A national pilot study: Clinical long-read sequencing of chromosomal rearrangements

Jesper Eisfeldt[1,2,3*], Adam Ameur[4,5*], Felix Lenner[4,5], Esmee Ten Berk de Boer[1,2,3], Marlene Ek[1,2], Josephine Wincent[1,2], Raquel Vaz[1], Jesper Ottosson[6], Tord Jonson[10,11], Sofie Ivarsson[10,11], Sofia Thunström[6], Alexandra Topa[6], Simon Stenberg[6], Anna Rohlin[6,7], Magnus Burstedt[8], Frida Nordin[9], Maria Sobol[4], Panagiotis Baliakas[4,5], Marie-Louise Bondeson[4,5], Ida Höijer[4,5], Kristine Bilgrav Saether[1,3], Lovisa Lovmar[6], Hans Ehrencrona[10,11], Malin Melin[4,5], Lars Feuk[4,5#], Anna Lindstrand[1,2,12#]

1. Department of Molecular Medicine and Surgery and Center for Molecular Medicine, Karolinska Institutet, Stockholm, Sweden.
2. Department of Clinical Genetics and Genomics, Karolinska University Hospital, Stockholm, Sweden
3. Science for Life Laboratory, Karolinska Institutet Science Park, Solna, Sweden.
4. Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden
5. Science for Life Laboratory, Uppsala University, Uppsala, Sweden
6. Department of Clinical Genetics and Genomics, Sahlgrenska University Hospital, Gothenburg, Sweden
7. Department of Laboratory Medicine, Institute for Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
8. Department of Medical Bioscience, Medical and Clinical Genetics, Umeå University, Umeå, Sweden.
9. Department of Pharmacology and Clinical Neurosciences, Umeå University, Umeå, Sweden.
10. Division of Clinical Genetics, Department of Laboratory Medicine, Lund University, Lund, Sweden
11. Department of Clinical Genetics, Pathology and Molecular Diagnostics, Office for Medical Services, Region Skåne, Lund, Sweden
12. Genomic Medicine Center Karolinska, Karolinska University Hospital, Stockholm, Sweden.

*#These authors contributed equally to this work

Correspondence:

Prof Anna Lindstrand (anna.lindstrand@ki.se); Prof Lars Feuk (lars.feuk@igp.uu.se)

ABSTRACT

Background: Clinical genetic laboratories often require comprehensive analysis of chromosomal rearrangements/structural variants (SVs) which can range from gross chromosomal events, such as translocations and inversions, to supernumerary ring/marker chromosomes, and small deletions or duplications. To fully understand the complexity of a specific event and its associated clinical consequences, it is imperative to locate the breakpoint junctions (BPJs) and to resolve the derivative chromosome structure. This task, however, often surpasses the capabilities of conventional short-read sequencing technologies. In contrast, emerging long-read sequencing techniques present a compelling alternative for clinical diagnostics.

Methods: Here, the Genomic Medicine Sweden Rare Diseases (GMS-RD) consortium validated HiFi Revio long-read whole genome sequencing (lrGS) for clinical digital karyotyping of SVs nationwide. The first 13 samples included in this study were collected from five of the seven Swedish regions with university hospitals. We established a national pipeline and a shared variant database for variant calling and filtering. The included validation samples cover a spectrum of simple and complex SVs including inversions, translocations and copy number variants.

Results: The results from the lrGS analysis match the reported karyotype for 11/13 individuals and known SVs were mapped at nucleotide resolution. A complex rearrangement on chromosome 15 was identified only through read depth analysis and a mosaic ring chromosome 21 remained undetected. Depending on input DNA quality, the average read length ranged from 8.3-17.6 kb and the coverage was >20x for all samples. *De novo* assembly resulted in a limited number of contigs per individual (N50 range 6-73 Mb) clearly separating the two alleles in most cases, enabling direct characterization of the chromosomal rearrangements.

Conclusions: We have within a national pilot study successfully demonstrated the utility of HiFi Revio lrGS as a clinical analysis of chromosomal rearrangements.

BACKGROUND

Although short-read genomic (srGS) analysis approaches such as exome and genome sequencing have been highly successful in identifying disease-causing genetic variants for diagnostic purposes, the primary focus of analysis remains on single nucleotide variations (SNVs) and small insertions-deletions (INDELs) (1). In contrast, the calling and interpretation of structural chromosomal rearrangements is still challenging (2, 3). Such events, collectively called structural variants (SVs), are defined as genetic variants larger than 50 bp. When SVs involve a repositioning of genetic material either within a chromosome or between chromosomes, they are also referred to as chromosomal rearrangements. These encompass both recurrent and non-recurrent copy number variations (CNVs, deletions and duplications) and balanced chromosome abnormalities (translocations, inversions, and insertions). Furthermore, there is growing evidence that complex SVs, which contain multiple breakpoint junctions or consist of more than one simple structural variation in *cis*, are more prevalent than previously assumed (4). A thorough analysis of such events detailing the DNA breakpoints at the nucleotide level is crucial for comprehending the disease-causing and rearrangement-generating mechanisms. This level of understanding is necessary for effective personalized clinical management and genetic counselling (5).

While srGS shows great promise as a first-line diagnostic test, with potential to capture a high proportion of pathogenic SVs previously identified by traditional methods (6), it still has limitations. The existing srGS SV pipelines lack the ability to produce high-quality genome assemblies necessary to resolve complex disease-causing SVs. In this context, long-read genome sequencing (lrGS) has emerged as a promising

alternative with potential to capture the complete scope of structural genomic variation in the human genome (7, 8). Indeed, lrGS provides a comprehensive approach for characterizing various types of SVs identified in a clinical genetic laboratory, whether discovered through traditional methods or srGS. With the introduction of lrGS using highly accurate consensus reads (HiFi reads), the quality of SNV and short indel calls from lrGS data has increased dramatically and enables characterization of the full spectrum of genetic variation, from SNVs to SVs as well as methylation patterns in a single sequencing experiment. Due to the several kb long sequence reads, lrGS offers unique possibilities to study regions in the human genome that are not easily captured by other genomics technologies, such as highly repetitive or homologous regions. Furthermore, sequencing of the native DNA also provides methylation information. With decreasing cost, it is increasingly attractive to use lrGS in unsolved rare disease cases. Yet, lrGS differs markedly from short-read sequencing and various factors should be considered before its clinical implementation. These include the quality and quantity of DNA starting material, intricate specifics related to sequencing including read quality, total coverage, and read length, as well as the complexities tied to data analysis, visualization, and the eventual clinical interpretation.

