Detection of Copy Number Variants using Targeted Next-Generation Sequencing Data from the Inherited Disease Panel

Carmen Phipson

Student number: 1863844

Supervisors:

Mrs Nkateko Mayevu & Dr Fiona Baine-Savanhu

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

A research report (in the format of a "submissible" paper) submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the Degree of Master of Science in Medicine (Genomic Medicine) by Research and Coursework.

Johannesburg, 2025

NATIONAL HEALTH LABORATORY SERVICE
School of Pathology, University of the Witwatersrand
DIVISION OF HUMAN GENETICS
Hospital Street, Johannesburg, 2001 I PO Box 1038, Johannesburg, 2000
[T] +27 11 489 9223 I [M] +27 78 080 8841 I [F] +27 11 489 9226 I [E] human.genetics@nhls.ac.za

5 December 2024

Dear Examiner,

Re: Research Report Submitted to the Faculty of Health Sciences by Carmen in Partial Fulfilment of the Requirements for the Degree of Master of Science in Medicine (Genomic Medicine)

Thank you for agreeing to examine my research report, titled *"**Detection of Copy Number Variants Using Targeted Next-Generation Sequencing Data from the Inherited Disease Panel**."* This study explores the potential for CNV detection in targeted next-generation sequencing (NGS) data, specifically from the Inherited Disease Panel.

The manuscript has been written in 'submissible' format and formatted according to the requirements of *BMC Medical Genomics*. This is an open-access journal that publishes research in genomics of human health and disease, including diagnostic techniques and systems analyses, which align with the themes of my study. The journal's author guidelines have been included in the additional files for your reference.

To facilitate review, the tables and figures have been included within the main manuscript document rather than as separate files. Additionally, the approved research protocol has been attached as an additional file to provide an extended background. This protocol has already undergone assessment and approval by the Faculty of Health Sciences and therefore not for examination. A table of contents has been included for easy navigation of the report.

Thank you for your time and consideration in assessing this work.

Yours sincerely,

Carmen Phipson (Candidate)

Nkateko Mayevu
(Supervisor)

Fiona Baine-Savanhu
(Supervisor)

DECLARATION

I, Carmen Phipson, declare that this research report (in the format of a "submissible" paper) is my own, unaided work. It is being submitted for the Degree of Master of Science in Medicine at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other university.

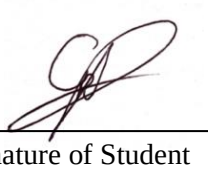
Carmen Phipson
Student number: 1863844

4th day of March, 2025, in Johannesburg

CONTRIBUTION OF CANDIDATE

Declaration: Student's contribution to the article and agreement of co-author(s)

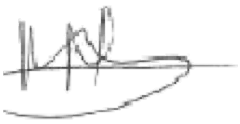
I, Carmen Phipson, student number 1863844, declare that this research report is my own work and that I contributed significantly towards the research findings presented in the manuscript below.



Signature of Student

5 December 2024

Date



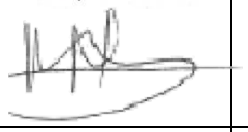
Signature of Supervisor

5 December 2024

Date

Agreement by co-authors: By signing this declaration, the co-authors listed below agree to the use of the article(s) by the student as part of her research report.

Article Title: **Detection of Copy Number Variants using Targeted Next-Generation Sequencing Data from the Inherited Disease Panel**

Authors	**Name**	**Signature**	**Date**
First Author	Carmen Phipson		5/12/2024
Second Author	Nkateko Mayevu		5/12/2024
Third Author	Fiona Baine-Savanhu		5/12/2024

PRESENTATIONS ARISING FROM THIS RESEARCH

1. YEGG Symposium, 26th October 2024, University of the Witwatersrand Public Health Auditorium (poster presentation)

2. SASHG Biennial Congress 2024, 27th–29th October 2024, Sun City, South Africa (poster presentation)

ACKNOWLEDGEMENTS

I would like to thank my supervisors, Fiona Baine-Savanhu and Nkateko Mayevu, for their direction and guidance throughout the year and for keeping me on track through regular meetings. I am also very grateful for their help in refining my manuscript before submission.

Secondly, I thank Next Biosciences for making arrangements for me to continue with my part-time job throughout my research and for being supportive and understanding of my busy schedule. This job has allowed me to cover my own tuition fees this year.

Lastly, I thank the various staff and teachers at the NHLS, especially Dr Maria Mudau, Dr Robyn Kerr, and Ms Monica Araujo, who were readily available to answer my questions, provide helpful resources, and give me advice.

Research Article
For submission to *BMC Medical Genomics*

Detection of Copy Number Variants using Targeted Next-Generation Sequencing Data from the Inherited Disease Panel

Carmen Phipson[1], Nkateko Mayevu[1], Fiona Baine-Savanhu[1]

[1]Division of Human Genetics, National Health Laboratory Service and School of Pathology, University of the Witwatersrand

Corresponding Author:
Carmen Phipson
Division of Human Genetics, Faculty of Health Sciences, University of the Witwatersrand
+27 72 703 4205
1863844@students.wits.ac.za / carmenrphipson@gmail.com

TABLE OF CONTENTS

ABSTRACT

Background: Copy number variants (CNVs) are linked to numerous genomic disorders, especially those involving dosage-sensitive genes. While chromosomal microarrays have been the standard method for CNV detection, next-generation sequencing (NGS) allows for CNV detection with improved resolution. The Inherited Disease Panel (IDP) is a custom Ion AmpliSeq™ panel targeting 500 genes associated with a range of single-gene disorders. This study aimed to develop a baseline for a CNV-detection workflow specific to samples sequenced using the IDP and to assess the workflow's utility as a complementary analysis tool.

Methods: This study used NGS data derived from patient samples that were negative for single nucleotide variants. Three custom read depth baselines were created in Ion Reporter™, namely: a 10-sample baseline, a 48-sample baseline, and a uniform-read-depth baseline using 18 samples with mean read depths of between 200x and 400x. Their performance was tested by incorporating each into a workflow and analysing five samples with known CNVs. The best-performing workflow was further applied to 78 samples. The resulting CNV calls were filtered by genes and confidence scores. Candidate CNVs were then classified according to the American College of Medical Genetics and ClinGen guidelines.

Results: The known CNVs were detected in all five validation samples. The workflow incorporating the uniform-read-depth baseline was selected for all subsequent analyses. Analysis of the IDP samples identified four potentially clinically relevant CNVs, including pathogenic deletions in the *TCOF1* and *NIPBL* genes, as well as a whole X-chromosome duplication. CNVs of interest all had high confidence and matching precision. Analyses of samples with mean read depths 200–400x showed improved specificity compared to samples outside this range.

Conclusions: This CNV-detection method demonstrated potential as a supplementary analysis tool, with a diagnostic yield of ~4.3% (3/70). A few limitations were identified, particularly with samples exhibiting aberrant CNV calls and genes with recurrent CNVs across multiple samples, although manual prioritization strategies were able to counteract these. This method may be useful as a screening tool; however, all clinically relevant CNVs require validation before reporting.

[333 words]

Keywords: copy number variants; next-generation sequencing; targeted gene panel

LIST OF ABBREVIATIONS

ACMG: American College of Medical Genetics

bp: base pairs

ClinGen: Clinical Genome

CNV: copy number variant

HI: haploinsufficiency

IGV: Integrative Genomics Viewer

IDP: inherited disease panel

Indel: insertion-deletion

ISCN: International System for Human Cytogenomic Nomenclature

kb: kilobases

Mb: megabases

NGS: next-generation sequencing

NHLS: National Health Laboratory Service

PCR: polymerase chain reaction

SNV: single nucleotide variant

TS: triplosensitivity

VUS: variant of unknown significance

LIST OF FIGURES

LIST OF TABLES

BACKGROUND

Copy number variants (CNVs) are structural variants characterised by the deletion or duplication of a segment of DNA ranging from 50 base pairs (bp) to several megabases (Mb). These variants make up a significant proportion of normal genomic variation (1) but have also been implicated in numerous genomic disorders (2). The primary mechanism by which CNVs lead to disease is through the loss or gain of haploinsufficient (HI) or triplosensitive (TS) genes, respectively. CNVs affecting such dosage-sensitive genes represent a significant portion of pathogenic variants in monogenic disorders (3), especially those related to developmental pathways that are highly conserved and less tolerant to dosage variation (4).

The current recommended method for detecting CNVs is chromosomal microarray, which is sensitive but limited in its resolution, as it depends on the genomic spacing of the probes used in its design (5). In recent years, next-generation sequencing (NGS) has emerged as a promising alternative for CNV detection, offering increased precision and higher throughput than chromosomal microarray (6). NGS-based assays offer higher coverage and enhanced resolution and are thus valuable for diagnosing hereditary diseases linked to smaller, often intragenic, CNVs (6,7). The most commonly used computational model for detecting CNVs in NGS data is the 'read depth' method, which assumes that the read depth of a chromosomal segment is proportional to the number of copies present in that sample(8). This method compares the read depth of a target region with that of a reference baseline to infer CNVs.

Previous CNV-detection methods using the read depth approach have been able to capture whole-gene CNVs and exon-level CNVs, including multi-exon, single-exon, and partial-exon variants (9). In a study on CNV detection in targeted gene panels, all of the known CNVs (n=36) were detected using a read depth approach, albeit with imperfect specificity (90.9%). The false-positive calls (n=126) all occurred in regions with very low coverage due to suboptimal capture by the kit or in repetitive sequence regions. Of these, 54 were systematic errors observed in the same genes across samples. The read depth method has also been shown to accurately predict exact copy numbers, which is particularly valuable for investigating copy number polymorphic genes (1). Across studies, the main limitation of the read depth method is that detection of CNVs is hindered in repetitive regions, homologous regions, and GC-rich regions that are challenging to sequence and map (1,9,10).

Ion AmpliSeq™ is a targeted sequencing technology that generates short-read sequencing data with high coverage, making it valuable for CNV detection using the read depth approach. The sequencing data is analysed using Ion Reporter™ software, in which results are often filtered by confidence and precision scores (11). Confidence scores indicate the certainty of a deviation from the expected copy number (2 for autosomes), while precision scores measure the certainty of the assigned copy number value (0 or 1 for a deletion; ≥3 for a duplication).

The inherited disease panel (IDP) is a custom-designed Ion AmpliSeq™ targeted sequencing panel used at the National Health Laboratory Service (NHLS) in Johannesburg. It includes multiple sub-panels of genes associated with genetic disorders, including hereditary cancers, congenital syndromes, musculoskeletal disorders, RASopathies, and others, comprising over 500 genes in total. Although the IDP is primarily used for the detection of single nucleotide variants (SNVs) and small indels (insertion-deletions), it holds potential for CNV analysis.

When using this platform for CNV detection, it is essential to have a reference baseline specific to the panel and representative of the target population. Currently, no such baseline exists for NGS data from the IDP. To ensure the clinical relevance of detected CNVs, a custom baseline must be developed, tailored to the target regions of the IDP and accounting for normal genomic variation in South African patients.

Therefore, this study aimed to create and compare multiple IDP-specific baselines and to develop a CNV-detection workflow using a custom baseline. The workflow was subsequently used to re-analyse data from previously sequenced samples. A secondary aim of the study was to assess the utility of this method in detecting clinically significant CNVs.

METHODS

Data Handling and Ethics

The study design was a secondary data analysis, investigating samples that were previously sequenced using the IDP and were negative for pathogenic/likely pathogenic SNVs.

NGS data was derived from the DNA of patients with a suspected genetic disorder referred for molecular genetic testing. All sample data was anonymised using project codes (e.g., CNV001) to maintain patient confidentiality.
The study was performed in accordance with ethics guidelines and approved by the University of the Witwatersrand's Human Research Ethics Committee (HREC) Medical under ethics clearance certificate number M240429 (Additional File 6).

Data processing and analyses were conducted using Ion Reporter™ software (Thermo Fisher Scientific), which is specifically designed for analysis of sequencing data from the Ion Torrent semiconductor platform. Sequence data was provided in the form of binary alignment map (BAM) files. A quality check was performed on the BAM files; those with a mean read depth of <100 and >500 were excluded.

Baseline Creation and Validation

Three read depth baselines were created using data from samples sequenced on the IDP. The baselines were created using only male samples to standardise the coverage of the X chromosome to a ploidy of 1. Firstly, a 10-sample baseline using randomly selected samples was created for a proof-of-concept analysis. A second baseline was created using 48 samples, in accordance with the Ion Reporter™ User Guide (12). These samples were selected based on their low likelihood of harbouring clinically significant CNVs i.e., each sample's suspected clinical diagnosis was reviewed on GeneReviews® (13) and those with a reported CNV yield of less than 10% were included. Lastly, a 'uniform-read-depth' baseline was created using samples with mean read depths between 200x and 400x, which is roughly the mean ($\bar{x} = 312$) ± 1 standard deviation ($s = 102$). A total of 18 samples within this mean read depth range were available for use in this baseline.

Each of the three baselines was incorporated into a separate workflow with default parameters and applied to a validation set, that is, five samples with known clinically significant CNVs. These analyses were done to verify that the workflow could identify true positives whilst limiting 'false calls,' i.e. those due to technical artefacts or common population variation. The workflow that exhibited the highest confidence scores for the known positives, paired with a relatively low number of total calls, was selected for all subsequent analyses.

Workflow Customization

The selected workflow used the human reference genome GRCh37 (hg19) for alignment and annotation and applied the most up-to-date panel design, IDPv4 (Additional File 7), to define target regions.

The workflow also included several adjusted parameters, including sensitivity set to 'low', which allows for enhanced specificity of CNV calls (i.e., fewer false calls) compared to the default 'medium' setting, and median absolute pairwise difference (MAPD) set to 0.4, which filters for less variability in read depth across adjacent amplicons (compared to the default of 0.5).

Sample Analyses

A total of 78 BAM files derived from samples sequenced using the IDP between 2019 and 2024 and negative for pathogenic SNVs were analysed using the custom CNV-detection workflow. These samples were selected based on an estimated higher chance of harbouring pathogenic CNVs i.e., the suspected clinical diagnoses were reviewed on GeneReviews® (13), and only those with a documented CNV-detection rate >10% were included. The mean read depth of each sample was obtained from the quality control files generated by Ion Reporter™.

For each sample's analysis, the total calls (defined as the number of calls with confidence >0), high-confidence calls (defined as calls with confidence >20), and calls with matching confidence and precision scores were recorded (11).

The CNVs were filtered for those with confidence scores greater than 20 and by genes associated with the suspected genetic condition, including those associated with the differential diagnoses, based on GeneReviews® (13). Shortlisted variants were inspected visually using the Integrative Genomics Viewer (IGV) desktop application (14) and compared with at least one other sample of similar mean read depth to rule out technical artefacts and common variants.

Classification of Candidate CNVs

Candidate CNVs were classified according to the American College of Medical Genetics (ACMG)-ClinGen guidelines (15). This process involved the use of VarSome (16) and public databases, including ClinVar (17) and ClinGen (18), to gather the required evidence. Patient phenotypes were reviewed in order to assess genotype-phenotype correlations, where necessary, as part of the classifications.

RESULTS

Baseline Validation

Three workflows were created using the custom baselines and tested for effectiveness with five CNV-positive samples. Each workflow detected the expected CNV in all of the validation samples, with consistency in the size and genomic coordinates called across all workflows. These CNVs had high confidence scores and corresponding precision scores (Table 1), which indicate equal certainty of the deviation from the normal copy number and certainty of the exact copy number state (11).

Table 1. CNVs detected in validation samples using three workflows incorporating the different baselines

Sample	Known CNV Detected?	ISCN (hg19)	Size	Confidence	Precision	Visible on IGV?
Pos1—homozygous 18.4 kb deletion in *IDS* (exons 1 – 8)						
Workflow 1	Yes	Xq28(148568380-148586757)x0	18.4 kb	159.88*	159.88	Yes
Workflow 2	Yes	Xq28(148568380-148586757)x0	18.4 kb	159.88*	159.88	
Workflow 3	Yes	Xq28(148568380-148586757)x0	18.4 kb	328.48*	328.48	
Pos2—1.3 Mb deletion including *NF1*						
Workflow 1	Yes	17p11.2q11.2(17713247-29701203)x1	11.9 Mb‡	475.1*	475.1	Yes
Workflow 2	Yes	17p11.2q11.2(17713247-29701203)x1	11.9 Mb	463.58*	463.58	
Workflow 3	Yes	17p11.2q11.2(17713247-29701203)x1	11.9 Mb	468.02*	468.02	
Pos3—heterozygous 11.6 kb deletion in *CHD7* (exon 2)						
Workflow 1	Yes	8q12.2(61653950-61655666)x1	1.7 kb	54.17*	54.17	Yes
Workflow 2	Yes	8q12.2(61653950-61655666)x1	1.7 kb	63.69*	63.69	
Workflow 3	Yes	8q12.2(61653950-61655666)x1	1.7 kb	40.81*	40.81	
Pos4—heterozygous 384 kb deletion including *RBM8A*						
Workflow 1	Yes	1q21.1(145414710-145509287)x1	94.6 kb†	116.71*	116.71	No
Workflow 2	Yes	1q21.1(145414710-145509287)x1	94.6 kb	126.94	104.59	
Workflow 3	Yes	1q21.1(145414710-145509287)x1	94.6 kb	150.77	115.81	
Pos5—heterozygous 3.1 Mb deletion including *RBM8A*						
Workflow 1	Yes	1q21.1(145414710-145509287)x1	94.6 kb†	76.62*	76.62	Yes
Workflow 2	Yes	1q21.1(145414710-145509287)x1	94.6 kb	126.24*	126.24	
Workflow 3	Yes	1q21.1(145414710-145509287)x1	94.6 kb	99.89*	99.89	

* These values represent the highest confidence with matching precision among the sample's analysis results

† This size denotes the genomic interval from *HJV* to *RBM8A.* The full size of the deletion cannot be determined using IDP-based sequencing data.

‡The 1.3 Mb loss in Pos2 involving the *NF1* gene was called as a 12 Mb loss between the *RAI1* gene located at 17p11.2 and the *NF1* gene located at 17q11.2.

