Nucleic Acids Research, 2025, **53**, gkaf247
https://doi.org/10.1093/nar/gkaf247
Genomics

Long-read whole-genome sequencing-based concurrent haplotyping and aneuploidy profiling of single cells

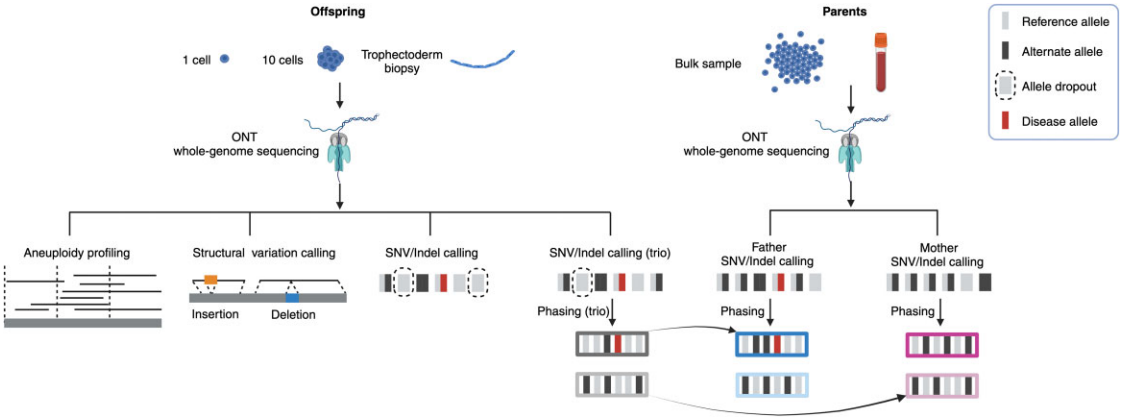
Yan Zhao [1], Olga Tsuiko[3], Tatjana Jatsenko[1], Greet Peeters[1], Erika Souche[1], Mathilde Geysens[1], Eftychia Dimitriadou[3], Arne Vanhie[2], Karen Peeraer[2], Sophie Debrock[2], Hilde Van Esch[3], Joris Robert Vermeesch[1,3,*]

[1]Laboratory for Cytogenetics and Genome Research, Department of Human Genetics, KU Leuven, 3000 Leuven, Belgium
[2]Leuven University Fertility Center, University Hospitals Leuven, Leuven 3000, Belgium
[3]Centre for Human Genetics, University Hospitals Leuven, Leuven 3000, Belgium
[*]To whom correspondence should be addressed. Email: joris.vermeesch@kuleuven.be

Abstract

Long-read whole-genome sequencing (lrWGS) enhances haplotyping by providing more phasing information per read compared to short-read sequencing. However, its use for single-cell haplotype phasing remains underexplored. This proof-of-concept study examines lrWGS data from single cells for small variant (single nucleotide variant (SNV) and indel) and structural variation (SV) calling, as well as haplotyping, using the Genome in a Bottle (GIAB) Ashkenazi trio. lrWGS was performed on single-cell (1 cell) and multi-cell (10 cells) samples from the offspring. Chromosome-length haplotypes were obtained by leveraging both long reads and pedigree information. These haplotypes were further refined by replacing them with matched parental haplotypes. In single-cell and multi-cell samples, 92% and 98% of heterozygous SNVs, and 74% and 78% of heterozygous indels were accurately haplotyped. Applied to human embryos for preimplantation genetic testing (PGT), lrWGS demonstrated 100% consistency with array-based methods for detecting monogenic disorders, without requiring phasing references. Aneuploidies were accurately detected, with insights into the mechanistic origins of chromosomal abnormalities inferred from the parental unique allele fractions (UAFs). We show that lrWGS-based concurrent haplotyping and aneuploidy profiling of single cells provides an alternative to current PGT methods, with applications potential in areas such as cell-based prenatal diagnosis and animal and plant breeding.

Graphical abstract



Introduction

Most mammalian genomes are diploid, consisting of one haploid set of chromosomes from each parent. Haplotyping reconstructs the unique nucleotide content of the two homologous chromosome sets known as haplotypes. This process is crucial because the haplotypes can have different functional roles [1]. Recent advancements in long-read sequencing technologies provide read lengths over 10 kb and accuracy comparable to next-generation sequencing (NGS) [2]. These improvements significantly enhance genome haplotyping, as individual long reads cover more heterozygous SNVs and provide haplotype information across extensive genomic regions, surpassing the capabilities of traditional single nucleotide polymorphism (SNP) arrays and short-read data. The main long-read sequencing technologies are the single molecule real-time (SMRT) sequencing from Pacific Bioscience (PacBio)

Received: September 23, 2024. **Revised:** March 13, 2025. **Editorial Decision:** March 14, 2025. **Accepted:** March 30, 2025

[3] and nanopore sequencing from Oxford Nanopore Technologies (ONT) [4]. Studies have highlighted the effectiveness of long-read sequencing in variant identification, haplotyping, and genome assembly. Wenger *et al.* demonstrated the ability of ∼28× PacBio high-fidelity (HiFi) reads for high-performance small variant (SNV and indel) calling and phased 99.64% of called variants. [5]. More recently, the Telomere-to-Telomere (T2T) Consortium created a complete human reference genome, T2T-CHM13, using PacBio HiFi reads and ONT ultralong reads [6].

