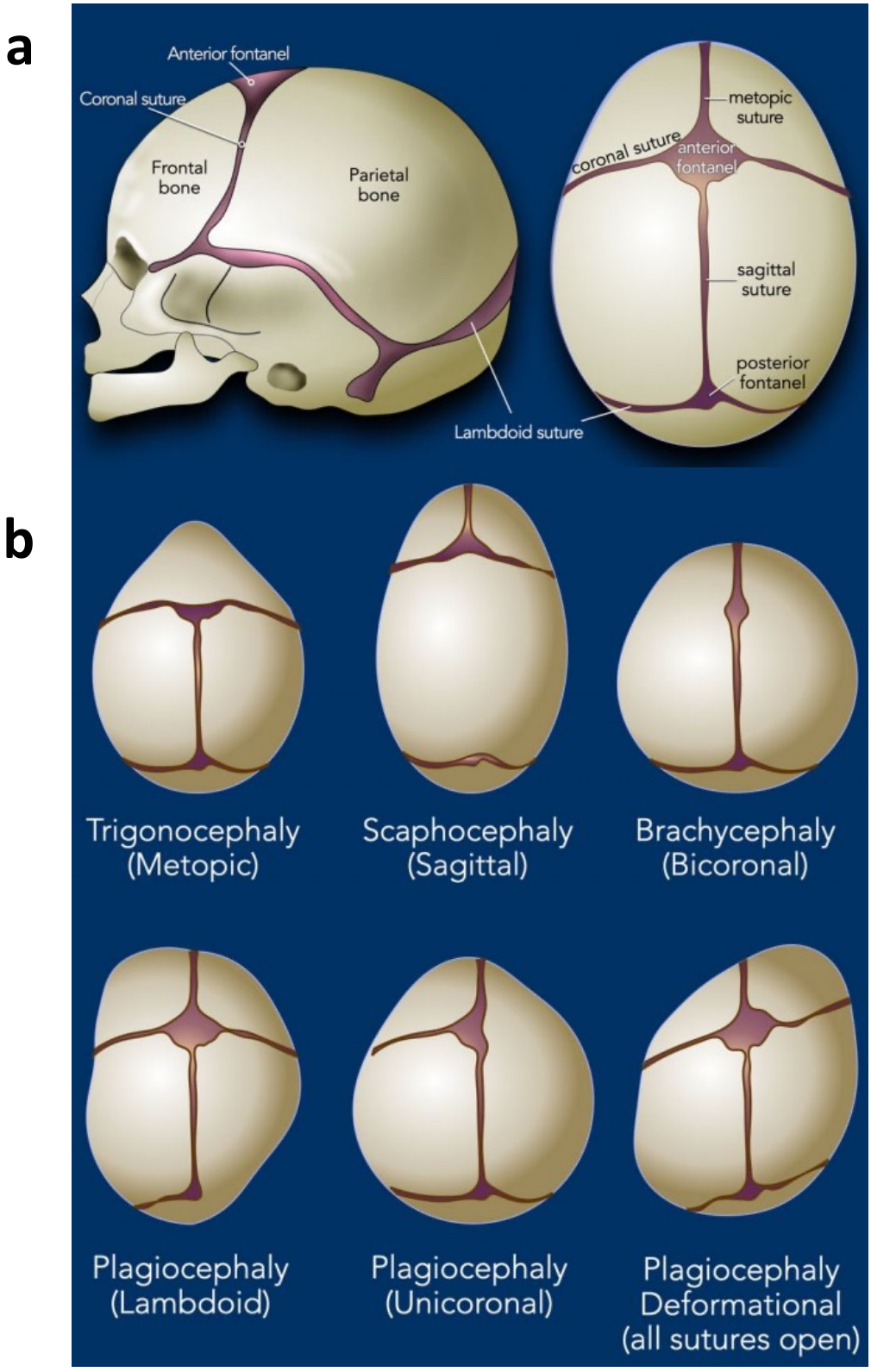


Figure 9 Craniosynostosis (CRS) is the premature fusion of one or more skull sutures. ***a****. Newborn skulls are separated by the metopic, sagittal, coronal, and lambdoid sutures.* ***b****. Premature fusion of these cranial sutures causes CRS and results in abnormal skull development. Figure extracted from radiologyassistant.nl/pediatrics/hip/craniosynosthosis (last accessed 30.10.2023).*[64]

The underlying causes of CRS can be genetic, environmental, or a combination of both. Environmental causes include 1) intrauterine constraint, such as abnormal foetal position and twinning; 2) teratogenic exposure, such as to alcohol[65,66], cigarettes[67-69], and some types of anti-depressant[70,71]. The best documented teratogenic cause of CRS is the anticonvulsant medication sodium valproate.[72] Genetically, the majority of CRS diagnoses can be made by screening seven genes, *EFNB1, ERF, FGFR2, FGFR3, SMAD6, TCF12* and *TWIST1*.[73]

Among the genetic cases of CRS, both loss-of-function and gain-of-function mechanisms are apparent. A loss-of-function mechanism has been attributed to haploinsufficient genes involved in craniofacial development, such as *TWIST1*, *EFNB1*, *TCF12*, and *SOX6*. For example, TWIST1 is an important transcription factor involved in craniofacial development through BMP signalling.[74] Heterozygous loss of *TWIST1* leads to defective skull formation and CRS seen in Saethre-Chotzen syndrome.[75] As well as loss-of-function point mutations, *TWIST1* DELs have been well documented in patients with CRS.[76-79]

Activating or gain-of-function mutations are seen in fibroblast growth factor (FGF) receptor-related CRS syndromes, such as Apert, Muenke, Pfeiffer, and Crouzon syndrome,[80] and represent a major mechanism in CRS. FGF differential expression is crucial for suture development and maintenance.[81] As illustrated in **Figure 10**, in normal suture development, mesenchymal cells (MSCs) near the osteogenic front differentiate into osteoblasts, producing new bones. This process requires relatively high level of FGF through FGF receptor 1 (FGFR1) signalling.[82] In contrast, MSCs in

the centre of the suture requires relatively low level of FGF expression to maintain the suture in a patent undifferentiated state.[83] In FGF-related CRS cases, gain-of-function point mutations in *FGFRs* result in an increase in FGF signalling, either through increased ligand-FGFR binding affinity, or by allowing ligand-independent FGFR activation.[81,84] In terms of SVs, CNV gains affecting the tandemly located *FGF3* and *FGF4* genes have been reported in patients with CRS, via suspected dosage sensitivity causing increased FGF signaling.[85,86]

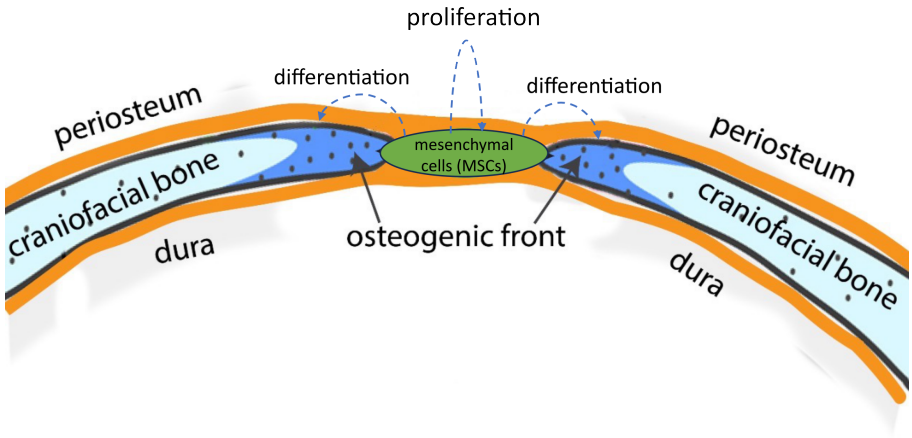


Figure 10 Mesenchymal cells (MSCs) provide essential niche for suture development. *FGF signalling is crucial for normal suture development through suture MSCs proliferation and differentiation. MSCs near the osteogenic front differentiate into osteoblasts while the MSCs in the centre of the suture remain undifferentiated, which keeps the suture unfused. Gain-of-function mutations in FGFRs and CNV gains in FGFs have been described to cause CRS due to abnormal increase in FGF signalling. Figure adapted from Zhao et al (2015).*[87]

In addition to the two well-established molecular mechanisms, a more nonspecific mechanism has been proposed to explain the rare association of many chromosomal abnormalities and large SVs with CRS.[80] For example, a study of 139 CRS patients revealed several nonspecific chromosomal abnormalities without directly affecting any known CRS causative gene.[88] These large variants vary in size, type, position, and associated clinical presentation, which together suggests that CRS is likely a secondary feature in the pathology. Nonspecific chromosomal abnormalities may

cause suboptimal brain development, which perturbs neural crest development and/or the normal growth-inducing strain on the sutures due to abnormal brain growth.[89] Additionally, looking in to the DECIPHER cohort, CNVs have been detected across every chromosome in patients with CRS, as shown in **Figure 11**. Hotspots can be seen, such as the 7p21.1 *TWIST1* locus (a known haploinsufficient gene), while the majority of CNVs recorded remain as VUSs, without any definitive causative gene association or molecular mechanism. Therefore, carefully selected cases with chromosomal abnormalities and SVs, may hold the potential to offer valuable insights into further understanding novel aspects of CRS pathology and craniofacial development.

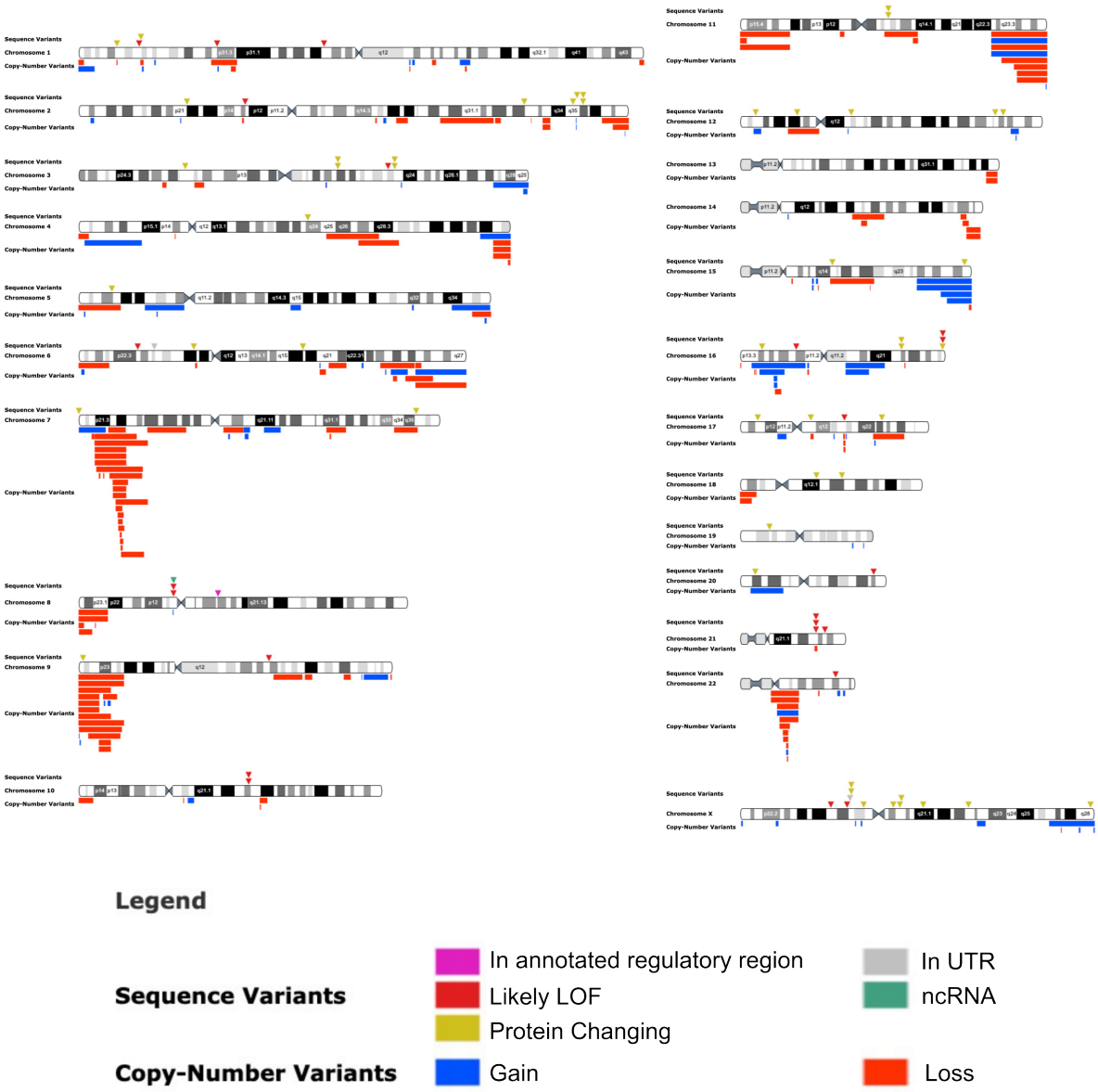


Figure 11 Summary of both sequence variants and CNVs in DECIPHER cases with CRS. *Figure extracted from DECIPHER variant browser by querying "craniosynostosis" (www.deciphergenomics.org/search/patients/browser?q=phenotype%3ACraniosynostosis, last accessed 31.10.2023)*

Despite significant efforts, a considerable number of CRS cases remain without a molecular diagnosis. Prior to this project, in the local study of 666 CRS patients, only 24% had a genetic diagnosis.[80,90] This number increases to 63% when looking at cases with a likely genetic underlying cause, such as cases with multiple sutures affected, positive family history, and/or syndromic features. Additionally, in the

100kGP CRS cohort of 114 cases that had been screened for common causes, only 13 cases had a positive genetic diagnosis.[90] Overall, this leaves a significant number of cases likely having an underlying genetic cause and yet without a molecular diagnosis despite previous efforts. Given the availability of enhanced SV detection technologies, improved SV interpretation resources, and the large volume of accessible high-quality data from the 100kGP, there is an excellent starting point and opportunity to identify and further understand clinically relevant SVs, with CRS as a model disease.

1.8 Summary, hypothesis, and aims

SVs account for a significant portion of genetic variation, arising through various DNA replication or repair-mediated mechanisms, including NAHR, NHEJ, and FoSTeS. Conventionally, SVs have been studied predominantly using physical genome mapping technologies, including karyotyping, FISH, and array-based approaches. While effective, these technologies usually have limited resolution. With improved resolution, the use of short-read technologies significantly expanded the scope of detectable SVs to include small/medium sized events, as well as CN neutral events which were previously challenging to detect. However, the DNA fragmentation requirement for short-read technologies remains a major limitation, as it relies on *post hoc* computational methods to reassemble large CPX events. Recent developments in 3GS technologies have further improved SV detection and characterisation through the use of long sequencing reads, while NGM technologies provided a much-needed upgrade to the conventional physical mapping technologies. These novel technologies offer more effective approaches for SV detection and characterisation, being able to examine intact DNA molecules with minimum fragmentation requirements. The

increasing power of SV detection has further facilitated the aggregation of SVs from both control (eg DGV) and case (eg DECIPHER) cohorts, enabling a more comprehensive understanding of SVs at the population level. With these technological and informational advancements, we are in a prime position to better characterise and understand clinically relevant SVs in human disease. CRS, as a highly genetically heterogeneous condition with a substantial, but minority fraction known to be caused by monogenic or chromosomal abnormalities, is an excellent test disease to seek further clinically relevant SVs and investigate their significance.

By exploring the available CRS cases in both local (Oxford-based) craniofacial malformations study and the 100kGP, and using both short-read and long-range technologies, I set out to identify and characterise clinically relevant SVs which may have been previously overlooked in patients. My hypothesis is that:

1. Pathogenic SVs may account for some of the CRS cases currently without a genetic diagnosis.
2. Some clinically relevant SVs have been overlooked due to the inherent disadvantage of short-read and conventional technologies in detecting SVs.
3. Long-range technologies, such as ONT and Bionano OGM, offer superior performance in SV detection compared to short-read sequencing technology.

To address these hypotheses, I analysed the SVs in the CRS cohort from the 100kGP, aiming to identify previously overlooked clinically relevant SVs in **Chapter 3** and **Chapter 4**. Complex SV candidates from the 100kGP analysis were further characterised using Bionano OGM and FISH in **Chapter 4** and **Chapter 5**. Additionally,

in **Chapter 5**, locally available 100kGP cases and several non-100kGP cases were analysed using Bionano OGM to identify clinically relevant SVs potentially overlooked by short-read technologies. Lastly, in **Chapter 6**, I evaluated the performance of the different SV detection technologies via a three-way comparison between Illumina WGS, ONT WGS, and Bionano OGM, based on the available data from the CRS cohort. Overall, this project aims to improve the overall SV diagnostic rate in human disease, while concurrently providing a better understanding in CRS mechanisms and molecular pathology.

Chapter 2 Methods

2.1 Ethics

The clinical studies were approved by the relevant Institutional Review Boards including Oxfordshire Research Ethics Committee (REC) B (C02.143), London–Riverside REC (09/H0706/20 for Genetic Basis of Craniofacial Malformation [GBoCM]), and East of England–Cambridge South REC (14/EE/1112 for 100kGP). Written informed consent was given by each child's parent or guardian to obtain samples for genetics research.

2.2 Array data analysis

Array data were collected from the 45 cases included in the Genetic Basis of Craniofacial Malformation study. The reported CNVs were re-examined and organised into a database and later cross-referenced with the whole exome sequencing (WES) and WGS results to assess the clinically utility of these approaches.

2.3 WES data analysis

WES data of 268 individuals from 142 families (51 solved) were obtained from previous work[91] by the Clinical Genetics Group and SV calling (in hg38 genome build) was performed by the Medical Research Council (MRC) Weatherall Institute of Molecular Medicine (WIMM) Centre for Computational Biology (CCB) using SavvyCNV.[92] Detailed protocol for WES data generation was described in Miller et al (2017).[91] The CNV calls were filtered by excluding CNVs with phred/width < 10 as recommended by the developer (https://github.com/rdemolgen/SavvySuite, last accessed 3rd Nov 2020). Segregating CNVs were analysed individually by querying the UCSC (University of California, Santa Cruz) genome browser with tracks including GENCODE v32 Comprehensive Transcript Set, GeneHancer Regulatory Elements

and Gene Interactions, comparison of Micro-C and In situ Hi-C protocols in H1-hESC and HFFc6[93], H3K27Ac Mark on 7 cell lines from ENCODE, Vertebrate Multiz Alignment & Conservation (100 Species), DGV: Structural Variation (CNV, Inversion, In/del), and Repeating Elements by RepeatMasker. In addition, the SNPs ratio within the SV region should also support the genotype, ie a DUP including a heterozygous SNP should have roughly double the number of supporting reads for the alt SNPs than the ref SNP or *vice versa.*

2.4 100kGP WGS data analysis

The majority of the 100kGP analysis has been summarised comprehensively in the resulting publication Hyder et al (2021).[90] WGS was performed on the Illumina platform (HiSeq X) and processed by GE. I extracted the primary cohort of CRS patients using the term "Craniosynostosis" in the "Normalised Specific Disease" data field of the 100kGP main programme v10. SVs from this CRS cohort were called in hg38 (263 samples) and hg19 (74 samples), and subsequent analyses were done in the respective genome version. I performed the WGS analysis in the GE Research Environment as a member of the musculoskeletal Genomics England Clinical Interpretation Partnership (GeCIP) under the Research Registry Project RR65. SV calling was performed using three SV callers: Manta[36], Canvas[34], and Lumpy[35]. Manta and Canvas calling was performed by GE; I later performed SV calling using LUMPYExpress (Lumpy v.0.2.13) as described in https://github.com/arq5x/lumpy-sv#lumpy-express-usage (last accessed 3rd Nov 2020).

An analysis pipeline (**Figure 12**) was designed to extract true and most likely pathogenic variants. A series of annotations was applied to the SV calls to: 1) exclude

likely common or false positive SV/CNV 2) categorise SVs based on segregation analysis and 3) prioritise most likely true positive and pathogenic calls for further functional studies. As described below, five annotations (VarCount, DGV, gap region, Coding region, and a CRS gene panel) were applied to the Canvas and Manta datasets, and a Union annotation was added for the Lumpy dataset. Post annotation analyses were carried out using relevant familial information, known disease associations, targeted screening, mosaicism detection, and the SVRare software.[94]

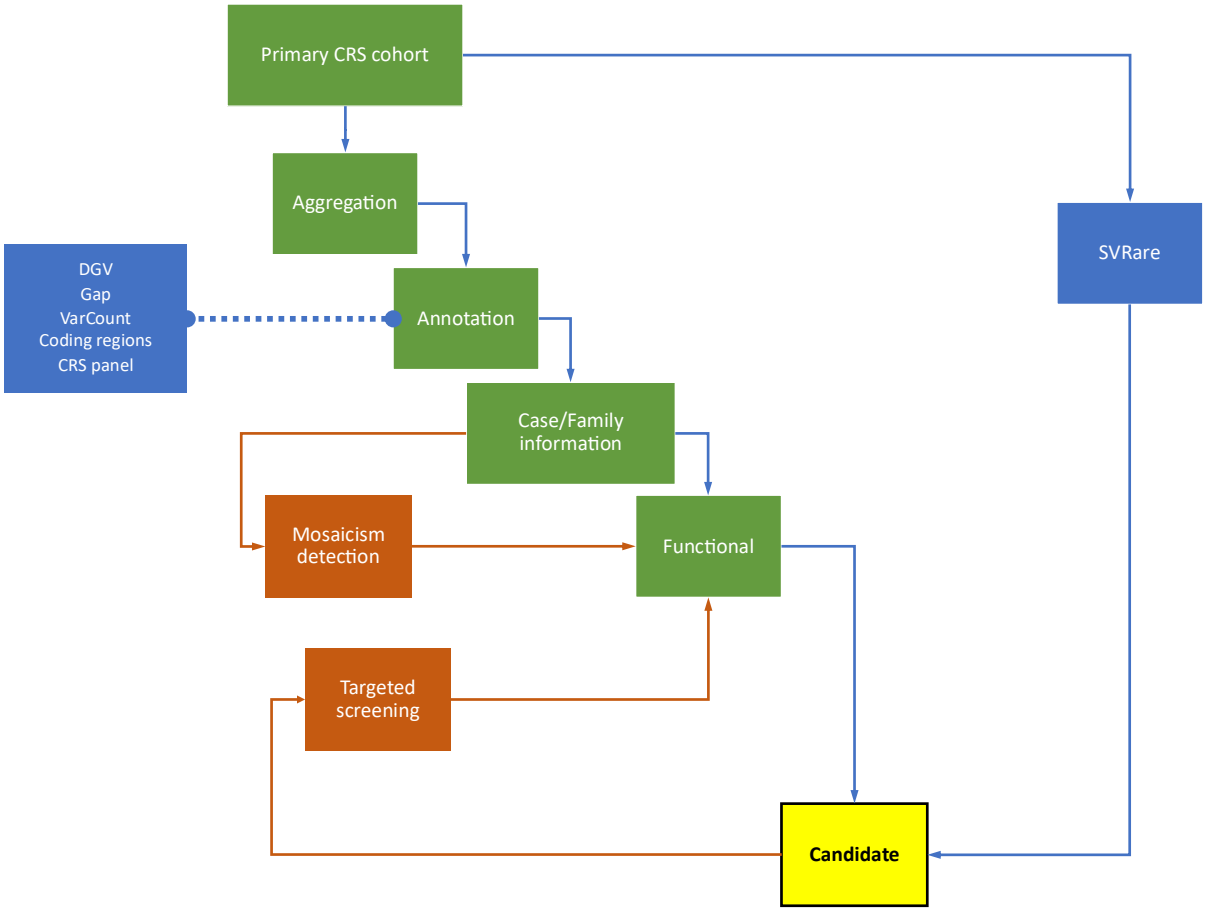


Figure 12 The 100kGP analysis pipeline was designed to identify SV candidates from WGS data in the primary CRS cohort. *The WGS data were first annotated with five different annotations, and then analysed with phenotypic and functional information from the literature and databases to generate a list of candidate SVs. Non-segregating SVs were further analysed to detect mosaicism. The pipeline was also used to screen prioritised regions from the candidates across the 100kGP rare disease cohort to explore SVs in human diseases beyond CRS. In addition, the pipeline's performance was complemented by using SVRare[94], which enhanced certain aspects of the analysis, such as singletons (individual cases without parental information) and INV analysis.*

2.4.1 VarCount annotation

This annotation aimed to filter out likely common or false positive SV/CNV based on the rarity of a specific call in the CRS cohort. The rationale is that high frequency variants are likely to be noise or common polymorphisms, while rare variants are of primary interest in our data. A particular challenge for this approach occurs when the same rearrangement is called with slightly differently breakpoints in different samples. Such calls are particularly likely to be associated with repetitive regions, assembly gaps, INVs, highly polymorphic regions, and most Canvas calls, since Canvas relies solely on the changes in read depth to call CNVs. Therefore, I applied a "fuzzy-end" matching strategy which determines SV calls as the same event using five pieces of information: chr, start, end, SV size (SV length in base pairs), and SV type, as illustrated in **Figure 13**.

Permissible length variation of ±10% was selected after testing 5%, 10%, 15%, and 20% of the SV size in the query. The results showed that 5% SV size was too short to match large events especially Mb-sized INVs, whilst ≥15% SV size failed to distinguish small (~100 bp) yet unique SVs in the same region. Using this strategy, each SV call was queried against the primary CRS cohort, and the number of matching calls for each query was deposited as the VarCount value of the queried call.

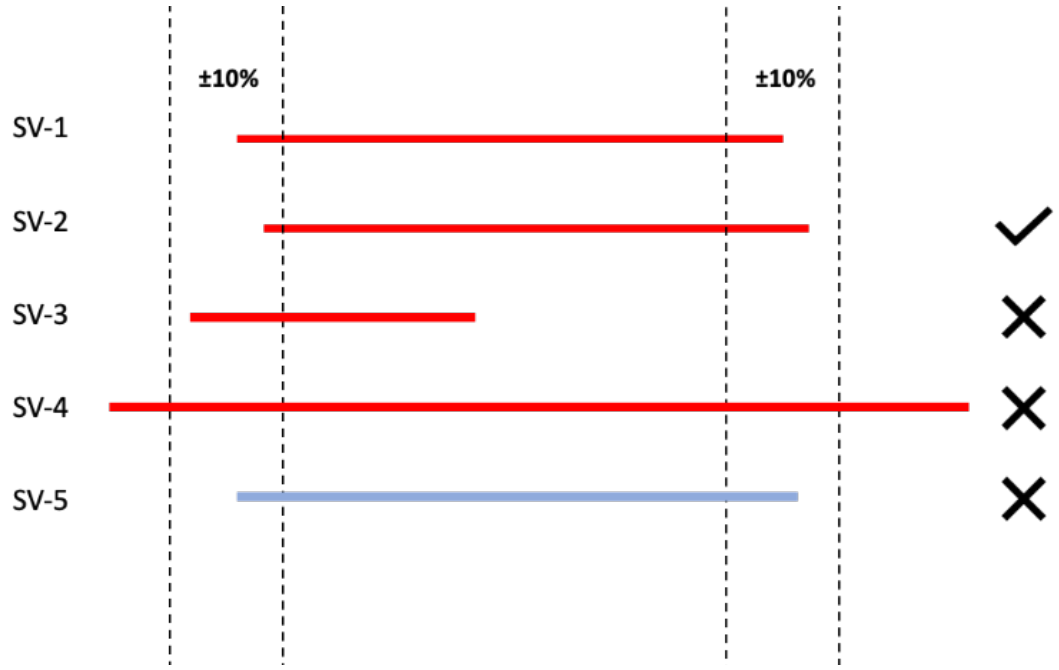


Figure 13 Soft "fuzzy" matching approach for identifying calls representing the same event in different samples*. SV calls were determined as the same event if they share the same chr, SV type, and the start and end lie within a range that is 10% of the SV size in base pairs, such as SV-1 and SV-2. SVs such as SV-3, SV-4, and SV-5 (a different event type) did not match all criteria and were therefore marked as representing different events.*

2.4.2 DGV annotation

The DGV database aggregates SV/CNV in >22,000 genomes of the healthy population,[8] and therefore, provides a valuable annotation for excluding common/benign SVs. An SV observed in DGV, hence in the healthy population, is highly unlikely to be pathogenic. Therefore, SVs were excluded if they were also observed in DGV. However, SVs were retained regardless of DGV annotation when they intersected with any of the ClinGen Dosage Sensitivity - Recurrent CNVs.[9] This ensured that no recurrent pathogenic CNVs escaped analysis owing to the stringent DGV filter.

DGV data were obtained from UCSC DGV Struc Var – dgvMerged. Similar to that of the VarCount, a fuzzy query (±20% SV size) was applied to determine if an SV call

was present in DGV, with a few adjustments to accommodate the unique features of DGV data. Firstly, a large proportion of DGV SVs were obtained from chromosomal microarray data; these SVs are likely to have less well-defined breakpoints compared to the SV calls from genome sequencing, and hence need a larger permissible variation at the breaks. Secondly, the precision of the DGV matching is more relaxed than that of the VarCount matching because different SV/CNV events within a given genomic region are likely to be benign when this region harbours similar benign rearrangements in the healthy DGV population. Overall, the ±20% SV size criterion was selected for the DGV query value after testing 10%, 15%, 20%, 25%, and 30%. In addition, to address the difference in variant terminology in DGV compared with Canvas/Manta, I applied the conversions as shown in **Table 2**. Note that Break end (BND) calls were excluded from annotation but retained for verification of complex event.

Table 2 DGV to Manta/Canvas call conversions

		Manta/Canvas call type			
		DEL	DUP	INS	INV
DGV SV Type	Complex	✓	✓	✓	✓
	Gain + loss	✓	✓		
	Deletion/loss	✓			
	Gain/duplication/tandem duplication		✓		
	Inversion				✓
	Insertion/mobile element insertion/novel sequence insertion			✓	

2.4.3 Gap regions annotation

The Gap annotation aims to exclude noise clusters at specific regions of the genome that are largely inaccessible by Illumina WGS. These include the short arms of acrocentric chromosomes, heterochromatin, telomeres, centromeres, gaps between contigs and scaffolds in chromosome assemblies. These regions were obtained from the UCSC genome table browser under the gap subset.[10] The SV calls were then queried against the gap subset using the midpoint of the call, as this gap annotation aims to remove false positive calls clustered around noisy regions. The midpoint of the SV is calculated as follows:

$$SV_{mid} = (SV_{END} - SV_{START})/2$$

On the same chromosome, an SV call was annotated as in the Gap region when SV_{mid} is between the start and the end coordinate of the gap region.

2.4.4 Coding regions annotation

This annotation allows SVs to be prioritised under the assumption that in general, protein altering SVs are more likely to be pathogenic compared to non-coding SVs, and hence have a priority in my analysis. The coding region data were extracted from UCSC table browser gene & gene predictions (GENCODE v32),[11] and the SV calls were annotated with the coding regions using an “intersection” query approach. Briefly, the SV break points were used when identifying an intersection between the SV and a coding region. As shown in **Table 3**, there are six possible relations between the SV (green) and a region of interest (blue). Of all possible relations, four are annotated as positive for SV-region intersection. SVs were prioritised when they intersected with at least one coding region.

Table 3 SV-coding region intersection criteria

SV/region position	Break requirement	Intersecting?
	V1 ≤ G1 & V2 ≤ G1	N
	V1 ≤ G1 & V2 ≥ G1	Y
	V1 ≥ G1 & V2 ≤ G2	Y
	V1 ≤ G2 & V2 ≥ G2	Y
	V1 ≥ G2 & V2 ≥ G2	N
	V1 ≤ G1 & V2 ≥ G2	Y

V1/V2 represents the start/end of an SV; G1/G2 represents the start/end of a region of interest (gene/exon)

2.4.5 CRS panel annotation

SV calls were further annotated using a panel of known CRS genes. These genes were extracted from the Genomics England PanelApp for Craniosynostosis (Version 2.2),[12] and the respective genomic locations were defined by Ensembl Genes 101 GRCh38.p13/GRCh37.p13. Genes are mostly defined from the 5′-UTR to the 3′-UTR by Ensembl, while sometimes extensive regulatory regions are included (for example, *TWIST1* is defined by Ensembl to include a ~96 kb regulatory region upstream of the 5′-UTR). The CRS panel was queried against the SV calls using the intersection approach outlined in **Table 3**. SVs intersecting with any known CRS gene were prioritised.

2.4.6 Lumpy concordant matches

The Lumpy callset was annotated to identify concordant calls with Manta/Canvas. The Lumpy calls were queried against Manta/Canvas calls using the same “Fuzzy” matching approach as described in **Figure 13**. From this annotation, the discordant calls aim to provide extra SVs due to the low sensitivity of Manta/Canvas, while the concordant calls provided more confident prioritisation for singleton samples without available segregation analysis.

2.4.7 Segregation analysis

Two overall types of segregation analyses were carried out under three different hypotheses:

1) CRS is caused by SVs that segregate with the phenotype. Under this assumption, segregating SVs were selected, such as inherited SVs in familial cases and *de novo* SVs in sporadic cases with unaffected parents.

2) Pathogenic SVs may cause CRS through complex mechanisms that are harder to detect, such as parental mosaicism or incomplete penetrance. For this hypothesis, non-segregating and yet compelling SVs were further analysed for evidence of pathogenicity through complex mechanisms. An example would be SVs affecting known CRS genes but inherited from an apparently unaffected parent.

3) Pathogenic SVs may have been missed by the default callers. To test this hypothesis, discordant calls from Lumpy were analysed, with the rationale being that Lumpy may be able to detect some of the pathogenic SVs missed by Manta/Canvas.

2.4.8 SV filtering and prioritisation

SVs were filtered and prioritised using the annotations and segregation analysis described above and summarised in **Table 4**. Filtering and prioritisation criteria were adjusted based on family structure and segregation pattern. The four filtering annotations provided hard exclusion criteria, whilst the three prioritisation strategies provided soft filters that reserved low priority SVs for further analyses when needed. For the caller concordance, all Manta/Canvas calls were analysed; concordant Lumpy calls were analysed for singletons while discordant Lumpy calls were analysed for all other cases.

Table 4 SV segregation, filtering, and prioritisation strategies

Segregation	SV Inheritance	Segregating				Non-segregating	
	Family structure	Sporadic Trio	Familial Duos	Complex	Singletons	Sporadic Trios	Sporadic Duos
Filtering	VarCount (non-ClinGen)[b]	= 1	= 2	= # of affected	= 1	= 2	
	DGV	Not observed in DGV or intersecting ClinGen recurrent CNVs					
	Gap	N					
	Concordance	M&C + discordant L	M&C + discordant L	M&C + discordant L	M&C + concordant L	M&C + discordant L	M&C + discordant L
Prioritisation	Coding regions	Coding SVs prioritised					
	CRS panel	SVs including CRS genes prioritised					
	Size	Large SVs prioritised					

Familial Duos are families with an affected child and an affected parent available in the 100kGP dataset; sporadic duos are families with an affected child and an unaffected parent available; sporadic trios are trio samples, with affected proband and unaffected parents; singletons are affected individuals without parental data in the 100kGP; complex cases are families with any other composition (multiple affected and/or with relative samples instead of parents). Recurrent ClinGen CNV calls were set with VarCount < 20. M&C = Manta&Canvas dataset; L = Lumpy dataset.

2.4.9 SV call verification and further analysis

The 100kGP analysis generated a large list of SVs, which required further manual examination to assess the evidence that they were real and to further assess their pathogenicity.

To verify the SV calls, the Integrative Genomics Viewer (IGV)[13] was used to assess supporting reads at the breakpoints; Samplot[14] was used to visualize the overall read depth change supporting the called SV.

To assess the pathogenicity of the true SV calls, databases and the literature were queried manually. From gnomAD, the constraint of candidate genes was assessed, particularly for loss-of-function SVs such as DELs; gnomAD SVs v2.1[15] provided further evidence on the frequency of the SV call in the gnomAD cohort. DECIPHER was queried to identify any similar SV findings from recorded cases, from which any independent cases with similar features would advance evidence of SV pathogenicity. True SVs were also assessed for conservation within the affected region and possible affected regulatory elements, using the UCSC genome browser with additional tracks including GeneHancer Regulatory Elements and Gene Interactions and Vertebrate Multiz Alignment & Conservation (100 Species). Known disease associations, particularly with skeletal abnormalities, were assessed using OMIM and literature searches.

2.4.10 Mosaicism detection

An SV mosaicism detection method was developed to investigate the presence of mosaicism in cases where the SVs of interest did not segregate with the observed phenotypes in the family, such as when the SV was inherited from an unaffected parent. Overall, this method uses the WGS read count from the SV region of interest, normalised via the whole chromosome, and compared with that of a control to see if an unbalanced SV is constitutional. The detailed method is as following: 1) the read count from the SV region of interest, C_{SV}, was extracted from the Binary Alignment Map (BAM) files of the individual of interest (usually the carrier parent) and a normal control (usually the non-carrier parent when available). 2) the read count of the whole chromosome, C_{chr}, of the SV was also extracted from the same individuals (eg carrier parent and control). 3) the ratio, R_{mos}, of the reads between the normalised SVs are compared against the control as the following:

$$R_{mos} = \frac{\frac{C_{SV}Carrier}{C_{SV}Carrier}}{\frac{C_{SV}Control}{C_{SV}Control}}$$

4) likelihood of the mosaicism was determined based on the R_{mos} and the SV type, as shown in **Table 5**. A point of caution was exercised when considering complex SVs and SVs in highly polymorphic/repetitive regions, as Illumina WGS may not be able to accurately map reads in these SVs/regions, resulting in inaccurate read counts and consequently R_{mos}.

Table 5 Mosaicism detection method based on the R_{mos} value

	R_{mos}	Mosaic
DEL/CNV loss	~50%	Constitutional
DEL/CNV loss	60%-90%	Likely mosaic
DUP/INS/CNV gain	~150%	Constitutional
DUP/INS/CNV gain	110%-140%	Likely mosaic

2.4.11 100kGP rare disease screening

Initial 100kGP analysis generated a list of SVs of interest. For the most compelling SVs, the corresponding regions were screened in the entire 100kGP rare disease cohort to detect any pathogenic SVs in non-CRS patients. Nine regions were selected and screened, as shown in **Table 6**.

Table 6 List of regions screened in the 100kGP rare disease cohort

Regions	Coordinates	Genome Build
FGF9 locus	chr13:21609875-22349400	hg38
	chr13:22184914-22923539	hg19
HOXC cluster	chr12:53980401-54265909	hg38
	chr12:54374184-54659693	hg19
KANSL1	chr17:46029916-46225389	hg38
	chr17:43360658-45315396	hg19
TWIST1 locus	chr7:17346121-19695497	hg38
	chr7:17385745-19735120	hg19
KCNJ16/2 locus	chr17:69912724-70477512	hg38
	chr17:67908866-68473653	hg19
BCL11B locus	chr14:97170412-100597846	hg38
	chr14:97636750-101064183	hg19
DSCAML1	chr11:117671416-117673929	hg38
	chr11:117539748-11544644	hg19
IL11RA locus	chr9:34150702-35161881	hg38
	chr9:34150700-35161878	hg19
ZIC1 locus	chr3:146660403-148134627	hg38
	chr3:146378191-147852414	hg19

2.4.12 SVRare analysis

SVRare is a toolkit developed to aggregate and prioritise rare, and therefore more likely disease-causing SVs within the 100kGP project.[94] SVRare generates a report for each family in the 100kGP. I extracted reports for all cases in the primary CRS cohort. The report consists of SV calls from Manta and Canvas, with each call annotated with several features, including:

1) number of matched calls in all (N_all)/patient with same disease (N_share_disease) /all other families (N_else),

2) coding regions disrupted (protein_coding_disrupted_genes),

3) known disease genes disrupted (interesting_HPO[Human Phenotype Ontology]_gene),

4) genotypes of the family members for the SV call, if abnormal,

5) genotypes of other unrelated participants for the SV call, if abnormal.

Similar to section **2.4.8**, SVRare reports were analysed using filtering and prioritisation strategies based on the family structure as shown in **Table 7**. Subsequently, each SV call was inspected manually in IGV to verify the call and assess their clinical relevance to the phenotypes in the family.

Table 7 SV segregation, filtering, and prioritisation strategies for the SVRare analysis

Segregation	SV Inheritance	Segregating				Non-segregating	
	Family structure[a]	Sporadic Trio	Familial Duos	Complex	Singletons	Sporadic Trios	Sporadic Duos
Filtering	N_all	= 1	= 2	= # of affected	= 1	= 2	
	N_else	= 0 (common SVs in 100kGP were excluded)					
Prioritisation	protein_coding_disrupted_genes	Coding SVs prioritised					
	interesting_HPO_gene	SVs including CRS genes prioritised					
	Size	Large SVs prioritised					

Familial Duos are families with an affected child and an affected parent available in the 100kGP dataset; sporadic duos are families with an affected child and an unaffected parent available; sporadic trios are trio samples, with affected proband and unaffected parents; singletons are affected individuals without parental data in the 100kGP; complex cases are families with any other composition (multiple affected and/or with relative samples instead of parents).

2.5 Breakpoint PCR & Dideoxy-sequencing

The precise molecular rearrangements of each SV were validated using breakpoint PCR and dideoxy-sequencing when samples were available. Primers were designed to span each individual breakpoint or multiple break points when possible. A list of PCR primers used is shown in **Table 8**. Standard PCR with FastStart Taq DNA Polymerase (Roche, FTAQ-RO) follows the modified protocol from New England Biolabs (NEB), as shown in **Table 9** and **Table 10**; LongAmp (NEB, M0323S) protocols, also from NEB, were shown in **Table 11** and **Table 12**. I performed all breakpoint PCR while all dideoxy-sequencing was performed by the MRC WIMM Sequencing facility. I analysed all dideoxy-sequencing data: sequencing chromatograms were visualised using 4Peaks (Nucleobytes), and break points were located using the UCSC BLAT tool (https://genome.ucsc.edu/cgi-bin/hgBlat, last accessed, 6th May 2023)[95].

Table 8 Primers used for breakpoint PCR including the expected product length

Case ID	Candidates	SV	Primer	Oligo 5'-3'	Product length (bp)	
27	*ABL1*, *etc.*	chr9 DUP	F	AATGGATCGGGGTGGGTATG	620	
			R	AGTTTTATTGCTGGGCGTGG		
1	*CERS6*	chr2 CON	F	AGAGGGTTGGCATTCTTTGCAG	595	
			R	GCACTAGGGAGATGCAGAAAGTTG		
10	*FGF9*	chr13 INS	F	CATAGAAAAATCTTGGTAATTGGC	403	
			R	TGACAGTAGCTGCTTAATTCAGG		
			F	GTATTGTCTACTTTAGATATCAGC	415	
			R	TCGTTAATATCTTACAACCATCC		
16	*PLCB4*	chr20 INS	F	TCAGCAAGGAAGCCATCAGG	485	
			R	GAAGTGCTGCTCCTTGAATTGC		
			F	GCTCTGCTAATAATGGTTTGGCTC	451	
			R	CTGAATGAAAACGTACCTGGGC		
			F	ACCATATTTAAGCTAGGCTCCC	505	
			R	TGAGCCTAAAAAAGATCCCCC		
			F	CTTTCACCTTCGCTGTGTTGG	691	
			R	CTCCATTTGCCTACCAATTTCCC		
2	t(16:17)(q23:q24)	16cen - 17qter	F	AATCACTCTATGGAGGGTCCAGC	197	
			R	CTTCCACTTATGCTTACCCCTGC		
		17cen – 16qter	F	CTGTCATTCCCACAGTACATATTGCC	474	
			R	GGGGACACAACCAAACCATATTAGC		
12	*TWIST1*	7pINV 5'	F	CTTTATAGGACCAACACATGGGTCTGCC	248	
			R	AAGTAGCTGGGACTACAGGCATGTGC		
		7pINV 3'	F	CATAGATGTCTGGGAGAGTCATCCTATGGG	2109	Long-Amp
			R	GGCCTCCTTTACCTGAGCATAATATGGG		
			S	TCACTTGAGCACAGGAGTTCAAGACC	-	
18	*HOXCs*	chr12 CPX	F	ACATGTCTCCCCACTGATCTCAAACG	3073	Long-Amp
			R	TAGGTAGGTAGGCAGCAAGACTTTGC		
21	*HOXCs*	chr12 DUP	F	GGAGGCAAGAAATGTTCTCCC	686	
			R	GGTTTGCTTCTAAGTTGCCCC		
22	t(4:7)(p21:p15)	7pter-4cen	F	CCAGTCTATCATTGTTGGACATTTGGG	483	
			R	GTAGCCAAATGGTGATGAAATTGGCAG		
31	*KIAA0825*	chr5 DEL	F	AATGACTTAGTCCTGGGACCG	167	
			R	GAAGATGAATAACCCAGAGAGCCC		

F: forward primer; R: reverse primer; S: internal primer for sequencing only and hence no PCR product; Long-Amp tag indicates that the PCR amplification followed the Long-Amp protocol in ***Table 11*** *and* ***Table 12****; all others follow the standard Taq PCR protocol in* ***Table 9*** *and* ***Table 10***

Table 9 Standard PCR setup using Taq DNA polymerase

	Volume (µl) per reaction
20 µg/µl DNA	1
10 µM Forward primer	1
10 µM Reverse primer	1
10 mM dNTPs	1
10X Standard Taq Reaction Buffer	2
***Taq* DNA Polymerase**	0.1
Nuclease-free water	13.9
Reaction volume	20

Table 10 PCR program with Taq DNA polymerase

Step	Temp	Time
Initial Denaturation	95°C	8 min
32 Cycles	95°C 60°C 72°C	30 s 30 s 1 min/kb
Final Extension	72°C	10 min
Hold	10°C	∞

Table 11 LongAmp PCR setup using LongAmp Taq DNA polymerase

	Volume (µl) per reaction
10 µM Forward Primer	0.5
10 µM Reverse Primer	0.5
20 µg/µl DNA	1
LongAmp *Taq* 2X Master Mix	10
Nuclease-free water	8
Reaction volume	20

Table 12 PCR program with LongAmp Taq DNA polymerase

Step	Temp	Time
Initial Denaturation	95°C	2 min
	95°C	30 s
30 Cycles	60°C	30 s
	65°C	50 s/kb
Final Extension	65°C	10 min
Hold	10°C	∞

2.6 Fluorescent in situ hybridisation (FISH)

A three coloured FISH was designed for case 10 to distinguish the alternatives using both the order and the distance between the FISH probes. Bacterial artificial chromosome (BAC) clones, shown in **Table 13**, were selected and ordered from BACPAC GENOMICS (https://bacpacresources.org/, CA). I performed BAC clone culture (**section 2.6.1**), DNA extraction (**section 2.6.1**), and nick translation (**section 2.6.2**). Metaphase cell harvest (**section 2.6.3**) was performed by Dr. Dagmara Korona. Slide preparation (section 2.6.4), treatments (**section 2.6.5**), probes processing (**section 2.6.6** and **2.6.7**) were carried out under the supervision and mostly by Jill Brown (MRC Molecular Haematology Unit [MHU], WIMM, Oxford).

2.6.1 BAC clone culture and DNA extraction.

A starter culture was generated by inoculating 5 ml of chloramphenicol Luria-Bertani (LB) medium (10 g/L Tryptone, 5 g/L Yeast extract, 10 g/L NaCl, and 12.5 µg/ml chloramphenicol) with a single colony from a streaked LB plate (from the ordered BAC clones), which was then incubated overnight at 37°C with shaking. The starter culture was used to inoculate 250 ml of chloramphenicol LB medium in a 1000 ml autoclaved flask, which was then incubated at 37°C with shaking overnight. The culture was harvested by centrifugation at 3,220 Relative Centrifugal Force (RCF) for 20 min at

4°C, and the supernatant was discarded. The cell pellet was resuspended in 50 ml of cold GET solution (Glucose 0.92% w/v, EDTA 10 mM, Tris 26 mM) by pipetting up and down with a 10 ml pipette until a homogeneous suspension was obtained. Next, 50 ml of room temperature NaOH/sodium dodecyl sulphate (SDS) solution (0.2 M NaOH, 1% SDS w/v) was added to the suspended cells and mixed by gentle inversion. The mixture was incubated at room temperature for 5 min, and then 50 ml of cold potassium acetate solution (3 M potassium acetate, 11.54% Glacial Acetic Acid v/v) was added and mixed by gentle inversion. The mixture was incubated on ice for 20 min. The mixture was then centrifuged at 3,220 RCF for 50 min at 4°C. The supernatant was carefully transferred to a new 250 ml conical bottle by filtering through Propax gauze to avoid disturbing the protein pellet. Ninety ml of isopropanol was added to the supernatant, mixed by gentle inversions, and incubated for 5 min. The mixture was again centrifuged at 3,220 RCF for 50 min at 4°C, and the supernatant was discarded. The DNA pellet was washed with 25 ml of 70% ethanol and then transferred to a 50 ml Falcon tube. The pellet was centrifuged at 3,220 RCF for 40 min at 4°C, and the supernatant was discarded. The pellet was air-dried for 30 min and then dissolved in 800 µl of PCR grade water overnight at 4°C. The concentration of the extracted DNA was quantified using the Qubit 4 Fluorometer following the manufacturer's guidelines.

Table 13 BAC clones for FISH of case 10

BACs	Genomic Coordinates hg19	Label	Backup?
RP11-441F14	chr13:22,855,636-23,039,514	Cy3	
RP11-316G23	chr13:22,796,525-22,982,810	Cy3	Backup
RP11-177H23	chr13:23,932,784-24,120,803	DIG	Backup
RP11-1095B20	chr13:23,323,910-23,587,290	DIG	
RP11-355O23	chr13:24,187,489-24,364,907	AF647	Backup
RP11-1013L17	chr13:24,223,379-24,423,559	AF647	

Cy3: Aminoallyl-dUTP-XX-Cy3; DIG: Digoxigenin-11-dUTP; AF647: Aminoallyl-dUTP-XX-AF647. Backups were second set of probes targeting the same genomic region in case the first set of probes failed.

2.6.2 Probe Preparation by Nick Translation

DNA probes were prepared using a modified version of the nick translation method. Firstly, to remove any RNA contamination, 1 µg of previously extracted BAC DNA was treated with 200 ng of RNase (Sigma, R4642-10MG) at 37°C for 30 min.

Nick translation was carried out in a 50 µl reaction volume, which included 1 µg of RNase-treated DNA, 5 µl of nick translation buffer (0.5 M Tris-HCl pH 8.0, 50 mM $MgCl_2$, and 0.5 mg/mL BSA), 5 µl of 0.1 M β-mercaptoethanol, 5 µl of dAGC Mix (0.05 M dATP, 0.05 M dGTP, and 0.05 M dCTP), 3 µl of DNase (Takara Bio, 2270A), 1 µl of hapten (1 nmol cy3/cy5/digoxigenin(DIG)-dUTP, Jena Bioscience, **Table 13**), and 20 units of DNA polymerase I (New England Biolabs, M0209S). The reaction mixture was incubated at 16°C for 2 hours.

Digestion efficiency was assessed by visualising the digested DNA on a 2% TBE agarose gel. The DNA smear should be less than 500 bp. The resulting digested DNA was cleaned using illustra G-50 columns (GE Healthcare), according to the manufacturer's instructions. Clean DNA probes were stored at -20°C.

2.6.3 Metaphase cell harvesting

I prepared all reagents/solutions required while metaphase cells were harvested by Dr. Dagmara Korona. More than 5 million induced pluripotent stem cells (iPSCs) were split from both the patient and a control line for harvesting metaphase cells. The cells were treated with Karyomax Colcemid (10 µg/mL, 15210-040 Gibco) to arrest cells at the metaphase stage of mitosis by inhibiting spindle microtubule formation. After incubation at 37°C for 1 hour, cells were washed with PBS, trypsinised, and spun down from the solution.

To create a single cell suspension, cells were incubated with 37°C hypotonic solution (sterile 37.5 mM KCl) for 15 min. To fix the cells, approximately 2 mL of -20°C fixative (3:1 methanol:glacial acetic acid) was added dropwise using a glass pipette while flicking the cells gently to avoid clumping. Additional fixative was added to make up 10 mL total volume, and the cells were incubated at -20°C for 30 min. After fixation, the cells were pulse spun to form pellets, which were stored at -20°C until further use.

2.6.4 Slide preparation

Harvested metaphase cells were fixed on to slides for hybridisation analysis. Cell pellets were taken out of the -20°C storage and resuspended in fresh fixative (3:1 methanol:glacial acetic acid). Microscope Slides (VWR® SuperFrost® BS 7011) were

pre-warmed by blowing open-mouthed. One drop of the cell suspension was added to the slide using a glass pastette. The drop was blown on again to encourage spreading of the cells. When the fixative starts to evaporate, another drop of fresh fixative (clean fixative without cell suspension) was added onto the centre of the previous drop. When dried, slides were checked with phase contrast microscopy, looking for well spread metaphase/interphase cells. The slides were then dried at room temperature (RT) for at least a week before proceeding to the next step.

2.6.5 Slide pre-treatment and chromosomal denaturation

Slide treatments were all performed in Coplin jars. Slides were incubated with 200µl of 100µg/ml RNAse (Sigma, R4642-10MG) at 37°C for 30min-1 hour. The slides two times in 2xSSC (Saline-Sodium Citrate buffer, diluted from 20xSSC, Sigma S6639-1L) at RT with agitation and then in 1xPBS at RT with agitation. After washing, the slides were dehydrated in 70%, 90%, 100% ethanol for 4 min each at RT, and then airdried.

Under a fume hood, slides were incubated in preheated formamide (70% formamide in 2xSSC at 72°C) in a Coplin jar for 5 min. Slides were immediately transferred to ice-cold 70%, 90%, and 100% ethanol sequentially for 4 min each. Slides were airdried and ready for hybridisation with probe.

2.6.6 Probe precipitation, denaturation, and hybridisation

To precipitate the probes, 3 µl Human Cot-1 DNA (Invitrogen, 15279011), 2 µl of Salmon Sperm DNA (Invitrogen, 15632011), 1 µl of 3M sodium acetate (pH5.3), and 22.5 µl of ice cold 100% ethanol were added to 5 µl of each probe (**Table 13**). The mixture was incubated at -20 °C overnight to precipitate the probes.

After incubation, the supernatant was carefully discarded. The pellet was washed in ice cold 70% ethanol, vortexed briefly, and centrifuged at 20,000 RCF at 4°C for 10 minutes. The washed pellet was collected by removing the supernatant and airdried at 37 °C for 5-10 min, while avoiding over-drying. The pellet was resuspended in 12 µl hybridisation mix (50 µl formamide, 10 µl milliQ water, 10 µl 10% Tween in 2xSSC, 10 µl 20xSSC at pH5.4, and 20 µl 50% dextran sulphate) at 37°C and shaken on Eppendorf ThermoMixer C at 1400 rpm for 5 min. The probe mixture was proceeded to denaturation at 90°C for 8 min, and then pre-annealing at 37°C for 20 min.

The pre-annealed probe mix was placed on the area to be hybridised on the slides (**section 2.6.5**), which was then covered with an appropriately sized coverslip, and sealed with bike repair glue. The sealed slides were incubated at 37 °C for 2 days.

2.6.7 Probe cleaning, blocking, and detection

The probe hybridised slides were cleaned with a series of washes. Firstly, the coverslips were removed by washing the slides with 2xSSC. The slides were then washed in 5 steps with: 50% formamide in 2xSSC pH7.0 at 45 °C for 3 times 4 min per wash; 2xSSC at 45 °C for 3 times 4 min per wash; 0.1xSSC at 60 °C for 3 times 4 min per wash; SSCT (Saline-Sodium Citrate buffer with Tween, 4xSSC with 0.05% Tween 20) at RT briefly.

For antibody detection for digoxigenin (DIG) probes, 100 µl of blocking solution (3% BSA in 4x SSC) were added to each slide, which was covered with parafilm and incubated at RT for 30 min, and then washed briefly in SSCT at RT. For the first layer for DIG detection, 100 µl sheep anti-DIG-FITC (anti-digoxigenin- fluorescein isothiocyanate, Roche, 11207741910, 1:50 diluted) was placed on the drained slides,

which were then covered with Parafilm and incubated at RT for 30 min. After the first layer incubation, the slides were washed in SSCT at RT 3 times for 3 min.

After drying the slides, the second layer of antibody solution, 100 µl Anti-Sheep Antibody FITC-Labelled (Vector Labs, FI-6000, 1:100 diluted), was added to the slides, which were then covered with Parafilm and incubated at RT for 30 min. After the second layer incubation, the slides were washed in SSCT at RT 3 times for 3 min, and once in PBS for 5 min at RT.

Finally, the slides were dehydrated in 70%, 90%, and 100% ethanol sequentially. The Slides were visualised on the DeltaVision Elite system by Jill Brown (MRC MHU, WIMM, Oxford). I carried out image processing and analysis using Fiji – ImageJ.

2.7 EBV-transformed lymphoblastoid cell lines generation

Fresh EDTA blood, drawn within 24 hours and kept at RT, was used to generate Epstein–Barr virus (EBV)-transformed lymphoblastoid cell lines. Briefly, 1.3 ml of blood was mixed with 1.3 ml phosphate-buffered saline (PBS, Gibco™, 10010023) and layered on top of 3 ml of Histopaque® (Sigma-Aldrich, 10771) without mixing. The mixture was then centrifuged at 652 RCF for 25 min. White blood cells (WBCs) from the middle phase of the centrifuged mixture were collected with a Pasteur pipette, being careful to avoid collecting too much Histopaque. The WBCs were washed by resuspending in 7 ml of PBS and then centrifuged at 200 RCF for 5 min to collect the pellet while discarding the supernatant. The wash step was repeated, and the washed pellet was resuspended in ~500 µl of residual liquid after pouring out most of the supernatant. The resuspended WBCs were incubated with EBV (obtained from the

WIMM Genome Engineering Facility) at 37°C for 90 min. The culture was then incubated at 37°C for two days with 50 µl of Phytohemagglutinin M form (Gibco™, 2269531) and 4 ml media (500 ml RPMI-1640 [Sigma-Aldrich, R0883], 90 ml foetal bovine serum (FBS) [Sigma-Aldrich, F7524], 6 ml Penicillin-Streptomycin [Sigma-Aldrich, P0781], and 6 ml 1x L-Glutamine [Gibco™, 2554705]). The culture was checked every 3 days and the medium was exchanged when needed. Mycoplasma infection was checked when necessary, using the Lonza Lucetta™ Luminometer following the manufacturer's guidelines. When the culture reached the desired density, the cell pellet was collected by centrifugation at 200 RCF for 5 min. The pellet was resuspended in 3 ml freezing medium (90% FBS and 10% Dimethyl sulfoxide [Sigma-Aldrich, D2650]) and snap frozen as 1 ml aliquots at -80°C, and then transferred to liquid nitrogen for long-term storage.

To regrow frozen EBV-transformed lymphoblastoid cell lines, cell suspension aliquots were quickly thawed in a 37°C water bath. Defrosted cells were washed with 5 ml of media and centrifuged at 200 RCF for 5 min, and the supernatant was discarded. The cell pellets were then resuspended in media and incubated at 37°C. The culture was checked every few days, and the media was exchanged when needed.

2.8 Whole genome sequencing using Oxford Nanopore Technologies (ONT)

2.8.1 DNA extraction for ONT sequencing

DNA for ONT sequencing was extracted from fresh/frozen human blood using the QIAGEN Gentra Puregene Blood Kit based on the ONT "High molecular weight gDNA extraction from whole rabbit blood – QIAGEN Gentra Puregene Blood Kit" protocol. Briefly, red blood cells (RBCs) from fresh or thawed (water bath at 37°C) frozen blood

were lysed using RBC Lysis solution (Gentra). The white blood cell (WBC) pellet was collected via brief Eppendorf centrifugation and the RBC lysate supernatant was discarded. The WBC pellet was resuspended, lysed with Cell Lysis Solution (Gentra), and treated with RNase A Solution (Gentra). Proteins pellets were precipitated using Protein Precipitation Solution (Gentra) and centrifugation. The supernatant was collected, and DNA was precipitated using 100% isopropanol by gentle inversion and centrifugation. The DNA pellets were washed with 70% ethanol and eluted using DNA Hydration Solution (Gentra) for 2 hours at 50°C. The Extracted DNA was sent to Wellcome Sanger Institute (WSI) for ONT sequencing on PromethION.

2.8.2 ONT data processing

ONT raw data were processed by the WSI bioinformaticians using the WSI's internal pipeline, including mapping via Minimap2 and SV calling using Sniffles/Sniffles2. The specific tool version differences and their implications to the ONT performance was discussed in detail in **Chapter 6**. The SV Variant Call Format (VCF) files and the BAM files were then analysed as part of the Illumina vs ONT vs Bionano OGM comparison.

I reanalysed data from two trios initially by performing trio-calling using the Sniffles2 inbuilt populational calling function. SV calls were deemed *de novo* when the read support for a variant call is >0 in the proband but = 0 in both parents. I later performed populational re-calling on all available 10 trios. Coding calls from Sniffles2 were extracted by intersecting with the same coding regions as described in **section 2.4.4** but using BEDTools-*intersect* this time. Interesting candidates were examined individually on IGV as well as against the respective Illumina WGS and Bionano OGM data.

2.9 Bionano Optical Genome Mapping (OGM)

2.9.1 Bionano high molecular weight (HMW) DNA extraction

Clinical samples were received as fresh blood samples in EDTA tubes transported on ice/cold packs from collaborating clinicians. Samples were kept at 4°C and snap frozen in -80°C as 650 µl aliquots within 5 days (usually 2 days) of being drawn. HMW DNA for OGM was extracted using the "SP Blood & Cell Culture DNA Isolation Kit v2, product number: 80042" following the "Bionano Prep SP Frozen Human Blood DNA Isolation Protocol, 30246, revision F" protocol. Briefly, frozen blood aliquots (650 µl) were thawed in a 37°C water bath for 2 min. WBC counting was performed on the aliquots using the NucleoCounter® NC-3000™ Advanced Image Cytometer (ChemoMetec) with Solution 13 (ChemoMetec, 910-3013) following the manufacturer's instructions. After counting, blood containing 1.5 million WBCs was prepared with Stabilising Buffer (Bionano) and Proteinase K (Bionano). The WBCs were lysed and digested with Buffer LBB (Bionano). The digestion was then stopped with phenylmethylsulfonyl fluoride (PMSF, Sigma) after the recommended time to prevent unwanted DNA degradation. HMW DNA was precipitated using isopropanol and then captured by the Nanobind Disk (Bionano). Captured DNA was washed with Wash Buffers (Bionano) and then eluted using Elution Buffer (Bionano). Controlled shearing was performed by pipetting eluted DNA carefully 6 times according to Bionano guidelines. DNA was then resuspended at RT overnight.

DNA quantification was performed using the Invitrogen™ Qubit™ 4 Fluorometer, following the Bionano-modified Broad Range (BR) Qubit double-strand DNA quantification protocol. Two µl of the extracted DNA was sampled from the left, middle, and right of the DNA solution using a Positive Displacement Pipette (Rainin™). The

sampled DNA was fragmented using a vortex mixer at maximum speed for 30 s. Qubit Dye and buffer were added to the fragmented HMW DNA and the Qubit standards. The standards and samples were measured using the Qubit™ 4 Fluorometer. The coefficient of variation (CV) was calculated for each sample using the concentration of the Qubit readings from the left, middle, and right side of the DNA solution. If CV > 0.30, DNA was further homogenised by gentle pipetting up and down 5 times with wide-bore tips following Bionano guidelines, incubated at RT overnight, and re-evaluated using Qubit; if CV < 0.30, DNA was stored at -4°C until further use.

2.9.2 Bionano direct labelling and staining (DLS)

Extracted DNA was processed for optical mapping using the "Direct Label and Stain (DLS) Kit, part number 80005" following the "Bionano Prep Direct Label and Stain (DLS) Protocol, 30206, revision G". Briefly, 750 ng of HMW DNA was labelled with DL Green fluorophore (Bionano) using the DLE-1 enzyme (Bionano) in the DLE-1 buffer (Bionano) for 2 hours at 37°C. The enzyme reaction was halted by incubating the mixture with Proteinase K (Qiagen) for 30 min at 50°C. Labelled DNA was cleaned by two repeats of the membrane absorption process using the Bionano DLS membranes to remove enzyme and excessive fluorophore. The DNA backbone was then stained with a blue DNA stain (Bionano) in the Flow Buffer (Bionano) and dithiothreitol (DTT, Bionano) at RT overnight with an hour of gentle inversion at 5 rpm. The labelled and stained DNA (DLS DNA) was quantified following a similar Bionano-modified Qubit protocol as described in section 2.9.1. while the High Sensitive (HS) Qubit protocol and reagents were used.

Samples matching the recommended concentration 4 - 12 ng/μl and CV < 0.3 were loaded immediately on to the Bionano Saphyr chip and machine. For samples matching 4 - 12 ng/μl but CV > 0.3, gentle homogenisation was carried out using wide-bore pipette tips according to Bionano guidelines. These samples were then incubated at RT overnight and re-evaluated using HS Qubit. Samples falling outside of 4 - 12 ng/μl with CV < 0.3 were discarded and the DLS process was repeated with a fresh aliquot of 750 ng HMW DNA.

2.9.3 Bionano Saphyr data collection

The DLS DNA was loaded into the Saphyr chip (version G1.2, two samples per chip) and mounted into the Saphyr machine, following the "Saphyr® System User Guide, Document number 30143, revision C". Briefly, a chip was equilibrated to RT for at least 30 min and unsealed for loading. Volumes of 8.5 and 11 μl of DLS DNA were loaded to the inlet and outlet of the chip, respectively. PCR grade water was added dropwise until the in/outlet water level was slightly convex. The chip was sealed immediately with the Saphyr clip and placed on to the arm of the Saphry machine for data collection.

The maximum amount of data (1.5 Tb) was collected when possible. The available amount of DNA data that could be collected each run is limited by many factors, including DNA quality/quantity, DLS efficiency, chip quality, and machine performance. In general, good quality DNA with intact continuous long-molecules will yield the highest amount of data and the best coverage.

2.9.4 Bionano OGM data analysis

All analyses for the Bionano OGM data were performed on the Bionano Access platform (version 1.6.1) with the Bionano Solve software (version 3.6.1_11162020). **Figure 14** shows an outline of the Bionano OGM analysis pipeline from collecting raw data to either generating novel candidate SVs or producing supporting evidence to resolve previously detected SVs. Data from each Bionano run was subjected to quality control (QC) to ensure run metrics met the recommended threshold. Depending on quality, each run was subjected to the *de novo* assembly analysis, ultra-long assembly analysis (molecule length filter + *de novo* assembly analysis), and/or rare variant analysis (RVA). Each pipeline calls a set of SVs, which were filtered and verified to produce SVs of interests.