ISCN, International System for Human Cytogenomic Nomenclature; IGV, Integrative Genomics Viewer

Of note, the analysis of Pos2 yielded a large deletion call (12 Mb), including *NF1* and spanning the centromere of chromosome 17 (Figure 1). Upon inspection of the region in IGV, a loss in read depth was seen in *NF1* (17q11.2) as well as in the final exon of *RAI1* (17p11.2) (Additional File 1, Figure S2). There is no coverage between the two genes as the IDP design does not include any of the intervening genes. Thus, while the actual deletion is just 1.3 Mb, including *NF1*, the algorithm called a 12 Mb deletion from *RAI1* to *NF1*.

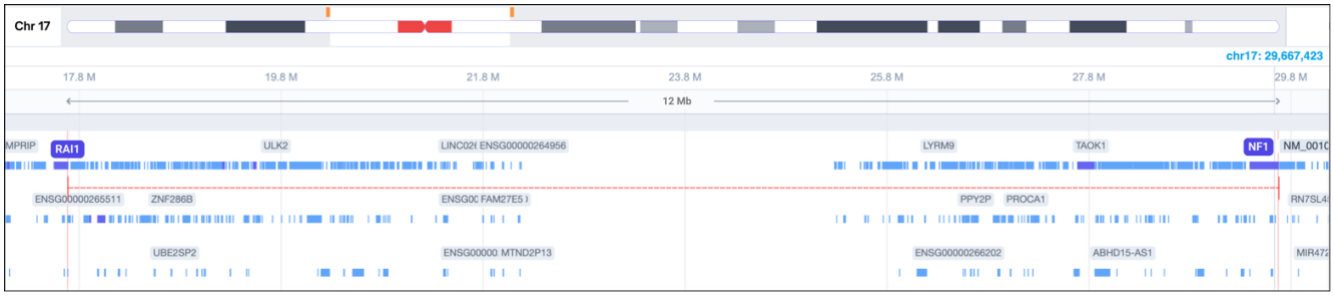


Figure 1. Image from VarSome genome browser showing the 12 Mb deletion called in sample Pos2. The deletion extends from the 3' end of the *RAI1* gene, up to and including the *NF1* gene, spanning the centromere.

In the analyses of Pos4, the deletion of *RBM8A* was the only CNV called with inconsistent confidence and precision scores; it was also not visible on IGV. This may result from variation in read depth of this region in the samples used to create the baseline and to compare in IGV. Such variability across samples may result from differences in primer hybridization, PCR amplification efficiency, homopolymers, repeat tracts, or template concentration in the library preparation, among other factors (19,20).

Having assessed the three workflows, workflow 3 was selected for all subsequent analyses based on its consistent calling of known CNVs with high confidence, along with a moderate number of total calls per sample, which suggests reasonable specificity (Additional File 2).

Sample Analyses

A total of 78 samples were analysed using workflow 3. Eight of the resulting analyses were excluded due to an excessive number of calls (>200 per sample). Seven of these were from samples that had mean read depths either <200x or >400x, suggesting that deviations in read depth from the baseline average result in a large number of false calls.
Of the remaining 70 samples, five yielded CNVs of interest. These were CNV calls that passed the confidence filter and sample-specific gene filters, and exhibited closely aligned confidence and precision values (within ~2% of each other). Notably, all but one of the shortlisted CNVs were identifiable by their reduction in read depth as visualised in IGV. The final shortlist consisted of five samples exhibiting eight variants in five genes. Among these were six intragenic deletions (22–49 kb), one whole-gene deletion (41.5 kb), and a duplication of the entire X chromosome (151.4 Mb) (Table 2).

Table 2. Shortlist of potential clinically significant CNVs

Sample	Suspected Diagnosis	CNV(s) of interest (ISCN)	Confidence	Precision	Size	Gene(s)	Visible on IGV?
CNV030	Treacher-Collins	5q32(149737222-149778716)x1	508.52*	508.52	41.5 kb	*TCOF1*	Yes
CNV044	Rett syndrome (M)	Xp22.33q28(2852798-154250851)x2	2850.75*	2816.12	151.4 Mb	Whole X chromosome	Yes
CNV151	Cornelia de Lange Syndrome	5p13.2(36955591-36985173)x1	56.39	56.39	29 kb	*NIPBL* (exons 3–10)	Yes
		5p13.2(36986173-37048880)x1	123.24	123.24	63 kb	*NIPBL* (exons 10–39)	Yes
CNV167	Ataxia telangiectasia	11q22.3(108098344-108124606)x1	76.37	76.37	26 kb	*ATM* (exons 3–13)	No
		11q22.3(108153574-108202254)x1	193.82*	193.82	48.7 kb	*ATM* (exons 26–50)	No
CNV181	Congenital Myopathy	6q13(75796202-75827279)x1	59.33	59.33	31 kb	*COL12A1* (exons 2–11)	Yes
		6q13q14.1(75890756-75912587)x1	66.84	66.84	21.8 kb	*COL12A1* (exons 47–66)	Yes

* These values represent the highest confidence with matching precision among the sample's analysis results

ISCN, International System for Human Cytogenomic Nomenclature; IGV, Integrative Genomics Viewer

To assess the reliability of detected CNVs, samples were grouped by similar mean read depth, and the results were compared to identify any CNVs shared between samples. The CNVs in *NIPBL*, *TCOF1*, and *COL12A1*, were unique to the respective samples in which they were detected, supporting their potential clinical significance (Figure 2). However, the CNVs identified in the *ATM* gene coincided with similar events in at least three other samples, all with high confidence and/or matching precision. Furthermore, no loss of read depth was visible on IGV (Additional File 1, Figure S9). These CNV calls were therefore deemed unreliable, perhaps due to inherent variability in sequencing depth of this region, or batch effects resulting from inconsistencies in the library preparation (20).

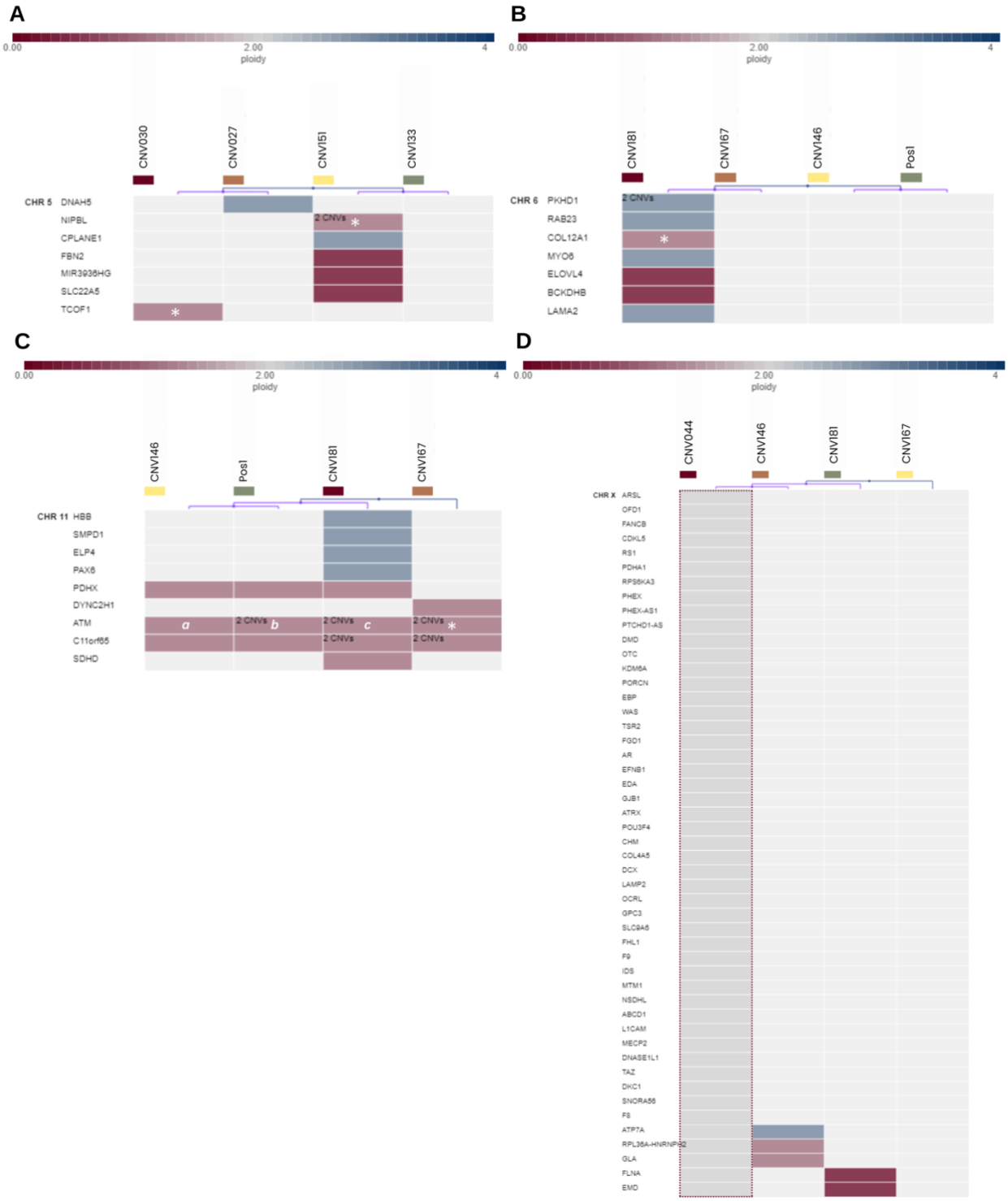


Figure 2. Heatmaps generated in Ion Reporter™ showing CNV calls in multiple samples. A) On chromosome 5, the deletion in *TCOF1* (*) is unique to sample CNV030, while the deletions in *NIPBL* (*) are unique to CNV151. B) On chromosome 6, the deletion in *COL12A1* is unique to CNV181(*). C) On chromosome 11, deletions in *ATM* are observed across four samples. D) A heatmap showing duplications (ploidy = 2) in each X-linked gene on the IDP in CNV044.

[a] deletion of 11q22.3(108099853-108236286), confidence/precision = 187.16/187.16

[b] deletion of 11q22.3(108124508-108200961), confidence/precision = 7.56/7.56; deletion of 11q22.3(108141745-108150369), confidence/precision = 2.45/2.45

[c] deletion of 11q22.3(108158275-108196114), confidence/precision = 62.56/62.56; deletion of 11q22.3(108119783-108129915), confidence/precision = 30.89/30.89

Variant interpretation and classification

CNV030—Deletion of the TCOF1 gene

Sample CNV030 corresponds to a patient with clinical features consistent with Treacher Collins syndrome. The analysis detected a deletion of the entire *TCOF1* gene (41.5 kb) (Figure 3), called with high confidence and matching precision. The decrease in read depth across this gene was clearly visible on IGV (Additional File 1, Figure S6). Applying the ACMG-ClinGen guidelines, this CNV obtained a total score of 1 due to the loss of a curated HI gene and its consistency with the Treacher Collins syndrome phenotype, thus classifying it as pathogenic. Although this syndrome is most commonly caused by sequence variants in *TCOF1* (21), four cases involving intragenic deletions have been reported in ClinVar (VCV003246423, VCV001072690.7, VCV000543946.8, VCV003246422.1).

Figure 3. Image from VarSome genome browser showing the full *TCOF1* deletion (41.5 kb) in sample CNV030.

CNV044—Unexpected whole X chromosome duplication

Sample CNV044 is that of a male patient with a clinical suspicion of Rett syndrome. The analysis detected a 151-Mb duplication of the whole X chromosome. While the confidence and precision scores were not concordant, the high confidence (2850.75) warranted further investigation. Inspection of the coverage data in IGV (compared to another male sample of similar mean read depth) revealed a consistent gain of read depth across all X-linked IDP genes, reaching *ARSL* at one end (Xp22) and *F8* at the other (Xq28). The copy number of *MECP2*, the gene typically associated with Rett Syndrome, was consistent with the rest of the genes on the X chromosome (Additional File 1, Figure S7). These findings suggest a whole X-chromosome duplication, resulting in XXY or Klinefelter syndrome.

The clinical features available for this case, namely developmental regression, aggressive behaviour, and dysmorphic features, are not consistent with the typical presentation of Klinefelter syndrome. An X chromosome duplication warrants an ACMG-ClinGen score of 1 due to its known association with Klinefelter Syndrome. Nevertheless, further clinical assessment and validation of the duplication is required.

CNV151—Adjacent intragenic deletions in the NIPBL gene

Sample CNV151 corresponds to a patient with a suspected diagnosis of Cornelia de Lange syndrome. The analysis detected two CNVs in the *NIPBL* gene with high confidence scores. The first deletion spans exons 3 to 10 (29 kb) and the second exons 10 to 39 (63 kb); together the two deletions remove the majority of the coding sequence (Figure 4). The loss of read depth in this region was visualised in IGV (Additional File 1, Figure S8). Combining these calls, the full 92 kb deletion was scored 1 and classified as pathogenic based on the large overlap with an established HI gene as well as the phenotype being highly specific to the gene. Furthermore, at least six similar or smaller deletions in *NIPBL* have been reported as pathogenic on ClinVar (VCV000583693.1, VCV001075076.7, VCV000543053.7, VCV000832382.5, VCV001460237.4, VCV003246419.1).

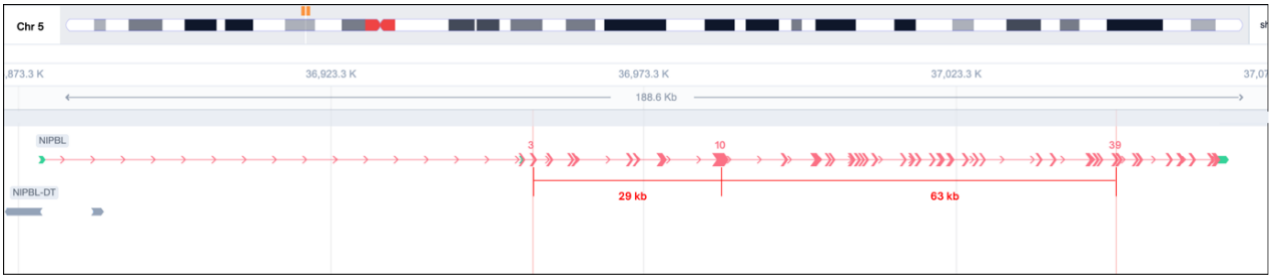


Figure 4. Image from VarSome genome browser showing two adjacent deletion calls in the *NIPBL* gene in sample CNV151. Exon numbers (in red) are indicated at the breakpoints.

CNV181—Disjointed deletion calls in the COL12A1 gene

CNV181 is a sample from a patient presenting with features of congenital myopathy. After filtering for 26 myopathy and dystrophy-related genes, the analysis produced two high-confidence deletions in the *COL12A1* gene, spanning exons 2–11 and 47–66, respectively (Figure 5). A third deletion with lower confidence scores (confidence = 9.83, precision = 9.83) was also detected, spanning exons 21–25.

Inspection of the coverage in IGV revealed reductions in read depth across the gene, suggesting that the whole gene may, in fact, be deleted (Additional File 1, Figure S10). Classification of the CNV was thus done under this assumption. It is unclear why the deletion was called in three parts and excluded many exons, but one theory is that the gene is subject to inconsistent depth of coverage due to instability of the sequencing reaction in that region, for example, due to homopolymers or suboptimal binding of primers.

Figure 5. Image from VarSome genome browser showing three deletion calls in the *COL12A1* gene in sample CNV181. The dotted red line indicates a deletion with confidence <10. Exon numbers are indicated at the breakpoints.

Mutations in the *COL12A1* gene are associated with Bethlem Myopathy 2, also known as myopathic-type Ehlers-Danlos syndrome (OMIM 616471). At least three intragenic deletions in *COL12A1* have been reported on ClinVar as pathogenic (VCV001459613.6, VCV001455959.5 and VCV001073075.8). The patient was said to exhibit some features consistent with Bethlem Myopathy 2, including delayed motor milestones, generalised muscle weakness, hypotonia, and facial dysmorphisms such as a high-arched palate and micrognathia. However, other prominent characteristics of Bethlem Myopathy 2—namely, brisk reflexes, joint hyperlaxity, joint contractures, and kyphoscoliosis—were not noted. Considering these clinical observations, alongside the lack of HI curation for *COL12A1*, this deletion was scored 0.7 and thus classified as a variant of unknown significance (VUS).

False positives, technical artefacts, and common variants

Across the 70 samples analysed, an average of 24 CNVs (range, 14–72) were called per sample. The samples had an average of 14 high-confidence calls (confidence >20) (range, 12–52); an average of 7 concordant calls (i.e., matching confidence and precision scores) (range, 0–47); and an average of 2 calls that were both confident and concordant (range, 0–10) (Figure 6A).

Concordant CNVs, even with lower confidence values, were considered "true calls," some of which are expected to be common population variants. On the other hand, non-concordant calls were attributed to technical artefacts, either from inconsistencies in sequencing data or limitations of the CNV-calling algorithm.