Recognizing the promise and challenges of lrGS, we embarked on a national pilot study using the PacBio Revio system. We leveraged on Genomic Medicine Sweden (GMS), a national initiative designed to implement precision medicine throughout the regionally organized healthcare system (9) with the aim to validate lrGS for clinical digital karyotyping of unresolved SVs nationwide. The included samples originated from five out of seven university hospital regions in Sweden and the DNA was not necessarily collected using lrGS-optimized protocols. This approach mirrored the

typical sample quality and preparation found in most of Sweden's hospitals and biobanks today, thereby giving us a realistic view of the potential and challenges of lrGS in clinical diagnostics.

METHODS

Study subjects

All Swedish university regions are affiliated to GMS-RD. In 2023 a nationwide lrGS pilot study was conducted, with five regions opting to send samples from individuals with seemingly complex SVs, in either a singleton or trio format. The multicenter study was approved by the Swedish Ethical Review Authority (2019-04746) and written informed consent was obtained from each participating individual or their respective legal guardians.

Altogether, 13 samples were sequenced including eight proband singletons, one duo, and one trio. For two samples a high molecular preparation had been performed prior to lrGS, while the remaining DNA samples were retrieved from clinical biobanks and obtained using standard extraction protocols. Details on the included individuals are given in Table 1.

Long-read genome sequencing

The DNA samples were fragmented to 15-20 kb using Megaruptor 3 (Diagenode). PacBio SMRTbell library construction was performed using the SMRTbell Template prep kit 3.0. SMRTbell libraries were size-selected either using AMPure beads or by the gel-based SageElf system (Sage Science). The library preparation procedure is described in the protocol “Preparing whole genome and metagenome sequencing libraries using SMRTbell prep kit 3.0” from PacBio. The SMRTbell library sizes and profiles were evaluated using Fragment Analyzer (Agilent Technologies). PacBio sequencing was performed on the PacBio Revio system with 24 h movie time. Each SMRTbell library was sequenced on a 25M SMRT cell.

Analysis of long-read genome data

The PacBio HiFi reads were aligned to both GRCh38 and the T2T-CHM13v2.0 reference using minimap2 (version 2.26) (10) and SAMtools (version 1.17) (11). SNVs and short indels were called with DeepVariant 1.5.0 (12) and phased using WhatsHap 1.7 (13). SVs were called using Sniffles (14), which was run according to previously published parameters (7), and CNV were called using HiFiCNV (0.1.6b) (https://github.com/PacificBiosciences/HiFiCNV). Quality metrics were gathered using fastqc 0.11.9 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), cramino 0.9.7 (15) and mosdepth 0.3.3 (16). Next, we generated phased assemblies using hifiasm (version 0.19.5-r587) (17). The resulting *de novo* assemblies were aligned to hg38 and T2T-CHM13 reference using minimap2 (version 2.26) (10), SVs and SNVs were called using HTSBOX (version r345) (11) and counted using bcftools (version 1.17) (11), and quality control was performed using QUAST (version 5.0.2) (18).

Identification and characterization of structural variants

Prior to SV characterization, the Sniffles SV calls were filtered based on size and allele frequencies. The size filtering was performed using bcftools (11), excluding all SV calls smaller than 2 kbp. The frequency filtering was performed using SVDB (19). First, a frequency database was constructed using the calls from all 13 individuals sequenced (Table 1). Next, we annotated the remaining calls using that database, and variants present in more than one unrelated individual were removed. Lastly, the remaining SV calls were inspected using IGV, and characterized as described previously (7).

RESULTS

High-quality long-read genome data is obtained from clinical DNA samples

All individuals were sequenced on one Revio 25M SMRTcell, generating at least 20X coverage of high-quality (HiFi) reads for all samples (range 19.9 - 32.1 X), with mean read length ranging from 8.3-17.6 kb (Figure 1; Additional file 1: Table S1). Two samples (P8.1 and P8.2) were obtained from high-molecular weight DNA extraction of fresh blood, enabling for a more efficient size selection during library preparation. These two samples generated the longest read lengths and >30 X coverage, indicating that DNA quality is of importance to obtain the best possible PacBio HiFi data from a single SMRTcell. Variant calling resulted in an average of 4.7 million single nucleotide variants (SNVs) and small indels per individual (Additional file 1: Table S2; Additional file 2: Document S2). This number of SNVs is similar to what is typically found by human srGS analysis. We further performed de novo assembly for all individuals and the contig N50 values are shown in Figure 1C.

Identification of chromosomal genomic rearrangements

In all but one individual, the lrGS analysis could identify the chromosomal genomic rearrangement(s) known to be present in the specific sample. The rearrangement in individual P10, that eluded detection, was mosaic thus hindering the SV detection (Table 1). The mosaic tri- and tetrasomy on chromosome 15 in individual P2 was characterized as a copy number gain (between 2 and 5) in the long-read CNV-calls, although a clear distinction between the tri- and tetrasomy could not be made. While the starting point could not be distinguished from the centromere, the location of the end point was approximated between chr15:32176000–32346000, with the exact breakpoint located within a segmental duplication (Additional file 1: Figure S1). The

remaining samples were then subdivided depending on whether a simple event or a complex rearrangement was present. The HGVS karyotypes for the resolved events are given in Table 2.

Translocations and inversions: Three translocations and one inversion were characterized in four individuals (Figure 2; Table 2**)** with a mother and daughter both carrying the same two balanced events (P7.1 and P7.3). All variants were fully resolved but the translocation between chromosomes X and 9, t(X;9) (individual P9), was only detectable using the Telomere-to-Telomere (T2T) assembly for analysis. A secondary analysis in hg38 revealed a call of a t(X;5) in that sample, indicating that the breakpoint 9 region is missing in hg38. In individual P1 the three structural variants (two deletions and one duplication) detected prior to lrGS were in fact two separate rearrangements; one 16p deletion and one unbalanced translocation between chromosome 4 and chromosome 9 (Figure 2).

Complex rearrangements: Six complex rearrangements were detected and resolved in five unrelated individuals, all of which were unbalanced (Table 2). Notably, the same inv(X) was detailed in a mother and daughter pair (P8.1 and P8.2). The highest complexity level was identified in individual P4 where two complex events were identified, one DUP-TRIP-QUAD-TRIP-DUP-DEL and one DEL-INV-DEL (Figure 3A; Table 2). The complexity found in P4 was tightly followed by P5, who carries a complex rearrangement consisting of two duplications, one deletion, as well as an inverted copy number neutral fragment (Figure 3B; Table 2). In contrast, the simplest rearrangements consisted of DEL-INV-DEL (P3 and P4) or DEL-INS-DEL (P6) (Additional file 1: Figure S2).

