


RESEARCH

Open Access



Combining chromosome conformation capture and exome sequencing for simultaneous detection of structural and single-nucleotide variants

Maria Gridina^{1,2,3,10*} , Timofey Lagunov^{1,2}, Polina Belokopytova^{1,2,3}, Nikita Torgunakov^{1,2}, Miroslav Nuriddinov^{1,2}, Artem Nurislamov^{1,2,10}, Lyudmila P. Nazarenko³, Anna A. Kashevarova³, Maria E. Lopatkina³, Stanislav Vasilyev³, Andrey Zuev³, Elena O. Belyaeva³, Olga A. Salyukova³, Aleksandr D. Cheremnykh³, Natalia N. Sukhanova³, Marina E. Minzhenkova⁴, Zhanna G. Markova⁴, Nina A. Demina⁴, Yana Stepanchuk^{1,2}, Anna Khabarova¹, Alexandra Yan^{1,2}, Emil Valeev^{1,2}, Galina Koksharova^{1,2,10}, Elena V. Grigor'eva¹, Natalia Kokh^{1,2,8}, Tatiana Lukjanova⁵, Yulia Maximova^{5,14}, Elizaveta Musatova⁶, Elena Shabanova⁷, Andrey Kechin^{2,8}, Evgeniy Khrapov⁸, Uliana Boyarskih⁸, Oxana Ryzhkova⁴, Maria Suntsova^{11,12}, Alina Matrosova^{11,12}, Mikhail Karoli¹⁰, Andrey Manakhov¹⁰, Maxim Filipenko⁸, Evgeny Rogaev^{10,13}, Nadezhda V. Shilova⁴, Igor N. Lebedev³ and Veniamin Fishman^{1,2,3,9,10*}

Abstract

Background Effective molecular diagnosis of congenital diseases hinges on comprehensive genomic analysis, traditionally reliant on various methodologies specific to each variant type—whole exome or genome sequencing for single nucleotide variants (SNVs), array CGH for copy-number variants (CNVs), and microscopy for structural variants (SVs).

Methods We introduce a novel, integrative approach combining exome sequencing with chromosome conformation capture, termed Exo-C. This method enables the concurrent identification of SNVs in clinically relevant genes and SVs across the genome and allows analysis of heterozygous and mosaic carriers. Enhanced with targeted long-read sequencing, Exo-C evolves into a cost-efficient solution capable of resolving complex SVs at base-pair accuracy.

Results Applied to 66 human samples Exo-C achieved 100% recall and 73% precision in detecting chromosomal translocations and SNVs. We further benchmarked its performance for inversions and CNVs and demonstrated its utility in detecting mosaic SVs and resolving diagnostically challenging cases.

Conclusions Through several case studies, we demonstrate how Exo-C's multifaceted application can effectively uncover diverse causative variants and elucidate disease mechanisms in patients with rare disorders.

*Correspondence:

Maria Gridina
gridinam@gmail.com
Veniamin Fishman
minja-f@yandex.com

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

Background

The genetic component is an essential factor in most health conditions. In particular, genetic variants dominate the field of rare diseases [1]. The most advanced approaches to human genome analysis, such as whole-genome long-read sequencing, are costly and thus cannot be broadly introduced in clinical practice (Additional File 1: Table S15). Routine genetic diagnosis methods include molecular cytogenetic analysis (such as array-based comparative genomic hybridization (aCGH)), karyotyping, and exome sequencing [2]. However, each of these methods is limited in resolution and spectrum of detectable variant types. For example, aCGH is a preferred method for the detection of copy number variants (CNVs), but fails to detect balanced structural variants (SVs) and single-nucleotide variants (SNVs) or short insertions and deletions (INDELs). Whole exome sequencing (WES) is powerful in detection of exonic SNVs, yet cannot identify SVs with breakpoints outside exome. Therefore, to choose an appropriate method of genomic diagnosis, it is required to predict the type of variant based on the patient's phenotype. This challenge asks for a very high qualification, and even for an experienced specialist it is not always possible to guess the genetic etiology of a disease based on the patient's phenotype [3]. Thus, there is a strong demand for novel methods of genetic assays with reasonable price and sensitivity to detect broad range of genomic variants.

The family of chromosome conformation capture methods, including the widely used Hi-C assay, utilizes proximity ligation techniques to analyze chromatin contacts across the genome [4]. Applications of these methods uncover essential properties of genome architecture, including characterization of chromatin compartments [5] and activity of loop extrusion machinery [6], through the analysis of patterns observed in Hi-C maps. Furthermore, studies by others [7] and our own research [8] have demonstrated that structural variants, including inversions and translocations, are detectable on Hi-C maps, suggesting the potential to repurpose Hi-C for the detection of structural variants. However, as a whole-genome assay, Hi-C requires significant sequencing depth to identify SNVs, which can be cost-prohibitive. Therefore, to effectively detect both SVs and SNVs affecting clinical genome, it is necessary to integrate Hi-C with WES. To address this challenge, we present a novel method based on chromatin conformation capture that can be used to simultaneously detect structural variants and point mutations in the human genome.

Methods

Human samples

We collect a cohort of 66 human samples, comprising 28 females, 37 males, and two immortalized cell line samples: K562 and A549. 58 donors were enrolled after visiting medical centers. Additionally, 6 healthy donors volunteer to enroll in the study as controls. All samples were collected from authorized medical centers listed in Additional File 1: Table S1. All samples were profiled with Exo-C, and 54 samples were profiled with at least one of the following methods: aCGH, WGS, or karyotyping. Methods applied to each sample are specified in the Additional File 1: Table S1. All these samples were used to benchmark Exo-C efficiency.

Five of these 66 samples were designed as Nanopore cohort based on the Exo-C analysis results; sample IDs and additional information about each sample is provided in Additional File 1: Table S6 and S1, respectively.

The study was approved by the local ethics committee of the Institute of Cytology and Genetics (protocol number 17, 16.12.2022) and the local ethics committee of the Tomsk National Research Medical Center (protocol number 15, 28.02.2023). An Informed Consent was obtained from all patients or their parents/representatives included in the study.

Collection of biological material

Blood samples for Exo-C were collected as described in [9]. Briefly, blood samples were collected in EDTA-coated tubes. Erythrocytes were lysed using RBC lysis buffer (BioLegend) according to the manufacturer's protocol. Following lysis, cells were washed twice with PBS, and resuspended in DMEM at a final concentration of 1×10^6 cells/mL.

iPS cells and fibroblasts were generated previously [10–12] and obtained from the Collective Center of ICG SB RAS “Collection of Pluripotent Human and Mammalian Cell Cultures for Biological and Biomedical Research” (<https://ckp.icgen.ru/cells/>; http://www.biores.cytogen.ru/brc_cells/collections/ICG_SB_RAS_CELL). Sample source for each case is provided in Additional file 1: Table S1.

Exo-C experiment

Exo-C protocol includes a 3 C part followed by exome enrichment. The 3C part was performed either following DNase I protocol [9], or S1 protocol [13]. Briefly, 2.5 million cells were fixed in 1% formaldehyde, lysed, and digested overnight with DNase I or S1 nuclease. Digested DNA ends were marked with biotin-14-dCTP and ligated overnight using T4 DNA ligase. Formaldehyde crosslinking was reversed by incubation with Proteinase

K at 65 °C overnight, followed by ethanol precipitation. Biotin from unligated ends was removed by T4 polymerase treatment. KAPA HyperPlus kit was used for NGS library preparation. Biotin-filled DNA fragments were pulled down using Dynabeads MyOne Streptavidin C1 beads before library amplification. Exome enrichment was performed using KAPA HyperExome Probes and KAPA HyperCap kit according to the manufacturer's instructions.

iPS cells generation and differentiation

Reprogramming of human peripheral blood mononuclear cells (PBMC) into iPS cells was performed by episomal vectors cocktail contained: reprogramming vectors expressed OCT4 (addgene #41,813), MYC and LIN28 (addgene #41,855), shRNA against p53 (addgene #41,856), SOX2 and KLF4 (addgene #41,814), EBNA1 (addgene # 41,857), according to the previously described method [14]. Briefly, 5×10^5 PBMC were transfected with 1 µg of each vector using the Neon Transfection System (Thermo Fisher Scientific) under optimized conditions (1650 V, 10 ms, 3 pulses). Transfected cells were plated onto mitomycin C-treated CD-1 mouse embryonic fibroblasts and incubated overnight in complete StemPro™-34 medium supplemented with cytokines. The next day, the medium was replaced with N2B27 medium (DMEM/F-12 with HEPES, 1% N-2 Supplement, 2% B-27™ Supplement, 1% GlutaMAX-I, 1% MEM NEAA, 1% Pen-Strep, and 0.1 mM 2-mercaptoethanol) containing 100 ng/mL bFGF, which was refreshed daily. On day 9, the medium was switched to iPS cells medium (DMEM/F12, 20% KnockOut Serum Replacement, 1% GlutaMAX-I, 1% MEM NEAA, 1% Pen-Strep, and 0.1 mM 2-mercaptoethanol) supplemented with 100 ng/mL bFGF. By day 16, colonies exhibiting typical iPS cells morphology were manually picked and expanded in iPS cells medium supplemented with 10 ng/mL bFGF. Cells were maintained at 37 °C in 5% CO₂ and passaged mechanically at a 1:5 split ratio.

As controls, we used two previously established iPS cell lines derived from a healthy male donor [10]. These lines were extensively characterized through cytogenetic analysis, immunofluorescence staining, and teratoma formation assays. Both the control iPSCs and those derived from Patient 10 (P10) were differentiated into primitive streak-like cells as described [15]. Briefly, iPSCs were maintained in mTeSR™1 medium (StemCell Technologies) under feeder-free conditions on Matrigel-coated plates. To enhance cell survival during passaging, 10 µM Y-27632 (a p160 Rho-associated coiled-coil kinase (ROCK) inhibitor) was added to the medium. For differentiation, cells were treated for 48 h with 6 µM CHIR99021 in DMEM supplemented with 100 µg/mL

ascorbic acid, inducing primitive streak formation in a monolayer culture.

Spectral karyotyping (SKY)

Multicolor FISH analysis for verification of predicted translocations was performed on patient's metaphase chromosomes using SkyPaint™ DNA Kit H-10 for Human Chromosomes (Applied Spectral Imaging, Israel). Visualization of hybridization signals was done using Olympus BX53 microscope (Olympus Life-science) equipped by HYPERSPECTRAL V8.1 SYSTEM SD 300 and CCD-камера: B/W CAMERA BV 300 (Applied Spectral Imaging, Israel). Results interpretation was performed by GenASIs software, V8.2 (Applied Spectral Imaging, Israel).

Chromosomal microarray analysis

Array-based comparative genomic hybridization (aCGH) was performed using SurePrint G3 Human CGH 8 × 60 K microarrays (Agilent Technologies, Santa Clara, CA, USA) with 41 kb overall median probe spacing according to the manufacturer's recommendations. Labeling and hybridization of the patient's and reference DNA (#5190–3797, Human 230 Reference DNA Female, Agilent Technologies, Santa Clara, CA, USA) were performed using enzymatic labeling and hybridization protocols, v.7.5 (Agilent Technologies, Santa Clara, CA, USA). Array images were acquired with an Agilent SureScan Microarray Scanner (Agilent Technologies, Santa Clara, CA, USA). Data analysis was performed using CytoGenomics Software, v.5.1.2.1 (Agilent Technologies, Santa Clara, CA, USA) and the publicly available Database of Genomic Variants (DGV) resources. Human genome assembly 19 (hg19) was used to describe the molecular karyotype revealed by aCGH.

In addition, for some samples the CytoScan HD array (Affymetrix, USA) was applied to detect the CNVs across the entire genome following the manufacturer's protocols. Microarray-based copy number analysis was performed using the Chromosome Analysis Suite software version 4.0 (Thermo Fisher Scientific Inc.). Detected CNVs were totally assessed by comparing them with published literature and the public databases: Database of Genomic Variants (DGV) [16], DECIPHER [17] and OMIM [18]. Genomic positions refer to the Human Genome February 2009 assembly (GRCh37/hg19).

Oxford nanopore breakpoints sequencing

For nanopore sequencing, high-molecular DNA from iPS cells and blood cell samples was isolated. WGS and AS libraries were prepared using the recommended protocol for LSK109 kit from the manufacturer (Oxford Nanopore, UK). Libraries for Cas9-targeted sequencing were

prepared using nCATS protocol [19] with modifications. The sgRNAs were divided into two distinct pools similar to the “tiling” approach: the first pool of sgRNAs targeted regions spanning 10–15 kb upstream and 5–7 kb downstream from the expected breakpoints, while the second pool targeted regions located 5–7 kb upstream and 10–15 kb downstream (Additional File 1: Table S8), the rest of the protocol was proceeded according to nCATS method. The regions selected for AS targets included flanking areas spanning 30–40 kb from the expected breakpoints (upstream and downstream) according to the official Oxford Nanopore guidelines. Sequencing was performed on a GridION device with FLO-MIN106D flow cells (Oxford Nanopore, UK). Reads with Q-score above 9 (> 90% accuracy) were taken into further analysis. Minimap2 [20] and NanoSV [21] software were used for the alignment and variant calling processes.

Transcriptome analysis

For transcriptome analysis, total RNA was extracted from the iPS cell differentiated to primitive streak cells and subjected to the oligo-dT-based RNA-sequencing. The resulting data were visually inspected in the IGV and analyzed using salmon [22] and standard DESeq2 pipeline [23].