Analysis results appeared to vary depending on the sequencing coverage of the sample. Samples with read depths in the 200–400x range (n=29), consistent with the baseline design, performed better than those outside of this range with a lower median and narrower range of confident true calls (Figure 6B). Samples with mean read depth <200 (n=14) or >400 (n=11) had a wider range of calls; a few samples had a very high number of confident and concordant calls (e.g. 50) which complicated the manual shortlisting process

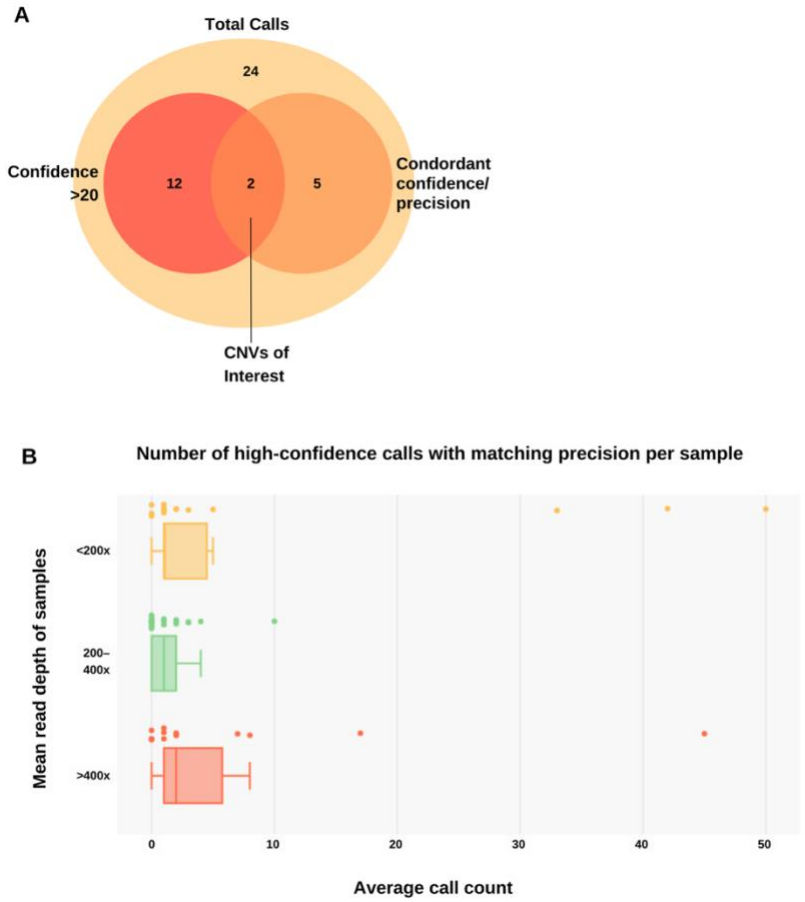


Figure 6. Average call counts per sample and effect of samples' read depth on analysis specificity. A) Venn diagram showing average call counts per sample, including those with confidence >20, those with concordant confidence and precision, and those with both (CNVs of interest/shortlisted CNVs). B) Number of calls that are confident and concordant, by mean read depth categories. Mean read depth (200–400x) in green, is that of the selected baseline used for all analyses. Each dot represents the number of calls in a single sample.

Furthermore, many CNV calls were shared across multiple samples, with some present in nearly all analyses. These represent regions in which CNV calls are unreliable, perhaps due to technical artefacts. Shared calls with identical genomic coordinates frequently had a copy number of 0 (Additional File 3), suggesting that they stem from areas of intrinsically decreased depth of sequencing. Similarly, shared calls with a copy number of 3 or more may be a result of an increased sequencing depth in that region.

DISCUSSION

This study resulted in the successful creation of a custom baseline to be used in the analyses of samples sequenced using the IDP. The baseline was incorporated into a CNV-detection workflow that identified all known CNVs (5/5) in the validation set. The workflow was used to analyse 70 samples for clinically significant CNVs. Pathogenic CNVs were detected in three samples and a VUS in another. The analysis of CNVs therefore increased the diagnostic yield by approximately 4.3% (3/70). The findings of this study demonstrate the ability of this type of NGS analysis to complement existing diagnostic methods, albeit with some shortcomings.

A custom baseline made up of numerous, diverse samples is important to ensure the accuracy and specificity of CNV analyses by capturing systematic variation (12). The baseline also accounts for common variation in a population in order to exclude recurring, neutral variants from the analysis results. While the minimum number of samples required to create a baseline is 10, increasing the number of samples enhances its efficacy by averaging the read depth across target regions to better reflect a true population average. Analyses performed using workflow 3, which incorporated a baseline with a defined range of mean read depths, demonstrated improved specificity. Samples within this mean read depth range exhibited fewer false calls and more reliable CNV detection overall. However, it is unknown whether this compromises the sensitivity of the workflow, which is a trade-off seen in other studies (22,23). Future work will have to include a larger validation set to measure the sensitivity and specificity of workflows using various baselines.

Filtering the CNV results by gene makes for an efficient approach to detect potentially clinically significant variants. However, this method is prone to confirmation bias, where a

high-confidence call in a known disease-associated gene can be assumed to be disease-causing. To avoid this, CNVs of interest should be evaluated within the context of all calls generated, specifically focusing on the confidence scores. Testing of the validation samples showed that the true clinically significant variant is always among the highest-scoring (confidence) CNV calls of each analysis. Lower-scoring CNV calls are more likely to be technical artefacts. Furthermore, all true deletions had concordant confidence and precision scores. This is supported by a validation set (n=31) from an Ion Torrent™ application note, in which all deletions had identical confidence and precision scores, while duplications and whole chromosome imbalances (e.g., Klinefelter syndrome) had mismatched confidence and precision scores (11). Therefore, prioritizing deletions with high confidence and concordant confidence and precision scores in the context of all unfiltered results is key for identifying true, potentially clinically significant CNVs.

Another effective strategy for assessing the reliability of CNVs is to compare the relevant CNV calls with those of multiple other samples. This was exemplified in the case of CNV161, where similar high-confidence CNVs in the *ATM* gene were observed across at least three other samples. The recurrence of similar CNVs in the same disease-related gene across unrelated samples suggests that calls in this region are unreliable. These false calls are possibly caused by deviations in read depth due to sequence polymorphisms disrupting PCR primer binding, which hinders amplification of the affected interval (24). Comparing CNVs across multiple analyses helps to flag problematic regions, reducing the risk of misinterpreting technical artefacts as clinically significant findings.

An unexpected finding was the whole X chromosome duplication in case CNV044, which corresponds to a male patient presenting with developmental regression, aggressive behaviour and dysmorphic features. Based on these clinical observations, the case was referred for testing for Rett syndrome (OMIM 312750), which is caused by mutations in the *MECP2* gene. An extra X chromosome in males causes Klinefelter syndrome/XXY syndrome (25). Klinefelter syndrome is typically associated with tall stature, small testes, and low testosterone with elevated gonadotropins; however, there is also a less severe form characterised by milder, less evident symptoms, which often goes undiagnosed (25). Classic Rett syndrome is mostly seen in females as their two X chromosomes allow for one functional copy of *MECP2*, while males with *MECP2* mutations on their single X chromosome rarely survive (26). However, in boys with Klinefelter syndrome, the presence

of an additional X chromosome may provide a second functional copy of *MECP2*. Consequently, it is not uncommon for males with Klinefelter syndrome who carry *MECP2* mutations to develop Rett-like symptoms (27). Therefore, this patient may be affected by an X chromosome duplication as well as a mutation in one copy of the *MECP2* gene. Clinical review of the patient is warranted in the context of this finding.

The investigation of the *COL12A1* deletion in sample CNV181 illustrates the importance of phenotype information as part of the classification of a CNV, especially in the case of a heterogeneous condition. The patient showed clinical features of a congenital myopathy, a broad term used for inherited muscular disorders characterised by hypotonia and weakness (28). Congenital myopathies are genetically and clinically heterogeneous and there are over 30 different genes underlying this group of disorders, with overlapping phenotypes (29). The *COL12A1* gene is associated with Bethlem Myopathy 2, also known as myopathic-type Ehlers-Danlos syndrome (OMIM 616471). Dosage curation has not been performed for this gene and therefore it does not have a ClinGen HI score (https://search.clinicalgenome.org/kb/genes/HGNC:2188). Ultimately, this deletion was classified as a VUS due to the lack of information. This status could change with the curation of a HI score for *COL12A1* or with information on the inheritance of the CNV (i.e., from family segregation analysis).

Short-read NGS methods face well-known challenges in detecting certain variant types, including small CNVs and CNVs in complex regions. For instance, highly similar sequences between disease genes and pseudogenes can lead to frequent read misalignment, affecting detection of clinically significant CNVs (30). Additionally, sequence polymorphisms interfering with PCR primer binding can lead to loss of amplification in that interval, and thus a false deletion call (24). In a pilot study by Lincoln et. al. (2021), the Ion Torrent™ sequencing workflow, with analyses performed in Ion Reporter™, detected 2 of 4 low-complexity CNVs (those affected by homopolymers or short tandem repeats) and 0 of 2 small CNVs, which they defined as less than 2 exons in size (22). Other NGS approaches, including Illumina paired-end sequencing, showed improvement in detection rate of small and/or complex CNVs, but no method could detect all types of variants. Lincoln et al. also reported that some of the ambiguous CNV calls were resolved only by long-read sequencing. Furthermore, this approach to CNV detection cannot determine CNV sizes due to the targeted panel design, which targets only exonic regions in selected genes. This was evident in sample

Pos2, where a 12 Mb deletion between *NF1* and *RAI1* was called in the analysis, while the actual event was a much smaller deletion affecting *NF1*. Conversely, analyses of samples CNV151 and CNV181 demonstrate the opposite problem: multiple smaller deletions were called within *NIPBL* and *COL12A1*, respectively, although these likely represent single, larger deletions. This fragmentation in CNV calls is likely due to inconsistent depth of coverage across the exons of the affected genes. Inaccurate sizing of a CNV can affect ACMG-ClinGen scoring and, ultimately, the classification of the CNV. These findings emphasize the importance of following up CNV calls from panel-based NGS data with alternative methods.

There are numerous scenarios which warrant confirmation of a CNV using an alternative method. These include 1) high-confidence CNVs found in an unreliable region (e.g., a gene that exhibits recurrent CNVs across samples), 2) CNVs in samples with overall aberrant CNV calling (e.g., >100 total calls), and 3) large CNVs (involving a whole gene or multiple genes) of which the actual size is unknown. For validation of CNVs, the choice of technology depends on the expected size of the CNV. Small CNVs (e.g., less than 100 kb) can be confirmed using quantitative PCR (qPCR), paired with Sanger sequencing to determine the breakpoints (31). This relatively simple and cost-effective approach could be used to confirm the deletions in *TCOF1* (41.5 kb) and *NIPBL* (92 kb). For intermediate-sized regions, such as the *COL12A1* gene (~122 kb), a more suitable technique would be a high-resolution chromosomal microarray which could confirm the deletion and approximate its size. Chromosomal microarray can also be used to confirm larger CNVs spanning multiple genes (5) or even whole chromosome aneuploidies, such as the X chromosome duplication in CNV044—although quantitative fluorescent PCR (QF-PCR) may be more practical in a diagnostics setting (32).

This study demonstrates how repurposing data from an established assay can enhance diagnostic yield at no additional cost. Specifically, samples that analysed for sequence variants can be re-examined for CNVs. The Ion AmpliSeq™ technology generates highly accurate reads with good coverage depth, enabling both the detection of sequence variants and the screening of CNVs using the read depth approach. This makes it a viable option for diagnostic laboratories in low- and middle-income countries where resources and assay choices are often limited.

A noteworthy limitation of this study is that the samples used to create the baseline were derived from patients with suspected genomic disorders. The Ion Reporter User Guide recommends that at least six normal samples, which are not affected by clinically significant CNVs, be included in the baseline (12). However, as this study used data from samples previously sequenced for diagnostic purposes, no general population samples were available. Furthermore, the use of a larger number of known positive samples is recommended for validation (9), but this was limited by the availability of IDP samples with known CNVs.

Conclusion

This study successfully created a custom baseline and CNV-workflow to be used for the analysis of samples sequenced using the IDP. Re-analysis of NGS data for CNVs increased the diagnostic yield by approximately 4.3% (3/70). This CNV-detection method thus shows promise as a complementary tool to existing diagnostic approaches. Some limitations are evident in this approach, with aberrant CNV calls and recurrent CNVs in multiple samples; however, strategies like filtering by confidence scores and cross-sample comparisons proved effective in identifying clinically significant CNVs. With refinement, this method could serve as a screening tool alongside the established sequence analyses. It is also advised that clinically significant CNVs be validated with an alternate method before being reported.

DECLARATIONS

Ethics approval

This study was approved by the Human Research Ethics Committee (HREC) Medical at the University of the Witwatersrand (certificate number M240429).

Consent for publication

n/a

Availability of data and materials

Data generated from the Ion Reporter analyses are available from the author upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

No funding was required for this study as all data was obtained from previous routine sequencing for diagnostic purposes.

Authors' contributions

CP performed all analyses and interpretation of results, including classifications of CNVs, wrote the manuscript, and created the tables and figures. FB and NM supervised the study, provided guidance and direction, confirmed classifications, and helped to edit the manuscript.

Acknowledgements

I would like to thank my supervisors for their direction and guidance throughout the year and for keeping me on track through regular meetings. I am also very grateful for their help in refining my manuscript before submission. Secondly, I thank Next Biosciences for making arrangements for me to continue with my part-time job throughout my research and for being supportive and understanding of my busy schedule. This job has allowed me to cover my own tuition fees this year. Thirdly, I thank the various staff and teachers at the NHLS, especially Maria Mudau, Robyn Kerr, and Monica Araujo, who were readily available to answer my questions, provide helpful resources, and give me advice.

REFERENCES

1. Alkan C, Kidd JM, Marques-Bonet T, Aksay G, Antonacci F, Hormozdiari F, et al. Personalized copy number and segmental duplication maps using next-generation sequencing. Nat Genet. 2009;41(10):1061–7.

2. Shaikh TH. Copy Number Variation Disorders. Curr Genet Med Rep. 2017;5(4):183–90.

3. Truty R, Paul J, Kennemer M, Lincoln SE, Olivares E, Nussbaum RL, et al. Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. Genetics in Medicine. 2019;21(1):114–23.

4. Rice AM, McLysaght A. Dosage-sensitive genes in evolution and disease. BMC Biol. 2017;15(1):78.

5. Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, et al. Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. The American Journal of Human Genetics. 2010;86(5):749–64.

6. Moreno-Cabrera JM, del Valle J, Castellanos E, Feliubadaló L, Pineda M, Brunet J, et al. Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. European Journal of Human Genetics. 2020;28(12):1645–55.

7. Zhang F, Gu W, Hurles ME, Lupski JR. Copy Number Variation in Human Health, Disease, and Evolution. Annu Rev Genomics Hum Genet. 2009;10(1):451–81.

8. Malekpour SA, Pezeshk H, Sadeghi M. PSE-HMM: genome-wide CNV detection from NGS data using an HMM with Position-Specific Emission probabilities. BMC Bioinformatics. 2017;18(1):30.

9. Singh AK, Olsen MF, Lavik LAS, Vold T, Drabløs F, Sjursen W. Detecting copy number variation in next generation sequencing data from diagnostic gene panels. BMC Med Genomics. 2021;14(1):214.

10. Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Res. 2009;19(9):1586–92.

11. Rhodes K, Hyland F, Konvicka K, Gottimukkala R, Altun G, Roth C, et al. CNV detection by Ion semiconductor sequencing. Life Technologies; 2014.

12. Life Technologies Corporation. Ion Reporter™ Software 5.20 USER GUIDE. MAN0028321. San Francisco: Thermo Fisher Scientific; 2023.

13. Adam MP, Feldman J, Mirzaa GM, Pagon RA, Wallace SE, Amemiya A, editors. GeneReviews® [Internet]. Seattle: Seattle (WA): University of Washington; 2024.

14. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24–6.

15. Riggs ER, Andersen EF, Cherry AM, Kantarci S, Kearney H, Patel A, et al. Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). Genetics in Medicine. 2020;22(2):245–57.

16. Kopanos C, Tsiolkas V, Kouris A, Chapple CE, Albarca Aguilera M, Meyer R, et al. VarSome: the human genomic variant search engine. Bioinformatics. 2019;35(11):1978–80.

17. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–7.

18. Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen — The Clinical Genome Resource. New England Journal of Medicine. 2015;372(23):2235–42.

19. Nishio S, Moteki H, Usami S. Simple and efficient germline copy number variant visualization method for the Ion AmpliSeq™ custom panel. Mol Genet Genomic Med. 2018;6(4):678–86.

20. Boeva V, Popova T, Lienard M, Toffoli S, Kamal M, Le Tourneau C, et al. Multi-factor data normalization enables the detection of copy number aberrations in amplicon sequencing data. Bioinformatics. 2014;30(24):3443–50.

21. Bowman M, Oldridge M, Archer C, O'Rourke A, McParland J, Brekelmans R, et al. Gross deletions in TCOF1 are a cause of Treacher–Collins–Franceschetti syndrome. European Journal of Human Genetics. 2012;20(7):769–77.

22. Lincoln SE, Hambuch T, Zook JM, Bristow SL, Hatchell K, Truty R, et al. One in seven pathogenic variants can be challenging to detect by NGS: an analysis of 450,000 patients with implications for clinical sensitivity and genetic test implementation. Genetics in Medicine. 2021;23(9):1673–80.

23. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, Guo H, et al. An Evaluation of Copy Number Variation Detection Tools from Whole-Exome Sequencing Data. Hum Mutat. 2014;35(7):899–907.

24. Kerkhof J, Schenkel LC, Reilly J, McRobbie S, Aref-Eshghi E, Stuart A, et al. Clinical Validation of Copy Number Variant Detection from Targeted Next-Generation Sequencing Panels. The Journal of Molecular Diagnostics. 2017;19(6):905–20.

25. Bonomi M, Rochira V, Pasquali D, Balercia G, Jannini EA, Ferlin A. Klinefelter syndrome (KS): genetics, clinical phenotype and hypogonadism. J Endocrinol Invest. 2017;40(2):123–34.

26. Leonard H, Silberstein J, Falk R, Houwink-Manville I, Ellaway C, Raffaele LS, et al. Occurrence of Rett Syndrome in Boys. J Child Neurol. 2001;16(5):333–8.

27. Vorsanova SG, Demidova IA, Kolotii AD, Kurinnaia OS, Kravets VS, Soloviev I V., et al. Klinefelter syndrome mosaicism in boys with neurodevelopmental disorders: a cohort study and an extension of the hypothesis. Mol Cytogenet. 2022;15(1):8.

28. Younger DS. Congenital myopathies. Handb Clin Neurol. 2023;195:533–561.

29. Gonorazky HD, Bönnemann CG, Dowling JJ. The genetics of congenital myopathies. Handb Clin Neurol. 2018;148:549-564.

30. Rojahn S, Hambuch T, Adrian J, Gafni E, Gileta A, Hatchell H, et al. Scalable detection of technically challenging variants through modified next-generation sequencing. Mol Genet Genomic Med. 2022;10(12):e2072.

31. Abildinova G, Abdrakhmanova Z, Tuchinsky H, Nesher E, Pinhasov A, Raskin L. Fast detection of deletion breakpoints using quantitative PCR. Genet Mol Biol. 2016;39(3):365–9.

32. Kooper AJ, Faas BH, Feenstra I, de Leeuw N, Smeets DF. Best diagnostic approach for the genetic evaluation of fetuses after intrauterine death in first, second or third trimester: QF-PCR, karyotyping and/or genome wide SNP array analysis. Mol Cytogenet. 2014;7(1):6.