Breakpoint junction analysis

In eleven of the rearrangements, the BPJs were characterized at nucleotide resolution, allowing for a detailed analysis of insertions, repeat elements and microhomology and disrupted genes (Table 3; Additional file 1: Table S3). Comparing the Sniffles SV calls to the calls from our *de novo* assembly pipeline, we find that the de novo workflow detects 30 of the 31 BPJs detected by Sniffles (Table 2, Table 3; Additional file 1: Table S3). Notably, we find a great diversity of such genetic signatures in the analyzed cases. In two rearrangements, the breakpoints contain matched repeat elements: L1 in the t(4;9) (P1) and various Alu elements in the inv(X) (P8.1 and P8.2). In individual P4 (Table 3) microhomology of 1-11 nucleotides (nt) was observed in five of the six BPJs and the final BPJ contains a nine nt insertion. Insertions larger than 2nt were detected in three additional rearrangements (Table 3); a complex translocational insertion (P6) (Additional File 1: Figure S2), a t(1;10) (P7.1 and P7.3) (Figure 2) and a t(X;9) (P9) (Figure 2). However, for P9, that warranted T2T analysis, we suspect that the 990 nt insertion results from a reference error. The final two rearrangements, a DEL-INV-DEL on chr22 (P3) (Additional File 1: Figure S2) and a complex DEL-INV-NML-DUP-NML-DUP on chr3 (P5, Figure 3B), showed blunt ends that contained neither microhomologies, matched repeats, or larger insertions (Table 3).

Characterization of background structural variants

Next, we assessed the SV burden in all the 13 samples of the GMS-RD lrGS cohort. On average, we detected 23 814 SVs per individual that were subdivided into complex (0.05%), deletions (41%), duplications (1%), insertions (51%), inversions (1%) and translocations (5%) (Figure 4A). In general, the detected SVs were small, with 82% of SV below 1000bp and a peak is observed at 300bp representing the *Alu* insertions and deletions (Figure 4B). Most variants are present in more than one of the 13

analyzed individuals and only 30% are unique (Figure 4C). Finally, when comparing to the SweGen SV database (20) containing background SVs in 1000 Swedish individuals assessed from srGS data, approximately 70% of the lrGS SV calls are novel and not present in the SweGen dataset (Figure 4D).

DISCUSSION

In this unique national collaborative project, we have validated lrGS as a clinical follow up analysis of SVs in Sweden. By leveraging expertise from different regions, we implemented comprehensive workflows - from sample collection and DNA prep to sequencing and data analysis – available to individuals nationwide.

While srGS shows great promise as a first-line diagnostic test, it has a limited diagnostic yield of approximately 40% in patients with suspected rare genetic conditions (1). This shortcoming is due to several factors. Firstly, our knowledge about benign and morbid genomic changes in humans is still limited. Many variants identified in the srGS data are not recognized as pathogenic, and thus, they are either filtered out or reported as variants of unknown significance. Secondly, certain types of variants, such as SVs, repeat expansions, and variants in complex and repetitive genomic regions are difficult to capture with short reads. However, lrGS is beginning to overcome these limitations, with SVs representing promising initial targets for lrGS in clinical diagnostics. In the context of chromosomal rearrangements, the ability to phase the long-reads into contigs by *de novo* assembly will likely increase the possibility to describe complex derivative chromosomes with breakpoints situated within regions difficult to sequence. One of the cases included here, P9, with a reciprocal translocation, t(X;9), was unresolved in GRCh38 but could be mapped in

the T2T assembly because the chromosome 9 BPJ was located within a sequence that was missing in GRCh38. We have previously shown that lrGS using the Oxford Nanopore Technologies platform was able to capture a centromeric translocation that was undetected by srGS, linked read GS, and optical genome mapping. Such cases highlight that limitations of the quality of the current reference genome (hg38) still hamper accurate variant annotation and detection of clinically relevant breakpoints (21). Graph genomes are one proposed solution to overcome the lack of diversity in current reference genomes, and our high-quality personal assemblies indicate that such reference may be generated locally in clinical settings. It is also clear that some regions of the genome are currently recalcitrant to both assembly and mapping based variant identification. Among the samples analyzed here, the mosaic partial tri- and tetrasomy on chromosome 15 was only detected by read depth analysis highlighting the importance of including such tools for a complete lrGS analysis.

For the two inv(X) individuals, we obtained new blood samples necessary for the extraction of high molecular weight DNA (HMW DNA). The difference in this new extraction was reflected in the sequencing statistics: the HMW DNA allowed a more stringent size selection of the library leading to a sequencing yield of 100.0 Gb and 96.7 Gb, respectively, with an average read length of 16.4kb, compared to the 11 samples where only regular clinical DNA was available, with an average sequencing yield of 78.4Gb and average read length of 11.9kb (Additional File 1: Table S1; Additional File 2: MultiQC report). Even so, the results from the latter samples are outstanding, and the assemblies were highly contiguous (median contig N50 was 46 Mb for HMW-DNA samples and 27 Mb for regular DNA) (Additional File 1: Table S2). In summary, we find that lower quality, clinically available DNA extracted using

standard methods can be used to produce high quality lrGS datasets and obtain clinically relevant information. However, HMW-DNA provides superior results.

In addition to being able to detect 12 of 13 chromosomal rearrangements, the lrGS SV analysis also increases our understanding of background SVs in the human genome. We present well known features, such as the presence of *Alu* insertions/deletions, resulting in a peak in the SV size distribution at roughly 330 bp (Figure 4B). Also, we show that a small database, consisting of only the 11 unrelated individuals in this cohort is sufficient to filter out the majority of SV calls as non-unique variants (Figure 4D). Overall, these results highlight the high performance of lrGS SV analyses.

However, we notice that our *de novo* assembly-based SV calling pipeline reports all except one BPJ (Additional files 1: Table S3), and that our assemblies are contiguous and yield a reasonable number of SNV calls (Supplementary Table 2). In aggregate, these results indicate that *de novo* assembly-based genome analyses are comprehensive and allow for the characterization of SVs alongside other variant types, such as SNV and repeat expansions. We envision further studies incorporating *de novo* assembly and graph genomes to characterize SVs, as well as genomic features such as centromeres, telomeres, and segmental duplications in population scales and in clinical settings. Previous studies have shown that compared to srGS, lrGS shows a three-fold increase in sensitivity for genetic variants within the size range of 7 bp to 1 kbp (22-24). Hence, in clinical diagnostics, lrGs will likely identify novel pathogenic variants in this size window even after an initial srGS analysis. Comparing the SV calls of the GMS-RD long-read cohort to that of SweGen (20), we discovered that

approximately 70% of the lrGS SV calls are not present in the SweGen SV database (Figure 4C).