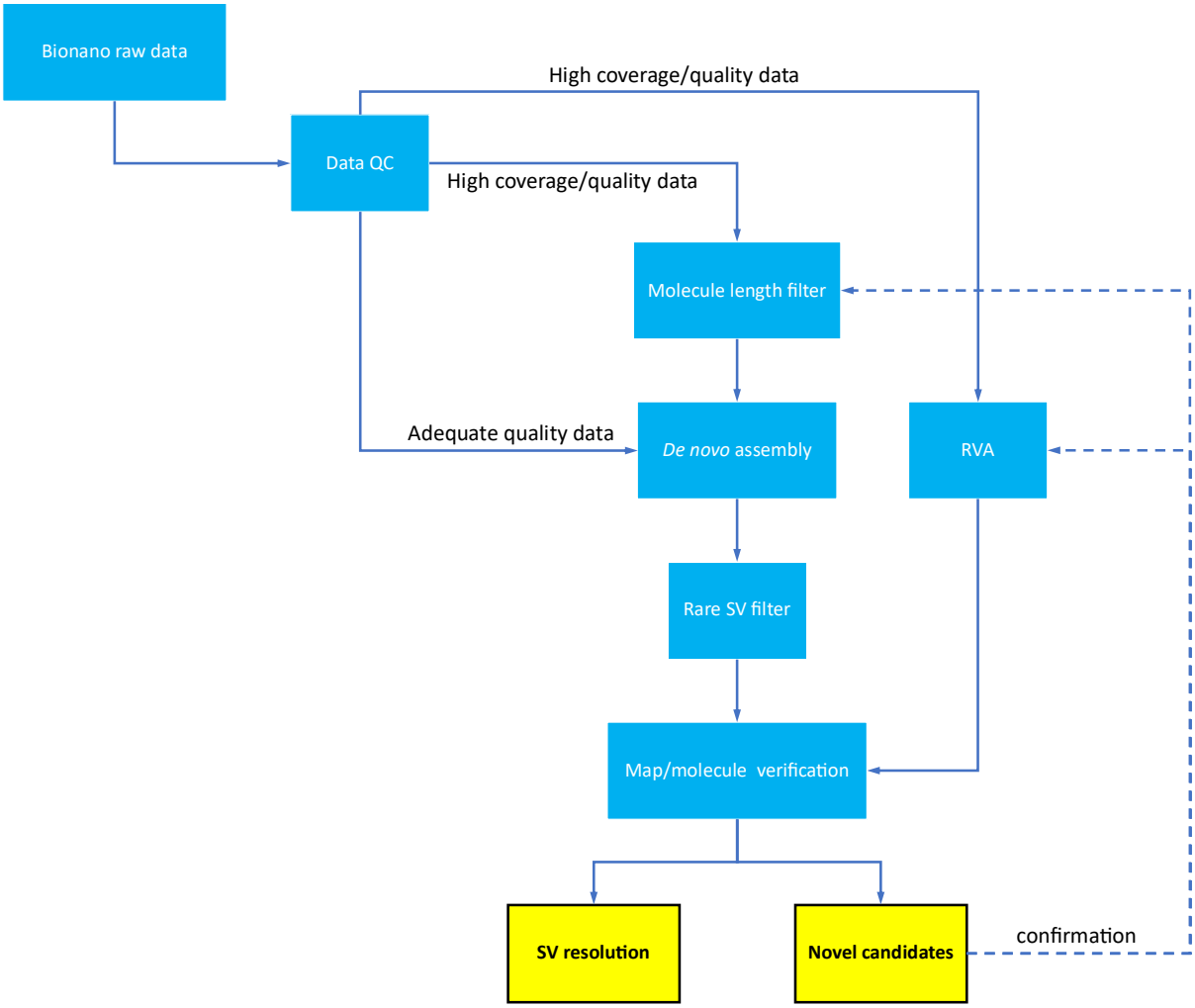


Figure 14 Bionano OGM analysis pipeline. *Data quality was assessed to determine the appropriate analysis route. De novo assembly was performed directly on Adequate quality data (meeting the minimum Bionano recommended threshold). High quality data, depending on required application, were either down-sampled first prioritising long molecules before de novo assembly (ultra-long assembly analysis), or directly fed into the RVA pipeline. The OGM analysis pipeline aimed to generate either novel SV candidates or supporting molecules for previously unsolved SVs. Some large or CPX candidates required further confirmation through re-analysis via either ultra-long assembly or RVA.*

2.9.4.1 Bionano data QC and analysis pipelines

Data metrics of each Bionano run were evaluated against the Bionano recommendation in the "Saphyr Molecule Quality Report Guidelines, Document Number: 30223, Revision: C" and the "Data Collection Guidelines, Document Number: 30173, Revision: E", as summarised in **Table 14**. The data metrics of each run were extracted from the Molecule Quality Report, and then evaluated against the Bionano

recommended metrics to determine the appropriate downstream analysis. High quality data, generally with high map rate and total DNA (coverage), went through the ultra-long assembly analysis and/or RVA when required. Data with Adequate quality went through the native *de novo* assembly analysis as there was usually not enough quality data collected for ultra-long assembly/RVA analysis.

De novo assembly analysis is the native analysis within the Bionano Access platform. *De novo* assembly aligns consensus raw OGM molecules to create long contigs called genome maps. The genome maps are then aligned directly to the reference genome to create a continuous re-assembled human genome. SV calling is performed as part of the pipeline by analysing the inconsistencies between the genome maps and the reference genome.

Ultra-long assembly analysis is a two-step approach aimed to enhance the native *de novo* assembly analysis to be able to correctly call and map large complex SVs. High quality data are required, as the first step of the analysis involves filtering on the raw molecules, which reduces total DNA content. Using the Bionano Molecules Filter function, molecules were filtered down to ~450 kb by prioritising long molecules, usually retaining > 300 kb molecules only. This allows most complex SVs to be called and mapped correctly.

RVA aims to detect low level and/or mosaic SVs. RVA usually requires the maximum amount of DNA data (1.5 Tb capped by the data collection process of the Bionano Saphyr system). RVA extracts only abnormal molecules/maps compared to the

reference, without any other down-sampling. RVA was performed when mosaic SVs were suspected, and when low level ultra-long molecules needed to be extracted as the supporting evidence for resolving large CPX SVs.

Table 14 Run metrics from Bionano recommendations to determine data quality

Metric	Recommendation for *de novo* assembly	Recommendation for RVA	Adequate quality	High quality
N50 (>= 20 kbp)	> 150 kb	> 150 kb	> 150 kb	> 150 kb
N50 (>= 150 kbp)	>230 kb	>230 kb	>230 kb	**~ 300 kb**
Map rate	>70%	>70%	>70%	**~90%**
Total DNA	> 450 Gb	**Max (1.5 Tb)**	> 450 Gb	**> 1 Tb**
Effective coverage	> 80X	**> 340X**	> 80X	**> 200X**
Average label density (>= 150 kb)	14 - 17 /100 kb	14 - 17 /100 kb	14 - 17 /100 kb	14 - 17 /100 kb
Site SD	< 0.25	< 0.25	< 0.25	< 0.25
integrity_num	< 20	< 20	< 20	< 20
Negative label variance (NLV)	< 15.0	< 15.0	< 15.0	< 15.0
Positive label variance (PLV)	< 10.0	< 10.0	< 10.0	< 10.0

Highlighted metrics (in blue) are different from the minimum native de novo assembly threshold.

2.9.4.2 SV filtering

Post-assembly data were analysed on the Bionano interactive viewer of the Bionano Access platform. Two filters were applied to the SV calls: SV mask and the Bionano internal control filter. SV mask is provided by Bionano containing a list of SVs calls in regions with unusually high variance, which are often concentrated around the centromeres and telomeres. SV masked regions are shown in **Supplementary Figure 1**. Molecules mapping in these regions are unreliable and the SV calls using these molecules are likely to be false positives. When applied, SV mask filters out any SV calls matching the ones in the mask.

Bionano also provided an internal control database, which aggregated a list of SV calls detected in 179 healthy individuals from different ethnic origins, as shown in **Supplementary Table 1**. Due to the small size of the control dataset, I applied a stringent filter to the SV calls, by retaining only SVs that had not been matched to any calls in the control dataset. The filter is applied as follows: 0% in both “SV in less than _% of the control samples with the same enzyme” and “SV in less than this _% of the control samples” in the interactive viewer of the Bionano Access platform.

2.9.4.3 *De novo* assembly prioritisation

After the previous filtering steps, an additional prioritisation step may be applied, to further reduce the number of SV calls for functional analysis. Three panels have been used for the prioritisation within the Bionano access viewer: hg38 gencode v34 basic exons (Bionano), a DDG2P (Developmental Disorders panel in the Gene2Phenotype database) panel v2.2 (Genomics England PanelApp, https://nhsgms-panelapp.genomicsengland.co.uk/panels/484/v2.2, last accessed 10.05.2023), and the CRS gene panel described in **section 2.4.5**. However, the prioritisation is not essential as the number of SVs post-filtering is likely to be manageable. Prioritisation was not applied during the comparison between Illumina, ONT, and Bionano to avoid filtering out interesting SVs.

2.10 Illumina WGS vs ONT WGS vs Bionano OGM

Eight trios with high quality data were fully analysed, comparing Illumina WGS, ONT WGS, and Bionano OGM. Bionano OGM analysis was performed only on the proband for most of the trios, as there was WGS data available from both Illumina and ONT.

2.10.1 Bionano reference "truth" callset benchmark

The Bionano rare SV call was set as the reference "truth" set. VCFs were generated from the Bionano Access software, after the rare SV filtering approach (0% match in the Bionano internal control database) described in **section 2.9.4.2**. The validity of each of the rare SV calls was checked with reference to Bionano molecules, ONT reads support, and Illumina WGS reads support. All read support was inspected manually on IGV. A true positive (TP) Bionano SV call was required to show reads/molecules support from at least two out the three technologies. The precision of the Bionano OGM is calculated as shown in **Table 15**. Certain SV type incompatibility was allowed for the precision calculation, specifically INSs/DUPs, CNVs, and complex SVs. This SV type difference was evaluated separately in **section 2.10.2**.

Table 15 Bionano OGM benchmarking error matrix

		Processed Bionano OGM calls
Reads/molecules support from 2/3 technologies	True calls	TP_{OGM}
	False calls	FP_{OGM}
		$Precision_{OGM} = \frac{TP_{OGM}}{TP_{OGM} + FP_{OGM}}$

2.10.2 Bionano SV type normalisation

To enable accurate comparison of the SV types called amongst the three technologies, Bionano SV types were normalised as follow:

CNV gain → DUP/INS

CNV loss → DEL

DUP ↔ INS

Others/complex events → DEL/INS/DUP

DUP/INS conversion was carried out due to Bionano OGM's inability to distinguish certain small-medium sized DUP/INS events. This process is further discussed in detail in **section 6.2**. Complex event normalisation was carried out for specific complex cases discussed in **Chapter 5**, and this process is also discussed in **section 6.2**.

2.10.3 ONT/Illumina sensitivity evaluation

The performance of ONT and Illumina WGS technologies were evaluated using the sensitivity metric as calculated in **Table 16**. A TP SV was deemed detectable via ONT when there was clear read support in the BAM file when examined on IGV. The SV calls were also examined in the annotated Manta & Canvas VCFs, Lumpy VCFs, and Sniffles/Sniffles2 VCFs to determine if the same SV call has been made by the 4 SV callers using data from their respective WGS technologies.

Table 16 ONT/Illumina WGS sensitivity matrix against TP Bionano OGM calls

	Illumina calls/ ONT calls/ ONT read support		
	Positive	Negative	
TP_{OGM}	TP	FN	$Sensitivity = \frac{TP}{TP + FN}$

FN = False negative

2.10.4 ONT *de novo* calling performance evaluation

The performance of *de novo* calling from ONT data was evaluated by calculating the false positive rate (FPR) of Sniffles/Sniffles2 *de novo* calls, as shown in **Table 17**. A TP SV was deemed inherited when the parental data showed reads support in ONT or Illumina WGS reads.

Table 17 ONT de novo calling performance evaluation

	ONT Sniffles/Sniffles2 calls		
	De novo calls (only in proband)	Inherited calls (in both proband and parent)	
False *de novo/ true inherited (positive read support in parents)*	FP	TN	$FPR_{de\ novo} = \frac{FP}{FP + RN}$

TN = true negative; FPR = false positive rate

Chapter 3 Results – 100kGP WGS analysis

3.1 Introduction

The initial hypothesis proposed that part of the diagnostic gap in patients with CRS could be explained by pathogenic SVs that had eluded detection through conventional approaches, and that the incorporation of WGS would unveil some of these missing diagnoses. Consequently, a total of 114 CRS cases were analysed as part of the 100kGP. SVs were called using Manta[36], Canvas[34], and Lumpy[35], followed by aggregation, filtering, prioritisation, and subsequent analysis to identify the most compelling candidate variants (**section 2.4**). A summary list of these compelling candidate SVs is presented in **Table 18**. Three clinically significant SVs (case 11, 32, and 33, marked with * in **Table 18**) were identified prior to my work in the cohort, and these were summarised in Hyder et al (2021).[90]

Among the candidates identified in this project, four were classified as pathogenic or likely pathogenic. However, multiple VUSs remained challenging in advancing the understanding of their molecular pathogenicity. In this chapter, I will provide an in-depth analysis of selected candidate SVs to highlight their potential implications.

Table 18 100kGP cases where a candidate SV was identified through WGS analysis

Case ID	Candidate gene(s)	Genome build	Candidate coordinate	SV size	SV type	Origin	Class	Array	WES	Segregation	Chapter
1	*CERS6*	hg19	chr2:169318436-169436287	118 kb	DEL/CON	DN	SVUS	X	X	X	3 & 5
21	*HOXCs*	hg19	chr12:54374184-54659693	286 kb	DUP	P	LP	X	NA	✓	3 & 4
24	*HDAC2 & COL10A1*	hg38	chr6:113743212-117175387	3.43 Mb	DEL	DN	P	✓	NA	✓	3
25	*GPC3*	hg38	chrX:133528230-133632279	104 kb	DEL	M	P	?	NA	✓	3
27	*ABL1, EXOSC2*, etc.	hg19	chr9:130574132-130769628	195 kb	DUP	M	SVUS	X	X	X	3
28	*NBN, OTUD6B,* etc.	hg38	chr8:85878104-91482989	5.6 Mb	DUP	P	SVUS	✓	X	X	3
29	*ARX* locus	hg38	chrX:24409886-25458349	1 Mb	DUP	?	LP	NA	NA	✓	3
30	*SHOC2*	hg19	chr10:110992774-111177529	184 kb	CPX	M	SVUS	✓	NA	X	3
31	*KIAA0825*	hg38	chr5:94132069-94180877	48 kb	DEL	P	SVUS	?	NA	X	3
10	*FGF9*	hg19	chr13:24428230-24428231	568 kb	CPX-INS	DN	SVUS	X	NA	✓	5
16	*PLCB4*	hg38	chr20:9780422-9780423	1.2 Mb	CPX-INS	M	SVUS	X	Partially	X	5
18	*HOXC* locus	hg38	chr12:53477870-54419319	941 kb	CPX-INS	M	LP	✓	NA	✓	4
23	*KCNJ2 & 12*	hg38	chr17:70413811-70791045	377 kb	DEL	M	P	?	NA	✓	5
26	*HOXCs*	hg38	chr12:53975383-54039954	65 kb	DEL	M	P	?	NA	✓	4
*11**	*FOXD3*	hg38	chr1:63129955-63141504	11.5 kb	DUP	M	P	?	X	✓	-
*32**	*TWIST1*	hg38	chr7:19056374-32575879	~13 Mb	CPX	M	P	?	?	✓	-
*33**	*ERF*	hg38	chr19:41952441-42266625	314 kb	DEL	DN	P	?	?	✓	-

Genome build was selected based on the data availability in the 100kGP, with preference for hg38 when available. Coordinates represent the overall affected genomic region, rather than the precise break points of the SV. SV type: DEL= deletion, DUP = duplication, CON = conversion, INS = insertion, CPX = complex/complex event. Origin column describes the origin of the SV in the proband: DN = de novo, M = maternal, P =

paternal. Array & WES columns indicate if the candidate SV was detected by these two conventional technologies: X = not detected, ✓ = detected, Partially = detected but not fully characterised, NA = not investigated using this technology. Segregation column indicates if the SV segregates with the relevant phenotype in the family: ✓ = segregates, X = does not segregate. Class column describes the determined pathogenicity of the candidate SV: SVUS = structural variant of unknown significance, LP = likely pathogenic, P = pathogenic. Question mark (?) indicates no information was available. Five cases (greyed out text) are findings described in later chapter(s) as they are either additional findings or complex cases requiring further analysis outside of the 100kGP. Cases solved by collaborators prior to this project are indicated with an asterisk.

3.2 Pathogenic and likely pathogenic SVs

From the 100kGP analysis, four additional pathogenic or likely pathogenic SVs (case 21, 24, 25, and 29) were successfully identified. Despite being highly clinically interesting events, three of these SVs were likely known diagnoses prior to the 100kGP analysis and subsequently rediscovered during my analysis of the 100kGP dataset. These cases serve as compelling examples demonstrating the effectiveness of my 100kGP analysis pipeline, while the low yield, four likely/pathogenic SVs from 144 cases, likely contrast the hypothesis that further missing diagnoses involving pathogenic SVs. Alternatively, the missing diagnostic gap may revolve around other candidates, such as complex SVs or elusive single nucleotide variants (SNVs).

Notably, one pathogenic SV involving the *HOXC* cluster (case 21), emerges as the most compelling novel candidate identified via the 100kGP analysis. This finding not only led to the discovery of two additional cases featuring *HOXC* SVs in the 100kGP, but also enabled an in-depth exploration of the potential molecular mechanisms of SVs at the *HOXC* locus and the role of these genes in craniofacial development. Work on these SVs is described separately in **Chapter 4**.

To begin this chapter, two rediscovery cases are discussed. These likely pathogenic SVs should have been detectable through array-based methods, owing to their size and the well-described genes they affect.

3.2.1 Case 24: large chr6 DEL – rediscovery

The first likely pathogenic candidate is from case 24, which is a trio family in the 100kGP with the proband presenting syndromic CRS, and abnormalities of the eye, nose, heart, and kidney. During the initial 100kGP WGS analysis, a 3.4 Mb *de novo* DEL (NC_000006.12:g.113743212_117175387del, hg38) was detected, as shown in **Figure 15**. This DEL includes several disease genes, notably *HDAC2* and *COL10A1*. Furthermore, Interstitial 6q21q22.1 Deletion Syndrome (region overlaps with the case 24 DEL) has been well characterised and has a variable syndromic phenotype[96] comparable to that of the case 24 proband. Taken together with the *de novo* nature of this DEL, this is highly likely to be the causative SV.

Interestingly, due to its size and location, this DEL should have been detectable by an array-based method, which would typically preclude the recruitment of this patient to the 100kGP. Indeed, through the Contact Clinician request in the 100kGP, it was revealed that this DEL had been previously detected using an array-based technique. Notably, no information on the prior identification of this deletion was available within the 100kGP research environment.

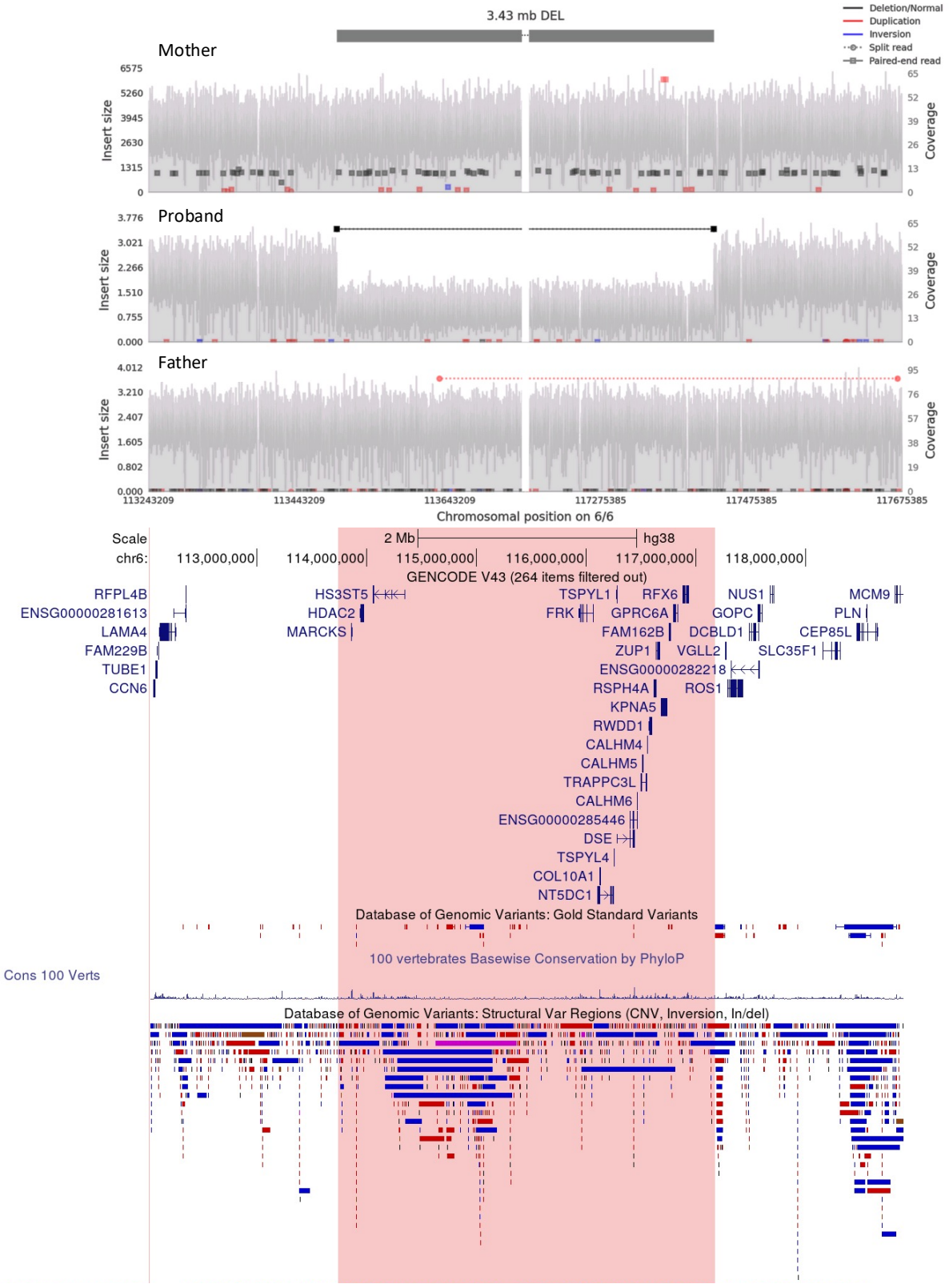


Figure 15 A large 3.4 Mb de novo DEL was identified on chr6 in the case 24 proband. *The DEL affects several genes of interest such as HDAC2 and COL10A1.* ***The upper panel*** *of the figure contains three sequence tracks from Samplot created using Illumina WGS data. A de novo deletion is shown, which can be observed in the proband's track from both the coverage (depth) drop over the 3.43 Mb region as well as the presence of abnormally linked*

paired/split reads (black line with square ends). Note that this particular Samplot figure is in split-view mode, which excludes a large central portion of the 3.4 Mb region to increase the readability of ultra-long SVs. ***The lower panel*** *is the respective region in the USCS genome browser, with tracks from top to bottom indicating: the scale bar, genomic coordinate, GENCODE genes, DGV Gold Standard SVs, PhyloP Conservation, and DGV all SVs. The two panels are aligned by genomic coordinates. Samplot figure exported from GE Airlock. Figure in hg38.*

3.2.2 Case 25: DEL affecting *GPC3* - rediscovery

A pathogenic candidate identified in case 25, which is a non-classic trio in the 100kGP, was also a rediscovery. This trio include the affected proband, the unaffected mother, and the unaffected maternal aunt, as shown in **Figure 16a**. The proband presented syndromic CRS with kidney and aorta abnormalities. In contrast to case 24, initial analysis of case 25 did not reveal any interesting candidates.

Subsequently, analysis focusing on non-segregating SVs identified a 104 kb DEL in all three family members, as shown in **Figure 16b**. The DEL affects the last two exons of *GPC3* (Glypican-3), a crucial regulator of both Wnt/hedgehog signaling pathways[97,98] and a known CRS gene associated with the X-linked condition - Simpson-Golabi-Behmel syndrome (SGB, OMIM312870).[99] Partial or whole gene deletions of *GPC3* have been well documented in the literature in SGB[100], supporting this DEL as the likely causative variant in case 25. Despite its smaller size compared to the case 24 DEL, further information was obtained from the clinician indicating this deletion was a known diagnosis; a different clinician, seemingly unaware of this fact, had undertaken the referral to 100kGP.

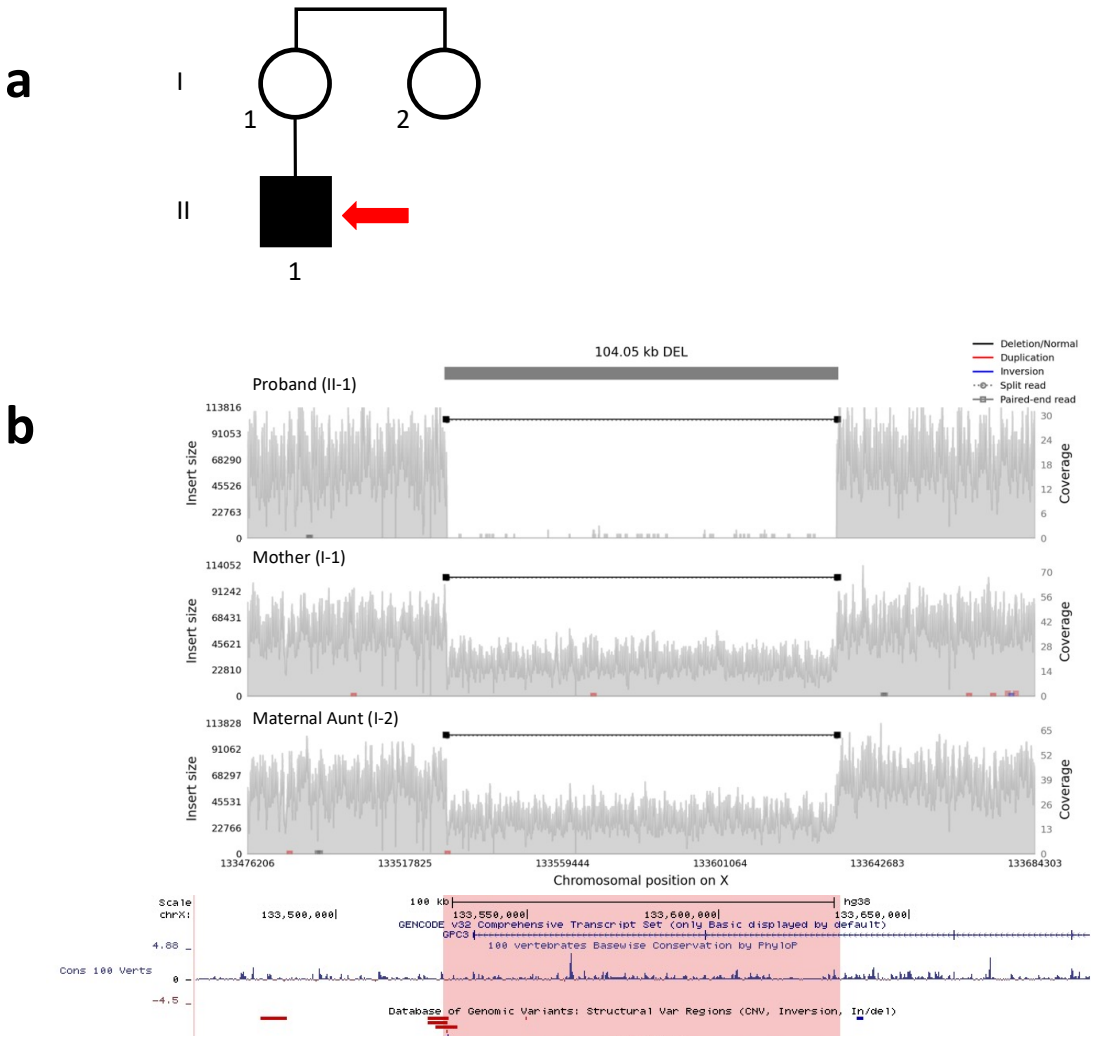


Figure 16 Hemizygous DEL affecting GPC3 in the proband. ***a***. *Only the affected proband (II-1), the unaffected mother (I-1), and the unaffected maternal aunt (I-2) were recruited to the 100kGP.* ***b***. *The hemizygous 104 kb DEL was identified in the proband, while the mother and the maternal aunt are unaffected heterozygous carriers. Samplot figure exported from GE Airlock. Figure in hg38.*

3.2.3 Case 29: DUP at *ARX* locus – novel diagnosis

The Case 29 SV was a novel finding in my analysis, and the SV is positioned in a well-documented locus associated with Intellectual Disability (ID). This SV demonstrated the difficulties in analysing singletons within the context of rare disease in the 100kGP.

Case 29 was recruited as a male singleton to the 100kGP due to global developmental delay and bicoronal & sagittal CRS. Parental information was not available. No SVs of interest were identified using the original WGS analysis pipeline (**section 2.4**), but in the later SVRare analysis (**section 2.4.12**), one interesting SV was detected, a 1.05 Mb DUP on chrX spanning the whole *ARX* gene and the last four exons of *POLA1*, as shown in **Figure 17**. This SV was identified by examining unique calls in SVRare (N_all = 1) affecting any genes in the DDG2P panel (Genomics England PanelApp). Detailed read analysis revealed a clean break junction suggesting a simple tandem DUP as NC_000023.11:g.24409886_25458349dup (hg38).

DUPs at the *ARX* locus have been extensively studied and linked to a spectrum of neurodevelopmental phenotypes, including ID, speech delay, hypotonia, and psychiatric abnormalities.[101] The *ARX* DUP phenotypes are also highly heterogeneous, ranging from normal to severe cognitive impairment depending on the size (41.1 kb - 3 Mb reported) and position of the DUPs. These previous cases suggest that case 29 is highly likely part of the *ARX* phenotypic spectrum, with the identified DUP being the underlying causative variant.

Further feedback obtained via the 100kGP Contact Clinician form suggested that this was in fact a novel diagnosis. A previous sub-telomeric screen and karyotype on the patient (prior to availability of array technology) had been normal. There was also an extended family history (male and female siblings with developmental delay and sagittal synostosis or scaphocephaly, offspring of a maternal half-brother), while array of this affected male family member was negative. Unfortunately, no DNA samples

were available for PCR confirmation and no further updates were obtained from the clinician subsequently.

It is important to recognise that this DUP eluded detection in the original WGS analysis pipeline. This was primarily due to the inherent challenges posed by singletons, which are difficult to analyse when the control group is relatively small, as was the case when only the CRS cohort was analysed. In contrast, SVRare uses the entire 100kGP rare disease cohort as a control group, significantly enhancing its analytical power in filtering through singleton SV calls in the absence of parental data.

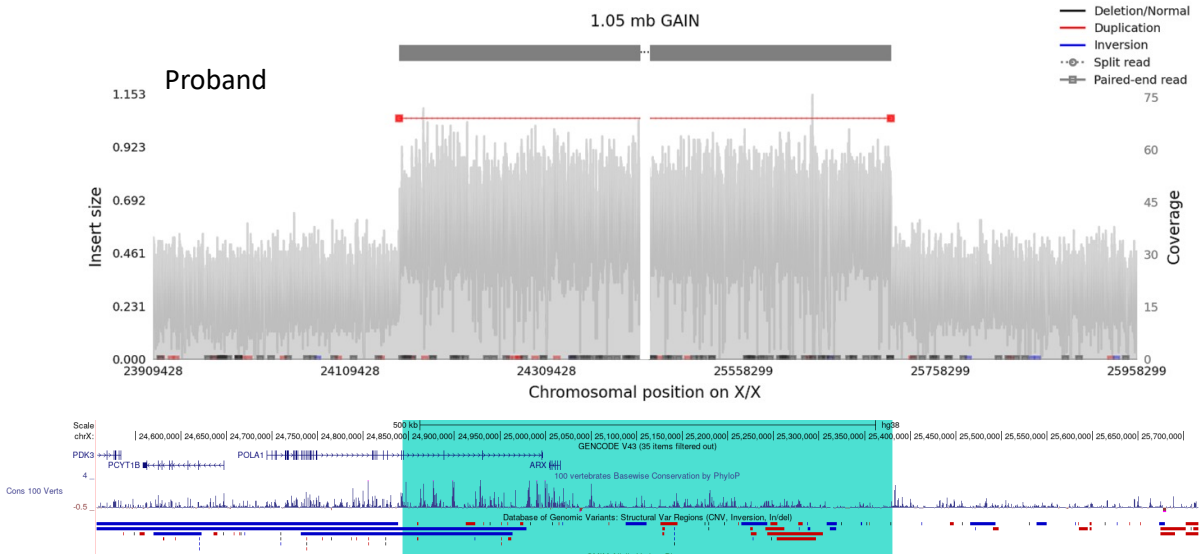


Figure 17 A DUP detected by SVRare at the ARX locus*. The DUP is ~1 Mb and includes the entire ARX gene and the last four exons of POLA1. Samplot figure exported from GE Airlock. Figure in hg38*

3.2.4 Case 21: DUP at *HOXC* gene cluster – novel diagnosis

The Case 21 SV represents a novel finding which led to the potentially new association of *HOXC* genes with craniofacial abnormalities. Subsequent detailed analysis post-100kGP is explored in **Chapter 4**.

Initially, from analysing non-segregating SVs, an inherited DUP was identified in the trio, as shown in **Figure 18**. Only the proband was annotated as affected in the 100kGP, marking this finding as a non-segregating SV. However, upon further exploration of the Human Phonotype Ontology (HPO) terms, it was noted that the father shared part of the syndromic features found in the proband, suggesting that the father is likely also affected (detailed in **Chapter 4**). This further strengthened the interest in this inherited SV.

From the coverage information in Samplot (**Figure 4**), it can be observed that the coverage depth for the father's DUP appears less pronounced than the child's DUP, that is, below the 1.5x coverage increase expected for a constitutional tandem DUP. This discrepancy raised the possibility of paternal mosaicism. The mosaic nature of the paternal DUP was confirmed via mosaicism analysis (described in **section 2.4.10)**, which demonstrated a lower-than-expected number of reads (<150%) in the paternal DUP region compared to the proband. This finding correlates well with the father's milder phenotype without CRS.

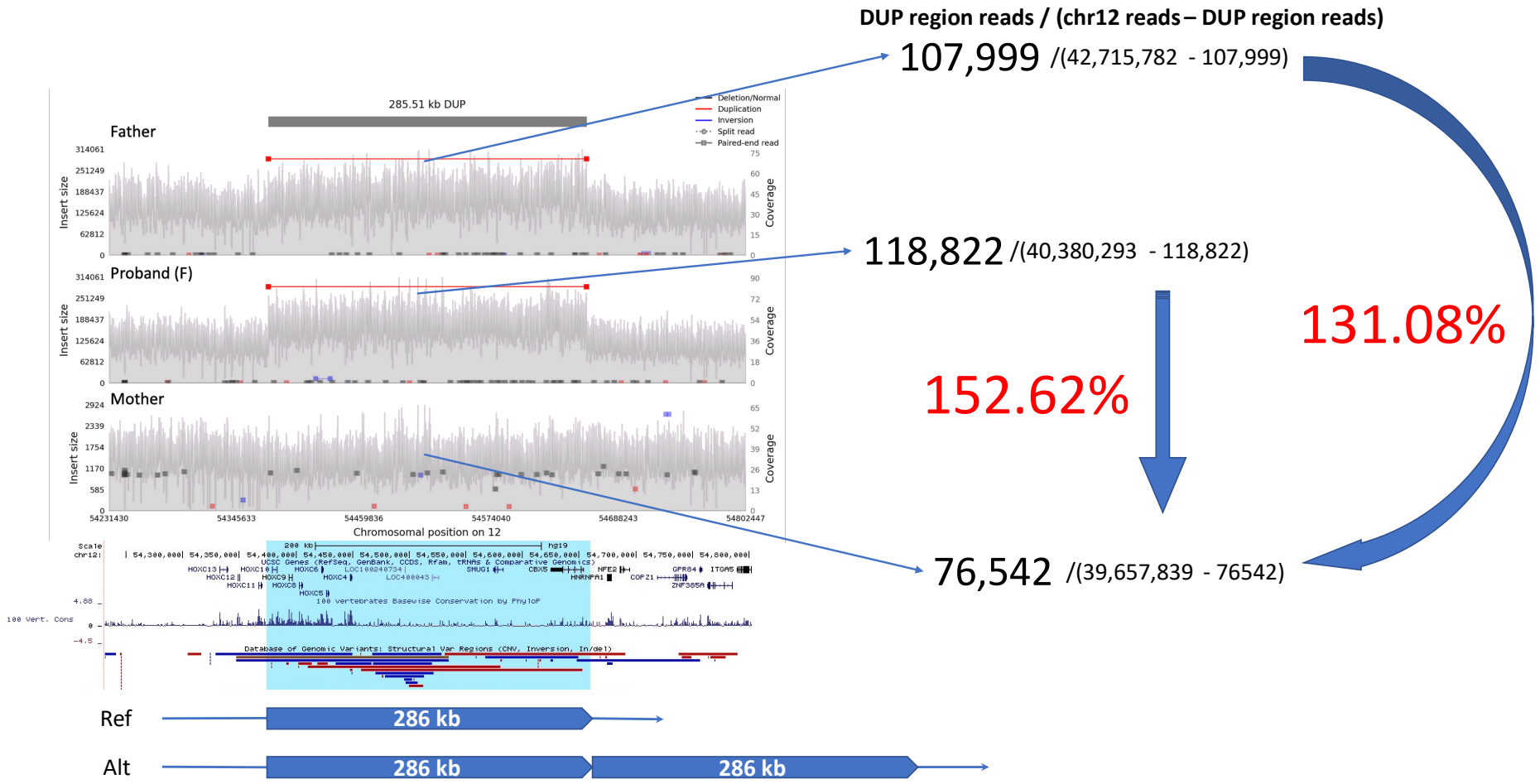


Figure 18 A paternally inherited 286 kb tandem DUP was identified in the case 21 proband. *The DUP is located at the HOXC cluster. Read depth analysis demonstrated that the SV is likely mosaic in the father. Samplot figure exported from GE Airlock. Figure in hg19.*

This finding stands out as the most compelling novel SV discovery in the CRS patients sequenced in the 100kGP, given that abnormalities of *HOXC* genes had not been associated with craniofacial abnormalities prior to this research. The finding led to a detailed functional study aimed at understanding the molecular mechanism of *HOXC* cluster function and regulation. This further analysis led to the identification of two additional cases with SVs affecting the *HOXC* cluster in the 100kGP. These are explored comprehensively in **Chapter 4**.

3.3 Structural Variants of Unknown Significance (SVUS)

In addition to the pathogenic or likely pathogenic SVs, a major finding arising from the 100kGP WGS analysis was the identification of several SVUS. These SVUS consist of SVs for which the functional significance and clinical implications remain unclear due to several factors, including a lack of functional support, a lack of segregation with phenotype in the family, or a combination of both. In this section, some example cases are discussed to elucidate each of the factors.

Two complex SVUSs (Case 10 & 16) could not be fully characterised using WGS data alone. To better understand these two complex SVs, Bionano OGM was employed. The detailed analysis of these two complex SVUS is discussed in **Chapter 5** along with other cases requiring Bionano OGM.

3.3.1 Case 1: chr2 DEL affecting *CERS6*

Case 1 SV is an example of a SVUS owing to the lack of functional support for the candidate gene, despite the *de novo* SV initially segregating with the phenotype. This

family was submitted to 100kGP as a trio comprising a sporadically affected child and unaffected parents, with no other information provided about the family history in the 100kGP RE.

WGS analysis identified a ~120 kb *de novo* DEL as shown in **Figure 19b**. The *de novo* nature of the DEL also matches the apparent sporadic phenotypes in the family. Importantly, no DELs of a similar size were observed in DGV or gnomAD SV, making this *de novo* SV a rare event in the healthy population.

As the family was local to Oxford, it was possible to obtain further phenotypic information and access DNA samples for analysis. The pedigree (**Figure 19a**) reveals that the proband II-3 has a similarly affected monozygotic twin sister II-2; both sisters presented with syndromic multi-suture CRS and Chiari malformation. It was therefore critical to demonstrate that the DEL was also carried by the twin sister. A breakpoint PCR was designed (**Table 8**) and performed, validating its presence in both twins as shown in **Figure 19e**.

Subsequent characterisation of the break junction using dideoxy-sequencing revealed a small 19 bp DEL located 42 bp upstream of the large ~120 kb DEL, as shown in **Figure 19c**. Mechanistically, I hypothesised that the small DEL triggered the formation of a secondary structural loop around the 19 bp sequence, leading to a DSB (**Figure 19g)**. This DSB may subsequently have been repaired likely through a FoSTeS mechanism, resulting in the large ~120 kb DEL, as evidenced by the microhomology at the break junction. This event, including both DELs, can therefore be categorised

as either a delins or a conversion (CON). Since CON holds priority over delins[102], this SV is described as:

NC_000002.11:g.169318436_169436287con169318455_169318499.

However, after re-evaluating the pedigree and analysis of the functional significance, this SV was downgraded to an SVUS for the CRS phenotypes in case 1. This re-classification was based both on lack of segregation to an additional affected family member, and the lack of clear functional relevance of the involved gene. Further clinical contact revealed that the older sister, II-1 (**Figure 5a**), has bilateral squamo-parietal synostosis and growth deficiency. The occurrence of this rare type of CRS suggests that she may have the same genetic basis for her disorder as her twin sisters, despite having a slightly milder phenotype. Crucially, II-1 does not share the same SV event, as demonstrated by the lack of PCR product using the same breakpoint primers shown in **Figure 19e**. When considering all three affected children in the family, a recessively inherited variant is more likely to be the underline genetic cause, rather than (as initially considered) a *de novo* mutation.

Functionally, there is little to no evidence from the literature to support the pathogenicity of a heterozygous *CERS6* loss-of-function SV in craniofacial abnormalities. *CERS6,* coding for ceramide synthase 6, is responsible for *de novo* ceramide synthesis[103], which is primarily involved in membrane lipid production and lipid metabolism. Currently no clinical disorder has been ascribed to alterations in this gene.

Furthermore, the case 1 SV is predicted to result in a loss-of-function copy of *CERS6* by removing the second and third exons of the canonical *CERS6* (**Figure 19c & d)**. However, *CERS6* has a pLI = 0.04, suggesting that it is likely to be haplosufficient and unlikely affected by a heterozygous loss-of-function SV. Similarly in an animal model, *Cers6*-knockout mice (het) show a reduced sphingolipid production, hind limbs clasping abnormality, and habituation deficit, with no skeletal/CRS related abnormalities [104]. Overall, the chr2 SV is likely an incidental finding of unknown significance to the CRS phenotypes observed in the case 1 family. Follow up work seeking recessive genetic causes has been planned for case 1 family, including trio-ONT sequencing.

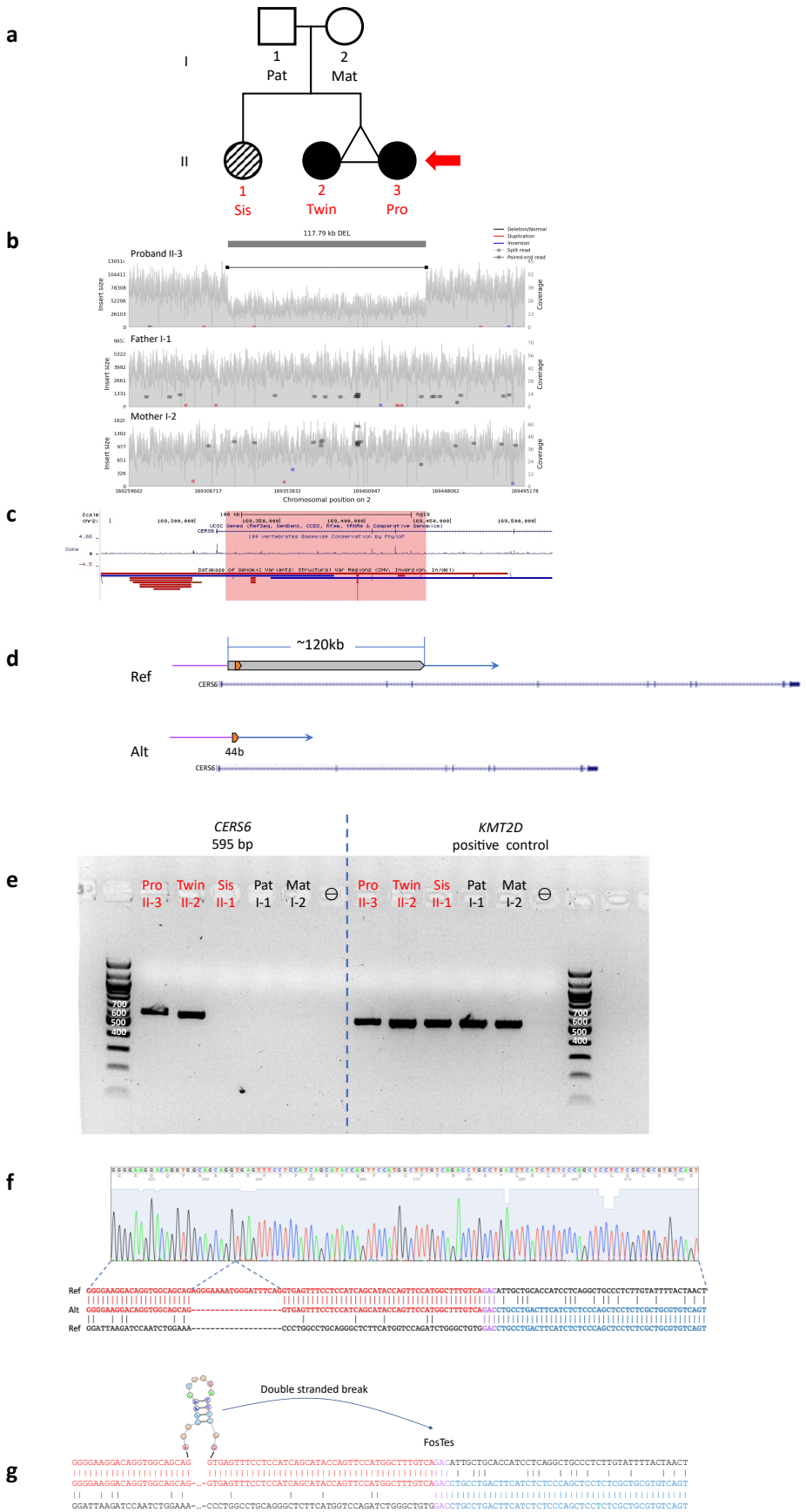


Figure 19 Case 1 is a family with three affected individuals and a chr6 de novo SV in CERS6 is a candidate causative event. a. *Three family members are affected including the*

monozygotic twins (II-2 & II-3) and an older sister (II-1) with slightly different clinical features. The parents (I-1 & I-2) are unaffected. ***b.*** *The SV was initially detected as a de novo DEL in the proband of the trio recruited to the 100kGP and is visualised using Samplot.* ***c.*** *The SV is located at the CERS6 locus, and there are no similar sized DELs recorded in DGV or gnomAD (not shown).* ***d.*** *This SV event replaces the ~120 kb region with a 44 bp region, which results in the removal of the second and the third exons of the canonical CERS6.* ***e.*** *Breakpoint PCR confirmed that the de novo SV is present in both twins (II-2 & 3), but not the older sister (II-1).* ***f.*** *Dideoxy-sequencing validated and further confirmed a 19 bp DEL immediately upstream of the ~120 kb DEL.* ***g.*** *The small DEL is hypothesised to have initiated the SV event by creating a secondary loop structure, causing a DSB, which was subsequently repaired via a FoSTeS mechanism evidenced by the microhomology at the break junction of the large DEL. ⊖ = negative control/ water. Samplot figure exported from GE Airlock. Figure in hg19.*

3.3.2 Case 31: small DEL affecting *KIAA0825*

Case 31 is an example where initial WGS data suggested a possible *de novo* SV affecting a known disease gene, however subsequent investigation indicated that the SV was non-segregating and should be classified as a SVUS. This case highlights the significance of having complete familial information in investigating the underlying cause of a rare phenotype.

Case 31 is both a 100kGP participant and a local case, with the proband presenting syndromic multi-suture CRS. In the 100kGP, only the proband and the mother were recruited and underwent WGS. The 100kGP analysis identified a ~49 kb DEL in the proband and not the mother (**Figure 20a**), suggesting possible *de novo* origin of the DEL. The SV was also interesting functionally, as it affects the last exon and the 3' UTR of gene *KIAA0825*, which is a known recessive disease gene for postaxial Polydactyly type A10 (OMIM 618498).[105] To verify the *de novo* origin of the DEL, breakpoint PCR was performed on the father using locally available genomic DNA. As shown in **Figure 20**, the DEL was confirmed to be inherited from the unaffected father.

Dideoxy-sequencing was performed to verify the breakpoint sequences. Using HGVS nomenclature, this event was determined as:

NC_000005.10:g.94132070_94180876delinsCAAA. Efforts to identify a second hit SNP were undertaken, with no success.

Overall, this DEL was classified as an SVUS due to the identification of an unaffected carrier despite affecting a known disease gene. This case highlights the importance of considering complete familial information when interpreting genetic findings, which can be challenging in the 100kGP dataset, as evidenced by a significant number of singletons and incomplete trios. Follow up work for case 31 may include OGM and ONT analysis pending on sample availability.

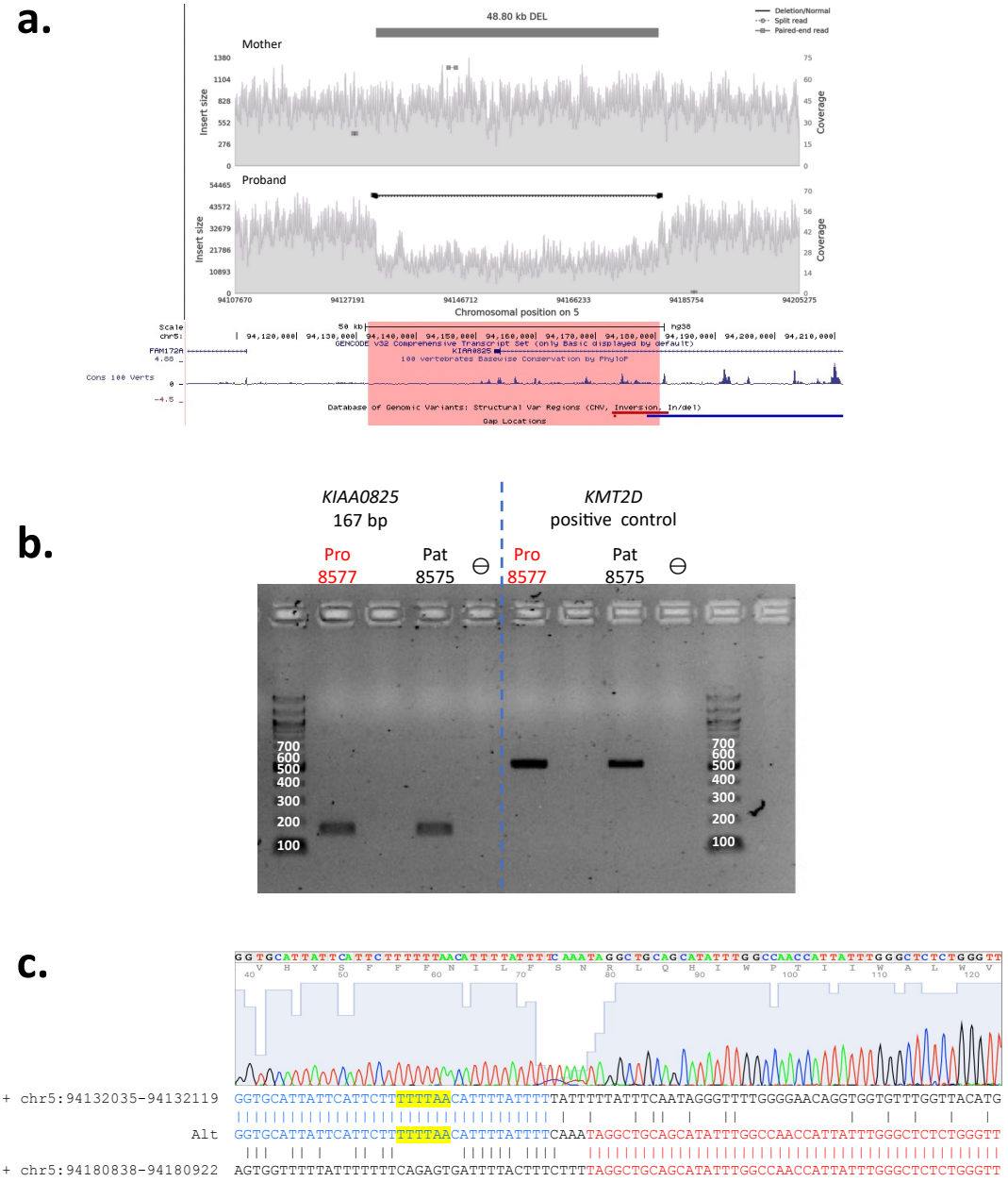


Figure 20 Chr5 DEL affecting the last exon of KIAA0825 – a known recessive disease gene for polydactyly. a*. WGS analysis identified the DEL affecting the last exon of KIAA0825 in the proband but not in the mother within the 100kGP.* ***b****. Breakpoint PCR confirmed the DEL is inherited from the father.* ***c.*** *Dideoxy-sequencing illustrates the "TTTTAA" sequence at the break, suggesting a DSB induced SV. ⊖ = negative control/ water. Samplot figure exported from GE Airlock. Figure in hg38.*

3.3.3 Case 27: chr9 DUP affecting *ABL1* locus

Case 27 is an example of a non-segregating SV classified as SVUS despite affecting several highly interesting disease genes. This case illustrates the challenges associated with explaining the unaffected carrier of a functionally likely pathogenic SV.

Case 27 involves a sporadic parent-child trio in the 100kGP with the sporadically affected proband presenting syndromic multi-suture CRS. The proband was documented to have capillary haemangiomas, supernumerary nipple, and both sagittal and right coronal CRS. Analysis of 100kGP data identified a 195 kb tandem DUP, which was inherited from the unaffected mother, as shown in **Figure 21a**. The DUP spans *FUBP3, PRDM12*, and the shorter isoforms of *ABL1*. The long canonical form of *ABL1* is only partially spanned by the DUP. Breakpoint PCR (**Figure 21b**) and dideoxy- sequencing (**Figure 21c**) from locally available samples further characterised the DUP as NC_000009.12:g.130574132_130769628dup. The break junction sequences showed an extensive homology due to the SINE repeats AluSq2 & FLAM_C, suggesting an Alu-mediated NAHR mechanism contributed to the DUP formation.

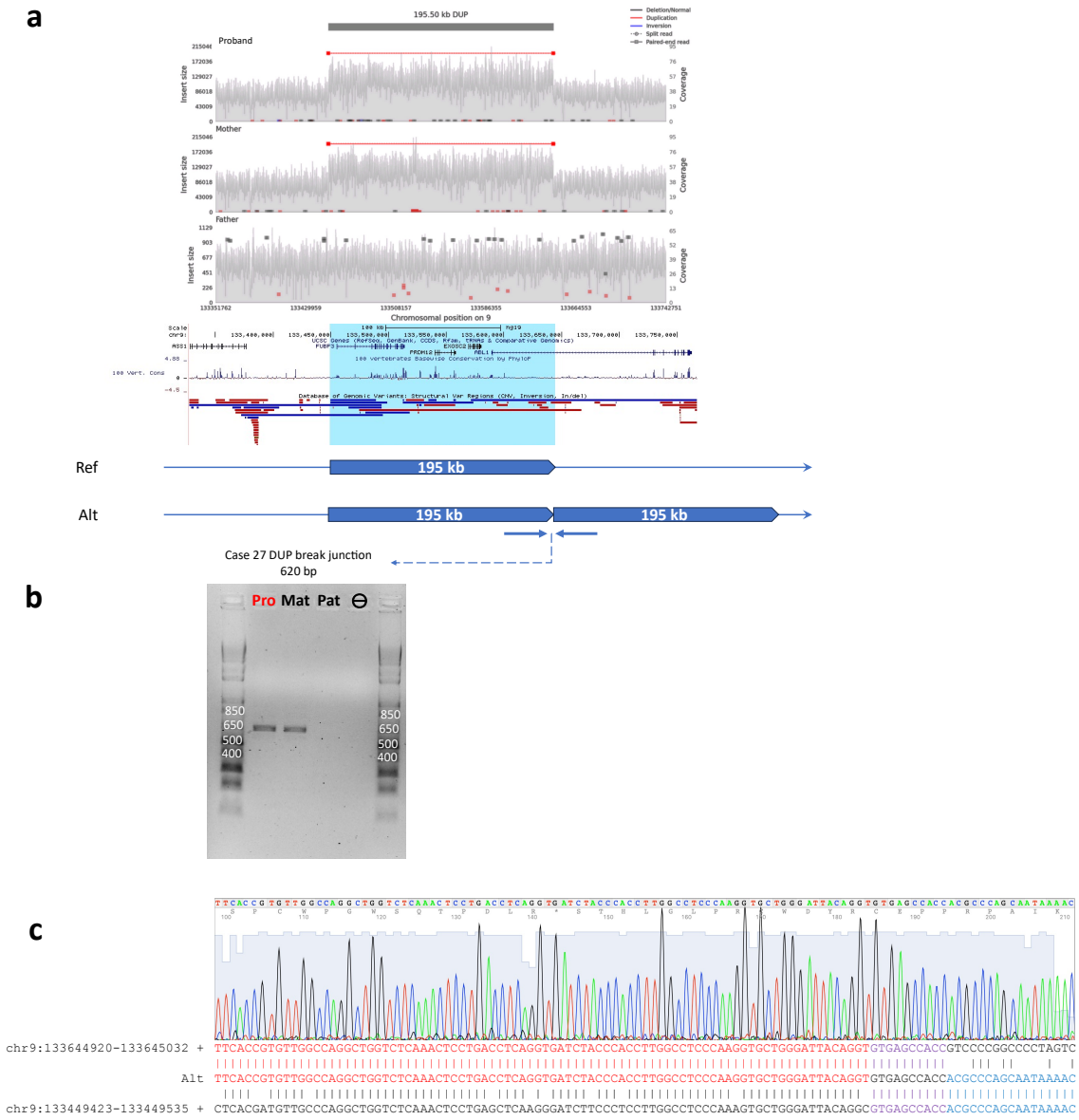


Figure 21 Chr9 tandem DUP affecting ABL1, FUBP3, EXOSC2, and PRDM12. a. *Analysis of WGS data detected a 195 kb tandem DUP. Using breakpoint PCR (**b**) and dideoxy-sequencing (**c**), the tandem DUP was confirmed to be inherited from the unaffected mother and the break point sequence showed an extensive homology, suggesting a NAHR-based mechanism. ⊖ = negative control/ water. Samplot figure exported from GE Airlock. Figure in hg19.*

Further information was sought from the referring clinician via the 100kGP contact clinician form. This revealed that the DUP was previously detected by array (SNP array Human OmniEurHD [2M markers]) but disregarded as it was inherited from the

unaffected mother. Further efforts to explain the mother's unaffected carrier status included mosaic analysis, but this did not yield any evidence of maternal mosaicism.

Functionally, the DUP directly affected four genes, including *EXOSC2*, *FUBP3*, *PRDM12*, and *ABL1*. Interestingly, three of these four genes are known disease genes: *EXOSC2* is linked to a recessive condition involving short stature, hearing loss, retinitis pigmentosa, and distinctive facies (OMIM 602238)[106]; *PRDM12* is linked to recessive hereditary sensory and autonomic neuropathy (OMIM 616488)[107]; *ABL1* is a known gene for a dominant condition of congenital heart defects and skeletal malformations syndrome (OMIM 617602)[108]. However, the case 27 proband did not present clinical features resembling the phenotypes typically associated with these disease conditions. The incompatibility between the candidate genes and the proband phenotype makes the clinical interpretation of this SV exceedingly challenging.

Overall, this tandem DUP was ultimately classified as SVUS, due to the lack of segregation, lack of gene disruption, and inconsistent phenotypes in relation to the affected disease genes. This means case 27 remains without a genetic diagnosis to date, despite extensive genetic investigations involving multiple analysis approaches.

3.4 Summary

This chapter presents the key findings from the search for causative SVs using the WGS analysis carried out on the 100kGP CRS cohort, which represent a current genetic diagnostic gap for patients with CRS. Our initial hypothesis suggested that pathogenic SVs may account for some of the missing genetic diagnosis in CRS

patients, and that WGS could offer an enhanced capability for detecting these pathogenic SVs compared to conventional diagnostic approaches. Three pathogenic SVs (affecting *TWIST1*, *ERF*, and a non-coding region of 1p31.3 located close to *FOXD3*) were already known to be present in the 100kGP WGS data at the outset of this work.[90]

However, the results were somewhat disappointing, with only four further pathogenic SVs identified, and two of these SVs are previously known diagnoses. The two known diagnoses, case 24 (3.4 Mb DEL) and case 25 (DEL affecting *GPC3*), both had already been extensively documented in the literature with similar SVs at the respective loci. Notably, the case 24 DEL, due to its large size, should have been easily detectable through array-based methods, suggesting that this case originally may not have been an intended target for 100kGP recruitment. These cases potentially highlight the challenges encountered during the participant recruitment phase of the 100kGP, such as the need for thorough screening out of cases with known diagnosis.

Two cases, case 21 and 29, obtained novel genetic diagnosis following the identification of pathogenic SVs. Case 29 DUP, at ~1 Mb in size, is likely detectable via conventional methods. Additionally, the affected *ARX* locus is well-documented in the literature, providing evidence to establish the DUP-phenotype connection. However, the DUP remained undetected as case 29 had not been investigated using an array, which was yet carried out when this case was initially assessed. Furthermore, as a singleton, this case highlighted the challenges when analysing WGS data in a rare phenotype without any familial information. On the other hand, the case 21 DUP,

the most compelling novel SV in the 100kGP CRS cohort, had eluded previous detection from array as well as targeted analysis of *HOX* genes – this family is discussed in detail in **Chapter 4**.

Putting all these cases together, the overall prevalence of pathogenic SVs in the 100kGP cohort is 7/114 (6.1%, 95% confidence interval [CI] 2.5% - 12.2%). However, it can be argued that three of the families (involving the 1p31.3 DUP, Chr6 DEL, and *GPC3* DEL rearrangements) should not have originally been enrolled into 100kGP, as the underlying pathology was in fact already known. Excluding these cases, the proportion of causative SVs solved by WGS reduces to 4/111 (3.6%, 95% CI 1.0% - 9.0%) – an unequivocal, but relatively modest yield.

Additionally, the 100kGP analysis yielded several SVUSs. These SVs showed promise of being causative in certain aspects but ultimately failed to demonstrate their clinical relevance due to two main reasons: a lack of segregation (cases 27, 28, 30 and 31) and the lack of functional relevance (case 1).

Two additional SVUSs identified in the 100kGP data (from cases 10 and 16, **Table 1**) have not been presented in this chapter. Both involved complex SVs with multiple break junctions, meaning they could not be fully characterised using WGS alone. Consequently, clinical interpretation for these two complex SVs could not be carried out within the framework of the available 100kGP data. To address this challenge, Bionano OGM was employed for further characterisation. Alongside the available 100kGP and local cases, Bionano OGM was also utilised to identify further possible

missing diagnoses, such as large complex SV eluding even sequencing-based technologies. The detailed Bionano OGM cases are discussed in **Chapter 5**.

Chapter 4 SVs involving the *HOXC* gene cluster at 12q

4.1 Introduction

As described in the previous chapter, one of the most interesting findings from my SV/CNV analysis in the 100kGP CRS cohort was the identification of a 286 kb tandem DUP on chromosome 12, that included six members of the *HOXC* gene cluster, together with the adjacent genes *SMUG1* and *CBX5*. A *de novo* origin in the more mildly affected father (phenotypic information from the RE) was suggested from CN analysis that indicated apparent mosaicism for the DUP identified in the father of case 21. A *de novo* origin, together with the absence of similar DUPs in DGV or gnomAD, suggested that this DUP could be the cause of the associated phenotype and that this might be related to disturbance of gene expression of members of the *HOXC* cluster, one of four *HOX* gene clusters that demonstrate collinear gene regulation in the human genome. In this chapter, I describe the further exploration of this hypothesis. I will (1) present the further molecular characterisation and phenotype in the family and (2) the identification of two additional *HOXC* rearrangements identified in the 100kGP data, one of which involved a larger duplication of the entire *HOXC* cluster. To further analyse this larger DUP, I used Bionano OGM, a methodology that is explored further in **Chapter 5**.

To understand the molecular and developmental context of this work, I will first review some of the remarkable properties of *HOX* gene clusters in development.

4.2 Evolution and biology of *HOX* gene clusters

HOXs are a large family of transcription factors containing a homeobox DNA binding domain.[109] The human genome contains four paralogous *HOX* clusters (A – D), with

each cluster accommodating between nine to 11 *HOX* genes[110], as shown in **Figure 22**. The four clusters are believed to have arisen from two sequential whole gene DUP events during early evolution of vertebrates, and contribute to the increased morphological complexity between species.[111-113] This concept has been well explored in the literature, providing a broad understanding of the underlying mechanisms to biodiversity.

HOX transcription factors play a key role in body segmentation along the cranio-caudal axis during embryonic development, and exhibit both spatial and temporal collinearities[114], as summarised in **Figure 22**. Spatial collinearity refers to the strong correlation between the chromosomal position of the *HOX* genes and their roles in each body segment along the cranio-caudal axis, while temporal collinearity describes the synchronisation between *HOX* gene position and the sequential activation of each *HOX* gene during development.[114] Due to the importance of the collinearities, *HOX* clusters in vertebrates are highly conserved to preserve the crucial elements of their genomic structures, such as gene order and intergenic distance.[115]

The *HOXC* cluster is one of the smaller *HOX* clusters containing nine *HOX* paralogues (*HOXC4-6* & *HOXC8-13*). Interestingly, the three paralogues (numbered 1-3) that show the most anterior (cranial) expression in the *HOXA*, *-B* and *–D* clusters, are absent in the *HOXC* cluster. Mutations in the *HOXC* cluster have been implicated in several human conditions, including Ectodermal dysplasia 9 (OMIM 614931)[116], arthrogryposis[117], and lower extremity malformations[118]. Notably, these conditions have originated from heterozygous loss-of-function mutations of the *HOXCs,* such as

deleterious SNVs (in *HOXA1*, *2*, *11*, and 13; *HOXB1* and *13*; *HOXC13*; *HOXD4*, *10*, and *13*) and CNV loss.[119] Meanwhile, the potential clinical consequences of *HOXC* DUPs and CNV gains in human diseases remain to be thoroughly investigated.