ADDITIONAL FILES

Additional File 1

Figures showing changes in read depth on IGV

Additional File 2

Table S1. Details of the validation analyses

Additional File 3

Table S2. Common CNV calls across samples

Additional File 4

Formatting Guidelines from *BMC Medical Genomics*

Additional File 5

Research Proposal

Additional File 6

Ethics Clearance Certificate

Additional File 7

IDPv4 Panel Contents

ADDITIONAL FILE 1

Figures showing changes in read depth on IGV

Table of Contents

Note for Figures S1–S11

Read depth information in the sample of interest is compared with that of a sample of similar mean read depth. Selected single nucleotide positions in relevant exons show read depths (total read count)—these should be compared between samples (one atop the other) while also taking into account the samples' mean read depths (shown on the left).

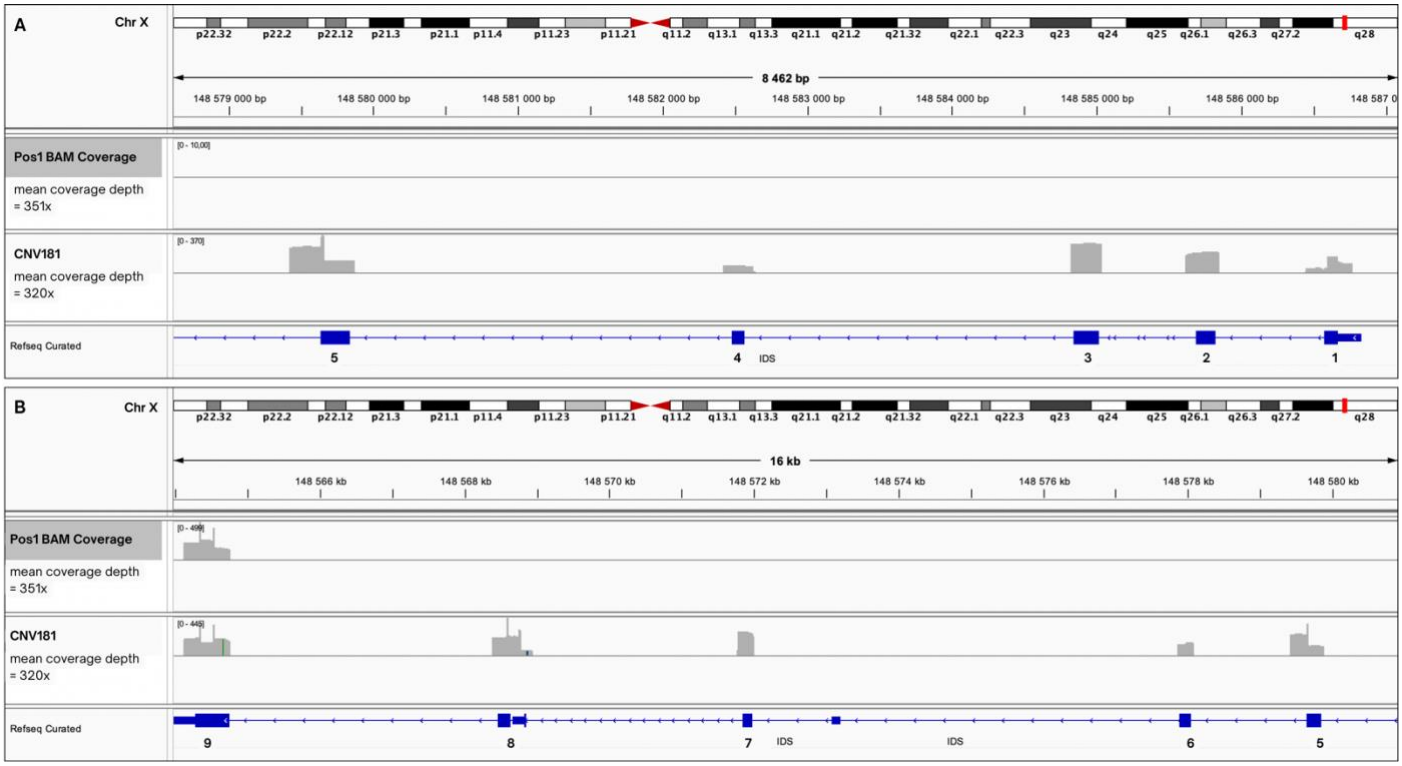


Figure S1. Pos1 BAM coverage viewed in IGV. Complete absence of coverage is seen in exons 1–5 (A) and exons 5–8 (B) of the *IDS* gene, indicating a homozygous deletion.

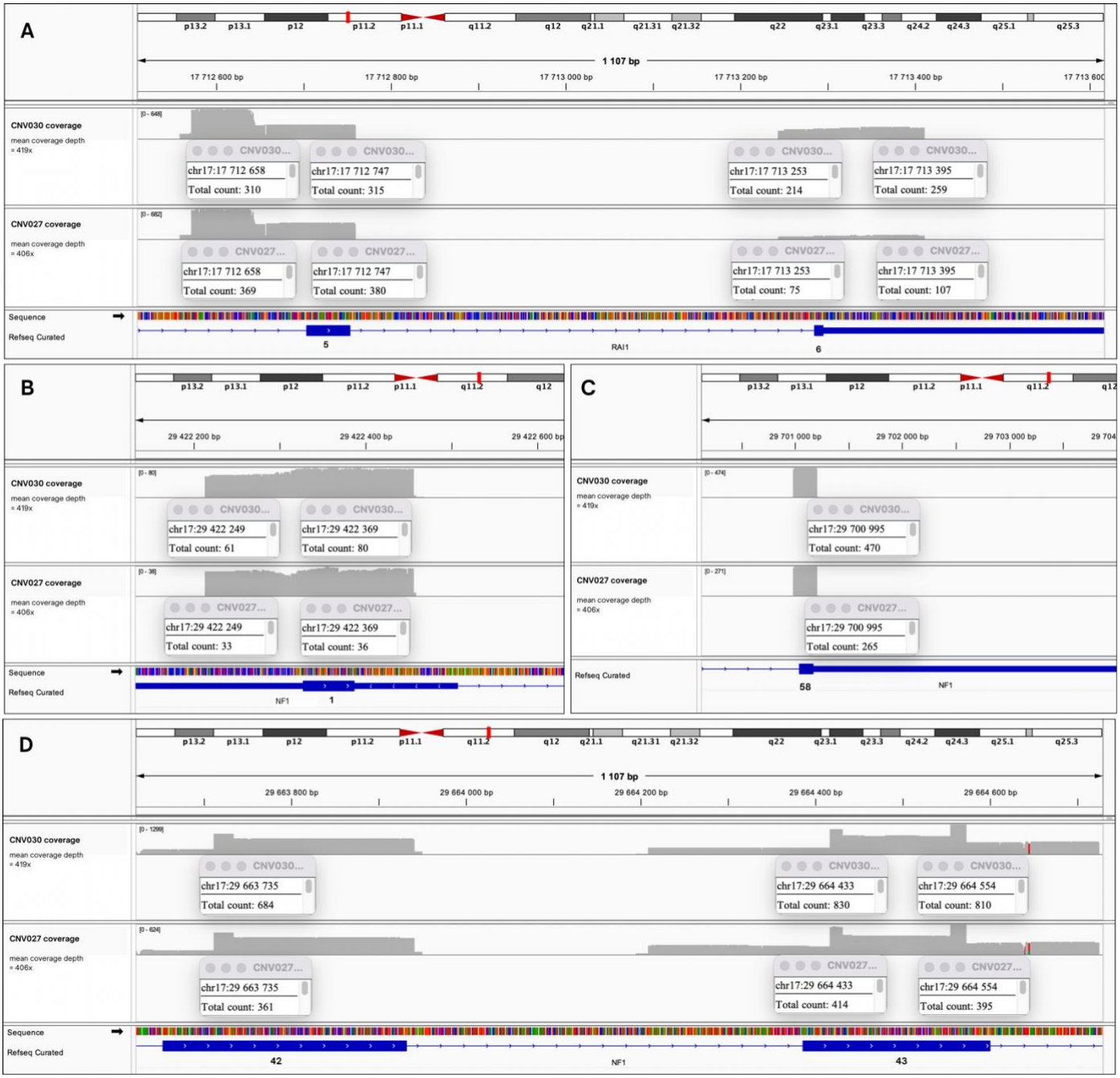


Figure S2. Pos2 BAM coverage viewed on IGV. Reduced read depth (by roughly 50%) is seen in exon 6 of *RAI1* (A) as well as throughout *NF1*, represented here by the first exon (B), the final exon (C), and intermediate exons 42 and 43 (D).

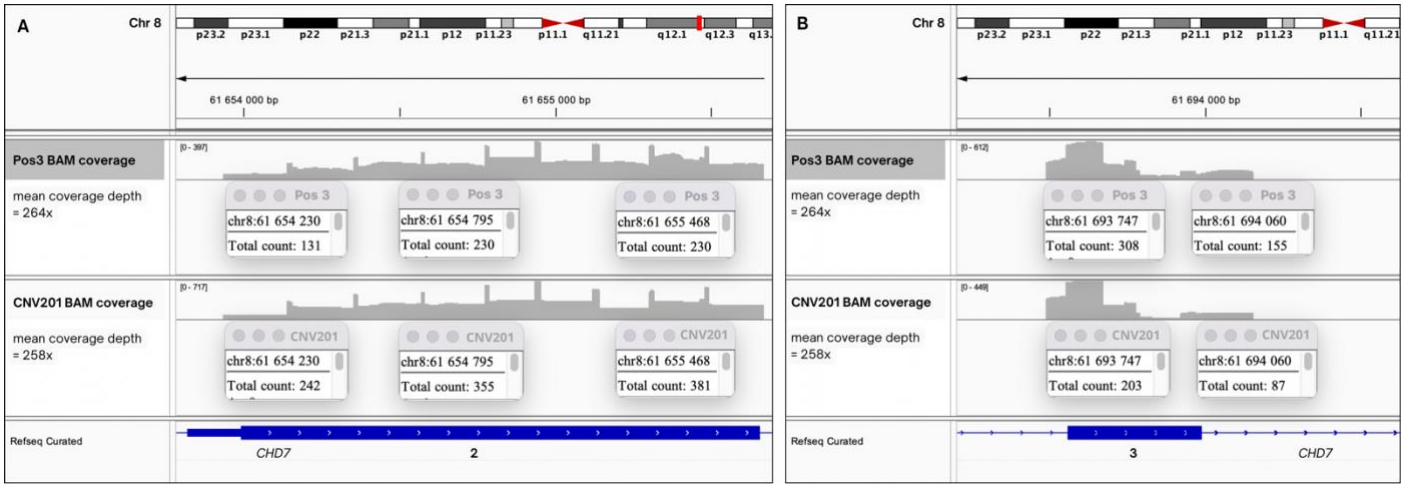


Figure S3. Pos3 BAM coverage viewed in IGV. Read depth in exon 2 of *CHD7* is decreased (roughly 50%) compared to that in CNV201 (A). For reference, the read depth of exon 3 is greater than that of CNV201 (B).

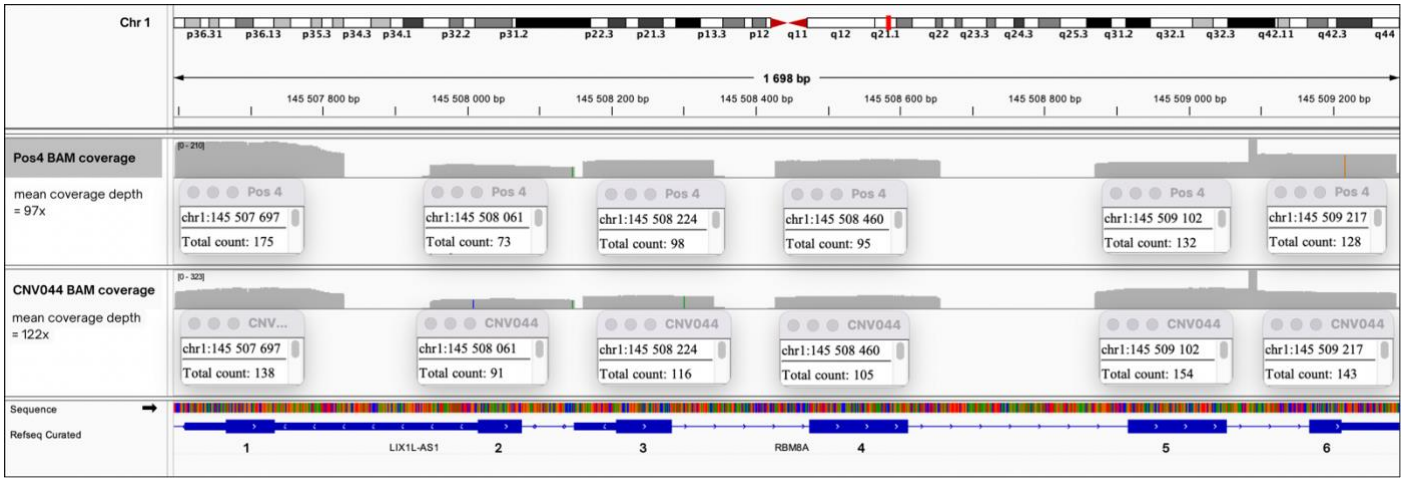


Figure S4. Pos4 BAM coverage viewed in IGV. No significant reduction in read depth is seen across the *RBM8A* gene compared to CNV044. The small decrease seen at most positions can be attributed to the lower mean read depth of Pos4 (97x compared to 122x).

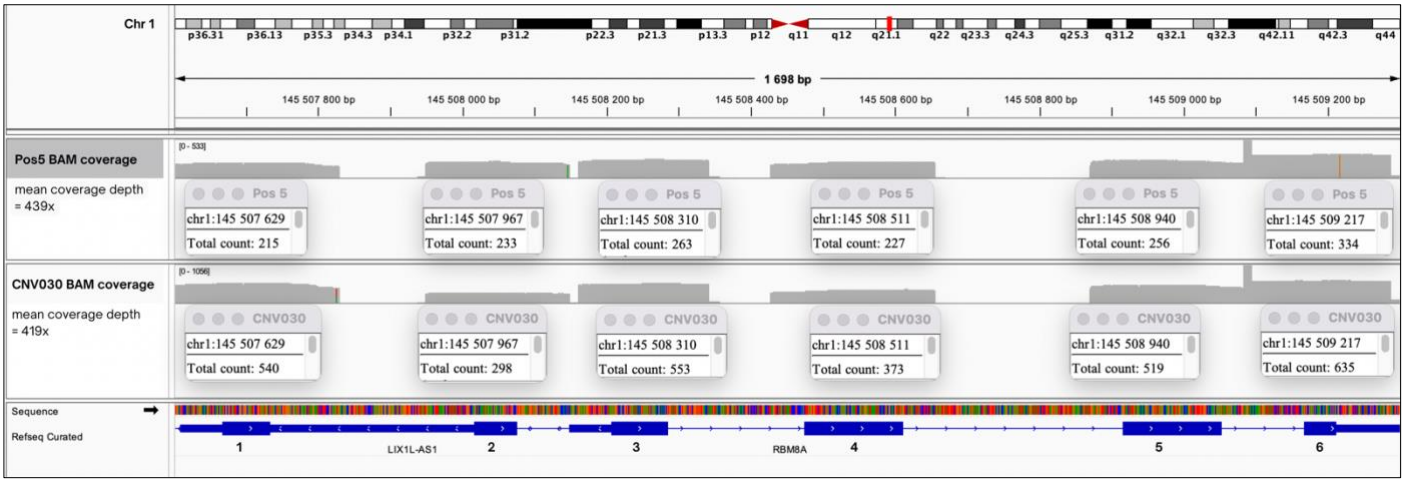


Figure S5. Pos5 BAM coverage viewed in IGV. Significant reduction in read depth can be seen across all six exons of the *RBM8A* gene.

Figure S6. CNV030 BAM coverage viewed on IGV. Reduced read depth can be seen in *TCOF1*, from exon 2 (A) to exon 25 (final exon)(B), as well as throughout the gene (C).

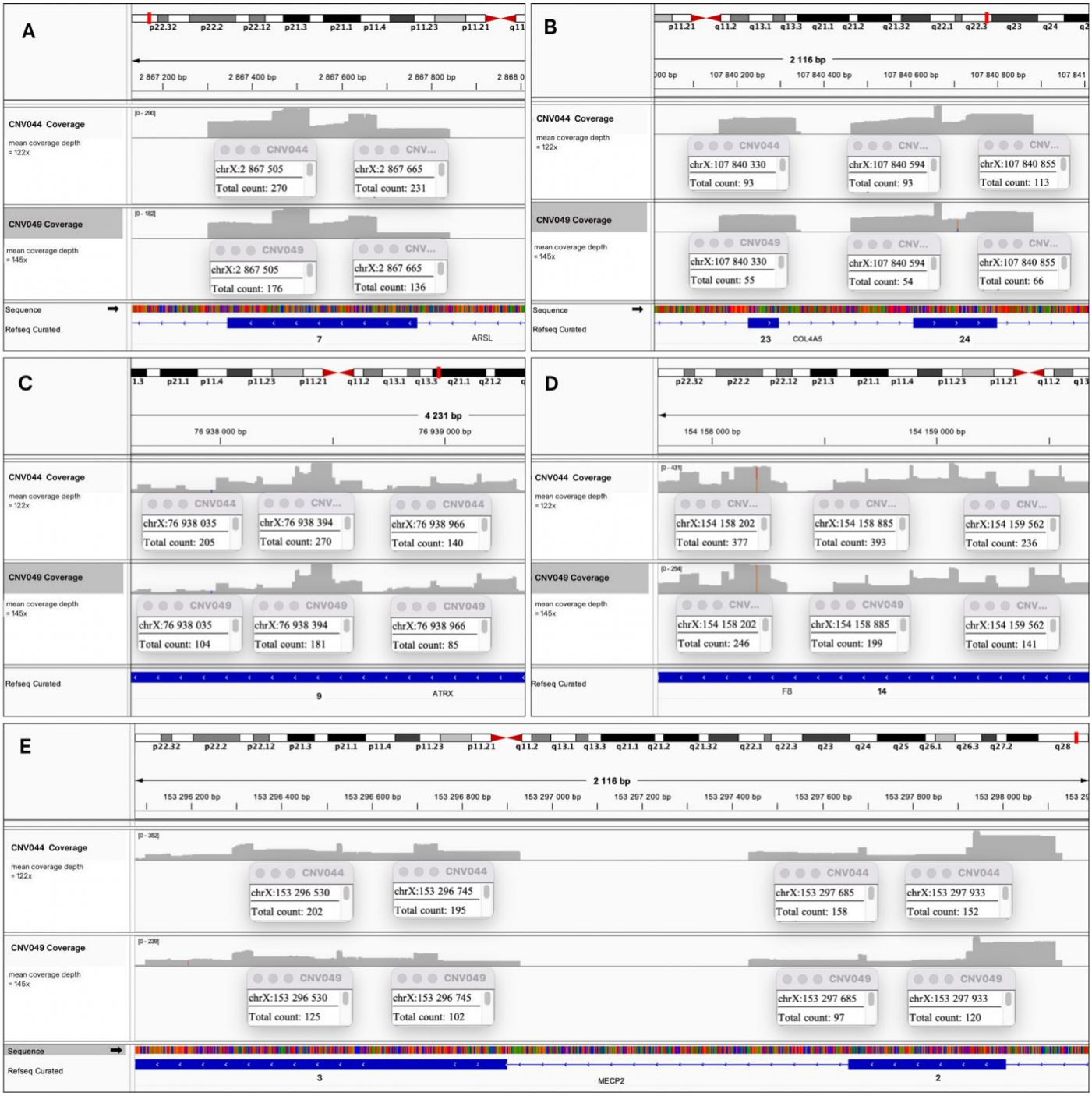


Figure S7. CNV044 coverage depth of selected genes on the X chromosome viewed on IGV. Increased read depth is seen across chromosome X, including in *ARSL* (A), *COL4A5* (B), *ATRX* (C), and *F8* (D). *MECP2* exhibits a gain in read depth consistent with the rest of the X-linked genes (E).