Both deletions and insertions show a peak at 300 bp representing mobile element variation (Figure 4B), and such variation is typically detected using specialized software in the lrGS data (25). Additionally, it is well known that the current lrGS callers report repeat expansions/retractions as insertions/deletions (14), while such variants may also be called using specialized tools in the lrGS data (26). Also we noted that the lrGS SV analysis results in ~100 duplications per individual (Figure 4A), which is low compared to the number of duplications detected from a srGS SV analysis (8, 27). Even if some of the srGS SV calls may be false positives these results indicate that lrGS SV analysis needs improvement. Overall, lrGS and srGS SV calling tools report different types and sizes of variation, and a more in-depth analysis is required to assess the strengths and weaknesses of these bioinformatics tools.

Analyzing the BPJ of the 12 rearrangements, we discovered a large diversity in breakpoint characteristics, highlighting the diversity and complexity of the underlying SV formation (Table 4). In particular, we discovered two rearrangements with matched repeat elements at the BPJs, a pattern indicative of non-allelic homologous recombination (NAHR) (28). In P4 (Figure 3A) we discovered stretches of microhomologies, as well as a large insertion (990 nt). These patterns are indicative of a replicative mechanism of formation, such as fork-stalling and template switching (FoSTeS). In three cases we discovered large insertions, that may indicate either a replicative mechanism or formation through repair mechanisms such as microhomology-mediated end-joining. It has been shown that Polθ may introduce such

insertions during the repair of double stranded breaks (29). The largest insertion was discovered at the BPJ of the t(X;9) in P9. This 990 nt long insertion consists mainly of TTCCA repeats that are also present at the breakpoint region at chromosome 9. Therefore, we hypothesize that this insertion may be due to background variation in this repeat, and not caused through the formation of the translocation. In future studies we strongly suggest analyzing SV in the context of trios or extended families, that allow high quality personal de novo assemblies be generated from the parents. Such assemblies will give a more accurate analysis of SV formation and genome dynamics of repetitive regions. Our results shown here indicate that this type of assemblies may be produced from the DNA obtained by current routine diagnostic workflows.

Conclusions

In summary, utilizing the PacBio Revio system we successfully analyzed 12 chromosomal rearrangements from Swedish probands. By coordinating our local efforts and working together in the GMS Rare Diseases consortium we were able to validate lrGS for digital karyotyping. However, challenges remain, such as the development of more sophisticated interpretation tools and comprehensive databases for filtering.

Abbreviations

Array-CGH, array comparative genomic hybridization; BPJ: breakpoint junction; chr: chromosome; CNV: copy number variant; FISH, fluorescence in situ hybridization; FoSTeS: fork-stalling and template switching; GMS: Genomic Medicine Sweden; HMW: high molecular weight; lrGS: long-read genome sequencing; MLPA, multiplex ligation-dependent probe amplification; nt: nucleotide; pos: position; SNV: single nucleotide variant; srGS: short-read genomic sequencing; SV: structural variant; T2T: Telomere-to-Telomere.

Acknowledgments

We are very grateful to the participating families. We would also like to thank UPPMAX for the use of computer infrastructure resources (project sens2022541) and the support from the National Genomics Infrastructure (NGI) Uppsala and Clinical Genomics Uppsala. Several authors of this publication are members of the European Reference Network on Rare Congenital Malformations and Rare Intellectual Disability ERN-ITHACA [EU Framework Partnership Agreement ID: 3HP-HP-FPA ERN-01-2016/739516]. We also acknowledge the Uppsala ALF office.

Funding

Funding was provided by Genomic Medicine Sweden as well as grants to AL from the Swedish Research Council (2017-02936 and 2019-02078), the Stockholm Regional Council and the Swedish Rare Diseases Foundation, and to LF from the Swedish Research Council (2017-01861) and Hjärnfonden (FO2022-0207).

Availability of data and materials

Analysis results supporting the conclusions of this article are included within the article and its additional files. The PacBio HiFi data and individual variant calls will be available through the European '1+ Million Genomes' Initiative that is being set up. As a clinical case data will be stored at the National Genomics Platform managed by Genomic Medicine Sweden, accession number GMS-RD_00002. Data will be made available upon reasonable request through the corresponding authors.

The following public databases and open-source software were used:

Genome Reference Consortium Human Build 38 (https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001405.26/)

The Telomere-to-Telomere consortium CHM13 project T2T-CHM13v2.0

(https://github.com/marbl/CHM13)

Assemblatron (https://github.com/J35P312/Assemblatron)

BCFtools (https://samtools.github.io/bcftools/bcftools.html) (11)

Cramino (https://github.com/wdecoster/cramino) (15)

Deepvariant (https://github.com/google/deepvariant) (12)

FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

Hifiasm (https://github.com/chhylp123/hifiasm) (17)

Hificnv (https://github.com/PacificBiosciences/HiFiCNV)

HTSBOX (https://github.com/J35P312/htsbox) (11)

Minimap2 (https://github.com/lh3/minimap2) (10)

Samtools (https://www.htslib.org/) (11)

Sniffles (https://github.com/fritzsedlazeck/Sniffles) (14)

SVDB (https://github.com/J35P312/SVDB) (19)

Mosdepth (https://github.com/brentp/mosdepth)

MultiQC (https://multiqc.info/)

Whatshap (https://whatshap.readthedocs.io/en/latest/) (13)

The SweGen dataset (https://swefreq.nbis.se/dataset/SweGen) (20)

Authors' contributions

JE, AA, MM, LL, HE, LF and AL conceptualized the study. JE, AA, FL, ETB, RV, LF and AL were major contributors to writing the manuscript. ME and IH performed extraction and QC analysis of DNA. JW, JO, TJ, SI, ST, AT, SS, AR, MB, FN, PB, MLB, LL, HE and AL provided patient samples and clinical information of patients. JE, AA, FL, ETB, KBS and LF performed bioinformatics analysis and development. All authors have read, edited and approved the final manuscript.

Ethics approval and consent to participate

The Ethical Review Board in Sweden approved the study (ethics permit number 2019-04746). Written consent to participate was provided by the subject or their legal guardians. The research conformed to the principles of the Helsinki Declaration.

Consent for publication

Written informed consent was obtained to publish.

Competing interests

AL has received honoraria from Pacific Biosciences

Supporting information

Additional file 1: Document S1 (Bioinformatic supplementary methods), Table S1 (Basic quality metrics of the samples included in this study), Table S2 (De novo assembly quality metrics), Table S3 (De novo assembly variant calling), Figure S1 (Read coverage variation in chromosome 15), Figure S2 (Subway plots of characterized rearrangements).