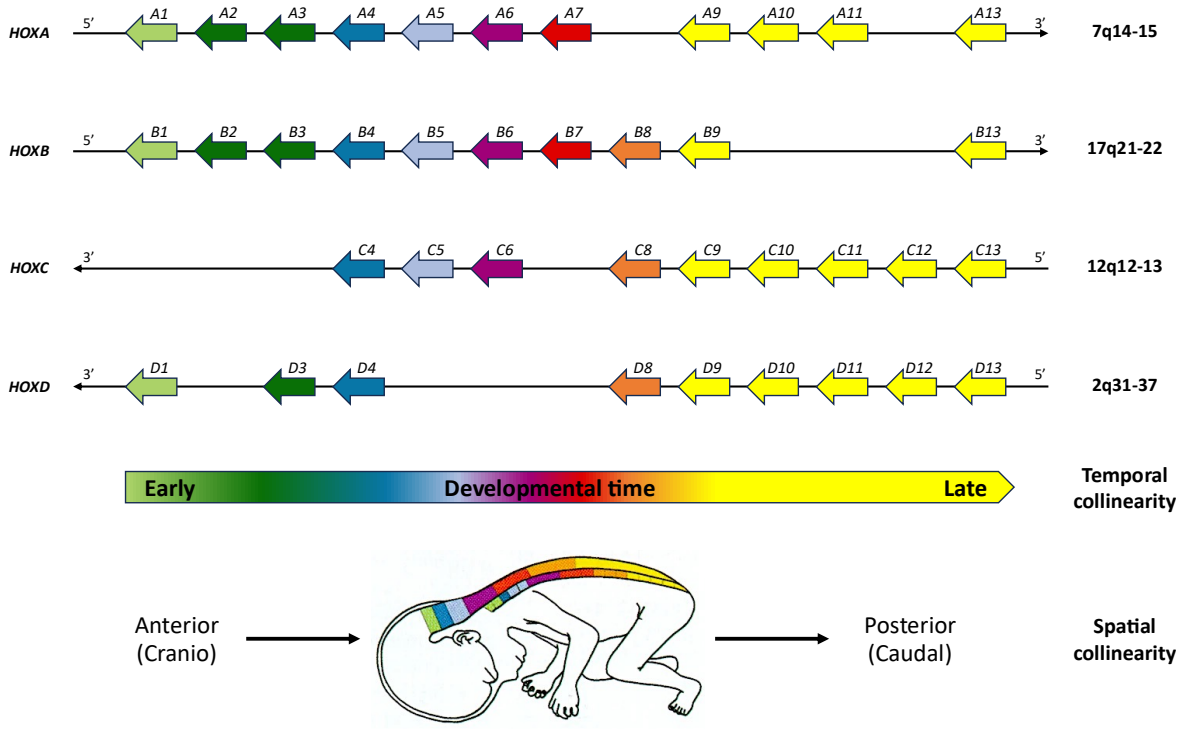


Figure 22 Four human HOX clusters in the human genome with a total of 39 HOX genes. *HOX genes exhibit both temporal and spatial collinearity in relation to developmental time and the cranio-caudal axis. Figure generated using data from the following sources: human foetus illustration and overall colour scheme adapted from Mark et al (1997)[120]; gene directionality from Hueber et al (2010)[110]; chromosomal positions from Lappin et al (2006)[121]; spatial and temporal collinearities from Afzal & Krumlauf (2022)[122].*

4.3 Case 21: *HOXC* DUP in craniofacial abnormality

Case 21 was the first case in which an SV affecting the *HOXC* cluster was identified. Following the identification of the rearrangement in the 100kGP data (**section 3.2.4**), a Contact Clinician request was completed within the GE RE. This led to further clinical information becoming available; it transpired that the family had already been recruited to the *Genetic Basis of Craniofacial Malformations* (GBoCM) study, and DNA samples

were already available in the lab for molecular analysis. In addition, a second affected daughter had been born (**Figure 23**), and a DNA sample from this individual was subsequently obtained as well.

The proband II-1 had originally presented at birth bicoronal CRS with oral clefting and Pierre-Robin sequence. Her facial features, as illustrated in **Figure 23**, were severally dysmorphic, including a broad forehead, hypertelorism, down-slanted s-shaped palpebral fissures, a low V-shaped frontal hairline, micrognathia, a long philtrum, a thin upper lip, and a rounded nasal tip. In addition, she exhibited moderate conductive hearing impairment associated with underdeveloped low-set ears (microtia) and abnormal auditory canals. She had previously documented raised intracranial pressure and had undergone a tracheostomy in the past. She also had a patent foramen ovale (PFO), which resolved spontaneously.

The father I-1 shares some of the proband's phenotypes, particularly relating to the ear and auditory abnormalities. He has severe conductive hearing impairment associated with fusion of the middle ear ossicles. However, his orofacial abnormalities are considerably milder in comparison to the proband. He has a tall prominent forehead and a high philtrum, but CRS has not been suspected or confirmed. He also had weak dental enamel which required several teeth to be removed.

The proband's sister II-2, born later, also presented with severe syndromic multisuture CRS as illustrated in **Figure 23**. A CT scan revealed fusion of the sagittal and both coronal sutures, along with possible bilambdoid synostosis, with the left lambdoid

fusion being more severe. Subsequent MRI scan at a later age showed a Chiari malformation with an irregular syrinx extending from t4/4 to the conus, with the worst extent at t10/11. She also had micrognathia that required multiple distraction operations. In terms of cardiac issues, she had a patent ductus arteriosus (PDA), PFO, as well as an aortic shelf. These cardiac issues were not problematic and did not require surgery.

Overall, both siblings had a bilateral presentation of severe craniofacial abnormalities with multisuture CRS. The father's less severe features of ear and auditory abnormalities likely represent a milder manifestation of the same pathological process, rather than originating from an independent mechanism.

The proband had previously undergone genetic screening using the Oxford Level 1 CRS panel, Filamin A (*FLNA*) testing, and *TCF12* testing, which were all negative. Further, a negative array analysis was obtained by the diagnostic lab in Manchester. Notably, the clinician had specifically requested investigation for *HOX* mutations due to the evident phenotype, but no positive findings were detected. Lastly, this case was also enrolled in the Solve-RD study[123], but no significant findings were obtained from this effort either. Consequently, the proband, father, and the unaffected mother were recruited to the 100kGP as a trio.

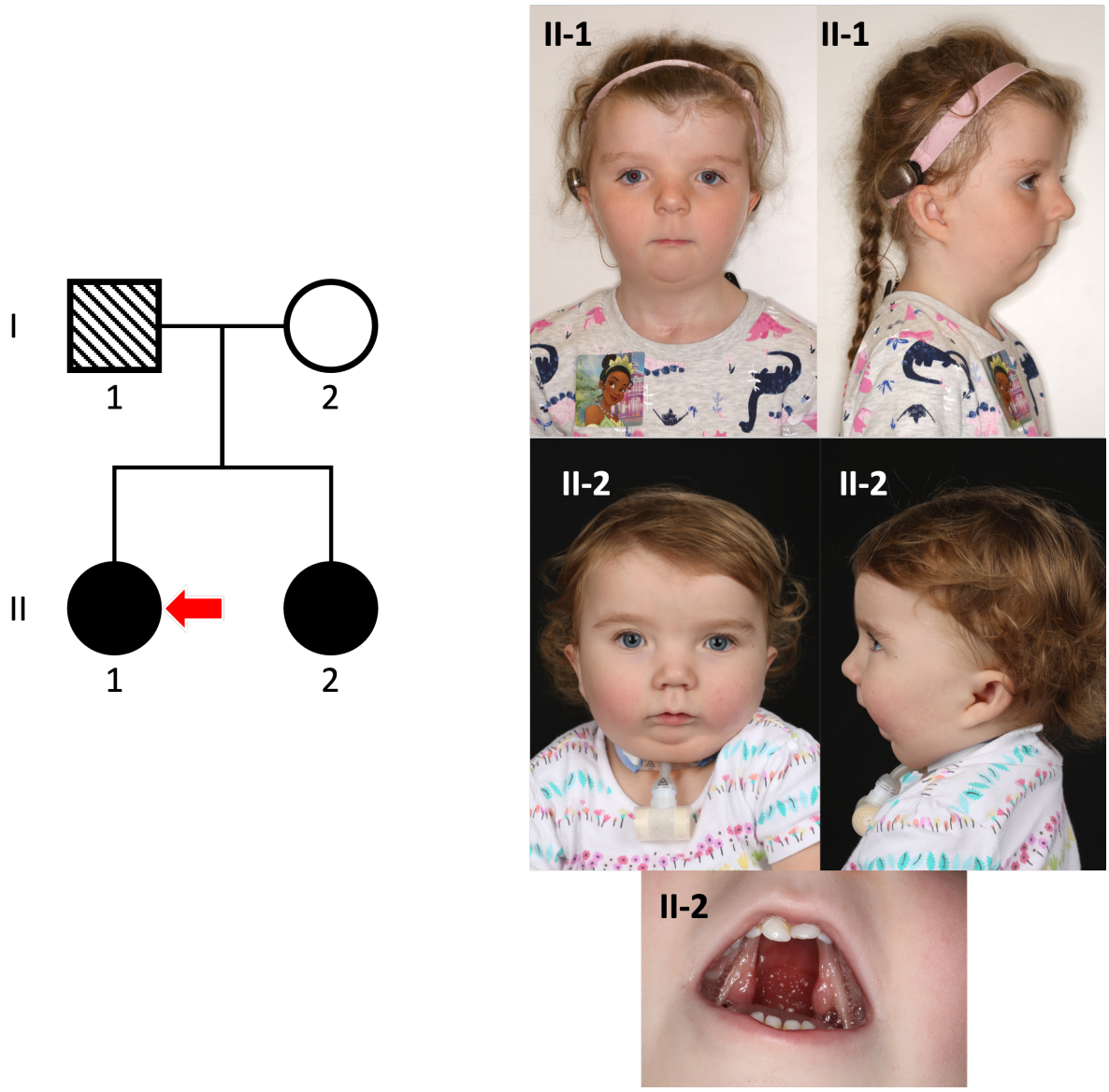


Figure 23 Case 21 family consists of three affected individuals. *The proband II-1 (red arrow) and her sister II-2 have syndromic CRS associated with ear abnormalities and hearing impairment. The father I-1 is considered to have a milder craniofacial phenotype, presenting ear abnormalities and hearing impairment without suspected or confirmed CRS. Photos of the proband II-1 was taken when she was aged 5.5 years. Photos of the sister II-2, taken at the age of 2 years before her jaw surgery, illustrated her severe micrognathia, wide cleft palate, dysmorphic, posteriorly rotated ears, and tracheostomy in situ. Permission to include clinical photographs was given by the patient and/or their parents.*

As described in the previous chapter (**section 3.2.4**), a 285 kb tandem DUP was detected in both the proband and the father. The break junction appeared simple, enabling facile design of primers to undertake breakpoint PCR. This enabled the break

junction of the DUP to be verified and confirmed to segregate in all three affected family members, as shown in **Figure 24b**. Dideoxy-sequencing of the break junction showed a three bp microhomology (**Figure 24c**), suggesting a microhomology mediated mechanism, such as MMBIR/MMEJ.[8] With the detailed break junction sequences, the DUP can be described as NC_000012.11:g.54374187_54659697dup (hg19).

Following this new finding, the initial diagnostic array was re-evaluated, showing that the DUP was indeed detectable, but was not reported due to the apparent DUP size on array falling near the reporting limit (typically at ~200 kb at the time). A new diagnostic SNP array successfully confirmed the DUP by the diagnostic lab following the 100kGP finding.

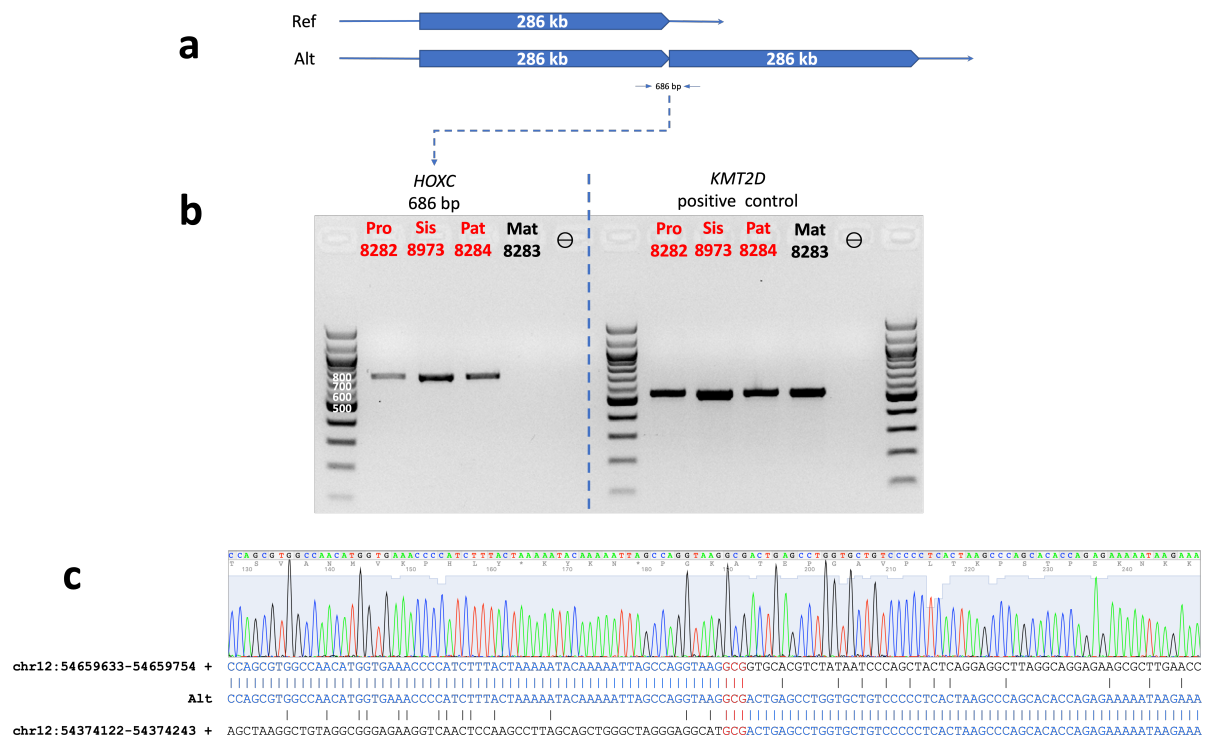


Figure 24 Case 21 HOXC DUP was verified using breakpoint PCR and dideoxy-sequencing. a*. Breakpoint PCR was designed to amplify the unique break junction created by the DUP.* ***b****. Break point PCR successfully amplified the break junction in the three affected individuals.* ***c****. Dideoxy-sequencing characterised the break junction with three bp microhomology, suggesting a microhomology mediated mechanism, such as MMBIR or MMEJ. Pro, proband; Sis, sister; Pat, father; Mat, mother. ⊖ = negative control/ water. Sequences mapped to hg19.*

In addition to the DUP findings, a segregating heterozygous SNV, rs148209145, was identified from the WGS data in the original analysis by the laboratory. This SNV was believed to potentially explain the observed deafness phenotype due to its effect on the *TECTA* gene (NC_000011.10:g.121166646T>G, p.Cys1818Gly), which is a known dominant disease gene associated with hearing loss in various populations.[124] However, the pathogenicity of this SNV within this particular family remains uncertain due to the observed low frequency in the healthy control population (2 alleles in gnomAD). Moreover, the deafness observed in the family is considered more likely associated with the familial hemifacial microsomia rather than stemming from an independent genetic cause.

4.4 Case 26: *HOXC* DEL in joint and limb malformation

Screening of the *HOXC* region in the 100kGP rare disease cohort subsequently identified two more cases with *HOXC* disruptions, with one of them being case 26. This case involves four family members recruited to the 100kGP, three of whom were affected, as shown in **Figure 25**. The family was recruited due to the diagnosis of familial arthrogryposis multiplex congenita (AMC). Within the family, all three affected individuals were annotated as affected by several prominent muscular dysfunctions, including hip contractures, joint contractures in both the upper and lower limbs, flexion contractures in the knees and elbows, bladder dysfunction, and camptodactyly of the fingers. In addition, the proband and her sister demonstrated motor delay, with the sister also presenting talipes equinovarus. Furthermore, the proband displayed delay in speech and language development. The father is unaffected within the familial context.

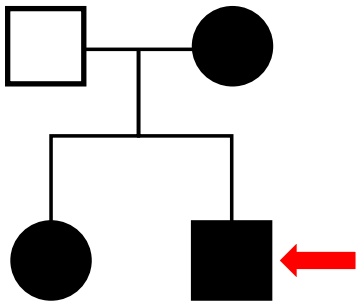


Figure 25 Case 26 is a family of four in the 100kGP with three affected individuals. *Red arrow indicates the proband.*

From the 100kGP data, a 64.5 kb DEL, NC_000012.12:g.53975383_54039954del (hg38), was identified in all three affected individuals, as shown in **Figure 26**. The DEL effectively removes *HOXC5*, *6*, *8*, & *10*, and partially, *HOXC4* & *11*. In the literature, DELs within the *HOXC* locus have been documented in multiple cases of arthrogryposis and lower extremity malformations[117,118,125]. The case 26 DEL aligns

well with the DELs reported in the literature, and therefore was classified as likely to be the causative SV. This SV was reported back to GE as a diagnostic finding, but no further update was received since.

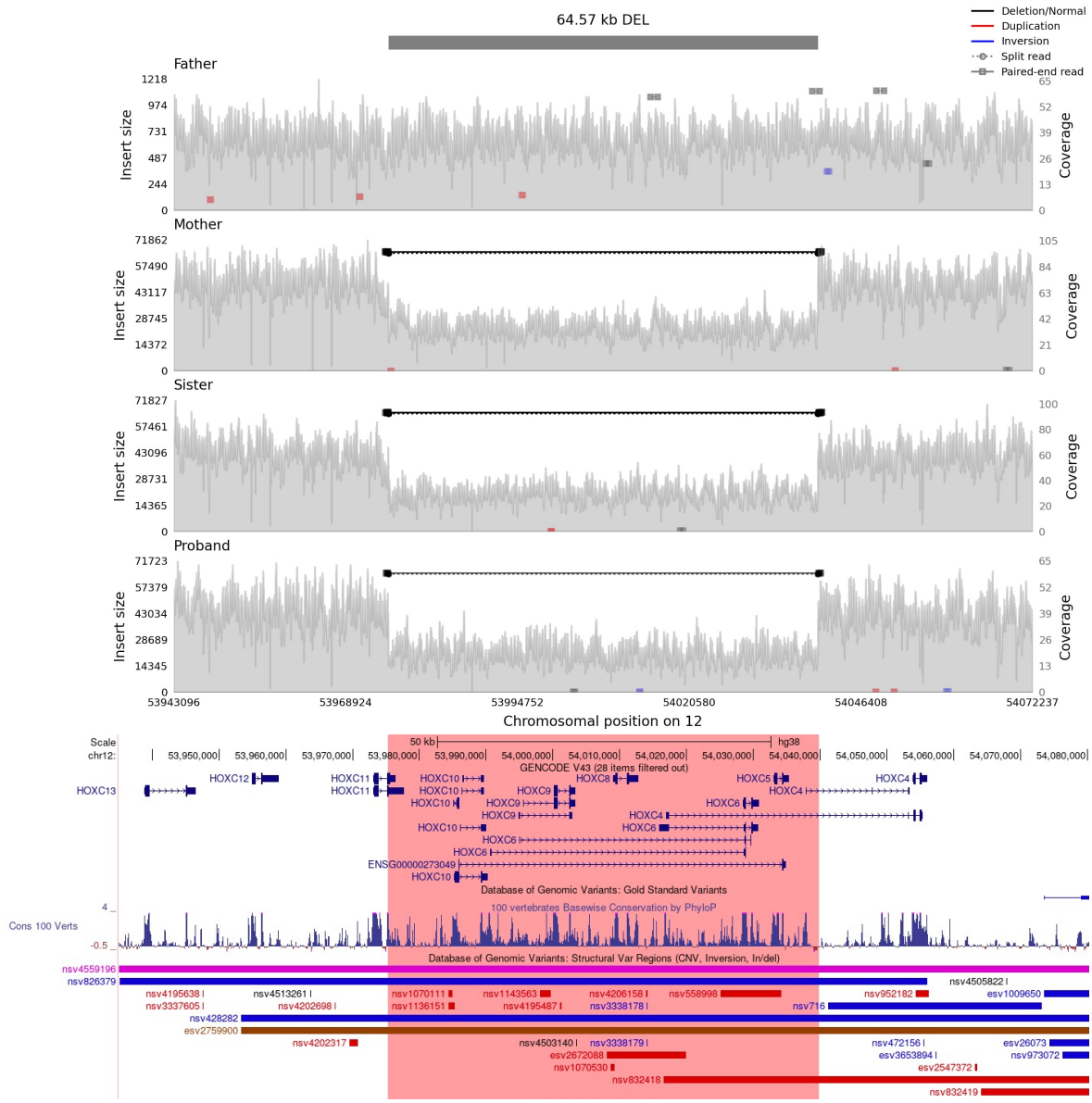


Figure 26 A small 64.5 kb DEL was identified in all three affected individuals in case 26. *The DEL affects six HOXC genes. DEL indicated by red shading. The bottom track in the lower panel indicates other SVs present in the DGV. Samplot figure exported from GE Airlock. Figure in hg38.*

4.5 Case 18: *HOXC* CPX split-DUP in congenital heart disease

Case 18 is the third case in the 100kGP where an SV affected the *HOXC* cluster. This was a local Oxford case recruited to the 100kGP as a singleton with congenital heart disease (CHD), without any CRS or other significant clinical features recorded (**Figure 27**). Analysis of the 100kGP data (described in further detail below) identified a ~944 kb DUP including the *HOXC* cluster (**Figure 29**), but there was no information in the 100kGP records that the referring clinician was aware of this DUP from previous investigations. Hence the local clinical records were obtained.

Figure 27 Case 18 proband with congenital heart disease without noticeable craniofacial abnormality. *A mild asymmetry can be seen between the nostrils. Permission to include clinical photographs was given by the patient and/or their parents.*

Examination of the case record unveiled an extensive family history of CHD tracing back at least three generations, as shown in **Figure 28**. The proband IV-6 was found to have a heart murmur at 2 weeks of age, which later progressed rapidly into a large ventricular septal defect (VSD), requiring a surgical repair at 6 weeks of age. Additionally, he presented with mild syndromic features, including mild hypertelorism, a flat nasal bridge, an asymmetric nasal septum, mild left facial palsy. He had an

abnormal epiglottis which required surgical intervention. He is also dyslexic and was investigated for autism spectrum disorder (ASD) and attention deficit hyperactivity disorder (ADHD) with no confirmation.

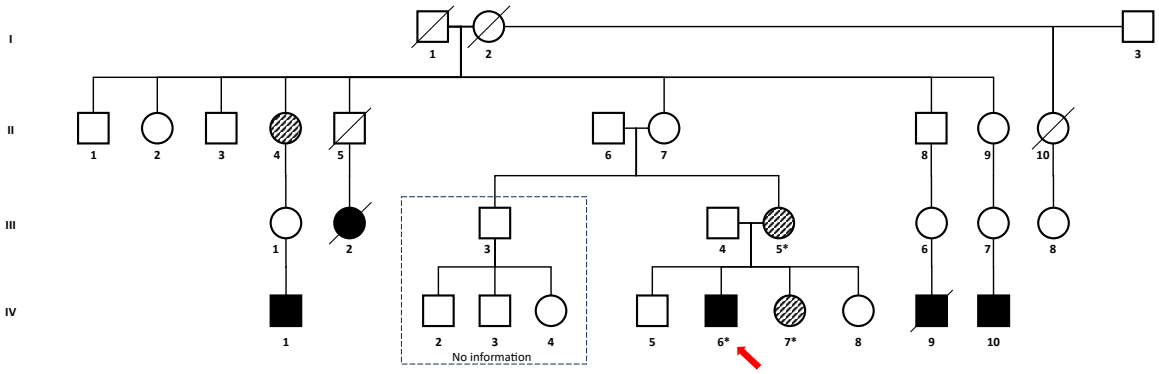


Figure 28 Case 18 is a family affected by congenital heart disease (CHD). *The proband IV-6 was recruited as a singleton to the 100kGP, while local clinical files indicated an extensive family history of congenital heart abnormalities tracing back three generations. A total of five family members (III-2, IV-1, IV-6, IV-9, and IV-10) presented with severe CHD. Only three individuals (marked with *) are known to have had genetic investigations. Dashed symbols indicate a mild presentation of CHD. The red arrow indicates the proband.*

The proband's sister IV-7 underwent a prenatal echocardiogram that revealed a moderate to large inlet-to-outlet VSD. Remarkably, this VSD resolved spontaneously without surgical intervention. She appeared to develop normally when reassessed at 1 year of age. The mother III-5, with a similar mild presentation of CHD to the sister, was found to have a murmur in early childhood, which also resolved spontaneously. Two other siblings IV-5 & 8 remained untested, and no information was available for the proband's uncle III-3 and his side of the family. The grandmother II-7 presented with a short stature, although no further clinical details were sought.

Additional clinical contact with the family yielded further information. An extensive family history of CHD was revealed from the maternal side of the family. Five more affected family members were identified: IV-1 was described to have "multiple holes

in the heart", with his grandmother II-4 needing a double heart valve replacement at the age of 40; III-2 died at the age of 6 weeks due to "three holes in the heart"; IV-9 underwent a foetal operation due to heart abnormalities, but unfortunately did not survive beyond birth; IV-10 underwent two open heart surgeries before the age of 7. This extensive family history of CHD strongly suggests an underlying genetic cause, likely originating from the second generation.

Several array-based (Agilent ISCA 60K) investigations were carried out for the proband IV-6, the sister IV-7, and the mother III-5, detecting a large 12q13.5 ~900 kb DUP in all of them. However, this DUP was determined as a SVUS/incidental, and the lab reported that "there is no need for further testing in the extended family" for this DUP. Owning to the lack of other candidate variants, the proband was recruited to both the DECIPHER study (case 385713) and the 100kGP.

Further scrutiny of the CNV gain in the 100kGP WGS data revealed that it was more complex than the smaller DUP characterised in **section 4.3**. Although, as shown in **Figure 29**, the CN gain region was identified to be ~944 kb, consistent with the array finding, the reads information annotated the SV as an inverted interlinked DUP. Further detailed reads analysis revealed two pairs of linked reads, rather than one pair of abnormal reads expected in a simple DUP, as shown in **Figure 30**. Within this context, three interlinked CNV gain segments emerged with two possible break junctions constructed as in **Figure 30**. By integrating the break junctions and the coverage information, four alternative configurations of the SVs could be constructed to explain the WGS data, as shown in **Figure 31**. The four alternatives can be

summarised into two scenarios, differentiated by the orientation of the longest segment (purple segment): either the purple segments are in tandem (as in Alt 1 & Alt 4), or are inverted (as in Alt 2 & Alt 3).

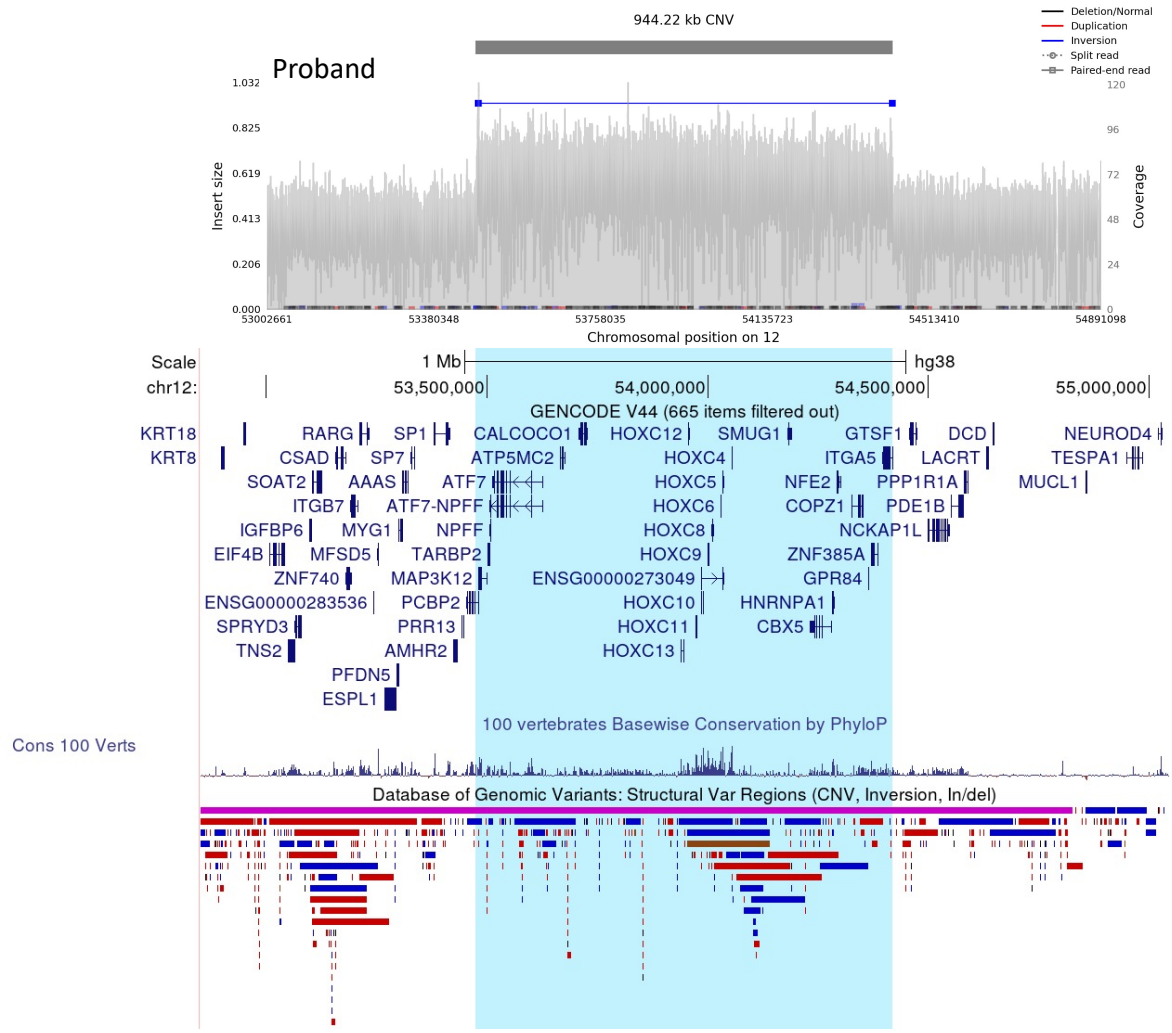


Figure 29 A large CNV gain was detected via WGS in case 18. *The HOXC cluster is fully contained within the CNV gain region. However, rather than supporting a tandem DUP, the reads information indicated an inverted DUP (note ends of the DUP segment are connected by a blue line). Lower panel: blue shading indicates the DUP event. Note that two genes, ITGA5 and PCBP2, are directed affected by the breakpoints. Samplot figure exported from GE Airlock. Figure in hg38.*

Figure 30 Detailed reads analysis of case 18 SV revealed the complex nature of this event. *Reads were examined in detail on IGV, where additional information at the break junctions revealed the complex nature of this large SV. Two pairs of linked reads were identified, with two possible break junctions constructed. Based on the read pairs and the coverages, the event is illustrated with three coloured segments: 1.6 kb in blue (1 extra copy), 2.4 kb in brown (2 extra copies), and the largest 937 kb in purple (1 extra copy). IGV figure exported from GE Airlock. Figure in hg38.*

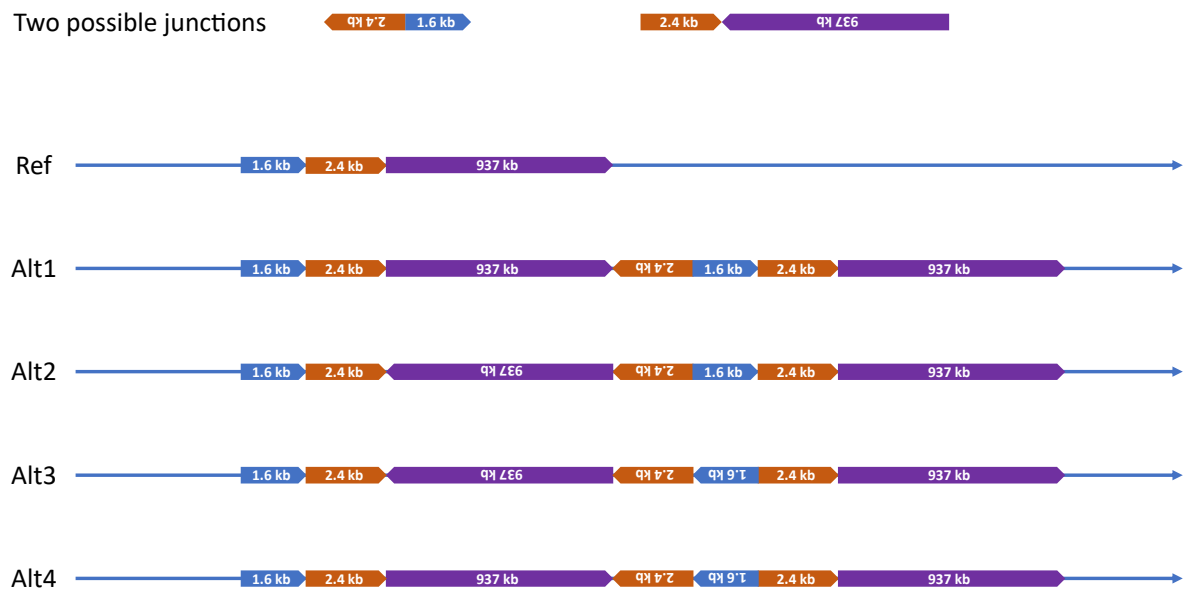


Figure 31 Four possible structures can explain the WGS data with the complex break junctions and the CNV gains. *Note that Alt1 and Alt4 have the purple segments in tandem, while Alt 2 and Alt3 have the purple segments inverted.*

To determine the precise configuration of the SV, Bionano OGM was performed using a fresh blood sample from the affected mother III-5, while a long-range PCR was performed on available genomic DNA from the affected sister IV-7 and the mother III-5. As shown in **Figure 32**, OGM successfully mapped across the break junctions between the two copies of the large purple segments, thereby illustrating a tandem-like structure in accordance with Alt 1 & Alt 4. However, the ~7 kb segment lacked sufficient labels to differentiate between the Alt 1 & 4 configurations. Alternatively, long range PCR was attempted to amplify a ~4 kb region in Alt 1, as shown in **Figure 33**; the same primers would fail in Alt 4 due to the incorrect orientations of the primers. As the PCR result demonstrated in **Figure 33**, a unique product of ~4 kb was successfully amplified in both affected members of the family. Together with dideoxy-sequencing, Alt 1 was confirmed as the true orientation of this SV, specifically as:

NC_000012.12:g.54419319_54419320ins[53479534_53481965inv;53477870_5441

9319] (hg38).

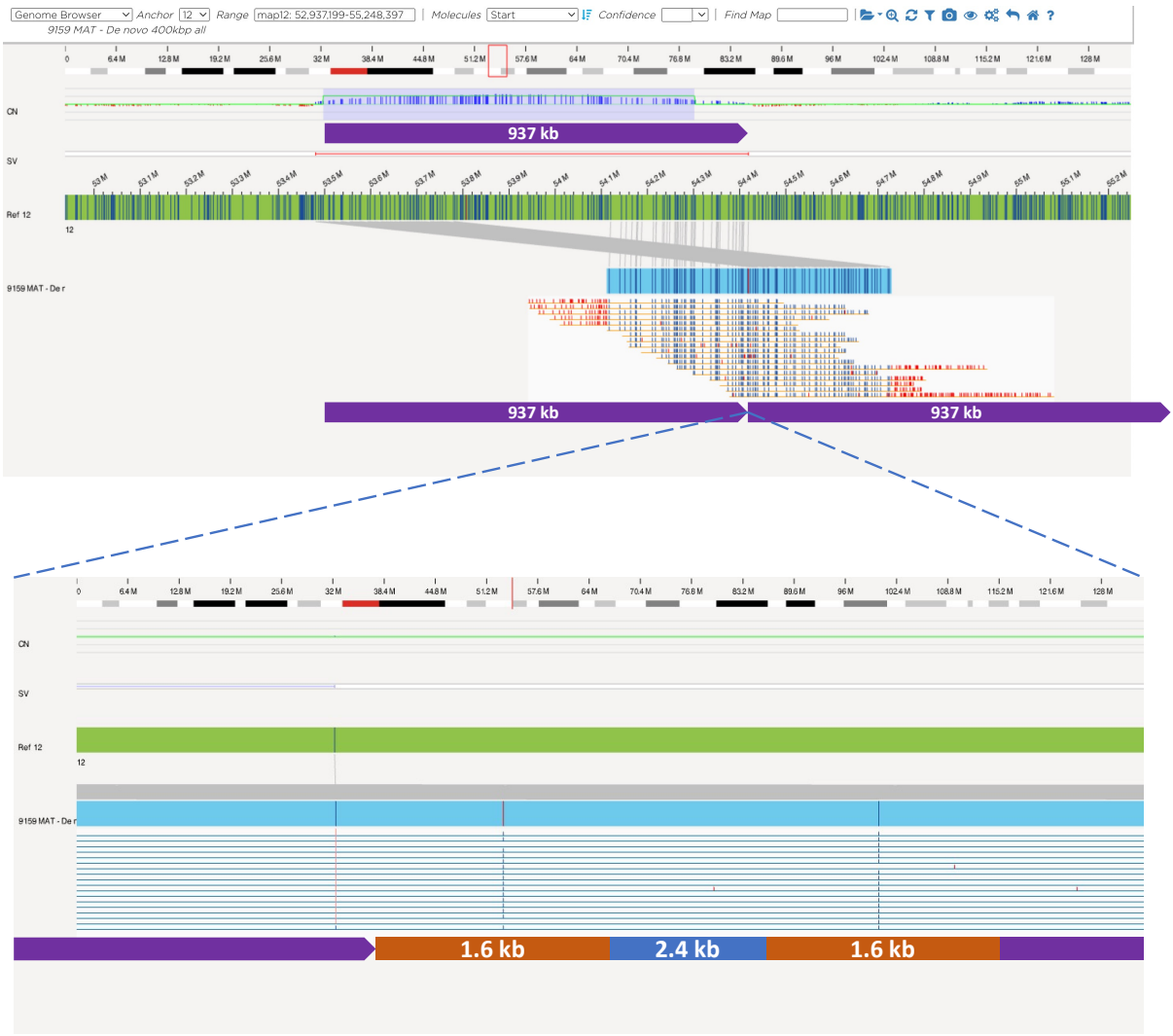


Figure 32 Bionano OGM was used to determine the orientations of the two largest purple segments in tandem. *The orientations of the smaller segments, blue and brown segment, cannot be characterised owing to the paucity of labels included within these segments.*

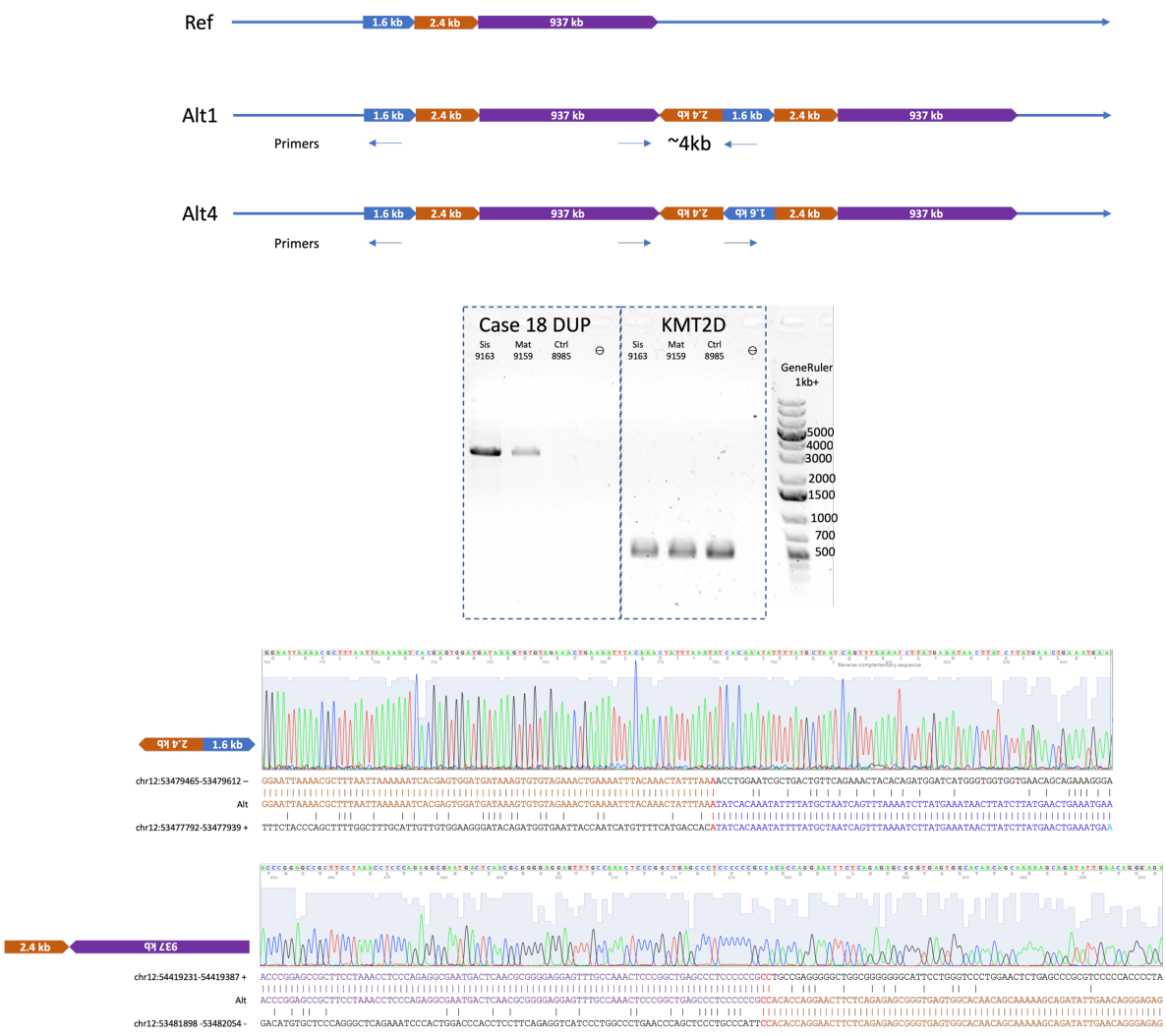


Figure 33 Long-range PCR and dideoxy-sequencing confirmed the Alt 1 structure as the true structure of this complex SV*. ⊖ = negative control/ water. Coordinates in hg38.*

4.6 Interpreting the clinical relevance of *HOXC* SVs

Deciphering the precise molecular mechanisms underlying the three cases with *HOXC* SVs is challenging, owing to the involvement of multiple critical genes within the *HOXC* locus. However, given the distinct clinical features among the three cases, it is possible to propose that these cases can be explained by substantially different mechanisms. Here, the different mechanisms underlying the three cases are explored to better understand the clinical interpretation of these SVs.

Case 26 is likely attributed to the known haploid insufficiency resulting from the *HOXC* cluster DELs. Previous reports have documented whole cluster DELs of *HOXC* in patients with arthrogryposis and lower limb malformation, as mentioned in **section 4.4**. In contrast, the case 26 DEL is more localised, affecting solely *HOXC5-11* and the long canonical *HOXC4* (the shorter *HOXC4* isoform remains intact).

This localised DEL offers valuable insights into the potential causative *HOXCs*, as certain *HOXCs* appear to tolerate DELs well. For example, examining DELs in the control population (DGV) has revealed that DELs affecting *HOXC4* (nsv832418 and esv2672088), *HOXC5* (nsv832418), *HOXC6* (esv2672088, ns558998, and nsv832418), and *HOXC8* (esv2672088) can be carried by heathy individuals (shown in **Figure 26** DGV track). Additional examples can be found in mouse models, such as the phenotypically normal *Hoxc8* targeted mice.[126] However, apparently contradictory evidence has also shown that no abnormality in limb development was observed in *Hoxc* double knock out mice, despite their lethality.[127] One reasonable hypothesis indicates that loss of a combination of *HOXC9*, *10*, and/or *11* are likely responsible for the arthrogryposis phenotype. However, it is crucial to approach this conclusion with caution, as DGV CNVs are often imprecisely described due to the limitations of array-based methods, which may lead to inaccurate identification of the affected *HOXC* genes. Similarly, in gnomAD SV, only intergenic and intronic DELs and INSs <1.5 kb have been recorded.

Overall, it is confident that the case 26 DEL is indeed pathogenic and responsible for the arthrogryposis phenotype in case 26 family. This conclusion aligns well with

previously reported *HOXC* DELs in patients. However, the exact causative gene and the precise molecular mechanism are difficult to understand. This would require further experiments that, given the focus of this thesis on CRS and craniofacial malformations, were considered beyond the scope of this work.

Unravelling the pathogenicity of the case 21 DUP is even more challenging. The initial hypothesis was that the pathogenicity could be attributed to the triploid sensitivity of the *HOXC* cluster. However, it appears evident that partial and whole *HOXC* cluster CNV gains can be well tolerated in healthy individuals. For example, in the control population (**Figure 34** DGV track), the DUP nsv826379 (121 kb) consists of the entire *HOXC* cluster, while the DUP nsv428282 (188 kb) covers the *HOXC* cluster except *HOXC13*. These two DGV DUPs were both identified in healthy individuals, suggesting that *HOXC* cluster is less likely to be triploid sensitive. In gnomAD SV, only small non-coding SVs and two large INVs (3.6 Mb and 64 Mb) were recorded.

The second hypothesis for the case 21 DUP pathogenicity revolves around the potential disruption of the TADs within the *HOXC* locus. As shown in **Figure 34**, the *HOXC* cluster is fully contained under the "*HOXC* TAD". Non-pathogenic DUPs and CNV gains, such as nsv3338179 and nsv428282 mentioned above, are all intra-TAD events located within the *HOXC* TAD. This suggests that intra-TAD DUPs are likely non-pathogenic. Similarly, case 18 SV can also serve as a control for the syndromic craniofacial phenotypes observed in case 21, given that case 18 family did not exhibit any syndromic craniofacial features. Notably, the case 18 SV completely covers the *HOXC* TAD, thereby creating a copy of the TAD without disrupting the structure of

either TAD copy (**Figure 34**). In contrast, the case 21 DUP extends beyond the 3' boundary of the *HOXC* TAD, resulting in an inter-TAD event. Consequently, this inter-TAD DUP disrupts the structure of the *HOXC* TAD to create a neo-TAD[59], possibly introducing an entirely different set of regulatory elements, potentially resulting in the misexpression of these vital developmental genes.

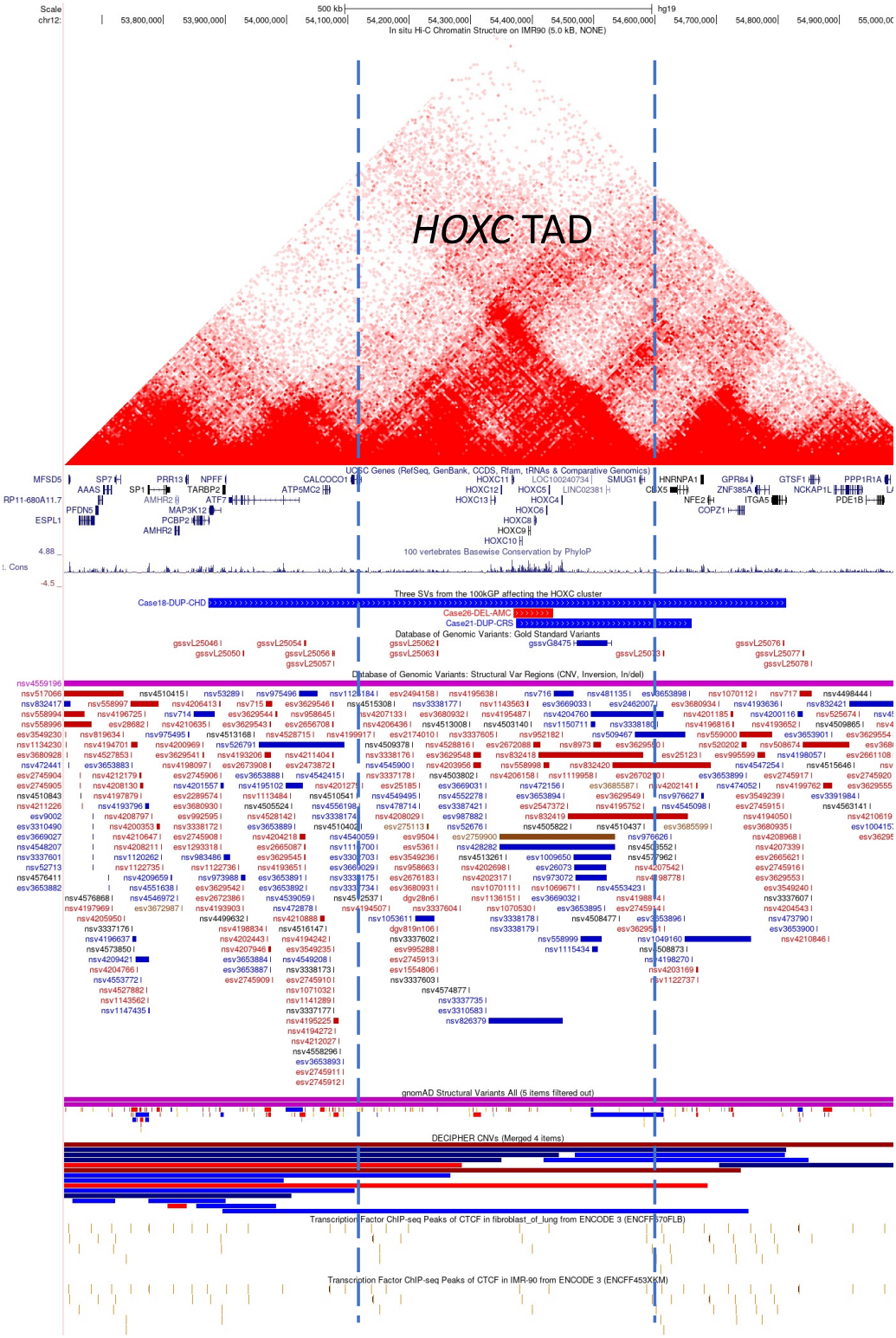


Figure 34 HOXC cluster and its genomic context. *Blue dash lines mark the boundaries of the HOXC TAD. Figure includes the Hi-C / TAD track on IMR90, UCSC gene track, PhyloP conservation track, the three HOXC SVs in the 100kGP, DGV SVs, gnomAD SVs, DECIPHER SVs, and ChiP-seq CTCF peaks. CHD = congenital heart disease; AMC: arthrogryposis multiplex congenita; CRS: craniosynostosis. Figure in hg19.*

To investigate this hypothesis further, DeepC was applied to predict the effects of case 18 and case 21 SVs on the *HOXC* cluster TADs. DeepC is a computational tool designed to understand the three-dimensional structure of chromatin in mammalian genomes, particularly focusing on TADs.[128] It works by inputting DNA sequence information from either reference or mutant sequences, and uses a deep leaning model to predict the chromatin interactions. As depicted in **Figure 35**, two distinct TADs, the *HOXC* TAD and *NFE2* TAD, were directly affected by the case 21 DUP. DeepC predicted, for case 21, a neo-TAD due to the merging of the partial *HOXC* TAD and the partial *NFE2* TAD. This further supports the hypothesis that the partial *HOXC* cluster contained within the neo-TAD may be mis-regulated, therefore yielding a pathogenic effect. In contrast, for case 18, both the *HOXC and NFE2* TADs are fully contained within the large DUP region. As predicted by DeepC, both the original the extra copy of the TAD remain separated and undisrupted. This undisrupted *HOXC* TAD again supports that the syndromic craniofacial abnormalities are likely associated only with inter-TAD disruptions in the *HOXC* locus. Overall, this scenario potentially parallels that of the well-known *KCNJ2*-*SOX9* loci, where inter/intra-TAD DUPs result in distinct clinical phenotypes.[61]

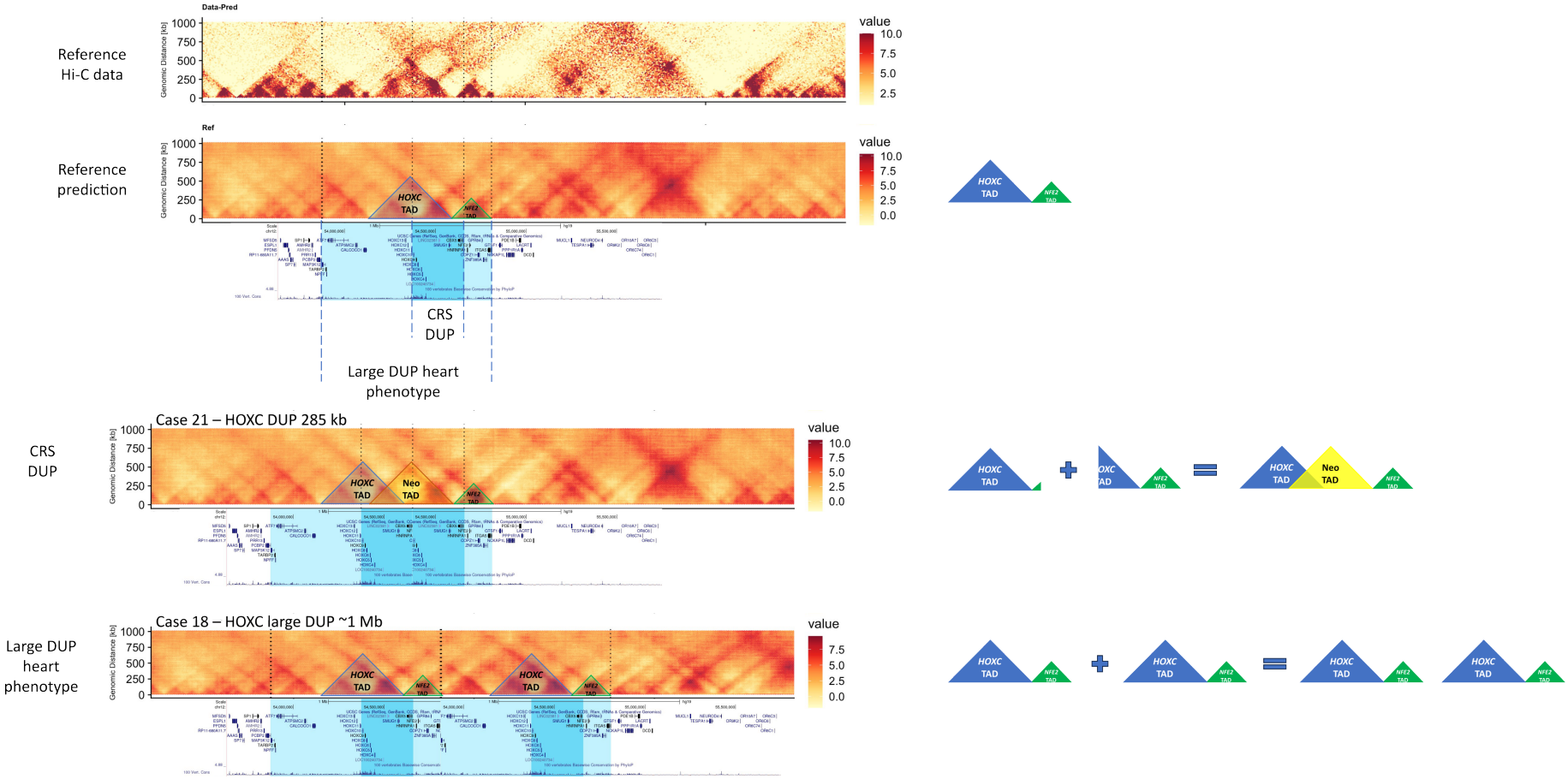


Figure 35 DeepC predicted TADs and interpretations for the HOXC cluster for the reference, case 18, and case 21 data. *Tracks are, from the top to bottom, HI-C data (IMR90), Ref DeepC prediction, Ref genes, case 21 prediction, case 21 genes, case 18 prediction, case 18 genes. Blue shading at the gene tracks highlights the two DUPs, with the darker blue shading highlights the small case 21 DUP (hence also the overlapping region between the two DUPs), while the whole blue shading (light + dark) highlights the large case 18 DUP. Figure in hg19.*

A third hypothesis relating to the case 21 SV pathogenicity adds to the second hypothesis mentioned above. This third hypothesis proposes that, the disruption of the *HOXC* TAD and regulation further interferes with the *HOXC* spatial collinearity, leading to abnormal expression of certain *HOXCs* along the cranio- caudal axis. During normal development, the *HOX* spatial collinearity dictates that the *HOXC* cluster is typically inactive in cranial tissues, as it lacks the first three *HOX* paralogues, "*HOXC1-3"* (**Figure 22**). Due to the TAD disruptions in case 21, the spatial collinearity of the *HOXC* cluster may have been disturbed, causing the *HOXCs* to be ectopically expressed more anteriorly in the cranio-caudal axis, potentially affecting craniofacial development. For example, it was previously shown in the mouse that ectopic *Hoxa2* expression in branchial arch 1 territory (corresponding to the developing maxilla and mandible) led to severe, dose-dependent facial malformations with partial homeotic transformation into branchial arch 2-like structures.[129]

In the literature, a previous study documented a mouse model with a similar mechanism whereby disruption of the *Hoxc* cluster organisation causes disruption of the *Hoxc* spatial collinearity. The mouse In(15)2Rl (hairy ears, *Eh*) mutation is a large 46 Mb inversion with the distal breakpoint positioned immediately after the *Hoxc* cluster, as shown in **Figure 36**. This INV effectively repositions the *Hoxc* cluster to an entirely different environment, without directly altering the coding region of the *Hoxc* - comparable to the consequence of case 21 SV. Interestingly, using the mouse model, Mentzer et al (2008)[130] also showed that, as the result of this INV, *Hoxc* 8 and 9, which are commonly expressed posteriorly in the cranio-caudal axis, are now expressed anteriorly in the ear, leading to the "hairy ear" phenotype in the *Eh* mouse. This study therefore provides a foundational basis for the hypothesis that case 21 DUP might

lead to the ectopic expression of *HOXCs* more anteriorly in the cranio-caudal axis, subsequently contributing to the craniofacial phenotype, including the unique ear abnormalities observed in the patient along with other syndromic features.

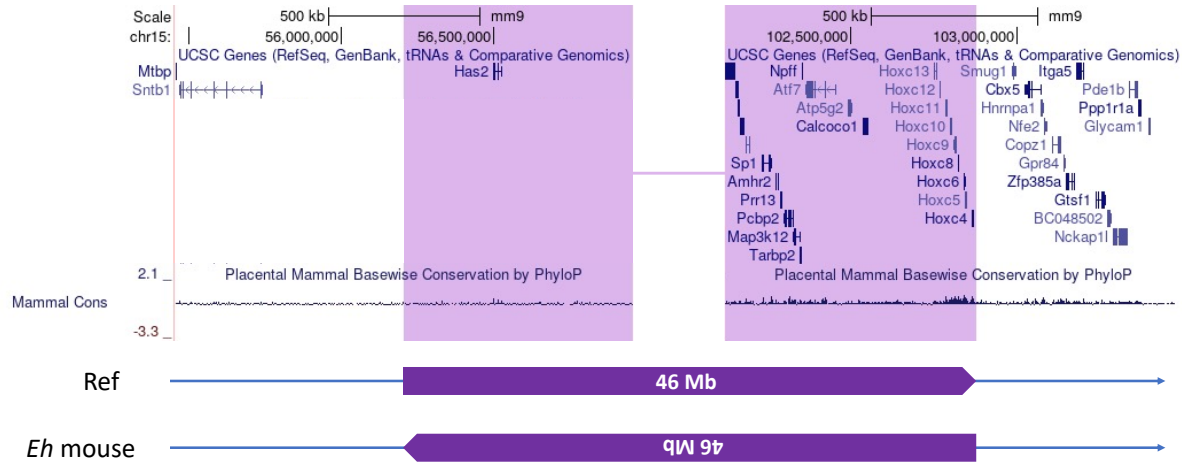


Figure 36 Eh mouse model contains a large 46 Mb INV affecting the genomic environment of the Hoxc cluster without directly affecting the coding region of the Hoxcs. *Precise break point coordinates in mm9 were generated using the breakpoint sequences obtained by Mentzer et al (2008).*[130] *Figure in mm9.*

This collinearity disruption, characterised by the ectopic expression of posterior *HOXs* affecting the anterior cranial tissues, can also be seen in other animal models. For example, in chickens, the ectopic expression of *HOXC10,* caused by an intronic DUP, was shown to affect the anterior tissues, leading to altered crest morphology and cerebral hernia.[131] Additionally, a similar effect was observed with ectopically expressed *HOXC8*, although it was not as pronounced as that seen with *HOXC10*.[132] In another chicken model, a large complex SV result in the ectopically expressed *HOXB8*, again leading to anterior effects, such as altered muffs and beard.[133] These animals models, both mice and chicken, provide compelling evidence that the disruption of the *HOX* collinearity can lead to ectopic expression of posterior *HOXs*, abnormally exerting an anterior effect along the cranio-causal axis. This further

supports the hypothesis that case 21 DUP results in similar effects on the *HOXCs* leading to the craniofacial phenotypes observed in the affected patients.

In contrast, the pathogenicity of the case 18 SV may stem from specific genes rather than the misregulation of the *HOXC* cluster. *HOX* genes are essential in cardiac development, particularly in processes such as cardiac progenitor patterning and artery formation.[134] Some specific *HOX* genes, such as *Hoxa1*, *Hoxa3*, and *Hoxb1*, have been associated with cardiac phenotypes in animal models.[135-140] Interestingly, while *HOX* genes have been well studied in the literature, *HOXCs* specifically have not been well explored for their roles in heart abnormalities or heart development. However, a notable distinction from the other two *HOXC*-related families presented in this chapter is that the case 18 DUP does not disrupt the fundamental structure of the *HOXC* cluster, or the relevant HOXC TAD. Despite having an extra copy of the entire *HOXC* cluster, the surrounding genetic environment remains largely intact for both copies of the *HOXCs* in case 18. Since *HOXC* DUPs seem to be tolerated in DGV as discussed previously, it is difficult to formulate a clear hypothesis of *HOXC* pathogenicity relating to the heart phenotype.

On the other hand, the large case 18 DUP affects not only the *HOXCs,* but also multiple other clinically relevant genes, as summarised in **Table 19**. These genes may contribute to the pathogenic mechanisms due to their susceptibility to altered gene dosage or misregulation prompted by the novel regulatory environment. For example, *ITGA5* is a compelling candidate due to its role in endocardial differentiation through the fibronectin signalling pathways in the heart, as demonstrated in animal models.[141]

As shown in **Figure 37**, *ITGA5* is regulated by several regulatory elements downstream near the *GTSF1* locus. In case 18, this genetic environment is altered due to the DUP, which substitutes the *GISF1* locus with the *MAP3K12* locus, introducing a new regulatory context for *ITGA5*. Another gene of interest, *PCBP2,* has previously been implicated in cardiac hypertrophy and heart failure.[142,143] Similar to *ITGA5*, *PCBP2* is also positioned at the boundary of the DUP (highlighted in **Figure 37**), which potentially alters the surrounding genetic environment of *PCBP2*. Lastly, *TARBP2* represents another boundary gene linked to cardiac hypertrophy and remodelling, evidenced by the cardiomyopathy and lethal heart failure resulting from the *Tarbp2* knock out mouse models.[144,145] Overall, while the specific contribution of *HOXC* genes remains complex and challenging to disentangle, the broader genomic context suggests that the pathogenicity of case 18 SV may be multifactorial, involving not only the *HOXC* cluster, but also other genes affected by the DUP.

Table 19 Genes directly affected by case 18 complex SV

Gene	pTriplo	pLI	LOEUF	sHet	pHaplo
MAP3K12	1	1	0.07	0.333	0.52
CBX5	1	0.93	0.35	0.118	0.76
PCBP2	0.99	1	0.23	0.073	0.79
HOXC13	0.99	0.42	0.66	0.026	0.9
HOXC9	0.99	0.01	1.15	0.03	0.86
HOXC8	0.99	0.17	0.73	0.17	0.83
HOXC6	0.98	0.51	0.6	0.154	0.87
HOXC10	0.96	0.17	0.73	0.154	0.83
HOXC11	0.95	0	1.14	0.006	0.72
ATF7	0.89	1	0.2	0.241	0.31
HOXC4	0.89	0.17	0.73	0.022	0.74
HNRNPA1	0.88	1	0.15	0.069	0.92
TARBP2	0.84	0.47	0.49	0.073	0.25
HOXC12	0.64	0	1.93	0.005	0.49
ITGA5	0.63	0.42	0.34	0.423	0.66
CALCOCO1	0.58	0	0.64	0.035	0.17
HOXC5	0.5	0	1.55	0.006	0.5
COPZ1	0.45	0.95	0.34	0.172	0.24
ZNF385A	0.43	0.23	0.57	0.008	0.25
NPFF	0.41	0	1.9	0.007	0.09
NFE2	0.39	0.67	0.52	0.155	0.18
SMUG1	0.23	0	1.08	0.005	0.07
GPR84	0.12	0	1.58	0.006	0.06
ATF7-NPFF	-	0.88	0.37	-	-
ATP5MC2	-	0	1.66	-	-
CISTR	-	-	-	-	-
RN7SKP289	-	-	-	-	-
HOXC13-AS	-	-	-	-	-
HOTAIR	-	-	-	-	-
HOXC-AS3	-	-	-	-	-
MIR196A2	-	-	-	-	-
HOXC-AS2	-	-	-	-	-
HOXC-AS1	-	-	-	-	-
MIR615	-	-	-	-	-
FAM242C	-	-	-	-	-
LINC02381	-	-	-	-	-
MIR3198-2	-	-	-	-	-
RN7SL390P	-	-	-	-	-
SCAT2	-	-	-	-	-
RNU6-950P	-	-	-	-	-
MIR148B	-	-	-	-	-
RN7SL744P	-	-	-	-	-

pTriplo: Predicted Probability of Triplosensitivity. pLI: Probability of Loss-of-function Intolerance. LOEUF: Loss-of-function Observed / Expected Upper bound Fraction. sHet: Selection coefficient of heterozygous loss-of-function variants. pHaplo: Predicted Probability of Haploinsufficiency. Values coloured coded in red-yellow-green scale, with red values indicating genes are most likely dosage sensitive. Data extracted from DECIPHER.45

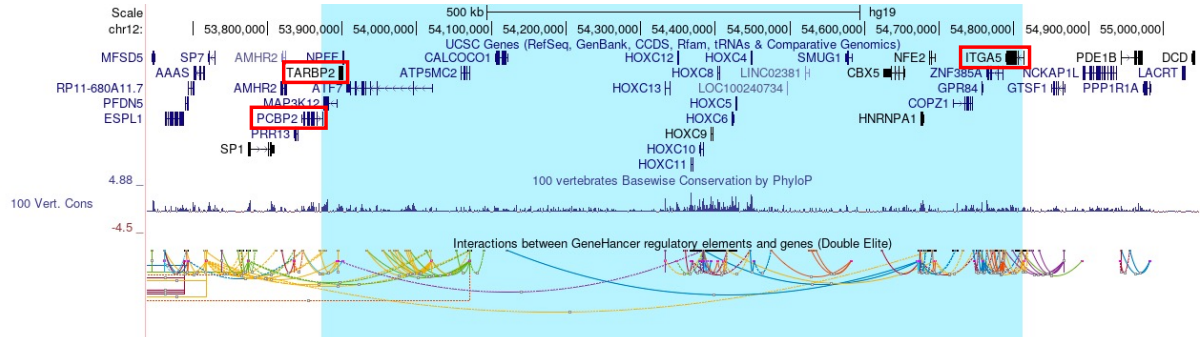


Figure 37 case 18 SV affects many compelling candidate genes other than the HOXCs. *The largest SV segment (purple segment in* ***Figure 30****) is highlighted in light blue. Genes at the DUP boundaries are most likely affected by disrupted regulation. Three compelling candidate genes were highlighted in the red boxes. UCSC genome browser tracks: gene track, PhyloP conservation track, GeneHancer regulatory element track. Figure in hg19.*

4.7 *HOXC* SVs and future direction

The *HOX* clusters have been investigated for their roles in embryonic development across multiple species. These genes play a crucial role in body segmentation and therefore are highly likely to be involved in genetic diseases. This chapter has investigated three cases of SV affecting the *HOXC* cluster. These cases, discovered from the 100kGP data, serve as excellent examples demonstrating the clinical heterogeneity in *HOXC*-related phenotypes, including syndromic craniofacial abnormalities (case 21), arthrogryposis (case 26), and congenital heart disease (case 18).