Figure S8. CNV151 coverage of the *NIPBL* gene viewed on IGV. Reduced read depth is seen in the first three exons (A), exon 10 (B), and up to but not including exon 39 (C).

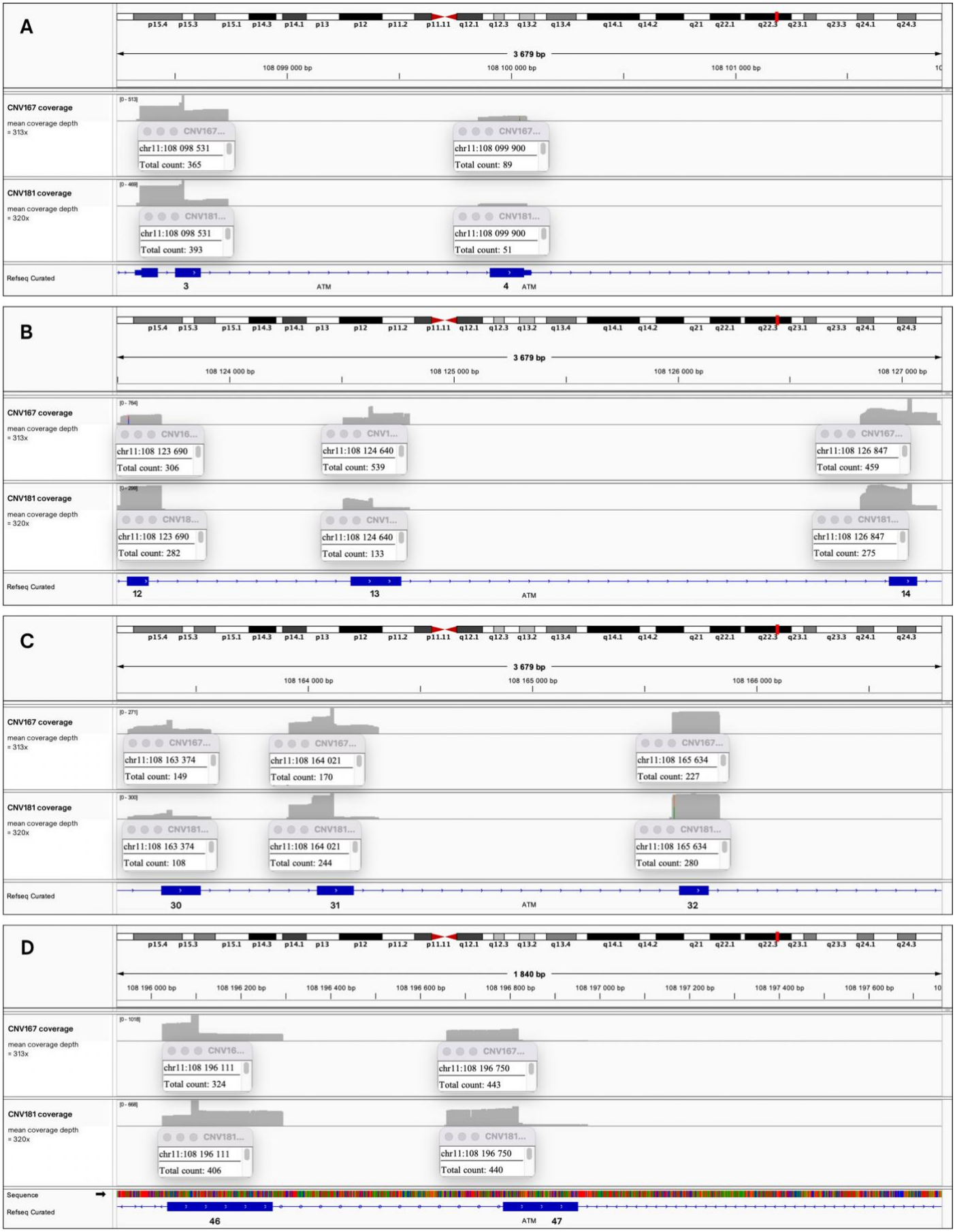


Figure S9. Coverage depth of *ATM* in sample CNV167 viewed in IGV. The expected reduced read depth is not visible at the 5' end (A), nor the middle (B, C), nor the 3' end (D) of the deletion calls in *ATM*.

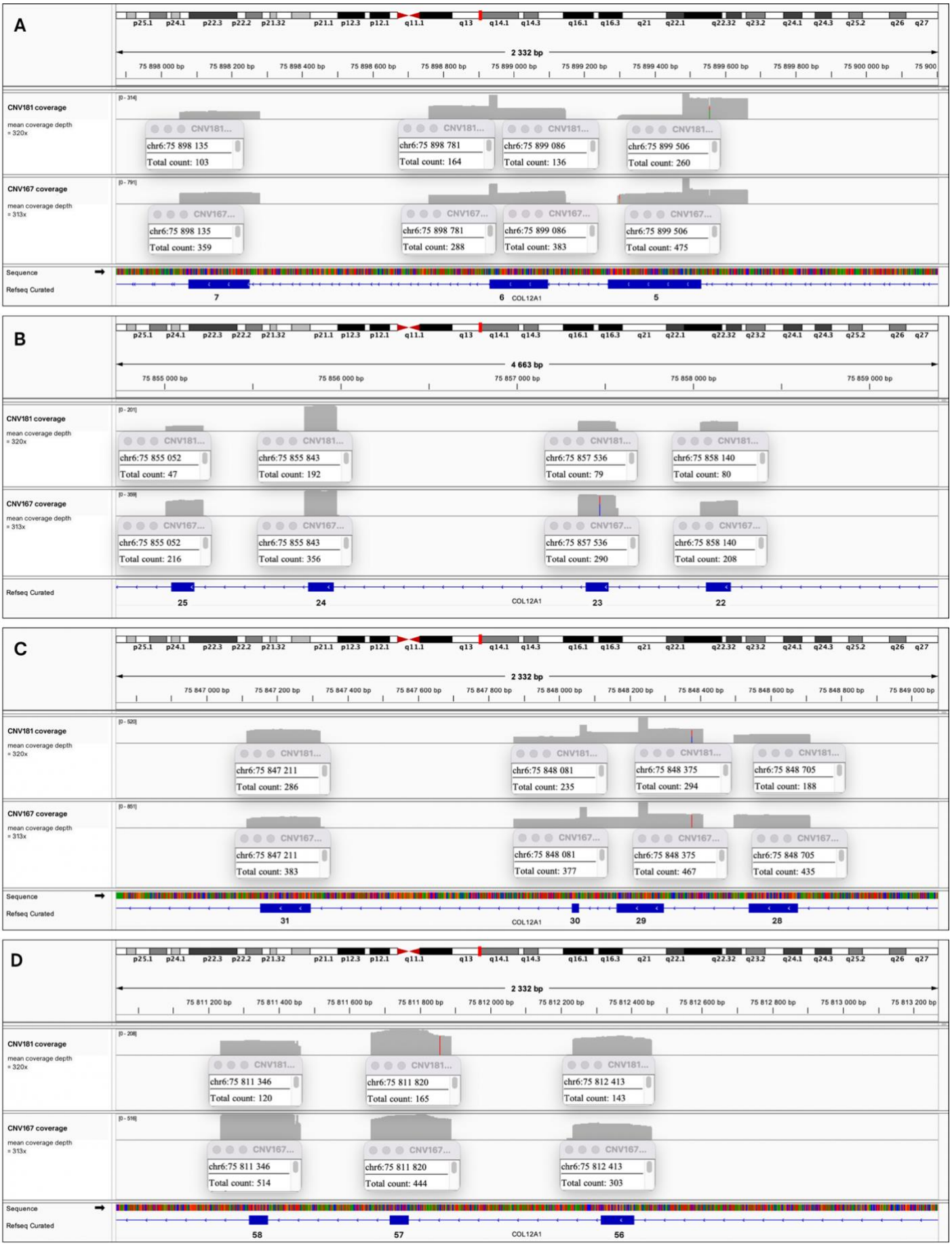


Figure S10. CNV181 coverage of the *COL12A1* gene viewed in IGV. A significant reduction in read depth is seen in the first call (here represented by exons 5, 6, and 7)(A), as well as in the lower confidence call (exons 22–25)(B). A similar loss of read depth is seen in the region between calls (here represented by exons 28–31)(C). Finally, a significant loss in read depth is observed in the last call (here represented by exons 56–58)(D).

ADDITIONAL FILE 2

Table S1. Details of the validation analyses

Sample	Known CNV Detected?	ISCN (hg19)	Size	Confidence	Precision	Visible on IGV?	# calls with confidence >20	# calls with matching confidence and precision	# calls with confidence >20 and matching precision
Pos1—homozygous 18.4 kb deletion in *IDS* (exons 1 – 8)							Pos1 mean coverage depth = 351		
Workflow 1	Yes	Xq28(148568380-148586757)x0	18.4 kb	159.88*	159.88		19	32	6
Workflow 2	Yes	Xq28(148568380-148586757)x0	18.4 kb	159.88*	159.88	Yes	20	33	6
Workflow 3	Yes	Xq28(148568380-148586757)x0	18.4 kb	328.48*	328.48		21	36	9
Pos2—1.3 Mb deletion including *NF1*							Pos2 mean coverage depth = 258		
Workflow 1	Yes	17p11.2q11.2(17713247-29701203)x1	11.9 Mb‡	475.1*	475.1		4	2	1
Workflow 2	Yes	17p11.2q11.2(17713247-29701203)x1	11.9 Mb	463.58*	463.58	Yes	19	14	12
Workflow 3	Yes	17p11.2q11.2(17713247-29701203)x1	11.9 Mb	468.02*	468.02		17	13	13
Pos3—heterozygous 11.6 kb deletion in *CHD7* (exon 2)							Pos3 mean coverage depth = 264		
Workflow 1	Yes	8q12.2(61653950-61655666)x1	1.7 kb	54.17*	54.17		1	32	3
Workflow 2	Yes	8q12.2(61653950-61655666)x1	1.7 kb	63.69*	63.69	Yes	15	37	3
Workflow 3	Yes	8q12.2(61653950-61655666)x1	1.7 kb	40.81*	40.81		14	31	2
Pos4—heterozygous 384 kb deletion including *RBM8A*							Pos4 mean coverage depth = 97		
Workflow 1	Yes	1q21.1(145414710-145509287)x1	94.6 kb†	116.71*	116.71		1	3	1
Workflow 2	Yes	1q21.1(145414710-145509287)x1	94.6 kb	126.94	104.59	No	13	34	0
Workflow 3	Yes	1q21.1(145414710-145509287)x1	94.6 kb	150.77	115.81		13	37	0
Pos5—heterozygous 3.1 Mb deletion including *RBM8A*							Pos5 mean coverage depth = 439		
Workflow 1	Yes	1q21.1(145414710-145509287)x1	94.6 kb†	76.62*	76.62		4	19	1
Workflow 2	Yes	1q21.1(145414710-145509287)x1	94.6 kb	126.24*	126.24	Yes	16	17	1
Workflow 3	Yes	1q21.1(145414710-145509287)x1	94.6 kb	99.89*	99.89		21	5	2

***** These values represent the highest confidence with matching precision among the sample's analysis results

† This size denotes the genomic interval from *HJV* to *RBM8A.* The full size of the deletion cannot be determined using IDP-based sequencing data.

‡The 1.3 Mb loss in Pos2 involving the *NF1* gene was called as a 12 Mb loss between the *RAI1* gene located at 17p11.2 and the *NF1* gene located at 17q11.2.

ADDITIONAL FILE 3

Table S2. Common CNV calls across samples (n=25)

Cytoband	CNV call (Genomic coordinates)	Observed sample count	Size	Genes
1p21.1	1p21.1(103363671-103364479)x0	4	808 bp	*COL11A1*
	1p21.1(103463774-103548621)x3	1*‡	84 kb	*COL11A1*
	1p21.1(103385789-103455156)x3	1‡	69 kb	*COL11A1*
	1p21.1(103453124-103488463)x3	1*‡	35 kb	*COL11A1*
	1p21.1(103347379-103573785)x3	1*‡	226 kb	*COL11A1*
	Total	**8**		
1p36.22	1p36.22(11167467-11319472)x0	**18***	152 kb	*MTOR*
1p36.33	1p36.33(2160084-2238368)x0	3*†	78 kb	*SKI*
	1p36.33p36.22(2237363-11994980)x1	1‡	9.7 Mb	*SKI, PEX10, MTOR, MTOR-AS1, PLOD1*
	Total	**4**		
1q21.1	1q21.1(145414710-145416941)x1	3‡	2 kb	*NBPF20, NBPF19, NBPF10, HJV*
	1q21.1(145414710-145507829)x1	2*‡	93 kb	*NBPF20, NBPF19, NBPF10, HJV, LIX1L-AS1, RBM8A*
	1q21.1(145414710-145509287)x1§	2*†	95 kb	*NBPF20, NBPF19, NBPF10, HJV, LIX1L-AS1, RBM8A*
	1q21.1(145414710-145507829)x3	2	93 kb	*NBPF20, NBPF19, NBPF10, HJV, LIX1L-AS1, RBM8A*
	1q21.1q21.2(145414710-149897966)x3	3‡	4.5 Mb	*NBPF20, NBPF19, NBPF10, HJV, LIX1L-AS1, RBM8A, SF3B4*
	Total	**12**		
1q22	1q22(155870085-155880787)x0	**21***	11 kb	*RIT1*
2p21	2p21(44502541-44547828)x0	19*	45 kb	*SLC3A1, PREPL*
	2p21(47690102-47710093)x1	1‡	20 kb	*MSH2*
	2p21(47635476-47693967)x1	1*‡	59 kb	*MSH2*
	Total	**21**		
2p23.3	2p23.3(25457126-25536924)x0	16*	80 kb	*DNMT3A*
	2p23.3(25457126-26781616)x0	3*†	1.3 Mb	*DNMT3A, OTOF*
	Total	**19**		
2q37.3	2q37.3(241808225-241818289)x0	**21*†**	10 kb	*AGXT*
6p21.33	6p21.33(31827435-31830594)x0	**21***	3 kb	*NEU1*
9p23	9p23(12693962-12709207)x0	**21*†**	15 kb	*TYRP1, LURAP1L-AS1*
11p15.4	11p15.4(6635565-6640709)x0	18	5 kb	*TPP1*
	11p15.4p15.1(6635565-17565882)x0	3*†	10.9 Mb	*TPP1, USH1C*
	Total	**21**		
12q13.12	12q13.12(49372324-49375549)x0	**15***	3 kb	*WNT1*
12q13.3	12q13.3(57637521-57643512)x0	**21***	6 kb	*STAC3*
17q22	17q22(56769935-56811621)x0	10*	42 kb	*RAD51C*
	17q22q23.2(56769935-59534125)x0	6*	2.7 Mb	*RAD51C, TBX4*
	Total	**16**		
19q13.11	19q13.11(33321357-33359485)x0	**20***	38 kb	*SLC7A9*
22q11.21	22q11.21q12.1(21336595-26877795)x0	**3*†**	5.5 Mb	*LOC107987389, LZTR1, SMARCB1, HPS4*

* High confidence (>50)

† Similar confidence and precision (±5%)

‡ Identical confidence and precision

§ Reported positive results

ADDITIONAL FILE 4

Formatting Guidelines

BMC Medical Genomics
https://bmcmedgenomics.biomedcentral.com/submission-guidelines/preparing-your-manuscript#preparing+main+manuscript+text

Preparing your manuscript

The information below details the section headings that you should include in your manuscript and what information should be within each section.
Please note that your manuscript must include a 'Declarations' section including all of the subheadings (please see below for more information).

Title page

The title page should:

- present a title that includes, if appropriate, the study design e.g.:
  - "A versus B in the treatment of C: a randomized controlled trial", "X is a risk factor for Y: a case control study", "What is the impact of factor X on subject Y: A systematic review"
  - or for non-clinical or non-research studies a description of what the article reports
- list the full names and institutional addresses for all authors
  - if a collaboration group should be listed as an author, please list the Group name as an author. If you would like the names of the individual members of the Group to be searchable through their individual PubMed records, please include this information in the "Acknowledgements" section in accordance with the instructions below
  - Large Language Models (LLMs), such as ChatGPT, do not currently satisfy our authorship criteria. Notably an attribution of authorship carries with it accountability for the work, which cannot be effectively applied to LLMs. Use of an LLM should be properly documented in the Methods section (and if a Methods section is not available, in a suitable alternative part) of the manuscript.
- indicate the corresponding author

Abstract

The Abstract should not exceed 350 words. Please minimize the use of abbreviations and do not cite references in the abstract. Reports of randomized controlled trials should follow the CONSORT extension for abstracts. The abstract must include the following separate sections:

- **Background:** the context and purpose of the study
- **Methods:** how the study was performed and statistical tests used
- **Results:** the main findings
- **Conclusions:** brief summary and potential implications

Keywords

Three to ten keywords representing the main content of the article.