Additional file 2: Document S2 (MultiQC report)

REFRENCES

1. Stranneheim H, Lagerstedt-Robinson K, Magnusson M, Kvarnung M, Nilsson D, Lesko N, et al. Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. Genome Med. 2021;13(1):40.
2. Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive evaluation and characterisation of short read general-purpose structural variant calling software. Nat Commun. 2019;10(1):3240.
3. Kosugi S, Momozawa Y, Liu X, Terao C, Kubo M, Kamatani Y. Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. Genome Biol. 2019;20(1):117.
4. Schuy J, Grochowski CM, Carvalho CMB, Lindstrand A. Complex genomic rearrangements: an underestimated cause of rare diseases. Trends Genet. 2022;38(11):1134-46.
5. Tesi B, Boileau C, Boycott KM, Canaud G, Caulfield M, Choukair D, et al. Precision medicine in rare diseases: What is next? J Intern Med. 2023;294(4):397-412.
6. Lindstrand A, Eisfeldt J, Pettersson M, Carvalho CMB, Kvarnung M, Grigelioniene G, et al. From cytogenetics to cytogenomics: whole-genome sequencing as a first-line test comprehensively captures the diverse spectrum of disease-causing genetic variation underlying intellectual disability. Genome Med. 2019;11(1):68.
7. Eisfeldt J, Pettersson M, Petri A, Nilsson D, Feuk L, Lindstrand A. Hybrid sequencing resolves two germline ultra-complex chromosomal rearrangements consisting of 137 breakpoint junctions in a single carrier. Hum Genet. 2021;140(5):775-90.
8. Eisfeldt J, Pettersson M, Vezzi F, Wincent J, Kaller M, Gruselius J, et al. Comprehensive structural variation genome map of individuals carrying complex chromosomal rearrangements. PLoS Genet. 2019;15(2):e1007858.
9. Fioretos T, Wirta V, Cavelier L, Berglund E, Friedman M, Akhras M, et al. Implementing precision medicine in a regionally organized healthcare system in Sweden. Nat Med. 2022;28(10):1980-2.
10. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.
11. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10(2).
12. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983-7.
13. Martin M, Patterson M, Garg S, Fischer SO, Pisanti N, Klau GW, et al. WhatsHap: fast and accurate read-based phasing. bioRxiv. 2016:085050.
14. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, von Haeseler A, Schatz MC. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018;15(6):461-8.
15. De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. Bioinformatics. 2023;39(5).
16. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 2018;34(5):867-8.
17. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18(2):170-5.
18. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34(13):i142-i50.

19. Eisfeldt J, Vezzi F, Olason P, Nilsson D, Lindstrand A. TIDDIT, an efficient and comprehensive structural variant caller for massive parallel sequencing data. F1000Res. 2017;6:664.
20. Ameur A, Dahlberg J, Olason P, Vezzi F, Karlsson R, Martin M, et al. SweGen: a whole-genome data resource of genetic variability in a cross-section of the Swedish population. Eur J Hum Genet. 2017;25(11):1253-60.
21. Eisfeldt J, Martensson G, Ameur A, Nilsson D, Lindstrand A. Discovery of Novel Sequences in 1,000 Swedish Genomes. Mol Biol Evol. 2020;37(1):18-30.
22. Chaisson MJ, Huddleston J, Dennis MY, Sudmant PH, Malig M, Hormozdiari F, et al. Resolving the complexity of the human genome using single-molecule sequencing. Nature. 2015;517(7536):608-11.
23. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et al. Corrigendum: Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 2018;28(1):144.
24. Chaisson MJP, Sanders AD, Zhao X, Malhotra A, Porubsky D, Rausch T, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat Commun. 2019;10(1):1784.
25. Keane TM, Wong K, Adams DJ. RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics. 2013;29(3):389-90.
26. Dolzhenko E, Bennett MF, Richmond PA, Trost B, Chen S, van Vugt J, et al. ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. Genome Biol. 2020;21(1):102.
27. Collins RL, Brand H, Karczewski KJ, Zhao X, Alfoldi J, Francioli LC, et al. A structural variation reference for medical and population genetics. Nature. 2020;581(7809):444-51.
28. Nazaryan-Petersen L, Eisfeldt J, Pettersson M, Lundin J, Nilsson D, Wincent J, et al. Replicative and non-replicative mechanisms in the formation of clustered CNVs are indicated by whole genome characterization. PLoS Genet. 2018;14(11):e1007780.
29. Carvalho CM, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nat Rev Genet. 2016;17(4):224-38.