The first major aspect discussed is the challenges in the interpretation of clinically relevant SVs in different contexts. Case 21 highlights the difficulties due to the mosaic nature of the affected father, the relatively small size of SVs, and the lack of known associations between craniofacial abnormalities and *HOXC* genes. This resulted in a (nearly) decade-long diagnostic odyssey for the family, while a proper diagnosis was

only identified by examining specifically non-segregating and likely mosaic SVs in the data.

Case 26 illustrated the challenges with existing panel-based screening method and the translation of research findings into clinical applications. Despite multiple cases in the literature linking *HOXC* DELs with arthrogryposis, the GE PanelApp (v5.0, 22 Mar 2023) has yet to register the *HOXCs* into the arthrogryposis panel. This is likely a significant contributing factor leading to this missed diagnosis. In contrast, case 18 SV's complexity and the lack of functional studies illustrated the difficulties in further interpreting and assessing the clinical relevance of the SV during initial diagnosis.

The second key focus has been on the molecular mechanisms underlying these SVs to better understand the regulatory landscape at the *HOXC* locus. TAD analysis using SVs from the three cases, control populations, literature, and animal models, suggested that inter-TAD SVs affecting *HOXC* regulation and collinearity are more likely pathogenic, causing syndromic phenotypes as shown in case 21 and the *Eh* mouse. Intra-TAD SVs without affecting *HOXC* collinearity are likely tolerated or lead to mild clinical features, such as described in case 18. Certain *HOXC* genes might also be haploinsufficient, as evidenced by case 26. However, establishing a specific genotype-phenotype correlation for *HOXC* genes remains incomplete.

Future studies hold promise to further elucidate the complexities of these cases. A mouse model could provide valuable insight for case 21, given the small and simple nature of the DUP, the conserved synteny between the human *HOXC* and the mouse

Hoxc clusters, and the prominent syndromic features that can be effectively studied in mice. The case 18 SV would be much more challenging to model in mice owing to the large size of the event. A more suitable approach may include transcriptomic analysis, aiming to identify key genes affected by the SV. Although beyond the scope of this project, follow up work is planned to obtain patient-derived induced pluripotent stem cells (iPSCs) from both case 18 and 21. These iPSCs will be differentiated into cranial neural crest, from which RNA sequencing (RNA-Seq) or long-range RNA-Seq analysis will be conducted to explore *HOXC* gene expression.

Chapter 5 Results – Long-range technologies and complex events

5.1 Introduction

In the previous chapter, I demonstrated the use of Bionano OGM to determine whether a large (~1 Mb) DUP including the *HOXC* cluster was oriented in direct or reverse orientation with respect to the original copy. In this chapter, I describe the use of the OGM technology to analyse a series of SVs of increasing complexity. This has enabled assessment of the strengths and limitations of the OGM approach and enabled me to progressively refine my methodology to tackle increasingly challenging SVs. I also used Bionano OGM to seek missing diagnoses in a series of unsolved CRS patients, 15 from the 100kGP and 5 from other sources of ascertainment. In total, 20 unrelated cases were studied, as summarised in **Table 20**. All probands presented with CRS, except case 19 which was a 100kGP case with generalised arterial calcification of infancy (GACI), analysed by OGM because it was apparent that this technology could resolve alternative interpretations of the genotype. **Supplementary Table 3** in the **Appendix** provides key metrics for the data quality obtained on each Bionano OGM run, where this contributed to the final dataset (poor quality runs are omitted).

Including the *HOXC* DUP described in **Chapter 4** (Case 18 in **Table 20**), seven of the 20 cases exhibited obvious abnormalities on OGM; of these, an abnormal array CGH had been found in two. As a result of OGM, four complex SVs were fully characterised, while for the remaining three, only partial characterisation could be achieved, necessitating further investigation through alternative technologies. Case 11 is not further discussed as this is a separated project led by Dr. Eduardo Calpena, where I only performed Bionano OGM for SV verification without any addition analysis. Overall, this chapter presents my analyses of five of the remaining six cases (2, 10, 12, 16,

and 19) harbouring major structural abnormalities, whereas the next Chapter, **Chapter 6**, explores the value of Bionano OGM in identifying occult SVs.

Table 20 Summary of the 20 cases analysed using Bionano OGM

Case ID	*100kGP*	*Notes*	*Candidates*	*Array*	*Resolution*	*Chapter section*
1	Y	Affected twin				
2	-	Bionano + FISH resolves the SV	t(16:17), *KCNJ2 & 16*	✓	Partial	5.5
3	Y	Paternal inheritance				
4	Y	*De novo* trio				
5	Y	Causative SNP identified by GE	*HIST1H1E* c.430dup	X	N/A	
6	Y	Mat inheritance, duo				
7	Y	*De novo* trio				
8	Y	*De novo* trio, Treacher Collins syndrome				
9	Y	*De novo* trio, non-syndromic				
10	Y	*De novo* trio	INS near *FGF9* vs *FOXP2* SNV	X	Full	5.4
11	Y	Work led by Dr. Eduardo Calpena	INS near *FOXD3*	X	Full	
12	-	Familial Saethre-Chotzen, duo	7pINV near *TWIST1*	X	Full	5.2
13	Y	Consanguineous, 1st cousin parents				
14	-	Complex family with 2 affected siblings				
15	-	*De novo* trio				
16	Y	Smaller segments resolved	Complex INS affecting *PLCB4*	X	Partial	5.6
17	Y	*De novo* trio				
18	Y	Bionano + long range PCR resolves the SV	Large *HOXC* complex DUP	✓	Partial	4.5
19	Y	Consanguineous	HOM INS near *ENPP1*	X	Full	5.3
20	-	*De novo* trio				

**Resolution column indicates whether Bionano OGM can fully characterise the candidate SV: "Full" category > Bionano OGM can fully characterise the SV without the need or other technologies; "Partial" category > Bionano OGM can only partially characterise the candidate SV; SNVs are not detectable by Bionano OGM and therefore marked as "N/A". Array column indicates if the candidate SV was detected or partially detected by array-based approaches: "✓" suggests candidate was at least partially detected by an array; "X" indicates the candidate was not detectable by array methods. No compelling candidate variants were identified in other cases. Blank cells: no information was available*

5.2 Case 12: Saethre-Chotzen syndrome

Case 12 is a local family with Saethre-Chotzen syndrome, characterised by an SV disrupting *TWIST1* regulation without directly affecting the coding region of *TWIST1*. Case 12 serves as a test scenario for the effectiveness of Bionano OGM, demonstrating its ability to identify causative SVs that had previously eluded other detection methods. Additionally, this case is a good example of how the disruption of long-range regulatory elements can lead to a clinical phenotype, providing valuable insights into the mechanisms underpinning SV pathogenesis.

Case 12, diagnosed with Saethre-Chotzen syndrome, presented with bicoronal synostosis and dysmorphic small ears, as shown in **Figure 38**. Her mother did not have CRS but presented with a similar appearance and ptosis, and was suspected to have the same diagnosis. However, no disruption of the *TWIST1* gene was found initially in the proband using several diagnostic approaches, including karyotype analysis, array CGH (Agilent 105k V2), targeted panel testing for multiple CRS related genes including *TWIST1*, WGS analysed as part of the WGS500 whole genomes project[146], and an exhaustive resequencing effort of a 2.3 Mb region (chr7:17346143-19695462, hg38) surrounding the *TWIST1* locus (unpublished). Notably, array data revealed a substantial 570 kb DUP on chr12 present in the mother, but this was not transmitted to the proband, and therefore did not segregate with the phenotype. WGS analysis did not detect any interesting CNVs, and the resequencing findings were still pending full analysis. This case represents a critical diagnostic gap, where a strong clinically suspected genetic diagnosis is evident, and yet pinpointing the causative variant remained elusive despite extensive genetic testing.

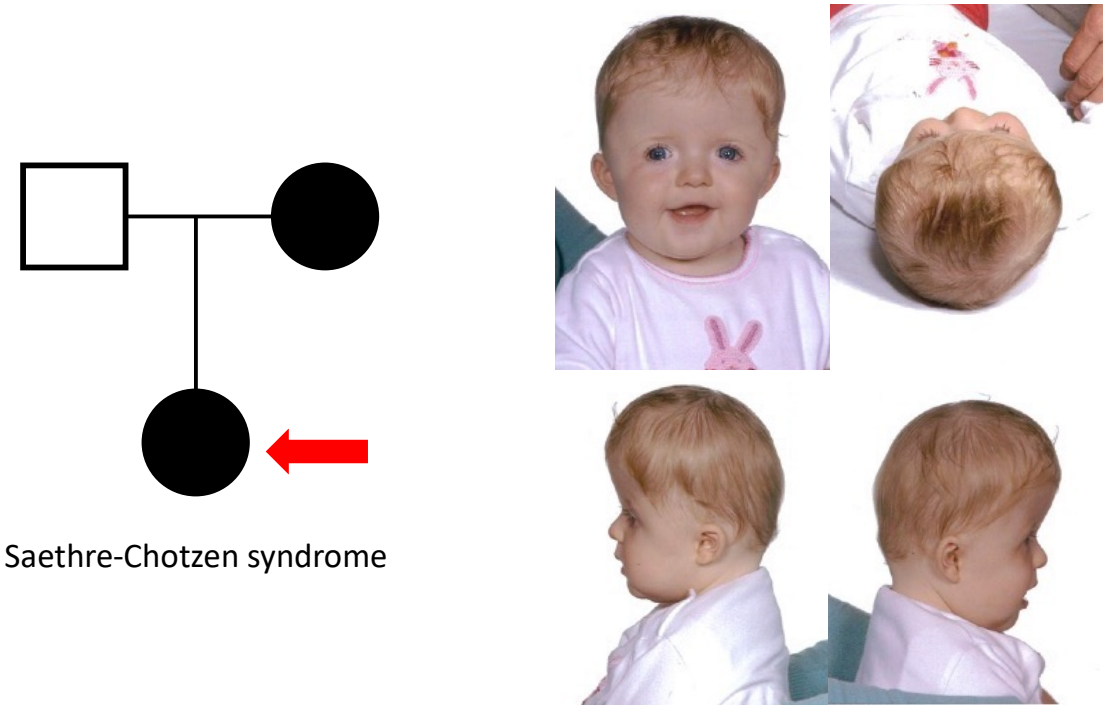


Figure 38 Case 12 presented familial Saethre-Chotzen syndrome where both the proband and the mother are affected. *Permission to include clinical photographs was given by the patient and/or their parents.*

One hypothesis was that a causative SV might underline the patient's phenotype, and due to the nature of short read and array-based technologies, this SV had eluded detection. Therefore, Bionano OGM was performed using a fresh blood sample acquired from the affected mother. *De novo* assembly and rare variant filtering identified 35 rare SVs in total. Among these, a striking ~4 Mb INV (chr7:14,573,406-18,885,342 INV, hg38) on chr7 stood out, as shown in **Figure 39a**. Upon closer inspection, the INV is located only ~200 kb upstream of *TWIST1* within the *TWIST1* regulatory region, as shown in **Figure 39b**.

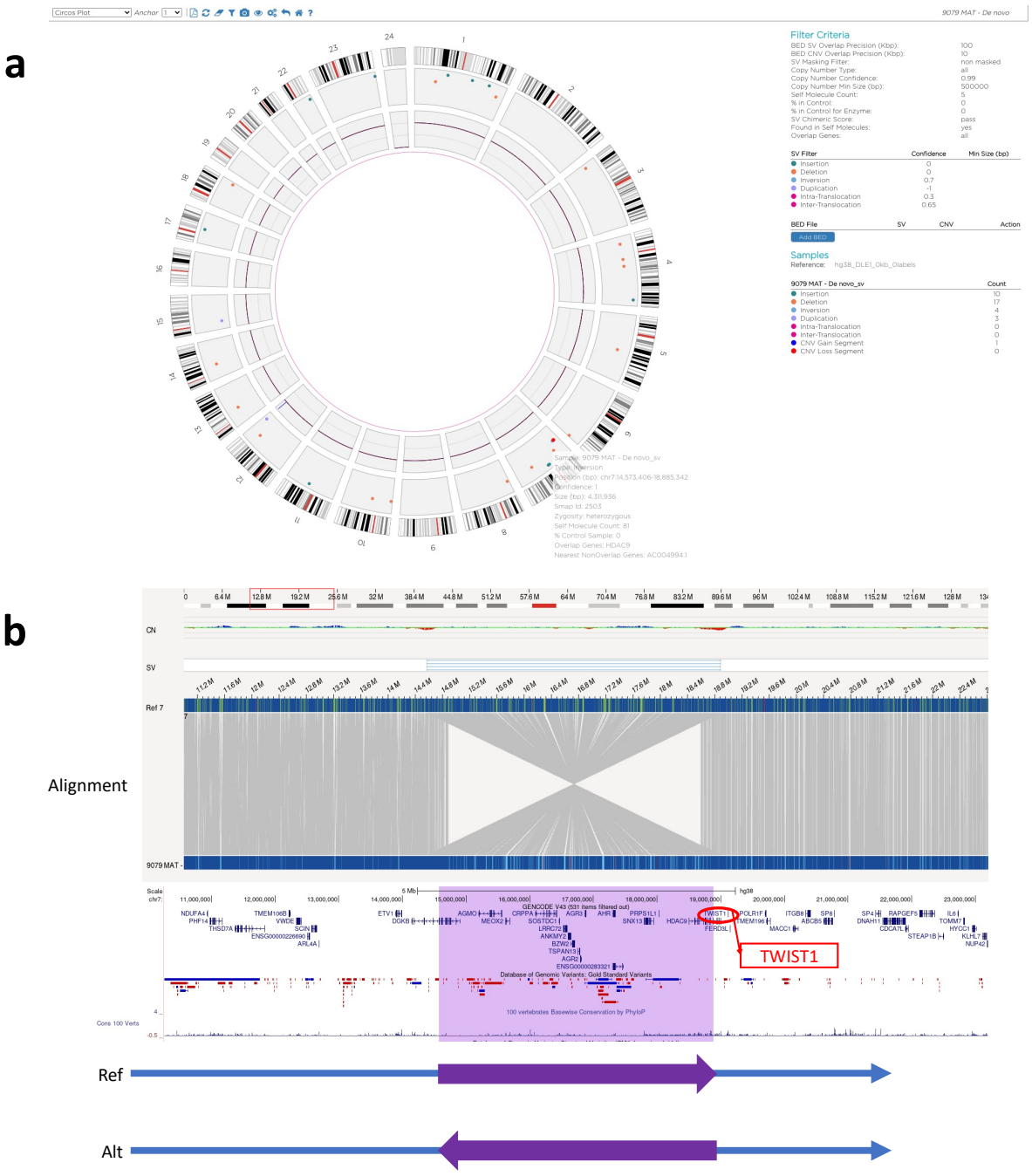


Figure 39 Bionano OGM analysis identified a large INV ~200 kb telomeric of TWIST1. a*. Circos plot from the Bionano OGM analysis showing that a total of 35 rare SVs were identified. Among these SVs, the large chr7 INV stood out.* ***b****. Local view of the Bionano data for the chr7 INV. Comparing the maternal labelling pattern on chr7 against the reference labelling pattern suggests a large INV has occurred in the maternal sample. The schematic illustrates the INV using a purple block and thick arrows to indicate the inverted region and the thinner blue arrow for reference chr7. Figure in hg38.*

Breakpoint PCRs were undertaken to verify the INV. The 3' break of the INV lies in a non-repetitive region with PCR & dideoxy-sequencing confirming the break, as shown in **Figure 40**. The 5' break, however, lies on top of a 291 bp AluSx repeat, followed by a 520 bp $(TA)_n$ repeat, that posed difficulties for PCR amplification. Initial PCRs were attempted using long-range enzymes (LongAmp & OneTaq) to span the repeats but were unsuccessful. It was hypothesised that the enzyme detaches and drops out when encountering large $(TA)_n$ simple repeats. Therefore, further primers were designed within the Alu repeat between the TAs and the 5' break, avoiding reading through the TA repeats, while risking generating non-specific products caused by the Alu element. Despite the suboptimal PCR conditions, a unique product was amplified in the proband and the mother, and not in the control, as shown in **Figure 40a.** Subsequent dideoxy-sequencing further confirmed the 5' break despite noisy baseline from the small amount of non-specific product, as shown in in **Figure 40b.** From the dideoxy-sequences, this INV event can be described as:

NC_000007.14:g.[14567813_14567814insG;14567814_18889829inv;18889830delinsTA].

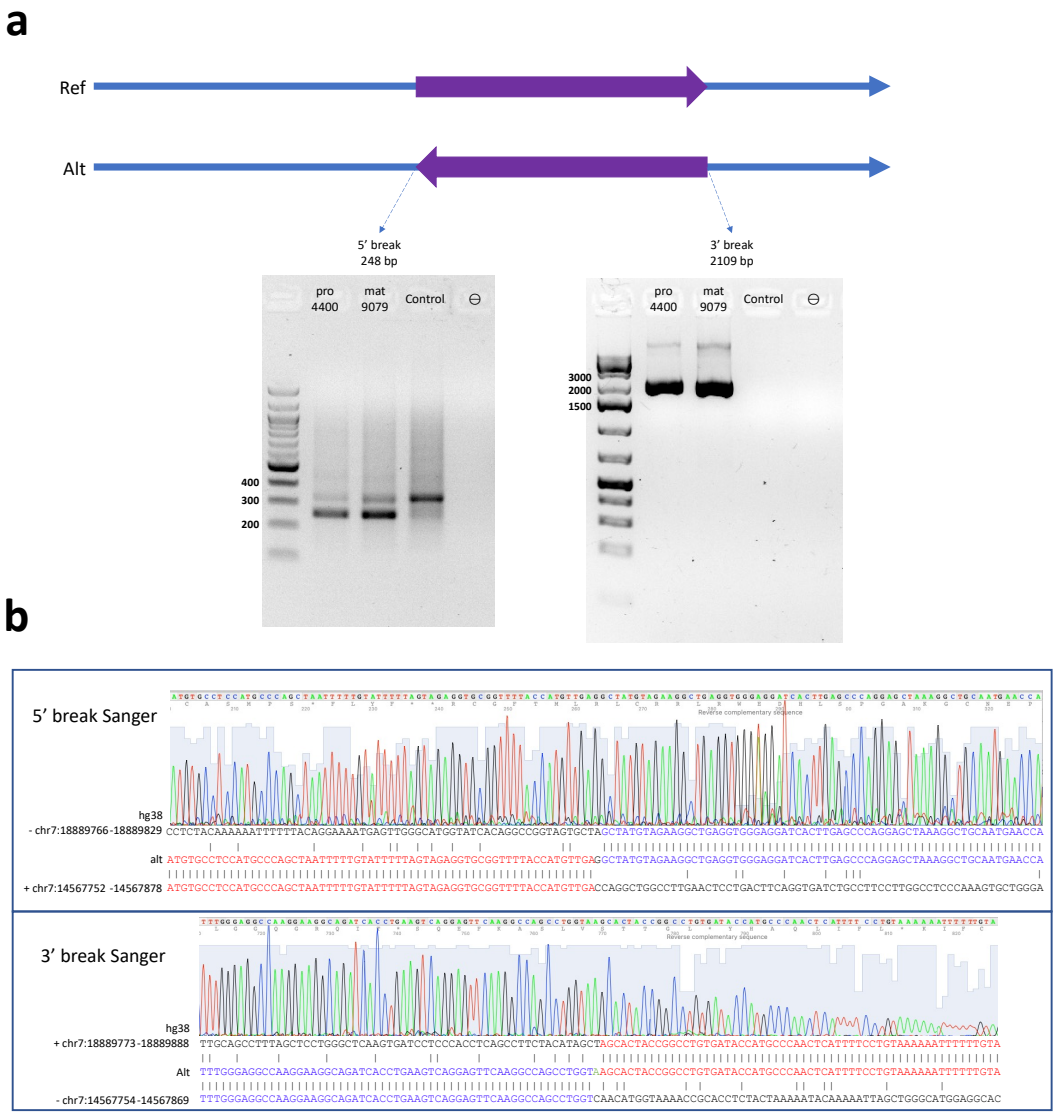


Figure 40 Breakpoint PCR and dideoxy-sequence verification of the INV. ***a****. Breakpoint PCR amplified unique break junctions in the proband and the mother.* ***b****. Subsequent dideoxy-sequencing characterised the detailed break junctions. ⊖ = negative control/ water. Figure in hg38.*

This case and another case (case 22**, Appendix, Supplementary Figure 2**) of Saethre-Chotzen syndrome characterised by WGS and confirmatory breakpoint PCR are crucial demonstrations that disruption of *TWIST1* long-range regulatory elements can cause Saethre-Chotzen syndrome. Both the balanced translocation (case 22) and INV (case 12) result in the repositioning of the regulatory elements upstream of the

TWIST1 locus. In the *HDAC9-TWIST1* region, six transcriptional enhancers have been established as critical regulators of *TWIST1* activity during craniofacial development, with an additional 15 potential enhancers proposed on mouse models.[147] These enhancers ensure robust *TWIST1* expression throughout craniofacial and limb development. Therefore, the misplacement of these enhancers is highly likely to be disruptive for *TWIST1* regulation in these two cases.

Similarly, craniofacial malformation has been well-documented in multiple cases with SVs affecting the *HDAC9-TWIST1* enhancers and the associated regulatory region, including DELs affecting *TWIST1* enhancers in the *HDAC9* coding regions[147,148] and translocations in the *HDAC9-TWIST1* intergenic region[149]. These cases led to the identification of a critical regulatory region of *TWIST1,* as highlighted in **Figure 41**. In comparison, case 12 INV and case 22 translocation break this critical regulatory region, and therefore likely disrupting *TWIST1* regulation, similar to other cases summarised in **Figure 41**.

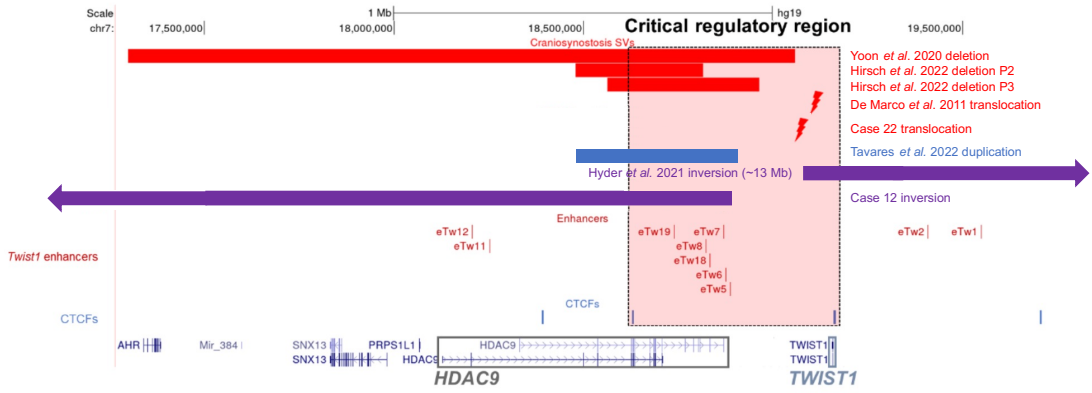


Figure 41 SVs affecting the TWIST1 critical regulatory region. *DELs, DUP, INVs, and translocations have been previously reported to cause craniofacial malformation without directly affecting the TWIST1 coding region. Figure adapted from Hirsch et al (2022)[147] with additional SVs from Yoon et al (2020)[148], De Marco et al (2011)[149], Tavares et al (2022)[150], and Hyder et al (2021)[90]. Figure in hg19.*

Importantly in the context of this chapter, Bionano OGM demonstrated its effectiveness in successfully identifying the causative SV in case 12, a task that had previously proven beyond the scope of several other detection methods. Upon retrospective analysis of the previous data, it became evident that Manta variant caller had indeed flagged two unpaired break ends (BNDs) from the targeted genome sequencing in case 12. However, these signals had not been further investigated during the diagnostic analysis due to the difficulties in interpreting unpaired BND calls, compounded by the high frequency of false positive BND calls made by the Manta caller. From the targeted resequencing data, the proximal break was identifiable with multiple abnormal reads. However, the distal break, due to the size of the INV, was beyond the captured region included in the experimental design. In comparison, OGM was both effective and efficient in the SV detection, setting a confident foundation for using this novel approach to tackle missing diagnoses in CRS.

5.3 Case 19: complex INS affecting *ENPP1*

Case 19 involves a non-CRS collaboration project with the Wellcome Center of Human Genetics (WCHG), and was recruited to test the robustness and effectiveness of Bionano OGM in characterising complex SVs.

Case 19 arises from a consanguineous family recruited to 100kGP, in which the proband was initially diagnosed as having *ABCC6* deficiency-induced generalised arterial calcification of infancy 2 (GACI 2; OMIM #614473), including necrotising enterocolitis, calcification of joints and wall of the descending aorta, supernumerary teeth, and vitamin D deficiency rickets. A heterozygous missense SNV in the recessive

ABCC6 was initially considered a strong candidate, however there was a lack of a second hit identified in this haploinsufficient gene. In the meantime, using SVRare, the WCHG collaborators identified a complex homozygous split-DUP possibly affecting a different gene associated with GACI.[151] Based on the read depth information, both parents were heterozygous carriers of this SV without any clinically significant phenotypes.

From the short-read WGS data, two regions of CNV gains were linked via paired/split reads information, as illustrated in **Figure 42a**. Two alternative structures, C19_alt1 & C19_alt2, could be constructed to explain the abnormal coverage and reads, as shown in **Figure 42b**. The two alternative CNV gains are located near either *LAMA5* or *ENPP1* locus. If C19_alt1 were true, no genes would be directly disrupted as the DUPs would be in an intergenic region near *LAMA2*. However, if C19_alt2 were true, *ENPP1* would have been broken between exon 1 & 2, causing a truncated gene/protein.

From the Bionano OGM data, a DUP was successfully called at chr6:131,838,080-131,854,801 (hg38) at the *ENPP1* locus. Closer scrutiny showed that the *LAMA5* locus was not disrupted, while the *ENPP1* locus consisted of inverted & interlinked DUPs exactly as predicted in C19_alt2, as shown in **Figure 42c**. This was further validated by examining individual molecules, where the C19_alt2 structure is supported by multiple spanning molecules as shown in **Figure 42d.** Upon the confirmation of the *ENPP1* disruption, the clinical diagnosis, and more importantly, the treatment of case 19 was updated to reflect best practice in GACI1 (OMIM #208000)/ ENPP1 deficiency.

This case is an excellent example demonstrating that Bionano OGM can be highly effective and efficient in detecting specific SVs. Due to the size of the SV at ~60 kb, a karyotype could not detect this SV, and only an ultra-high-resolution array could do so reliably. Short-read technologies, such as Illumina WGS, successfully identified the SV but failed to resolve between the two alternative hypotheses. Long-read technologies, such as PacBio and ONT, would likely be effective in resolving this SV, provided they met the crucial requirement of including multiple independent reads able to completely span the two DUPs, which were sized at 23 kb (*LAMA2*) and 17 kb (*ENPP1*). However, even if able to fulfil this requirement, the run time and data analysis may not have been as efficient for PacBio or ONT as with Bionano OGM. Quantitative reverse transcription PCR (RT-qPCR) was another technique attempted by the WCHG collaborators that failed to resolve the SV in timely fashion initially. In comparison, Bionano OGM successfully resolved the SV within 3 days, including DNA extraction, labelling, Bionano runtime, and analysis. This illustrated that Bionano OGM can be both effective and efficient in a clinical setting to resolve medium-sized CPX SVs and provide crucial molecular diagnosis to patients. Further information about this case and the molecular analysis were described in detail in Moore et al (2023)[151].

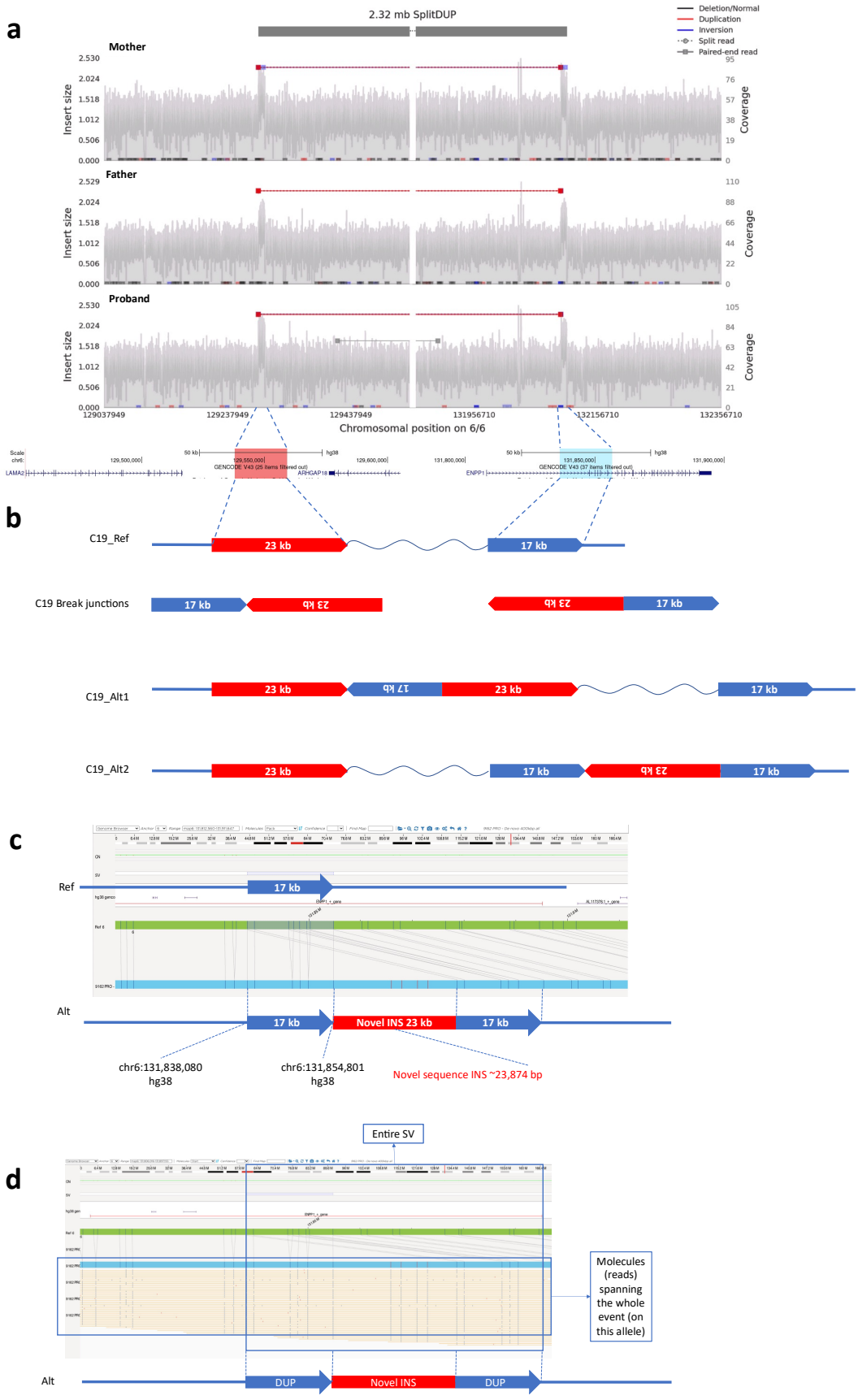


Figure 42 The chr6 SV in Case 19 was identified by WGS and subsequently fully characterised by Bionano OGM. a*. The SV was initially identified by WGS, showing two*

interlinked DUPs at either the intergenic region near LAMA2, or within ENPP1. Samplot highlighted two regions of CNV gain linked by abnormal reads. ***b****. Two possible alternative configurations of the SV can explain the WGS data. Alt 1 depicts a 40 kb INS at the intergenic region near LAMA2. Alt 2 shows the 40 kb INS positioned within ENPP1, breaking its continuity.* ***c****. Bionano OGM generated an abnormal map including both SV loci, with the labelling pattern supporting the Alt 2 configuration.* ***d****. Detailed analysis and alignment of individually labelled molecules showing molecules spanning the entire SV event, providing irrefutable evidence for the Alt 2 configuration. Samplot figure exported from GE Airlock. Figure in hg38.*

5.4 Case 10: CPX INS at *FGF9 locus*

Case 10 is a trio family enrolled in the 100kGP in which I identified two potential contributors to the phenotype – a *de novo* single nucleotide deletion in *FOXP2*, and two *de novo* DUPs on chromosome 13, sized at 323 and 244 kb. This section describes the further investigation and characterisation of the DUPs, and a functional investigation of whether these SVs might be contributing to the phenotype.

The proband presented with sporadic syndromic sagittal CRS, whilst her parents are unaffected. Her phenotype is characterised by sagittal CRS, hearing loss, low-set ears, hypoplastic ear canals, developmental delay, linear cutis aplasia on the anterior skull, and speech difficulties (**Figure 43**). She underwent multiple otorhinolaryngology interventions for bone-anchored hearing aids. However, she continues to experience speech and language difficulties with limited vocabulary. Notably, her comprehension skills are slightly better than her expressive abilities.

The proband, when assessed at the age of 5 years, was under regular monitoring for her narrow-elongated skull after initial surgical intervention, although no further surgical interventions were planned. She also presented significant visual acuity asymmetry, necessitating the use of corrective glasses and patches, with regular

optometry reviews. Additionally, she experienced issues related to constipation and continued to require nappies owing to inconsistent recognition of stimuli. Physical examination places her stature in the 6th percentile for her age, and her stocky build in the 90th percentile. Despite these challenges, she maintains good overall general health.

Figure 43 Case 10 is a family with sporadic syndromic CRS. *The proband presented sagittal synostosis, hearing loss, low set ears, hypoplastic ear canals, developmental delay, linear cutis aplasia on the forehead, and speech and language problems. Permission to include clinical photographs was given by the patient and/or their parents.*

5.4.1 SV identification from WGS

Array CGH was conducted by the clinical laboratory, which only detected a paternally inherited 43 kb DUP on 16q12.2q13. This was deemed a likely incidental finding owing to the paternal inheritance. Without a clear molecular diagnosis, the case 10 family was recruited to the 100kGP. From the 100kGP WGS data, a CPX interlinked split-DUP was detected on chr13, containing a **323 kb segment** (referred to as **proximal DUP**), a **244 kb segment** (referred to as **distal DUP**), and a small **147 bp INV**, as shown in **Figure 44**. The “Contact Clinician" process was followed within the 100kGP

RE and the family trio was recruited to GBoCM study. Stored DNA was obtained for PCR characterisation of the breakpoints. Subsequent breakpoint PCR and dideoxy-sequencing verified the two possible break junctions of this SV, as shown in **Figure 45**. The extensive homology at the **distal DUP-proximal DUP** junction suggests a NAHR mechanism. Conversely, the lack of homology implies that NHEJ mechanisms may have been responsible for the generation of the other break junction. Lastly, the origin of the duplicated segments was investigated by extracting individual SNPs from the two large segments. As illustrated in **Figure 46**, the two paternal copies of the **323 kb proximal DUP** in the proband originated from two different alleles of the father, while the two paternal copies of the **244 kb distal DUP** in the proband originated from the same paternal allele. This further demonstrated the intricate molecular mechanism of this complex SV.

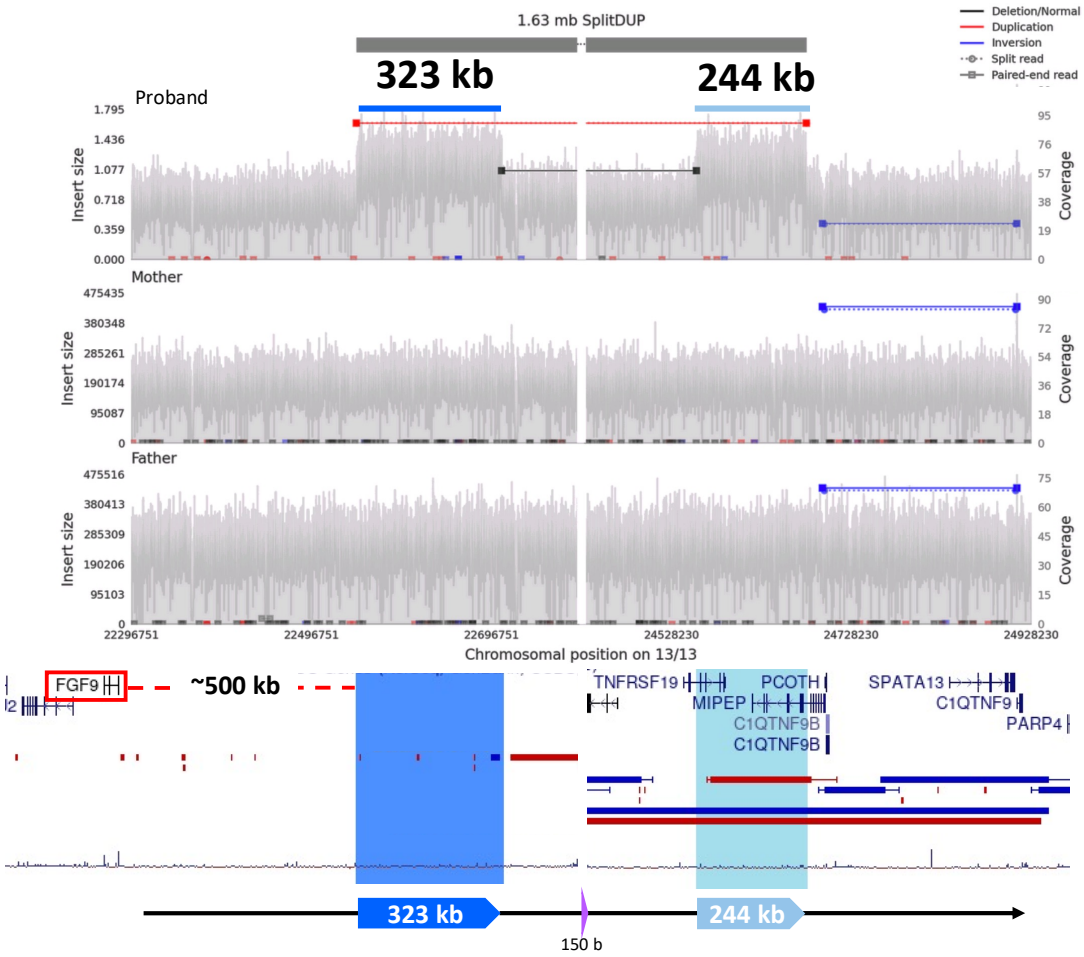


Figure 44 100kGP WGS analysis detected a de novo interlinked split-DUP on chr13. *Bioinformatic identification of this rearrangement was achieved by Dr Eduardo Calpena before the start of my DPhil studies, while I later discovered the complex nature of this event. Three segments were detected at 150 bp, 323 kb, and 244 kb, illustrated as* ***purple****,* ***dark blue****, and* ***light blue*** *blocks, respectively. The* ***323 kb segment*** *is ~500 kb downstream of the candidate gene FGF9. Samplot figure exported from GE Airlock. Figure in hg19.*

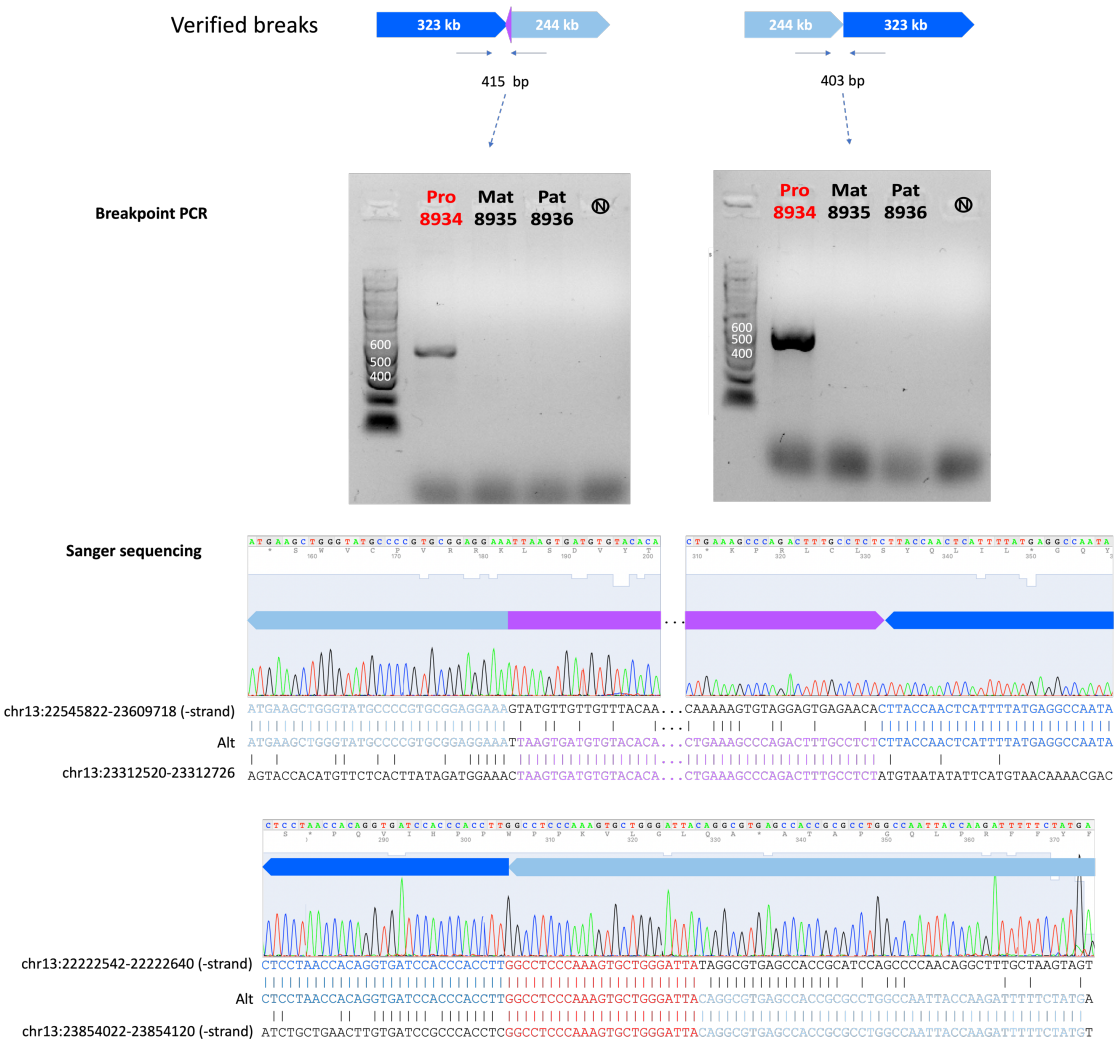


Figure 45 Break point PCR and dideoxy-sequencing verified the two break junctions. *Breakpoint PCR produced unique products in the proband, confirming the de novo nature of both events. Ⓝ = negative control/ water. Sequence coordinates were mapped to hg19.*

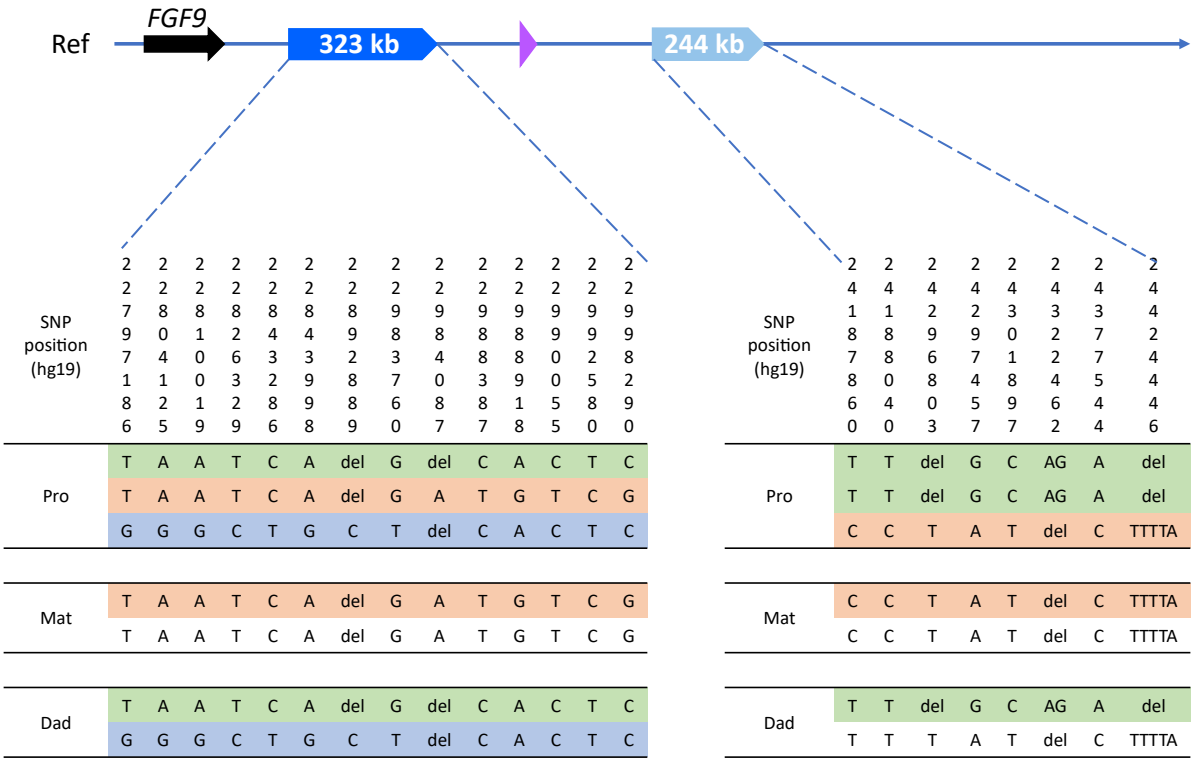


Figure 46 The origin of the two large, duplicated segments were investigated. *SNPs in the duplicated segments were collected. The proband was shown to have three alleles due to the extra copy number, while the parents have two alleles each. By matching the genotype of the proband to the parents, it was shown that the two paternal copies of the* ***323 kb proximal DUP*** *in the proband originated from two different alleles of the father, while the two paternal copies of the* ***244 kb distal DUP*** *in the proband originated from the same paternal allele. The same shading indicates the alleles presented the same genotype.*

The **proximal DUP** (chr13:22222612-22545849, hg38) is located within an intergenic region approximately 500 kb downstream of the candidate gene *FGF9*. *FGF9*, encoding one of the fibroblast growth factors (FGFs), is a known disease gene for multiple synostoses syndrome 3 (SYNS3) and CRS, whereby LoF mutations severely hinder the interaction between the mutant FGF9 and its receptor FGFR3, resulting in impaired FGF signalling.[152,153] More specifically relevant to the sagittal CRS phenotype in the proband, Rodriguez-Zabala et al (2017) described a family in which both the proband and the father exhibited sagittal CRS. However, the case 10 phenotype diverged from the Rodriguez-Zabala et al (2017) family in terms of the lack

of any joint, palate, or eye abnormalities. This comparison suggests the possibility that case 10 may share some features of *FGF9*-related phenotypes. However, the dissimilarities observed between case 10 and other *FGF9* cases imply that the effect of *FGF9* on case 10's pathology may involve molecular mechanism beyond simple LoF.

The **distal DUP** (chr13:23609688-23854091, hg38) contains segments of *TNFRSF19* and *MIPEP*. *MIPEP* is a known recessive disease gene associated with Combined Oxidative Phosphorylation Deficiency 31 (COXPD31, OMIM 617228)[154]. The clinical relevance of these two genes appears limited given their low pLI scores, low to moderate pTriplo scores (less dosage sensitive), and the absence of any documented shared phenotypes between these genes and the case 10 proband. In addition, a similar CNV gain to the **distal DUP**, Esv3631529, has been previously reported in the 1000 Genomes Consortium (*via* DGV[155]) without any clinical features. This suggest that the distal DUP on its own has been carried by a healthy control in the general population. However, this does not preclude any positional or regulatory effect the **distal DUP** may have when linked with the **proximal DUP**.

The **small INV** (chr13:23312552-23312698, hg38) corresponds to a LINE repeat - L1PB4 (5935-6081). The **small INV** is positioned within the intron of *SGCG*, adjacent to *SACS*. Interestingly, both *SACS* and *SGCG* are known recessive disease genes associated with Charlevoix-Saguenay type Spastic Ataxia (OMIM 270550) and Muscular Dystrophy (OMIM 253700), respectively[156-158]. Similar to that of the **distal DUP**, the **small INV**, when considered in isolation, is likely a benign polymorphism

due to its intronic location, the relatively short sequence span, and the lack of known phenotypes shared with the Case 10 proband.

Based on the verified breaks and the two-fold coverage changes, three hypothetical configurations of the event can be constructed to explain the WGS data, as illustrated in **Figure 47**. The first two configurations suggest only one allele is affected in the proband, while the other allele remains as reference. These two configurations, Alt 1 and 2, suggest that the SV event occurred either near the *FGF9* locus, approximately 500 kb downstream of *FGF9*, or at the *MIPEP* locus, which is 2 Mb away from *FGF9*. A third hypothetical configuration, Alt 3, implies that the event occurred in *trans*, whereby both proband's alleles were disrupted. In this configuration, one allele involves a tandem DUP, and the other allele contains a CPX event consisting of DEL-INV-DEL. However, it's important to note that the *trans* configuration is biologically highly unlikely due to the *de novo* nature of the event. Consequently, for the candidate gene *FGF9*, all three configurations, Alt 1, Alt2, and Alt3, could potentially have distinct regulatory ramifications owning to their positional effect on the adjacent genomic environment.

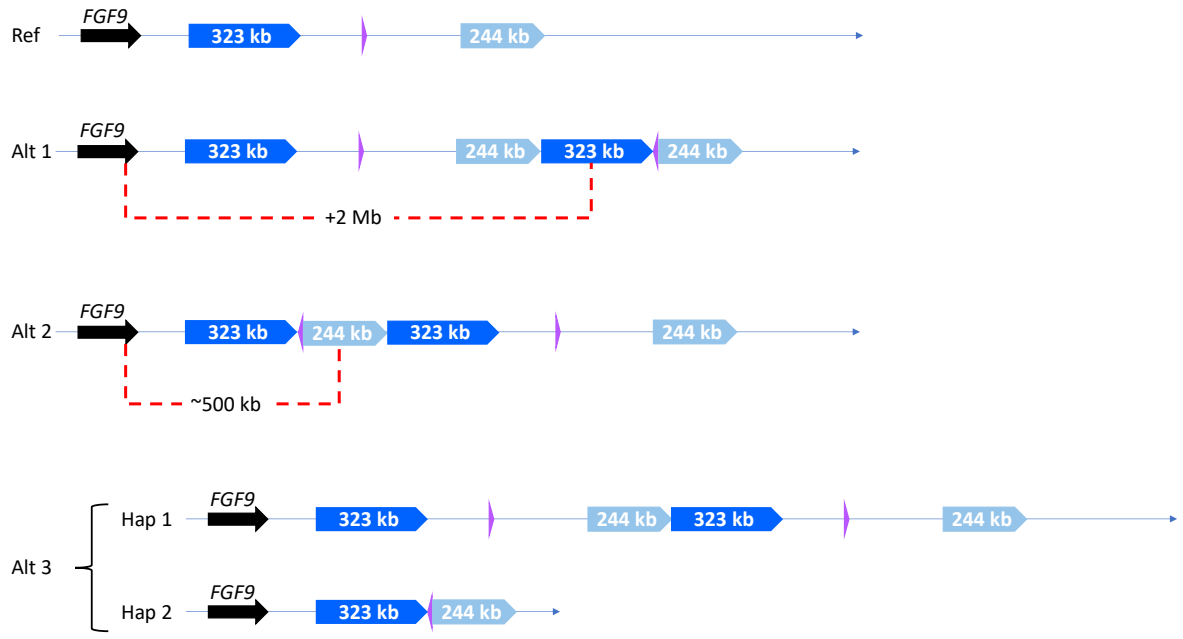


Figure 47 Three alternative configurations for case 10 CPX SV. *Alt 1, Alt 2, and Alt 3, can explain the WGS abnormal reads and coverages on chr13 near the FGF9 locus. In Alt 1 and Alt 2, only one of the proband's allele is affected, while in Alt 3, both alleles from the proband could be affected.*

5.4.2 SV characterisation with Bionano OGM

Bionano OGM, having shown a promising outcome in characterising the complex SV from case 19, was employed to determine the true configuration of this SV event. The Bionano OGM data, as shown in **Figure 48**, successfully constructed the abnormal map of interest at the SV locus on chr13 in case 10. Also illustrated in **Figure 48**, five molecules were detected to completely span the critical informative segment, comprising the smaller (244 kb) of the two DUP segments. Based on these five informative molecules, a recognisable pattern emerges as **proximal DUP-distal DUP-proximal DUP**. When compared to the three possible configurations (**Figure 47**), this specific pattern is exclusively found in the Alt 2 configuration. Note that the size of the small INV (~150 bp) is significantly smaller than the inter-label distance, making its detection impossible for the current Bionano OGM technology. From this analysis,

Bionano OGM strongly supports the Alt 2 configuration as the likely SV carried by the proband.

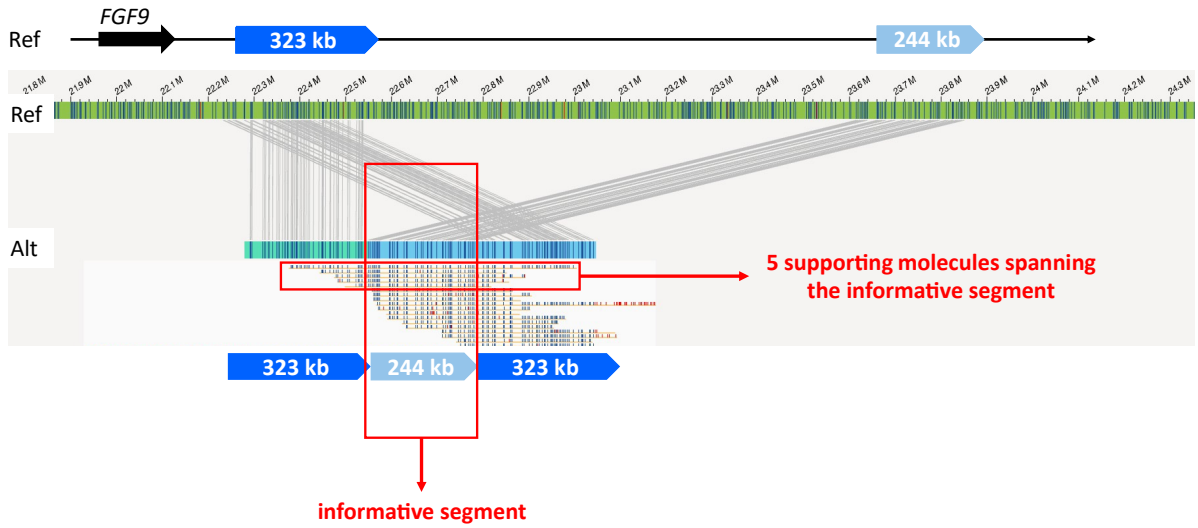


Figure 48 Bionano OGM detected five molecules successfully spanning the 244 kb informative segments. *By comparing to the reference labels and illustration, a pattern unique to Alt 2 can be constructed as* ***323 kb-244 kb-323 kb****. Figure in hg38.*

5.4.3 SV characterisation with FISH

However, considering the relatively limited support with only five reads from the Bionano data, FISH was then undertaken under the supervision of Jill Brown (HMU, MRC WIMM) to validate the SV configuration in accordance with the diagnostic guidelines. To enable cellular studies, further consent was obtained to enrol the patient in the iPSC-CRS study and a fresh blood sample was obtained for iPSC derivation. This was then used for FISH and subsequent RNA-Seq (**section 5.4.6**) studies. A three coloured FISH was designed, as shown in **Figure 49**. FISH was carried out using iPSCs derived from the case 10 proband by Dr Dagmara Korona, in comparison with a random healthy control. The three expected signals can be summarised as follows:

- Reference: OOO
- Alt 1 OOOOO + Ref
- Alt 2 OOOOO + Ref
- Alt 3 OOOOO + OO

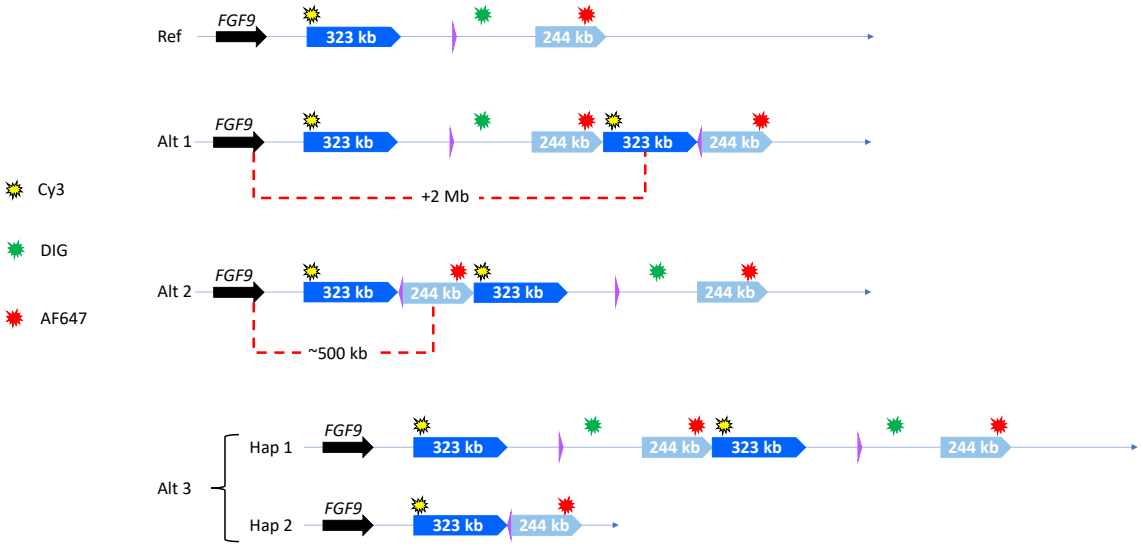


Figure 49 A three coloured FISH was designed to verify the true configuration of the case 10 SV.

A total of 334 cells and 645 allele signals were examined to determine the true configuration of the case 10 SV. The majority (65%) of the signals were non-informative, which were overlapped signals, overexposed signals, large regions of noise signals (contamination), or signals from likely replicating cells. Replicating cells accounted for most of the non-informative signals. This is likely due to the fast-growing nature of the patient-derived iPSCs. From the control sample, 90% informative signals supported the reference pattern, one signal had a pattern resembling alt 2, and 9%

informative signals did not support either the ref or the alt patterns. Two example cells showing the reference pattern (OOO) are shown in **Figure 50**.

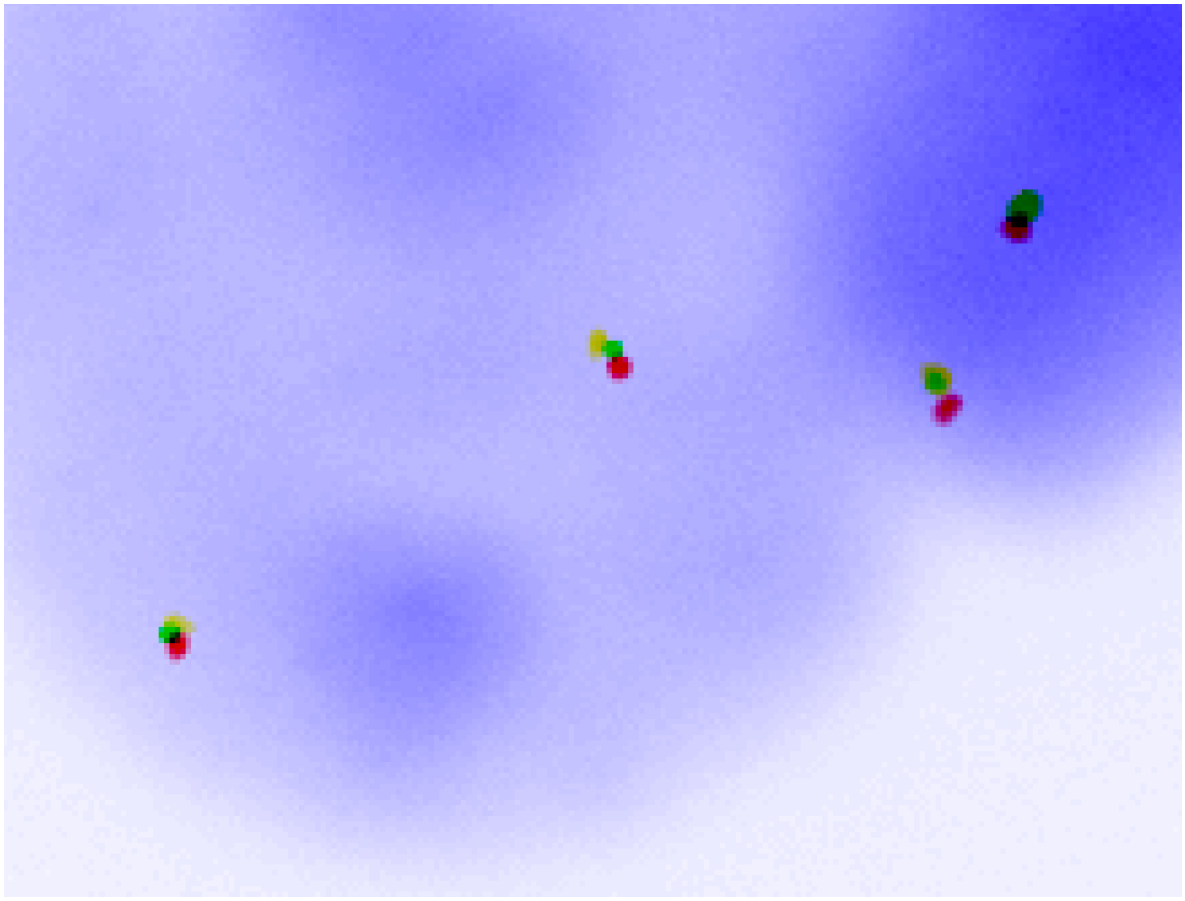


Figure 50 An image from the FISH analysis of the control sample. *Figure shows two cells (four allele signals) consistent with the expected reference signal pattern (OOO).*

From the patient samples, 47% informative signals were consistent with the reference pattern, suggesting the patient is likely heterozygous, ruling out the possibility of the alt 3 configuration with the *trans* structure where both alleles are affected. Another 31% signals showed an alt 2 or likely alt 2 pattern, and only one signal was most likely the alt 1 interpretation. A total of 21% signals were not consistent with either ref or the alts, and these were likely partially non-informative signals (label dropout) leading to the appearance of a completely different pattern. **Figure 51** shows four example FISH

images supporting the Alt 2 configuration, with each cell containing one reference signal (OOO) and one Alt 2 signal (OOOOO). Only one informative cell from the patient samples showed a reference pattern on both alleles, suggesting either the presence of contamination or extremely low level of mosaic cells that are reference.

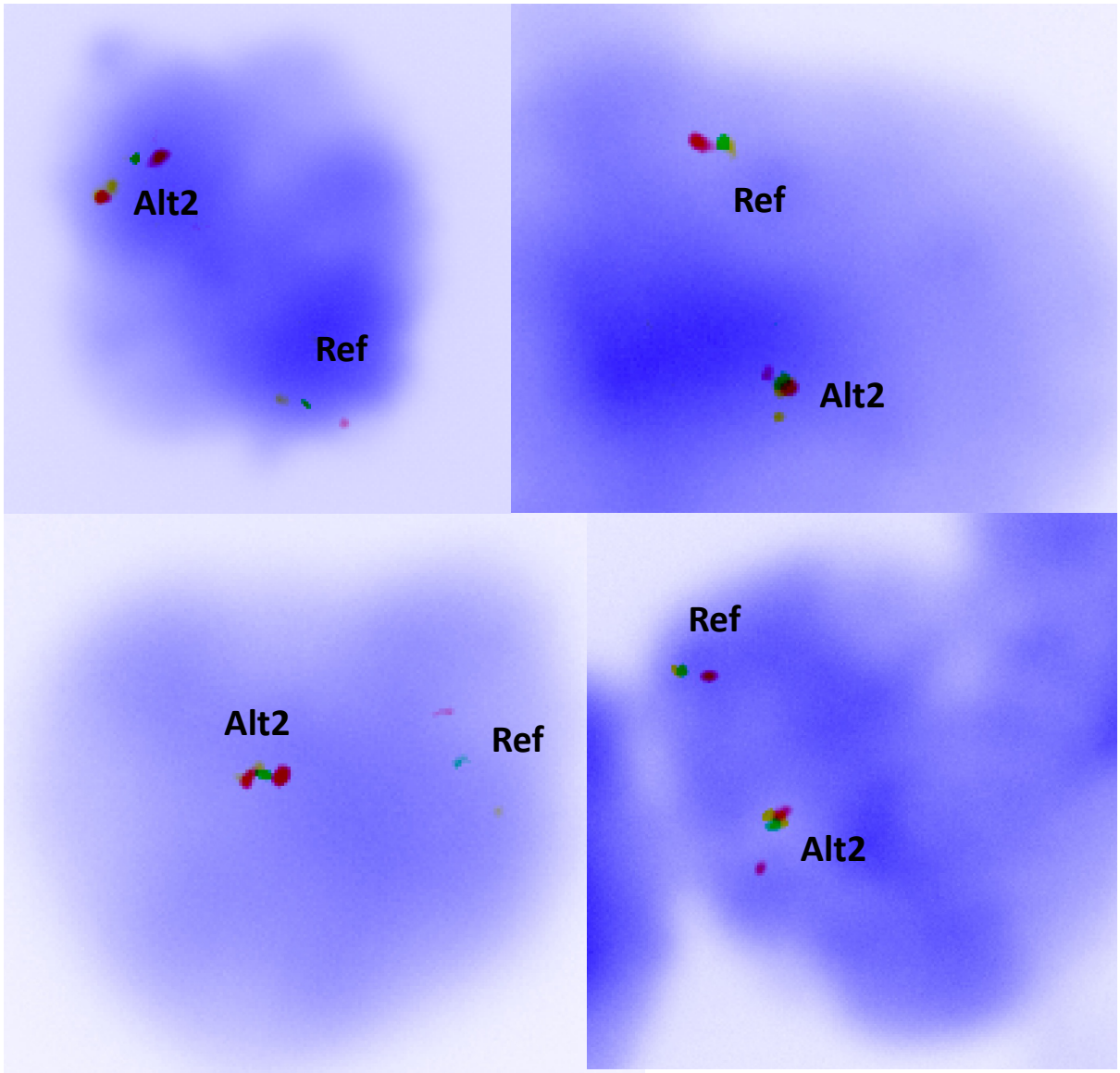


Figure 51 Four example FISH images from the patient cell line. *Figure shows cells presenting signals with one ref pattern (OOO) and one Alt 2 pattern (OOOOO).*

Overall, the FISH result is consistent with the Bionano OGM data, suggesting Alt 2 is the true configuration carried by the proband. With the confirmed configuration, it is therefore important to understand the functional impact of Alt 2 on the candidate gene *FGF9*.