Exo-C data processing

All data were processed using a pipeline based on the juicer toolbox [24], which outputs a list of pairs, hic and cool maps made from pairs by cooler [25], as well as quality statistics described in [9]. All analyzes were performed based on the human hg19 genome build.

Read coverage distribution analysis

To allow comparison between coverage of Hi-C data produced by S1, DNase I, DpnII enzymes, fastq files were trimmed to 50 bp of read length and subsetting to 65 mln reads. Equalized data were aligned with BWA (v 0.7.17) [26] on human hg38 genome build and converted to read coverage tracks with deepTools 3.5.1 bamCoverage (with the option `–binSize 1`) [27]. To construct profile plot of read density across KAPA HyperExome capture probes we excluded the segments that have DpnII sites in them and then computed coverage on $\pm 1\,000$ bp distance from center of selected capture probe segments with deepTools 3.5.1 computeMatrix reference-point (with the option `–binSize 1`) and visualized it with deepTools 3.5.1 plotProfile.

To build distributions of read coverage depth exome-wide we calculated coverage sums in each capture probe segment for the same coverage tracks of the equalized data using pyBigWig 0.3.18 [27] and NumPy 1.21.6 [28]. The histograms of distributions were plotted using matplotlib 3.5.3 [29] and seaborn 0.13.0 [30].

Structural variants calling

We performed structural variants calling for all samples listed in Additional file 1: Table S1.

Translocations

For translocation calling, one of two control samples was used (male control or female control). A control sample is the sum of samples of the same sex (22 female samples and 30 male samples). We combine the results of two methods for translocation calling: gradient pattern search method (works better for translocations of fragments above ~ 100 kb) and line pattern search method (works better for translocations of fragments from ~ 20 to ~ 200 kb).

Gradient pattern search method For each resolution of the Exo-C map (resolutions: 4096 kb, 256 kb, 16 kb, descending order), we define “dots of interest” (DOIs) as elements of Exo-C contact matrix that fulfill the following conditions:

- 1) DOI must have non-zero Hi-C count in both sample and control;
- 2) the difference of logarithms of binomial probability mass function of sample and control for the DOI is below threshold (different for each resolution);

$$DAr_{ij} = \log_{10} \left(\text{Bin} \left(C_{ij}^s, D^s, \frac{C_{ij}^c}{D^c} \right) \right) - \log_{10} \left(\text{Bin} \left(C_{ij}^c, D^c, \frac{C_{ij}^s}{D^s} \right) \right) < Thr(res) \quad (1)$$

where *Bin* is binomial probability mass function, C_{ij}^s , C_{ij}^c is counts at coordinates (i, j) for sample (*s*) and control (*c*); and D^s , D^c is sequencing depth for sample and control. To compute D^s and D^c , we excluded from both control and sample Exo-C maps those dots, which contain zero value in either case or control; $\frac{C_{ij}^c}{D^c}$ is contact (i, j) probability estimation; and *Thr(res)* is threshold that depends on map resolution.

$$Thr(res) = \begin{cases} -4.0, & res = 4096kb \\ -2.0, & res = 256kb \\ -1.0, & res = 16kb \end{cases}$$

Next, for each DOI we test the hypothesis that this DOI represents contact of translocation breakpoints. Following this hypothesis, the DOI is formed by loci that are at distance 1 bin from breakpoint, their neighbors are at distance 2 bins from breakpoint, and etc. Under this assumption, each dot near DOI receives the contact probability according to the distance from the breakpoint, $P_{art_cis}(m)$. Under alternative assumption, there is no translocation, and dots near DOI are interchromosomal contacts with trans-contact probability $P_{control_like}$.

We estimate $P_{art_cis}(m)$ and $P_{control_like}$ based on dependence of contact frequencies from distance and other control sample statistics, see details in Supplementary Note 1.

Next, we compute the cumulative sum of $P_{control_like}$ and $P_{art_cis}(m)$ for all dots near DOI. For DOI with coordinates a_0, b_0 , we identify neighboring dots located

$$ArLL_j = \log_{10} \left(\text{Bin} \left(\sum_{i \in ChrA} C_{ij}^c, D_{ChrA-ChrB}^c, \frac{\sum_{i \in ChrA} C_{ij}^c}{D_{ChrA-ChrB}^c} \right) \right) - \log_{10} \left(\text{Bin} \left(\sum_{i \in ChrA} C_{ij}^s, D_{ChrA-ChrB}^s, \frac{\sum_{i \in ChrA} C_{ij}^c}{D_{ChrA-ChrB}^c} \right) \right), j \in Chr \quad (3)$$

within each of four rectangles anchored at DOI. We define upper right rectangle as all dots with coordinate $a_0 + i, b_0 + j, i \in [0, k], j \in [0, n]$; upper left rectangle as all dots with coordinate $a_0 - i, b_0 + j, i \in [0, k], j \in [0, n]$; and similarly lower left and lower right rectangle. For all k and n values between 1 and 32 and each rectangle location (one of: upper right, upper left, lower right, lower left) we compute *ArLG* (Artifacts Level for Gradient pattern search) statistics (Additional File 2: Fig. S2E):

$$ArLG = \sum_{i,j \in W} \left(\log_{10} \left(\text{Bin} \left(C_{ij}^s, D^s, P_{art_cis}(m) \right) \right) - \log_{10} \left(\text{Bin} \left(C_{ij}^s, D^s, P_{control_like} \right) \right) \right) \quad (2)$$

where W is window, i.e., set of indices specific for the combination of rectangle location, k and n parameters; $\text{Bin} \left(C_{ij}^s, D^s, P_{art_cis}(m) \right)$ is probability for contact (i, j) to belong to the binomial distribution expected for cis-contacts of neighboring loci with distance m ; and $\text{Bin} \left(C_{ij}^s, D^s, P_{control_like} \right)$ is probability for contact (i, j) to belong to the binomial distribution expected for trans-contacts; for $P_{art_cis}(m)$ and $P_{control_like}$, see calculation details in Supplementary Note 1.

For each DOI, we keep only one window that gives the highest *ArLG* score.

After that, three filtering steps were applied:

- 1) $ArLG > 0$ (means that it's more likely for contacts in a window to belong to the cis-contacts than to trans contacts);
- 2) If one DOI belongs to the window of another DOI, only the DOI (with its window) that has greater *ArLG* is passed;
- 3) $\log_{10}(ArLG) > 0.917 * \log_{10}(contacts_sum) - 0.92$, where *contacts_sum* is sum of contacts in a window (constants were estimated from modeled translocations);

The DOIs and their windows, which passed all below steps, are considered as translocations.

Line pattern search method First steps of Line pattern search method are similar to the previous one (until dots of interest are defined; resolutions: 256 kb, 16 kb, 1 kb). For each DOI, we compute the following metric

$$D_{ChrA-ChrB}^s = \sum_{i \in ChrA, j \in ChrB} C_{ij}^s \quad (4)$$

$$D_{ChrA-ChrB}^c = \sum_{i \in ChrA, j \in ChrB} C_{ij}^c \quad (5)$$

where $\sum_{i \in ChrA}$ means the summation by Exo-C contact matrix columns that belongs to the $ChrA$; $D_{ChrA-ChrB}^s$ and $D_{ChrA-ChrB}^c$ are the sum of contacts for sample and control that belong to the chromosome pair $ChrA - ChrB$

(after same bins cutting, see Supplementary Note 1). For the column that is less likely to belong to sample binomial than to control binomial, the $ArLL > 0$. Every two columns were combined as one translocation when their $ArLL > 6$ and distance between them less than 5 bins. For the new column combinations the *ArLLs* were recalculated (with summation by combined columns). Additionally, the same measure was calculated by rows for every column combination to find the maximum (position that is nearest to insertion). At the end, the sample contact sum dependent filter was applied: $\log_{10}(ArLL) > \exp(1.73 * \log_{10}(contacts_sum))$ (constant was estimated from modeled translocations).

Inversions

For inversion calling, we used the same control samples as for translocation calling. We called inversions at different resolutions (1 MB, 250 kb, 100 kb, 10 kb) independently and then merge all these predictions. To speed up the calculation, we have determined the minimum and maximum inversion sizes at each resolution that we are looking for (Additional File 1: Table S7).

For each resolution at the first stage, we computed the matrix of difference between control and experiment sample as described above in (Eq. 1) formula. Further, for each dot (i, j) , we computed the metric that shows how probable it is that this dot is an inversion breakpoint. This metric allows finding “butterfly”-like inversion pattern at hi-c map. We called it “sweet” metric:

$$\text{SweetMetric} = \text{LeftSum} + \text{RightSum} \quad (6)$$

where

$$\text{LeftSum} = \sum_{i,j \in W_{\text{left}}} \text{D}Ar_{ij} \quad (7)$$

$$\text{RightSum} = \sum_{i,j \in W_{\text{right}}} \text{D}Ar_{ij} \quad (8)$$

where $\text{D}Ar_{ij}$ is the difference of logarithms of binomial probability mass function of sample and control for the dot (Eq. 1). W_{right} and W_{left} are isosceles triangles with the vertex in (i, j) dot (Additional File 2: Fig. S2B) and edges equal to sweet size. We used different sweet sizes for different resolutions (Additional file 1: Table S7).

Then we computed z-score value for each dot. We picked up the threshold for each resolution and chose all z-score values lower than this threshold. After that, we merged all overlapping coordinates at current resolution using the (i, j) coordinates of dot with the lowest z-score value as a final inversion breakpoints. Finally, we merged all predicted inversions at different resolutions and clarified each breakpoint to 10 Kb resolution. To do this we iteratively calculated “sweet” metric for dots around 10 bins from breakpoint until we reached the highest resolution 10 Kb.

CNV genotyping: Exo-C data were aligned to the hg19 human genome assembly [31] using BWA (v 0.7.17) [26]. Then, BAM file sorting and indexing was performed using samtools (v 1.6) [32]. We used the following tools for CNV detection based on sequence alignments: CNVkit (v 0.9.10) [33], CoNIFER (v 0.2.2) [34], GATK (v 4.4.0.0) [35]. All tools were used with default parameters. We used 65 samples in all tool's runs, out of which 19 or 17 (after filtration of CNVs that are less than 100,000 bp) samples formed lists of predictions to compare with control lists. Additionally, in CNVkit run one sample (P31) out of 65 was used as a normal to construct reference. We used DGV Gold database to exclude tool's predictions that are possibly related to population CNV data and are not sample-specific [16].

Estimation of algorithms performance

When scoring performance of computational algorithms for SV calling in clinical samples, we used SV calls from validated methods (aCGH, karyotyping) as ground truth. Thus, we did not include reported SVs beyond the

resolution limit of the methods (3–5 Mb for karyotyping, ~100 kb for aCGH) in the scoring. For A549 and K562 samples we adjusted our benchmarking strategy: for SVs, we excluded translocations and inversions smaller than 3 kb, the detection threshold of Exo-C; for CNV benchmarking, we disregarded regions that do not overlap with exons and INDELs smaller than 150 bp, which is under the read length threshold.

EagleC performance estimation

We run EagleC [36] with standard parameters (threshold equals to 0.9) and calculate F1-score for each type of rearrangement for 65 human samples.

In the case of translocations, we calculated all predicted rearrangements between a pair of chromosomes involved in a known translocation as one true positive value. For predicted translocations between pairs of chromosomes that do not have known rearrangement, we used the following statistics. If EagleC predicted one translocation for a pair of chromosomes, or if there were several predicted translocations between one pair of chromosomes with the distance between breakpoints more than 3 Mb, we counted each of the reported translocations as one false positive value. If the distance between breakpoints was less than 3 Mb, they were not included in assessment.

In the case of inversions, we defined true positive values as predicted rearrangements with breakpoints overlapping the 5 Mb region around known inversion breakpoints. All other predicted inversions we calculated as false positive values if the distance between breakpoints was more than 3 Mb.

Finally, in the case of CNVs, we defined true positive values as predicted rearrangements with breakpoints overlapping the 250 kb region around actual breakpoints. All other predicted CNVs we defined as false positive values if the distance between breakpoints was more than 500 kb.

Scoring algorithm for inversion calling developed in this paper

To estimate performance of the algorithm for inversion calling, we used F1-score. We deleted repeated inversions between samples. We called all predicted inversions as true positive values if they meet following conditions:

$$1) \text{chrom}_{\text{pred}} = \text{chrom}_{\text{real}}$$

Where $\text{chrom}_{\text{pred}}$ is chromosome of predicted inversion breakpoints and $\text{chrom}_{\text{real}}$ is chromosome of real inversion from experimental or modeled data.

$$2) \begin{cases} \text{brp1}_{\text{pred}} \in [\text{brp1}_{\text{real}} - \text{err}; \text{brp1}_{\text{real}} + \text{err}] \\ \text{brp2}_{\text{pred}} \in [\text{brp2}_{\text{real}} - \text{err}; \text{brp2}_{\text{real}} + \text{err}] \end{cases}$$

Where *brp1* and *brp2* are coordinates of inversion breakpoints for real and predicted data. *err* = 3 Mb for experimental data and *err* = 300 Kb for simulated data.

Other predicted inversions we called false positive values if their size was more than 3 Mb for experimental data and more than 20 kb for simulated data.

We defined all known inversions, which weren't predicted, as false negative values.