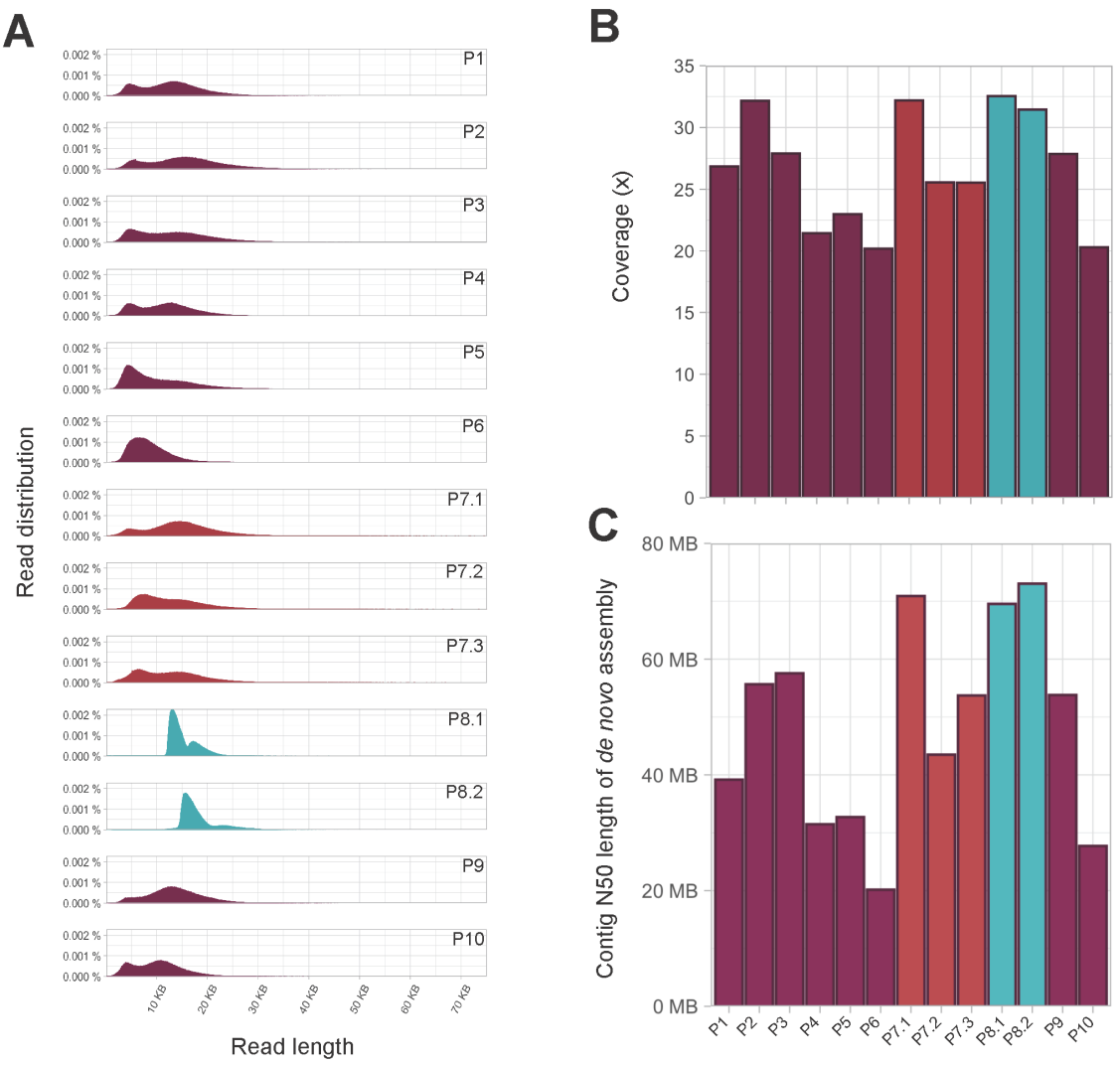

Figure 1. Quality measures of lrGS. Read length distribution (left). Coverage (top right) and N50 of *de novo* assembly (bottom right).

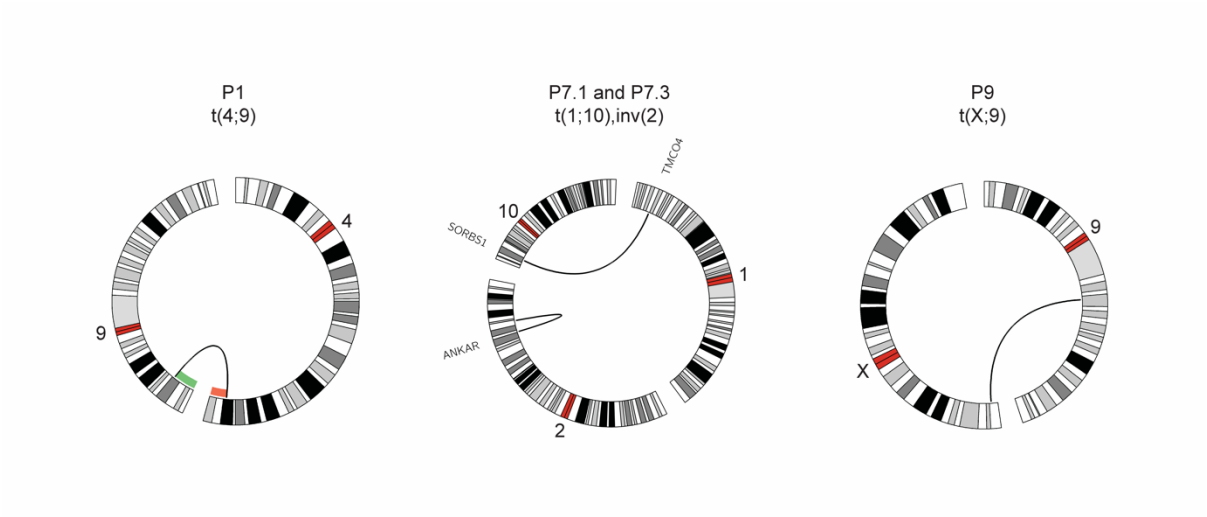

Figure 2: Translocations and inversions identified with lrGS. Circus plots of rearrangements detected in three cases using lrGS: a t(4;9) in P1, a t(1;10) and inv(2) (P7.1 and P7.3) and a t(X;9) (P9). A green/red line indicates copy number gain/loss, respectively. Genes disrupted are indicated at the breakpoint.

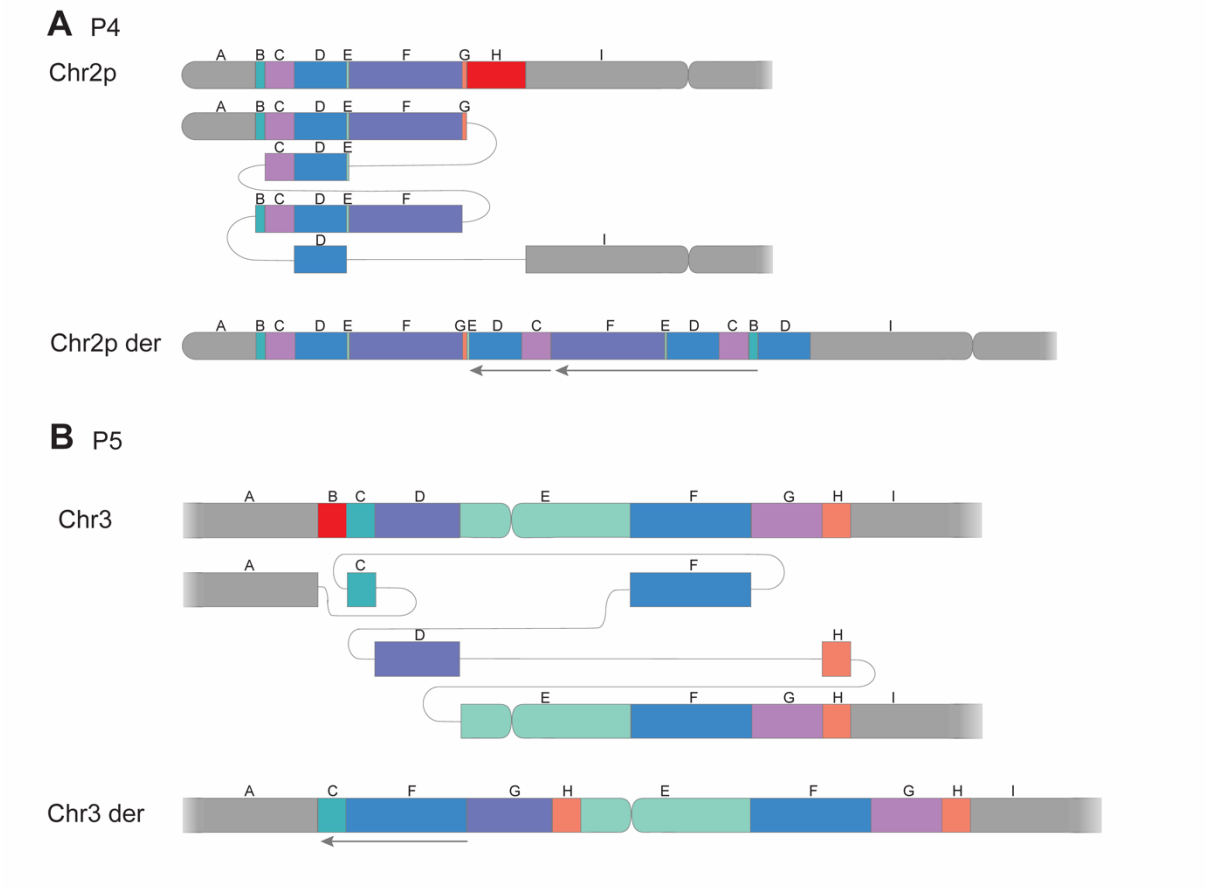

Figure 3: Subway plots of two complex rearrangements: A) A complex dup-trip-quad-trip-dup-del on chromosome 2 observed in P4. B) A clustered CNV on chromosome 3 detected in P5. Deleted segments shown in red, arrows mark inverted segments.