Main Text

Background

The Background section should explain the background to the study, its aims, a summary of the existing literature and why this study was necessary or its contribution to the field.

Methods

The methods section should include:

- the aim, design and setting of the study
- the characteristics of participants or description of materials
- a clear description of all processes, interventions and comparisons. Generic drug names should generally be used. When proprietary brands are used in research, include the brand names in parentheses

Results

This should include the findings of the study including, if appropriate, results of statistical analysis which must be included either in the text or as tables and figures.

Discussion
This section should discuss the implications of the findings in context of existing research and highlight limitations of the study.

Conclusions
This should state clearly the main conclusions and provide an explanation of the importance and relevance of the study reported.

List of abbreviations
If abbreviations are used in the text they should be defined in the text at first use, and a list of abbreviations should be provided.

Declarations
All manuscripts must contain the following sections under the heading 'Declarations':

- Ethics approval
- Consent for publication
- Availability of data and materials
- Competing interests
- Funding
- Authors' contributions
- Acknowledgements
- Authors' information (optional)

Ethics approval and consent to participate
Manuscripts reporting studies involving human participants, human data or human tissue must:

- include a statement on ethics approval and consent (even where the need for approval was waived)
- include the name of the ethics committee that approved the study and the committee's reference number if appropriate

Availability of data and materials
All manuscripts must include an 'Availability of data and materials' statement. Data availability statements should include information on where data supporting the results reported in the article can be found including, where applicable, hyperlinks to publicly archived datasets analysed or generated during the study. By data we mean the minimal dataset that would be necessary to interpret, replicate and build upon the findings reported in the article. We recognise it is not always possible to share research data publicly, for instance when individual privacy could be compromised, and in such instances data availability should still be stated in the manuscript along with any conditions for access.
Data availability statements can take one of the following forms (or a combination of more than one if required for multiple datasets):

- The datasets generated and/or analysed during the current study are available in the [NAME] repository, [PERSISTENT WEB LINK TO DATASETS]
- The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.
- All data generated or analysed during this study are included in this published article [and its supplementary information files].
- The datasets generated and/or analysed during the current study are not publicly available due [REASON WHY DATA ARE NOT PUBLIC] but are available from the corresponding author on reasonable request.
- Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.
- The data that support the findings of this study are available from [third party name] but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of [third party name].
- Not applicable. If your manuscript does not contain any data, please state 'Not applicable' in this section.

Competing interests
All financial and non-financial competing interests must be declared in this section.

See our editorial policies for a full explanation of competing interests. If you are unsure whether you or any of your co-authors have a competing interest please contact the editorial office.
Please use the authors initials to refer to each authors' competing interests in this section.
If you do not have any competing interests, please state "The authors declare that they have no competing interests" in this section.

Funding
All sources of funding for the research reported should be declared. If the funder has a specific role in the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript, this should be declared.

Authors' contributions
The individual contributions of authors to the manuscript should be specified in this section. Guidance and criteria for authorship can be found in our editorial policies.
Please use initials to refer to each author's contribution in this section, for example: "FC analyzed and interpreted the patient data regarding the hematological disease and the transplant. RH performed the histological examination of the kidney, and was a major contributor in writing the manuscript. All authors read and approved the final manuscript."

Acknowledgements
Please acknowledge anyone who contributed towards the article who does not meet the criteria for authorship including anyone who provided professional writing services or materials.
Authors should obtain permission to acknowledge from all those mentioned in the Acknowledgements section.
See our editorial policies for a full explanation of acknowledgements and authorship criteria.
If you do not have anyone to acknowledge, please write "Not applicable" in this section.
Group authorship (for manuscripts involving a collaboration group): if you would like the names of the individual members of a collaboration Group to be searchable through their individual PubMed records, please ensure that the title of the collaboration Group is included on the title page and in the submission system and also include collaborating author names as the last paragraph of the “Acknowledgements” section. Please add authors in the format First Name, Middle initial(s) (optional), Last Name. You can add institution or country information for each author if you wish, but this should be consistent across all authors.
Please note that individual names may not be present in the PubMed record at the time a published article is initially included in PubMed as it takes PubMed additional time to code this information.

Authors' information
This section is optional.
You may choose to use this section to include any relevant information about the author(s) that may aid the reader's interpretation of the article, and understand the standpoint of the author(s). This may include details about the authors' qualifications, current positions they hold at institutions or societies, or any other relevant background information. Please refer to authors using their initials. Note this section should not be used to describe any competing interests.

References
Examples of the Vancouver reference style are shown below.
See our editorial policies for author guidance on good citation practice

Web links and URLs: All web links and URLs, including links to the authors' own websites, should be given a reference number and included in the reference list rather than within the text of the manuscript. They should be provided in full, including both the title of the site and the URL, as well as the date the site was accessed, in the following format: The Mouse Tumor Biology Database. http://tumor.informatics.jax.org/mtbwi/index.do. Accessed 20 May 2013. If an author or group of authors can clearly be associated with a web link, such as for weblogs, then they should be included in the reference.

Example reference style:
Article within a journal
Smith JJ. The world of science. Am J Sci. 1999;36:234-5.
Article within a journal (no page numbers)
Rohrmann S, Overvad K, Bueno-de-Mesquita HB, Jakobsen MU, Egeberg R, Tjønneland A, et al. Meat consumption and mortality - results from the European Prospective Investigation into Cancer and Nutrition. BMC Medicine. 2013;11:63.
Article within a journal by DOI

Slifka MK, Whitton JL. Clinical implications of dysregulated cytokine production. Dig J Mol Med. 2000; doi:10.1007/s801090000086.

Article within a journal supplement

Frumin AM, Nussbaum J, Esposito M. Functional asplenia: demonstration of splenic activity by bone marrow scan. Blood 1979;59 Suppl 1:26-32.

Book chapter, or an article within a book

Wyllie AH, Kerr JFR, Currie AR. Cell death: the significance of apoptosis. In: Bourne GH, Danielli JF, Jeon KW, editors. International review of cytology. London: Academic; 1980. p. 251-306.

OnlineFirst chapter in a series (without a volume designation but with a DOI)

Saito Y, Hyuga H. Rate equation approaches to amplification of enantiomeric excess and chiral symmetry breaking. Top Curr Chem. 2007. doi:10.1007/128_2006_108.

Complete book, authored

Blenkinsopp A, Paxton P. Symptoms in the pharmacy: a guide to the management of common illness. 3rd ed. Oxford: Blackwell Science; 1998.

Online document

Doe J. Title of subordinate document. In: The dictionary of substances and their effects. Royal Society of Chemistry. 1999. http://www.rsc.org/dose/title of subordinate document. Accessed 15 Jan 1999.

Online database

Healthwise Knowledgebase. US Pharmacopeia, Rockville. 1998. http://www.healthwise.org. Accessed 21 Sept 1998.

Supplementary material/private homepage

Doe J. Title of supplementary material. 2000. http://www.privatehomepage.com. Accessed 22 Feb 2000.

University site

Doe, J: Title of preprint. http://www.uni-heidelberg.de/mydata.html (1999). Accessed 25 Dec 1999.

FTP site

Doe, J: Trivial HTTP, RFC2169. ftp://ftp.isi.edu/in-notes/rfc2169.txt (1999). Accessed 12 Nov 1999.

Organization site

ISSN International Centre: The ISSN register. http://www.issn.org (2006). Accessed 20 Feb 2007.

Dataset with persistent identifier

Zheng L-Y, Guo X-S, He B, Sun L-J, Peng Y, Dong S-S, et al. Genome data from sweet and grain sorghum (Sorghum bicolor). GigaScience Database. 2011. http://dx.doi.org/10.5524/100012.

Preparing figures

When preparing figures, please follow the formatting instructions below.

- Figures should be numbered in the order they are first mentioned in the text, and uploaded in this order. Multi-panel figures (those with parts a, b, c, d etc.) should be submitted as a single composite file that contains all parts of the figure.
- Figures should be uploaded in the correct orientation.
- Figure titles (max 15 words) and legends (max 300 words) should be provided in the main manuscript, not in the graphic file.
- Figure keys should be incorporated into the graphic, not into the legend of the figure.
- Each figure should be closely cropped to minimize the amount of white space surrounding the illustration. Cropping figures improves accuracy when placing the figure in combination with other elements when the accepted manuscript is prepared for publication on our site. For more information on individual figure file formats, see our detailed instructions.
- Individual figure files should not exceed 10 MB. If a suitable format is chosen, this file size is adequate for extremely high quality figures.
- **Please note that it is the responsibility of the author(s) to obtain permission from the copyright holder to reproduce figures (or tables) that have previously been published elsewhere.** In order for all figures to be open access, authors must have permission from the rights holder if they wish to include images that have been published elsewhere in non open access journals. Permission should be indicated in the figure legend, and the original source included in the reference list.

Preparing tables

When preparing tables, please follow the formatting instructions below.

- Tables should be numbered and cited in the text in sequence using Arabic numerals (i.e. Table 1, Table 2 etc.).

- Tables less than one A4 or Letter page in length can be placed in the appropriate location within the manuscript.
- Tables larger than one A4 or Letter page in length can be placed at the end of the document text file. Please cite and indicate where the table should appear at the relevant location in the text file so that the table can be added in the correct place during production.
- Larger datasets, or tables too wide for A4 or Letter landscape page can be uploaded as additional files. Please see [below] for more information.
- Tabular data provided as additional files can be uploaded as an Excel spreadsheet (.xls) or comma separated values (.csv). Please use the standard file extensions.
- Table titles (max 15 words) should be included above the table, and legends (max 300 words) should be included underneath the table.
- Tables should not be embedded as figures or spreadsheet files, but should be formatted using 'Table object' function in your word processing program.
- Color and shading may not be used. Parts of the table can be highlighted using superscript, numbering, lettering, symbols or bold text, the meaning of which should be explained in a table legend.
- Commas should not be used to indicate numerical values.

If you have any questions or are experiencing a problem with tables, please contact the customer service team at info@biomedcentral.com.

Preparing additional files

As the length and quantity of data is not restricted for many article types, authors can provide datasets, tables, movies, or other information as additional files.
All Additional files will be published along with the accepted article. Do not include files such as patient consent forms, certificates of language editing, or revised versions of the main manuscript document with tracked changes. Such files, if requested, should be sent by email to the journal's editorial email address, quoting the manuscript reference number. Please do not send completed patient consent forms unless requested.
Results that would otherwise be indicated as "data not shown" should be included as additional files. Since many web links and URLs rapidly become broken, BioMed Central requires that supporting data are included as additional files, or deposited in a recognized repository. Please do not link to data on a personal/departmental website. Do not include any individual participant details. The maximum file size for additional files is 20 MB each, and files will be virus-scanned on submission. Each additional file should be cited in sequence within the main body of text.
If additional material is provided, please list the following information in a separate section of the manuscript text:

- File name (e.g. Additional file 1)
- File format including the correct file extension for example .pdf, .xls, .txt, .pptx (including name and a URL of an appropriate viewer if format is unusual)
- Title of data
- Description of data

Additional files should be named "Additional file 1" and so on and should be referenced explicitly by file name within the body of the article, e.g. 'An additional movie file shows this in more detail [see Additional file 1]'.

ADDITIONAL FILE 5

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

Detection of Copy Number Variants Using Targeted Next-Generation Sequencing Data from the Inherited Disease Panel

Research Proposal

Carmen Phipson (1863844)
MSc(Med) Genomic Medicine
University of the Witwatersrand
Division of Human Genetics

Supervisors:
Nkateko Mayevu
Dr Fiona Baine-Savanhu

TABLE OF CONTENTS

1. Introduction

1.1. Copy number variation in the human genome

Copy number variants (CNVs) are a subtype of structural variants, namely deletions and duplications, that consist of 50 base pairs (bp) to several megabases (Mb) of DNA (Shaikh, 2017). Within this size range, these variants can occur within an exon, encompass multiple exons, or span entire genes. CNVs contribute a substantial amount of genomic variability between individuals, accounting for more variation than single nucleotide polymorphisms (SNPs) and insertion-deletions (indels) combined (Sudmant et al., 2015).

Besides contributing to normal variation, CNVs have also been linked to genomic disorders (Stankiewicz & Lupski, 2010). Using a gene panel for hereditary cancer syndromes, cardiovascular, neurological, and paediatric disorders, Truty et al. (2019) showed in a study cohort of >143 000 people that CNVs comprised almost 10% of all clinically significant results (i.e., pathogenic or likely pathogenic).

The primary mechanism behind CNV-mediated disease pathogenesis is the alteration in copy number of one or more dosage-sensitive genes within large microdeletions or microduplications (typically 1–3 Mb)(Shaikh, 2017). This is true for 22q11.2 deletion syndrome, Williams-Beuren syndrome, Prader-Willi and Angelman syndrome, Charcot-Marie tooth disease type 1/hereditary neuropathy with pressure palsies, and Sotos syndrome (Shaikh, 2017). Alternatively, an intragenic CNV may mimic a disruptive point mutation, effectively deactivating the gene and causing haploinsufficiency. Notably, exonic or multi-exonic CNVs within dosage-sensitive genes constitute a significant proportion of mutations in monogenic disorders (Shaikh, 2017). Deletions and duplications can also disrupt a regulatory element in a non-coding region, affecting the expression of a gene (Harel & Lupski, 2018). Genes affected negatively by CNVs are often those involved in development; they are highly conserved and do not tolerate changes in gene dosage (Rice & McLysaght, 2017).

1.2. CNV detection

1.2.1 The diagnostic standard

Since recognizing CNVs as a cause of several hereditary diseases, their analysis has become an important step in a diagnostic strategy. The recommended first-tier method for CNV detection is array comparative genomic hybridization (CGH) (Miller et al., 2010); a microarray technique that uses oligonucleotide probes to scan the whole genome for imbalances. Although array CGH is sensitive, it is limited by probe size and genomic distance between probes (Miller et al., 2010). Thus, it is better suited to finding large gains or losses of DNA copies (>50 kb) and often cannot detect the exact breakpoints of these variants (Perry et al., 2008). Array CGH is the standard test for CNVs at the Division of Human Genetics, NHLS/Wits in Johannesburg, where the technique has a reliable resolution of about 200 kb (personal communication).

1.2.2 Next-generation sequencing

In the last two decades, it has become increasingly more feasible to utilize next-generation sequencing (NGS) data for the detection of CNVs. This method has the advantage of increased throughput at a lower cost compared to microarrays (Moreno-Cabrera et al., 2020), and it has a significantly higher resolution due to the calling of each nucleotide in a target region. Thus, NGS data becomes particularly useful when screening for small CNVs (affecting one or a few exons), as it can detect exact breakpoints with high accuracy (Singh et al., 2021). This is important since many hereditary diseases are associated with small deletions (70 bp–10 kb)(Zhang et al., 2009). Another important application of this data is the investigation of novel CNVs and classification according to their exact size and genomic content (Corsini et al., 2022)

Most research that relies on NGS data uses a 'read depth' approach (Malekpour et al., 2017), in which a ratio of coverage depth of target regions to a normal baseline is calculated to infer CNVs (Figure 1). Figure 1 depicts the use of read depth data to infer deletions and duplications.

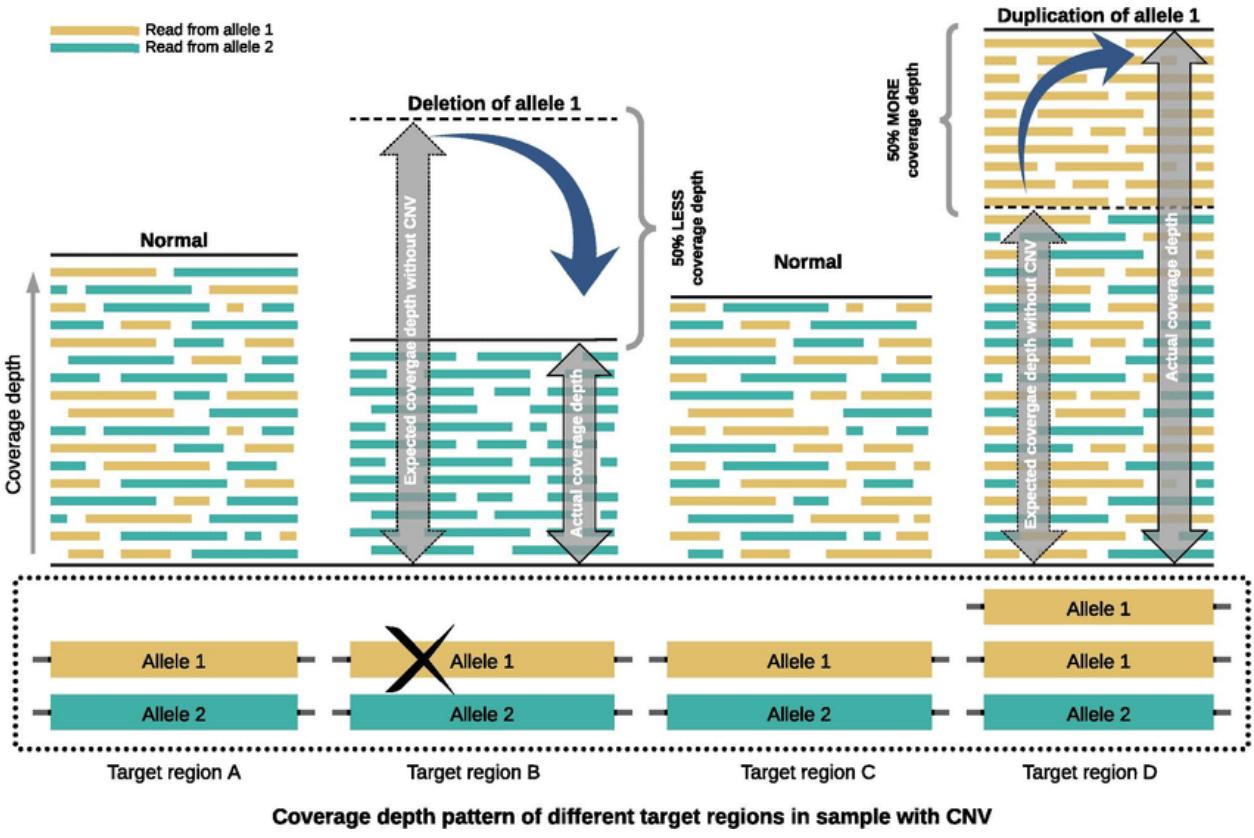


Figure 1. The principle of CNV detection using coverage depth information. The normal coverage depth may vary between regions, as shown for target regions A and C. Deletion of allele 1 in target region B reduces the coverage in that region by 50% (i.e., to 1/2), compared to the normal coverage depth for the region. Duplication of allele 1 in target region D increases the coverage in that region by 50% (i.e., to 3/2), again compared to the normal coverage depth for the region (Singh et al., 2021).