Algorithm for translocations calling

F1-score was used for estimation performance of translocation calling. Two different methods of true positives (TP) calculations were used for experimental samples and simulated samples. For experimental samples, we mark a call as TP if the chromosome pair was correct. For simulated samples, we mark a call as TP if the predicted intervals of translocation for both chromosomes intersected with the simulated translocation interval. For fragment of the chromosome *i* of the length *L* inserted into position *x* of chromosome *j* we define simulated translocation interval answer as rectangle on Exo-C contact map containing all contacts of translocated fragment with loci on chr_j with coordinates $x \pm 5 * L$ (Additional File 2: Fig. S2 C). If several calls were marked as TP and belong to the same chromosome pair, all of them were counted as single TP. All calls that were not marked as TP were marked as false positives (FP). If FP answers in more than 3 samples pointed at exactly the same coordinates on the same chromosome pair, they were filtered out. All translocations that were not called were marked as false negatives (FN).

CNVkit, CoNIFER, GATK

Preparation of known CNVs sets To compose CNVs control lists for tools evaluation, we classified all CNVs known to be in our samples by the source of information. This way we obtained two lists: high-confidence (clinically relevant CNVs confirmed by qPCR or other methods) and low-confidence (reported in aCGH data but not confirmed by orthogonal methods). After that we excluded repeats within lists, i.e. removed all CNVs that have intersecting boundaries. We also produced lists with filtered out CNVs whose size is less than 100,000 bp.

We evaluated the quality of predictions in terms of precision and recall. We defined prediction as TP if its coordinates intersected with CNVs from control lists (allowing errors of 100,000 bp, i.e., each control CNV border was extended by 100,000 bp). We also evaluated optional conditions: counting as TP only predictions which intersections with control had Jaccard Index greater than 0.5; exclusion of FP which intersections with DGV gold

database had Jaccard Index greater than 0.5, or which coordinates were inside those of DGV gold database entries. Jaccard Index was calculated as: (range of intersection)/(range of union).

Precision was calculated as a proportion of all predictions marked as TP out of all predictions. To calculate recall we marked all TP, which correspond to one known CNV, as one TP, and assessed the proportion of such predictions out of known CNVs.

Single nucleotide variants calling and annotation

SNV calling, annotation, and filtration was performed as described previously [37]. Reads were aligned to the human reference genome (hg19) using BWA-MEM algorithm (BWA v.0.7.17) [26]. PCR-duplicates were removed with PicardTools [38]. Base recalibration and haplotype calling were performed using GATK v.3.3 [39]. SNVs were annotated using ANNOVAR [40] software and filtered based on the following criteria: genotype quality score >20, maximum population allele frequency <0.01%, not reported as benign in ClinVar or Leiden Open Variation Database.

After filtering, the pathogenicity of each variant was assessed according to the recommendations of the American College of Medical Genetics and Genomics (ACMG) and the Association for Molecular Pathology [41].

Estimating the percentage of cells carrying a translocation in the cell mixture

To estimate the percentage of cells carrying a translocation in the in vitro cell mixtures, we used SNV data. We used GATK to call SNVs in Exo-C data for the mixture and in the NGS datasets for both types of cells present in the mixture, separately. We then identified homozygous SNVs unique to the cell type carrying a translocation using *bcftools isec* [42] and calculated reference and alternative allele counts for each of them in the Exo-C mixture data using *samtools mpileup* [32]. The final percentage was calculated by dividing the sum of alternative allele counts by the sum of all allele counts for each of the chosen SNV.

Comparison of SNV recall rate between Exo-C and reference NGS methods

To evaluate Exo-C's SNV recall rate, we compared Exo-C K562 data we generated to public reference datasets for K562, namely WGS data from [7], Hi-C data from [43] and [44] and Repli-seq data from [45]. We called SNVs in all of these datasets using GATK and then filtered out the SNVs outside the exome regions using *bcftools*.

We employed the following strategy for SNV calling comparison. First, we select one of four SNV datasets as “reference.” Next, we use the remaining three datasets to construct Golden Standard SNV set. This was done by intersecting the SNV calls from the three selected datasets. We then intersected each of four versions of the Golden Standard SNVs with one remaining reference SNV dataset and with Exo-C SNV dataset. The percentage of SNVs in the intersection is shown in Additional File 1: Table S9. All intersections were performed using *bcftools isec*.

SNV calling benchmark in A549 was performed similarly, however only one public dataset was used [7] and all SNVs overlapping target enrichment regions from this dataset were defined as A549 Golden Standard.

Results

Development of the Exo-C: a chromosome conformation capture assay with enriched exome representation

Recently, chromosome conformation capture techniques such as Hi-C were utilized to detect structural variants in the genomes of patients with hereditary diseases [46]. We propose several modifications to this method, allowing to achieve high and uniform coverage of exome sequences while retaining information about chromatin contacts genome-wide (Fig. 1A and 1B). The most essential protocol modifications compared to conventional Hi-C method [47] include using an improved version of DNase I or S1 Hi-C [9, 13] and introducing an exome enrichment step based on a common clinical exome panel (Methods). We designed the resulting protocol as Exo-C, *Exome-captured Hi-C* analysis.

Compared with conventional Hi-C, Exo-C produces similarly high-quality data (Fig. 1B; Additional File 1: Table S1). Expectedly, using S1 or DNase I enzymes allows obtaining much more uniform exome coverage compared to DpnII (Fig. 1C and 1E). Moreover, Exo-C results in the exome enrichment level only slightly below conventional exome capture data, obtained using the

same enrichment panel (Fig. 1D). These results were reproduced by applying Exo-C to fresh blood samples, cultured fibroblasts or iPS cells (Additional File 1: Table S1). Finally, comparative analysis of SNVs called from various K562 and A549 short-read sequencing dataset reveals that Exo-C performs comparably to other established methods in SNV detection (Additional File 1: Table S9). Concordant with these results, when applying Exo-C to find rare clinically-significant variants in a cohort of patients with monogenic diseases (detailed below), we were able to validate 11 out of 11 detected variants using Sanger sequencing (Additional File 1: Table S13). Thus, combining chromosome conformation capture with exome enrichment allows obtaining high-quality data to study SNV in exome and chromatin contacts genome-wide.

Exo-C profiling of 66 human samples and 750 computational models of chromosomal rearrangements

To study the efficiency of the Exo-C, we applied it to a cohort of 66 human samples, comprising 28 females, 37 males, and two cell line samples: K562 and A549.

Prior to Exo-C analysis, the majority of these samples underwent specific molecular and/or cytogenetic profiling (refer to Fig. 2A and Additional File 1: Table S1). The cohort encompassed a diverse array of genetic anomalies: 23 balanced translocations, 6 inversions (excluding clinically irrelevant pericentric inversions in heterochromatic regions, undetectable by short-read NGS), 86 CNVs, and several exonic SNVs of clinical significance (detailed in Fig. 2A and Additional File 1: Table S1). Additionally, the study included 6 healthy donors with normal karyotype as controls subjects. A subset of the probands presented complex SVs, previously characterized by FISH and conventional karyotyping (see below). The aggregated Exo-C dataset that includes all samples, derived from approximately 2 billion reads, incorporates about 1 billion Hi-C contacts, making it one of the most extensive human capture Hi-C datasets currently available.

(See figure on next page.)

Fig. 1 Exo-C achieves high and uniform coverage of exome sequences while retaining information about chromatin contacts genome-wide.

A Exo-C Experiment Scheme. The process begins with chromatin fixation and isolation, akin to conventional 3C methods. Subsequently, the chromatin is digested using sequence-agnostic enzyme (DNase I or S1 nuclease). After proximity ligation, products are first selected to contain ligation junctions and then exonic sequences are enriched using hybridization-based target capture panel. The resulting sequences facilitate the identification of structural variations (SVs) across the genome, single nucleotide variations (SNVs) in the exome, and alterations in spatial contacts of promoters. **B** Representative interaction heatmap of iP65 cells obtained using Exo-C protocol (above the diagonal line) and S1 Hi-C protocol (below the diagonal line). Red lines correspond to the enriched exonic sequences. **C** Exome-wide histograms of read coverage depth. Sums of read coverage depth are calculated for segments of all exome capture probes. **D** Comparative analysis of exome coverage enrichment in Exo-C (blue) and Whole Exome Sequencing (WES, orange) across various batches. Each data point represents an individual sample (Exo-C and WES samples are independent). The Y-axis quantifies the enrichment ratio, defined as the average coverage within the exome relative to the average coverage outside of the exome. **E** Read density across segments of exome capture probes, which do not intersect with DpnII restriction sites. The center of the plot is the center of each selected exome capture probe region

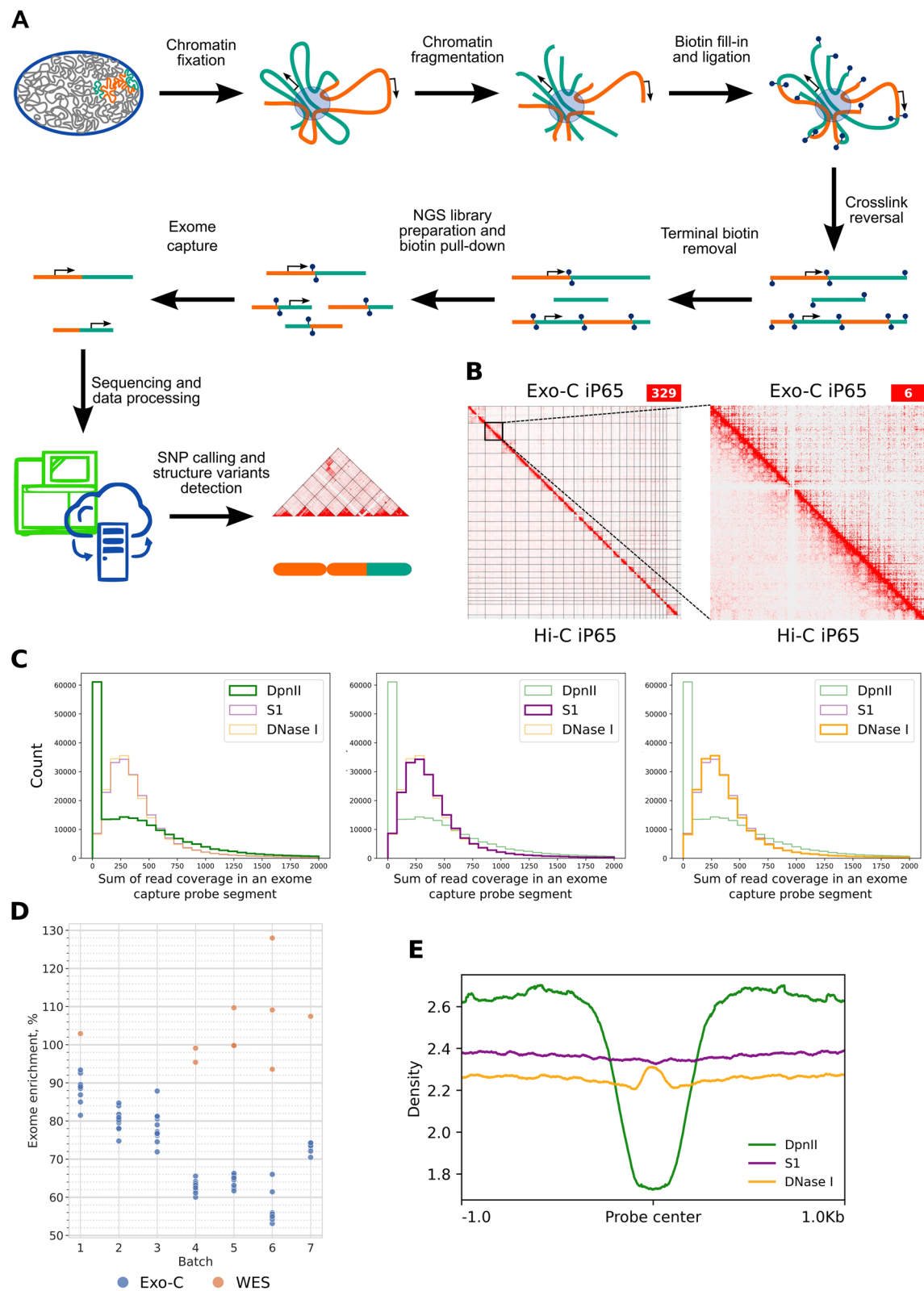


Fig. 1 (See legend on previous page.)

In this study, balanced SVs were initially identified via microscopy-based methods, typically involving rearranged fragments exceeding several megabases in size and the breakpoints were defined at a resolution of cytobands (detailed in Additional File 1: Table S1). To evaluate the Exo-C protocol's effectiveness at finer resolutions, we expanded our analysis to include a range of SV sizes. This was achieved by simulating Exo-C maps for SVs of varying dimensions, utilizing the Charm [48] framework for in silico generation.

Examination of Exo-C maps readily revealed megabase-scale SVs, as evidenced in Fig. 2B. Despite a pronounced bias towards exome sequences, these large-scale SVs induce marked alterations in chromatin contact patterns across entire chromosomes. In contrast to WGS data, where only reads adjacent to the breakpoint are indicative of SVs, Exo-C maps demonstrate altered contact patterns at loci situated millions of base pairs from the breakpoint, thereby signifying the presence of SVs (refer to Fig. 2B).

Despite being visually distinguishable, SVs were not called using the state-of-the-art chromosome conformation capture data analysis tool such as EagleC (*F1*-score for translocations equals to 0.15, *F1*-score for inversions equals to 0.03, *F1*-score for CNV equals to 0 since no CNV were reported). To perform unbiased, automatic calling of chromosomal rearrangements and allow detection of smaller SVs that are not always visible on whole-genome Exo-C maps, we developed dedicated computational algorithms detecting translocations and inversions, and benchmarked existing CNV-detecting tools (Methods).