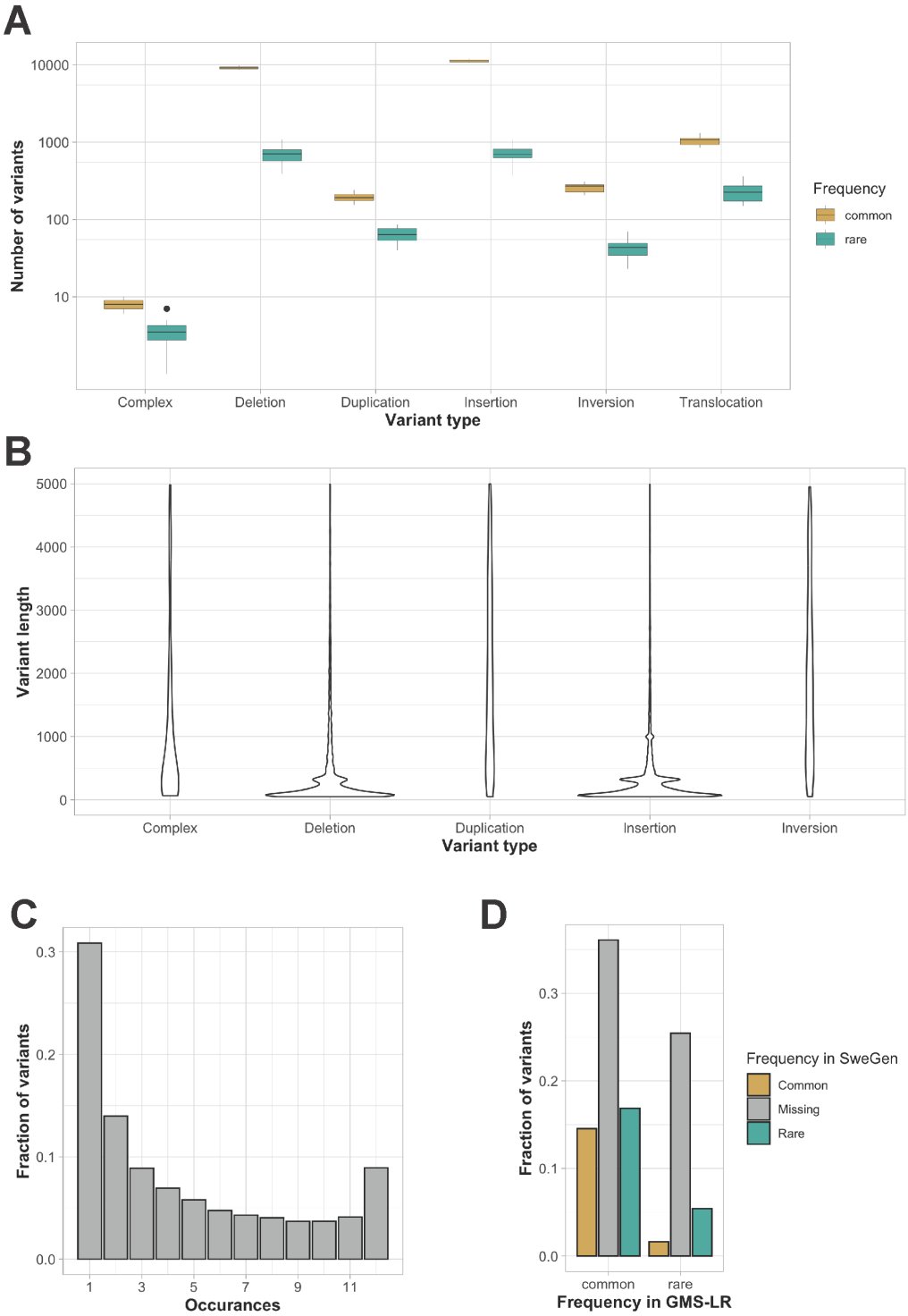

Figure 4: Characterization of background SV. A) Boxplot illustrating the number of SVs per type (common in yellow and rare in green). B) Violin plot of SV length per SV type (excluding SV > 5kbp). C) Allele frequency histogram. D) Comparison of allele frequencies between the GMS-lr cohort and SweGen srGS SV database.

Table 1: Karyotypes and mode of ascertainment of included cases.

Case ID	Investigations prior to lrGS	Karyotype	Ascertainment
P1	srGS	seq[GRCh38] 16p12.1p11.2(28462001_29430800)x1 dn seq[GRCh38] 4q34.3q35.2(181412301_190043700)x1 dn seq[GRCh38] 9p24.3p23(200001_13038400)x3 dn	Syndrome
P2	srGS	seq[GRCh38] 15q11.1q13.2(19799420-30095350)x3-4 dn seq[GRCh38] 15q13.2-q13.3(30521460-32201830)x2-3 dn	Syndrome
P3	srGS	seq[GRCh38] 22q13.33q13.33 (50624868_50626276)x0	Syndrome
P4	Karyotype SNP array	arr[GRCh37] 2p25.2p25.1(5689487_7379378)x3 arr[GRCh37] 2p25.2(5689487_6397700)x4 arr[GRCh37] 2p25.2p25.1(6397750_7379378)x3 arr[GRCh37] 2q21.2(133714063_134181189)x1 arr[GRCh37] 2q22.1(138458639_141918297)x1	Syndromic craniofacial condition
P5	Karyotype SNP array	arr[GRCh37] 3p14.3p14.2(54554106_59390517)x1 arr[GRCh37] 3q24(145318138_147828077)x3	Motor delay
P6	srGS MLPA	seq[GRCh37] Xq28(152990161_153045093)x1	Adrenoleukodystrophy
P7.1	Karyotype FISH	47,XX,t(1;10)(p36.2;q24),inv(2)(q32.2q33.2),i(?)(p10)	Recurrent miscarriages
P7.2	None	Father of P7.1	Healthy
P7.3	None	Mother of P7.1	Healthy

P8.1	Karyotyping FISH Array-CGH srGS linked read GS	46,X,inv(X)(p22.31q28)	Short stature
P8.2	Karyotyping FISH Array-CGH	46,X,inv(X)(p22.31q28), mother of P8.1	Healthy and of normal height
P9	Karyotyping FISH Panel	47,XX,t(X;9)(p22;q12)[27]/46,X,t(X;9)(p22;q12)[3]	Trombocytopenia (50×10^9 - 74×10^9/L), early menopause, learning difficulties, attention difficulties, epicanthus
P10	Karyotype FISH SNP-array	46,XY,r(21)(p11q22)[9]/46,XY,del(21)(q22.3)[4]/46,XY[12] arr[GRCh37]21q22.3(47431031_48100155)	Infertility and oligospermia

lrGS, long-read genome sequencing; srGS, short read genome sequencing; SNP array, single nucleotide polymorphism array; MLPA, multiplex ligation-dependent probe amplification; FISH, fluorescence in situ hybridization; Array-CGH, array comparative genomic hybridization.