5.4.4 Functional analysis: DeepC

To delve deeper into the regulatory implications of Alt 2 on the candidate gene *FGF9*, DeepC[128] was employed to predict the effect of the Alt 2 SV on nearby TADs associated with *FGF9*. **Figure 52** illustrates the true Hi-C data (IMR-90 lung fibroblasts) and the DeepC predictions for both the reference and the Alt 2 sequences. Firstly, in comparison against the data track, the algorithm in reference track successfully predicted the two significant TAD boundaries between the *FGF9* locus and the *MIPEP* locus. This attests to the precision of DeepC within these relevant regions. Secondly, from the data/reference track, *FGF9* was shown to reside under the same TAD (referred to as the *FGF9* TAD) as the **dark blue proximal DUP segment**. This is crucial information, as any disruption of this *FGF9* TAD has the potential to impact *FGF9* regulation. Finally, in the Alt 2 track, the *FGF9* TAD is notably expanded due to the presence of the two inserted segments. This suggests that expression of *FGF9* could be affected due to disruption of the regional TAD structure because of the SV in Alt 2 configuration.

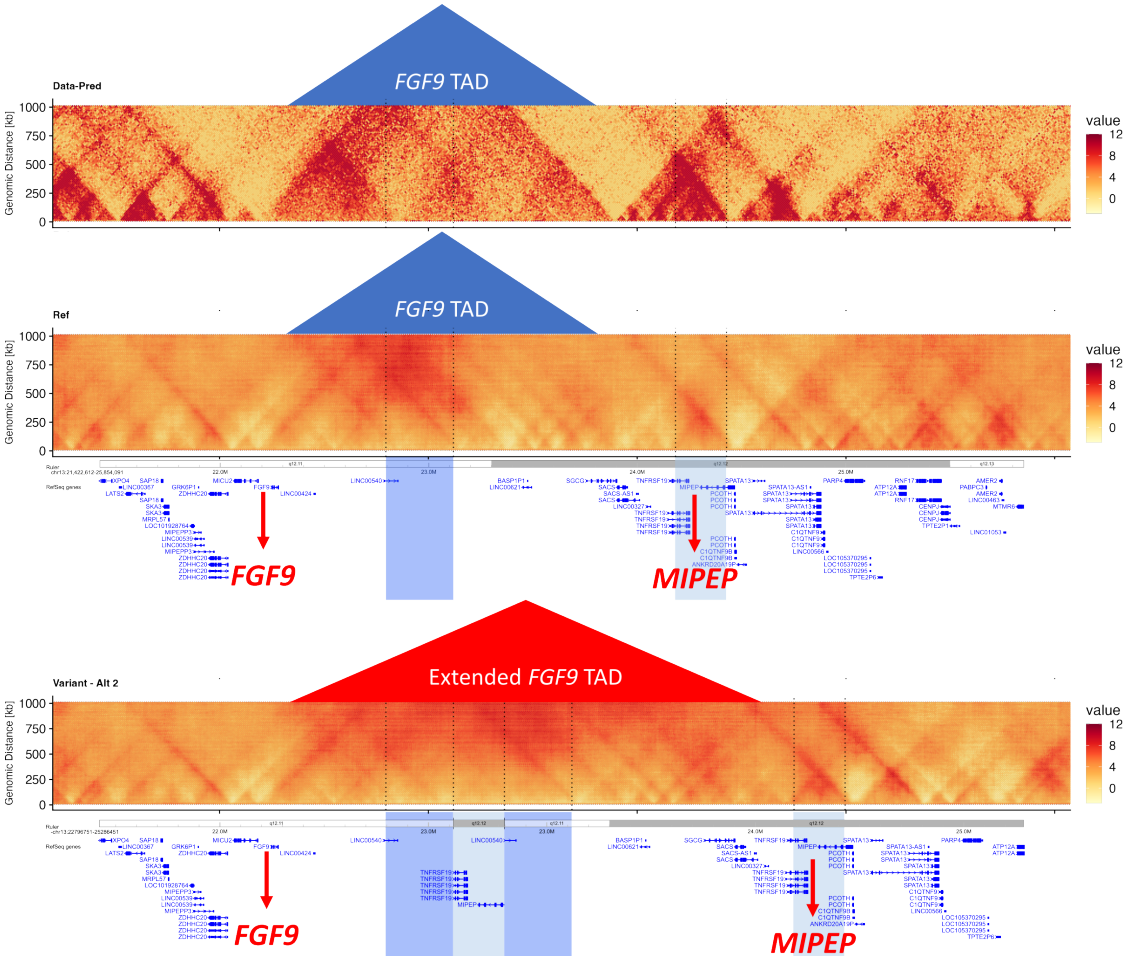


Figure 52 DeepC prediction suggests an expansion of the FGF9 TAD due to the SV in Alt 2 compared to the reference. *Prediction was carried out using IMR-90 line data. Figure in hg19.*

5.4.5 A competing hypothesis: *FOXP2* frameshifting DEL

A competing hypothesis in case 10 involves the identification of a *de novo* single nucleotide DEL, which was detected through GE's analysis pipeline. I verified this DEL using PCR and dideoxy- sequencing in the case 10 family, as shown in **Figure 53**. This specific DEL is NC_000007.14:g.114642616del (hg38, chr7:114642616del) and is located in the coding region of the *FOXP2* gene. *FOXP2* is a known dominant (monoallelic) disease gene for speech-language disorders, particularly childhood apraxia of speech.[159] Additional features may include speech development delay, oral-motor function difficulties, lower average intelligence quotient (IQ), as well as fine and

gross motor impairments. Strikingly, these features align closely with the observed phenotypes in case 10.

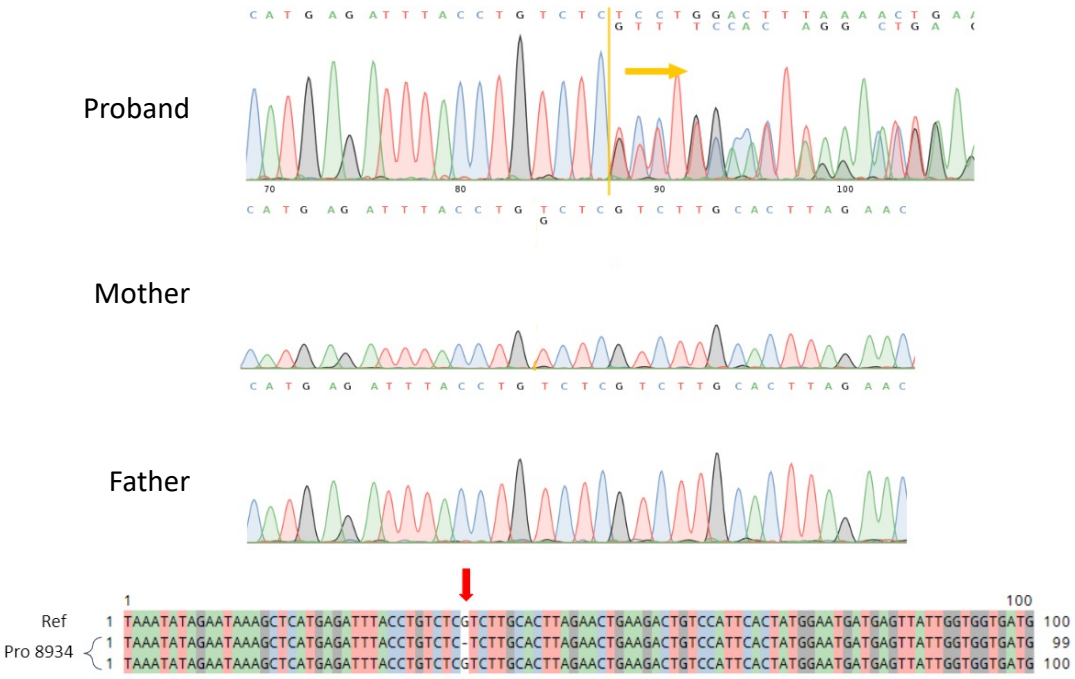


Figure 53 Dideoxy- sequencing analysis of the de novo single nucleotide DEL affecting FOXP2 in case 10. *Double peak in the proband suggests a DEL event, where the parents showed reference sequences.*

However, the association between *FOXP2* and CRS is more difficult to interpret. For example, a C-T transition (NC_000007.14:g.114642616C>T) leading to a stop-gain variant (c.982C>T, R328X) has been documented in three family members with verbal dyspraxia.[160] Despite the loss-of-function *FOXP2, no* craniofacial anomalies were observed. In contrast, another case featuring a frameshift DEL, c.484del, p.(Gln162Asnfs*100), was detected in a family with both speech-language disorder as well as syndromic sagittal CRS.[161]

Within the context of case 10, the *FOXP2* frameshift matches well with her speech and language difficulties. Furthermore, the sagittal CRS is also consistent with the phenotype detailed in Tonne et al (2022).[161] Nevertheless, case 10 appears to have a higher degree of cognitive impairment than typically expected in *FOXP2*-related cases. This discrepancy raises the potential for either a secondary cause contributing to her ID, or a high level of clinical heterogeneity within *FOXP2*-related disorders. Overall, the presence of this *de novo* single nucleotide DEL, in addition to the *de novo* chromosome 13 SV, adds a substantial layer of intricacy to the process of relating genotype to phenotype in this patient.

5.4.6 Functional analysis: RNA-seq

Simultaneously, RNA-seq was carried out to scrutinize the transcriptome at different stage of development, using patient-derived iPSC differentiated to neural crest cells and osteoblasts (prepared by Dr Dagmara Korona). While the data from other stages are still in progress, here I present my preliminary analysis of the data obtained from late-stage neural crest cells. **Figure 54a** summarises the transcriptomic overview of the data distribution, excluding genes encoded on the X and Y chromosomes and mitochondrially-encoded (MT) genes. Six genes have been highlighted in the volcano plot, comprising *FGF9*, four other genes (*MIPEP, SPATA13, SACS, SGCG*) adjacent to the SV, and the competing candidate *FOXP2*. Despite significant deviations from the control samples (padj < 0.05), none of the highlighted genes exhibited a pronounced level of differential expression in comparison to the broader transcriptome. A more localised view is shown in **Figure 54b**, highlighting genes within ± 5 Mb vicinity of the SV. Similar to the broader transcriptomic view, the differential expression of *FGF9* was not prominent even in the local context.

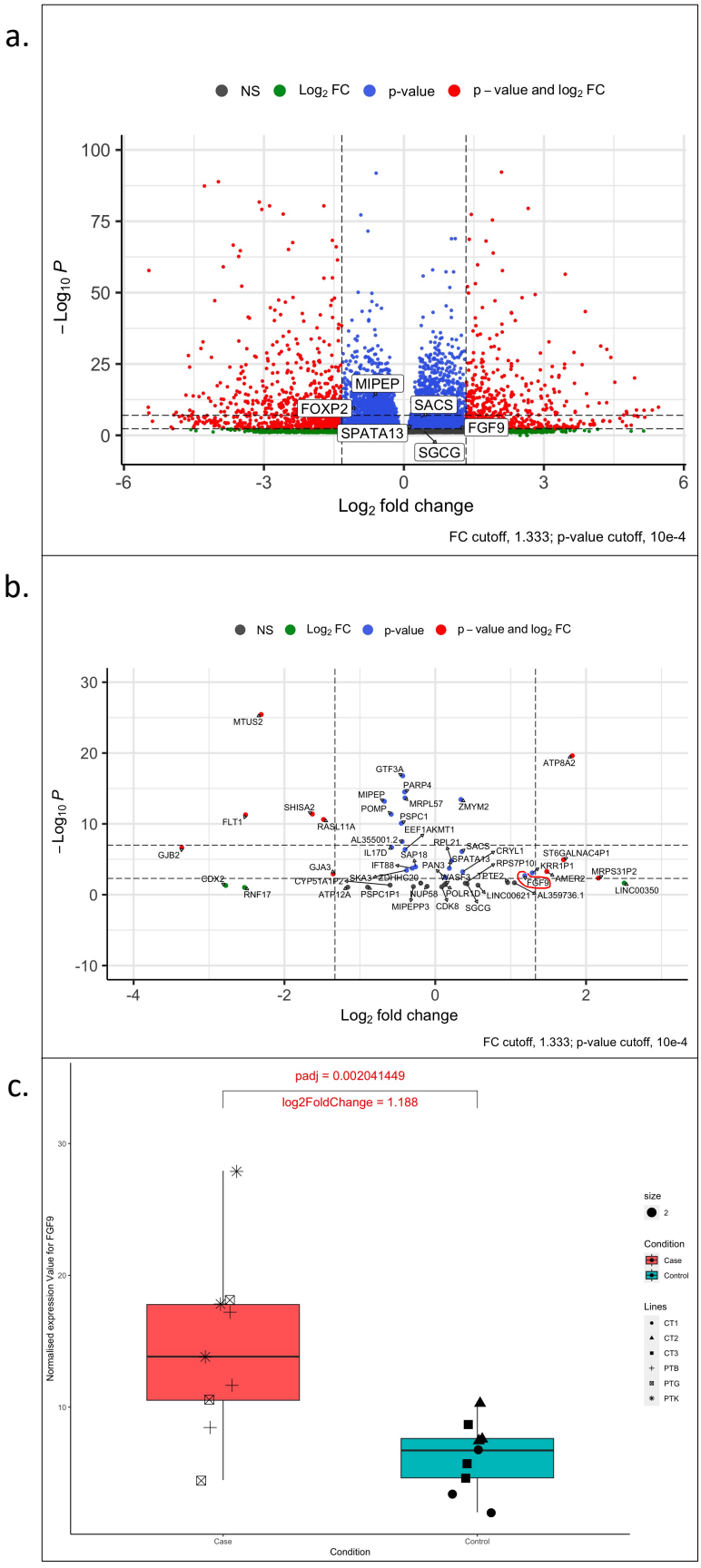


Figure 54 Bulk RNA-seq data visualised using volcano and box plot. a*. The transcriptome overview of the differential expression analysis, excluding genes from chrX, Y, and MT.* ***b****.*

*Local view of the differential expression analysis at the SV locus on chr13. In **a** and **b**, the X axis is log-transformed change of normalised expression level in the case compared to the control, and the Y axis is the log transformed adjusted p value. **c**. Differential expression analysis of FGF9. FC = fold change, NS = not significant, CT1-3: three technical replicates of each of three different control iPSC lines, PTB/G/K: three technical replicates of three separately picked iPSC clones derived from the patient (case10 proband).*

Further focusing on the specific gene, **Figure 54c** shows only a marginal 2.4-fold (padj = 0.002) upregulation of *FGF9*. This finding might be attributed to a composite result of misexpression, where *FGF9* is upregulated in certain cell types, while concurrently down regulated in others. Additionally, the large variance across the three biological replicates of the patient iPSCs (ie independently picked clones) strikingly contrasts with the more consistent expression profiles observed in the three biological replicates of the control line. This pattern is also consistent with the observation that the patient lines exhibit a significantly accelerated growth compared to the control lines, which may suggest the necessity of cell cycle correction in future analysis. Subsequently, this fast growth rate potentially introduces divergence in the developmental stages amongst the three patient lines, contributing to the large variance of the *FGF9* expression in the patient lines. Consequently, it is possible to hypothesise that *FGF9* expression is highly specific during various developmental stages, with the effect of the SV becoming apparent only in the later stages. The forthcoming RNA-seq analysis at a more advanced differentiated cell stage, specifically in osteoblasts, holds some potential to understand the effect of the chr13 SV on the temporal dynamics of *FGF9* expression.

Gene ontology analysis was also conducted to investigate pathways enriched in the differentially expressed genes. As illustrated in **Figure 55**, six pathways were shown

to be enriched within the differentially expressed transcriptome of the case 10 patient iPSC lines. Amongst the six pathways, the Cadherin and Wnt signalling pathways stood out as most significant with the lowest false discovery rate (FDR) values. Both Cadherin and Wnt signalling pathways have shown crucial roles in bone formation, particularly in processes such as osteoblast differentiation and calcium-dependent cell adhesion.[162] In addition, the Wnt signalling pathway has been implicated in promoting abnormal endochondral ossification in sagittal CRS[163] - a phenotype that holds particular relevance in the context of case 10. Most importantly, these findings are consistent with the patient's sagittal CRS phenotype, and therefore may facilitate the identification of additional causally contributing variants among the differentially regulated pathways in future analysis.

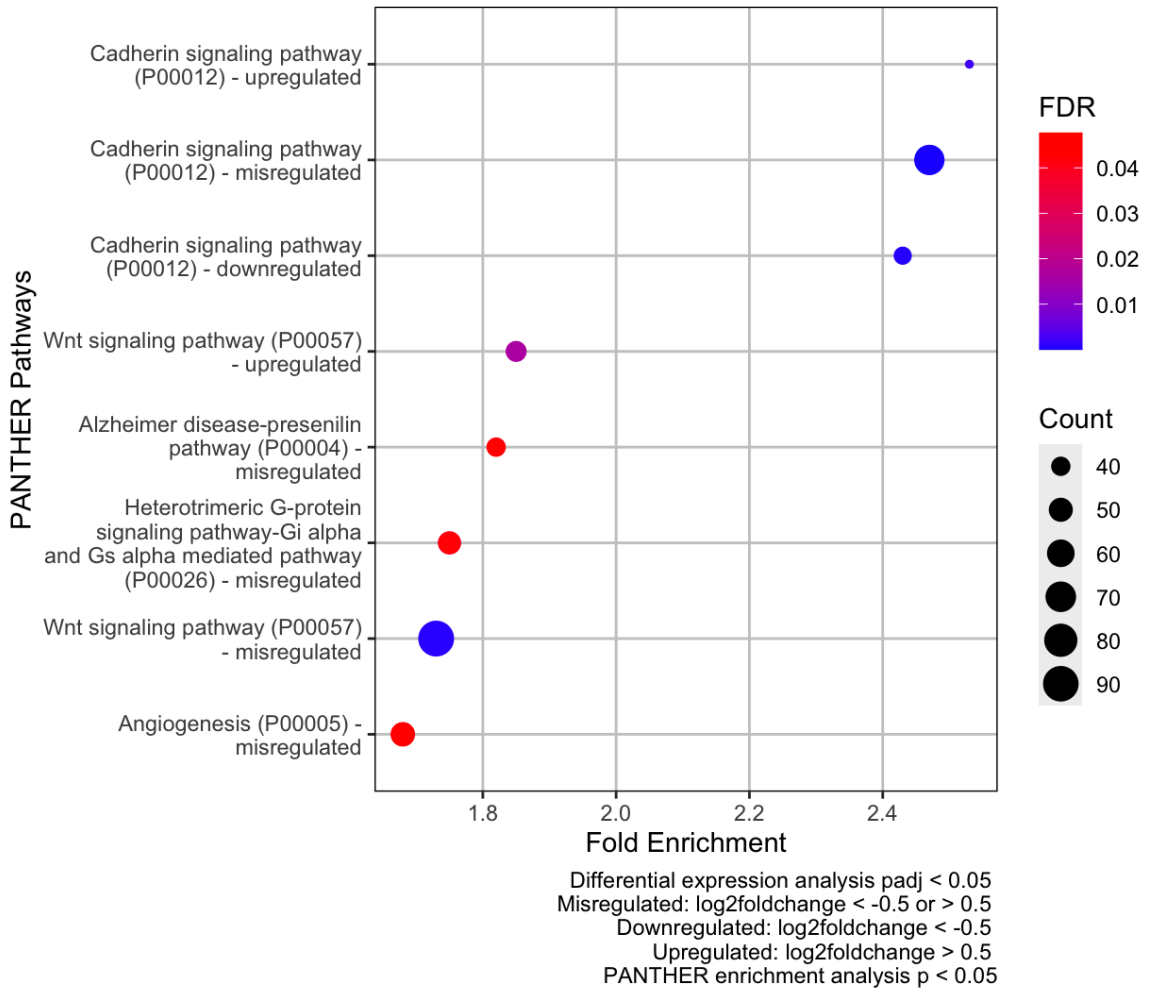


Figure 55 Gene ontology enrichment analysis was carried out against the PANTHER[164] pathway database with Fisher's Exact test. *X, Y, and Mitochondrial genes were excluded from the analysis. Data visualised in R. FDR = false discovery rate. Count = number of genes featured in the corresponding pathway.*

5.4.7 Case 10 summary: navigating the dual dilemma of competing hypotheses

In summary, I have presented a case involving syndromic sagittal CRS with speech-language disorder. The investigation unveiled two compelling *de novo* variants. The first variant was a *de novo* single nucleotide DEL affecting the speech gene, *FOXP2*. The other variant was a complex 567 kb SV consisting of interlinked split-DUPs near a known CRS gene *FGF9*. Due to the complex nature and the size of this SV, short-read data were insufficient to fully characterise this event. Aiming to tackle the short-

read challenge, I employed Bionano OGM, which successfully mapped five informative molecules, with each of them spanning over 250 kb to fully characterise the SV. This result highlighted Bionano OGM's ability to generate ultra-long molecules - a pivotal asset in understanding large complex SVs. Simultaneously, FISH analysis aligned with the SV configuration suggested by the Bionano.

To understand the effect of the SV on *FGF9*, DeepC prediction was carried out, and illustrated a potential expansion of the *FGF9* TAD. Subsequent bulk RNA-seq was used to examine *FGF9* and *FOXP2* expression, as well as the broader transcriptome of iPSC-derived neural crest cells from the case. Differential expression analysis identified a modest change in expression of both *FGF9* ($log_2FC = 1.2$, padj = 0.002) and *FOXP2* ($log_2FC = -1$, padj = 2.37E-10) in the patient line compared to the control. The *FOXP2* expression change is reasonable, as it is likely attributed to nonsense mediated decay caused by the frameshift SNV. Gene ontology analysis highlighted two pathways, cadherin and Wnt signalling pathways, that were enriched amongst the differentially expressed genes in the patient lines. These two pathways are critical in bone formation, with Wnt signalling pathway previously implicated in sagittal CRS[163], consistent with the proband's phenotype.

However, it has been increasingly challenging to disentangle the two competing hypotheses to identify the true causative variants, or a combined pathogenic effect from both, for the proband's phenotypes. The difficulty mainly arises from both candidates constituting rare *de novo* variants in a case with rare phenotypes. Extensive functional studies may be required. The ongoing osteoblast RNA-seq holds

the potential to understand the changes in the temporal dynamic expression of *FGF9* and *FOXP2*. Single cell RNA-seq may help explore the hypothesis that *FGF9* disruption specific affects cell types crucial in craniofacial development. A mouse model is usually another approach to establish causality, although its feasibility hinges on evaluation of the preliminary data, owing to the cost associated with the generation of an SV specific model and the potential for inconclusive results.

Overall, case 10 presented an intriguing scenario where Bionano OGM has been crucial in charactering a large complex SV, and subsequently in establishing the correct clinical interpretation of the SV. Concurrently, a competing hypothesis featuring a *FOXP2* SNV introduces an additional challenge to uncover the true causative variant behind the patient's phenotype. Current functional data strongly suggests the pathogenicity of the *FOXP2* SNV, while whether the complex *de novo* SV near *FGF9* adds additional pathogenicity remains unclear. This further emphasises the need for additional functional studies to gain a better understanding of the contributions of the two candidate variants.

5.5 Case 2: CPX SV at *KCNJ2* & *16* locus

Case 2 is a collaboration project between Great Ormond Street Hospital (GOSH) and Oxford involving a patient with an unusual syndromic CRS phenotype. Bionano OGM was applied to characterise the candidate complex SV involving large segments from both chr16 & chr17. While Bionano OGM successfully yielded informative findings, the complexity and the size of this SV pushed the boundaries of what Bionano OGM can

fully characterise. In such cases, orthogonal technologies, such as FISH, demonstrate potential advantages over Bionano OGM.

5.5.1 Case 2: syndromic CRS with unusual gum phenotype

Case 2 is a trio family with sporadic syndromic multi-suture CRS in the proband. The proband had three sutures affected: both coronal sutures and the sagittal suture. The patient was monitored, but craniofacial surgery was not deemed necessary since there was no signs of cosmetic deformation or evidence of raised intracranial pressure. The other defining syndromic features in this case are hypertrichosis and gum hypertrophy (**Figure 56**), which are rare features in syndromic CRS. To address the gum abnormalities, the proband underwent multiple gum debulking surgeries. Moreover, he had developed a marked pectus carinatum and a connective tissue disorder-like phenotype with pes planus. The parents are unaffected and in good general health. In the proband more than a decade ago (2011), the cytogenetics lab initially detected via array CGH (Nimblegen 135 K WG CGH v.3.1) three CNVs:

- 302.47 kb CNV gain at 17p11.2, affecting *FAM83G, GRAP, PRPSAP2,* and *SLC5A10*;
- 1,905.29 kb CNV gain at 16q23.1, affecting *ADAMTS18*, *CLEC3A*, *KIAA1576*, *NUDT7*, and *WWOX*;
- and 478.73 kb CNV gain at 17q24.3, affecting *KCNJ16* and *KCNJ2*.

The diagnostic lab further undertook quantitative polymerase chain reaction (qPCR) for the 16q CNV and identified the *de novo* nature of this event, while parental testing for the 17 CNVs were not conducted. These three CNVs were eventually classified as

VUSs due to the lack of known diseases associations. Notably, the patient's karyotype was normal, and no FISH studies were undertaken.

Subsequently, this family was recruited to GBoCM and trio WES was performed to seek other candidate variants. However, no plausibly causative SNVs were identified. During my initial analysis of these exome data using the SavvyCNV variant caller, I reidentified the 16q and 17q CNVs, and found, based on coverage data, that the 17q CNV had also arisen *de novo*. Using the SNP data from WES, the two *de novo* CNV gains on 16q and 17q both were of maternal origin. Notably, the 17q CNV is positioned near two regions of interest: the *ABCA* gene cluster, which has been associated with hypertrichosis[165,166], and the *KCNJ-SOX9* locus, which has been implicated in craniofacial development[167]. In addition to these interesting regions, a question arises stemming from the biologically unlikely scenario of acquiring two independent *de novo* coding CNVs in one generation. The 17p CNV was deemed benign and no longer considered in this project due to the significant number of similar sized CNVs recorded in DGV for this locus.

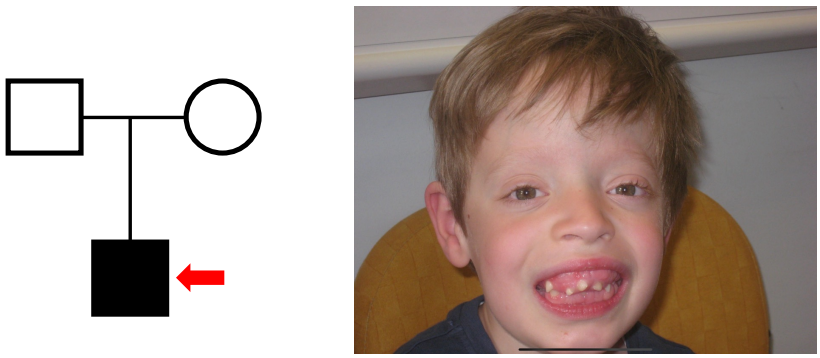


Figure 56 Case 2 proband has syndromic multi-suture synostosis with hypertrichosis and gum hypertrophy. *The parents are unaffected. Permission to include clinical photographs was given by the patient and/or their parents.*

5.5.2 WGS revealed the complex nature of the event

To further characterise these two *de novo* events, WGS was performed on the Illumina platform. The abnormal WGS coverage confirmed the presence of the two CN gains on chr16 & 17, as shown in **Figure 57**. The 17q CNV region, referred to as "**17DUP**" and illustrated as a **light blue block**, contains two genes *KCNJ16* and *KCNJ2*, flanked on both sides by apparent gene deserts, as shown in **Figure 57a**. Multiple DELs shown in DGV of these *KCNJs* hinted at potential loss-of-function tolerance of this region, although the lack of SVs in gnomAD may suggest an alternative interpretation. The other CNV region on chr16, referred to as "**16DUP**" and illustrated as a **dark blue block** (**Figure 57b**), contains two known disease genes. *ADAMTS18* has been implicated in microcornea, myopic chorioretinal atrophy, and telecanthus (OMIM 615458). On the other hand, *WWOX* is an extremely large gene spanning over 1 Mb and been associated with various malignancies[168,169] as well as a recessive condition - developmental and epileptic encephalopathy (OMIM 616211)[170]. Importantly, the 16DUP lies within one of the common fragile sites (CFSs) – FRA16D. CFSs are unstable genomic hotspots for DNA breakage during replicative stress, as observed in metaphase chromosomes.[171] DELs and translocations frequently occur at CFSs, aligning with the large number of CNVs recorded in DGV for the 16DUP region (**Figure 57a**). Within the FRA16D, a long stretch of $(AT)_{34}$ repeats, named Flex1, has been shown to greatly increase the genome instability at the site. This contributes significantly to nearby CNVs and DNA breakage.[172]

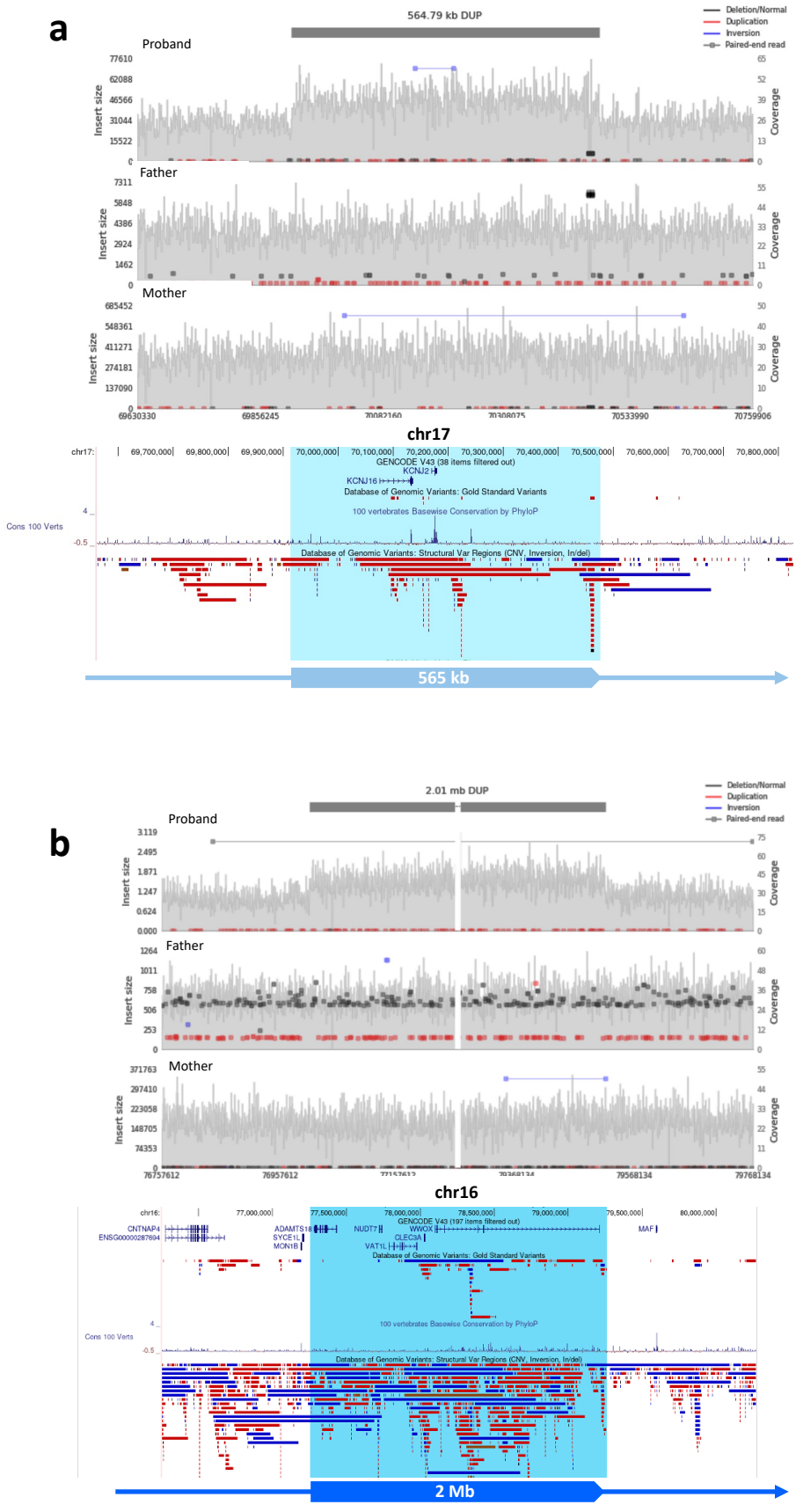


Figure 57 Two CNV gains were confirmed from WGS data. a. *The* ***chr17 CNV gain*** *containing two genes, KCNJ16 and KCNJ2. In contrast to the* ***chr16 DUP****, the affected chr17*

locus contains only DELs and barely any DUPs in DGV. ***b.*** *The large* ***chr16 CNV gain*** *containing multiple genes, while also highly polymorphic due to the large number of CNVs recorded in DGV. Chromosome 16 and the* ***chr16 DUP*** *are illustrated as a* ***dark blue arrow*** *and a* ***dark blue block****; chromosome17 and the* ***chr17 DUP*** *are illustrated as a* ***light blue arrow*** *and a* ***light blue block****. Figure in hg38.*

WGS data further identified crucial information on these SVs that was previously unknown from array or WES data. Based on the split reads and paired reads, it was determined that the two CNVs, **16DUP** and **17DUP**, are linked, as shown in **Figure 58a**. This suggests a single CPX event had occurred rather than two independent DUPs. Breakpoint PCR (**Figure 58b**) and dideoxy- sequencing (**Figure 58c**) subsequently confirmed and characterised the break junction. Notably, the 2 bp of microhomology at the 16-17 break suggests a NHEJ mechanism, and the inverted 6 bp at the 17-16 break suggests a FoSTeS/MMBIR mechanism. Given the breakpoint sequence and the nature of the CFS, it is plausible to suggest that this complex SV arose due to the innate genome instability at FRA16D. This instability would lead to a DSB at chr16, which was repaired via FoSTeS/MMBIR at the chr17 locus.

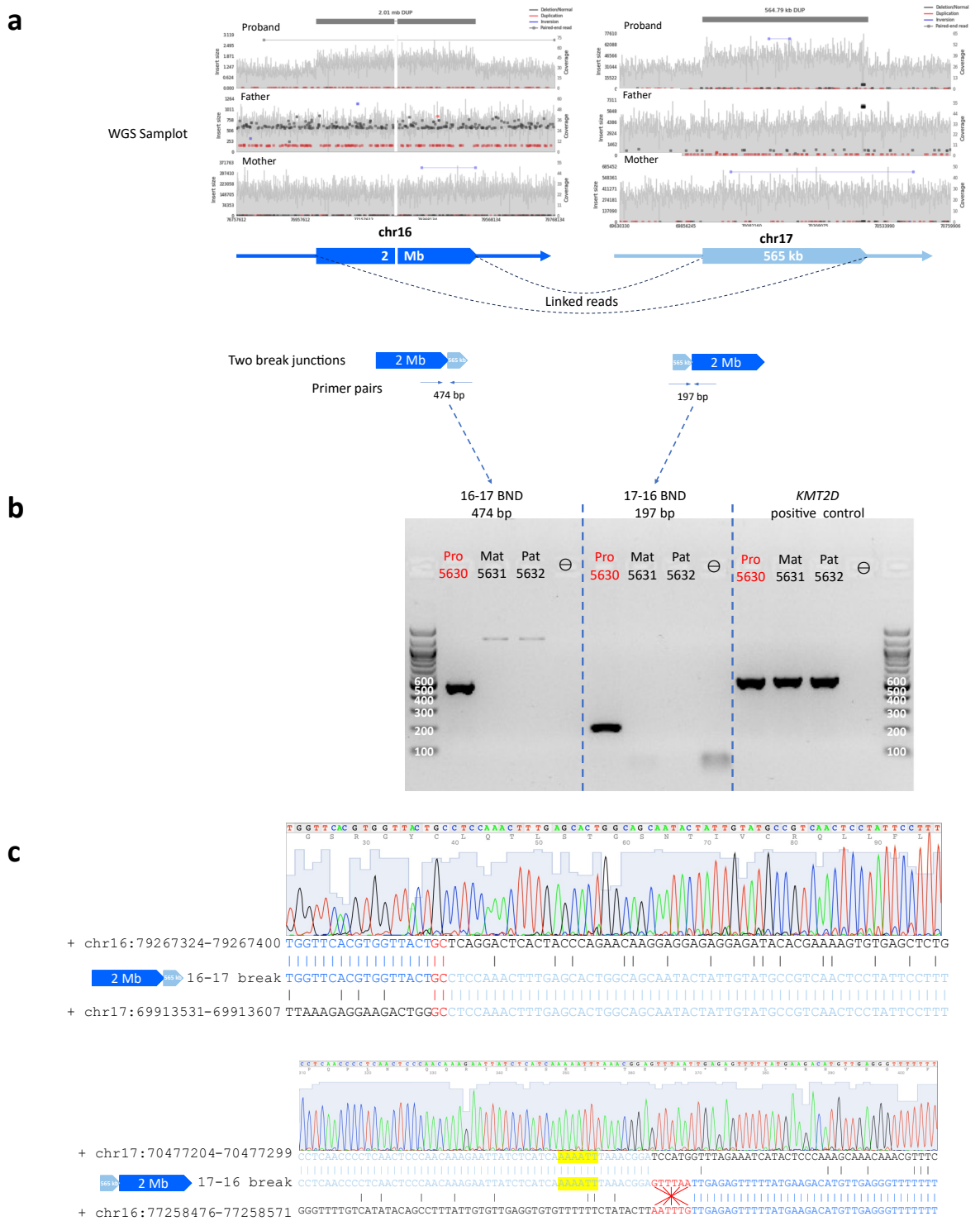


Figure 58 The complex nature of case 2 SV was revealed by WGS and verified by breakpoint PCR and dideoxy- sequencing. a*. WGS generated paired/split reads linking the two CNV segments together to form two different break junctions.* ***b.*** *Breakpoint PCR successfully amplified the two break junctions in the proband and not the parents. Note that non-specific bands can be seen in the parents.* ***c****. Subsequent dideoxy- sequencing characterised the break junction sequences. Figure in hg38.*

Based on the two confirmed break junctions and the CN information, three alternative configurations can be constructed to explain the WGS data, as shown in **Figure 59**. Alt 1 illustrates that the patient carries a normal chr16 and an abnormal chr17 where the **16DUP** segment is inserted between two copies of the **17DUP**. The Alt 2 configuration is the opposite, whereby the patient carries a normal chr17 and an abnormal chr16, with the **17DUP** segment inserted between two copies of **16DUP**. Finally, Alt 3 suggests that both chr16 and chr17 are abnormal due to a translocation between the qters of the two chromosomes. Importantly, the predicted lengths of the translocated 16qter and 17qter telomeric regions are very similar (13.1 Mb and 13.4 Mb, respectively), so that Alt3 would not necessarily be excluded because the patient's karyotype had been reported as normal.

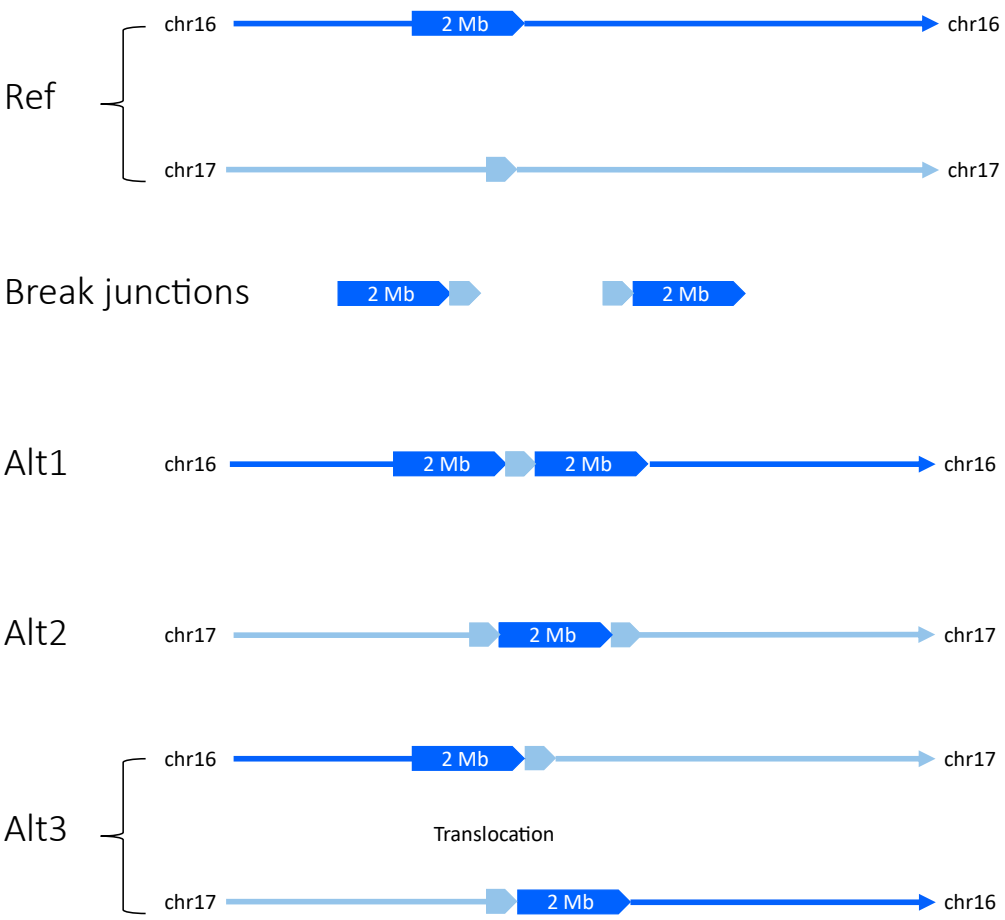


Figure 59 For case 2, three alternative hypotheses can be constructed to explain the WGS data. *Alt1 illustrates a senario where only chr16 is affected, with* ***the small piece of chr17 (light blue box)*** *inserted between* ***the duplicated chr16 (dark blue box)****. Alt2 is the opposite senario, where only chr17 is affected, with* ***the small piece of chr16*** *inserted between the* ***duplicated chr17 pieces****. Alt3 illustrates a translocation alternative, where the qters of chr16 and 17 are switched.*

5.5.3 Resolving case 2 complex SV requires both Bionano and FISH

Therefore, determining the true configuration of the SV became a pivotal prerequisite before evaluating its clinical significance. To attempt to address this, I carried out Bionano OGM, while the GOSH diagnostic lab performed FISH. To differentiate between Alt1 and the other two alternatives, Alt2/Alt3, DNA must be read as one single unbroken molecule, and span the informative **light blue 17DUP** segment of ~540 kb. This length greatly exceeds the advertised average Bionano molecule length, posing a considerable challenge and serving as an excellent test to push the limits of the

technology. To further differentiate between Alt 2 and Alt 3, the ~2 Mb **dark blue 16DUP** must be spanned by intact DNA molecules. This is likely impossible even with long-range technologies such as Bionano OGM. In contrast, FISH has the potential to resolve all three structures, depending on the experimental design.

Bionano OGM was repeated three times to obtain high quality data (Coverage: 426.68x, map rate: 92.1%, N50 ≥150 kb: 0.3529 Mb, see **Supplementary Table 3**), maximising the yield of long informative molecules at the **17DUP** locus. RVA (**section 2.9.4.1**) was carried out to visualise informative molecules by extracting all abnormal (non-reference) molecules. As shown in **Figure 60**, Bionano OGM and RVA yielded five molecules apparently spanning the whole length of the **17DUP**. Subsequent manual realignment of these five molecules (**Figure 61**) revealed that only three molecules were truly informative. The three informative molecules all support the same labelling pattern as the **16DUP-17DUP-17Ref** configuration, as depicted in **Figure 61**. By comparing to the three alternative configurations in **Figure 59**, it becomes evident that the **16DUP-17DUP-17Ref** pattern deduced from the three informative is present in Alt 2 and Alt 3, but not Alt 1. However, no molecule was found to span the **16DUP** ~2 Mb region in the Bionano data to further differentiate between Alt 2 and Alt 3.

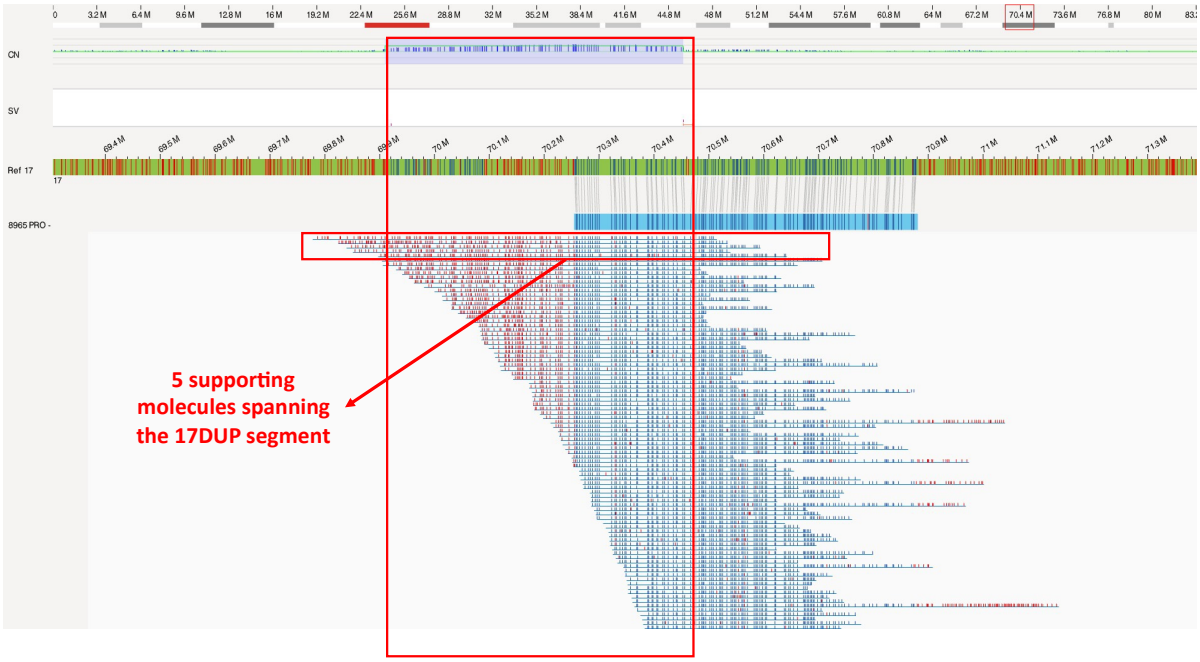


Figure 60 RVA analysis extracted 5 molecules spanning the 17DUP region in the case 2 proband. *The blue rectangle is the consensus map aligning to the Ref generated by the native Bionano analysis pipeline. The vertical red box highlights the 17DUP region, which the molecules must span to be informative. The horizontal red box highlights the five potentially informative molecules as they span the 17DUP region. Figure in hg38.*

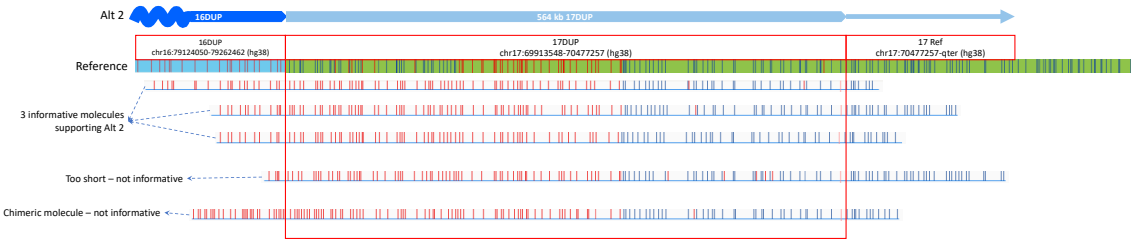


Figure 61 Manual realignment of each of the five supporting molecules detected by RVA (Figure 60). *Three molecules are truly informative molecules where the labelling patterns are consistent with the expected labelling pattern of **16DUP**-17DUP-17Ref (supporting break junction). One molecule, having only four labels in the **16DUP** region, was too short to be informative. One molecule has a high labelling density, suggesting a chimeric molecule (two molecules stuck together and read as one by the machine), and therefore not informative. Figure in hg38.*

Concurrently, FISH analysis was carried out by GOSH collaborators. A two-coloured FISH was initially designed as illustrated in **Figure 62a**. The result, **Figure 62b**, showed that the 17DUP (red signal) has been abnormally inserted into chr16 (green

signal), which is consistent with the Alt 1 structure (**Figure 59**). This finding was subsequently issued as a diagnostic report for the patient. However, this conclusion conflicts with the Bionano OGM analysis above, which excluded Alt 1. Upon closer examination, it becomes evident that Alt 3 shares the same FISH signal pattern with Alt 1, with green and red signals on the same chromosome. Consequently, the **Figure 62a** FISH design is incapable of differentiating between the Alt 1 and Alt 3 configurations. Therefore, a reciprocal FISH design was requested to verify if the **16DUP** is also present in 17q, testing whether the translocation is the true configuration.

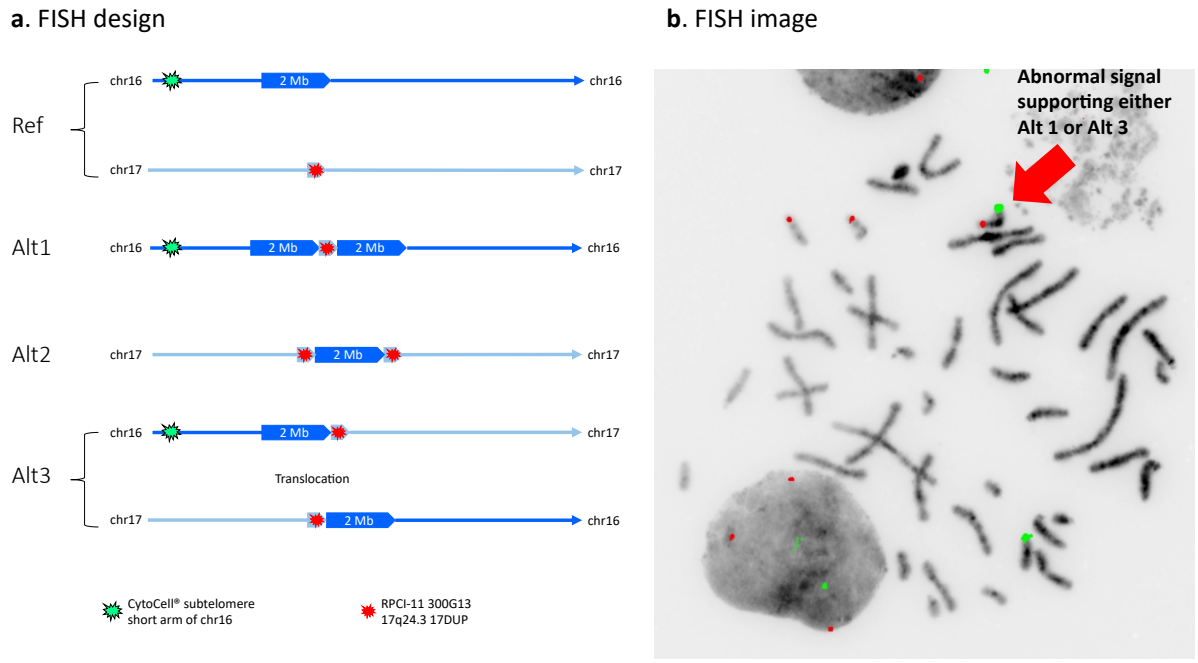


***Figure 62 FISH was carried out by GOSH diagnostic lab**. **a.** A two-coloured FISH was designed, which a green probe on the short arm of chr16 and a red probe on the 17DUP region. **b.** FISH result showed abnormal chromosomes with both green and red signals, suggesting either Alt 1 or Alt 3 is the true structure.*

Most recently, the reciprocal FISH was performed by GOSH, with the reciprocal design as illustrated in **Figure 63**. When integrating the information from both FISHs and

Bionano OGM, the only configuration that can explain the results is Alt 3, which is a reciprocal translocation occurring between chr16 and chr17, that, unusually, is associated with a net DUP of the donor chromosome at the site of each translocation breakpoint. Using the precise breakpoint sequence from the dideoxy- sequencing (**Figure 58c**), this translocation SV can be described as:

NC_000017.11:g.70477257_qterdelins[GTTTAA;NC_000016.10:g.77258535_qter]

NC_000016.10:g.79267341_qterdelins[NC_000017.11:g.69913548_qter]

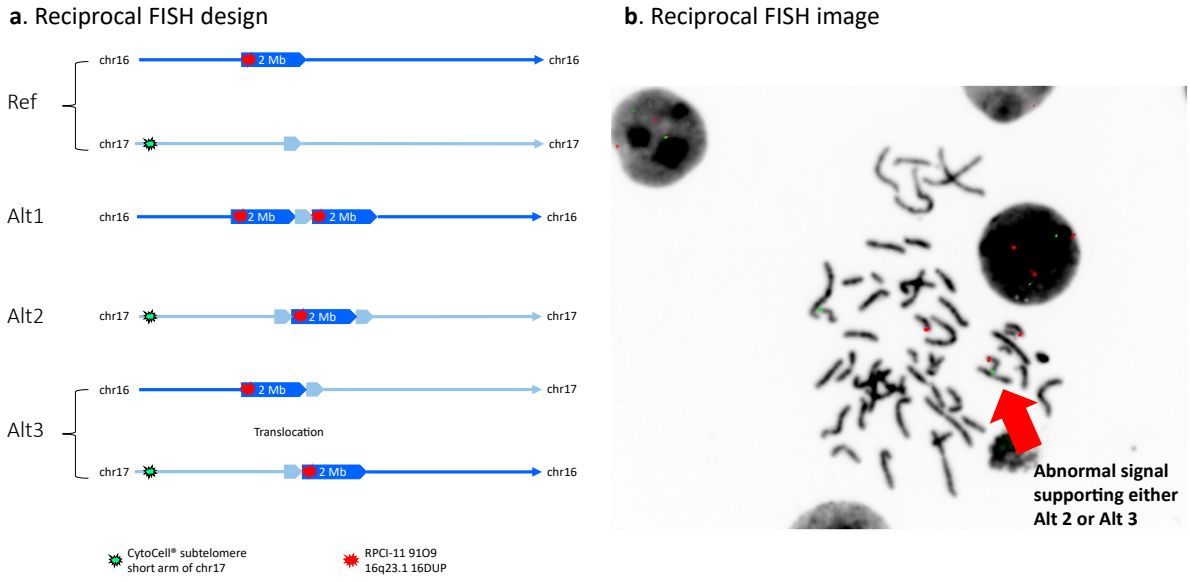


Figure 63 The reciprocal FISH was designed and carried out by GOSH as the initial FISH in Figure 62 cannot fully characterise all three Alts. a. *The reciprocal FISH was designed, which a green probe on the short arm of chr17 and a red probe on the* ***16DUP*** *region.* ***b.*** *FISH result showed abnormal chromosomes with both green and red signals, suggesting either Alt 2 or Alt 3 is the true structure. In combine with previous FISH from* ***Figure 62****, Alt 3 translocation can be confirmed as the true structure of the SV.*

5.5.4 Assessing the clinical relevance of the *KCNJ* SVs

With the confirmation of Alt 3 configuration, a more accurate assessment of the clinical relevance for the case 2 SVs can be undertaken. Within the context of syndromic CRS,

the proband presented an unusual yet highly specific phenotype - hypertrichosis with gingival hyperplasia. Phenotypically, three entries of generalised hypertrichosis (HTC) have been recorded in OMIM as HTC1 (OMIM#145701), HTC2 (OMIM#307150), and HTC3 (OMIM#135400). However, gingival hyperplasia, the distinct phenotype in case 2, is only present in HTC3. Interestingly, the critical locus for HTC3 also overlaps with the **17DUP** locus in case 2, suggesting a possible connection between HTC3 and case 2.

The HTC3 critical region, located at 17q24.3-q24.3, primarily consists of two groups of interesting genes: the *ABCA* cluster and two *KCNJ* genes. Past literature has highlighted the significance of *ABCA* genes in the HTC3 phenotype. For example, DeStefano et al (2014)[165] investigated a case of consanguineous HTC3 using WES and identified a causative homozygous variant, c.4320+1G>C (rs199753304 G>C), affecting a splice donor site of *ABCA5*. However, this G>C change has been observed in 3 alleles in the gnomAD dataset, along with additional 5 other changes at the same position. Despite the lack of homozygous SNPs at this position in gnomAD, the presence of multiple heterozygous SNPs at the same position somewhat weakens the strength of *ABCA5* causality in this rare phenotype.

From the SV/CNV perspective, multiple cases with heterozygous rearrangements at the *ABCA-KCNJ* loci have been implicated in HTC3, as shown in **Figure 64**. Amongst the example cases, Sun2009-KK-INVDUP[166], Hayashi2017-DEL[173], Afifi2015-DEL[174], DECIPHER-2641105-DUP, and case 2 all presented with both gingival hyperplasia and hypertrichosis, while the remaining cases in **Figure 64** presented only hypertrichosis without gum abnormalities. Significant effort has been made for multiple DEL cases in **Figure 64** to investigate the consequences of heterozygous loss-of-

function in *ABCA* gene(s). However, multiple recorded DELs in the control population (DGV) suggest that the *ABCA* locus can tolerate localised DELs. In addition, the presence of hypertrichosis and gum abnormalities in cases with SVs that do not intersect with *ABCA*, such as DUP, INVDUP, and the case 2 translocation (**Figure 64**), further suggests that the loss of *ABCAs* may not be essential for HTC3 phenotypes.

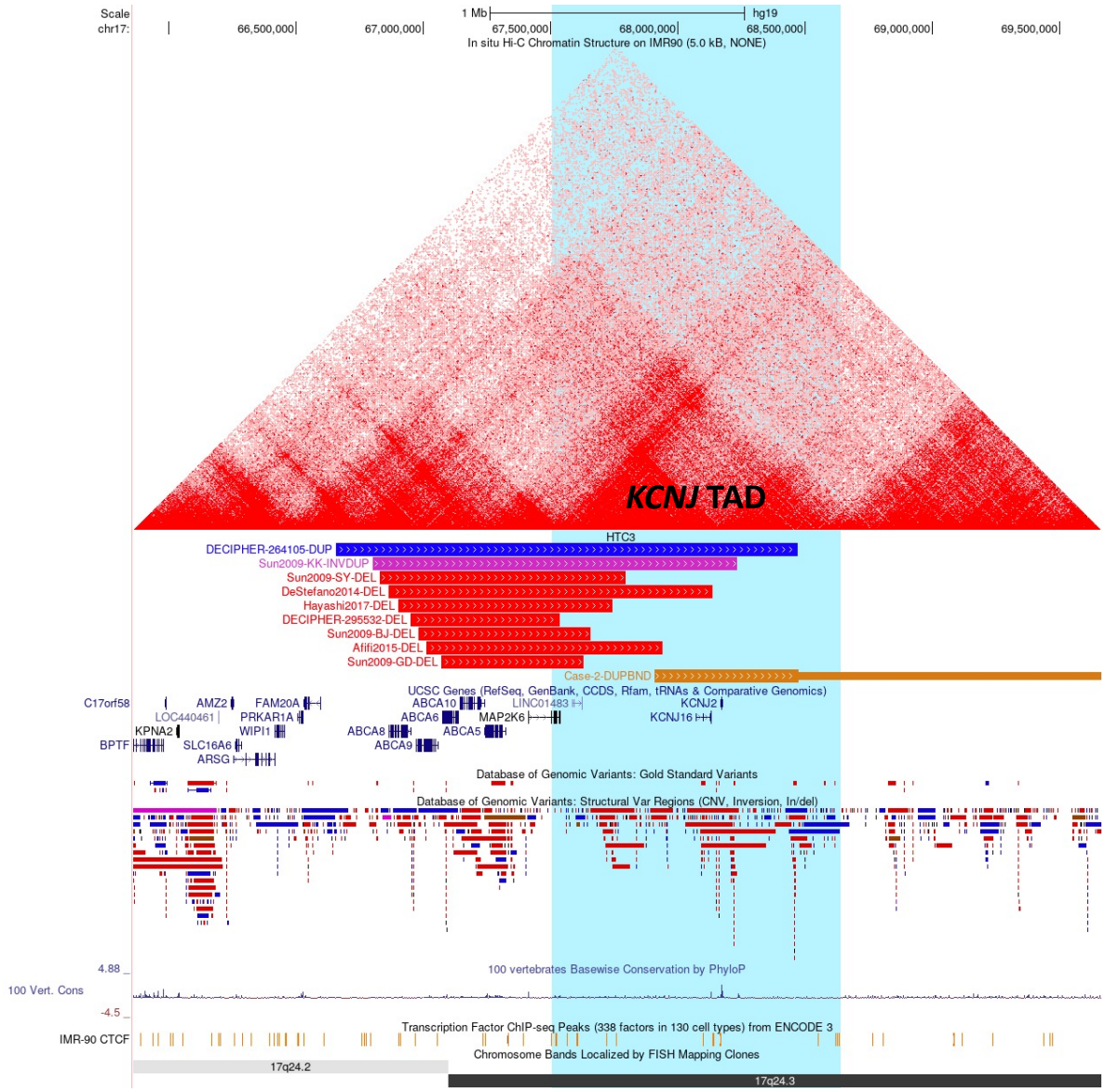


Figure 64 Several cases in the literature have implicated SV/CNVs at the KCNJ locus with the hypertrichosis and gingival hypertrophy phenotype. *SV/CNV extracted from Sun et al (2009)[166], DeStefano et al (2014)[165], Afifi et al (2015)[174], Hayashi et al (2017)[173], and DECIPHER. The case 2 DUP-translocation is shown as* ***the brown segment****. The main KCNJ TAD is highlighted in light blue background, containing two visible smaller sub-TADs. Figure in hg19.*

The addition of case 2 introduces an interesting alternative mechanism, implicating the disruption of the *KCNJ* TAD (blue highlight in **Figure 64**) as a potential explanation for the pathogenicity of the *ABCA-KCNJ* SVs in HTC3. Among the cases featured in **Figure 64**, the disruption of the *KCNJ* TAD appears consistent in all, due to either the DEL or the rearrangement of one of the two *KCNJ* TAD boundaries. This TAD disruption has not been adequately addressed in the literature in HTC3 cases. One additional example (not included in **Figure 64**) is described by Kim et al (2007)[175], part of the Developmental Genome Anatomy Project (DGAP)[176], detailing a t(3;17)(p14.3;q24.3) in a patient with both hypertrichosis and gingival hyperplasia. In this study, extensive effort was allocated to characterise and investigate the chr3 break and the nearby genes, while the chr17 break at the *KCNJ* locus received considerably less attention. For this case, evaluating the TAD disruption is challenging without knowing the precise break on chr17, but the case may be another instance where the *KCNJ* SV is misinterpreted.

One particularly interesting example in the literature described a complex *KCNJ* SV, which was highly likely misinterpreted due to the technological limit of array assays. Fantauzzo et al (2012)[177] documented a familial case of HTC3 with mild gingival hyperplasia, and the genetic cause was investigated using an array approach. At the *KCNJ* locus on chr17, the study identified four rare DUPs of varying sizes: 391 kb, 66 kb, 1.2 Mb and 35 kb. Subsequent qPCR analysis suggested the potential triplication of the 391 kb segment. FISH was carried out showing an inverted orientation of the large 1.2 Mb segment. However, this FISH design was unable to adequately capture the likely CPX nature of this SV event. Overall, the combined array, qPCR, and FISH result for this case highly resembles the pattern of complex SVs described in several

cases here (case 2, 10, 16, and 19). Regrettably, the potential complexity of the SV was not appreciated from the initial report due to the limited long-read capability at the time (2012). Based on this case, it can be hypothesised that misclassified CPX SVs are more prevalent than expected due to the frequent use of array before clinical WGS was widely available. A retrospective review may further identify previously misclassified complex SVs, when focusing on cases with multiple rare *de novo* or segregating CNVs.

To assess the *KCNJ* TAD in the case 2 SV in silico, DeepC prediction was carried out based on the Alt 3 configuration. As shown in **Figure 65**, the reference track illustrates the extent of the *KCNJ* TAD, containing two smaller sub-TADs, spanning over both boundaries of the 17DUP region. In the Alt 3 SV, the translocation affects both copies of the *KCNJ* TADs in the mutant allele. Specifically, in **Figure 65** in the Alt 3 chr17 cent - chr16 qter track, the 3' of the *KCNJs* TAD extends into the 5' of the 16DUP region. Similarly, in the chr16 cent – chr17 qter track, the 5' of the *KCNJs* TAD extends into the 3' of the 16DUP region. Interestingly, there is a well-defined TAD boundary between the *ABCA* and *KCNJ* loci, effectively preventing the *ABCA* TAD from being affected by the case 2 translocation. Lastly, DUPs downstream of the *KCNJs* do not seem to produces craniofacial phenotypes, as well-documented in several cases of Cook syndrome.[61,178] This further indicates that the critical region for case 2 and HTC3 likely resides upstream of the *KCNJs*.