Benchmarking Exo-C callers on A549 and K562 cells

First, we evaluated Exo-C caller performance using information about SVs in K562 and A549 genomes reported previously [7]. In A549 cells Exo-C caller detected all four reported translocations and made six additional calls, which were initially classified as false positives, resulting in 100% recall and 40% precision. In contrast, EagleC tool applied for SVs detection in Exo-C data reported only one out of four known translocations along with 5 additional calls. Notably, Exo-C maps showed a complex SV pattern involving chromosomes 10, 14, 19, 18, and 20 (Additional File 2: Fig. S5). Intriguingly, a segment of chromosome 19 identified by the Exo-C translocation caller had been previously reported as duplicated in A549 cells [7]. These observations suggest that the six reported breakpoints may represent a complex genomic rearrangement that was not resolved in previous studies. Thus, the aforementioned 40% precision of Exo-C may be an underestimate.

In K562 cells, Exo-C detected 12 out of 18 known translocations, with an additional 55 calls classified as false positives (precision ~18%, recall ~67%). We note that for more than half of 18 known genomic rearrangements in K562 cells, fragments involved in translocations are smaller than 500 kb or have an unknown insertion size, indicating limited recall for small insertions (see detailed analysis below and in Fig. 2E). Many false positives in K562 also display complex SV patterns, indicating a likely underestimation of precision (Additional File 2: Fig. S5). Benchmarking EagleC confirms that the tool cannot handle Exo-C data as efficient as dedicated caller, detecting 4 out of 18 known translocations, with 2 additional calls.

(See figure on next page.)

Fig. 2 Benchmarking Exo-C against conventional and molecular methods of karyotype and genome analysis (namely karyotyping, aCGH or WGS).

A Graphical summary of the 66 samples analyzed in this study. Each sample was examined using the Exo-C protocol, with most undergoing additional orthogonal profiling methods. **B** Representative example of translocation detected using Exo-C: contacts map of P136 sample P136, which exhibits a chr4-chr9 translocation (above diagonal), and P114, with a normal karyotype (below diagonal). A schematic representation of the translocation and its breakpoint coordinates ([hg19] chr4:174,328,001–174,336,000; chr9:108,424,001–108,432,000) is also included. **C** Performance of the translocation caller. This graph displays the *F1* score, a measure of the translocation caller's performance. Each data point corresponds to a subset of samples, filtered according to their deviation from the cohort average *cis*-score. The *X*-axis specifies the maximum allowable deviation for each data point, while the *Y*-axis represents the observed *F1* score for the corresponding filtered sample set. **D** The distribution of *cis*-scores across samples is shown, with vertical lines indicating the average for XX and XY genotypes. The shaded area highlights the optimal *cis*-score range ($\pm 8\%$ from average). **E** Percentage of True Positive (TP) and False Negatives (FN) calls for modeled translocations of varying lengths. Red bars denote TP calls, while blue bars represent FN. The *Y*-axis shows the percentage of TP and FN calls for modeled translocations of varying lengths. **F** Representative example of inversion detected by Exo-C: contacts map of P137 sample (with an inversion on chromosome 7; above diagonal) and P114 sample (with normal karyotype; below diagonal). **G** *F1*-score of the inversion-calling algorithm, based on a dataset of simulated inversions of different lengths. **H** The number of TP calls for simulated inversions, grouped by length intervals, each containing 20 simulations. **J** Distributions of CNV types and length classes in the low-confidence list of CNVs present in analyzed samples. Bars represent the count of amplifications (blue) or deletions (green) in each length class. **K** Precision and recall of CNV predictions by GATK, CoNIFER, and CNVkit tools. Color of markers represent evaluation of tools based on low- or high-confidence lists of CNVs with or without CNV size filtration; shape of markers represents evaluation of tools with or without predictions, intersected with known CNVs for less than 50% (in terms of Jaccard Index) and with or without common CNVs (DGV Gold database). Note that precision and recall outcomes in analyzes excluding or including DGV Gold CNVs are almost similar, therefore corresponding markers may not be fully visible due to overlap

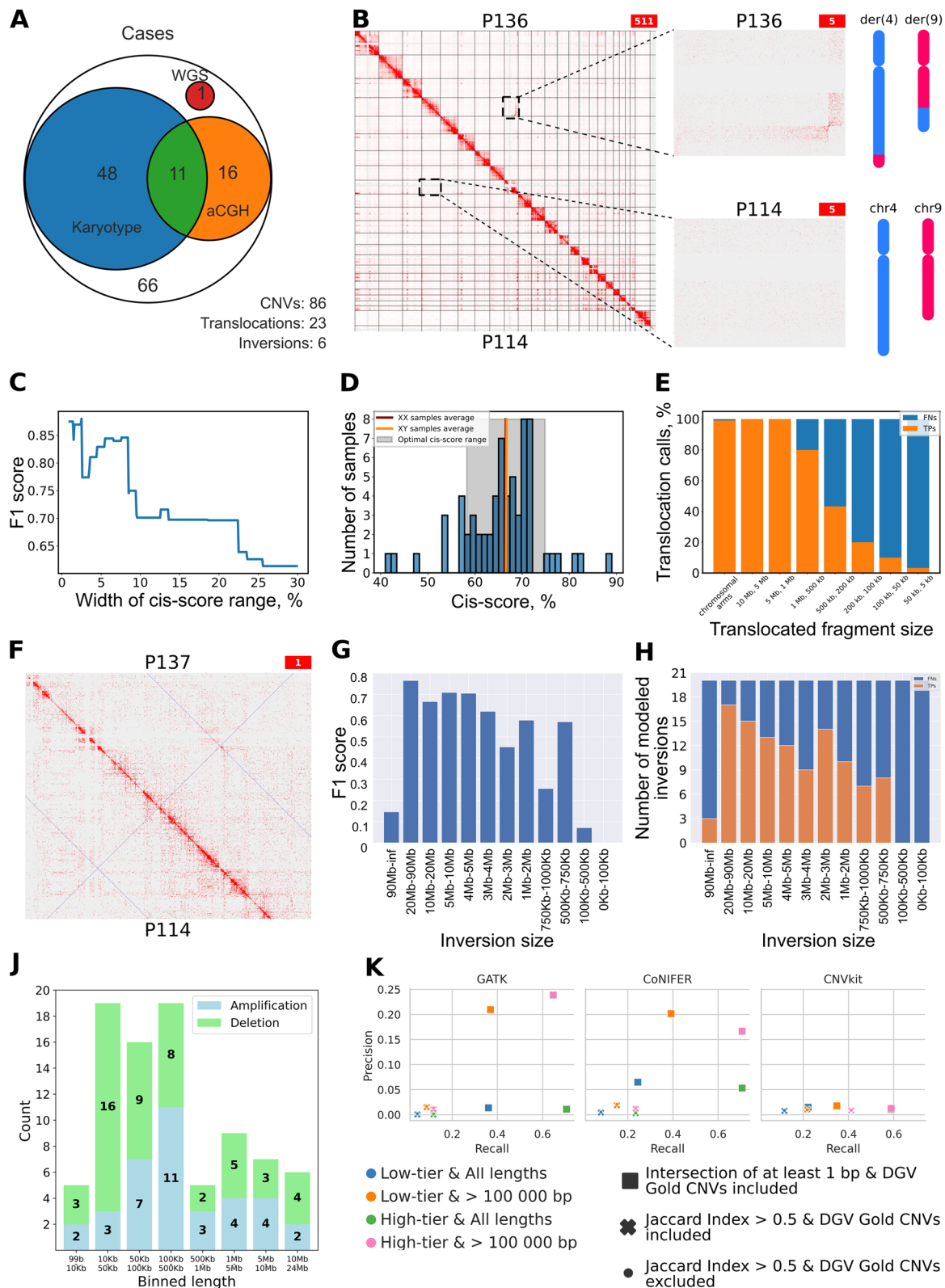


Fig. 2 (See legend on previous page.)

In A549 cells, where no inversions were previously reported, Exo-C made 21 inversion calls, nine of which coincided with breakpoints of known deletions or duplications. This suggests that while the inversion caller accurately identified regions with atypical Hi-C patterns, it misclassified them as inversions. In K562 cells, Exo-C detected two out of four reported inversions, along with 19 false positives, nine of which were associated with known deletions, duplications, or unclassified SVs. EagleC did not find any reported inversions for K562 along with four false positives.

Finally, for A549 and K562 cells, we observed CNV detection precision between 0 and 0.18 (the highest precision was observed for CNVkit) and recall from 0 to 0.375 (the highest recall was again observed for CNVkit). Detailed results for each of the callers are presented in the Additional File 1: Table S16. These results are on par with the data obtained for clinical samples as described below and in Fig. 2E.

Automatic detection of chromosomal translocations in patient cohort

Our methodology for translocation calling demonstrates robust precision and recall in human samples, as shown in Fig. 2C (list of human samples for this analysis is provided in Additional File 1: Table S1). In patient samples previously analyzed by conventional karyotyping, our method achieved a Precision of 36%, Recall of 100%, with 25 True Positive (TP) calls, 45 False Positive (FP) calls, and no False Negatives (FN). We observed that the quality of experimental data, quantified by the cis-to-trans contacts ratio, significantly influences the caller's performance. Higher precision is noted in samples with metrics akin to control samples (Fig. 2C and 2D).

Exo-C map for each FP call was examined visually, and for 12 FP calls we observed a clear translocation pattern (Additional File 2: Fig. S1; Additional File 1: Table S2). Given that some predicted insertion sizes fell below the detection limit of microscopy, (the smallest detected by our algorithm was ~20 kb, Additional File 2: Fig. S1), it is plausible that these FP calls might represent genuine SVs, undetected by conventional diagnostic techniques. Importantly, in many cases translocation breakpoints were resolved at high resolution, allowing primer design for PCR-validation of the breakpoints (see examples below). For breakpoints where we were not able to design primers, we performed FISH with insertion-specific probes. This approach validated 5 out of 6 tested FP translocation calls using orthogonal methods (PCR or Spectral Karyotyping), as elaborated in the “Cases” section.

Upon refining our analysis by filtering 17 samples based on their “cis/trans” ratio ($58\% < \text{Cis} < 75\%$; Fig. 2C

and 2D) and reclassifying the validated translocations as True Positives, the revised metrics showed a Precision of 73%, Recall of 100%, with 22 TP, 8 False Positives, and no False Negatives (Additional File 1: Table S10). It's important to note that this efficiency may be an underestimation, as some FP calls could not be further verified due to the unavailability of additional sample material for orthogonal testing.

To investigate the impact of translocated segment size on method sensitivity, we employed the Charm in silico modeling framework to create Exo-C contact maps for 330 simulated translocations or insertions, with sizes ranging from 10 kb to 140 Mb (detailed in the “Methods” section). As anticipated, the recall rate of translocations increased with the size of the insertions (Fig. 2E). Detection rates increased from approximately 5% for the smallest insertions (5–50 kb) to around 45% for larger events (200–500 kb), achieving 100% sensitivity for fragments exceeding 1 Mb. The overall F1-score for all modeled cases remained 0.77, with a Precision of 87%, Recall of 68%, comprising 226 true positives, 34 false positives, and 104 false negatives.

Automatic detection of inversions in patient cohort

Inversions, like translocations, manifest as distinct contact pattern changes on Hi-C maps (Fig. 2F). Of the six inversion cases known in clinical samples, our method successfully detected three, alongside 220 false positive inversions, yielding a precision of 10% and a recall of 50% (Additional File 1: Table S11). Notably, the sensitivity of inversion detection is closely tied to the quality of experimental data, quantified by the cis-to-trans contacts ratio, similar to translocation calling. Remarkably, the false positive count significantly reduces from 220 to 33 when applying the same cis-to-trans filter to the samples, while the recall remains consistent at 50% for this subset. Further investigation reveals that some false-positive calls correspond to inversions undetectable by cytogenetic analysis but verifiable through alternative methods. For instance, a 2 Mb inversion identified in P10 data was initially reported as a false positive. Subsequent validation using long-read sequencing, as detailed below, confirmed this inversion and its correlation with the observed phenotype in the P10 case. Consequently, the actual precision of the inversion detection algorithm is likely higher than initially estimated, suggesting an overestimation of false positive calls.

We next assessed the algorithm's performance across a range of inversion sizes, from 100 kb to several megabases, utilizing 240 in silico generated inversions (Fig. 2G and 2H). This benchmarking, using the series of simulated Exo-C maps, yielded an average F1-score of 0.49 for all inversion lengths (Precision: 56%, Recall: 45%,

True Positives: 107, False Positives: 85, False Negatives: 133). The analysis revealed a positive correlation between the algorithm's performance and the size of the inversion, with larger inversions yielding higher scores. However, it was noted that the algorithm encounters challenges in detecting inversions exceeding 90 Mb, primarily due to their breakpoints being situated at the chromosomal ends, which complicates the recognition of inversion patterns on Hi-C maps.

Detection of CNVs in patient cohort

Copy-number variants (CNVs) modify the representation of genomic loci in NGS datasets, enabling their identification through analysis of read coverage distribution. We hypothesized that Exo-C results should allow analysis of genomic coverage similar to the whole-exome data. To evaluate this, we tested existing CNV prediction tools on Exo-C datasets. Utilizing available aCGH data, we compiled a list of CNVs, which we further categorized into high-confidence (clinically relevant CNVs validated by qPCR or other methods) and low-confidence groups (identified in aCGH data but lacking orthogonal confirmation) (detailed in Methods; Additional File 2: Fig. S2 A and Additional File 1: Table S3). This categorization resulted in 46 CNVs in the high-confidence list and 86 in the low-confidence list. The distribution of these CNV types and their length classes is illustrated in Fig. 2J.