Table 2: HGVS results of cases with identified BPJs subdivided into simple and complex rearrangements.

Sample ID	HGVS karyotype	SV structure
Translocations and inversions		
P1	chr4:g.qter_181414917delins[chr9:g.pter_13043905] chr16:g.28442001_29508000del	t(4;9) DEL
P7.1 P7.3	chr1:g.19783172_qterdelins[chr10:g.pter_95395335] chr10:g.pter_19783105delins[chr1:g.95395328_qter]	t(1;10)
	chr2:g.189694535_202576085inv	inv(2)
P9	chr9:g.pter_(75862011)delins[chrX:g.3044233_qter] chrX:g. 3044228_qterdelins[chr9:g.pter_(75862011)]	t(X;9)
Complex rearrangements		
P3	chr22:g.50618055_50626277delins[chr22:g.50624361_50624868inv]	DEL-INV-DEL
P4	chr2:g.7246507_7734100delins[chr2:g.5620671_6248808inv; chr2:g.5549092_7241289inv;chr2:g.5858296_6246303]	DUP-TRIP-QUAD-TRIP-DUP-NML-DEL
	chr2:g.132954122_141163928delins[chr2:g.133434126_137703180inv]	DEL-INV-DEL
P5	chr3:g.54518907_80981660delins[chr3:g.59407900_63614153inv; chr3:g.135767535_135891958;chr3:g.63614153_80981660inv; chr3:g.145600155_148110490]	DEL-INV-NML-DUP-NML-DUP
P6	chrX:g.153724706_153724706delins[chr9:g.97596764_. inv]	DEL-INS-DEL
P8.1	chrX:g.9768910_154113036delins[chrX:g.9420014_154208530inv]	DUP-INV-DUP
P8.2	chrX:g.9768910_154113036delins[chrX:g.9420014_154208530inv]	DUP-INV-DUP

Table 3: Junction characteristics of identified rearrangements, one case was only detected in T2T

Sample ID	BPJs	ChrA	posA	chrB	posB	Strand	Gene posA	Gene posB	Breakpoint insertions	Repeat A	Repeat B	Microhomology
Hg38												
P1	1	4	181414917	9	13043905	+/-	-	-	-	L1	L1	-
	1	16	28455071	16	29476612	+/+	-	-	-	*AluJo*	*SVA_D*	-
P3	2	22	50618055	22	50624868	+/-	-	***ARSA***	2nt	tigger1	MIR	-
		22	50624361	22	50626277	+/-	***ARSA***	***ARSA***	-	*Alu*sp	-	2nt
P4	4	2	5549092	2	5858296	+/-	*XR_001739261.1*	-	-	-	L1PA7	4nt
		2	5620671	2	7241289	+/+	-	-	9nt	-	-	-
		2	6246303	2	7734100	+/+	-	-	-	L1M5	-	2nt
		2	6248808	2	7246507	+/-	-	-	-	L1ME4b	-	3nt
	2	2	132954122	2	137703180	+/-	*NCKAP5*	-	-	MLT1K	L1PA16	1nt
		2	133434126	2	141163928	+/-	*NCKAP5*	*LRP1B*		-	-	11nt
P5	5	3	54518907	3	63614153	+/-	*CACNA2D3*	*SYNPR*	-	-	MamRep38	-
		3	59407900	3	135767535	+/-	*XR_002959675.1*	-	-	-	MER101	1nt
		3	63614153	3	145600155	+/-	*SYNPR*	-	-	MamRep38	L2A	-

		3	80981660	3	135891958	+/-	-	-	-	MSTA-int	L2a	1nt
		3	80981660	3	148110490	+/+	-	-	1nt	MSTA-int	MER58A	-
P6	2	9	97596764	X	153779639	+/-	*TMOD1*	-	-	-	-	1nt
		9	97598236	X	153724706	+/-	*TMOD1*	*BCAP31*	12nt	*Alu*SX1	-	-
P7.1	4	1	19783172	10	95395335	+/-	*TMCO4*	*SORBS 1*	7nt	MIRb	-	-
		1	19783105	10	95395328	+/-	*TMCO4*	*SORBS 1*	5nt	MIRb	-	-
		2	189694535	2	202576085	+/-	-	*ANKAR*	1nt	-	L1PA17	-
		2	189694533	2	202576083	+/-	-	*ANKAR*	-	-	L1PA17	1nt
P7.2	-	-	-	-	-	-	-	-	-	-	-	-
P7.3	4	1	19783172	10	95395335	+/-	*TMCO4*	*SORBS 1*	7nt	MIRb	-	-
		1	19783105	10	95395328	+/-	*TMCO4*	*SORBS 1*	5nt	MIRb	-	-
		2	189694535	2	202576085	+/-	-	*ANKAR*	1nt	-	L1PA17	-
		2	189694533	2	202576083	+/-	-	*ANKAR*	-	-	L1PA17	1nt
P8.1	2	X	9420014	X	154113037	+/-	XR_0017 55796.1	-	28nt	*Alu*Jr	*Alu*Sx1	-
		X	9768909	X	154208530	+/-	GPR143	-	31nt	*Alu*Sz6	*Alu*Jo	-
P8.2	2	X	9420014	X	154113037	+/-	XR_0017 55796.1	-	28nt	*Alu*Jr	*Alu*Sx1	-
		X	9768909	X	154208530	+/-	GPR143	-	31nt	*Alu*Sz6	*Alu*Jo	-

T2T												
P9	2	9	75862011*	X	3044233	+/-	-	-	-	(AATGG)n	ERVL-MaLR	2nt
		9	75862011*	X	3044228	+/-	-	-	990 nt	(TTCCA)n	ERVL-MaLR	-

BPJ, breakpoint junction; chr, chromosome; pos, position; nt, nucleotide; Bold text indicates a disease-causing gene. *=estimated location