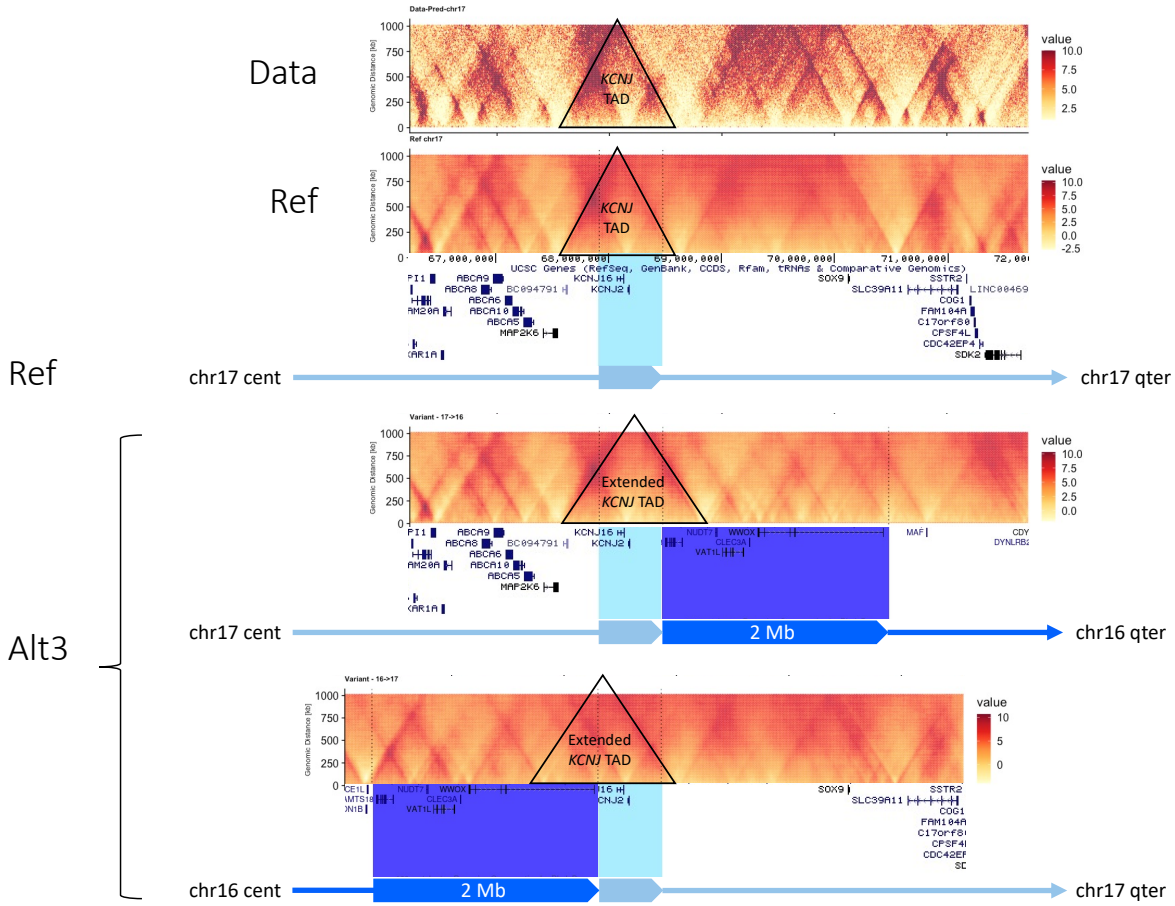


***Figure 65 DeepC prediction on the case 2 SV Alt 3 configuration**. The KCNJ TAD marked in black triangles are altered in both alleles of Alt 3 compared to the Ref. Figure in hg19.*

Furthermore, a closer examination of the *KCNJs* themselves revealed a significant interest in these genes due to their relevance with three known diseases. The first particularly relevant condition is Zimmerman-Laband syndrome (ZLS), characterised by distinct features such as gingival fibromatosis, hypertrichosis, ID, and abnormalities in soft cartilages and digits.[179] Phenotypically, there are notable similarities amongst ZLS, HTC3, and case 2, particularly the hypertrichosis and gingival abnormalities. Moreover, ZLS is known to be caused by mutations in several K^+ channel genes, including *KCNH1*, *KCNK3*, and *KCNN4*.[180] In ZLS, the gain-of-function mutations in these K^+ channels have been shown to lead to increased K^+ conductance, ultimately result in the channelopathy-related phenotypes.[180] This particular association between

the K^+ channelopathy and the phenotype of the gingival abnormality with hypertrichosis lends support to the hypothesis that *KCNJs* may be the potential underlying cause of the HTC3-case 2 phenotype.

The second relevant disease is Andersen-Tawil syndrome (ATS, OMIM# 170390), known to be caused by *KCNJ2* dominant negative mutations suppressing the K^+ function.[181] Interestingly, mandibulomaxillary hypoplasia and incomplete skull mineralisation, two phenotypes seemingly opposite to the gingival hypertrophy and CRS in case 2, have been documented in ATS.[182,183] By examining the molecular mechanisms and phenotypical consequence of ATS and ZLS together, one hypothesis emerges, suggesting that HTC3 is likely a similar channelopathy to ZLS, caused by disruption of *KCNJ* regulation.

In the third relevant disease, *KCNJ2* gain-of-function mutations cause cardiac phenotypes without syndromic or dysmorphic features.[184,185] This further refines the hypothesis, where the craniofacial features in case 2 may be attributed to the overexpress or mis-expression of *KCNJs*, differing from the gain-of-function mutations associated with a cardiac phenotype and the dominant negative mechanism underlying ATS.

In conclusion, the analysis of the case 2 SV challenges the conventional understanding of the underlying genetics of hypertrichosis and gum hypertrophy – HTC3. While the literature has commonly pointed towards the *ABCA* genes as the primary contributor, my analysis offers a different perspective, focusing on the

overlooked adjacent *KCNJ* locus. Specifically, disruption of the *KCNJ* TAD is seen in all cases discussed, providing a compelling alternative explanation for HTC3 phenotypes. This is further supported by the DeepC analysis of case 2 SV, as well as the intricate relationship amongst the three known K^+ channelopathies. This novel perspective highlights the need for revaluating the role of the two *KCNJ* genes and their TAD in the broader context of this highly specific phenotype characterised by hypertrichosis and gum hypertrophy.

5.5.5 Unbalanced reciprocal translocation

The unique structure of this SV highlights the importance of understanding the mechanisms that gave rise to this event. Like this SV, seemingly balanced translocations can in fact harbour complex structures at their break junctions. As shown in **Figure 66**, several models have been proposed previously to illustrated these complex breaks, including possible balanced/unbalanced translocation junctions with overlapping CNV gain/loss.[186] Examples of such translocations have been observed in cancer cell lines[186] and somatic tissues[187], displaying translocations in model **a-c** (**Figure 66**). The case 2 SV further serves as an example for model **d** (**Figure 66**), which involves tandem DUPs at the break junctions of both the donor and acceptor chromosomes. Note that this model **d** has only been theorised but is yet to be reported in the literature.

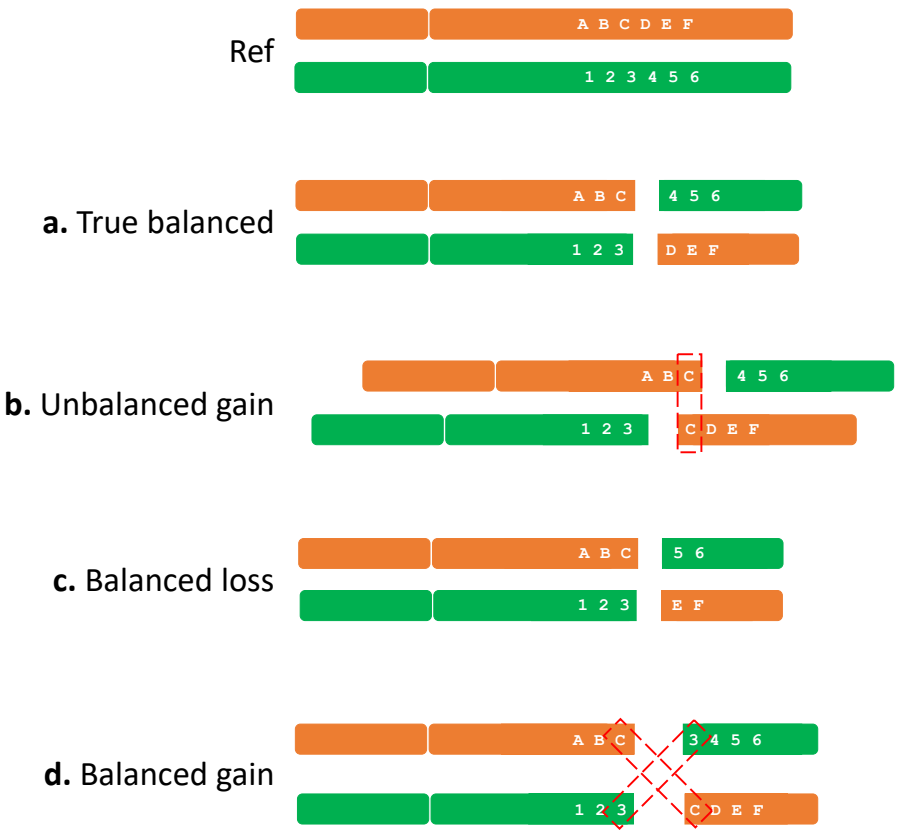


Figure 66 Seemingly balanced translocations may contain complex structures at the break junctions. *Multiple models have been proposed by Howarth et al (2011)[186] with example translocations demonstrating* ***a****,* ***b****, and* ***c*** *structure in cancer cell lines. Case 2 provides an example for structure* ***d****, which consists of tandem DUPs on both the donor and acceptor breaks of the chromosomes. Figure adapted from Howarth et al (2011)[186] and Wang et al (2022).[187]*

Examining the information scars at the break junction provided insight into the mechanism of how case 2 SV might have arisen, as summarised in **Figure 67**. Briefly, replication bubbles were firstly responsible as a source of tandem DUPs on 16q and 17q.[186] Next, the short $(AT)_n$ sequence caused a DSB between the newly created 17qDUPs. Subsequently, break-induced repair was initiated, causing the 17qDUP to integrate into the 16qDUP via FoSTeS. Notably, the 16qDUP locus is a known CFS, predisposed to breaks and fork stalling, further supporting the FoSTeS mechanism here. Lastly, the remaining two overhangs from the breaks on 16q and 17q joined via NHEJ to complete the reciprocal translocation.

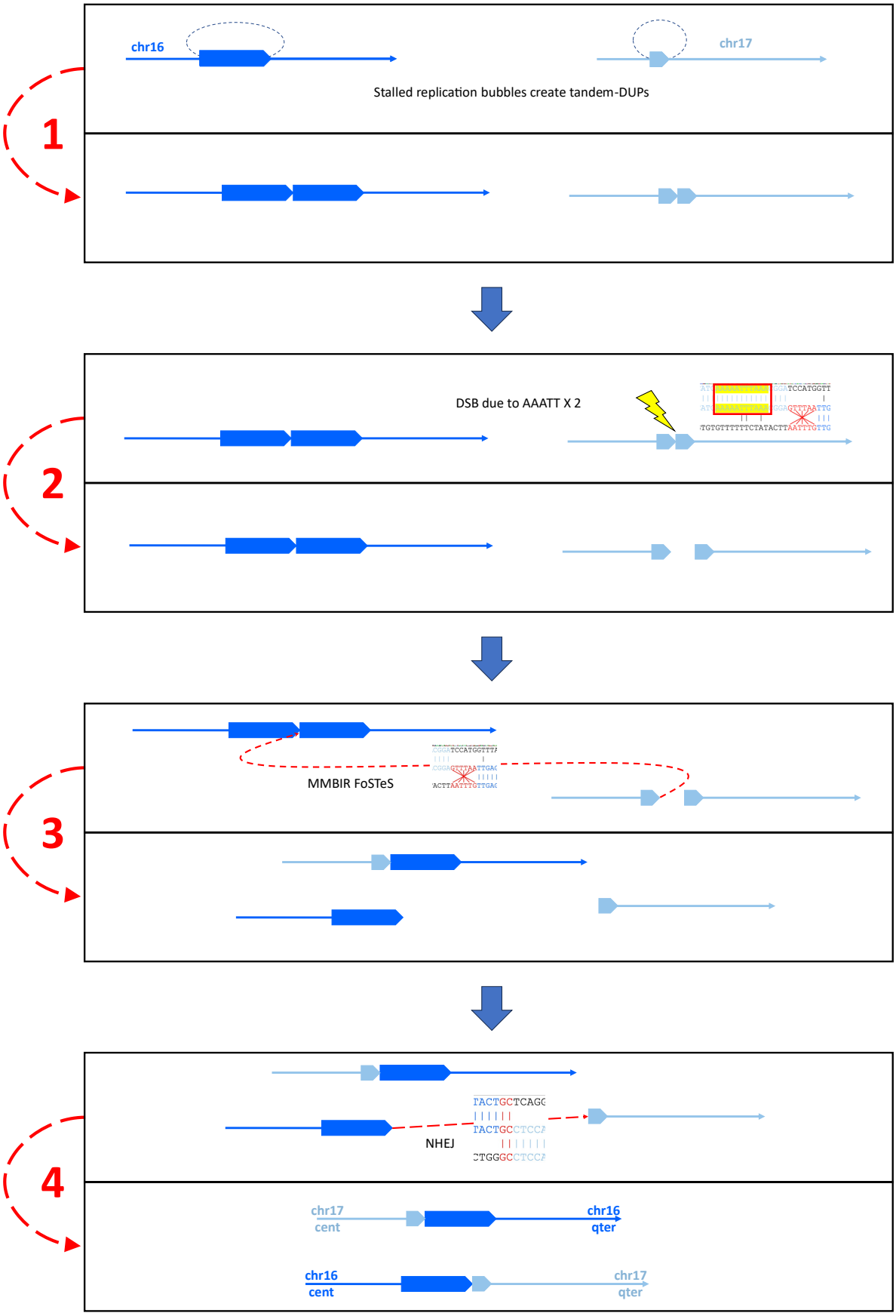


Figure 67 Proposed mechanism of how the case 2 SV might have arisen. *Firstly, tandem DUPs were created due to stalled replication bubbles, as proposed by Howarthet al (2011).*[186]

*Secondly, DSB occurred between the two **17DUPs** as predisposed by the TTTAA sequences. In step 3, break induced repair initiates at the 17q break, while switching template to between the **16DUPs** via FoSTeS, as supported by the 6 bp micro-INV at the break junction. Lastly, the remaining breaks join via NHEJ, as evidenced by the 2 bp microhomology at the break junction.*

This unusually structure highlights the considerable challenges it presents in the clinical setting. As evidenced by the initial diagnostic report, the array-based approach failed to comprehend the translocation nature of the event. Additionally, the karyotype analysis appeared normal, as the translocated qters are almost identical in length. WES and target sequencing would also face difficulties as the breaks are all located in intergenic regions. WGS emerged as the first technology capable of elucidating the inter-chromosomal nature of the event. However, a total of three possible configurations of the SV (**Figure 59**) from WGS would have resulted in, at best, an ambiguous report, and at worst, an incorrect molecular diagnosis. In contrast, Bionano OGM was more informative, while diagnostical use of Bionano OGM in general is not currently approved within the NHS. FISH is the only method theoretically capable of characterising the case 2 SV. However, an incomplete FISH design alone, as illustrated in **Figure 62** and discussed in **5.5.3**, has led to the diagnostic lab initially issuing an erroneous report, which failed to appreciate the translocation nature of the event. Such a misdiagnosis could have serious reproductive implications for case 2 family, as it would not account for the possibility of unbalanced translocations in future generations.

5.5.6 Approaching the detection limit of Bionano OGM

Case 2 also demonstrated the upper size limit for informative molecules of Bionano OGM technology. Compared to case 10 (section **5.4**), the informative molecules

required for case 2 were substantially longer (> 540 kb for partial resolution, > 2 Mb for full resolution), making the experimental process considerably more challenging. The ability for Bionano OGM to generate long informative molecules strongly relies on the quality of the input DNA, which is dependent on several factors during sample preparation including the sample type, storage/transport conditions, DNA extraction, and the DNA labelling process.

For case 2, I performed a total of three complete Bionano runs, including sample re-extraction each time. These repeats were undertaken to achieve the best quality data to maximise the chance of generating long informative molecules. Amongst the three repeats, run 1 failed to collect sufficient data (low coverage) owing to issues with sample processing and chip quality. Run 2 reached the recommended minimum data coverage but failed to identify any informative molecules. Run 2 did, nevertheless, characterise the two break junctions (**Figure 59**), consistent with those deduced from the WGS data. However, with minimum coverage and lack of ultra-long molecules, run 2 faced the same challenge as short-read data when it came to resolving large CPX SVs. Run 3, despite successfully collecting a maximum amount of high-quality data, only generated three informative molecules over 540 kb. This suggests that reliably generating reads over ~500 kb is likely approaching the upper limit of the current Bionano OGM technology.

5.6 Case 16: CPX SV involving *PLCB4*

One of the most complex SV this project has encountered is the case 16 SV on chr 20. Case 16 presented a localised chromothripsis-like event that is yet to be fully

characterised and understood despite intense scrutiny from multiple genomic technologies. On the other hand, this case did provide an interesting type of SV, some components of which are invisible to all cytogenetic technologies with the exception of FISH.

Case 16 is a family trio (**Figure 68a**) recruited to both the 100kGP and GBoCM. The proband presented with exorbitism, arachnodactyly, and syndromic multisuture recurrent CRS. The parents were unaffected suggesting a possible *de novo* underlying cause. The unaffected maternal grandmother was later recruited to the local study to further understand the transmission of the candidate variants. Previously, the family had undergone both arrayCGH (BlueGnome CytoChip ISCA 8x60K v2.0) and WES analysis, yielding no conclusive diagnostic results. Review of the previous WES and array data as part of my initial analysis suggested that the chr20 abnormality was indeed detected by both technologies, but the SV had not been further investigated due to being inherited from the unaffected mother.

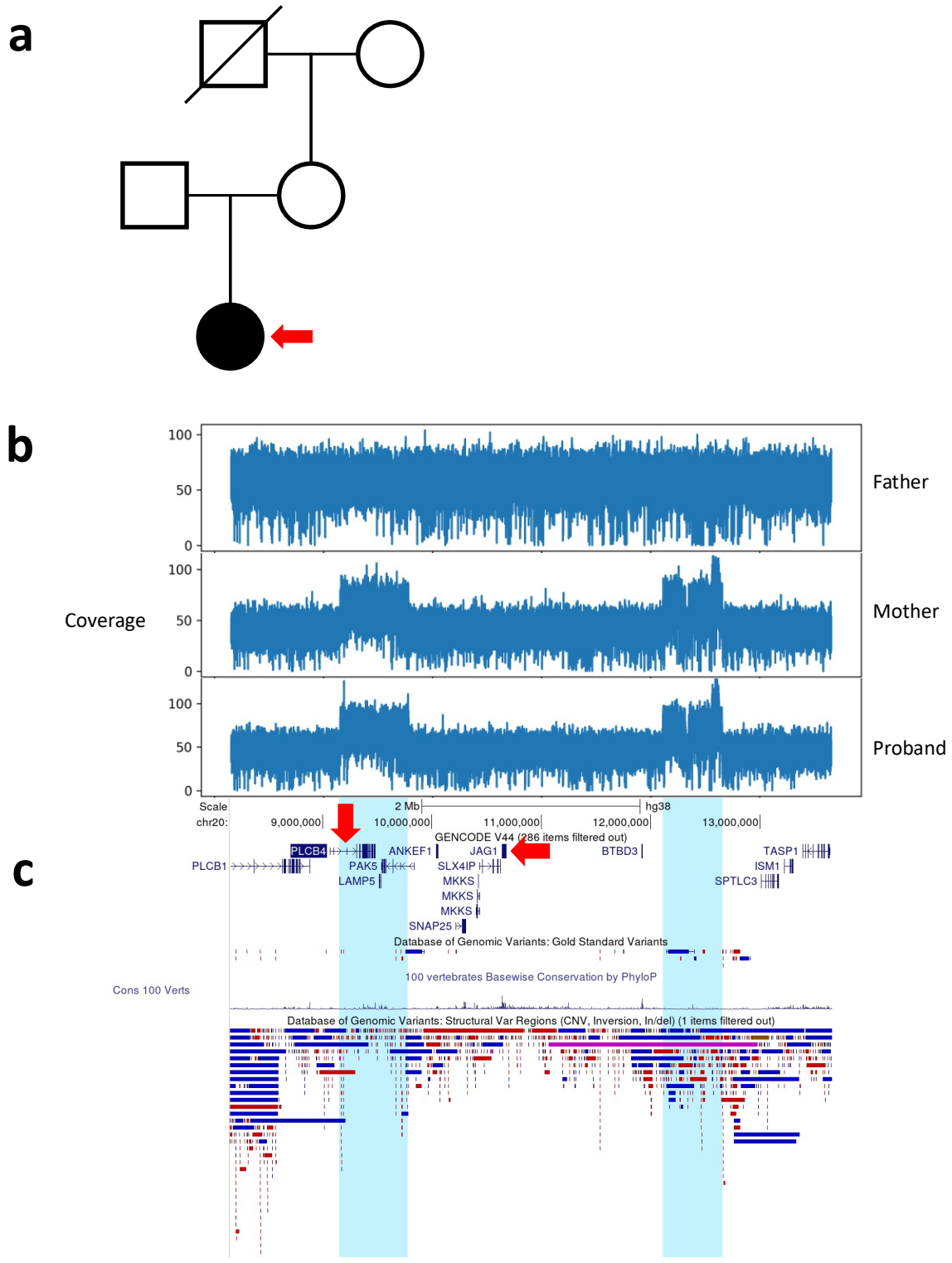


***Figure 68 The case 16 trio was recruited to the 100kGP due to syndromic multisuture CRS**. **a.** Four family members were involved in the study – the proband and the parents were included as part of the 100kGP WGS analysis, while the maternal grandmother was later recruited to the local study to investigate the transmission origin of the candidate SV. Only the proband was affected. **b.** An overview of the large SV spanning over 3 Mb on chr20 identified by WGS analysis. **c**. Genetic context of the SV, with the two candidate genes ,PLCB4 and JAG1, highlighted by red arrows. WGS coverage plot exported from GE Airlock. Figure in hg38*

5.6.1 SV characterisation using WGS, PCR, and dideoxy- sequencing

From the 100kGP WGS analysis, a large DUP call (**Figure 68b)** was initially of a high interest due to spanning over two known diseases genes (highlighted in **Figure 68c**) implicated in craniofacial disorders: *PLCB4* in Auriculocondylar syndrome (OMIM#614669) and *JAG1* in Alagille syndrome 1 (OMIM#118450). This event, however, was inherited from the unaffected mother, making it either an incidental finding or an SV with a complex pathogenic mechanism.

WGS provided additional read and coverage information that had eluded detection by the WES and array analysis, uncovering the CPX nature of this event as shown in **Figure 69**. The detailed abnormal coverages suggested that this event occurred across two distinct loci: a 626 kb CN gain at the distal *PLCB4* locus and a second 543 kb CN gain at the proximal *BTBD3* locus. Integrating both the coverage and the linked reads, four break junctions were identified, splitting the two CN gains into five segments. These segments are illustrated using differently coloured blocks, comprising **brown (B)**, **yellow (Y)**, **green (G)**, **red (R)**, and **purple (P)**, as detailed in **Figure 69**. The red segment appears to have a higher CN than the others, indicating a likely triplication. Using a maternal DNA sample, breakpoint PCR followed by dideoxy- sequencing verified and characterised all four break junctions, as shown in **Figure 70**. With the breaks accurately defined, the SV was shown to directly affect only three genes: *PLCB4, LAMP5,* and *PAK5*.

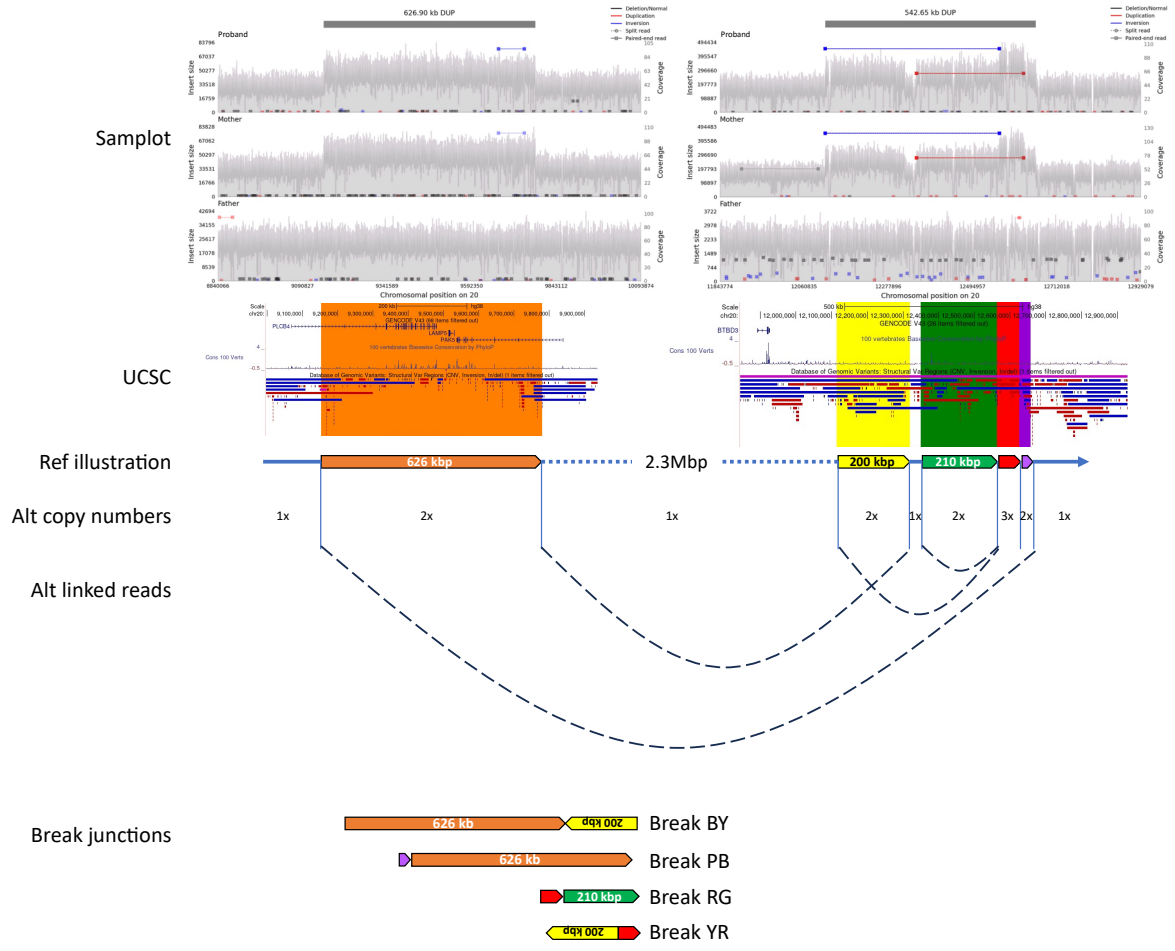


Figure 69 WGS analysis revealed the complex nature of the case 16 SV*. From the coverage and linked reads information, five segments of abnormal copy number (CN) are illustrated using the following colour scheme: Brown (2x chr20:9153518-9780418), Yellow (2x chr20:12115100-12319713), Green (2x chr20:12350710-12564054), Red (3x 3x chr20:12564055-12626497), and Purple (2x chr20:12626498-12657753). Three genes were shown to be directly affected by the SV: PLCB4, LAMP5, and PAK5. Samplot figures exported from GE Airlock. Figure and all coordinates are in hg38.*

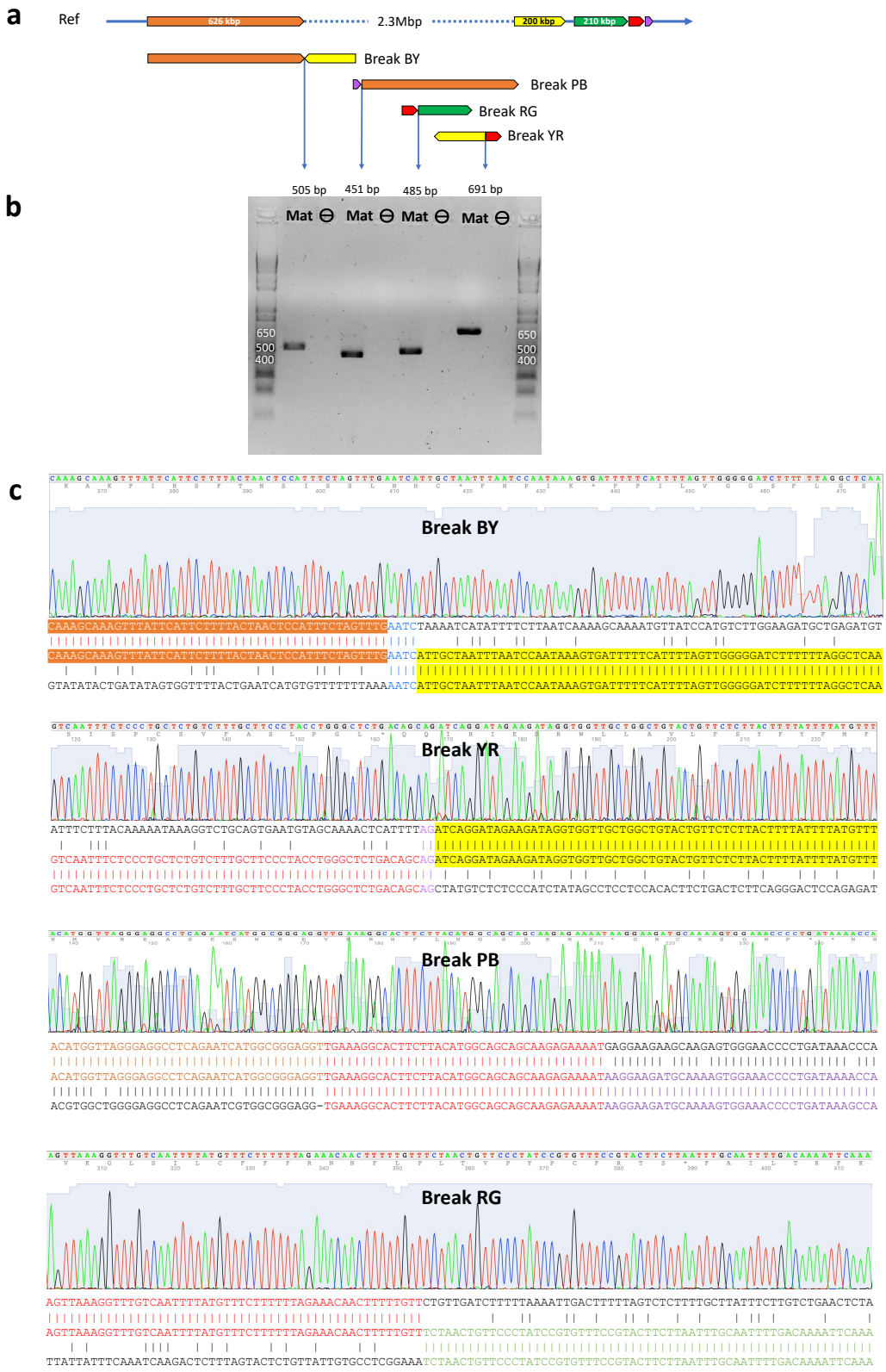


Figure 70 Breakpoint PCR and dideoxy- sequencing verified the four breaks detected by WGS in case 16. a*. The reference (Ref) illustration for the loci of interest and the four possible break junctions deduced from WGS (**Figure 69**). **b**. Breakpoint PCR result showing that all four breaks can be amplified in the maternal DNA sample. **c**. Dideoxy- sequencing characterised in detail all four amplified breaks.*

Using the verified breaks and the CN information, the SV can be reconstructed with a total of 12 alternative configurations, as illustrated in **Figure 71**. These alternatives posed an interesting clinical challenge in the context of CPX SV detection, stemming from the fact that these 12 different alternatives are indistinguishable by most genomic technologies. This challenge arises from two fundamental features of this event. Firstly, these 12 alternatives differ from each other by only one or two INVs. For example, Alt 2 can be derived from Alt 1 via a single INV between the two yellow segments. Likewise, the remaining alternatives can all be derived from at most two INVs from Alt 1. This leads to the second feature, whereby current technologies, such as WGS and array, cannot distinguish between these alternatives, potentially leading to an oversight of the INV events. Further supporting evidence comes from the fact that these inverted-duplicated segments likely predispose each generation to more INVs owing to the significant homology between duplicated copies. Similar cryptic INVs have also been described in less complex DUP-INV-DUP SVs, and the surprising abundance of the cryptic SV signatures in the array data further demonstrated the need for sequencing based and long-range technologies.[188]

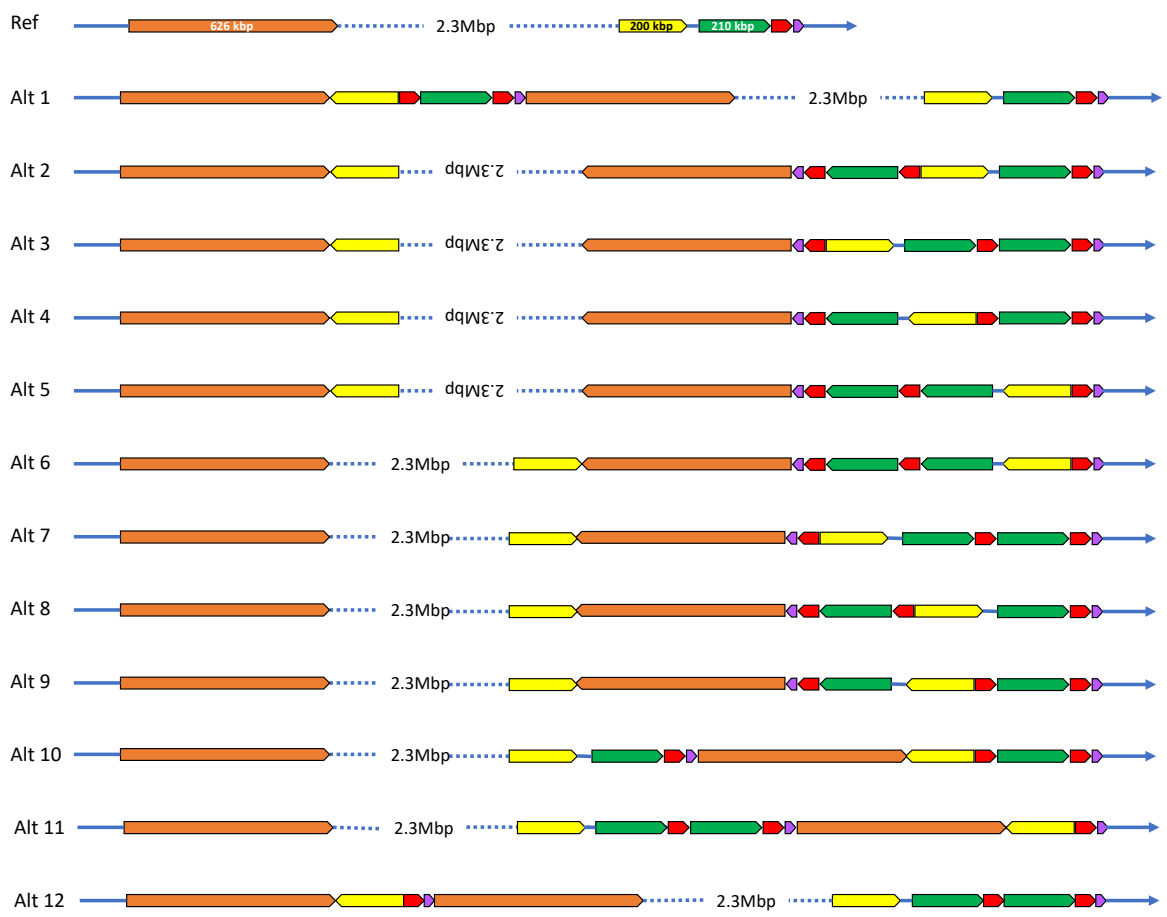


Figure 71 A total of 12 possible alternative configurations can explain the WGS data of the case 16 SV*. Illustrations follow the same colour schematics from* ***Figure 69****.*

More importantly from a clinical perspective, these different alternatives could bear completely different biological consequences, with some alternatives likely benign, and others being pathogenic due to the repositioning of critical genes and regulatory elements. These crypitic INVs offer a key hypothesis that can potentially explain the unaffected carriers in case 16. For example, although the mother in case 16 is an unaffected carrier, it is possible that she carries a different configuration of the SV compared to the proband. Furthermore, the copy neutral 2.3 Mb segment is inverted in Alt 2-5, which potentially places genes within this region (eg *JAG1*) under a new regulatory control. This further emphasises the importance of fully characterising the true SV configuration in each of the carriers in case 16.

5.6.2 SV characterisation using Bionano OGM

Bionano OGM was undertaken to characterise the SV and to identify potential structural differences amongst the case 16 carriers. While Bionano OGM was successful in narrowing down the number of possible alternatives in case 16, full characterisation of the SV was unachievable. In the following section, I will discuss detailed analysis of the Bionano OGM results for the proband, the parents, and the maternal grandmother.

The maternal SV was effectively characterised using Bionano OGM, however there were only three molecules in support. As shown in **Figure 72a**, Bionano successfully generated abnormal maps at both affected loci, separated by the 2.3 Mb normal CN region. At the distal *BTBD3* locus, two informative molecules were detected to suggest a **G-Y-R-G-R-P-B** structure (**Figure 72b)**. At the proximal *PLCB4* locus, only one informative molecule was detected and suggested a **Y-B-P** structure (**Figure 72c)**. Among the 12 possible alternatives, only Alt 8 contains the same pattern as observed in the maternal Bionano result, thus suggesting that the mother is likely carrying the Alt 8 SV configuration. Based on the Bionano OGM and the dideoxy- sequencing, the HGVS nomenclature for Alt 8 was annotated to be:

NC_000020.11:g.12319713_12319714ins[9153518_9780418inv;12350710_12657753inv;12564055_12626497inv;12115100_12319713].

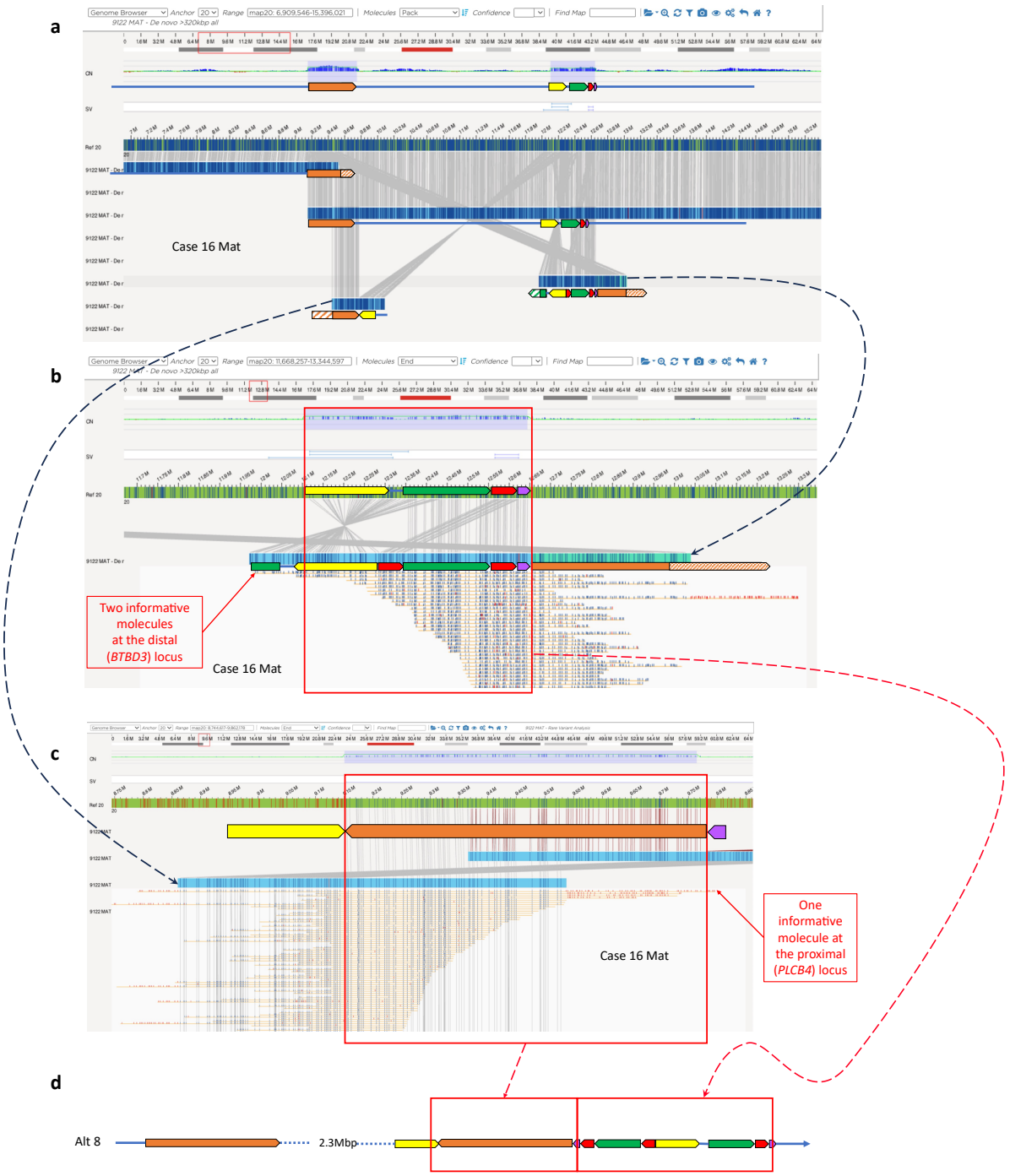


Figure 72 Bionano OGM data from the unaffected mother in case 16. *Red boxes indicate the regions that must be spanned by intact molecules for the data to be informative.* ***a****. An overview of the two loci of interest showing that abnormal maps were successfully constructed by the Bionano analysis at both loci.* ***b****. Two molecules successfully span the shorter segments at the distal locus, generating a pattern of* ***G-Y-R-G-R-P-B****.* ***c****. Only one intact molecule was able to span the 620 kb region to suggest the other informative pattern of* ***Y-B-P****.* ***d****. Only the Alt 8 configuration contains both patterns deduced from the Bionano result. Hashed segments represent the part not shown by the specific Bionano map (likely assemble into the next/previous map instead). Black dashed lines link up the zoomed view of a specific map in the overview. Red dashed lines connect the patterns deduced from Bionano to the corresponding segments in Alt 8. Figure in hg38.*

In contrast, for the proband, OGM was only able to narrow down the number of Alt possibilities, without achieving a fully confident characterisation of the SV. As shown in **Figure 73a**, similar to the maternal data, OGM successfully generated two abnormal maps at the two loci of interest. At the proximal *PLCB4* locus, one informative molecule suggested the maternal pattern: **G-Y-R-G-R-P-B**, as shown in **Figure 73b**. At the distal *BTBD3* locus, one molecule successfully spanned the informative brown segment (**Figure 73c**), while the high labelling density at the distal end of the molecule indicated that this was a chimeric molecule. Chimeric molecules are usually non-informative, as there are multiple molecules stuck together and falsely read as a single molecule. However, despite this, a few labels outside of the chimeric regions suggested the same maternal **Y-B-P** structure, shown in **Figure 73c**. Overall, if the chimeric molecule was included, the proband SV would be Alt 8, identical to the maternal SV; however, if only the confident informative molecule was considered, the proband SV could be either Alt 2 or Alt 8.

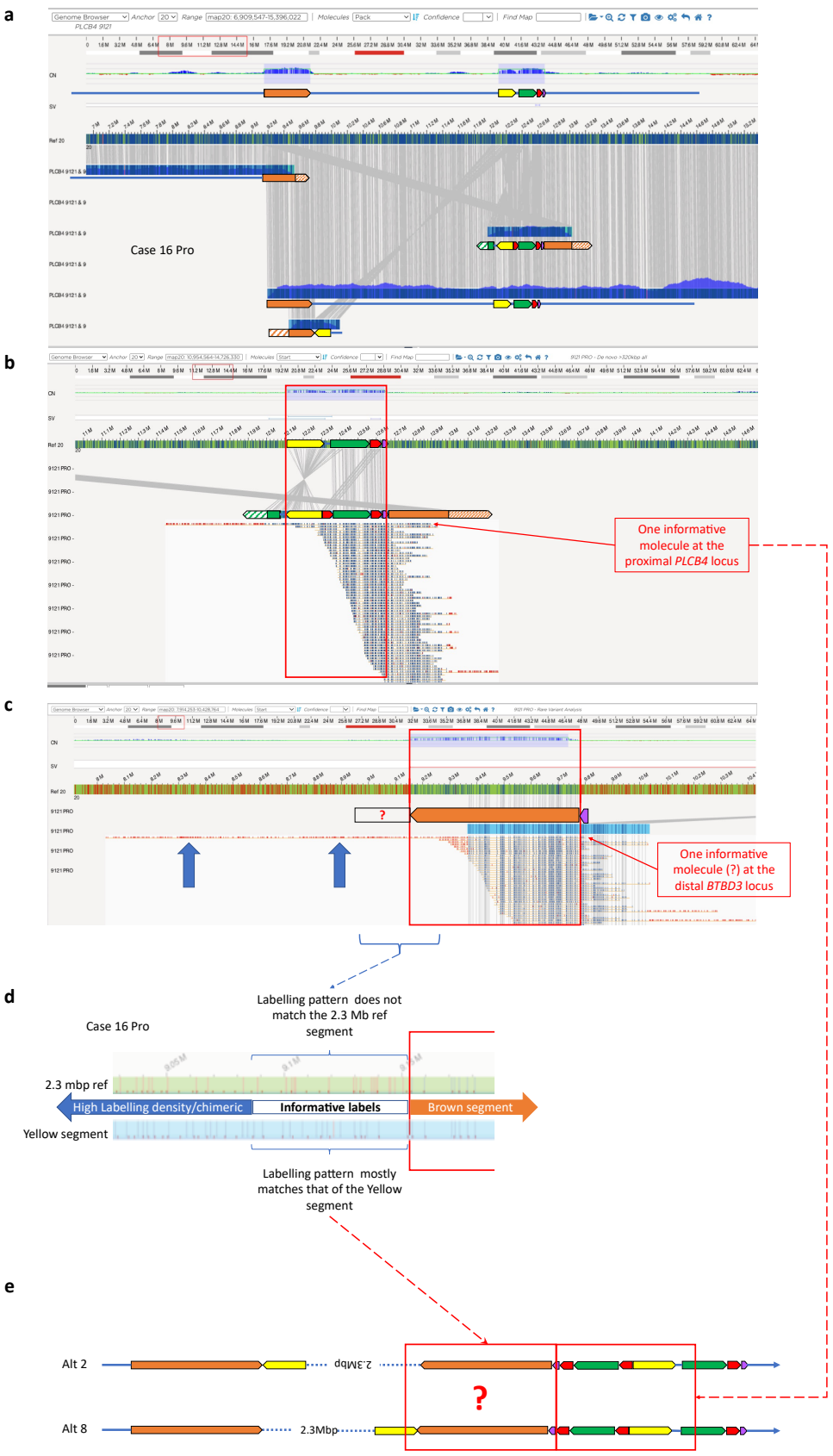


Figure 73 Bionano OGM data for the case 16 proband. ***a***. *Overview of the two loci of interest showing that abnormal maps were successfully constructed by the Bionano analysis at both loci.* ***b***. *Only one molecule fully spanned the combined segment at the distal locus, generating*

*a pattern of **G-Y-R-G-R-P-B**. **c**. One extremely long molecule (size) was detected to span the proximal 620 kb segment. However, regions of high labelling density (blue arrows) indicated that this was a chimeric molecule whereby multiple molecules have stuck together and been read as one. **d**. A further attempt was made to salvage informative labels from this chimeric molecule. The molecule was overlayed on top of two possible maps, the reference map in green, and the Alt 8 map in blue. Longer bars are labels from the maps, while the shorter dots are labels from the molecule. Labels on the right-hand side of the molecule align perfectly to the brown segment. The left-hand side of the molecule is chimeric and therefore non-informative. The section of the molecule in between, however, aligns much better to the Alt 8 map in blue, rather than the reference map in green. This suggests that the proband's configuration is Alt 8. However, this conclusion must be approached with caution, since the evidence was extracted from a molecule that is typically discarded. Red boxes indicate the regions that must be spanned by intact molecules for the molecules to be informative. Red dashed lines connect the patterns deduced from Bionano to the corresponding segments in either Alt 2 or Alt 8. Figure in hg38.*

Lastly, the clinically normal maternal grandmother was recruited to investigate the origin of this CPX event (the maternal grandfather is deceased). Breakpoint PCR was performed, which demonstrated that she also carried the chr20 CPX SV, as shown in **Figure 74a**. Subsequently, Bionano OGM was undertaken, yielding two abnormal maps with discernible patterns at the two loci, shown in **Figure 74b**, narrowing down the possible configuration to Alt 2 and Alt 8. In contrast, however, no further informative molecules were obtained at the distal *PLCB4* locus to further resolve between Alt 2 and 8.

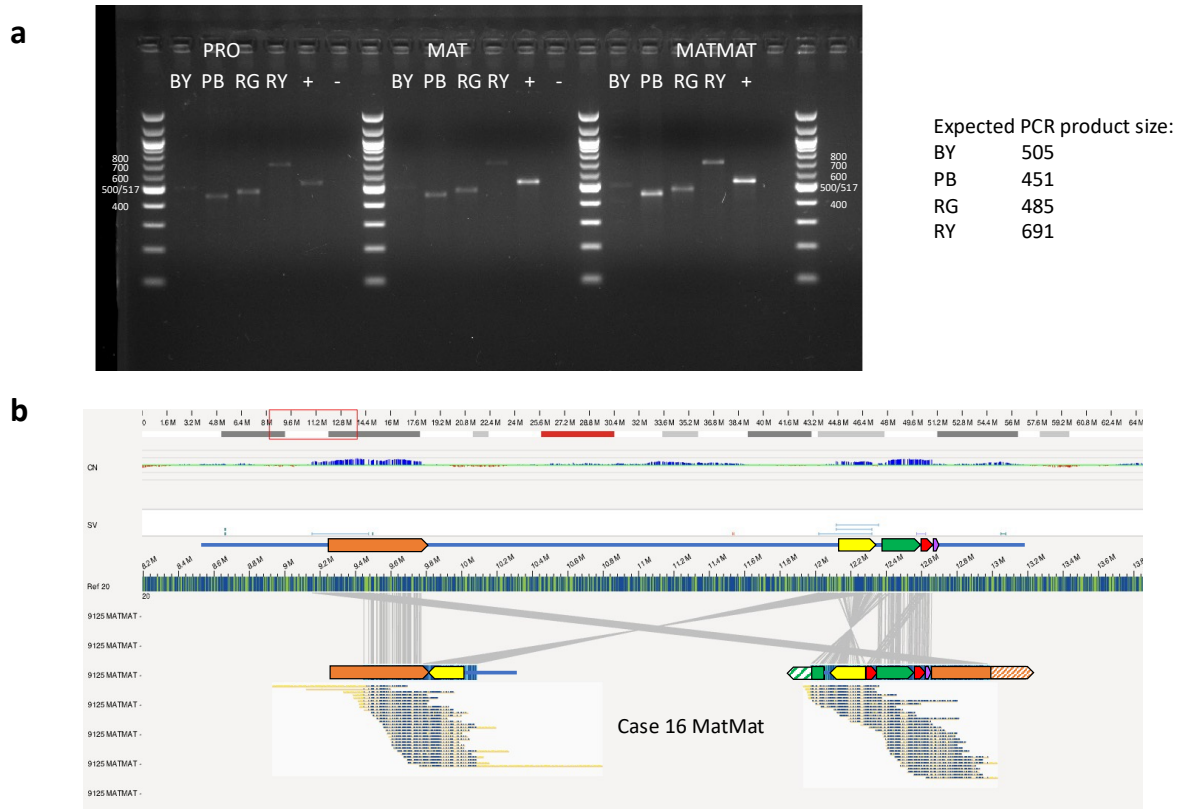


Figure 74 The case 16 maternal grandmother was recruited later to investigate the origin of the CPX SV using PCR and Bionano OGM. a*. Break point PCR demonstrates that the maternal grandmother carries the same SV event. See* ***Figure 70b*** *for the previous breakpoint PCR.* ***b****. An overview of the Bionano OGM result for the maternal grandmother, showing two abnormal maps. In contrast to the results from the proband and the mother, no intact molecules were long enough to span through either locus entirely. Some shorter molecules were informative to narrow down the maternal SV configuration to either Alt 2 or Alt 8. Figure in hg38.*

In summary, Bionano OGM successfully generated informative molecules that spanned the shorter segments of the SV for all three carriers in the family, effectively narrowing down the potential alternative configurations in each carrier. The proband SV could be either Alt 2 or Alt 8, with part of a chimeric molecule suggesting Alt 8. The maternal grandmother's SV could also be Alt 2 or Alt 8, but no extra-long molecules were detected to differentiate further. As for the mother, the SV is likely Alt 8, supported by one ultra-long molecule. Similar to case 2 and case 10, FISH, amongst

all genomic technologies, stands the best chance to be able fully characterise this CPX SV.

5.6.3 Case 16 variant interpretation

Assessing the clinical significance of this complex event hinges crucially upon the full resolution of the SV, which is a task regrettably beyond the capabilities of the long-range technology used. Nevertheless, with the preliminary result from the Bionano OGM, initial interpretation was achieved by considering the directly affected genes, as well as the potential regulatory effect from either Alt 2 or Alt 8 configurations present in the family.

Amongst the genes directly affected by the SV, the most compelling candidate gene is *PLCB4* due to its known association with craniofacial phenotypes. Pathogenic *PLCB4* variants are known to cause a craniofacial abnormality, Auriculocondylar syndrome 2 (ARCND2; OMIM 614669), through either a dominant-negative or a biallelic recessive loss-of-function mechanism.[189,190] This suggests the possibility of a shared genetic basis between case 16 and ARCND2 cases due to the involvement of *PLCB4* in case 16. However, it is crucial to highlight that ARCND2 is associated with a distinct syndromic phenotype characterised by ear and mandibular anomalies - features that have not been observed in Case 16. Furthermore, ARCND2 is associated with loss-of-function mutations in *PLCB4*, contrasting with case 16 where the SV retains at least one functional copy of the *PLCB4* in the affected allele, and thus possibly diluting the loss-of-function consequences. The other two affected genes, *LAMP5* and *PAK5*, have not been associated with human diseases. Overall, this

suggests that, if the SV were indeed pathogenic, its underlying mechanism is likely distinct from ARCND2.

PLCB4 misregulation is an alternative hypothesis for the case 16 SV pathogenicity. The EDNRA-PLCB4 signalling pathway plays an essential role in neural crest cell patterning and differentiation during development of the facial bones[191,192], and the disruption of this signalling axis has been studied in animal models with craniofacial defects.[193,194] This underlines the significance of correct *PLCB4* regulation and expression in craniofacial development.

In the case 16 SV, both copies of the *PLCB4* on the rearranged chr20 may be located in new regulatory contexts, bringing distant regulatory elements closer to the gene. Note that MANE (Matched Annotation from National Center for Biotechnology Information and European Bioinformatics Institute) select model of *PLCB4* (**Figure 69**) extends beyond the brown segment. However, the 5' exons of the MANE *PLCB4* have extremely low expression levels as recorded in Genotype-Tissue Expression (GTEx). Therefore, in the subsequent schematics for the functional analysis, the shorter and mostly broadly expressed transcript of *PLCB4* was used, where the coding region was fully contained in the brown segment. To predict the effect of the case 16 SV on the *PLCB4* regulation, DeepC prediction was carried out for the two possible SV configurations, Alt 2 and Alt 8 (**Figure 75**). As shown in the hi-C data track, *PLCB4* spans a boundary between two TADs (referred to as the 5'TAD and the 3'TAD in **Figure 75**), implying that proper *PLCB4* regulation may require interaction from both TADs. In Alt 8, one copy of the *PLCB4* remains in its native genomic environment,

maintaining the correct TAD interactions. However, the second copy the *PLCB4* in Alt 8 is inserted into a completely new genomic environment, causing both TADs to be affected as demonstrated by the DeepC prediction in **Figure 75**. In contrast in Alt 2, both copies of *PLCB4* partially lose their native genomic environment, with one copy's 5'TAD affected, while in the other copy the 3'TAD is affected.

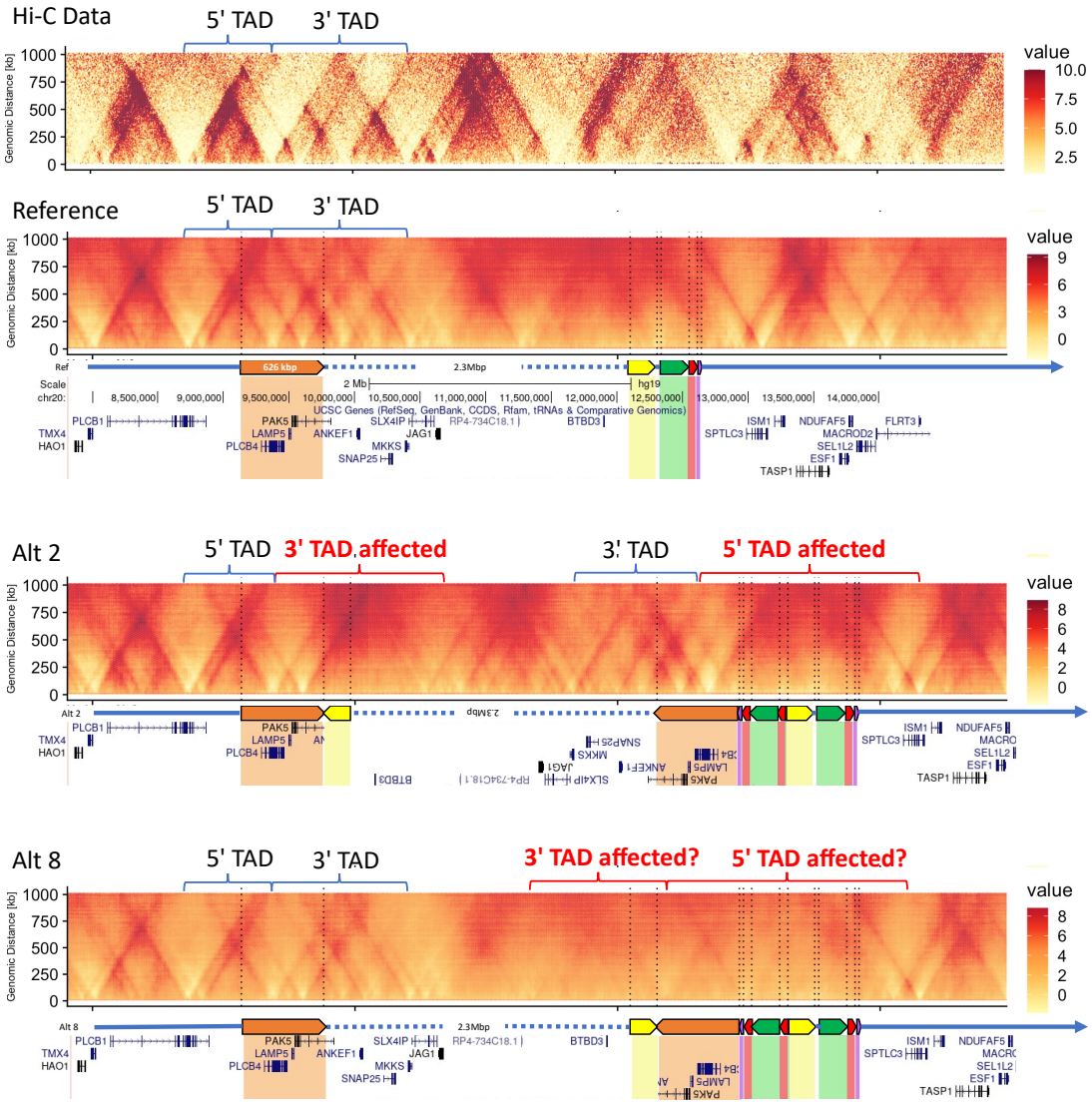


Figure 75 DeepC predictions for the case 16 SV. *The prediction includes the reference, Alt 2, and Alt 8. In the Hi-C data track (IMR-90 cell line), a TAD boundary can be seen right on top of the PLCB4 gene, suggest PLCB4 is regulated by both TADs – the 5' TAD and the 3' TAD, marked in blue brackets. The reference prediction, despite showing more pronounced interactions, is able to predict details of the TADs at the PLCB4 locus. The Alt 2 track shows that both copies of PLCB4 have altered TADs, while Alt 8 shows that at least one copy of the PLCB4 TAD remains functionally native. Likely altered TADs are marked in red brackets. Schematics and gene track colour scheme follow the previous figures. Figure in hg19.*

Integrating these predictions from DeepC, one hypothesis arises to potentially explain the unaffected carriers in case 16, as depicted in **Figure 76**. This hypothesis suggests that the unaffected family members carry Alt 8, whereby one copy of *PLCB4* remains functionally native in the mutant allele. In contrast, a cryptic INV is hypothesised to have occurred in the proband, either constitutionally or mosaically, causing the SV to be rearranged into Alt 2 in the proband. This cryptic INV consequently disrupts the TAD interactions for both copies of *PLCB4* on the mutant allele in the proband. To test this hypothesis, FISH could be used to further decipher the cryptic INV between the generations. Alternatively, transcriptome analysis of relevant cell types could determine if *PLCB4* is truly affected in the proband compared to the unaffected carriers.

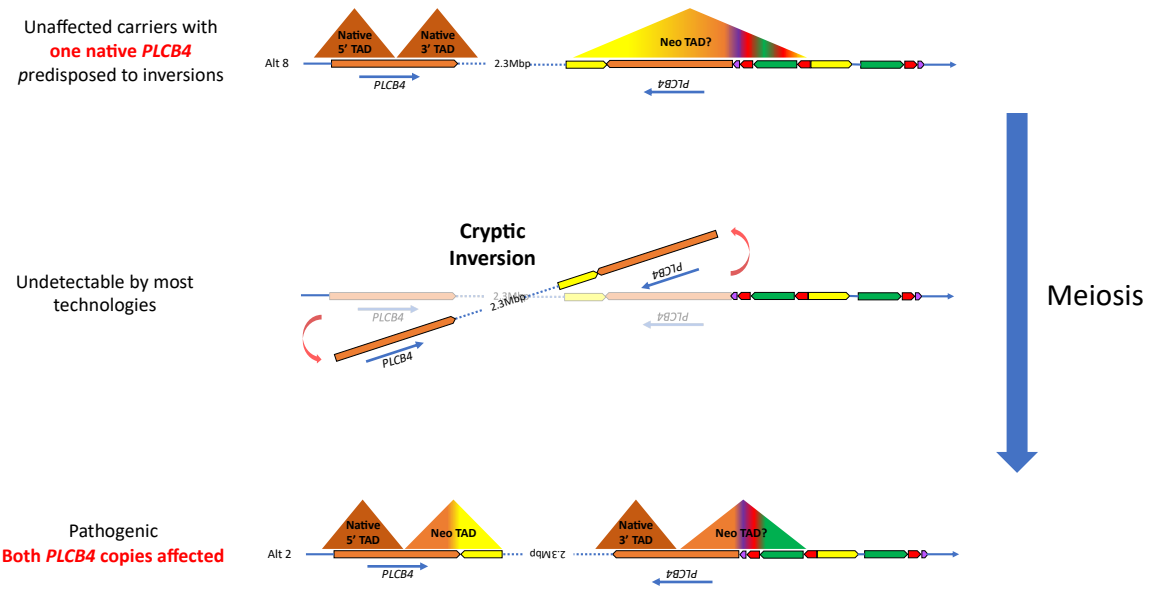


Figure 76 One hypothesis can potentially explain the pathogenic mechanism in the case 16 family. *This hypothesis proposes that, a cryptic INV occurred in the proband, producing the Alt2 SV configuration, and causing TADs to be altered for both copies of PLCB4.*

In summary, I investigated a syndromic CRS case characterised by a highly complex SV near the *PLCB4* locus. The presence of asymptomatic carriers in the family posed an interesting challenge for variant interpretation. Detailed WGS analysis suggested the involvement of cryptic INVs, which gave rise to 12 hypothetical configurations to explain the WGS data.

To tackle the cryptic INVs, Bionano OGM was used to characterise the SVs in the case 16 family. However, spanning the 620 kb informative segment with intact DNA molecules proved to be challenging even for Bionano OGM. Nonetheless, Bionano OGM still effectively characterised shorter segments of the SV, therefore narrowing down the possible configurations to only Alt 2 and Alt 8. Additionally, the full trio for this case were analysed using Bionano OGM, and no other compelling SV/CNVs, *de novo* or inherited, were identified.

Functional analysis was carried out in silico using DeepC, determining that Alt 2 is more likely pathogenic than Alt 8, owing to TAD disruption of both copies of *PLCB4.* This suggested an interesting hypothesis, whereby a cryptic INV could have occurred in the proband, causing rearrangement of the inherited benign Alt 8 into the more pathogenic Alt 2.

Overall, this case presented a highly complex SV that supplements the existing knowledge on cryptic INVs documented in the literature. This further highlights the importance of applying long-range technologies, such as Bionano OGM, to understand these clinically relevant CPX SVs.

5.7 Discussion and Summary: challenges and future directions

5.7.1 Uncovering clinically relevant SVs

Examining underlying genetic causes through the lens of CRS revealed several SVs that were previously overlooked due to both technological limitations and computational challenges. Conventional technologies, such as array CGH, are effective in detecting simple CNVs, yet they can lead to inaccurate classification of CPX SVs. Case 2 (chr16-17 translocation, **section 5.5**) and case 16 (chr20 CPX SV with cryptic INVs, **section 5.6**) are excellent examples, demonstrating that simple CNV gains detected by array can, in fact, represent much more complex events. Crucially, the clinical significance of a variant can be drastically different depending on whether the event is interpreted as a simple CNV or a complex SV. This is also evident in case 19 (*ENPP1* complex INS, **section 5.3**), where the full characterisation of the SV led to a revised diagnosis and consequently altered clinical management for the patient. In addition, case 12 (chr7 large INV, **section 5.2**) serves as a prompt reminder that some SVs, such as INVs, simply cannot be detected by array-based technologies.

In contrast, sequencing-based technologies offer enhanced detection capability by generating additional information for CPX SVs compared to array, such as read information for detailed break junction characterisation. However, sequencing-based technologies often face major computational challenges in identifying and interpreting the substantial volume of SV calls. These challenges can lead to missed SV diagnoses. For example, highlighted in case 12, the causative INV was in fact detectable in the WGS data, while large number of false positive INV/BND calls significantly hindered the detection of the true pathogenic INV. Furthermore, despite offering additional

information, short-read technology still faces limitations in fully characterising large CPX SVs, such as in case 2, 10 (CPX INS near *FGF9*, **section 5.4**), 16, and 19.