Leveraging the compiled dataset alongside corresponding Exo-C data, we conducted a benchmarking study of three CNV callers—CNVkit, CoNIFER, and GATK—originally designed for WES data analysis. Despite not being tailored for Exo-C data, these tools demonstrated the capacity to identify over 60% of CNVs in our samples (square markers in Fig. 2K; Additional File 1: Table S4). However, the overlap between CNVs detected by these tools and those reported by aCGH was often partial. When predictions were filtered to include only those with a Jaccard Index greater than 50% (cross markers in Fig. 2K), the performance of GATK and CoNIFER, in both precision and recall, significantly decreased. This filtration method only marginally impacted CNVkit's performance, possibly due to its inherently lower precision prior to any filtration.

Implementing an additional filtration layer, which excludes CNVs commonly found in the population as per the DGV Gold database, significantly reduces the number of false positive predictions. This effect is particularly evident in the GATK results, where the count of small CNV predictions (less than 100,000 bp) is reduced by nearly 70%. A similar, albeit less marked, reduction is observed across all evaluated tools, both for various CNV tiers and size-based filtration (as detailed in Additional File 1: Table S4 and Additional File 1: Table S12). However,

the overall precision of these tools, initially modest, does not exhibit a substantial absolute increase post-filtration (indicated by dot markers in Fig. 2K, closely aligning with cross markers).

We hypothesized that CNVs from the high-confidence tier of clinically-relevant variants would exhibit more comprehensive detection, given their lower likelihood of being common population variants. Corresponding results affirmed this hypothesis, revealing that high-confidence tier CNVs were indeed predicted with greater recall across all assessed tools (Fig. 2K). The precision for high-confidence CNVs surpassed that of the low-confidence tier exclusively in the case of GATK.

Exo-C can be employed to detect mosaic translocation carriers

Exo-C demonstrates exceptional precision and recall in detecting translocations, particularly excelling in cases involving large, megabase-scale translocations. This suggests a potential for detecting SVs even in a minor subset of cells, a capability crucial for applications in cancer research or germ-line mosaicism studies. To investigate this potential, we employed our translocation calling methods in both *in silico* and *in vitro* experiments simulating “mosaic cell populations.” In the *in silico* experiment, we merged random reads from pairs of samples, one carrying known structural variants and the other with a normal karyotype, in varying proportions (as detailed in Additional File 1: Table S5). We proportionally adjusted the data inputs so that 10% to 90% of the reads in the merged dataset originated from the sample with translocations. As depicted in Fig. 3A, our method can reliably identify heterozygous translocations in samples with a fraction above 80%, and some heterozygous translocations are detectable when the mix contains only 40% of data from the translocation-carrying sample. It is important to note that, since we mixed data from heterozygous translocation carriers, the actual fraction of alleles with the structural variant is half of the indicated data fraction.

In the *in vitro* experiment, we generated cell mixtures from P62 + P114 samples (where P62 exhibits t(1;4) and P114 has a normal karyotype) with 10.3%, 36.6%, and 42% of cells carrying the translocation; and from P69 + P114 samples (P69 with t(2;6) and t(7;11), P114 normal) with 14.5% translocation-carrying cells. The exact proportion of cells with chromosomal rearrangements was determined based on the SNV analysis. Applying Exo-C to these samples corroborated previously described findings (Fig. 3A). Although our bioinformatic tool did not detect translocations in samples with less than 36% SV-carrying cells, visual analysis suggests that the translocation pattern is still discernible at a 14.5% fraction (Fig. 3B). These

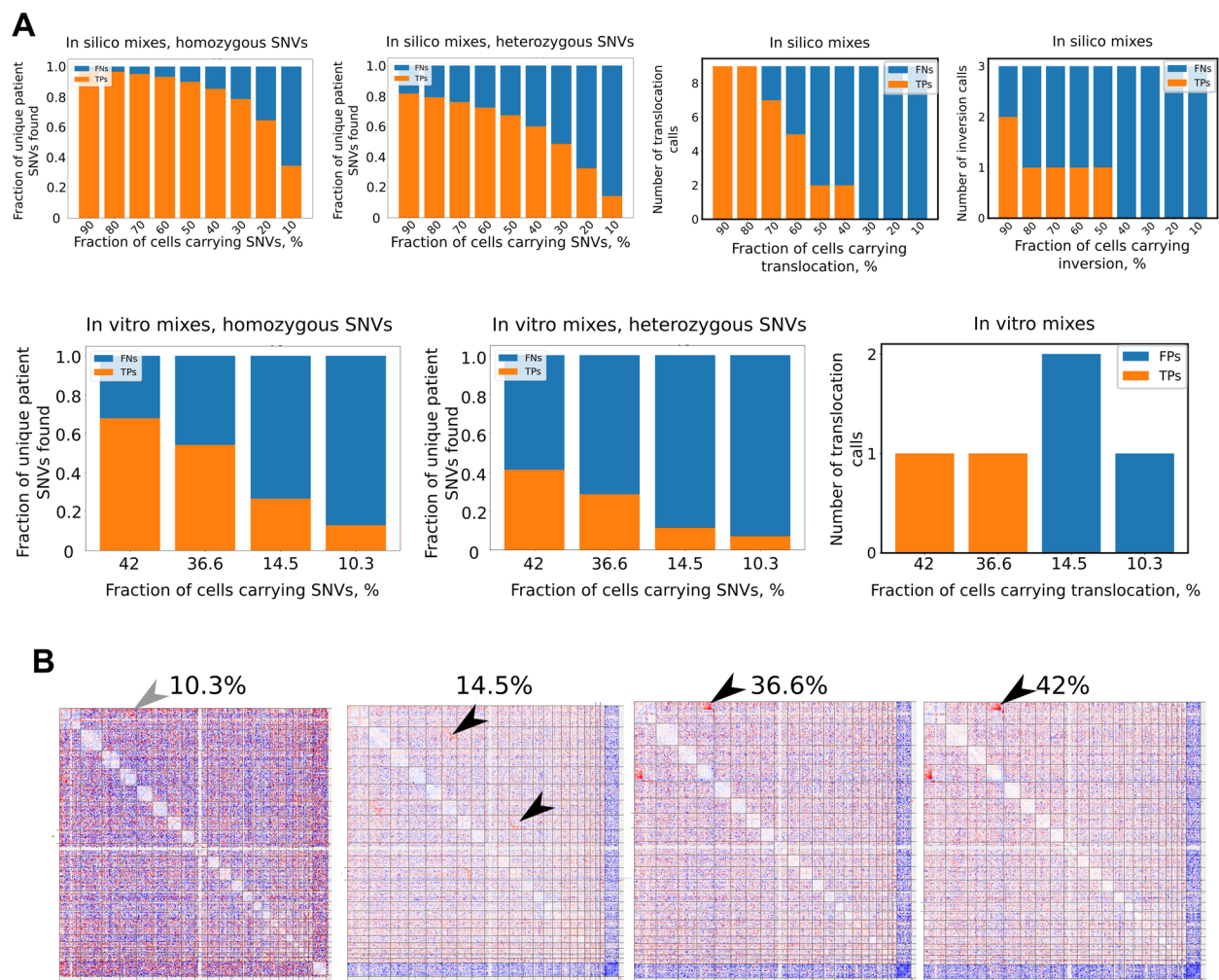


Fig. 3 Detecting mosaic translocation carriers using Exo-C. **A, B** Recall of the variant callers for mosaic carriers. Y-axis show number of TP calls for samples with different fraction of translocation-carrying cells (%) for in silico and in vitro experiments. Red bars indicate True Positive calls, blue bars False Negative. **B** Representative interaction Exo-C heatmap (Log(Observed/Control); Observed, mixed samples; Control, patient without translocations). Black arrows indicate visually detectable chromosomal rearrangements, gray arrows—expected rearrangements

observations lead us to speculate that further refinement of computational tools could enhance the precision of translocation detection in mosaic cell populations.

We next performed similar analysis for inversions, CNVs, and SNVs. We analyzed in silico mixes derived from inversion carriers (P52, P82, P137) and donors with normal karyotypes (P114, P115), detailed in Additional File 1: Table S5. Using Exo-C, we identified 2 out of 3 inversions when the data fraction from an inversion carrier was above 90%, and one inversion was detectable when the fraction was above 40% (Fig. 3A).

For SNV detection benchmark, we utilized the same Exo-C libraries from cell mixtures previously employed in the search for mosaic translocations. After pinpointing unique SNVs for patients P62 and P69, we compared

these findings to calls made on mixed cell populations (Fig. 3, A). The detection rates were up to 40% for heterozygous SNVs and up to 65% for homozygous SNVs in mixtures with a patient's cell percentage of up to 42%. These results are consistent with expectations, as both Exo-C and WES, with our typical coverage level of around 30X, are not optimal for reliably detecting clinical SNVs in cell mixtures, which would require a coverage of 200X [49]. To study SNVs detectability for higher fraction of target cells, we in silico mixed reads from the libraries of patients P114 and P62 to create pseudosamples with proportions of P62 reads ranging from 10 to 90%. SNVs were then called on both the mixed and original libraries. We identified variants unique to patient P62 and assessed how many unique SNVs from this set were detectable in

the mixed libraries, analyzing heterozygous and homozygous SNVs separately (results provided in the Fig. 3). Our analysis successfully identified over 90% of homozygous SNVs in mixes containing at least 60% of the patient's cells. For heterozygous SNVs, we achieved a detection rate of up to 80%.

We next employed the same pseudosamples described in the SNV section, along with analogous combinations of P69 and P114, P37 and P31, and P109 and P31 (Additional File 1: Table S5) to benchmark CNV callers. We utilized the same three tools (GATK, CNVkit, CoNIFER) and assessed predictions against a low-confidence list of known CNVs. Given their poor performance on our original clinical samples, these tools demonstrated limited capability to differentiate mosaic CNVs. For most pseudosamples, there were no true positive predictions. Notably, GATK was able to identify one out of seven CNVs in the P62 sample when fraction of its reads exceed 50%. Furthermore, all three tools successfully detected two out of eleven CNVs present in the P109 sample with an 80% fraction. Analyzing *in vitro* experiment, we identified only one known CNV via GATK in the P62 sample at the 36.6% fraction of cells carrying the CNV.

Overall, this analysis highlights capacity of Exo-C to detect mosaic translocation, however for other mosaic variants (inversions, CNVs, SNVs) its capability is limited.

Augmenting Exo-C with targeted Oxford Nanopore sequencing allows cost-efficient detection of structural variants with base-pair resolution

Although Exo-C shows high sensitivity in SVs detection, variants cannot be resolved with single-nucleotide resolution, and in some cases resolution is too low to design primers for breakpoints amplification and sequencing. Yet, precision of the Exo-C method is high enough for using cost-efficient targeted sequencing of the detected translocation breakpoints. To benchmark this approach, we selected 5 samples which we define as Nanopore cohort; sample IDs and additional information about each sample is provided in Additional File 1: Table S6 and S1, respectively. These 5 cases involving 23 Exo-C-identified breakpoints, with 17 confirmed via orthogonal techniques. Only two junctions were validated by PCR with single-nucleotide resolution, others were confirmed by karyotyping (Additional File 1: Table S6). We applied whole-genome nanopore sequencing (WGS), adaptive sampling (AS) and modified (see the “[Methods](#)” section) nanopore Cas9-targeted sequencing (nCATS) [19] approaches to this cohort (Fig. 4A and 4B). We confirm that nCATS greatly improves coverage of the breakpoint regions (Fig. 4C), whereas substantially deeper sequencing would be required to achieve the same breakpoint

coverage without enrichment. The adaptive sampling approach, while displaying a balanced performance compared to previous methods, offers the advantage of validating breakpoints without requiring specialized library preparation such as the production of sgRNAs for nCATS. It is important to note that although the coverage of target regions in AS is lower compared to nCATS, it presents a viable option for breakpoint validation. For both enrichment and WGS we limit sequencing depth to the same value (up to 1.5X coverage, Additional File 1: Table S6) to ensure reasonable cost of analysis in future clinical use.

We analyzed sequencing results both manually (aiming to detect any chimeric reads in a vicinity of the Exo-C-derived breakpoints) or automatically using the NanoSV [21] software. The results demonstrate our ability to validate the majority of translocations identified using Exo-C through nanopore sequencing, achieving single-nucleotide resolution (Fig. 4D and Additional File 1: Table S6). In particular, we were able to confirm and resolve at single-nucleotide level 18 out of 23 Exo-C-detected breakpoints (78%) and 4 breakpoints which were not validated previously by orthogonal methods were validated in this assay (Fig. 4D). As expected, targeted sequencing methods (AS and nCATS) outperformed nanopore WGS in effectively covering breakpoint regions. Automatic breakpoint detection with NanoSV was effectively possible only in regions with high coverage, highlighting the possibility for use in combination with Cas9 enrichment.

Importantly, the targeted sequencing approaches require only 20–30% of the resources of a MinION or GridION flow cell, which allows cost-effective breakpoint verification for less than \$300 per sample. This affordability underscores the possibility of implementing targeted nanopore sequencing as a valuable complement to Exo-C, presenting an economical solution for breakpoint validation in clinical applications. These findings emphasize the potential of combining Exo-C with targeted nanopore sequencing as an efficient and budget-friendly tool for clinical use.

Application of Exo-C technology in resolving clinical cases

This study encompassed several patients with congenital diseases who had previously eluded molecular diagnosis. We demonstrate how Exo-C can identify rare genomic variants in these cases, and subsequent analyses elucidate their relevance to disease pathogenesis.