There are three major NGS approaches: whole-genome sequencing (WGS), whole-exome sequencing (WES), and targeted panel-based sequencing. The read depth method is especially useful for targeted sequence panels (see Section 1.3), as it makes use of region-specific read depths and can detect exact copy numbers, including deletions, duplications, and triplications (Singh et al., 2021).

1.3. Targeted sequencing panels

While WGS and WES provide comprehensive coverage of the genome and all exons, respectively, targeted sequencing focuses on selected genomic regions of interest. Targeted sequencing is performed using panels of a select set of genes that have known associations with a disease or phenotype. This process involves the amplification of target sequences, allowing for accurate sequencing with excellent coverage depth. With these advantages, targeted sequencing panels have become common in diagnostic labs, and are useful for detecting CNVs affecting part of an exon, a single exon, or a few exons (Moreno-Cabrera et al., 2020; Truty et al., 2019). For diagnosis of inherited diseases, targeted panels are as effective as exome sequencing, but with practical advantages, such as lower cost and reduced data processing (Saudi Mendeliome Group, 2015).

When using targeted gene sequencing for CNV detection, the quality of the results depends on a similar average read depth of the query sample and the selected pool of reference samples (Singh et al., 2021). When creating a baseline for reference, it is therefore necessary to select samples with a similar depth of coverage.

1.3.1 The inherited disease panel

The inherited disease panel (IDP) is a targeted sequencing panel custom-designed and utilized by the NHLS in Johannesburg. It includes multiple sub-panels of genes which are associated with hereditary disorders, e.g., cancer, cardiovascular diseases, ciliopathies, connective tissue disorders, differences of sex development, dermatological conditions, hearing loss, RASopathies, and gastrointestinal, haematology, metabolic, neuromuscular, ophthalmology, overgrowth, respiratory, skeletal disorders, in total comprising over 500 genes (personal communication). The resulting sequence data is typically used to detect SNPs and small indels, although it has strong potential for the investigation of CNVs as well.

2. Aims and Objectives

This project aims to identify and classify clinically significant CNVs from NGS data generated from the IDP, a custom targeted gene sequencing panel. In doing so, it can be gauged whether this method of CNV detection is applicable for routine diagnostic use at the NHLS.

Study objectives:

i) Create a custom copy number baseline using sequence data from 50 samples sequenced using the IDP.
ii) Create a CNV analysis workflow using the copy number baseline and validate it using 3–5 samples with known CNVs.
iii) Apply the CNV detection workflow to analyse up to 100 IDP-sequenced samples
iv) Use ACMG-ClinGen guidelines to classify candidate CNVs, with the use of public genomic databases, relevant literature, and patient clinical descriptions.

3. Methods

The study design is a secondary data analysis, in which IDP-sequenced samples that were previously negative for pathogenic SNVs will be investigated for CNVs.

Processing and analysis of the data will be performed in Ion Reporter (Thermo Fisher Scientific). This software is designed for analysis of data output from Ion Torrent semiconductor sequencing and is compatible with that generated by Ion AmpliSeq targeted panels, such as the IDP. The Ion AmpliSeq™ workflow, which includes a CNV module, calls copy number results down to the gene and sub-gene-level.

3.1. Data handling

Two types of data will be used in this study: 1) sequence data from up to 150 samples from patients that underwent IDP-based sequencing at the Division and 2) the corresponding clinical descriptions (patient phenotypes) where necessary. The raw sequence data, which has been de-identified, will be made available in the form of BAM files. The patient phenotypes will be accessed from the Clinical REDCap database by the project supervisors (N. Mayevu and F. Baine-Savanhu) and anonymized before being provided for use.

Access to the data has been granted by the Head of the Division of Human Genetics.

3.2. Baseline Construction

A read depth baseline will be constructed that will serve as a reference specific to IDP-sequenced samples and which is representative of normal copy number variation in the South African population. Fifty samples that were sequenced using the IDP will be selected based on mean read depth (>150X) and uniformity of coverage (similarity of coverage depth across the samples; e.g., >90%) (Corsini et al., 2022). Differences in read depth among normal samples used to create the baseline can introduce noise (Singh et al., 2021).

Therefore, a baseline created using samples of uniform coverage depth will allow for more unambiguous CNV calling in the query samples. Furthermore, the baseline samples must be those generated using the same platform as the query samples (i.e. the IDP) so that the normal read depths at each target region are comparable. In this way, the baseline will be a true, representative reference for reliable CNV calling.

Using these samples, the mean coverage depth of target genomic regions will be calculated and normalized to the total number of reads. Ion Reporter aggregates this data to establish a baseline coverage depth for each region, which serves as a reference for comparison during CNV detection in the samples of interest.

3.3. Workflow customization and validation

In Ion Reporter, a custom workflow will be created for CNV detection, as opposed to the default workflow which utilizes preset parameters and a pre-computed generic baseline. The workflow will be specific to the IDP by importing the relevant custom panel design from Ampliseq.com and specifying the target regions (Thermo Fisher, 2020). In a custom workflow, not only can a custom baseline be implemented, but specific annotations (hg19), filters (e.g., confidence score, frequency, predicted effect, etc.), and parameters (e.g., sensitivity, specificity) can be selected.

Using the 50-sample IDP-specific baseline as a reference, Ion Reporter will perform CNV detection, variant calling, and annotation. This is performed using a Hidden Markov Model, which uses read depth information compared to the baseline read depth to infer CNVs at each target region, similar to the principle depicted in Figure 1 (Life Technologies, 2014).

This workflow will be validated using five CNV-positive samples (previously detected by manual inspection of the sequence data). Called CNVs will be visually inspected using the built-in integrative genomics viewer (IGV), in which segments that exhibit copy number gain or loss are shown along the length of the chromosome.

3.4. Multi-sample analysis

The validated workflow will be applied to up to 100 IDP-sequenced samples from 2019 to 2024. Samples to be analysed will include those with a similar read depth to those used in the baseline (>150X) and those that did not have a clinically significant variant (SNV or indel) identified.

Identified variants can be filtered based on technical criteria such as read coverage, allele frequency, P-value, confidence levels, and other specific annotations. This filtering can exclude low-confidence variant calls, artefacts of sequencing, and common (normal) variants in the population so that candidate CNVs for classification can be extracted.

3.5. Classification of candidate CNVs

The American College of Medical Genetics (ACMG) in collaboration with the Clinical Genome Resource (ClinGen) project has developed professional standards for the classification and reporting of CNVs (Riggs et al., 2020). This is a quantitative, evidence-based scoring system which classifies variants as benign, likely benign, variant of unknown significance (VUS), likely pathogenic, or pathogenic. The system involves categories of evidence such as genomic content, number of genes, overlap with established/predicted haploinsufficient regions, matches with published literature or databases, and family history. Within each category, multiple options are possible, each with suggested points for addition or subtraction.

This process will involve the use of public databases, such as dbVAR, ClinVAR, and DECIPHER, and a study of relevant literature to see if a similar CNV has been previously described. The patients' clinical descriptions (phenotypes) will also be reviewed to determine a possible genotype-phenotype relationship and evaluate overlap with that described in the literature, thus supporting a classification.

3.6. Feedback of Clinically Significant Results

If a CNV is found to be pathogenic or likely pathogenic, the variant will be validated by standard diagnostic testing protocols following the study and returned to the referring clinician so that the variant (and potentially a diagnosis) can be reported to the affected patient/guardian.

4. References

Corsini, S., Pedrini, E., Patavino, C., Gnoli, M., Lanza, M., & Sangiorgi, L. (2022). An Easy-to-Use Approach to Detect CNV From Targeted NGS Data: Identification of a Novel Pathogenic Variant in MO Disease. *Frontiers in Endocrinology*, *13*. https://doi.org/10.3389/fendo.2022.874126

Harel, T., & Lupski, J. R. (2018). Genomic disorders 20 years on—mechanisms for clinical manifestations. *Clinical Genetics*, *93*(3), 439–449. https://doi.org/10.1111/cge.13146

Life Technologies. (2014). CNV detection by Ion Semiconductor sequencing. https://assets.thermofisher.com/TFS-Assets/LSG/brochures/CNV-Detection-by-Ion.pdf

Malekpour, S. A., Pezeshk, H., & Sadeghi, M. (2017). PSE-HMM: genome-wide CNV detection from NGS data using an HMM with Position-Specific Emission probabilities. *BMC Bioinformatics*, *18*(1), 30. https://doi.org/10.1186/s12859-016-1296-y

Miller, D. T., Adam, M. P., Aradhya, S., Biesecker, L. G., Brothman, A. R., Carter, N. P., Church, D. M., Crolla, J. A., Eichler, E. E., Epstein, C. J., Faucett, W. A., Feuk, L., Friedman, J. M., Hamosh, A., Jackson, L., Kaminsky, E. B., Kok, K., Krantz, I. D., Kuhn, R. M., … Ledbetter, D. H. (2010). Consensus Statement: Chromosomal Microarray Is a First-Tier Clinical Diagnostic Test for Individuals with Developmental Disabilities or Congenital Anomalies. *The American Journal of Human Genetics*, *86*(5), 749–764. https://doi.org/10.1016/j.ajhg.2010.04.006

Minikel, E. (2014). A comparison of tools for calling CNVs from sequence data. https://www.cureffi.org/2014/01/17/comparison-of-tools-for-calling-cnvs-from-sequence-data/

Moreno-Cabrera, J. M., del Valle, J., Castellanos, E., Feliubadaló, L., Pineda, M., Brunet, J., Serra, E., Capellà, G., Lázaro, C., & Gel, B. (2020). Evaluation of CNV detection tools for NGS panel data in genetic diagnostics. *European Journal of Human Genetics*, *28*(12), 1645–1655. https://doi.org/10.1038/s41431-020-0675-z

Perry, G. H., Ben-Dor, A., Tsalenko, A., Sampas, N., Rodriguez-Revenga, L., Tran, C. W., Scheffer, A., Steinfeld, I., Tsang, P., Yamada, N. A., Park, H. S., Kim, J.-I., Seo, J.-S., Yakhini, Z., Laderman, S., Bruhn, L., & Lee, C. (2008). The Fine-Scale and Complex Architecture of Human Copy-Number Variation. *The American Journal of Human Genetics*, *82*(3), 685–695. https://doi.org/10.1016/j.ajhg.2007.12.010

Rice, A. M., & McLysaght, A. (2017). Dosage-sensitive genes in evolution and disease. *BMC Biology*, *15*(1), 78. https://doi.org/10.1186/s12915-017-0418-y

Riggs, E. R., Andersen, E. F., Cherry, A. M., Kantarci, S., Kearney, H., Patel, A., Raca, G., Ritter, D. I., South, S. T., Thorland, E. C., Pineda-Alvarez, D., Aradhya, S., & Martin, C. L. (2020). Technical standards for the interpretation and reporting of constitutional copy-number variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics (ACMG) and the Clinical Genome Resource (ClinGen). *Genetics in Medicine*, *22*(2), 245–257. https://doi.org/10.1038/s41436-019-0686-8

Saudi Mendeliome Group. (2015). Comprehensive gene panels provide advantages over clinical exome sequencing for Mendelian diseases. *Genome Biology*, *16*(1), 134. https://doi.org/10.1186/s13059-015-0693-2

Shaikh, T. H. (2017). Copy Number Variation Disorders. *Current Genetic Medicine Reports*, *5*(4), 183–190. https://doi.org/10.1007/s40142-017-0129-2

Singh, A. K., Olsen, M. F., Lavik, L. A. S., Vold, T., Drabløs, F., & Sjursen, W. (2021). Detecting copy number variation in next generation sequencing data from diagnostic gene panels. *BMC Medical Genomics*, *14*(1), 214. https://doi.org/10.1186/s12920-021-01059-x

Stankiewicz, P., & Lupski, J. R. (2010). Structural Variation in the Human Genome and its Role in Disease. *Annual Review of Medicine*, *61*(1), 437–455. https://doi.org/10.1146/annurev-med-100708-204735

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stütz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., … Korbel, J. O. (2015). An integrated map of structural variation in 2,504 human genomes. *Nature*, *526*(7571), 75–81. https://doi.org/10.1038/nature15394

Truty, R., Paul, J., Kennemer, M., Lincoln, S. E., Olivares, E., Nussbaum, R. L., & Aradhya, S. (2019). Prevalence and properties of intragenic copy-number variation in Mendelian disease genes. *Genetics in Medicine*, *21*(1), 114–123. https://doi.org/10.1038/s41436-018-0033-5

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. *Annual Review of Genomics and Human Genetics*, *10*(1), 451–481. https://doi.org/10.1146/annurev.genom.9.081307.

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

R14/49 Miss Carmen Rose Phipson M240429

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)

CLEARANCE CERTIFICATE NO. M240429

NAME: (Principal Investigator)	Miss Carmen Rose Phipson
DEGREE:	Master of Science Medicine in Genomic Medicine
STUDENT/STAFF NO:	1863844
DEPARTMENT:	School of Pathology Department of Human Genetics NHLS, Braamfontein
PROJECT TITLE:	Detection of Copy Number Variants Using Targeted Next-Generation Sequencing Data from the Inherited Disease Panel
DATE CONSIDERED:	19/04/2024
DECISION:	Approved unconditionally
CONDITIONS:	
NOTE:	Ethics Online System ref no MED24-02-387
SUPERVISOR:	Mrs Nkateko Mayevu and Dr Fiona Baine-Savanhu
APPROVED BY:	Paul Ruff Paul Ruff (Sep 3, 2024 18:33 GMT+2) Prof P Ruff, Chair, HREC (Medical)
DATE OF APPROVAL:	03/09/2024 **EXPIRY DATE:** 03/09/2029

This clearance certificate is valid for 5 years from date of approval. Extension may be applied for.

DECLARATION OF INVESTIGATORS

I/we fully understand the conditions under which I am/we are authorized to carry out the above-mentioned research and I/we undertake to ensure compliance with these conditions. Should any departure be contemplated, from the research protocol as approved, I/we undertake to resubmit the application to Committee. **I agree to submit a yearly progress report** in July each year until study is closed. Failure to submit annual report will result in ethics clearance certificate suspension. The date for annual re-certification will be one year after the date of convened meeting where the study was initially reviewed. In this case, the study was initially reviewed in April and will therefore be due in the month of April each year. Unreported changes to the application invalidate the clearance given by the HREC (Medical). Email signed copy of this ethics clearance certificate prior to commencing with the study hrec-medical.researchoffice@wits.ac.za

Principal Investigator Signature

15/09/2024
Date

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Additional File 7

NATIONAL HEALTH LABORATORY SERVICE
School of Pathology, University of the Witwatersrand
Division of Human Genetics

DIVISION OF HUMAN GENETICS

Hospital Street, Johannesburg, 2001 | PO Box 1038, Johannesburg, 2000
[T] +27 489 9223 | [M] +27 72 432 3043 | [E] human.genetics@nhls.ac.za

Inherited Disorders Panel (IDP)

Version 4

Please note:

- Requesting testing of the Inherited Disorders Panel should be based on a strong clinical suspicion of a particular disorder
- When requesting testing, please ensure that you have completed a National Health Laboratory Service test request form
- Please quote the sub-panel code that you want to request **AND** state the order in which you would like the genes to be analysed

Table of contents

Cancer Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Breast and gynae	*ATM, BARD1, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, DICER1, EPCAM, FANCC, MLH1, MSH2, MSH6, MUTYH, NF1, PALB2, PMS2, PTEN, RAD51C, RAD51D, SDHB, SDHD, STK11, TP53*
	Breast and gynae (guidelines)	*ATM, BARD1, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, EPCAM, MLH1, MSH2, MSH6, NF1, PALB2, PMS2, PTEN, RAD51D, STK11, TP53*
	Breast cancer	*ATM, BARD1, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, FANCC, MUTYH, NF1, PALB2, PTEN, RAD51D, STK11, TP53*
	Breast cancer (guidelines)	*ATM, BARD1, BRCA1, BRCA2, CDH1, CHEK2, NF1, PALB2, PTEN, STK11, TP53*
	Colorectal cancer	*APC, ATM, BLM, BMPR1A, CDH1, CHEK2, ENG, EPCAM, FLCN, GREM1, MLH1, MSH2, MSH6, MUTYH, PMS2, POLD1, POLE, PTEN, SMAD4, STK11, TP53*
	Colorectal cancer (guidelines)	*APC, BMPR1A, CHEK2, EPCAM, GREM1, MLH1, MSH2, MSH6, MUTYH, PMS2, POLD1, POLE, PTEN, SMAD4, STK11, TP53*
	Familial Gastrointestinal Stromal Tumour	*KIT, NF1, SDHA, SDHC, SDHD*
	Gastric cancer	*APC, BMPR1A, CDH1, EPCAM, KIT, MLH1, MSH2, MSH6, NF1, PMS2, SDHA, SDHB, SDHC, SDHD, SMAD4, STK11, TP53*
	Hyperparathyroidism	*CDKN1B, MEN1, RET*
	Melanoma	*BRCA1, BRCA2, CDK4, CDKN2A, MITF, PTEN, RB1, TERT, TP53*
	Myelodysplastic syndrome/leukaemia	*ATM, BLM, BRCA1, BRCA2, BRIP1, CHEK2, EPCAM, HRAS, MLH1, MSH2, MSH6, NF1, PALB2, PMS2, TERT, TP53*
	Nervous system/brain cancer	*APC, BARD1, DICER1, EPCAM, EZH2, GPC3, HRAS, MEN1, MLH1, MSH2, MSH6, NF1, NF2, PMS2, PRKAR1A, PTCH1, PTEN, RB1, TP53, TSC1, TSC2, VHL*
	Paediatric haematologic malignancies	*ATM, BLM, EPCAM, HRAS, MLH1, MSH2, MSH6, NF1, PMS2, TERT, TP53*
	Paediatric nervous system/brain tumours	*APC, DICER1, EPCAM, HRAS, MEN1, MLH1, MSH2, MSH6, NF1, NF2, PMS2, PRKAR1A, PTCH1, PTEN, RB1, TP53, TSC1, TSC2, VHL*