5.7.2 Bionano OGM: overcoming technological and computational constraints

Bionano OGM, in comparison to cytogenetic and short-read technologies, addresses both technological and computational limitations in SV analysis. For example, case 12 and 19 SVs were characterised extremely efficiently using the Bionano OGM, providing much needed diagnostic speed compared to the other technologies. This efficiency can be partially attributed to the robust internal control database and the integrated analysis pipeline of Bionano OGM. These features can effectively filter out the noisy SVs with minimal computational effort - a distinct advantage that sets it apart from the often hugely inflated datasets generated by sequencing-based approaches.

The technological limitation of Bionano OGM was thoroughly assessed using cases 2, 10, and 16. In the case 10 SV, the minimum informative segment (~242 kb) measures just under the advertised molecule detection limit of the Bionano OGM (~300 kb). This SV was successfully characterised by Bionano and later confirmed via FISH. Cases 2 and 16 were more challenging for the Bionano, as the longest informative segments for these two SVs exceeded the advertised detection limit of Bionano OGM. However, with good quality DNA, Bionano OGM did produce ultra-long informative molecules at ~700 kb – 1 Mb to span some of the informative segments. However, the limited depth, usually only one or two ultra-long molecules, leads to conclusions that are considerably less confident. One possible solution includes repeating multiple Bionano

runs to achieve a high accumulative depth, while the associated cost may be practically undesirable.

The overall potential for Bionano OGM to be implemented in the diagnostic setting is promising due to its performance. However, the major limitation of the Bionano OGM is its requirement for higher quality input samples compared to other genomic technologies in order to achieve optimal performance. For example, Bionano OGM requires fresh blood as the source material, which must be processed immediately or stored/transported as snap-frozen blood. However, rapid degradation of fresh blood significantly affects data quality, as reflected in the low performance observed in several of our initial Bionano runs. Alternatively, snap-frozen blood may prove difficult to transport and store, given that access to dry ice and -80°C freezers is not commonly available in diagnostic settings. Subsequent DNA extraction and processing are also exceptionally delicate procedures that ultimately determine whether ultra-long informative molecules can be generated. Lastly, widely available genomic DNA cannot be used for Bionano OGM, necessitating the re-collection of fresh blood for potentially every patient – a luxury that might not be available for most cases. These limitations collectively demonstrate the complexities associated with implementing Bionano OGM in real-world diagnostic setting despite its exceptional performance in SV analysis.

5.7.3 Addressing the remaining genetic diagnostic gaps in CRS

From the 20 families investigated using Bionano OGM, 12 cases persisted without a genetic diagnosis despite clear suspicions of an underlying genetic component based on their phenotypes. These particular cases were also comprehensively scrutinised

for SNPs and small indels as part of the 100kGP or local studies by Dr Rebecca Trim (née Tooze) during her thesis research. Yet no pathogenic variants, small or large, have been identified to date. These unsolved cases raise an interesting question about the genetic diagnostic gap in CRS. Several remaining aspects of SV analysis may guide future research to tackle these unsolved cases.

One aspect is the consideration of novel diseases genes that could contribute to the cases with missing diagnosis. Since CRS is a highly genetically heterogeneous condition involving multiple stages and pathways of craniofacial development, it is reasonable to expect novel disease genes to be identified in the future. One example here is case 21 (**Chapter 4 section 4.3**), as *HOXC* genes were not previously extensively studied in craniofacial abnormalities nor implicated in CRS. Furthermore, recent literature has revealed an additional 16 genes newly associated to CRS[73], further highlighting the potential for discovering novel contributors to this condition.

A second aspect stems from the fact that the extensive non-coding parts of the genome remain largely a mystery for clinical genomic studies. Conventionally, variant prioritisation often places the non-coding region lower on the list due to the challenges associated with interpreting these variants and the sheer quantity of non-coding variant calls. However, a few SVs have shown that rearrangements involving non-coding regions may be pathogenic. For example, case 11 (an unpublished ongoing parallel project in the research group led by Dr Eduardo Calpena) revealed a non-coding DUP affecting a critical neural crest enhancer of *FOXD3* leading to CRS, validated using mice models. Case 10 follows a similar trend, where a non-coding

variant may affect the adjacent candidate gene *FGF9*. Case 23 SV (chr17 DEL near *KCNJ* locus, **Appendix, Supplementary Figure 3** and **Supplementary Figure 4**) represents another instance where non-coding SVs affecting the TAD boundaries at *KCNJ* locus are likely pathogenic, contributing to the cleft palate phenotype in case 23. Enhancing our understanding of the structures and functions of the non-coding genome has the potential to further narrow down the diagnostic gap in CRS and other human diseases by enabling better evaluation and interpretation of non-coding SVs. Encouragingly, efforts in this direction are already evident, particularly using the latest the Telomere-to-Telomere (T2T) genome.[195,196]

Lastly, reduced penetrance and mosaicism might be another aspect contributing to the diagnostic gap in CRS cases. Together with the **Chapter 3** findings, several rare candidate SVs remain VUSs due to an unexplained healthy carrier in the family. For example, the case 16 candidate stands out as one of the most complex and most extensively investigated SVs in the project, given the potential effect on a known craniofacial disease gene *PLCB4*. However, despite intense scrutiny, explaining the unaffected carriers in the family remains challenging. One hypothesis was discussed in **section 5.6.3**, suggesting the possible presence of cryptic INVs, causing the proband to carry a more pathogenic configuration of the candidate SV. Other examples of reduced penetrance of CRS cases have been reported, such as in cases with pathogenic *SMAD6* mutations.[197] Case 21 (*HOXC* case, **Chapter 4**) again is a comparable example due to the paternal mosaicism, where the father's milder (but definite) phenotype was annotated as unaffected in the 100kGP. Consequently, the analysis initially focused on *de novo* variants, overlooking inherited SVs. Overall, reduced penetrance and mosaic SVs are still incredibly challenging to detect and

interpret, and therefore likely to remain as the last resort unless otherwise evidenced phenotypically. Nevertheless, a crucial step is that the clinical features must be investigated thoroughly so that the correct segregation pattern can be established to facilitate the accurate interpretation of candidate SVs.

Chapter 6 Result – Comparative analysis of Bionano OGM, Illumina WGS, and ONT WGS

6.1 Introduction

Compared with OGM, orthogonal ONT based long-read sequencing technologies have been steadily gaining ground since the introduction of their first sequencer, the MinION, nearly a decade ago.[198] ONT-based sequencing offers greater read lengths compared to conventional short-read technologies, but at the cost of higher error rates at certain genomic features, such as homopolymers.[199] With the longest read length amongst read-based technologies, ONT is well-equipped to bridge over gaps, repeats, and other challenging regions of the genome. In comparison to OGM technology, as described in **Chapter 4** and **Chapter 5**, ONT, despite generally having a shorter N50 (note that N50 > 1 Mb can be achieve through specific protocol), maintains sequence readability - a critical information loss in OGM. Overall, in the investigation of clinically relevant SVs, ONT was hypothesised to be able to bridge the gap between OGM and short-read technologies, offering at least a tenfold longer read length while retaining sequence readability. This chapter aims to examine the performance of ONT compared to Illumina WGS and Bionano OGM – the two technologies I have employed to investigate the genetic bases of patients with CRS.

As part of a 100kGP-pilot program, 35 individuals (13 families) from the CRS cohort were sequenced using the ONT PromethION by our GE collaborators at the Sanger Institute (led by Prof Greg Elgar). The subsequent bioinformatic analysis pipeline, including alignment and SV calling, was carried out by the collaborators in three different batches. These three batches differ by flow-cells used and the tool versions in the pipeline, as shown in **Table 21.** The most notable difference in the bioinformatics analysis between the three batches was the improved variant calling capability from the new version of Sniffles SV caller, which is discussed in **section 6.3**.

Table 21 Major differences between the three batches in ONT analysis

	BATCH 1	BATCH 2	BATCH 3
# of samples	11	17	7
ONT Flow-cell	R9	R9	R10
Minimap2	2.20-r1061	2.24-r1122	2.24-r1122
Sniffles	1.0.11	1.0.11	2.0.6
Samtools	1.11	1.11	1.11
Tabix	1.9	1.9	1.9

Minimap2 is for read alignment and mapping; Sniffles/Sniffles2 is the SV variant caller used specifically for ONT data; Samtools and Tabix are used for bam file processing, such as sorting and indexing.

Out of the 13 families, eight trios were included in the comparative analysis of the three technologies, as summarised in **Figure 77**. A three-way comparison was carried out for these eight trios, as they have been thoroughly investigated using all three technologies – Illumina WGS, ONT WGS, and Bionano OGM. As described in **Chapter 2**, Bionano OGM SV calls were set out to be the reference callset, and the capability of Illumina and ONT to detect and to call SVs using their respective SV callers was assessed against the OGM callset.

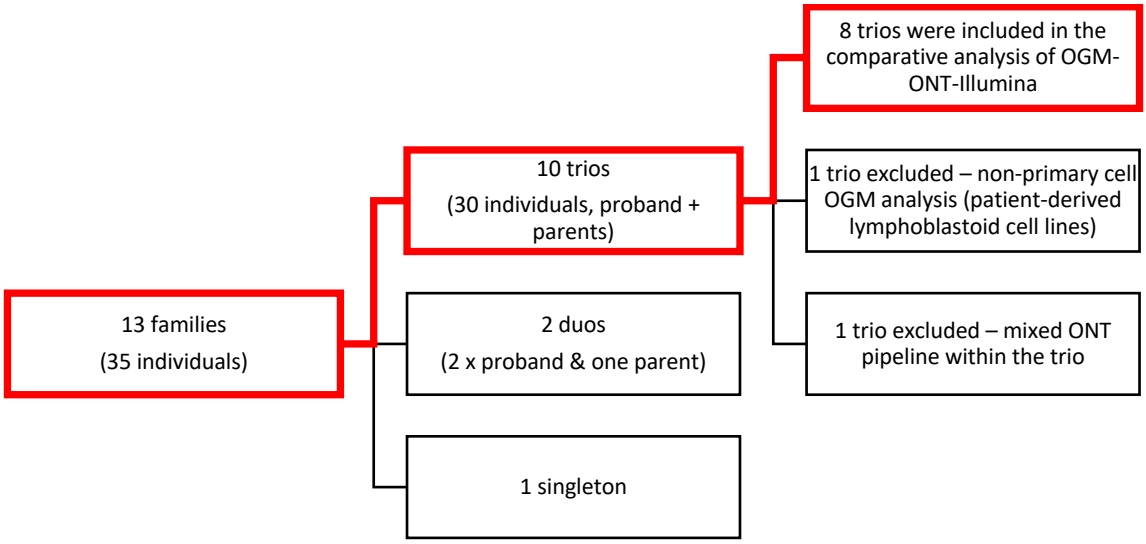
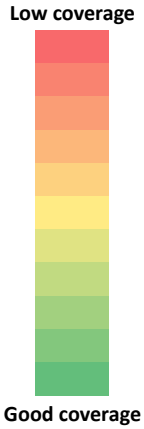


Figure 77 A total of 35 individuals from 13 CRS families were analysed on the ONT PromethION as part of the 100kGP pilot program*. Out of the 10 trios, eight trios were included in the comparative analysis between OGM, ONT WGS, and Illumina WGS. Two trios were excluded since OGM of one trio is from non-primary cells, while the other trio contained mixed ONT analysis pipeline due samples in different bathes. Add a bit more info here on why the 2 trios were excluded.*

The ONT run quality is summarised in **Table 22**. While most cases achieved coverage comparable to that of the Illumina WGS at 30x, several cases, particularly Case 5, had poor coverage that was very likely to affect the SV detection performance. The N50 ranged from 16 kb to 49 kb, providing a good middle ground between Illumina WGS and Bionano OGM. Interestingly, no correlation was observed between the N50 and yield (coverage) in this dataset.

Table 22 ONT data quality for all 35 individuals, including coverage and N50

CaseID	Sample	Estimated bases produced (Gb)	Estimated N50 (kb)	Estimated coverage (X)	Comparative analysis
1	Proband	122.29	18.76	38	N, not trio
	Mother	99.23	19.85	31	
3	Proband	156.92	26.62	49	Y
	Mother	104.41	24.59	33	
	Father	115.23	28	36	
4	Proband	84.33	29.38	26	Y
	Father	90.91	27.75	28	
	Mother	95.94	24.76	30	
5	Proband	9.94	27.72	3	Y
	Father	37.93	19.73	12	
	Mother	13.31	22.26	4	
6	Proband	57.52	27.23	18	N, not trio
	Mother	52.03	46.92	16	
7	Proband	95.96	21.89	30	N, mixed pipeline
	Mother	49.94	30.66	16	
	Father	35.71	32.48	11	
8	Proband	48.88	22.6	15	Y
	Father	55.98	34.31	17	
	Mother	46.88	18.96	15	
9	Proband	56.59	30.91	18	Y
	Father	47.56	24.24	15	
	MAP	73.86	18.73	23	
10	Proband	57.84	29.2	18	N, non-primary cell source
	Father	48.16	34.73	15	
	Mother	46.4	35.89	15	
11	Mother	107.64	37.38	34	N, not trio
13	Proband	85.31	29.92	27	Y
	Father	82.81	35.21	26	
	Mother	92.52	28.3	29	
16	Proband	73.77	19.24	23	Y
	Mother	85.13	18.69	27	
	Father	85.45	19.07	27	
17	Proband	53.16	18.53	17	Y
	Mother	77.93	20.56	24	
	Father	70.96	16.15	22	



De novo SV analysis using the ONT data was planned, hypothesising that ONT may detect SVs missed by Bionano OGM and Illumina WGS. However, initial examination of the Batch 1 ONT data identified poor performance in SV calling. Despite the significant improvement from the enhanced analysis pipeline used for the latest batch of data, a comprehensive SV analysis for novel candidates in the ONT data was ultimately beyond the scope of the project.

Instead, this chapter serves as a preliminary investigation, delving into the strengths and limitations of SV detection by ONT in a clinical setting compared to conventional and orthogonal technologies, such as Illumina WGS and Bionano OGM, respectively.

6.2 Bionano benchmarks

From the eight trios, a total of 234 rare SV/CNV events were called within the Bionano Access software, as described in **section 2.9.4**. These rare calls predominantly consist of DEL and INS, with a small number of DUP (**Table 23**). CNVs, by Bionano Access definition, are events larger than 500 kb, and the complex (CPX) SV refers to the chr20 event, which originally consists of multiple INV & DUP calls.

To normalise these event types to be comparable with Illumina and ONT calls, a reclassification was made as summarised in **Table 23**. The three CNVs were reclassified into DUPs/DELs based on sequence evidence from Illumina and ONT. For the chr20 CPX SV, several INV & DUP calls were made – these were reclassified into a single INS event to represent the most likely configuration as discussed in **section 5.6**. Notably, no true INV or BND were called as rare SVs in these 8 trios of Bionano.

Table 23 OGM SV types were normalised into DEL, tandem DUP, and INS.

	CNV	CPX	DEL	Tandem DUP	INS
OGM Calls	3	1	136	12	82
Reclassified/normalised	0	0	137	36	61

Total of 234 calls were collected from and benchmarked for Bionano OGM.

The most significant reclassification was made for several INS calls from the Bionano Access. This was primarily due to the limitation imposed by the Bionano OGM's labelling density, which typically averages around 6-7 kb between any two labels. Consequently, while Bionano OGM can detect an increased distance between the labels, indicating a possible gain of genetic material, it cannot pinpoint the source of the extra material if there is an inadequate number of informative labels.

For example, as illustrated in **Figure 78**, Bionano Access made an INS call, which is likely a tandem DUP based on the read information from both Illumina and ONT. The occurrence of this specific type of misclassification in the Bionano callset is summarised in **Figure 79**. Out of the 94 possible DUP/INS events, Bionano Access made 71 correct calls, resulting in an 75% accuracy at DUP/INS calling. Specifically, when calling INS, Bionano Access is 72% precise, with 28% false discoveries (true DUPs falsely called as INS); when calling DUPs, Bionano Access is only 34% sensitive, as the remaining 66% DUPs are misclassified as INSs (false negative rate/miss rate).

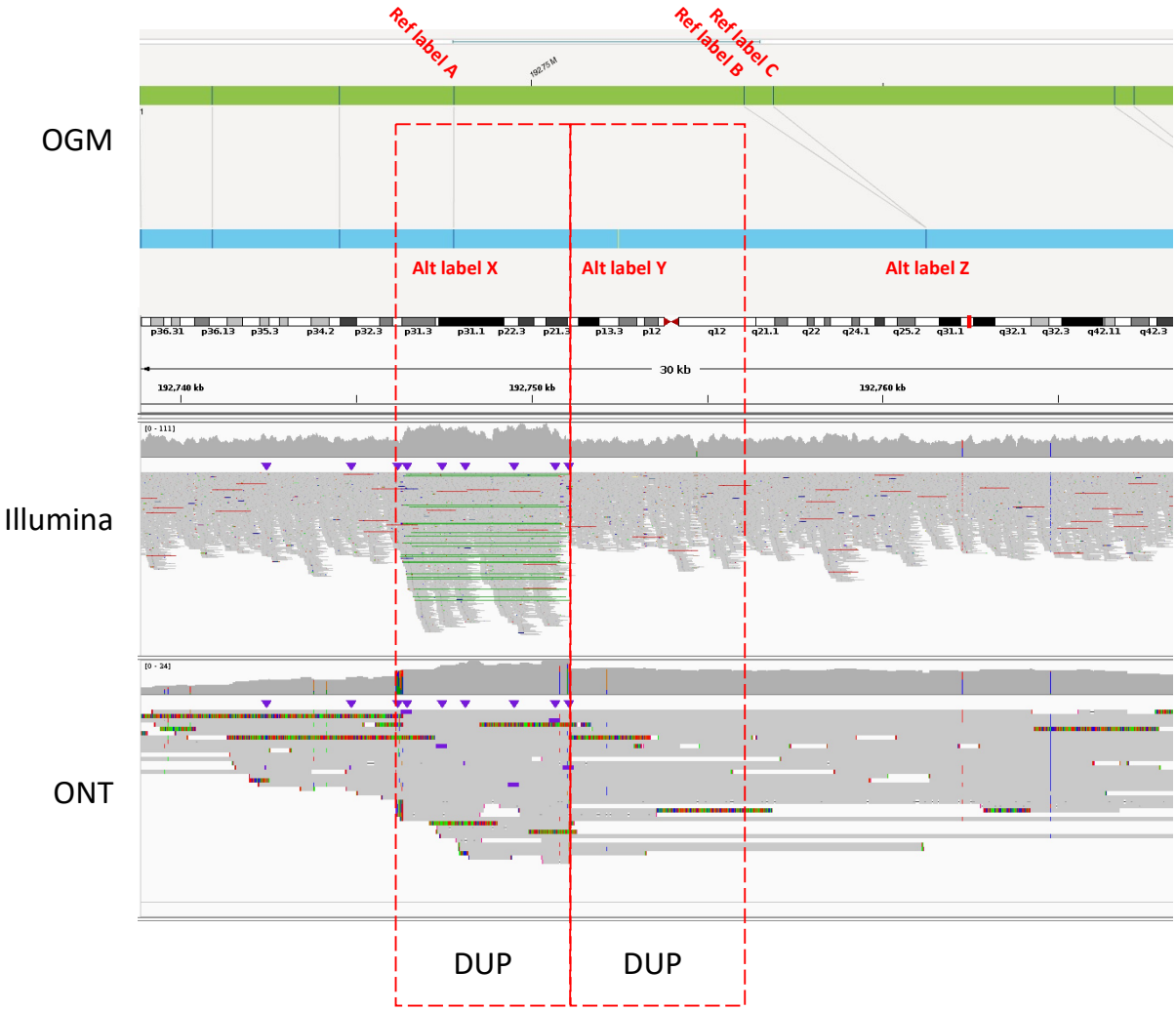


Figure 78 Example of an INS call made by the Bionano Access software. *However, when assessed against the WGS data, this INS was reclassified as a (tandem) DUP. Three data tracks from OGM, Illumina, and ONT are aligned based on reference coordinates in hg38. Evidence of the DUP can be observed particularly clearly in the Illumina track, where paired reads (green lines marking linked reads) show the pattern of a tandem DUP. In the OGM tracks, three labels of interest are highlighted for both the reference (Ref) map and the patient allele (Alt). Based on the Illumina and ONT data, the two copies of the duplicated segments are marked using two red boxes. Therefore, it can be deduced that labels X and Y represent two copies of the same label due to the DUP, both mapping to label A. Labels B and C map to label Z, possibly involving a dropout or merging of B and C. IGV figures exported from GE Airlock. Figure in hg38.*

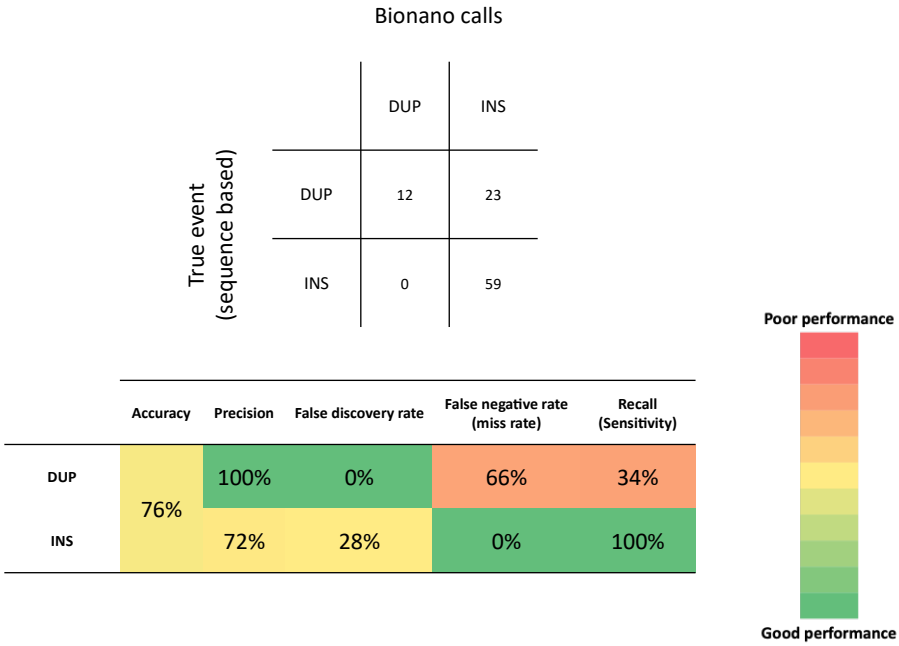


Figure 79 Confusion matrix and metrics for the INS and DUP events called by the Bionano Access pipeline. *A true event is determined based on read evidence from ONT and Illumina WGS. Only reclassifications between DUP and INS were considered, ie CPX events reclassified into INSs were excluded here.*

Without considering the SV types, out of the 234 OGM calls, 12 calls were deemed false or uncertain due to the following reasons:

- Seven calls were near or at centromeric regions characterised by highly polymorphic genotypes for all three technologies;
- one call was made at a challenging locus on chrY;
- one call was a small DEL of ~500 bp, near the detection limit;
- one call was a large CNV without depth/coverage support from any of the three technologies.

The remaining 222 calls were considered as true events, as evidenced by the consensus derived from OGM molecules and sequencing reads. Overall, this makes Bionano OGM ~95% precise at detecting rare SV events > ~500 bp, setting the

Bionano OGM callset as a reliable “truth set” for benchmarking the performance of Illumina and ONT WGS.

6.3 Illumina and ONT WGS performance – Overview

Illumina and ONT performances were evaluated using the set of normalised SVs from Bionano OGM as discussed above (**section 6.2**). Batch 1 and 2 were evaluated separately from batch 3 for the ONT data, due to differences in the analysis pipelines and flow-cells, as described in **section 6.1**. There was no difference in the analysis pipeline for Illumina data between “Batches” as all processes were carried out by Genomics England as part of the 100kGP. The overall performance of Illumina and ONT is summarised in **Figure 80**.

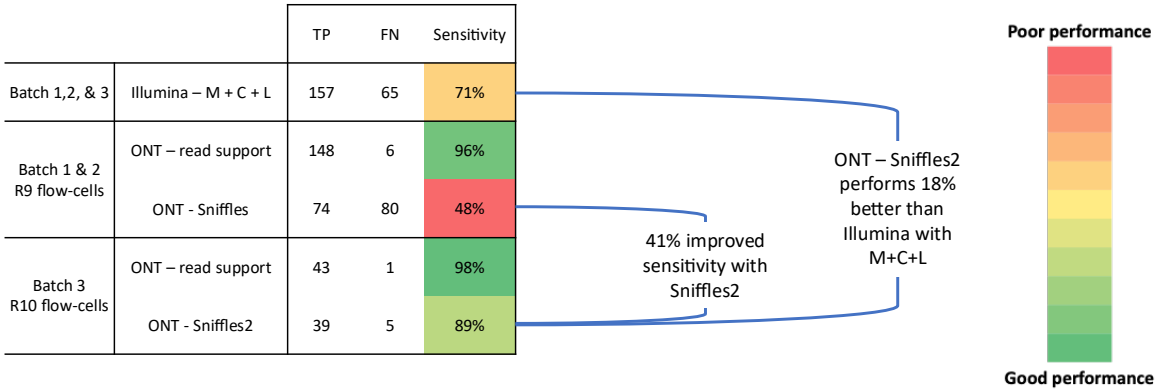


Figure 80 Performance evaluation for Illumina and ONT compared to the truth set of the normalised Bionano SV calls. *Illumina SV calls consists of a union of three variant callers, Manta (M), Canvas (C), and Lumpy (L). Due to the suboptimal performance of Sniffles in Batch 1 & 2, ONT read support was manually evaluated to determine if the reduced performance originated experimentally or computationally. Unpaired breaks (BNDs) were considered as negative calls.*

One critical observation was the minimal difference in the ONT performance purely based on read support between Batch 1 & 2 (R9 flow-cells) and Batch 3 (R10 flow-

cells). This suggests that the new R10 flow-cells provide no particular benefit in SV detection. This finding is reasonable as the primary improvement for the R10 flow-cells is the reduced base-calling error, which is mostly irrelevant for SV detection. Consequently, this suggest that the improved ONT performance in SV detection is likely due to computational modifications, specifically the change to Sniffles2 as the SV caller for ONT (**Table 21**).

As highlighted in **Figure 80**, the original Sniffles underperformed significantly with only 48% sensitivity, missing more than half of the OGM SVs in batch 1 & 2. In comparison, Sniffles2 improved with 89% sensitivity, exceeding even the performance of Illumina data with the union of three callers. This improvement in Sniffles2 may be attributed to the added repeat awareness and a coverage-adaptive filter when calling SVs.[200] This demonstrates that ONT indeed has high potential to enhance SV calling compared to Illumina, if performed with the improved Sniffles2 SV caller.

6.4 Illumina and ONT WGS performance – SV types

To gain a deeper understanding of the performance difference between ONT and Illumina, SV calls were stratified by the three detected SV types. **Figure 81** summarises the sensitivity of ONT and Illumina for the three SV types. As shown, one of the most striking differences between ONT and Illumina is the INS sensitivity, where ONT-Sniffles2 is 57% more sensitive than Illumina with three callers combined. It is also interesting to highlight that, in data from the original Sniffles, all three types of SVs underperformed equally, with DEL performing marginally better, likely due to the

larger number of calls made as DELs. This indicates that Sniffles underperforms in all SV types, and there's likely no type-specific issues.

		SV type	TP	FN	Sensitivity
Batch 1, 2, & 3	Illumina - M + C + L	DEL	115	19	86%
		DUP	29	1	97%
		INS	13	45	22%
Batch 1 & 2 R9 flow-cells	ONT - read support	DEL	93	1	99%
		DUP	23	2	92%
		INS	32	3	91%
	ONT - Sniffles	DEL	52	42	55%
		DUP	9	16	36%
		INS	13	22	37%
Batch 3 R10 flow-cells	ONT - read support	DEL	26	0	100%
		DUP	4	0	100%
		INS	13	1	93%
	ONT - Sniffles2	DEL	25	1	96%
		DUP	3	1	75%
		INS	11	3	79%

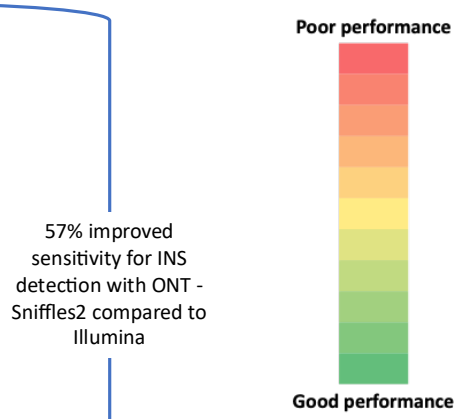


Figure 81 SV detection evaluation for ONT and Illumina compared to Bionano OGM SV calls, stratified into the three detected SV types.

The surprisingly good ONT performance in INS detection is likely attributable to the long-read nature of ONT, providing additional anchoring capability for reads affected by INS events. As shown in **Figure 82**, an example INS of 2.7 kb can be clearly observed from ONT reads, as the ONT long reads can confidently span the entire INS sequences. In contrast, Illumina WGS cannot effectively anchor the short reads fully contained within the INS, resulting in failure to map these reads to the correct locus. Bionano OGM, on the other hand, detected the extra 2.7 kb sequence but could not pinpoint the source of this INS event due to the limit of labelling density. Analysis of the ONT reads further revealed that the inserted sequences consist of 2.5 kb of SVA_E

and a short $(GAGGGA)_n$ repeat, similar to the region of chr1:48,381,684-48,384,402 in the hg38 reference.

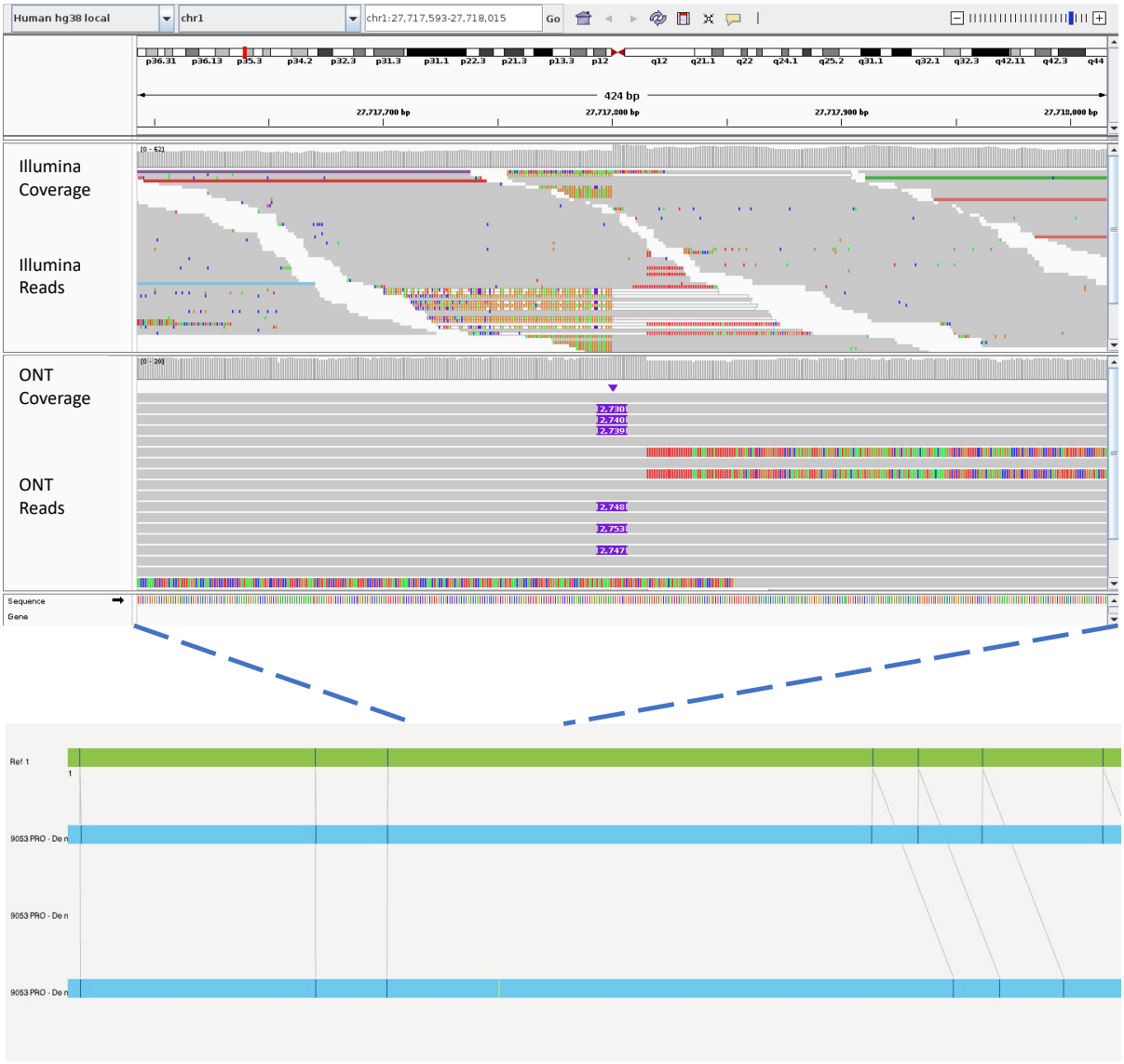


Figure 82 Example INS captured by ONT but not detected by Illumina WGS. *Purple boxes in the ONT track clearly highlight the presence of the 2.7 kb INS. Illumina short-reads struggle to anchor properly due to the size of the INS event. OGM was able to detect the INS, while the lack of labels cannot accurately pinpoint the origin of the inserted sequences. IGV figure exported from GE Airlock. Figure in hg38.*

6.5 ONT performance – *de novo* calling

The originally intended application for the ONT data in my project was to conduct trio analysis, looking for *de novo* SVs that might have been overlooked by both Illumina WGS and Bionano OGM. Therefore, the performance of *de novo* calling in the trios was evaluated. As shown in **Figure 83**, the first two batches demonstrated poor performance in *de novo* calling, with 27% of the inherited SVs lacking the parental call, consequently being falsely classified as *de novo*. This poor performance of the initial ONT data rendered trio analysis uninformative due to the large portion of false *de novo* calls. In contrast, the latest batch data with Sniffles2 improved significantly, with only 8% inherited SVs falsely classified as *de novo*. This improvement is likely the result of the improved overall sensitivity of Sniffles2, as described in **section 6.3** and **Figure 80**.

		FP *de novo* (parental call missing)	TN *de novo* (true inherited calls)	False positive rate (probability of false *de novo* calls)
Batch 1 & 2	ONT-Sniffles	19	55	26%
Batch 3	ONT-Sniffles2	3	36	8%

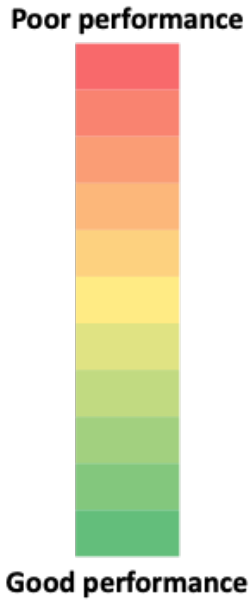


Figure 83 De novo calling performance improved in batch 3 using Sniffles2 compared to the first two batches using Sniffles. *False positive (FP) de novo calls are SVs inherited from a parent, where the parental SV was not called by Sniffles or Sniffles2. True negative (TN) de novo calls are inherited SVs called correctly in both the proband and the parent. False positive rate was calculated as FP/(FP+TN). One TP de novo variant was called in Batch 3 and not included in this table.*

To gain an understanding of the scale and feasibility of potential *de novo* analysis in the future, a test analysis was carried out for the latest two trios, cases 16 & 17 in batch 3. Overall, 268 (out of 38,260 total calls from case 16) and 414 (out of 38,466 total calls from case 17) *de novo* calls were made by Sniffles2 for case 16 and 17, respectively. These calls have at least one read support for the proband and none for both parents. Out of the *de novo* calls, 5 (case 17) and 3 (case 16) were shown to intersect with a coding region (**section 2.10.4**). Three of the coding calls were discarded as they affect highly polymorphic regions of the *MUC* genes. The remaining 5 candidate SVs are summarised in **Table 24**. After individual inspection on IGV, these 5 remaining candidates cannot be verified true, based on the minimum number of supporting reads. BND reads posed significant challenges to understand, as these are reads with partially soft clipped sequences without reciprocal/corresponding reads from the other direction. The 5 candidate calls were also cross-examined using Illumina WGS reads and Bionano OGM molecules, which were all inconclusive; Illumina presented soft clipped reads without mate reads; Bionano OGM showed no abnormal label or labelling distance. Overall, these calls may require further experimental work, such as PCR, to fully understand their nature.

Table 24 five coding, potentially de novo candidates from the ONT-Sniffles2 analysis

Case	Chromosome	Position	SV length	SV type	Sniffles2-ref reads	Sniffles2-alt reads
17	chr12	52517168	.	BND	2	2
17	chr16	69961666	6377	INS	5	3
17	chr4	39308656	.	BND	4	2
16	chr15	43572861	.	BND	2	2
16	chr19	1993025	70128	INS	1	11

Ref reads are reads classified as reference reads by Sniffles2 while alt reads are reads classified as variant reads.

Following the improved *de novo* calling from Sniffles2, re-calling and *de novo* analysis of coding regions was performed for the entire ONT cohort of 10 trios (**Figure 77**). This exhaustive analysis revealed a total of 13 unique (allele account = 1 in the entire cohort) *de novo* coding calls. As observed in the initial *de novo* analysis on case 16 & 17 (**Table 24**), these 13 *de novo* calls were either one-ended BND calls or calls with minimum read support, requiring further PCR examination to verify. Although this verification falls outside the scope of this project, it is indeed a promising direction for future research endeavour.

6.6 Discussion

In this study, I conducted a comprehensive comparative analysis of eight trios amongst three different genomics technologies: Illumina WGS, ONT WGS, and Bionano OGM. Using the Bionano OGM callset as the truth set, the performance of WGS and ONT was assessed relative to that of the Bionano OGM. Each of the SV calls was manually curated and examined using all three technologies, providing a thorough and detailed performance assessment in a validated set of rare SVs.

Several key findings emerged from the comparative analysis:

1. Bionano OGM is highly precise, with 95% of rare SVs called as true events. However, Bionano OGM also misclassified 65% of tandem DUP as INS, due to insufficient resolution for smaller DUPs, usually less than ~10 kb.
2. Illumina WGS, with the union of three SV callers, successfully called 71% of events from the Bionano callset, with the most poorly detected SV type being INS at only 22% sensitivity.

3. ONT performance is highly dependent on the variant caller, with the original Sniffles performing poorly at 48% sensitivity and Sniffles2 a significant improvement at 89% sensitivity.
4. ONT performs much better at INS detection compared to Illumina.

These finding have important implications for detecting clinically relevant SVs. The high ~95% precision of Bionano OGM highlights its advantages in clinical settings, particularly where minimising false positive findings is crucial for the prompt and efficient delivery of clinically relevant results. Additionally, it is encouraging that Illumina and ONT WGS were able to identify 71% and 89% of the true Bionano SVs, respectively. This suggests that the OGM technology indeed provides added value, particularly in uncovering overlooked SVs from WGS.

INS detection remains one of the most significant challenges for Illumina WGS due to the limitation of short reads. Given that Illumina is the most common WGS approach in the clinical setting, there is likely a systematic underrepresentation of INS in clinically relevant SVs. Similar challenges with INS detection via Illumina have been reported in the literature, where Manta was benchmarked to have only ~8.6% sensitivity in real Illumina datasets.[201] Given the truth dataset used by Kosugi et al (2019)[201] consists of only PacBio SVs and existing DGV SVs from the samples, the reported sensitivity may not fully represent the diversity of INS in the human genome. In contrast, ONT-Sniffles2 likely offers the best solution for INS detection among the three technologies assessed, leveraging the advantage of long reads to achieve much higher sensitivity

compared to Illumina, while maintaining sequence readability, which may be crucial to understand the true pathogenicity of INS events.

The major limitation of the comparative analysis in my project is the lack of false negative assessment of the Bionano truth set. This limitation arises due to the challenges in identifying SVs missed by Bionano OGM. A complementary analysis to address this would involve deriving a truth set from the Illumina and/or ONT WGS data. However, given the project's focus on manually curating an accurate set of clinically relevant rare SVs, it was impractical to undertake the same approach for the WGS data, primarily due to the substantial number of SV calls, likely mostly false positive calls, in the datasets.

The truth set issue seems to be highly prevalent when benchmarking different genomic technologies. Previous attempts have used simulated data by introducing artificial SVs into the test genome.[201] However, these efforts have revealed a near perfect SV calling performance in the simulated data, which does not align with assessments from real data. More recent efforts, such as the Genome in a Bottle (GIAB) Consortium, have aimed to compile a comprehensive truth set of SVs identified by a diverse range of genomic technologies, including Illumina, ONT, and PacBio.[202] Bionano OGM was also included, albeit primarily for verifying the SV length rather than for novel discoveries. Further comparative analyses of Bionano OGM performance against GIAB data could provide an interesting perspective. Another source of missed SV calls could stem from the inherent discrepancies in the hg38 reference genome. The recent Telomere-to-Telomere (T2T) assembly was undertaken to improve the accuracy of

the reference genome.[195] However, it is important to note that many existing pipelines and databases for variant interpretation are yet to be fully migrated to the T2T assembly, making it less informative in the clinical setting.

One further interesting observation is that there were only two *de novo* SV events in the 234 rare Bionano callset. This aligns with the consensus in the literature, stating that there is likely <1 *de novo* SV per generation.[203,204] However, within genomic regions accessible by sequencing technologies, there does seem to be a certain level of false positive *de novo* calls, as evidenced by the ONT-Sniffles false positive *de novo*. Similar observations were made from gnomAD SVs, where the apparent 3.0-7.4% *de novo* SVs were believed to be predominantly false negatives in the parents and/or false positive in the children.[205] In my project, ONT with Sniffles2 performed well to reduce the false *de novo* calls by minimising false negative parental calls, while the false negative SVs in the children are yet to be assessed. Using the inbuilt trio-calling function in Sniffles2, 682 *de novo* calls were made from two ONT trios, making it ~350 *de novo* calls per trio. Given the consensus of expecting <1 *de novo* SV per generation, the number of *de novo* calls by Sniffles2 is highly overinflated with false positive calls. However, when compared to the 3.0-7.4% *de novo* rate in gnomAD, Sniffles2-ONT performed better with only ~0.9% *de novo* rate (682 *de novo* calls out of 76,826 total calls from two trios), suggesting either a lower false positive rate or a higher false negative rate for ONT compared to short reads in gnomAD, though the latter (higher false negative for ONT) is less likely, as supported by the better sensitivity using ONT-Sniffles2 compared to Illumina as described in **Figure 80**.

Furthermore, the full trio analysis using the Bionano OGM, despite the small sample size, did not yield any *de novo* SV using the inbuilt trio analysis pipeline. These pieces

of evidence all support the hypothesis that *de novo* SVs affecting coding regions or functionally critical non-coding regions are exceedingly rare events in the human genome, emphasising their importance in clinical investigations.

With the improved Sniffles2 calling, *de novo* analysis was carried out to assess the feasibility of future SV analysis using ONT data in the clinical setting. This analysis initially focused on the latest two trios, yielding a manageable number of candidates calls as shown in **Table 24**. Following this robust *de novo* analysis, the entire cohort of 10 trios were re-analysed using Sniffles2, yielding a total of 13 potential *de novo* calls affecting coding regions. However, these calls could not be verified even with supporting data from Illumina WGS and Bionano OGM. Additionally, the low number of reads, both reference and alternative in most calls (**Table 24**), suggests that these calls likely represent ONT specific artefacts. Future PCR analysis is required to verify and understand these variant sequences.

It is important to acknowledge that the decision to analyse Sniffles2 calls in the coding region only was due to time constraints at the end of the project, but it does mirror the time and resource limitation typically encountered in clinical settings. However, similarly to the WGS analysis, focusing on the non-coding regions to look for small-medium sized INS holds great potential for identifying candidate pathogenic variants. This strategy may be particularly promising as these small to medium sized INSs are most likely to be missed by both Illumina WGS and Bionano OGM.

Overall, in this preliminary comparative analysis of the three datasets, I evaluated the performance of short- and long-range technologies in the clinical setting, with a focus

on identifying rare/*de novo* SVs that are likely to have clinical implications. Bionano OGM stands out as a highly precise technology, despite its challenges in occasionally misclassifying small INS/DUP. Illumina WGS demonstrated good sensitivity (71%) compared to Bionano OGM using the union of three variant callers and establishing a robust basis for the WGS analysis detailed in **Chapter 3**. ONT WGS initially underperformed significantly, but the improved analysis pipeline with the enhanced Sniffles2 SV caller set out a strong foundation for future implementation of ONT as a viable approach in clinical SV detection.

Chapter 7 Closing Remarks

7.1 Introduction

The landscape of clinical genetic research has been profoundly shaped by genomic technologies over the last six decades. These technologies have evolved from the original microscopic visualisation of the physical structure of DNA molecules and array-based approaches, to the current widely used short-read sequencing methods. Diagnostic efforts have also shifted from the early focus on whole chromosomal abnormalities, such as trisomy 21, to an emphasis on small sequence variants, such as SNPs and indels. Consequently, much of the research effort has been directed towards aggregating and understanding small sequence variants under 50 bp. On the other hand, variants larger than 50 bp, collectively referred to as SVs, pose inherent challenges in aggregation and interpretation due to their size and type variability. Large scale population projects, such as the 100kGP, have dedicated considerable effort in examining small variants, while SVs remained largely unexplored at the start of the thesis research. Therefore, I set out to address this gap of SVs in clinical genetic research, focusing on a specific craniofacial disorder – CRS.

At the beginning of my research, several objectives were envisaged, focusing on identifying and characterising pathogenic SVs using both short-read and long-range genomic technologies. To achieve this, I firstly analysed the short-read data from the 100kGP to identify clinically relevant SVs in the CRS cohort. I then applied long-range technologies, including ONT and Bionano OGM, to identify SVs that had eluded short read detection and to characterise CPX SVs. Lastly, with data available from all three genome technologies (Illumina WGS, ONT WGS, and Bionano OGM), I evaluated their performance in SV detection specifically in a clinical setting.

7.2 Research summary

My research addressed the above objectives in 4 results chapters. The first result chapter, **Chapter 3**, summarised my analysis of the 100kGP data. From the 114 CRS cases in the 100kGP, 7 likely pathogenic SVs were identified, as summarised in **Table 25**, with the prevalence of 6.1% (95% exact binomial probability confidence interval: $0.02504 \leq p \leq 0.12243$). This modest yield contrasted with the hypothesis that many undiagnosed cases might involve pathogenic SVs. Amongst the 4 discoveries (**Table 25**), 2 cases (**section 3.2.1** and **3.2.2**) were rediscoveries, suggesting that 100kGP recruitment processes were not always highly rigorous. One case (**section 3.2.3**) involved a novel diagnosis (*ARX* DUP) previously missed as the case predated the widespread use of clinical array and sequencing technologies. One final case (**section 3.2.4**) offered the most research interest, since the affected region, the *HOXC* gene cluster, is a novel disease locus not previously associated with CRS. This case led to extensive investigation regarding the clinical relevance of the *HOXC* cluster in syndromic craniofacial abnormalities in **Chapter 4**. The remainder of **Chapter 3** summarised the significance of VUSs in my initial 100kGP analysis, highlighting 3 SVs classified as VUSs due to multiple reasons.

Table 25 Seven likely/pathogenic SVs were identified in the CRS cohort of 144 cases

	SVs	Discussed in
Rediscoveries	*HDAC2* DEL	section 3.2.1
	GPC3 DEL	section 3.2.2
Confirmations	*ERF* DEL	Hyder et al (2021)[90]
	FOXD3 DUP	Hyder et al (2021)[90]
	TWIST1 CPX/INV	Hyder et al (2021)[90]
Novel discoveries	*ARX* DUP	section 3.2.3
	HOXC DUP	section 3.2.4 & Hyder et al (2021)[90]

Rediscoveries were positive independent identifications through my 100kGP analysis without knowing they were previously established; Confirmations were known diagnoses used as positive controls in my analysis; Novel discoveries were previously unknown.

Building on the *HOXC* case, **Chapter 4** explored three 100kGP cases with completely different clinical presentations, all carrying SVs affecting the *HOXC* gene cluster. These three cases, particularly the two DUP cases, presented an interesting contrast: the smaller DUP case exhibited more severe syndromic craniofacial features, while the larger DUP family presented with congenital heart disease without any other major clinical features. Clinical interpretation of these cases was carried out collectively by examining several aspects of the SV, including gene content, TADs disturbance, and the *HOXC* collinearity disruption. Additionally, to characterise the complex break ends of the large DUP, Bionano OGM was firstly introduced, laying the groundwork for **Chapter 5**.

Chapter 5 explored the use of Bionano OGM to investigate the remaining 100kGP and local CRS cases that remained without a genetic diagnosis. The first two cases (**sections 5.2** and **5.3**) demonstrated the efficiency and effectiveness of Bionano OGM in identifying and characterising pathogenic SVs previously missed by conventional technologies. Three subsequent cases pushed the OGM technology to its limit with 3

large CPX SVs. Full characterisation of these 3 SVs required unfragmented reads/molecules over the sizes of ~250 kb, ~500 kb, and ~620 kb. Bionano OGM made an excellent attempt in characterising, though not fully, these CPX SVs, with the practical limit shown to be able to span ~600 kb regions with intact DNA molecules. FISH was employed as a complementary technology for the full characterisation of two of these SVs. As to the clinical yield, Bionano contributed to the identification of crucial novel diagnoses in three cases described in **sections 5.2**, **5.3**, and **5.5**. For the remaining SVs described in **sections 5.4** and **5.6**, further functional analysis is warranted to fully assess their clinical relevance to the CRS phenotype in families.

In the final result chapter, **Chapter 6**, I evaluated the performance of the three technologies employed in this thesis, Illumina WGS, ONT WGS, and Bionano OGM. Bionano OGM was shown to be highly precise with 95% rare SVs calls verified with reads from Illumina and/or ONT. Due to the low false positive rate of Bionano OGM, I evaluated the performance of Illumina and ONT WGS in SV detection in the clinical context, with Bionano OGM data as a reference SV set. The key findings include:

- Illumina WGS with the union of three variant callers achieved 71% sensitivity to detect rearrangements in the reference OGM callset.
- The initial ONT-Sniffles combination showed poor performance with only 48% sensitivity, while the improved ONT-Sniffles2 analysis delivered 89% sensitivity, surpassing the Illumina baseline.
- ONT demonstrated superior INS detection compared to Illumina.
- *De novo* re-calling using Sniffles2 identified several potential SV candidates in the coding region, requiring further experimental validation such as break junction PCR.

Overall, **Chapter 6** highlights the improved ONT analysis with Sniffles2 as an additional approach for detecting specific types of SVs, especially INSs, that might have been missed by both Bionano and Illumina, offering the potential for novel diagnoses for unsolved cases of CRS.

7.3 Technologies and SV detection

Recent advances have emphasised the pursuit of longer reads from different technologies, indeed presenting a compelling avenue for genomic investigation as evidenced in my research. However, while long-reads have proved effective, challenges persist in efficiently analysing novel long-range genomic data. Reflecting on the strategies I have taken to address these challenges, two aspects stood out as pivotal in determining the efficiency of a technology for SV analysis, particularly in a clinical context: useability and versatility.

Useability refers to the ease of analysing and interpreting the raw data, especially crucial in clinical settings, where complex data structure often hinders efficient result delivery. When it comes to WGS, both ONT and Illumina, data processing involves highly specialised, computationally heavy tasks - from raw data to generating the finalised, fully annotated, VCFs. Subsequent filtering and prioritisation processes can also be challenging due to the significantly inflated number of entries in the VCFs. In contrast, the Bionano OGM interface provided a significantly superior user experience, particularly in clinical applications. The Bionano Access platform generates highly accessible results and plots as illustrated in **Chapter 5**. Regarding data volume, instead of relying on externally aggregated control data, Bionano Access contains a

highly effective internal control database of common rearrangements, generated with data using the same technology, which significantly reduced the final output volume to a manageable number (~20-50) in each individual genome. These features overall make the OGM technology much more usable in the clinical setting compared to WGS technologies for SV detection. While enhanced useability was beneficial, Bionano OGM might have incurred increased costs and technical complexities compared to well-established sequencing pipelines, and, most importantly, a compromise in the versatility.

Versatility in this context refers to the extent to which customisation can be applied to the analysis pipeline of a technology for specific situations. This was a significant limitation I have encountered during my Bionano analysis. Although allowing some level of modification, most Bionano pipelines were internal and could not be significantly adjusted. In a few complex cases, alternative methods had to be adapted, such as filtering by molecule length before assembly to force the program into using only the ultra-long molecules for a specific locus, regardless of the overall assembly quality. In contrast, for short-read WGS data, a wide range of open-source tools is available for various situations and analytical goals.

Furthermore, for Bionano in the data presentation phase, I was limited to Circos plots for the overview, and map/molecule plots for detailed analysis and interpretation. Features allowing for manipulation of the presentation of specific molecules were lacking, specifically for filtering molecules by length and position. These features could have significantly improved SV interpretation for several cases, particularly for large

CPX events where only a few informative molecules were present. This lack of versatility in data presentation has indeed made result delivery much more challenging for other researchers and clinicians. In comparison, read-based technologies offered significantly more versatility in data presentation due to mostly open access data and expanded supporting tools.

Another aspect affecting the versatility of Bionano OGM is its unique VCF output. While not strictly necessary for data interpretation in Bionano, VCF files can be generated, containing the SVs of interest. This feature can be particularly useful for analysing large cohorts or when comparing with VCFs from other technologies. However, the exported VCF from Bionano seems to contain highly conservative SV calls, where the start and the end of the event were estimated to be much larger than predicted within the Bionano Access platform. While this was manageable during my research as each Bionano call was examined individually during the comparison of the three technologies, future work in large cohorts may encounter challenges due to this discrepancy.

Overall, Bionano OGM stands out as a promising technology for clinical SV detection. Nevertheless, emerging technologies, such as EGM (Electronic Genome Mapping) by Nabsys, present compelling alternatives. EGM, similar to OGM, also preserves HMW (high molecular weight) DNA molecules that are labelled for detection. However, instead of using fluorescent labels and optical detection as in OGM, EGM adds tags to the DNA to influence the electrical signal (**Figure 84**) when the tagged DNA passes through single molecule detectors.[206] Nabsys EGM advertises the capability to

generate molecules > 100 kb, similar to that of the Bionano OGM, yet with a resolution down to 300 bp, slightly improved compared to Bionano OGM's 500 bp. This potential improvement in resolution may be due a higher tag density in EGM. The cost of EGM is advertised at a slightly lower entry level, while additional costs associated with reagents, detectors (chips), and service charges remain unknown, albeit suspected to be in a comparable range to Bionano OGM. Compared to OGM, EGM is still less well known, particularly in the clinical setting. Nonetheless, EGM has been involved in various projects, notably contributing to the most recent effort in generating a reference set of germline large DELs and INSs.[202] Overall, Nabsys EGM holds significant potential as an excellent alternative modern cytogenetic technology, offering promising features that might make it a strong contender in the field.

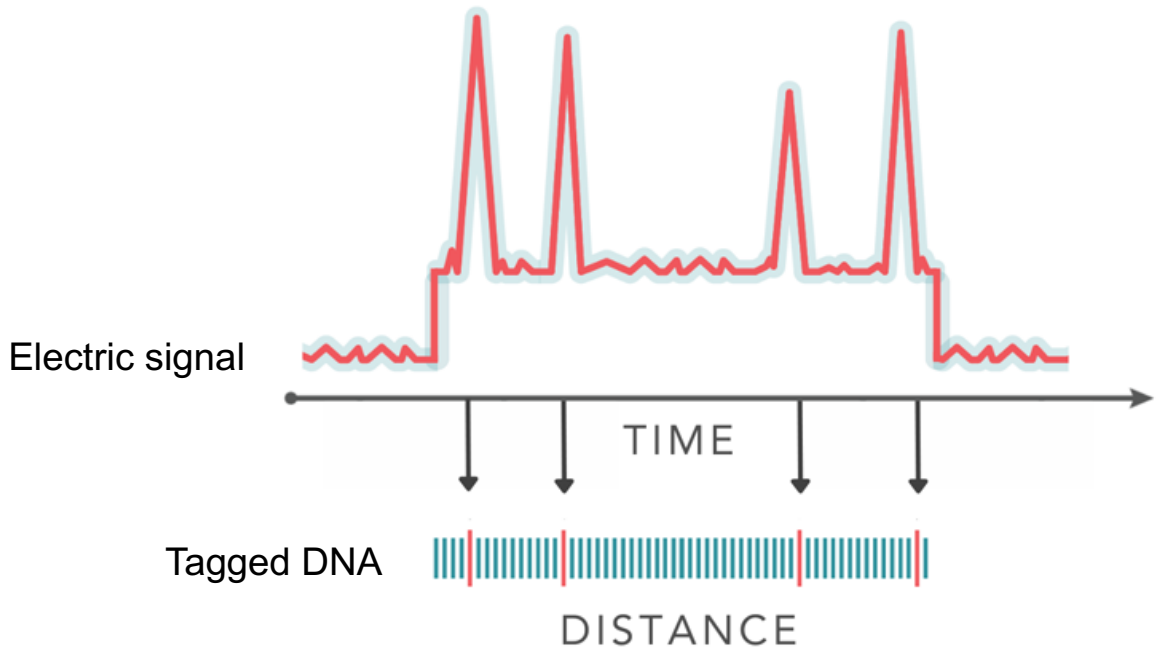


Figure 84 Nabsys EGM reads the electrical signal caused by tags on DNA molecules to generate a consensus map to compare to the reference map. *Figure adapted from https://nabsys.com/informatics/ (last accessed 29.11.2023).*

7.4 Future work

Throughout this 4-year project, despite interruptions cause by the global pandemic, the overall theme of my thesis has surprisingly maintained the general direction envisioned at the start of my research. Nevertheless, certain adjustments were made, which opened up excellent opportunities for future work. Here I have provided a list of potential directions based on their likelihood of yielding productive outcome and practical feasibility:

Work in progress

1. Functional studies on case 21 with the small *HOXC* DUP and craniofacial abnormalities using patient derived iPSCs and a mouse model are being undertaken. Similarly, functional studies for case 18 (large *HOXC* CPX DUP) are also underway with iPSC and mouse model approaches, although this will be significantly more challenging to achieve due to the size and the complex nature of the event. These are ongoing experiments led by Dr. Dagmara Korona.

2. Additional functional analysis is continuing for the case 10 SV affecting the *FGF9* locus. These analyses include RNA-seq in differentiated iPSCs to examine the effect of the SV at different developmental stages and capture-C on iPSCs to examine the effect on genomic 3D conformation. These studies are led by Dr. Dagmara Korona.

Sample dependent works

1. Verification of segregation of the large *HOXC* DUP in the extended family of case 18. A PCR test has been designed (described in **section 4.5**) that can be performed using routine genomic DNA samples (including saliva). This work, however, is dependent on sample availability from the additional affected individuals of the extended family. To date, no further affected family members or intermediate relatives have volunteered to provide a sample for testing.

2. Full characterisation of the case 16 CPX SV on chr20 (**section 5.6**) with FISH. This work is essential to further explore the hypothesis of the occurrence of cryptic INVs across generations and the potential pathogenicity of this chr20 CPX SV. I have already prepared two sets of FISH probes, but additional fresh blood samples would be required from both the proband and the mother of case 16.

Potential additional work

1. Literature review of SVs at *ABCAs/KCNJ2/16* locus related to hypertrichosis and gingival hyperplasia. As discussed in **section 5.5.4,** there is likely an incorrectly characterised genotype-phenotype correlation due to the potential effect of DELs on mis-expression of adjacent *KCNJ* genes, which appears to have been overlooked in the literature. Collaborators at GOSH may have already identified an additional case relevant to this hypothesis.

2. Analysis of the recalled data from ONT-Sniffles2, in combination with work already done in **section 6.5**, may identify additional clinically relevant SVs. Break junction PCR is required to further characterise and understand the candidate *de novo* coding SVs identified in **section 6.5**.

3. Additional annotation may be included to further analyse the 100kGP data. For example, conservation score annotation was a feature left out from both my 100kGP analysis and the SVRare tool at the time. Based on the ongoing work carried out by Dr Eduardo Calpena, some highly clinically relevant SVs can be identified by looking for SVs affecting highly conversed regions systematically. Another annotation feature would be to include CTCF sites, looking for SVs affecting regions with high density of CTCF sites. This is to look for SVs potentially exerting a long-range effect without affecting known genes or regulatory elements.

7.5 Conclusion

This work investigated clinically relevant SVs in patients with CRS, using both short-read and long- range technologies. While some SVs led to definitive diagnoses in several cases, other VUSs, particularly large CPX SVs, offered great research interest in understanding the mechanisms of SVs and their implications for gene function and regulation. This work further evaluated three different SV detection technologies, Illumina WGS, ONT WGS, and Bionano OGM, regarding their performance in SV detection and analysis.

While the technologies to identify and characterise SVs are constantly improving and becoming more automatable, the functional interpretation of these SVs is expected to remain highly challenging, as each *de novo* SV is often a unique event. In this thesis, I have illustrated the application of computational prediction methods for functional analysis, especially through TADs. However, aggregation of abnormal genotypes, along with animal models and in vitro studies, are likely to continue to be required to investigate causality for the foreseeable future.

In conclusion, this thesis set the groundwork for analysing SVs in a clinical setting, focusing on rare constitutional diseases such as CRS. The findings highlight the potential of integrating multiple genomic technologies to enhance the identification and characterisation of SVs, hence improving diagnostic capability and deeper understanding of complex genomic alternation in human diseases. Methods established in this research will also serve as a cornerstone for future advances in the field of clinical genetics and contribute new insights for individuals affected by genetic conditions associated with SVs.