Patient P10: detecting inversion which causes morbid gene disruption

An 8-month-old girl was referred for evaluation because of developmental delay, hypotonia, feeding problems, delayed motor milestones, and abnormal phenotype including microcephaly, almond-shaped palpebral

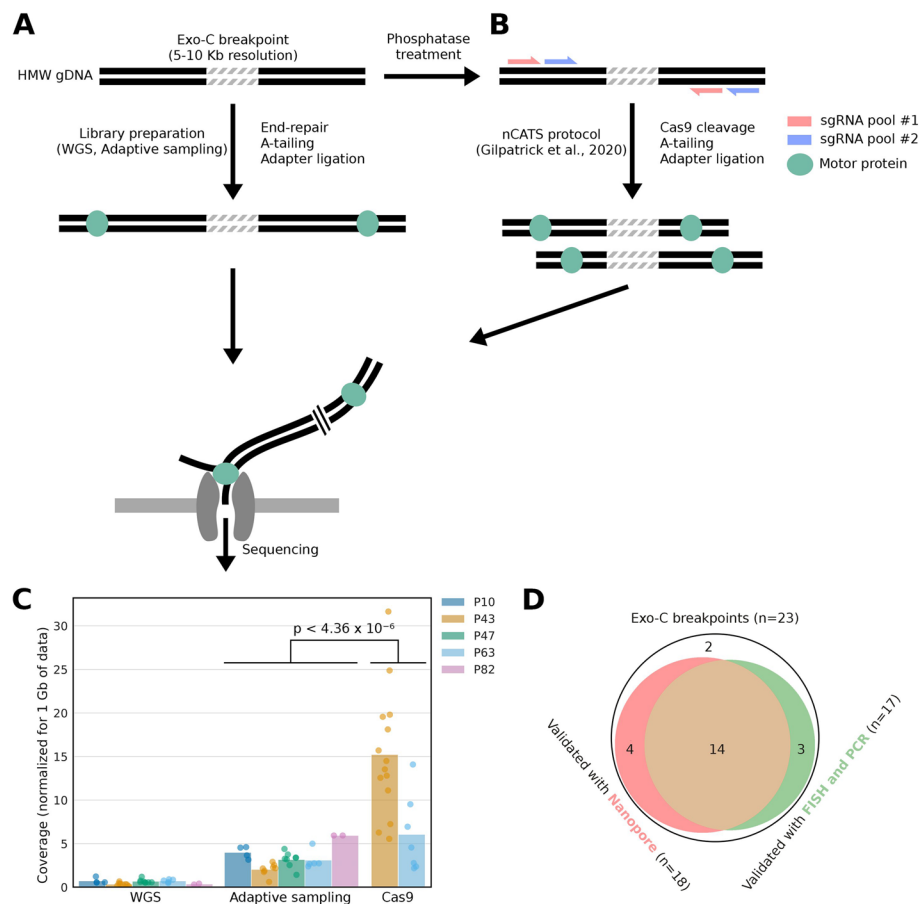


Fig. 4 Combination of Exo-C and nanopore sequencing enables detection of breakpoints with single-nucleotide resolution. **A** Assembly of Nanopore ligation libraries for WGS and adaptive sampling. **B** Enrichment of breakpoint regions with Cas9 utilizing a combination of tiling and nCATS approaches. **C** Comparison of coverage in breakpoint regions between different Nanopore library types (normalized for 1 Gb of data, coverage calculated for 5 kb bins). **D** Graphical representation of breakpoints detection by Exo-C, nanopore sequencing, and other methods

fissures, low-set ears, posteriorly angulated ears, wide nasal bridge, fifth-finger clinodactyly, pes valgus.

Karyotyping revealed a de novo, apparently balanced translocation between chromosomes 5 and 10 in the patient: 46,XX,t(5;10)(q11.2;q11.2)dn. aCGH did not detect unbalanced chromosomal rearrangements. Utilizing Exo-C, we resolved the genomic breakpoints of this translocation with high resolution (Fig. 5A). This resolution enabled the amplification and sequencing of the junction region (Fig. 5B). The sequencing data revealed that the breakpoint on chromosome 10 intersects the *PCDH15* gene, while the breakpoint on chromosome 5 is situated in an intergenic area proximal to the *RNF180* gene promoter. Intriguingly, neither of these genes appear to correlate with the patient's clinical phenotype.

In our analysis of genes proximal to the translocated fragments junction, we identified the *DKK1* gene approximately 1.5 Mb from the chromosome 10 breakpoint (Fig. 5A). We hypothesized that *DKK1* misregulation

might contribute to the disease phenotype. Supporting this, both low-input Hi-C and the original Exo-C data revealed altered chromatin interactions near the translocation breakpoints, affecting the *DKK1* promoter (Fig. 5A). To test the hypothesis of *DKK1* gene misregulation, we derived induced Pluripotent Stem Cells from the patient's peripheral blood and analyzed gene expression in these cells and their differentiated primitive streak derivatives (Additional File 2: Fig. S3 A and B). Given the known high expression of *DKK1* in primitive streak cells, we conducted digital PCR analysis on patient-derived iPS cells and primitive streak cells. Contrary to our hypothesis, the results indicated no significant difference in *DKK1* expression between the patient and control cells, challenging the hypothesis of *DKK1* misregulation (Additional File 2: Fig. S3 C).

Since translocation between chromosomes 5 and 10 cannot explain P10 phenotype, we apply Exo-C variant callers to identify other clinically relevant genomic

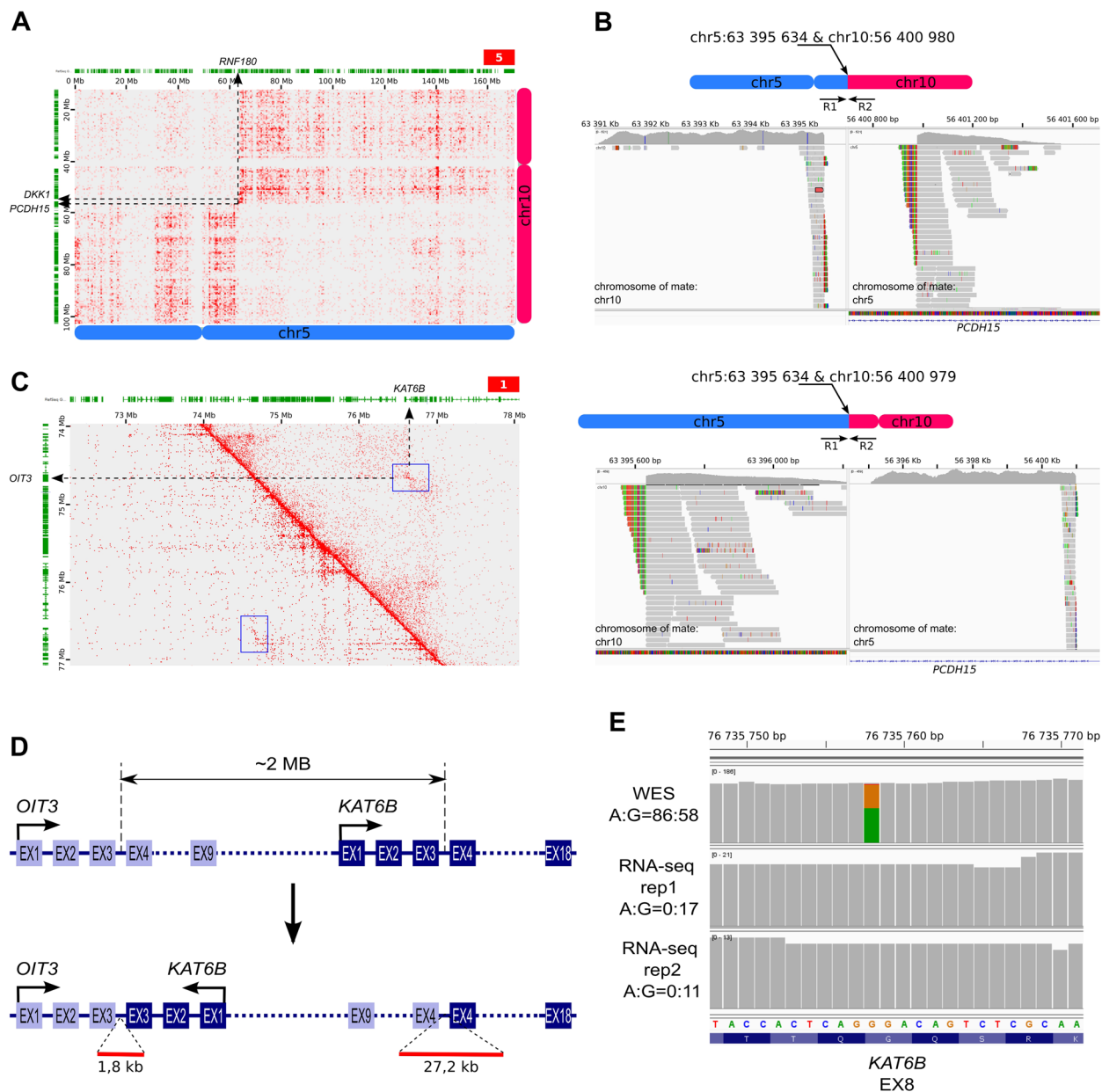


Fig. 5 Detection and functional characterization of balanced translocation and inversion using Exo-C technique. **A** An Exo-C contact map indicating balanced translocation pattern for P10 case. **B** IGV screenshot showing alignment of reads obtained by amplification and sequencing of the t(5;10) junction. **C** Hi-C (above diagonal) and Exo-C (below diagonal) maps indicating the presence of balanced translocation in P10 genome. **D** Scheme of the inversion detected using Exo-C and Hi-C data. Genomic breakpoints of inversion between *OIT3* (breakpoint coordinate—chr10:74,666,324) and *KAT6B* (breakpoint coordinate—chr10:76,646,623) were identified by 1778 bp and 27,189 bp nanopore sequencing reads (red lines), spanning of third introns *OIT3* and *KAT6B*. **E** IGV screenshot showing allelic representation of single-nucleotide variant in the *KAT6B* exon 8 according to the Exo-C sequencing data and transcriptome sequencing data

variants. No pathogenic SNVs were identified; however, the Exo-C inversion caller detected a ~ 2 Mb inversion on chromosome 10, situated approximately 20 Mb from the previously identified translocation breakpoints. Given the significant genomic distance between breakpoints, we propose that translocation and inversion

occur as two independent mutations on chromosome 10. The inversion, undetected in karyotyping due to its size, was corroborated by low-input Hi-C analysis, which confirmed the inversion pattern (Fig. 5C). Additionally, the presence of the inversion was supported by split-reads identified in nanopore sequencing data (see

above). Based on these data, inversion breakpoints can be mapped within the 3rd intron of the *KAT6B* gene and the 3rd intron of the *OIT3* gene, notably disrupting the *KAT6B* gene (Fig. 5D).

Loss-of-function variants in the *KAT6B* gene are known to cause SBBYSS syndrome (OMIM #603736) in an autosomal-dominant manner, potentially explaining the patient's phenotype. To evaluate the hypothesis of *KAT6B* gene disruption, we conducted transcriptome analysis in iPS cells. This analysis revealed abnormal splicing products consistent with the predicted inversion breakpoint within the 3rd intron of *KAT6B* (Additional File 2: Fig. S3D). Furthermore, based on Exo-C data, we identified a benign heterozygous SNV in exon 8 of the *KAT6B* gene (Fig. 5E). RNA-seq data, however, indicated the expression of only one allele of this gene (Fig. 5E). These findings collectively confirm that only one functional copy of the *KAT6B* gene is expressed, correlating with the patient's observed clinical manifestations.

Thus, Exo-C and Hi-C analysis allowed us to characterize two balanced structural variants and develop hypotheses about causative genes implicated in the disease. Pursuing this line of investigation, we were able to characterize inversion affecting the *KAT6B* gene.

Patient P47

An 8-year-old boy was examined by a geneticist regarding developmental delay (low height and weight) and multiple exostoses. The first exostosis on the rib was noted at birth, the second appeared on the arm at the age 2.5 years. At the time of examination, the patient had multiple exostoses. Dysmorphic features like dolichocephaly, epicanthus and right dysplastic ear were noted.

Conventional karyotyping of G-banded chromosomes revealed a balanced translocation between chromosomes 8, 11, and 21: 46,XY,t(8;11;21)(q23;q22;q21), aCGH did not reveal any unbalanced chromosomal rearrangements.

In Exo-C assay of this case, we did not detect any pathogenic SNVs. However, the Exo-C data analysis confirmed a complex apparently balanced chromosomal rearrangement involving chromosomes 8, 11, and 21. This analysis accurately resolved the breakpoints of these rearrangements and additionally identified two translocations: one between chromosomes 20 and 21, and another involving fragments of chromosomes 11, 8, and 20 (Fig. 6A and 6B). These findings were subsequently validated using Spectral Karyotyping (SKY) (Fig. 6C and 6E). Although SKY analysis confirmed SV structure predicted by Exo-C, it failed to detect insertion of chromosome 8 fragment into chromosome 11. Therefore, we designed a chromosome

8 probe based on Exo-C prediction and confirmed the insertion using FISH analysis (Additional File 2: Fig. S4). Furthermore, we were able to validate Exo-C results using Oxford Nanopore long-read sequencing.

Notably, one breakpoint was located within the *EXT1* gene (Fig. 6B and 6E), a known causative gene for autosomal-dominant hereditary multiple exostoses (OMIM #133700). The disruption of this gene elucidates the patient's clinical phenotype. This case exemplifies how the delineation of translocation breakpoints via Exo-C allows the classification of balanced chromosomal translocation as pathogenic.