	Paediatric solid tumours	*APC, BLM, BMPR1A, DICER1, DIS3L2, EPCAM, FH, GPC3, HRAS, MEN1, MLH1, MSH2, MSH6, NF1, NF2, PMS2, PRKAR1A, PTCH1, PTEN, RECQL4, RET, SDHA, SDHAF2, SDHB, SDHC, SDHD, SMAD4, STK11, TP53, TSC1, TSC2, VHL, WT1*
	Pancreatic cancer	*ATM, BRCA1, BRCA2, BRIP1, CHEK2, EPCAM, FANCA, MLH1, MSH2, MSH6, PALB2, PMS2, PRSS1, RAD51C, RAD51D, TP53*
	Paraganglioma-pheochromocytoma	*FH, MEN1, NF1, RET, SDHA, SDHAF2, SDHB, SDHC, SDHD, VHL*
	Renal/urinary tract	*DICER1, DIS3L2, EPCAM, FH, FLCN, GPC3, MITF, MLH1, MSH2, MSH6, PALB2, PMS2, PTEN, SDHB, SDHC, SDHD, TP53, TSC1, TSC2, VHL, WT1*
	Sarcoma	*APC, BLM, CDKN2A, DICER1, EPCAM, FH, HRAS, KIT, MLH1, MSH2, MSH6, NF1, PMS2, PRKAR1A, PTCH1, RB1, RECQL4, SDHA, SDHB, SDHC, SDHD, TP53, TSC1, TSC2*
	Schwannomatosis	*SMARCB1; LZTR1*
	Thyroid cancer	*APC, CHEK2, DICER1, MEN1, PRKAR1A, PTEN, RET, SDHB, SDHD, TP53*
	Wilms tumour panel	*DIS3L2, GPC3, WT1*

Cardiovascular Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Aortopathy	*ACTA2, BMP1, COL1A2, COL3A1, COL5A1, COL5A2, FLNA, COL1A1, ELN, FBN1, FBN2, MYH11, SKI, SMAD2, SMAD3, SMAD4, TGFB3, TGFBR1, TGFBR2*
	Arrhythmogenic right ventricular cardiomyopathy	*DSC2, DSG2, PKP2, TMEM43*
	Barth syndrome	*TAZ*
	Brugada Syndrome	*KCNH2, KCNQ1, SCN5A*
	Catecholaminergic polymorphic ventricular tachycardia	*CASQ2, RYR2*
	Generalised arterial calcification of infancy	*ENPP1*
	Hereditary haemorrhagic telangiectasias	*ACVRL1, ENG, SMAD4*
	Hereditary transthyretin amyloidosis	*TTR*

	Hypertrophic and dilated cardiomyopathy	*ACTC1, BAG3, CAV3, LAMP2, LMNA, MYBPC3, MYH7, MYL2, MYL3, NKX2-5, TCAP*
	Hypertrophic cardiomyopathy due to glycogen storage disorder	*PRKAG2, TNNI3, TNNT2, TPM1*
	Pulmonary hypertension	*ACVRL1, BMPR1B, BMPR2, ENG, KCNK3, SMAD9, TBX4*

Ciliopathies Panel

1) Skeletal ciliopathies

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Sensenbrenner syndrome	*WDR19*
	Short rib thoracic dysplasia ± polydactyly group	*IFT80, DYNC2H1, TTC21B, WDR19, NEK1, EVC, EVC2*
	Weyers acrofacial dysostosis	*EVC2*

2) Other ciliopathies

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Alstrom syndrome	*ALMS1*
	Bardet-Biedl syndrome	*BBS1, BBS10, BBS12, BBS2, CEP290, MKKS, MKS1, TRIM32,*
	Leber congenital amaurosis	*CEP290*
	Autosomal dominant polycystic kidney disease	*PKD1, PKD2*
	Autosomal recessive polycystic kidney disease	*PKHD1*
	Nephronophthisis	*CEP290, NPHP1, NPHP3, WDR19, TMEM67*
	Senior-Loken syndrome	*NPHP1, WDR19, CEP290*
	Meckel Gruber syndrome	*CEP290, MKS1, NPHP3, RPGRIP1L, TMEM67, TMEM216*
	Joubert syndrome	*AHI1, CEP290, CPLANE1, CSPP1, INPP5E, KIAA0586, NPHP1, NPHP3, OFD1, RPGRIP1L, TMEM216, TMEM67*
	Primary ciliary dyskinesia	*CCDC39, CCDC40, DNAH11, DNAH5, DNAI1, RSPH1*

Connective Tissue Panel

Sub-panel code	Sub-panel name	**Genes analysed** (alphabetical order)
	Aortopathy	*ACTA2, BMP1, COL1A2, COL3A1, COL5A1, COL5A2, FLNA, COL1A1, ELN, FBN1, FBN2, MYH11, SKI, SMAD2, SMAD3, SMAD4, TGFB3, TGFBR1, TGFBR2*
	Ehlers Danlos syndrome	*COL1A1, COL1A2, COL3A1, COL5A1, COL5A2*
	Homocystinuria	*ABCD4, CBS*
	Loeys-Dietz syndrome	*SMAD2, SMAD3, TGFBR1, TGFBR2*
	Marfan syndrome	*FBN1*
	Pseudoxanthoma elasticum	*ABCC6*
	Shprintzen-Goldberg syndrome	*SKI*
	Stickler syndrome	*COL2A1, COL11A1*

Differences of Sex Development Panel

Sub-panel code	Sub-panel name	**Genes analysed** (alphabetical order)
		AR, ATRX, CHD7, FRAS1, POR, SOX9, WT1

Dermatology Panel

1) Hypopigmentation conditions

Sub-panel code	Sub-panel name	**Genes analysed** (alphabetical order)
	Cole disease	*ENPP1*
	Dyskeratosis congenita	*DKC1, TINF2*
	Griscelli syndrome	*MYO5A, RAB27A*
	Hermansky-Pudlak syndrome	*HPS1, HPS3, HPS4, HPS6*
	Oculocutaneous albinism	*OCA2, TYR, TYRP1*
	Waardenburg syndrome	*MITF, PAX3, SOX10*

2) Neurocutaneous disorders

Sub-panel code	Sub-panel name	**Genes analysed** (alphabetical order)
	Congenital melanocytic nevus syndrome	*BRAF* (somatic), *NRAS* (somatic)
	Neurofibromatosis type 1 and Legius syndrome	*NF1, SPRED1*
	Neurofibromatosis type 2	*NF2*

	Tuberous sclerosis complex	*TSC1, TSC2*
	Von Hippel Lindau	*VHL*

3) Other

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	CHILD (Congenital hemidysplasia with ichthyosiform erythroderma and limb defects) syndrome	*NSDHL*
	Ectodermal dysplasia	*EDA, EDAR, EDARADD, TP63, WNT10A, WNT1*
	Epidermolysis bullosa	*KRT14, KRT5, COL7A1, DSP, LAMA3, LAMB3, LAMC2*
	Goltz syndrome (focal dermal hypoplasia)	*PORCN*
	Gorlin syndrome (nevoid basal cell carcinoma syndrome)	*PTCH1*
	Harlequin icthyosis	*ABCA12*
	KID (keratitis-ichthyosis-deafness) syndrome	*GJB2*
	Lamellar icthyosis	*TGM1*
	Lipoid proteinosis	*ECM1*
	Pseudoxanthoma elasticum	*ABCC6*
	Xeroderma Pigmentosum	*ERCC2, POLH, XPA, XPC*

Gastrointestinal Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
		ABCB11, ABCB4, ATP8B1, BMPR1A, JAG1, MPV17, MUTYH

Haematology Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Haemophilia	*F8, F9*
	Hereditary haemochromatosis	*HAMP, HFE, HFE2, SLC40A1, TFR2*
	Hereditary haemorrhagic telangiectasia	*ACVRL1, ENG, SMAD4*

	Sickle cell disease/Beta-thalassemia	*HBB*
	Barth syndrome	*TAZ*

1) Bone marrow failure syndromes

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Diamond-Blackfan anaemia	*RPL5, RPS19*
	Fanconi anaemia	*BRCA1, BRCA2, ERCC4, FANCA, FANCB, FANCC, FANCD2, FANCE, FANCF, FANCG, RAD51D, RAD51C, SLX4*
	Dyskeratosis congenita	*TINF2; DKC1*

2) Gray hair syndromes

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Chediak-Higashi	*LYST*
	Gricelli syndrome	*MYO5A, RAB27A*

Hearing Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Syndromic hearing loss	*ADGRV1, BTD, COL4A5, EYA1, GJB2, MYO7A, PCDH15, SALL1, SIX1, SLC26A4, TCOF1*
	Non-syndromic hearing loss	*GJB6, MYO6, OTOF, POU3F4, TMC1, TMIE, TMPRSS3*

Metabolic Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Aminoacidopathies (Organic acidaemias)	*OTC, ASL, BCKDHA, BCKDHB, CBS, GCDH, ABCD4, FAH, MMAA, MMAB, MMACHC, MUT, PCCA, PCCB, TAZ, AMT, GLDC*
	Canavan disease	*ASPA*
	Citrullinemia type I	*ASS1*
	Congenital disorders of glycosylation	*PMM2*
	Cystinuria	*SLC3A1, SLC7A9*
	Electron transfer defects	*ETFDH*
	Familial hypercholesterolemia	*APOB, LDLR, PCSK9*

	Fatty acid oxidation disorders	*ACADM, ACADVL, SLC22A5*
	Galactosaemia	*GALK1, GALT*
	Glycogen storage disorders	*AGL, G6PC, GAA, GBE1, GALNS, PYGM, SLC37A4*
	Maple syrup urine disease	*DBT*
	Menkes	*ATP7A*
	Metachromatic leukodystrophy	*ARSA, PSAP*
	Mitochondrial	*PDHA1, PDHX, POLG*
	Peroxisomal disorders	*ABCD1, EBP, PEX1, PEX10, PEX12, PEX26, PEX6, PEX7*
	Phenylketonuria	*PAH*
	Primary hyperoxaluria type 1	*AGXT*
	Smith-Lemli-Opitz	*DHCR7*
	Wilson disease	*ATP7B*

1) Lysosomal storage diseases

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Cystinosis	*CTNS*
	Fabry disease	*GLA*
	Gaucher	*GBA*
	Krabbe disease	*GALC*
	Lysosomal acid lipase deficiency	*LIPA*
	MPS	*ARSB, IDUA, IDS, GALNS, GLB1, GNS, GUSB, HGSNAT, NAGLU, SGSH*
	Mucolipidosis type 2	*GNPTAB*
	Mucolipidosis type 3A/B	*AGA*
	Mucolipidosis type 4	*MCOLN1*
	Neuronal ceroid-lipofuscinosis	*CLN3, TPP1*
	Nieman-Pick disease	*NPC1, NPC2, SMPD1*
	Pycnodysostosis	*CTSK*
	Tay-Sachs	*HEXA*

Neuromuscular Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Acrocallosal syndrome	*GLI3, KIF7*

	Adult-onset degenerative disorders	*LRRK2, PARK2, PINK1, PSEN1, PSEN2, SNCA, SOD1, VCP*
	Arthrogryposis	*CHRNG, ECEL1, MYBPC1, MYH3, MYH8, TNNI2, TNNT3, TPM2*
	Ataxia with oculomotor apraxia type 2	*APTX*
	Congenital muscular dystrophies and dystroglycanopathies	*B3GALNT2, B3GNT2, DAG1, FKRP, FKTN, LARGE1, POMGNT1, POMT1, POMT2, TMEM5, ISPD, LAMA2, CHKB, COL12A1, ITGA7*
	Dravet syndrome	*SCN1A*
	Duchenne/Becker Muscular Dystrophy	*DMD*
	Holoprosencephaly	*DHCR7, FGFR1, GLI2, PTCH1, SHH, SIX3, TGIF1*
	Limb girdle muscular dystrophy and related disorders	*ANO5, CAPN3, DMD, DYSF, EMD, FHL1, FKRP, LMNA, SGCA, SGCB, SGCD, SGCG, TNPO3, LAMA2*
	Lissencephaly	*DCX, DYNC1H1, LAMB1, PAFAH1B1, POMGNT2, RELN, TUBA1A*
	Myopathies	*ACTA1, BAG3, COL6A1, COL6A2, COL6A3, CRYAB, DES, FHL1, GNE, LDB3, MEGF10, MTM1, MYH7, MYOT, RYR1, STAC3, TOR1A, TPM2, TAZ, CACNA1S, COL12A1*
	Myotonia congenita	*CLCN1*
	Neuropathy (including Charcot-Marie Tooth disease)	*GJB1, MFN2, MPZ, PMP22, SH3TC2, SLC12A6, TTR*
	Periodic paralysis	*CACNA1S*
	Rett/Angelman and related syndromes	*ATRX, CDKL5, MECP2, MEF2C, SLC9A6, UBE3A, ZEB2*
	X-linked hydrocephalus	*L1CAM*

Ophthalmology Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
		ABCA4, ADGRV1, CDH23, CHM, CLNR1, ELOVL4, FRAS1, GUCY2D, OCRL, PAX6, RB1, RPE65, USH1C, USH1G, USH2A, WHRN, EYA1

Overgrowth Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)

	Overgrowth and macrocephaly syndromes	*AKT3, DIS3L2, DNMT3A, EZH2, GLI3, GPC3, MTOR, NF1, NPR2, NSD1, OFD1, PIK3R2, PTEN, SPRED1*

RASopathies Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Cardiofaciocutaneous syndrome	*BRAF, MAP2K1, MAP2K2*
	Costello syndrome	*HRAS*
	Neurofibromatosis type 1 and Legius syndrome	*NF1, SPRED1*
	Noonan syndrome	*KRAS, LZTR, PTPN11, RAF1, RIT1, SOS1*

Respiratory Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Alpha-1-antitrypsin deficiency	*SERPINA1*
	Cystic fibrosis	*CFTR*
	Primary ciliary dyskinesia	*CCDC39, CCDC40, DNAH11, DNAH5, DNAI1, RSPH1*
	Pulmonary hypertension	*ACVRL1, BMPR1B, BMPR2, ENG, KCNK3, SMAD9, TBX4*
	Pulmonary veno-occlusive disease (PVOD)/pulmonary capillary haemangiomatosis (PCH)	*EIF2AK4*

Skeletal Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Camurati Engelman	*TGFB1*
	Chondrodysplasia punctata	*AGPS, ARSE, EBP, LBR, NSDHL, PEX7*
	Chondrogenesis	*FGFR3, LMX1B, SLC26A2*
	Craniosynostosis	*EFNB1, ERF, FGFR1, FGFR2, FGFR3, POR, RAB23, RECQL4, TCF12*
	Ehlers Danlos syndrome	*COL1A1, COL1A2, COL3A1, COL5A1, COL5A2*
	Epiphyseal	*COMP, MATN3*
	Hypo- and Hyperphosphatasia	*ALPL, CTSK, DMP1, EXT1, FGF23, PHEX, SLC34A3, ENPP1*
	Hypoplasia	*DOCK6, ESCO2, SOX9, TBX3, TBX5, WNT10A, WNT3*
	Jarcho-Levin syndrome	*MESP2*
	Metaphyseal	*ANKH, COL10A1, PTH1R*

	Osteo	*ACVR1, DYNC2H1, FLNA, FLNB, GNAS, IFT80, KIF7, NEK1, NPR2, PCNT, PRKAR1A, ROR2, RUNX2, SF3B4, SLC35D1, SOST, SOX9, TRIP11, TRPS1, TRPV4, TTC21B, WNT5A*
	Osteogenesis imperfecta	*BMP1, COL1A1, COL1A2, FKBP10, WDR19, WNT1*
	Pycnodysostosis	*CTSK*
	Schwartz-Jampel syndrome	*HSPG2*
	Other	*GPC6*

Single Gene Disorders Panel

Sub-panel code	**Sub-panel name**	**Genes analysed** (alphabetical order)
	Aarskog syndrome	*FGD1*
	Alpha-1-antitrypsin deficiency	*SERPINA1*
	Branchiootorenal spectrum disorders	*EYA1*
	Canavan disease	*ASPA*
	CHARGE syndrome	*CDH7*
	Coffin Lowry syndrome	*RPS6KA3*
	Cohen syndrome	*VPS13B*
	Congenital nephrotic syndrome	*NPHS1, NPHS2, WT1*
	Cornelia de Lange syndrome	*NIPBL*
	Familial Mediterranean fever	*MEFV*
	Kabuki syndrome	*KDM6A, KMT2D*
	Lipoid proteinosis	*ECM1*
	Miller syndrome	*DHCDH*
	Porphyria Variegata	*PPOX*
	Rubinstein-Taybi syndrome	*CREBBP, EP300*
	Severe combined immunodeficiency	*ADA*
	Smith-Magenis syndrome	*RAI1*
	TAR syndrome	*RBM8A*
	van der Woude syndrome	*GRHL3, IRF6*
	Wiskott-Aldrich Syndrome	*WAS*

Updated: November 2023

Manuscript Body incl. Abstract.pdf

ORIGINALITY REPORT

3%	2%	2%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	ntnuopen.ntnu.no Internet Source	1%
2	"Abstracts from the 54th European Society of Human Genetics (ESHG) Conference: e-Posters", European Journal of Human Genetics, 2022 Publication	1%
3	Submitted to University of Witwatersrand Student Paper	<1%
4	bmcmedgenomics.biomedcentral.com Internet Source	<1%
5	www.frontiersin.org Internet Source	<1%
6	wiredspace.wits.ac.za Internet Source	<1%
7	"Abstracts", The Journal of Molecular Diagnostics, 2016 Publication	<1%