References

1. Ataman AD, Vatanoğlu-Lutz EE, Yıldırım G. Medicine in stamps: history of Down syndrome through philately. *J Turk Ger Gynecol Assoc*. 2012;13(4):267-9. doi:10.5152/jtgga.2012.43

2. Lejeune J, Turpin R, Gautier M. [Chromosomic diagnosis of mongolism]. *Arch Fr Pediatr*. 1959;16:962-3.

3. Clegg JB, Weatherall DJ, Contopolou-Griva I, Caroutsos K, Poungouras P, Tsevrenis H. Haemoglobin Icaria, a new chain-termination mutant which causes α thalassaemia. *Nature*. 1974/09/01 1974;251(5472):245-247. doi:10.1038/251245a0

4. Lander ES, Int Human Genome Sequencing C, Linton LM, et al. Initial sequencing and analysis of the human genome. *Nature*. Feb 2001;409(6822):860-921. doi:10.1038/35057062

5. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. Article. *Nature Genetics*. Jan 2010;42(1):30-U41. doi:10.1038/ng.499

6. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nature Reviews Genetics*. Feb 2006;7(2):85-97. doi:10.1038/nrg1767

7. Ho SVS, Urban AE, Mills RE. Structural variation in the sequencing era. *Nature Reviews Genetics*. Mar 2020;21(3):171-189. doi:10.1038/s41576-019-0180-9

8. Burssed B, Zamariolli M, Bellucco FT, Melaragno MI. Mechanisms of structural chromosomal rearrangement formation. *Molecular Cytogenetics*. Jun 2022;15(1)23. doi:10.1186/s13039-022-00600-6

9. Murakami H, Keeney S. Regulating the formation of DNA double-strand breaks in meiosis. *Genes & Development*. Feb 2008;22(3):286-292. doi:10.1101/gad.1642308

10. Reiter LT, Hastings PJ, Nelis E, De Jonghe P, Van Broeckhoven C, Lupski JR. Human meiotic recombination products revealed by sequencing a hotspot for homologous strand exchange in multiple HNPP deletion patients. *American Journal of Human Genetics*. May 1998;62(5):1023-1033. doi:10.1086/301827

11. Verges L, Molina O, Gean E, Vidal F, Blanco J. Deletions and duplications of the 22q11.2 region in spermatozoa from DiGeorge/velocardiofacial fathers. *Molecular Cytogenetics*. Nov 2014;786. doi:10.1186/s13039-014-0086-3

12. Lieber MR. The mechanism of human nonhomologous DNA end joining. *Journal of Biological Chemistry*. Jan 2008;283(1):1-5. doi:10.1074/jbc.R700039200

13. Pawlowska E, Blasiak J. DNA Repair-A Double-Edged Sword in the Genomic Stability of Cancer Cells-The Case of Chronic Myeloid Leukemia. *International Journal of Molecular Sciences*. Nov 2015;16(11):27535-27549. doi:10.3390/ijms161126049

14. Gollin SM. Mechanisms leading to nonrandom, nonhomologous chromosomal translocations in leukemia. *Seminars in Cancer Biology*. Feb 2007;17(1):74-79. doi:10.1016/j.semcancer.2006.10.002

15. Gelot C, Magdalou I, Lopez BS. Replication Stress in Mammalian Cells and Its Consequences for Mitosis. *Genes*. Jun 2015;6(2):267-298. doi:10.3390/genes6020267

16. van Poppelen NM, Yavuzyigitoglu S, Smit KN, et al. Chromosomal rearrangements in uveal melanoma: Chromothripsis. *Genes Chromosomes & Cancer*. Sep 2018;57(9):452-458. doi:10.1002/gcc.4

17. Caspersson T, Farber S, Foley GE, et al. CHEMICAL DIFFERENTIATION ALONG METAPHASE CHROMOSOMES. *Experimental Cell Research*. 1968;49(1):219-+. doi:10.1016/0014-4827(68)90538-7

18. Seabrigh.M. RAPID BANDING TECHNIQUE FOR HUMAN CHROMOSOMES. *Lancet*. 1971;2(7731):971-&.

19. Langersafer PR, Levine M, Ward DC. IMMUNOLOGICAL METHOD FOR MAPPING GENES ON DROSOPHILA POLYTENE CHROMOSOMES. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*. 1982;79(14):4381-4385. doi:10.1073/pnas.79.14.4381

20. Kallioniemi A, Kallioniemi OP, Sudar D, et al. COMPARATIVE GENOMIC HYBRIDIZATION FOR MOLECULAR CYTOGENETIC ANALYSIS OF SOLID TUMORS. *Science*. Oct 1992;258(5083):818-821. doi:10.1126/science.1359641

21. SolinasToldo S, Lampel S, Stilgenbauer S, et al. Matrix-based comparative genomic hybridization: Biochips to screen for genomic imbalances. *Genes Chromosomes & Cancer*. Dec 1997;20(4):399-407. doi:10.1002/(sici)1098-2264(199712)20:4<399::aid-gcc12>3.0.co;2-i

22. Armour JAL, Sismani C, Patsalis PC, Cross G. Measurement of locus copy number by hybridisation with amplifiable probes. *Nucleic Acids Research*. Jan 2000;28(2):605-609. doi:10.1093/nar/28.2.605

23. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Research*. Jun 2002;30(12)e57. doi:10.1093/nar/gnf056

24. Margulies M, Egholm M, Altman WE, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. Sep 2005;437(7057):376-380. doi:10.1038/nature03959

25. Eid J, Fehr A, Gray J, et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*. Jan 2009;323(5910):133-138. doi:10.1126/science.1162986

26. Vranken C, Deen J, Dirix L, et al. Super-resolution optical DNA Mapping via DNA methyltransferase-directed click chemistry. *Nucleic Acids Research*. Apr 2014;42(7)e50. doi:10.1093/nar/gkt1406

27. Pös O, Radvanszky J, Styk J, et al. Copy Number Variation: Methods and Clinical Applications. *Applied Sciences-Basel*. Jan 2021;11(2)819. doi:10.3390/app11020819

28. Chen P, Sun ZP, Wang JW, et al. Portable nanopore-sequencing technology: Trends in development and applications. *Frontiers in Microbiology*. Feb 2023;141043967. doi:10.3389/fmicb.2023.1043967

29. Schena M, Shalon D, Davis RW, Brown PO. QUANTITATIVE MONITORING OF GENE-EXPRESSION PATTERNS WITH A COMPLEMENTARY-DNA MICROARRAY. *Science*. Oct 1995;270(5235):467-470. doi:10.1126/science.270.5235.467

30. McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genetics*. Oct 2008;40(10):1166-1174. doi:10.1038/ng.238

31. Biesecker LG, Spinner NB. A genomic view of mosaicism and human disease. *Nature Reviews Genetics*. May 2013;14(5):307-320. doi:10.1038/nrg3424

32. Miller DT, Adam MP, Aradhya S, et al. Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. *American Journal of Human Genetics*. May 2010;86(5):749-764. doi:10.1016/j.ajhg.2010.04.006

33. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research*. Jun 2011;21(6):974-984. doi:10.1101/gr.114876.110

34. Roller E, Ivakhno S, Lee S, Royce T, Tanner S. Canvas: versatile and scalable detection of copy number variants. *Bioinformatics*. Aug 2016;32(15):2375-2377. doi:10.1093/bioinformatics/btw163

35. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. *Genome Biology*. 2014;15(6)R84. doi:10.1186/gb-2014-15-6-r84

36. Chen XY, Schulz-Trieglaff O, Shaw R, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. Apr 2016;32(8):1220-1222. doi:10.1093/bioinformatics/btv710

37. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nature Reviews Genetics*. Oct 2020;21(10):597-614. doi:10.1038/s41576-020-0236-x

38. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics Proteomics & Bioinformatics*. Oct 2015;13(5):278-289. doi:10.1016/j.gpb.2015.08.002

39. Amarasinghe SL, Su S, Dong XY, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*. Feb 2020;21(1)30. doi:10.1186/s13059-020-1935-5

40. Lam ET, Hastie A, Lin C, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology*. Aug 2012;30(8):771-776. doi:10.1038/nbt.2303

41. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Research*. Jan 2014;42(D1):D986-D992. doi:10.1093/nar/gkt958

42. Collins RL, Brand H, Karczewski KJ, et al. A structural variation reference for medical and population genetics. *Nature*. May 2020;581(7809):444-451. doi:10.1038/s41586-020-2287-8

43. Audano PA, Sulovari A, Graves-Lindsay TA, et al. Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell*. Jan 2019;176(3):663-+. doi:10.1016/j.cell.2018.12.019

44. Duclos A, Charbonnier F, Chambon P, et al. Pitfalls in the Use of DGV for CNV Interpretation. *American Journal of Medical Genetics Part A*. Oct 2011;155A(10):2593-2596. doi:10.1002/ajmg.a.34195

45. Firth HV, Richards SM, Bevan AP, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. Article. *American Journal of Human Genetics*. Apr 2009;84(4):524-533. doi:10.1016/j.ajhg.2009.03.010

46. Fuller ZL, Berg JJ, Mostafavi H, Sella G, Przeworski M. Measuring intolerance to mutation in human genetics. *Nature Genetics*. May 2019;51(5):772-+. doi:10.1038/s41588-019-0383-1

47. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Article. *Nature*. Aug 2016;536(7616):285-+. doi:10.1038/nature19057

48. Valentijn LJ, Bolhuis PA, Zorn I, et al. THE PERIPHERAL MYELIN GENE PMP-22/GAS-3 IS DUPLICATED IN CHARCOT-MARIE-TOOTH DISEASE TYPE-1A. *Nature Genetics*. Jun 1992;1(3):166-170. doi:10.1038/ng0692-166

49. Patel PI, Roa BB, Welcher AA, et al. THE GENE FOR THE PERIPHERAL MYELIN PROTEIN-PMP-22 IS A CANDIDATE FOR CHARCOT-MARIE-TOOTH DISEASE TYPE-1A. *Nature Genetics*. Jun 1992;1(3):159-165. doi:10.1038/ng0692-159

50. Pentao L, Wise CA, Chinault AC, Patel PI, Lupski JR. CHARCOT-MARIE-TOOTH TYPE-1A DUPLICATION APPEARS TO ARISE FROM RECOMBINATION AT REPEAT SEQUENCES FLANKING THE 1.5 MB MONOMER UNIT. *Nature Genetics*. Dec 1992;2(4):292-300. doi:10.1038/ng1292-292

51. Li J, Parker B, Martyn C, Natarajan C, Guo JS. The PMP22 Gene and Its Related Diseases. *Molecular Neurobiology*. Apr 2013;47(2):673-698. doi:10.1007/s12035-012-8370-x

52. Nora EP, Lajoie BR, Schulz EG, et al. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*. May 2012;485(7398):381-385. doi:10.1038/nature11049

53. McCord RP, Kaplan N, Giorgetti L. Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function. *Molecular Cell*. Feb 2020;77(4):688-708. doi:10.1016/j.molcel.2019.12.021

54. Han JL, Zhang ZL, Wang K. 3C and 3C-based techniques: the powerful tools for spatial genome organization deciphering. *Molecular Cytogenetics*. Mar 2018;1121. doi:10.1186/s13039-018-0368-2

55. Szalaj P, Plewczynski D. Three-dimensional organization and dynamics of the genome. *Cell Biology and Toxicology*. Oct 2018;34(5):381-404. doi:10.1007/s10565-018-9428-y

56. Li GL, Cai LY, Chang HD, et al. Chromatin Interaction Analysis with Paired-End Tag (ChIA-PET) sequencing technology and application. *Bmc Genomics*. Dec 2014;15S11. doi:10.1186/1471-2164-15-s12-s11

57. Hansen AS. CTCF as a boundary factor for cohesin-mediated loop extrusion: evidence for a multi-step mechanism. *Nucleus*. Jan 2020;11(1):132-148. doi:10.1080/19491034.2020.1782024

58. Yu WB, He B, Tan K. Identifying topologically associating domains and subdomains by Gaussian Mixture model And Proportion test. *Nature Communications*. Sep 2017;8535. doi:10.1038/s41467-017-00478-8

59. Spielmann M, Lupianez DG, Mundlos S. Structural variation in the 3D genome. *Nature Reviews Genetics*. Jul 2018;19(7):453-467. doi:10.1038/s41576-018-0007-0

60. Gordon CT, Tan TY, Benko S, FitzPatrick D, Lyonnet S, Farlie PG. Long-range regulation at the SOX9 locus in development and disease. *Journal of Medical Genetics*. Oct 2009;46(10):649-656. doi:10.1136/jmg.2009.068361

61. Franke M, Ibrahim DM, Andrey G, et al. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*. Oct 2016;538(7624):265-+. doi:10.1038/nature19800

62. Despang A, Schopflin R, Franke M, et al. Functional dissection of the Sox9-Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nature Genetics*. Aug 2019;51(8):1263-+. doi:10.1038/s41588-019-0466-z

63. Cornelissen M, den Ottelander B, Rizopoulos D, et al. Increase of prevalence of craniosynostosis. *Journal of Cranio-Maxillofacial Surgery*. Sep 2016;44(9):1273-1279. doi:10.1016/j.jcms.2016.07.007

64. Bomer-Skogstad J, Dremmen M, Mathijssen I, Smithuis R. Craniosynostosis. Accessed 30.10.2023, https://radiologyassistant.nl/pediatrics/hip/craniosynosthosis

65. Richardson S, Browne ML, Rasmussen SA, et al. Associations Between Periconceptional Alcohol Consumption and Craniosynostosis, Omphalocele, and Gastroschisis. *Birth Defects Research Part a-Clinical and Molecular Teratology*. Jul 2011;91(7):623-630. doi:10.1002/bdra.20823

66. Zeiger JS, Beaty TH, Hetmanski JB, et al. Genetic and environmental risk factors for sagittal craniosynostosis. *Journal of Craniofacial Surgery*. Sep 2002;13(5):602-606. doi:10.1097/00001665-200209000-00002

67. Carmichael SL, Ma C, Rasmussen SA, et al. Craniosynostosis and maternal smoking. *Birth Defects Research Part a-Clinical and Molecular Teratology*. Feb 2008;82(2):78-85. doi:10.1002/bdra.20426

68. Hoyt AT, Canfield MA, Romitti PA, et al. Associations between maternal periconceptional exposure to secondhand tobacco smoke and major birth defects. *American Journal of Obstetrics and Gynecology*. Nov 2016;215(5)613.e1-11. doi:10.1016/j.ajog.2016.07.022

69. Hackshaw A, Rodeck C, Boniface S. Maternal smoking in pregnancy and birth defects: a systematic review based on 173 687 malformed cases and 11.7 million controls. *Human Reproduction Update*. Sep-Oct 2011;17(5):589-604. doi:10.1093/humupd/dmr022

70. Alwan S, Reefhuis J, Rasmussen SA, Olney RS, Friedman JM, Natl Birth Defects Prevention S. Use of selective serotonin-reuptake inhibitors in pregnancy and the risk of birth defects. *New England Journal of Medicine*. Jun 2007;356(26):2684-2692. doi:10.1056/NEJMoa066584

71. Reefhuis J, Devine O, Friedman JM, Louik C, Honein MA. Specific SSRIs and birth defects: bayesian analysis to interpret new data in the context of previous reports. *Bmj-British Medical Journal*. Jul 2015;351h3190. doi:10.1136/bmj.h3190

72. Lajeunie E, Barcik U, Thorne JA, El Ghouzzi V, Bourgeois M, Renier D. Craniosynostosis and fetal exposure to sodium valproate. *Journal of Neurosurgery*. Nov 2001;95(5):778-782. doi:10.3171/jns.2001.95.5.0778

73. Tooze RS, Calpena E, Weber A, Wilson LC, Twigg SRF, Wilkie AOM. Review of Recurrently Mutated Genes in Craniosynostosis Supports Expansion of Diagnostic Gene Panels. *Genes*. Mar 2023;14(3)615. doi:10.3390/genes14030615

74. Tischfield MA, Robson CD, Gilette NM, et al. Cerebral Vein Malformations Result from Loss of Twist1 Expression and BMP Signaling from Skull Progenitor Cells and Dura. *Developmental Cell*. Sep 2017;42(5):445-+. doi:10.1016/j.devcel.2017.07.027

75. Goddard LM, Kahn ML. A BMPy Road for Venous Development. *Developmental Cell*. Sep 2017;42(5):435-436. doi:10.1016/j.devcel.2017.08.016

76. Motegi T, Ohuchi M, Ohtaki C, et al. A CRANIOSYNOSTOSIS IN A BOY WITH A DEL(7)(P15.3P21.3) - ASSIGNMENT BY DELETION MAPPING OF THE CRITICAL SEGMENT FOR CRANIOSYNOSTOSIS TO THE MID-PORTION OF 7P21. *Human Genetics*. 1985;71(2):160-162.

77. Wang C, Maynard S, Glover TW, Biesecker LG. MILD PHENOTYPIC MANIFESTATION OF A 7P15.3P21.2 DELETION. *Journal of Medical Genetics*. Jul 1993;30(7):610-612. doi:10.1136/jmg.30.7.610

78. Kress W, Schropp C, Lieb G, et al. Saethre-Chotzen syndrome caused by TWIST 1 gene mutations: functional differentiation from Muenke coronal synostosis syndrome. *European Journal of Human Genetics*. Jan 2006;14(1):39-48. doi:10.1038/sj.ejhg.5201507

79. Schluth-Bolard C, Till M, Labalme A, et al. TWIST microdeletion identified by array CGH in a patient presenting Saethre-Chotzen phenotype and a complex

rearrangement involving chromosomes 2 and 7. *European Journal of Medical Genetics*. Mar-Apr 2008;51(2):156-164. doi:10.1016/j.ejmg.2007.12.003

80. Wilkie AOM, Johnson D, Wall SA. Clinical genetics of craniosynostosis. *Current Opinion in Pediatrics*. Dec 2017;29(6):622-628. doi:10.1097/mop.0000000000000542

81. Teven CM, Farina EM, Rivas J, Reid RR. Fibroblast growth factor (FGF) signaling in development and skeletal diseases. *Genes Dis*. Dec 1 2014;1(2):199-213. doi:10.1016/j.gendis.2014.09.005

82. Iseki S, Wilkie AOM, Morriss-Kay GM. *Fgfr1* and *Fgfr2* have distinct differentiation- and proliferation-related roles in the developing mouse skull vault. *Development*. Dec 1999;126(24):5611-5620.

83. Morriss-Kay GM, Iseki S, Johnson D. Genetic control of the cell proliferation-differentiation balance in the developing skull vault: roles of fibroblast growth factor receptor signalling pathways. *Novartis Found Symp*. 2001;232:102-16; discussion 116-21. doi:10.1002/0470846658.ch8

84. Zhao XL, Erhardt S, Sung KH, Wang J. FGF signaling in cranial suture development and related diseases. *Frontiers in Cell and Developmental Biology*. Jun 2023;111112890. doi:10.3389/fcell.2023.1112890

85. Grillo L, Greco D, Pettinato R, et al. Increased *FGF3* and *FGF4* gene dosage is a risk factor for craniosynostosis. *Gene*. Jan 2014;534(2):435-439. doi:10.1016/j.gene.2013.09.120

86. Jehee FS, Bertola DR, Yelavarthi KK, et al. An 11q11-q13-3 duplication, including *FGF3* and *FGF4* genes, in a patient with syndromic multiple craniosynostoses. *American Journal of Medical Genetics Part A*. Aug 2007;143A(16):1912-1918. doi:10.1002/ajmg.a.31863

87. Zhao H, Chai Y. Stem Cells in Teeth and Craniofacial Bones. *Journal of Dental Research*. Nov 2015;94(11):1495-1501. doi:10.1177/0022034515603972

88. Ittleman BR, McKissick J, Bosanko KA, Ocal E, Golinko M, Zarate YA. Less common underlying genetic diagnoses found in a cohort of 139 individuals surgically corrected for craniosynostosis. *American Journal of Medical Genetics Part A*. Feb 2018;176(2):487-491. doi:10.1002/ajmg.a.38532

89. Twigg SRF, Wilkie AOM. A Genetic-Pathophysiological Framework for Craniosynostosis. *American Journal of Human Genetics*. Sep 2015;97(3):359-377. doi:10.1016/j.ajhg.2015.07.006

90. Hyder Z, Calpena E, Pei Y, et al. Evaluating the performance of a clinical genome sequencing program for diagnosis of rare genetic disease, seen through the lens of craniosynostosis. *Genetics in Medicine*. Dec 2021;23(12):2360-2368. doi:10.1038/s41436-021-01297-5

91. Miller KA, Twigg SRF, McGowan SJ, et al. Diagnostic value of exome and whole genome sequencing in craniosynostosis. *Journal of Medical Genetics*. Apr 2017;54(4):260-268. doi:10.1136/jmedgenet-2016-104215

92. Laver TW, Franco ED, Johnson MB, et al. SavvyCNV: genome-wide CNV calling from off-target reads. *bioRxiv*. 2019:617605. doi:10.1101/617605

93. Krietenstein N, Abraham S, Venev SV, et al. Ultrastructural Details of Mammalian Chromosome Architecture. *Molecular Cell*. May 2020;78(3):554-+. doi:10.1016/j.molcel.2020.03.003

94. Yu J, Szabo A, Pagnamenta AT, et al. SVRare: discovering disease-causing structural variants in the 100K Genomes Project. *medRxiv*. 2021:2021.10.15.21265069. doi:10.1101/2021.10.15.21265069

95. Kent WJ. BLAT - The BLAST-like alignment tool. *Genome Research*. Apr 2002;12(4):656-664. doi:10.1101/gr.229202

96. Tassano E, Mirabelli-Badenier M, Veneselli E, et al. Clinical and molecular characterization of a patient with interstitial 6q21q22.1 deletion. *Molecular Cytogenetics*. Apr 2015;831. doi:10.1186/s13039-015-0134-7

97. Capurro MI, Xu P, Shi W, Li FC, Jia A, Filmus J. Glypican-3 inhibits Hedgehog signaling during development by competing with patched for Hedgehog binding. *Developmental Cell*. May 2008;14(5):700-711. doi:10.1016/j.devcel.2008.03.006

98. Capurro MI, Xiang YY, Lobe C, Filmus J. Glypican-3 promotes the growth of hepatocellular carcinoma by stimulating canonical Wnt signaling. *Cancer Research*. Jul 2005;65(14):6245-6254. doi:10.1158/0008-5472.can-04-4244

99. Veugelers M, De Cat B, Muyldermans SY, et al. Mutational analysis of the GPC3/GPC4 glypican gene cluster on Xq26 in patients with Simpson-Golabi-Behmel syndrome: identification of loss-of-function mutations in the GPC3 gene. *Human Molecular Genetics*. May 2000;9(9):1321-1328. doi:10.1093/hmg/9.9.1321

100. DiMaio MS, Yang H, Mahoney MJ, McGrath J, Li PN. Familial GPC3 and GPC4-TFDP3 deletions at Xq26 associated with Simpson-Golabi-Behmel syndrome. *Meta Gene*. Feb 2017;11:147-151. doi:10.1016/j.mgene.2016.08.008

101. Poeta L, Malacarne M, Padula A, et al. Further Delineation of Duplications of ARX Locus Detected in Male Patients with Varying Degrees of Intellectual Disability. *International Journal of Molecular Sciences*. Mar 2022;23(6)3084. doi:10.3390/ijms23063084

102. den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Human Mutation*. Jun 2016;37(6):564-569. doi:10.1002/humu.22981

103. Mizutani Y, Kihara A, Igarashi Y. Mammalian Lass6 and its related family members regulate synthesis of specific ceramides. *Biochemical Journal*. Aug 2005;390:263-271. doi:10.1042/bj20050291

104. Ebel P, Dorp KV, Petrasch-Parwez E, et al. Inactivation of Ceramide Synthase 6 in Mice Results in an Altered Sphingolipid Metabolism and Behavioral Abnormalities. *Journal of Biological Chemistry*. Jul 2013;288(29):21433-21447. doi:10.1074/jbc.M113.479907

105. Ullah I, Kakar N, Schrauwen I, et al. Variants in KIAA0825 underlie autosomal recessive postaxial polydactyly. *Human Genetics*. Jun 2019;138(6):593-600. doi:10.1007/s00439-019-02000-0

106. Di Donato N, Neuhann T, Kahlert AK, et al. Mutations in EXOSC2 are associated with a novel syndrome characterised by retinitis pigmentosa, progressive hearing loss, premature ageing, short stature, mild intellectual disability and distinctive gestalt. *Journal of Medical Genetics*. Jun 2016;53(6):419-425. doi:10.1136/jmedgenet-2015-103511

107. Chen YC, Auer-Grumbach M, Matsukawa S, et al. Transcriptional regulator PRDM12 is essential for human pain perception (vol 47, pg 803, 2015). *Nature Genetics*. Aug 2015;47(8):962-962. doi:10.1038/ng0815-962b

108. Chen CA, Crutcher E, Gill H, et al. The expanding clinical phenotype of germline ABL1-associated congenital heart defects and skeletal malformations syndrome. *Human Mutation*. Oct 2020;41(10):1738-1744. doi:10.1002/humu.24075

109. McGinnis W, Krumlauf R. HOMEOBOX GENES AND AXIAL PATTERNING. *Cell*. Jan 1992;68(2):283-302. doi:10.1016/0092-8674(92)90471-n

110. Hueber SD, Weiller GF, Djordjevic MA, Frickey T. Improving Hox Protein Classification across the Major Model Organisms. *Plos One*. May 2010;5(5)e10820. doi:10.1371/journal.pone.0010820

111. Lemons D, McGinnis W. Genomic evolution of Hox gene clusters. *Science*. Sep 2006;313(5795):1918-1922. doi:10.1126/science.1132040

112. Carroll SB. HOMEOTIC GENES AND THE EVOLUTION OF ARTHROPODS AND CHORDATES. *Nature*. Aug 1995;376(6540):479-485. doi:10.1038/376479a0

113. Simakov O, Marletaz F, Yue JX, et al. Deeply conserved synteny resolves early events in vertebrate evolution. *Nature Ecology & Evolution*. Jun 2020;4(6):820-+. doi:10.1038/s41559-020-1156-z

114. Papageorgiou S. Hox Gene Collinearity: From A-P Patterning to Radially Symmetric Ani-mals. *Current Genomics*. 2016;17(5):444-449. doi:10.2174/1389202917666160616082436

115. Wagner GP, Amemiya C, Ruddle F. Hox cluster duplications and the opportunity for evolutionary novelties. *Proceedings of the National Academy of Sciences of the United States of America*. Dec 2003;100(25):14603-14606. doi:10.1073/pnas.2536656100

116. Lin ZM, Chen Q, Shi L, et al. Loss-of-Function Mutations in HOXC13 Cause Pure Hair and Nail Ectodermal Dysplasia. *American Journal of Human Genetics*. Nov 2012;91(5):906-911. doi:10.1016/j.ajhg.2012.08.029

117. Hancarova M, Simandlova M, Drabova J, et al. Chromosome 12q13.13 deletions involving the HOXC gene cluster: Phenotype and candidate genes. *European Journal of Medical Genetics*. Mar 2013;56(3):171-173. doi:10.1016/j.ejmg.2012.12.003

118. Alvarado DM, McCall K, Hecht JT, Dobbs MB, Gurnett CA. Deletions of 5 ' HOXC genes are associated with lower extremity malformations, including clubfoot and vertical talus. *Journal of Medical Genetics*. Apr 2016;53(4):250-255. doi:10.1136/jmedgenet-2015-103505

119. Quinonez SC, Innis JW. Human HOX gene disorders. *Molecular Genetics and Metabolism*. Jan 2014;111(1):4-15. doi:10.1016/j.ymgme.2013.10.012

120. Mark M, Rijli FM, Chambon P. Homeobox genes in embryogenesis and pathogenesis. *Pediatric Research*. Oct 1997;42(4):421-429. doi:10.1203/00006450-199710000-00001

121. Lappin TR, Grier DG, Thompson A, Halliday HL. HOX genes: seductive science, mysterious mechanisms. *Ulster Med J*. Jan 2006;75(1):23-31.

122. Afzal Z, Krumlauf R. Transcriptional Regulation and Implications for Controlling Hox Gene Expression. *Journal of Developmental Biology*. Mar 2022;10(1)4. doi:10.3390/jdb10010004

123. Zurek B, Ellwanger K, Vissers L, et al. Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. *European Journal of Human Genetics*. Sep 2021;29(9):1325-1331. doi:10.1038/s41431-021-00859-0

124. Yasukawa R, Moteki H, Nishio S, et al. The Prevalence and Clinical Characteristics of TECTA-Associated Autosomal Dominant Hearing Loss. *Genes*. Oct 2019;10(10)744. doi:10.3390/genes10100744

125. Jonsson DI, Ludvigsson P, Aradhya S, et al. A de novo 1.13 Mb microdeletion in 12q13.13 associated with congenital distal arthrogryposis, intellectual disability and mild dysmorphism. *European Journal of Medical Genetics*. Jun-Jul 2012;55(6-7):437-440. doi:10.1016/j.ejmg.2012.03.001

126. Blackburn J, Rich M, Ghitani N, Liu JP. Generation of Conditional Hoxc8 Loss-of-Function and Hoxc8 -> Hoxc9 Replacement Alleles in Mice. *Genesis*. Oct 2009;47(10):680-687. doi:10.1002/dvg.20547

127. Suemori H, Noguchi S. Hox C cluster genes are dispensable for overall body plan of mouse embryonic development. *Developmental Biology*. Apr 2000;220(2):333-342. doi:10.1006/dbio.2000.9651

128. Schwessinger R, Gosden M, Downes D, et al. DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nature Methods*. doi:10.1038/s41592-020-0960-3

129. Kitazawa T, Minoux M, Ducret S, Rijli FM. Different Ectopic Hoxa2 Expression Levels in Mouse Cranial Neural Crest Cells Result in Distinct Craniofacial Anomalies and Homeotic Phenotypes. *Journal of Developmental Biology*. Mar 2022;10(1)9. doi:10.3390/jdb10010009

130. Mentzer SE, Sundberg JP, Awgulewitsch A, et al. The mouse hairy ears mutation exhibits an extended growth (anagen) phase in hair follicles and altered Hoxc gene expression in the ears. *Veterinary Dermatology*. Dec 2008;19(6):358-367. doi:10.1111/j.1365-3164.2008.00709.x

131. Liu ZW, Chu JY, Li P, Zhao QQ, Li SJ, Mou CY. HOXC10 intronic duplication is associated with unsealed skull and crest in crested chicken with cerebral hernia. *Gene*. Oct 2022;840146758. doi:10.1016/j.gene.2022.146758

132. Tsai DY, Chen JJ, Su PC, et al. Chicken HOXC8 and HOXC10 genes may play a role in the altered skull morphology associated with the Crest phenotype. *Journal of Experimental Zoology Part B-Molecular and Developmental Evolution*. 2023 Apr 2023;doi:10.1002/jez.b.23194

133. Guo Y, Gu XR, Sheng ZY, et al. A Complex Structural Variation on Chromosome 27 Leads to the Ectopic Expression of HOXB8 and the Muffs and Beard Phenotype in Chickens. *Plos Genetics*. Jun 2016;12(6)e1006071. doi:10.1371/journal.pgen.1006071

134. Roux M, Zaffran S. Hox Genes in Cardiovascular Development and Diseases. *Journal of Developmental Biology*. Jun 2016;4(2)14. doi:10.3390/jdb4020014

135. Lufkin T, Dierich A, Lemeur M, Mark M, Chambon P. DISRUPTION OF THE HOX-1.6 HOMEOBOX GENE RESULTS IN DEFECTS IN A REGION CORRESPONDING TO ITS ROSTRAL DOMAIN OF EXPRESSION. *Cell*. Sep 1991;66(6):1105-1119. doi:10.1016/0092-8674(91)90034-v

136. Roux M, Laforest B, Capecchi M, Bertrand N, Zaffran S. Hoxb1 regulates proliferation and differentiation of second heart field progenitors in pharyngeal mesoderm and genetically interacts with Hoxa1 during cardiac outflow tract development. *Developmental Biology*. Oct 2015;406(2):247-258. doi:10.1016/j.ydbio.2015.08.015

137. Godwin AR, Stadler HS, Nakamura K, Capecchi MR. Detection of targeted GFP-Hox gene fusions during mouse embryogenesis. *Proceedings of the National Academy of Sciences of the United States of America*. Oct 1998;95(22):13042-13047. doi:10.1073/pnas.95.22.13042

138. Chisaka O, Capecchi MR. REGIONALLY RESTRICTED DEVELOPMENTAL DEFECTS RESULTING FROM TARGETED DISRUPTION OF THE MOUSE HOMEOBOX GENE HOX-1.5. *Nature*. Apr 1991;350(6318):473-479. doi:10.1038/350473a0

139. Chisaka O, Kameda Y. Hoxa3 regulates the proliferation and differentiation of the third pharyngeal arch mesenchyme in mice. *Cell and Tissue Research*. Apr 2005;320(1):77-89. doi:10.1007/s00441-004-1042-z

140. Kameda Y, Watari-Goshima N, Nishimaki T, Chisaka O. Disruption of the Hoxa3 homeobox gene results in anomalies of the carotid artery system and the arterial baroreceptors. *Cell and Tissue Research*. Mar 2003;311(3):343-352. doi:10.1007/s00441-002-0681-1

141. Schumacher JA, Wright ZA, Owen ML, Bredemeier NO, Sumanas S. Integrin alpha 5 and Integrin alpha 4 cooperate to promote endocardial differentiation and heart morphogenesis. *Developmental Biology*. Sep 2020;465(1):46-57. doi:10.1016/j.ydbio.2020.06.006

142. Zhang YJ, Si Y, Ma N, Mei J. The RNA-binding protein PCBP2 inhibits Ang II-induced hypertrophy of cardiomyocytes though promoting GPR56 mRNA degeneration. *Biochemical and Biophysical Research Communications*. Aug 2015;464(3):679-684. doi:10.1016/j.bbrc.2015.06.139

143. Zhang YJ, Si Y, Ma N. Meis1 promotes poly (rC)-binding protein 2 expression and inhibits angiotensin II-induced cardiomyocyte hypertrophy. *Iubmb Life*. Jan 2016;68(1):13-22. doi:10.1002/iub.1456

144. Ding J, Chen JH, Wang YQ, et al. Trbp regulates heart function through microRNA-mediated Sox6 repression. *Nature Genetics*. Jul 2015;47(7):776-+. doi:10.1038/ng.3324

145. Huang XH, Li JL, Li XY, et al. miR-208a in Cardiac Hypertrophy and Remodeling. *Frontiers in Cardiovascular Medicine*. Dec 2021;8773314. doi:10.3389/fcvm.2021.773314

146. Taylor JA, Paliga JT, Wes AM, et al. A Critical Evaluation of Long-Term Aesthetic Outcomes of Fronto-Orbital Advancement and Cranial Vault Remodeling in Nonsyndromic Unicoronal Craniosynostosis. *Plastic and Reconstructive Surgery*. Jan 2015;135(1):220-231. doi:10.1097/prs.0000000000000829

147. Hirsch N, Dahan I, D'Haene E, et al. HDAC9 structural variants disrupting TWIST1 transcriptional regulation lead to craniofacial and limb malformations. *Genome Research*. Jul 2022;32(7):1242-1253. doi:10.1101/gr.276196.121

148. Yoon JG, Hahn HM, Choi S, et al. Molecular Diagnosis of Craniosynostosis Using Targeted Next-Generation Sequencing. *Neurosurgery*. Aug 2020;87(2):294-302. doi:10.1093/neuros/nyz470

149. De Marco P, Raso A, Beri S, et al. A de novo balanced translocation t(7;12)(p21.2;p12.3) in a patient with Saethre-Chotzen-like phenotype downregulates TWIST and an osteoclastic protein-tyrosine phosphatase, PTP-oc. *European Journal of Medical Genetics*. Sep-Oct 2011;54(5):E478-E483. doi:10.1016/j.ejmg.2011.05.007

150. Tavares VLR, Guimaraes-Ramos SL, Zhou Y, et al. New locus underlying auriculocondylar syndrome (ARCND): 430 kb duplication involving *TWIST1* regulatory elements. *Journal of Medical Genetics*. Sep 2022;59(9):895-905. doi:10.1136/jmedgenet-2021-107825

151. Moore AR, Jing Y, Yang P, et al. Use of genome sequencing to hunt for cryptic second-hit variants: analysis of 31 cases recruited to the 100 000 Genomes Project. *Journal of Medical Genetics*. 2023:jmg-2023-109362. doi:10.1136/jmg-2023-109362

152. Rodriguez-Zabala M, Aza-Carmona M, Rivera-Pedroza CI, et al. FGF9 mutation causes craniosynostosis along with multiple synostoses. *Human Mutation*. Nov 2017;38(11):1471-1476. doi:10.1002/humu.23292

153. Wu XL, Gu MM, Huang L, et al. Multiple Synostoses Syndrome Is Due to a Missense Mutation in Exon 2 of FGF9 Gene. *American Journal of Human Genetics*. Jul 2009;85(1):53-63. doi:10.1016/j.ajhg.2009.06.007

154. Eldomery MK, Akdemir ZC, Vogtle FN, et al. MIPEP recessive variants cause a syndrome of left ventricular non-compaction, hypotonia, and infantile death. *Genome Medicine*. Nov 2016;8106. doi:10.1186/s13073-016-0360-6

155. Mills RE, Walter K, Stewart C, et al. Mapping copy number variation by population-scale genome sequencing. *Nature*. Feb 2011;470(7332):59-65. doi:10.1038/nature09708

156. Baets J, Deconinck T, Smets K, et al. Mutations in SACS cause atypical and late-onset forms of ARSACS. *Neurology*. Sep 2010;75(13):1181-1188. doi:10.1212/WNL.0b013e3181f4d86c

157. Fanin M, Hoffman EP, Angelini C, Pegoraro E. Private beta- and gamma-sarcoglycan gene mutations: Evidence of a founder effect in northern Italy. *Human Mutation*. 2000;16(1):13-17. doi:10.1002/1098-1004(200007)16:1<13::aid-humu3>3.0.co;2-v

158. Duncan DR, Kang PB, Rabbat JC, et al. A novel mutation in two families with limb-girdle muscular dystrophy type 2C. *Neurology*. Jul 2006;67(1):167-169. doi:10.1212/01.wnl.0000223600.78363.dd

159. Morgan A, Fisher SE, Scheffer I, Hildebrand M. FOXP2-related speech and language disorders. 2017;

160. MacDermot KD, Bonora E, Sykes N, et al. Identification of FOXP2 truncation as a novel cause of developmental speech and language deficits. *American Journal of Human Genetics*. Jun 2005;76(6):1074-1080. doi:10.1086/430841

161. Tonne E, Due-Tonnessen BJ, Vigeland MD, et al. Whole-exome sequencing in syndromic craniosynostosis increases diagnostic yield and identifies candidate genes in osteogenic signaling pathways. *American Journal of Medical Genetics Part A*. May 2022;188(5):1464-1475. doi:10.1002/ajmg.a.62663

162. Marie PJ, Haÿ E. Cadherins and Wnt signalling: a functional link controlling bone formation. *Bonekey Rep*. Apr 17 2013;2:330. doi:10.1038/bonekey.2013.64

163. Behr B, Longaker MT, Quarto N. Differential activation of canonical Wnt signaling determines cranial sutures fate: A novel mechanism for sagittal suture craniosynostosis. *Developmental Biology*. Aug 2010;344(2):922-940. doi:10.1016/j.ydbio.2010.06.009

164. Thomas PD, Ebert D, Muruganujan A, Mushayahama T, Albou LP, Mi HY. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science*. Jan 2022;31(1):8-22. doi:10.1002/pro.4218

165. DeStefano GM, Kurban M, Anyane-Yeboa K, et al. Mutations in the Cholesterol Transporter Gene ABCA5 Are Associated with Excessive Hair Overgrowth. *Plos Genetics*. May 2014;10(5)e1004333. doi:10.1371/journal.pgen.1004333

166. Sun M, Li N, Dong W, et al. Copy-Number Mutations on Chromosome 17q24.2-q24.3 in Congenital Generalized Hypertrichosis Terminalis with or without Gingival Hyperplasia. *American Journal of Human Genetics*. Jun 2009;84(6):807-813. doi:10.1016/j.ajhg.2009.04.018

167. Long HK, Osterwalder M, Welsh IC, et al. Loss of Extreme Long-Range Enhancers in Human Neural Crest Drives a Craniofacial Disorder. *Cell Stem Cell*. Nov 2020;27(5):765-+. doi:10.1016/j.stem.2020.09.001

168. Gao G, Smith DI. WWOX, large common fragile site genes, and cancer. *Experimental Biology and Medicine*. Mar 2015;240(3):285-295. doi:10.1177/1535370214565992

169. Palakodeti A, Han Y, Jiang YW, Le Beau MM. The role of late/slow replication of the FRA16D in common fragile site induction. *Genes Chromosomes & Cancer*. Jan 2004;39(1):71-76. doi:10.1002/gcc.10290

170. Oliver KL, Trivisano M, Mandelstam SA, et al. WWOX developmental and epileptic encephalopathy: Understanding the epileptology and the mortality risk. *Epilepsia*. 2023 Mar 2023;doi:10.1111/epi.17542

171. Li SB, Wu XH. Common fragile sites: protection and repair. *Cell and Bioscience*. Mar 2020;10(1)29. doi:10.1186/s13578-020-00392-5

172. Zhang H, Freudenreich CH. An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in S-cerevisiae. *Molecular Cell*. Aug 2007;27(3):367-379. doi:10.1016/j.molcel.2007.06.012

173. Hayashi R, Yoshida K, Abe R, Niizeki H, Shimomura Y. First Japanese case of congenital generalized hypertrichosis with a copy number variation on chromosome 17q24. *Journal of Dermatological Science*. Jan 2017;85(1):63-65. doi:10.1016/j.jdermsci.2016.10.010

174. Afifi HH, Fukai R, Miyake N, et al. De Novo 17q24.2-q24.3 microdeletion presenting with generalized hypertrichosis terminalis, gingival fibromatous hyperplasia, and distinctive facial features. *American Journal of Medical Genetics Part A*. Oct 2015;167(10):2418-2424. doi:10.1002/ajmg.a.37185

175. Kim HG, Higgins AW, Herrick SR, et al. Candidate loci for Zimmermann-Laband syndrome at 3p14.3. *American Journal of Medical Genetics Part A*. Jan 2007;143A(2):107-111. doi:10.1002/ajmg.a.31544

176. Higgins AW, Alkuraya FS, Bosco AE, et al. Characterization of apparently balanced chromosomal rearrangements from the developmental genome anatomy project. *American Journal of Human Genetics*. Mar 2008;82(3):712-722. doi:10.1016/j.ajhg.2008.01.011

177. Fantauzzo KA, Kurban M, Levy B, Christiano AM. Trps1 and Its Target Gene Sox9 Regulate Epithelial Proliferation in the Developing Hair Follicle and Are Associated with Hypertrichosis. *Plos Genetics*. Nov 2012;8(11)e1003002. doi:10.1371/journal.pgen.1003002

178. Kurth I, Klopocki E, Stricker S, et al. Duplications of noncoding elements 5 ' of SOX9 are associated with brachydactyly-anonychia. *Nature Genetics*. Aug 2009;41(8):862-863. doi:10.1038/ng0809-862

179. Shirian S, Shahabinejad H, Saeedzadeh A, et al. Zimmermann-Laband syndrome: Clinical and cytogenetic study in two related patients. *J Clin Exp Dent*. May 2019;11(5):e452-e456. doi:10.4317/jced.55214

180. Bauer CK, Schneeberger PE, Kortum F, et al. Gain-of-Function Mutations in KCNN3 Encoding the Small-Conductance Ca2+-Activated K+ Channel SK3 Cause Zimmermann-Laband Syndrome. *American Journal of Human Genetics*. Jun 2019;104(6):1139-1157. doi:10.1016/j.ajhg.2019.04.012

181. Marrus SB, Cuculich PS, Wang W, Nerbonne JM. Characterization of a novel, dominant negative KCNJ2 mutation associated with Andersen-Tawil syndrome. *Channels*. Nov-Dec 2011;5(6):500-509. doi:10.4161/chan.5.6.18524

182. Yoon G, Oberoi S, Tristani-Firouzi M, et al. Andersen-Tawil syndrome: Prospective cohort analysis and expansion of the phenotype. *American Journal of Medical Genetics Part A*. Feb 2006;140A(4):312-321. doi:10.1002/ajmg.a.31092

183. Andersen ED, Krasilnikoff PA, Overvad H. INTERMITTENT MUSCULAR WEAKNESS, EXTRASYSTOLES, AND MULTIPLE DEVELOPMENTAL ANOMALIES - NEW SYNDROME. *Acta Paediatrica Scandinavica*. 1971;60(5):559-+. doi:10.1111/j.1651-2227.1971.tb06990.x

184. Priori SG, Pandit SV, Rivolta I, et al. A novel form of short QT syndrome (SQT3) is caused by a mutation in the KCNJ2 gene. *Circulation Research*. Apr 2005;96(7):800-807. doi:10.1161/01.RES.0000162101.76263.8c

185. Xia M, Jin QF, Bendahhou S, et al. A Kir2.1 gain-of-function mutation underlies familial atrial fibrillation. *Biochemical and Biophysical Research Communications*. Jul 2005;332(4):1012-1019. doi:10.1016/j.bbrc.2005.05.054

186. Howarth KD, Pole JCM, Beavis JC, et al. Large duplications at reciprocal translocation breakpoints that might be the counterpart of large deletions and could arise from stalled replication bubbles. *Genome Research*. Apr 2011;21(4):525-534. doi:10.1101/gr.114116.110

187. Wang YA, Zhao ZH, Fu XY, Li SF, Zhang QY, Kong XD. Detection of a Cryptic 25 bp Deletion and a 269 Kb Microduplication by Nanopore Sequencing in a Seemingly Balanced Translocation Involving the LMLN and LOC105378102 Genes. *Frontiers in Genetics*. Aug 2022;13883398. doi:10.3389/fgene.2022.883398

188. Brand H, Collins RL, Hanscom C, et al. Paired-Duplication Signatures Mark Cryptic Inversions and Other Complex Structural Variation. *American Journal of Human Genetics*. Jul 2015;97(1):170-176. doi:10.1016/j.ajhg.2015.05.012

189. Kanai SM, Heffner C, Cox TC, et al. Auriculocondylar syndrome 2 results from the dominant-negative action of PLCB4 variants. *Disease Models & Mechanisms*. Apr 2022;15(4)dmm049320. doi:10.1242/dmm.049320

190. Kido Y, Gordon CT, Sakazume S, et al. Further Characterization of Atypical Features in Auriculocondylar Syndrome Caused by Recessive PLCB4 Mutations. *American Journal of Medical Genetics Part A*. Sep 2013;161(9):2339-2346. doi:10.1002/ajmg.a.36066

191. Ivey K, Tyson B, Ukidwe P, et al. G alpha(q) G alpha(11) proteins mediate endothelin-1 signaling in neural crest-derived pharyngeal arch mesenchyme. *Developmental Biology*. Mar 2003;255(2):230-237. doi:10.1016/s0012-1606(02)00097-0

192. Abe M, Ruest LB, Clouthier DE. Fate of cranial neural crest cells during craniofacial development in endothelin-A receptor-deficient mice. *International Journal of Developmental Biology*. 2007;51(2):97-105. doi:10.1387/ijdb.062237ma

193. Pritchard AB, Kanai SM, Krock B, et al. Loss-of-function of Endothelin receptor type A results in Oro-Oto-Cardiac syndrome. *American Journal of Medical Genetics Part A*. May 2020;182(5):1104-1116. doi:10.1002/ajmg.a.61531

194. Gordon CT, Weaver KN, Zechi-Ceide RM, et al. Mutations in the Endothelin Receptor Type A Cause Mandibulofacial Dysostosis with Alopecia. *American Journal of Human Genetics*. Apr 2015;96(4):519-531. doi:10.1016/j.ajhg.2015.01.015

195. Nurk S, Koren S, Rhie A, et al. The complete sequence of a human genome. *Science*. Apr 2022;376(6588):44-+. abj6987. doi:10.1126/science.abj6987

196. Yulia M, Philip MB, Yongqing H, et al. Resolution of ring chromosomes, Robertsonian translocations, and complex structural variants from long-read sequencing and telomere-to-telomere assembly. *bioRxiv*. 2023:2023.09.07.555775. doi:10.1101/2023.09.07.555775

197. Calpena E, Cuellar A, Bala K, et al. SMAD6 variants in craniosynostosis: genotype and phenotype evaluation. *Genetics in Medicine*. Sep 2020;22(9):1498-1506. doi:10.1038/s41436-020-0817-2

198. Wang YH, Zhao Y, Bollas A, Wang YR, Au KF. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology*. Nov 2021;39(11):1348-1365. doi:10.1038/s41587-021-01108-x

199. Kolmogorov M, Billingsley KJ, Mastoras M, et al. Scalable Nanopore sequencing of human genomes provides a comprehensive view of haplotype-resolved variation and methylation. *Nature Methods*. 2023 Sep 2023;doi:10.1038/s41592-023-01993

200. Moritz S, Luis FP, Christopher MG, et al. Comprehensive Structural Variant Detection: From Mosaic to Population-Level. *bioRxiv*. 2023:2022.04.04.487055. doi:10.1101/2022.04.04.487055

201. Kosugi S, Momozawa Y, Liu XX, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biology*. Jun 2019;20117. doi:10.1186/s13059-019-1720-5

202. Zook JM, Hansen NF, Olson ND, et al. A robust benchmark for detection of germline large deletions and insertions. *Nature Biotechnology*. Nov 2020;38(11):1347-+. doi:10.1038/s41587-020-0538-8

203. Sudmant PH, Rausch T, Gardner EJ, et al. An integrated map of structural variation in 2,504 human genomes. *Nature*. Oct 2015;526(7571):75-+. doi:10.1038/nature15394

204. Werling DM, Brand H, An JY, et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nature Genetics*. May 2018;50(5):727-+. doi:10.1038/s41588-018-0107-y

205. Collins RL, Brand H, Karczewski KJ, et al. A structural variation reference for medical and population genetics. *Nature*. May 2020;581(7809):444-+. doi:10.1038/s41586-020-2287-8

206. John SO, Anthony C, Jennifer RD, et al. High-Definition Electronic Genome Maps from Single Molecule Data. *bioRxiv*. 2017:139840. doi:10.1101/139840

207. Jakobsen LP, Ullmann R, Christensen SB, et al. Pierre Robin sequence may be caused by dysregulation of SOX9 and KCNJ2. *Journal of Medical Genetics*. Jun 2007;44(6):381-386. doi:10.1136/jmg.2006.046177

208. Gordon CT, Attanasio C, Bhatia S, et al. Identification of Novel Craniofacial Regulatory Domains Located far Upstream of SOX9 and Disrupted in Pierre Robin Sequence. *Human Mutation*. Aug 2014;35(8):1011-1020. doi:10.1002/humu.22606

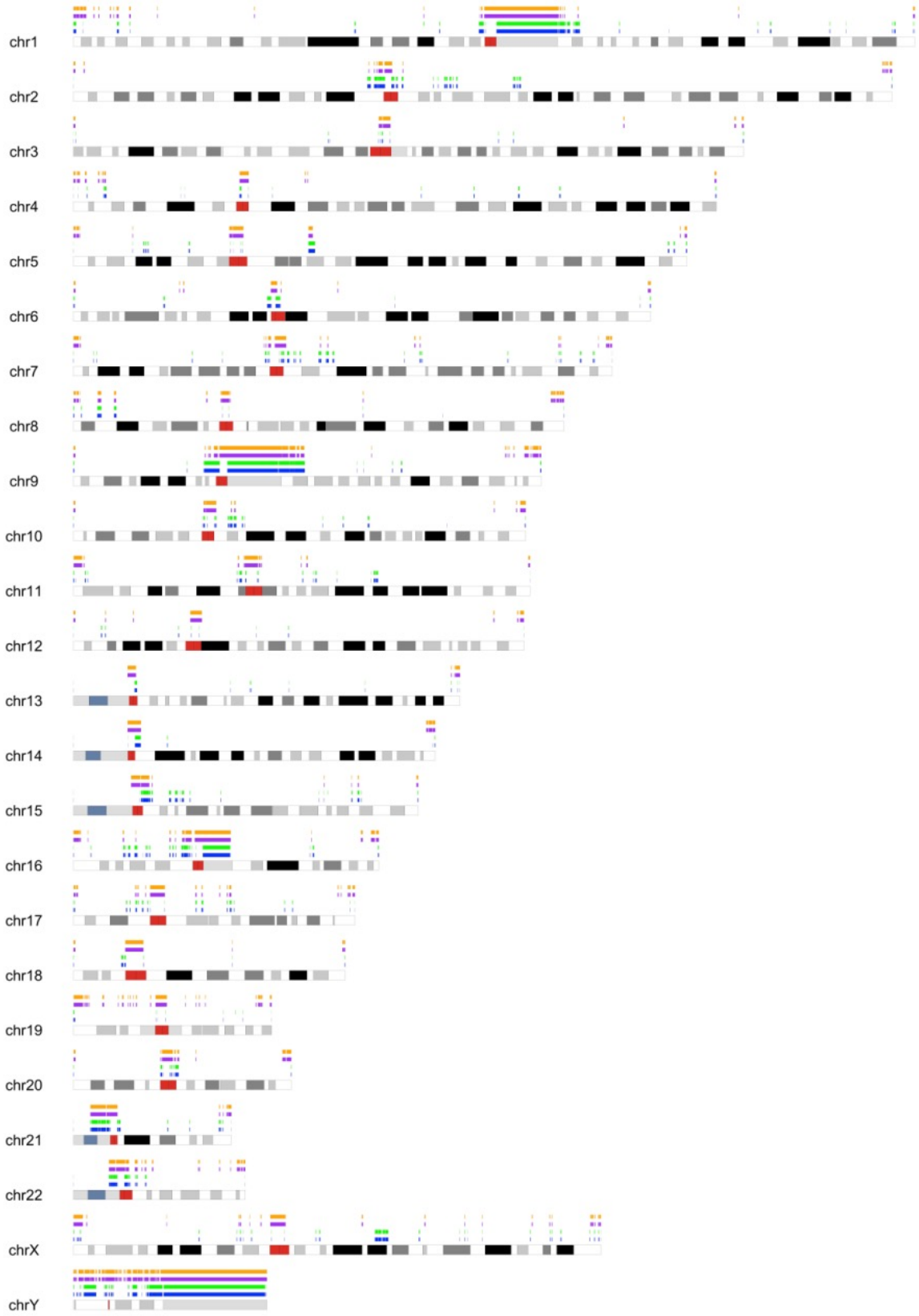
209. Benko S, Fantes JA, Amiel J, et al. Highly conserved non-coding elements on either side of SOX9 associated with Pierre Robin sequence. Article. *Nature Genetics*. Mar 2009;41(3):359-364. doi:10.1038/ng.329

Appendix

Supplementary Table 1 Bionano internal control database contains 179 healthy individuals

Classification	Count
African (AFR)	45
Admixed American (AMR)	16
East Asian (EAS)	17
European (EUR)	44
South Asian (SAS)	15
Unknown	43

Table adapted from "Bionano Solve Theory of Operation: Variant Annotation Pipeline, Document Number: 30190, Document Revision: J" available from https://bionano.com/



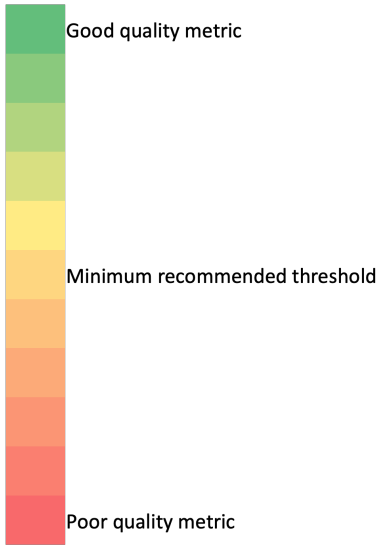
Supplementary Figure 1 *SVs & CNVs included in the Bionano SV mask in hg38. Figure extracted from "Bionano Solve Theory of Operation: Structural Variant Calling, Document Number: 30110, Document Revision: K" available from https://bionano.com/*

Supplementary Table 2 cases in local CRS ID or sample ID (if CRS ID is not available)

Case ID	CRS ID
1	537
2	347
3	L16
4	510
5	M196
6	S250
7	566
8	169
9	290
10	S474
11	9073
12	SC46
13	685
14	M29
15	9115
16	302
17	M44
18	9159
19	9162
20	9165
21	608
22	SC89
23	-
24	-
25	-
26	-
27	427
28	-
29	-
30	-
31	633

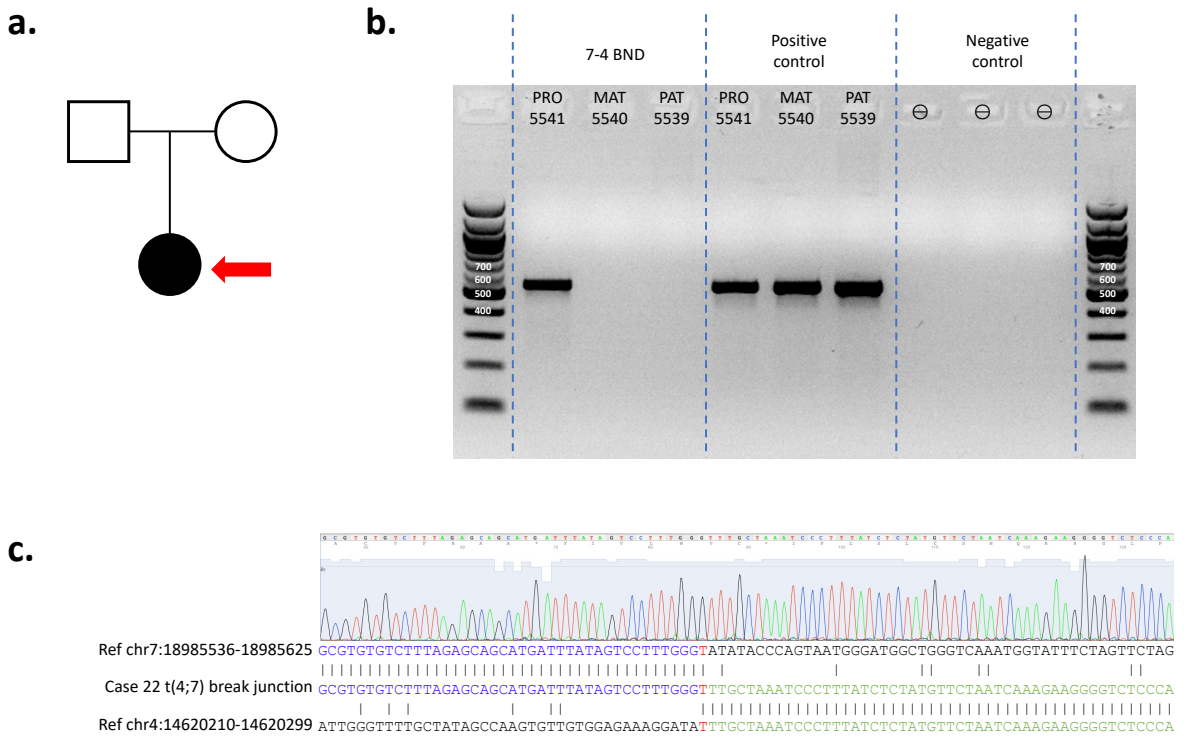
Supplementary Table 3 Key metrics for the data quality obtained on each Bionano OGM run

CRS ID	Case ID	Data quality	Total DNA (>= 20 kbp), Gbp	Total DNA (>=150 kbp), Gbp	N50 (>= 20 kbp), kbp	N50 (>=150 kbp), kbp	Average label density (>= 150 kbp), /100 kbp	Map rate (>= 150 kbp)	Effective coverage (X)	Molecule integrity number	Positive label variance (PLV)	Negative label variance (NLV)
537	1	H	1786.7	1512.7	328.9	370.9	14.6	92.6%	422.67	0.09	3.0%	8.5%
347	2	H	1926.2	1504.0	292.2	352.9	15.1	92.1%	426.68	0.09	3.0%	8.0%
L16	3	M	944.3	768.1	336.4	409.9	17.9	81.5%	195.79	0.09	2.6%	9.0%
510	4	M	675.6	507.8	265.5	328.9	14.3	84.1%	132.08	0.08	2.2%	14.6%
M196	5	L	965.9	510.7	159.8	267.4	15.3	51.2%	77.88	0.12	3.0%	18.6%
S250	6	H	2393.2	1519.6	195.5	269.6	15.8	92.5%	438.15	0.08	2.7%	6.0%
566	7	M	752.8	604.6	302.6	356.3	15.4	92.7%	173.54	0.07	4.3%	6.0%
169	8	H	3189.4	1527.1	125.6	409.5	16.4	90.8%	432.94	0.07	2.9%	6.4%
290	9	H	2163.0	1529.0	239.2	317.3	16.1	89.9%	417.7	0.1	2.4%	7.7%
S474	10	M	1371.4	867.1	217.1	348.9	18.0	79.5%	217.75	0.17	2.7%	7.5%
9073	11	M	1986.5	1522.5	311.6	399.7	20.1	73.2%	348.12	0.12	2.8%	10.7%
SC46	12	H	2364.6	1521.9	193.5	257.6	15.0	91.2%	426.99	0.06	2.7%	7.6%
685	13	H	681.3	542.0	272.6	320.2	15.8	93.4%	157.1	0.08	2.4%	7.2%
M29	14	H	1582.4	1222.3	285.2	349.5	15.5	91.2%	344.51	0.08	2.6%	9.0%
9115	15	H	1821.3	1511.0	342.0	397.7	17.6	86.4%	407.66	0.09	2.1%	7.9%
302	16	H	1855.6	1525.6	290.6	333.5	15.6	94.7%	450.36	0.07	2.3%	6.8%
M44	17	H	1960.1	1508.4	257.2	308.6	15.5	92.9%	433.17	0.07	2.3%	7.7%
9159	18	H	1995.7	1529.4	267.6	333.0	16.3	89.8%	428.91	0.09	2.7%	7.7%
9162	19	H	2013.0	1538.3	266.7	326.0	15.6	91.3%	435.4	0.1	2.4%	7.7%
9165	20	H	1965.1	1431.0	277.1	379.9	14.9	87.9%	388.88	0.09	2.9%	12.4%
	Recommended minimum for *de novo* assembly				> 150	> 230	14-17	>70%	>80	<20	<10%	<15%
	High quality for both RVA and *de novo* assembly				> 150	> 230	14-17	~90%	>200	<20	<10%	<15%



Case 22 – additional finding from case 12

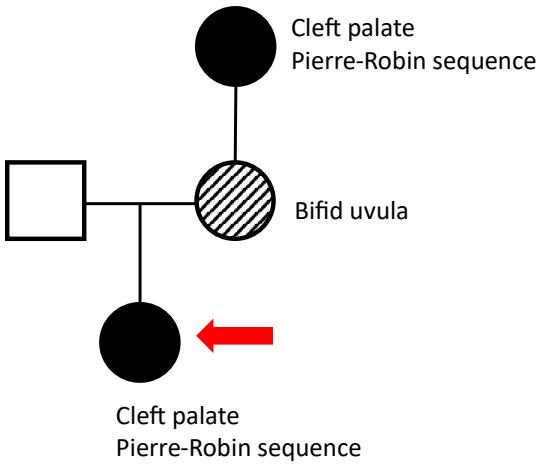
The second case of Saethre-Chotzen syndrome diagnosis arose from a balanced 4-7 translocation detected previously via amniotic testing in case 22. As shown in **Supplementary Figure 2a**, Case 22 is a trio family where only the proband is affected, suggesting a *de novo* origin of the translocation event. Breakpoint PCR successfully amplified the unique break junction presented *de novo* in the proband and not the in parents, as shown in in **Supplementary Figure 2b**. Dideoxy- sequencing subsequently characterised the break junction sequences as shown in in **Supplementary Figure 2c**.



Supplementary Figure 2 Case 22 is a trio family where the proband was diagnosed with Saethre-Chotzen syndrome due to a 4-7 balanced translocation. a. *case 22 proband is affected while the parents are unaffected.* ***b.*** *Breakpoint PCR successfully amplified the break junction in the proband but not the parents. Positive and negative controls were included.* ***c.*** *Dideoxy- sequencing confirmed and characterised the detail break junction of the 4-7 translocation. Figure in hg38.*

Additional clinical outcome: Case 23

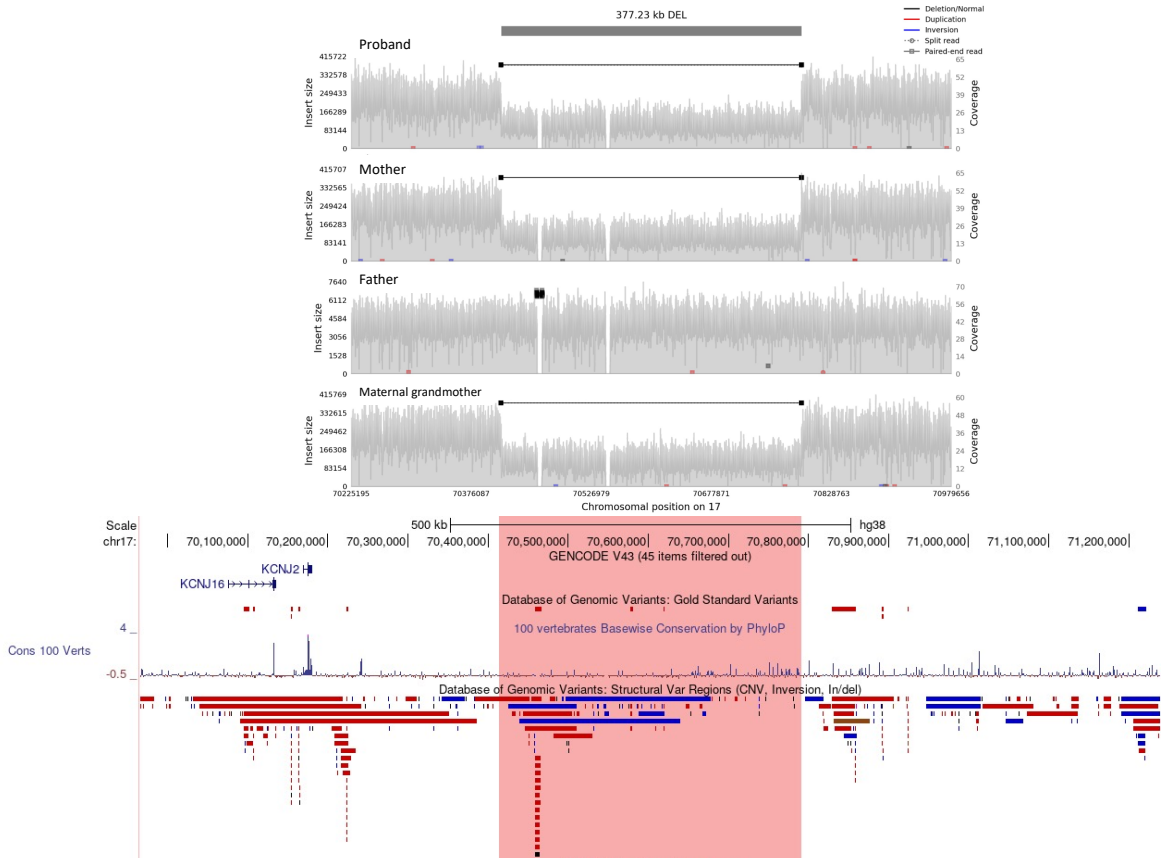
To identify further supporting cases in the 100kGP with SVs in the *KCNJ* locus, Case 23 was discovered as an additional finding. For this case, three affected and one unaffected family members were recruited to the 100kGP because of their cleft palate phenotype, as shown in **Supplementary Figure 3**. As recorded in the 100kGP, the proband and the maternal grandmother both presented cleft palate and Pierre-Robin sequence, while the father and the mother were documented as unaffected. Upon investigating the HPO terms, the mother was found to have a mild form of the cleft palate, presenting as a bifid uvula. This suggests a dominant inheritance pattern in the family, where the cleft palate phenotype and the underlying genetic cause are transmitted sequentially from the maternal grandmother to the mother, and subsequently to the proband.



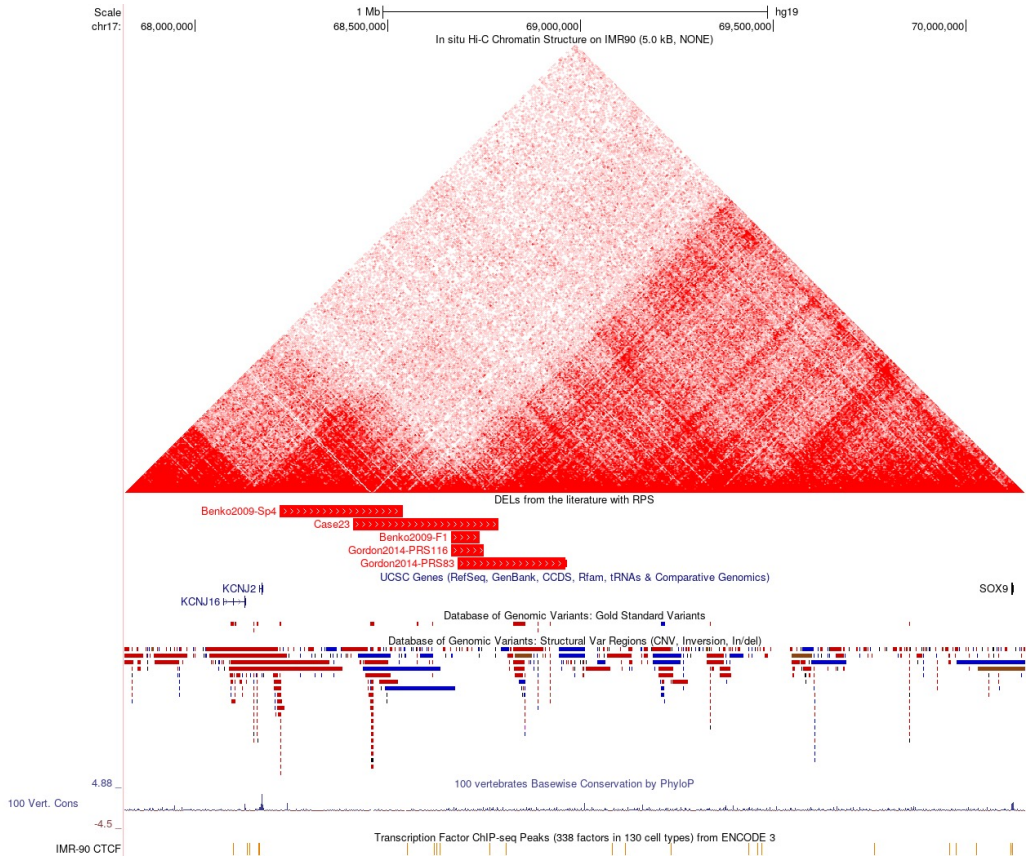
Supplementary Figure 3 case 23 is a family with three affected individuals recruited to the 100kGP project due to their cleft palate phenotype. *The mother has a mild form of clefting presented as a bifid uvula.*

By screening the 17DUP region, a 377 kb DEL was found to perfectly segregate with the cleft palate phenotype in case 23, as shown in **Supplementary Figure 4**. The DEL affects a critical regulatory region at the *KCNJs-SOX9* loci and is consistent with other

DELs in the literature in patients with Pierre-Robin sequence, as shown in **Supplementary Figure 5**.[207-209] This finding was reported back to GE as a clinical diagnosis and submitted as a CNV region for clefting panel in the PanelApp. However, no update from GE since and the case remain without a molecular diagnosis in the 100kGP to date.



Supplementary Figure 4 A 323 kb DEL on chr17 was found in family 23 and perfectly segregates with the cleft palate phenotype. *Samplot figure exported from GE Airlock. Figure in hg38.*



Supplementary Figure 5 Three example deletions in patients with Pierre-Robin sequence in the literature showing that the case 23 deletion is likely pathogenic. *Data from Gordan et al (2014)[208] and Benko et al (2009).[209] Figure in hg19.*