Patient P123

A 41 years old male was under examination after he has a stroke of the left MCA (middle cerebral artery) of ischemic type, with right-sided hemiparesis and gradual regression of focal disorders. Repeated episode of cerebrovascular accident occurred after six months, with hemiplegia of the left arm and leg, facial asymmetry. CT revealed a hematoma in the right hemisphere measuring 20 cm³, against the background of lesions arteriopathy with subcortical infarcts. The consequences of multiple repeated cerebral infarctions of the deep parts of both hemispheres of the white matter of the brain of varying duration were revealed. Neurometabolic therapy was applied with positive effects. Patient previously reported migraine-like headaches since age 20. The patient has a 17-year-old daughter who complains of migraine-like headaches.

While Exo-C analysis did not reveal any pathogenic SVs or CNVs, it identified a SNV rs1555729452 (Cys222 Tyr) within the *NOTCH3* gene, a known causative gene for cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy (CADASIL; OMIM #125310). Classified as pathogenic based on the American College of Medical Genetics and Genomics (ACMG) guidelines, this variant accounts for the patient's clinical presentation. This case underscores the capability of Exo-C data to pinpoint pathogenic SNVs within the exome, showcasing a performance comparable to traditional exome sequencing methods.

Discussion

In this study, we demonstrate the utility of Exo-C, a combination of modified chromosome conformation capture assay and exome enrichment, in elucidating a broad spectrum of clinically significant genomic variants in patients with monogenic and chromosomal diseases. With cost comparable to standard exome sequencing, Exo-C notably enhances it in the detection of balanced translocations and inversions. Thus, Exo-C facilitates the

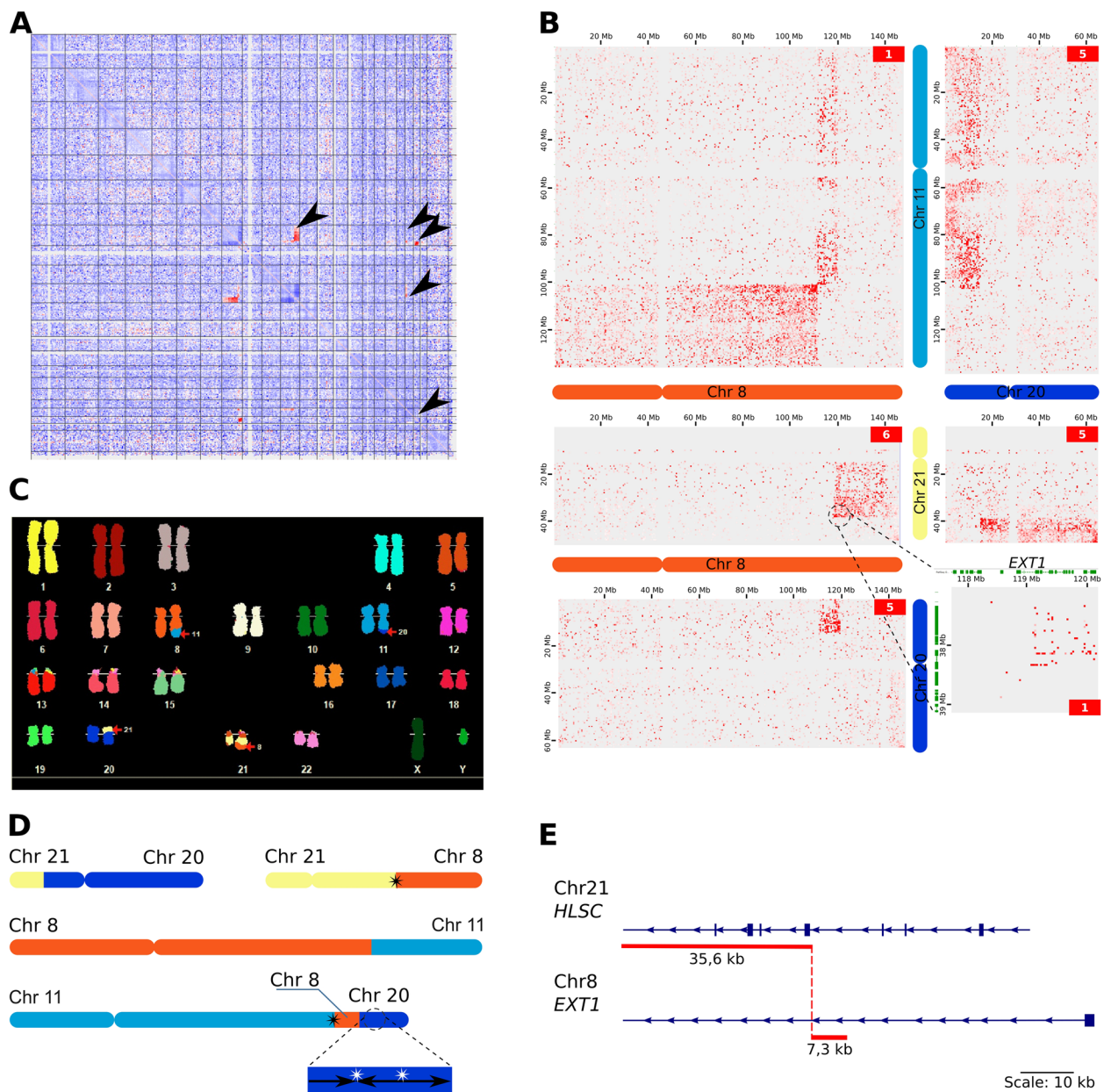


Fig. 6 Resolving breakpoints of complex chromosomal rearrangement through Exo-C analysis provides evidence of its pathogenicity. **A** Exo-C contacts normalized as $\text{Log}(\text{Observed}/\text{Control})$, where Observed is contact counts for P47 which exhibits complex SV, and control is Contact counts for control sample without translocations. Black arrows indicate Exo-C contacts at translocation breakpoints. **B** Interchromosomal Exo-C contact maps for P47 sample supporting SV structure. **C** SKY results supporting SV structure. **D** Schematic representation of SV detected in P47 cells based on Exo-C analysis. Asterisks indicate breakpoints identified with single-nucleotide resolution by ONT sequencing. **E** Genomic breakpoint of translocation between *EXT1* (breakpoint coordinate—chr8:119,069,497) and *HLSC* (breakpoint coordinate—chr21:38,321,275) was identified by a 41,312 bp sequencing read (red line), spanning 7.3 kb of first intron *EXT1* and 35.6 kb of second intron *HLSC*

simultaneous identification of SVs and SNVs, surpassing the capabilities of traditional WES.

Complementing the experimental application of Exo-C, this publication introduces a suite of computational tools specifically tailored for Exo-C data

analysis, focusing on SV calling. The development of automate callers is crucial for the clinical implementation of this technique as well as for an unbiased evaluation of its performance. Existing computational solutions, primarily designed for genome-wide cancer

Hi-C datasets, fall short when applied to Exo-C data. Addressing this gap, we have developed and rigorously benchmarked computational tools that facilitate automated Exo-C-based SV detection, thereby enabling users to process their data with greater efficiency and precision.

While Exo-C significantly enhances standard exome analysis, it is important to acknowledge its limitations. A primary constraint is the diminishing recall of Exo-C SV callers for smaller-sized SVs. The tool exhibits relatively modest performance in identifying translocations and inversions under 100 kb. Intergenic events smaller than 100 kb remain challenging to reliably identify without resorting to impractically high sequencing depths. However, it is noteworthy that smaller translocations, and potentially inversions, overlapping with exonic sequences are more detectable due to the significantly higher coverage of these fragments compared to the genomic average.

A second limitation of Exo-C concerns CNV detection. The precision and recall of CNV callers on Exo-C data are relatively low, mirroring the performance observed with WES data [50, 51]. This is partly because, for CNV calling, Exo-C data analysis was limited to genomic coverage assessment. The current effectiveness could be enhanced by incorporating analyses of specific contact frequency alterations attributable to CNVs. However, insights from studying contact frequency changes induced by inversions suggest that significant improvements in detecting CNVs smaller than 100 kb are unlikely. Consequently, aCGH and low-coverage WGS currently offer a more favorable balance between cost and the quality of CNV detection.

Considering the strengths and limitations of the Exo-C technique, its application can be particularly advantageous in several diagnostic scenarios. Firstly, Exo-C is well-suited for patients with karyotypic abnormalities of undetermined significance detected through microscopy-based methods. Here, Exo-C can precisely resolve breakpoints to an extent that facilitates the identification of genes disrupted by SVs. Additionally, Exo-C is capable of detecting exonic SNVs, testing alternative hypothesis about the nature of genomic variant underlining the patient's phenotype. From a cost perspective, Exo-C is only about \$100 more expensive than WES but significantly more cost-effective compared to short-read WGS (+\$600) or Nanopore sequencing (+\$7800), as detailed in Additional File 1: Table S15. Consequently, we advocate for the adoption of Exo-C as the primary diagnostic tool for patients with clinically ambiguous karyotypic abnormalities.

Secondly, Exo-C can serve as an alternative to standard WES for primary screening in patients suspected of harboring disease-causing SNVs. Exo-C can be adapted to various hybridization-based capture panels, enabling custom assays tailored to different patient groups. Exo-C not only delivers the same insights as WES but also identifies submicroscopic SVs within or adjacent to genes of interest.

Thirdly, Exo-C is valuable following aCGH analysis, particularly when a CNV affects one copy of a clinically relevant haplosufficient gene. In such cases, Exo-C can aid in identifying loss-of-function mutations in the second allele of the implicated gene.

Recent advancements in research indicate the potential of chromatin interactions for SNV phasing. This presents another promising application for the Exo-C method: distinguishing between cis and trans configurations in compound heterozygous scenarios. However, realizing this application will necessitate the development of specialized bioinformatic tools tailored for Exo-C data analysis.

Finally, Exo-C offers valuable opportunities to explore the regulatory effects of previously identified SVs, utilizing data on chromatin contact frequencies to gain insights into their functional implications. A unique aspect of Exo-C is its ability to analyze spatial interactions of gene promoters, which are crucial for understanding diseases caused by missregulation of gene expression [52]. While computational predictions of alterations in chromatin contacts exist [53], their accuracy remains insufficient to replace empirical data [54]. Consequently, the chromatin interaction patterns unveiled by Exo-C offer valuable insights for variant interpretation, bridging a critical gap in genomic analysis.

In the latter part of this manuscript, we illustrate this through three use-cases, demonstrating Exo-C's capability in identifying novel pathogenic variants and guiding the investigation into their pathogenicity.

The integration of Exo-C with other techniques of molecular analysis can significantly enhance its capacities. Here, we highlight integration of Exo-C with targeted Oxford Nanopore sequencing, enabling the reconstruction of complex SVs with single base-pair resolution in a cost-effective manner.

Conclusions

The Exo-C technique, as detailed here, broadens the scope for identifying causative variants in patients with genetic diseases, offering significant potential in genetic diagnostics and research.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13073-025-01471-3>.

Additional file 1: Table S1. List of samples analyzed by Exo-C. Table S2. New translocations detected by the Exo-C and their validation status. "N/A" indicates none of the methods was applicable for the sample. Table S3. Low-tier list of CNVs present in Exo-C samples. Table S4. Results of CNV prediction tools evaluation with different conditions. Table S5. Simulating/mosaic cell populations. Table S6. Validation of breakpoints found by Exo-C with Nanopore sequencing. Table S7. Sizes of called inversions at different resolutions. Table S8. List of primers used for sgRNA generation. Table S9. Table for Exo-C and reference NGS data SNV precision comparison. Table S10. List of translocations found by Exo-C and their validation status. "NOT" indicates the rearrangement was not validated by the alternative method; "N/A" indicates the method was not applicable for the sample. Table S11. List of inversions found by Exo-C and their validation status. "NOT" indicates the rearrangement was not validated by the alternative method; "N/A" indicates the method was not applicable for the sample. Table S12. List of CNVs found by Exo-C and their validation status. Table S13. List of rare clinically-significant SNVs found by Exo-C and their validation status. "N/A" indicates the method was not applicable for the sample. Table S14. Primer sequences for obtaining a locus-specific DNA probe for the CSMD3 gene in 8q23.3 region. Table S15. Costs required to achieve comparable depth of the exome coverage. Table S16. Precision/Recall of CNV for A549 and K562 cells

Additional file 2: Figure S1. Translocations called by Exo-C caller, but not conventional methods. Figure S2. Detection of SVs using Exo-C data. Figure S3. iPS cells obtained from P10 lymphocytes. Figure S4. FISH results supporting SV structure in P47 patient cells. Figure S5. Examples of translocations called by Exo-C caller, but not conventional methods in K562 and A549

Acknowledgements

We acknowledge center for collective usage of computational facilities of the Institute of Cytology and Genetics (supported by budget project FWNR- 2022 - 0019) and HPC cluster of the Novosibirsk State University (Project #2019 - 0546, supported by FSUS № FSUS-2024-0018) for providing access to the computational resources.

Short-read sequencing of Exo-C libraries was partially supported by a grant from the Ministry of Science and Higher Education of the Russian Federation (Agreement No. 075 - 15 - 2022 - 310 dated April 20, 2022). Nanopore sequencing was supported by the Ministry of Science and Higher Education of the Russian Federation (Agreement 075 - 10 - 2021 - 093 (project GEN-RND- 2017)). Chromosomes and DNA samples were obtained from "Biobank of the population of Northern Eurasia" of the Research Institute of Medical Genetics, Tomsk NRMС (<http://medgenetics.ru/Biobank/>). Chromosomal microarray analysis (Agilent Technology), FISH and SKY were performed on Core Facility "Medical Genomics" (Tomsk NRMС, Russia).

We thank physicians for clinical investigations of patients and all families for participation in the study.

Code availability

Translocation and inversion calling tools developed in this study are publicly available on GitHub: <https://github.com/genomech/ExoC/> [55].

Authors' contributions

MG performed genomic experiments with help from YS, AY, and AK; EVG and MG obtained and differentiated iPS cells; TL performed computational analysis with help from PB, NT, EV, GK, and MN; AN performed targeted Oxford Nanopore sequencing with help from MK, and AM; MEL performed aCGH; SV and AZ performed FISH experiments; ADCh performed SKY; NNS performed conventional karyotyping; AK, EK, UB, OR, MS, AM, and MF performed NGS; LPN, AAK, ZGM, NAD, NK, TL, YM, EM, ES, EOB, OAS and MEM enrolled samples to cohort and interpreted the patient data; VF conceived the study and supervised the work with co-supervision from INL, NVS, and ER. VF and MG drafted the manuscript and all authors read and approved the final manuscript.

Funding

The study was supported by the Russian Science Foundation (#21–65–00017), <https://rscf.ru/project/21-65-00017/>. Study of P123 and Exo-C profiling of cancer cell lines A549 and K562 was supported by the grant of the state program of the "Sirius" Federal Territory "Scientific and technological development of the "Sirius" Federal Territory" (Agreement №26–03, 27/09/2024).

Data availability

Raw human sequencing data are classified as personal information and cannot be shared, in accordance with Russian regulations on personal data protection. Processed Exo-C data is available under accession GEO: GSE253950. (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE253950>) [56]. Raw data for K562 and A549 cells are available under accessions SRX23360397. (<https://www.ncbi.nlm.nih.gov/sra/SRX23360397>) [57] and SRX27828095 (<https://www.ncbi.nlm.nih.gov/sra/SRX27828095>) [58], respectively.

Declarations

Ethics approval and consent to participate

The recruitment of the cohort in this study was conducted in strict accordance with the principles of the Declaration of Helsinki and the International Conference on Harmonization Good Clinical Practice (ICH-GCP) guidelines. The study was approved by the local ethics committee of the Institute of Cytology and Genetics (protocol number 17, 16.12.2022) and the local ethics committee of the Tomsk National Research Medical Center (protocol number 15, 28.02.2023). Informed Consent was obtained from all patients or their parents/representatives included in the study.

Consent for publication

Written informed consent for publications was obtained from all patients or their parents/representatives for all materials included in the study. The consent permits the publication of clinical information while explicitly prohibiting the disclosure of personal identifiers, including participant names and raw (unprocessed) genomic sequencing data, in accordance with Russian legislation on personal data protection.

Competing interests

The authors declare no competing interests.

Author details

¹Institute of Cytology and Genetics, 10, Prospekt Akademika Lavrent'yeva, Novosibirsk 630090, Russia. ²Novosibirsk State University, 1, Pirogova Str, Novosibirsk 630090, Russia. ³Research Institute of Medical Genetics, Tomsk National Research Medical Center of the Russian Academy of Sciences, 10, Nab. Ushai, Tomsk 634050, Russia. ⁴Research Centre for Medical Genetics, Moscow 115522, Russia. ⁵Center for Family Care and Reproduction, 1 Kiyevskaya Str, Novosibirsk 6300136, Russia. ⁶Genetics and Reproductive Medicine Center, "GENETICO" PJSC, Moscow 119333, Russia. ⁷North-Western State Medical University named after I.I. Mechnikov, Saint-Petersburg 191015, Russia. ⁸Institute of Chemical Biology and Fundamental Medicine, Novosibirsk 630090, Russia. ⁹Artificial Intelligence Research Institute, Moscow, Russia 121170. ¹⁰Sirius University of Science and Technology, Sirius Federal Territory, Sochi 354340, Russia. ¹¹Sechenov First Moscow State Medical University, Moscow 119435, Russia. ¹²Endocrinology Research Center, Moscow 117292, Russia. ¹³UMass Chan Medical School, Worcester 01655, USA. ¹⁴Novosibirsk State Medical University, Novosibirsk, 630091, Russia.

Received: 4 May 2024 Accepted: 10 April 2025

Published online: 07 May 2025

References

1. Chung BHY, Chau JFT, Wong GK-S. Rare versus common diseases: a false dichotomy in precision medicine. *NPJ Genom Med*. 2021;6:19.
2. Katsanis SH, Katsanis N. Molecular genetic testing and the future of clinical genomics. *Nat Rev Genet*. 2013;14:415–26.
3. Miller CE, Krautscheid P, Baldwin EE, Tvrdik T, Openshaw AS, Hart K, et al. Genetic counselor review of genetic test orders in a reference laboratory reduces unnecessary testing. *Am J Med Genet A*. 2014;164A:1094–101.

4. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*. 2009;326:289–93.
5. Kantidze OL, Razin SV. Weak interactions in higher-order chromatin organization. *Nucleic Acids Res*. 2020;48:4614–26.
6. Kabirova E, Nurislamov A, Shadskiy A, Smirnov A, Popov A, Salnikov P, et al. Function and Evolution of the Loop Extrusion Machinery in Animals. *Int J Mol Sci*. 2023;24: 5017.
7. Dixon JR, Xu J, Dileep V, Zhan Y, Song F, Le VT, et al. Integrative detection and analysis of structural variation in cancer genomes. *Nat Genet*. 2018;50:1388–98.
8. Lukyanchikova V, Nuriddinov M, Belokopytova P, Taskina A, Liang J, Reijnders MJMF, et al. Anopheles mosquitoes reveal new principles of 3D genome organization in insects. *Nat Commun*. 2022;13:1960.
9. Gridina M, Mozheiko E, Valeev E, Nazarenko LP, Lopatkina ME, Markova ZG, et al. A cookbook for DNase Hi-C. *Epigenetics Chromatin*. 2021;14:15.
10. Gridina MM, Matveeva NM, Fishman VS, Menzorov AG, Kizilova HA, Beregovoy NA, et al. Allele-Specific Biased Expression of the CNTN6 Gene in iPSC Cell-Derived Neurons from a Patient with Intellectual Disability and 3p26.3 Microduplication Involving the CNTN6 Gene. *Mol Neurobiol*. 2018;55:6533–46.
11. Khabarova AA, Pristayzhnyuk IE, Nikitina TV, Gayner TA, Torkhova NB, Skryabin NA, et al. Induced pluripotent stem cell line, ICAGi001-A, derived from human skin fibroblasts of a patient with 2p25.3 deletion and 2p25.3-p23.3 inverted duplication. *Stem Cell Res*. 2019;34: 101377.
12. Gridina MM, Nurislamov AR, Minina JM, Lopatkina ME, Drozdov GV, Vasilyev SA, et al. Generation of iPSC cell line (ICGi040-A) from skin fibroblasts of a patient with ring small supernumerary marker chromosome 4. *Stem Cell Res*. 2022;61: 102740.
13. Gridina G, Popov A, Shadskiy A, Torgunakov N, Kechin A, Khrapov E, et al. Expanding the list of sequence-agnostic enzymes for chromatin conformation capture assays with S1 nuclease. *Epigenetics Chromatin*. 2023;16:48.
14. Grigóeva EV, Kopytova AE, Yarkova ES, Pavlova SV, Sorogina DA, Malakhova AA, et al. Biochemical characteristics of iPSC-Derived Dopaminergic Neurons from N370S GBA variant carriers with and without parkinson's disease. *Int J Mol Sci*. 2023;24:4437.
15. Eskildsen TV, Ayoubi S, Thomassen M, Burton M, Mandegar MA, Conklin BR, et al. MESP1 knock-down in human iPSC attenuates early vascular progenitor cell differentiation after completed primitive streak specification. *Dev Biol*. 2019;445:1–7.
16. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014. <https://doi.org/10.1093/nar/gkt958>. 42 Database issue:D986–992.
17. Firth HV, Richards SM, Bevan AP, Clayton S, Corpas M, Rajan D, et al. DECIPHER: Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *The American Journal of Human Genetics*. 2009;84:524–33.
18. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). <https://omim.org/>. Accessed 2 Apr 2025.
19. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. *Nat Biotechnol*. 2020;38:433–8.
20. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–100.
21. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. *Nat Commun*. 2017;8:1326.
22. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14:417–9.
23. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.
24. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. 2016;3:95–8.
25. Abdennur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics*. 2020;36:311–6.
26. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.
27. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res*. 2016;44:W160–165.
28. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. *Nature*. 2020;585:357–62.
29. Hunter JD. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*. 2007;9:90–5.
30. Waskom ML. seaborn: statistical data visualization. *Journal of Open Source Software*. 2021;6:3021.
31. Church DM, Schneider VA, Graves T, Auger K, Cunningham F, Bouk N, et al. Modernizing reference genome assemblies. *PLoS Biol*. 2011;9: e1001091.
32. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10: giab008.
33. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput Biol*. 2016;12: e1004873.
34. Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, et al. Copy number variation detection and genotyping from exome sequence data. *Genome Res*. 2012;22:1525–32.
35. Babadi M, Fu JM, Lee SK, Smirnov AN, Gauthier LD, Walker M, et al. GATK-gCNV enables the discovery of rare copy number variants from exome sequencing data. *Nat Genet*. 2023;55:1589–97.
36. Wang X, Luan Y, Yue F. EagleC: A deep-learning framework for detecting a full range of structural variations from bulk and single-cell contact maps. *Sci Adv*. 2022;8: eabn9215.
37. Ivanoshchuk DE, Shakhshneider EV, Rymar OD, Ovsyannikova AK, Mikhailova SV, Fishman VS, et al. The Mutation Spectrum of Maturity Onset Diabetes of the Young (MODY)-Associated Genes among Western Siberia Patients. *J Pers Med*. 2021;11:57.
38. Picard Tools - By Broad Institute. <https://broadinstitute.github.io/picard/>. Accessed 9 Nov 2024.
39. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303.
40. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38: e164.
41. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American college of medical genetics and genomics and the association for molecular pathology. *Genet Med*. 2015;17:405–24.
42. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011;27:2987–93.
43. Moquin SA, Thomas S, Whalen S, Warburton A, Fernandez SG, McBride AA, et al. The Epstein-Barr virus episome maneuvers between nuclear chromatin compartments during reactivation. *J Virol*. 2018;92:e01413–7.
44. Ray J, Munn PR, Vihervaara A, Lewis JJ, Ozer A, Danko CG, et al. Chromatin conformation remains stable upon extensive transcriptional changes driven by heat shock. *Proc Natl Acad Sci U S A*. 2019;116:19431–9.
45. Wang Y, Zhang Y, Zhang R, van Schaik T, Zhang L, Sasaki T, et al. SPIN reveals genome-wide landscape of nuclear compartmentalization. *Genome Biol*. 2021;22:36.
46. Melo US, Schöpflin R, Acuna-Hidalgo R, Mensah MA, Fischer-Zirnsak B, Holtgrewe M, et al. Hi-C Identifies Complex Genomic Rearrangements and TAD-Shuffling in Developmental Diseases. *Am J Hum Genet*. 2020;106:872–84.
47. Belaghzal H, Dekker J, Gibcus JH. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods*. 2017;123:56–65.
48. Nuriddinov MA, Belokopytova PS, Fishman VS. Charm is a flexible pipeline to simulate chromosomal rearrangements on Hi-C-like data. 2023;2023.11.22.568374. <https://doi.org/10.1101/2023.11.22.568374>
49. Chen Z, Yuan Y, Chen X, Chen J, Lin S, Li X, et al. Systematic comparison of somatic variant calling performance among different sequencing depth and mutation frequency. *Sci Rep*. 2020;10:3501.

50. Gabrielaite M, Torp MH, Rasmussen MS, Andreu-Sánchez S, Vieira FG, Pedersen CB, et al. A Comparison of tools for copy-number variation detection in germline whole exome and whole genome sequencing data. *Cancers (Basel)*. 2021;13:6283.
51. Gordeeva V, Sharova E, Babalyan K, Sultanov R, Govorun VM, Arapidi G. Benchmarking germline CNV calling tools from exome sequencing data. *Sci Rep*. 2021;11:14416.
52. Fishman VS, Salnikov PA, Battulin NR. Interpreting chromosomal rearrangements in the context of 3-dimensional genome organization: a practical guide for medical genetics. *Biochemistry (Mosc)*. 2018;83:393–401.
53. International Nucleome Consortium. 3DGenBench: a web-server to benchmark computational models for 3D Genomics. *Nucleic Acids Res*. 2022;50:W4–12.
54. Belokopytova P, Fishman V. Predicting genome architecture: challenges and solutions. *Front Genet*. 2020;11:617202.
55. Torgunakov N. Pipelines for searching chromosomal rearrangements as inversions and translocations in Exo-C data. <https://github.com/genomech/ExoC/>
56. Torgunakov N. A novel approach for simultaneous detection of structural and single-nucleotide variants based on a combination of chromosome conformation capture and exome sequencing.
57. Torgunakov N. SRX23360397 A novel approach for simultaneous detection of structural and single-nucleotide variants based on a combination of chromosome conformation capture and exome sequencing.
58. Torgunakov N. SRX27828095 A novel approach for simultaneous detection of structural and single-nucleotide variants based on a combination of chromosome conformation capture and exome sequencing. <https://www.ncbi.nlm.nih.gov/sra/SRX27828095>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.