In addition to enabling haplotyping and genome assembly for bulk samples, long-read sequencing also holds potential for facilitating haplotyping at the single-cell level. Genetic analysis for single cells is challenging due to the necessity of whole genome amplification (WGA), typically using techniques like multiple displacement amplification (MDA). WGA can introduce technical errors, including allele dropout (ADO) due to the failure to amplify one allele, false-positive errors resulting from polymerase infidelity, and coverage nonuniformity caused by uneven amplification [7]. Due to WGA artifacts, haplotype-based analysis of single cells is crucial for areas like preimplantation genetic testing for monogenic disorders (PGT-M). In this context, embryos from couples at risk of transmitting genetic disorders to their offspring are tested for the inheritance of disease alleles using DNA from trophectoderm (TE) biopsies containing 5–10 cells or from single blastomere biopsies. Several genome-wide haplotyping methods for single cells have been developed, including karyomapping [8], siCHILD [9], One preimplantation genetic testing (PGT) [10] Haploseek [11], and scGBS [12]. These methods utilize SNP arrays or NGS data for genotyping, which provide minimal or no haplotype information. As a result, genetic phasing is applied, requiring DNA samples from prospective parents and first-degree relative(s) (Carvalho *et al.*, 2020), which are not always available. Furthermore, in cases of *de novo* mutations (DNMs) in prospective parents, the variant loci cannot be phased through genetic phasing. Long-read data has the potential to directly phase both parents and embryo biopsies, including DNMs in the parents, without requiring relatives. Initial explorations focused on targeted long-read sequencing. For instance, Wu *et al.* conducted haplotype linkage analysis for the HBB gene by phasing the parents and TE biopsies using SMRT reads [13]. Similarly, Tsuiko *et al.* explored both SMRT and ONT data in preclinical workup to infer the parental origin of DNMs [14]. More recently, long-read whole-genome sequencing (lrWGS) has been employed for phasing parental genomes. Zhang *et al.* utilized ∼30× PacBio long-read data for phasing the parents and conducted reference-free PGT-M for three monogenetic diseases [15]. However, the effectiveness of lrWGS for phasing single cells and its application in generic PGT remains unexplored. Hård *et al.* assessed variant calling and genome assembly with lrWGS data from single cells [16], but the limited sequencing depth of ∼5× HiFi reads constrains a thorough exploration of its potential for clinical applications.

Beyond haplotyping, SNP arrays and NGS-based single-cell haplotyping methods enable concurrent haplotype-aware aneuploidy profiling, facilitating the identification of chromosomal abnormalities in single cells and determining their mechanistic origins. This has significant clinical implications, as chromosomal abnormalities can arise during human gametogenesis and are common in early embryogenesis [17, 18]. PGT for aneuploidy (PGT-A) prevents the transmission of chromosomally abnormal embryos and enhances the *in vitro* fertilization (IVF) success rate [19]. Long-read sequencing has proven valuable for detecting aneuploidies [20] and segmental imbalances [21] in embryo biopsies. However, to our knowledge, the potential of long-read data to infer the mechanistic origins of aneuploidies in embryo biopsies has not yet been explored.

Here, we present the first comprehensive analysis of lrWGS data from human single cells at an adequate depth of ∼24× for SNV, indel, and structural variation (SV) calling, as well as haplotyping. Using a Genome in a Bottle (GIAB) trio consisting of HG002 (offspring), HG003 (father), and HG004 (mother) for benchmarking, we demonstrate the feasibility of lrWGS data for concurrent haplotyping and aneuploidy profiling of single cells without requiring additional phasing references. The clinical proof-of-concept application was validated in two PGT families, achieving 100% diagnostic concordance with SNP array-based PGT results (Fig. 1). This lrWGS-based PGT approach surpasses current methods with reduced clinical work-up, fewer family members involved, and a more comprehensive genomic analysis that integrates direct variant detection, haplotyping, and aneuploidy assessment. Furthermore, our data analysis strategy for concurrent haplotyping and aneuploidy profiling of single cells can be applied to other areas of single-cell genome analysis, such as cell-based prenatal diagnosis and animal and plant breeding.

Materials and methods

lrWGS for the GIAB trio

Lymphoblastoid cell line of the offspring (HG002) from the GIAB Ashkenazi trio, consisting of HG002 (offspring), HG003 (father), and HG004 (mother) (Coriell Institute), was cultured in Dulbecco's modified Eagle's medium/F12, complemented with 10% fetal bovine serum (FBS) (Gibco). Three single-cell samples (one cell each, denoted as sc, sc_2, and sc_3) and one multi-cell sample (10 cells) were manually collected into 0.2 ml tubes and DNA from these samples was amplified by MDA using the REPLI-g single cell kit (QIAGEN). ONT libraries were prepared on 3 μg of amplified material using the SQK-LSK114 kit, following ONT's recommendations for WGA library preparations. A bulk sample was collected in parallel, and DNA was extracted using the genomic DNA purification kit (Monarch). Libraries were then prepared on 3 μg of bulk DNA using the SQK-LSK114 kit. All libraries were loaded (30 fmol) on the PromethION device for sequencing with R10.4.1 flow cells. Details regarding the base calling model and base caller used for each cell line sample are provided in Supplementary Table S1. lrWGS datasets for the father (HG003) and mother (HG004) were obtained from the Oxford Nanopore Open Datasets. Specifically, BAM files were downloaded from the AWS storage bucket at s3://ont-open-data/giab_lsk114_2022.12/ and downsampled to ∼24×, using samtools view (v 1.9) [22], comparable to coverage of the offspring. Additionally, the lrWGS dataset for the offspring (HG002) was also downloaded from the same source to serve as a biological replicate of our HG002 bulk sample.

lrWGS for human PGT families

This study was approved by the Ethical Committee of UZ/KU Leuven (S68291). The parents consented to the use of residuary Human Bodily Material for scientific research at the start of their IVF treatment. IVF and universal PGT-related

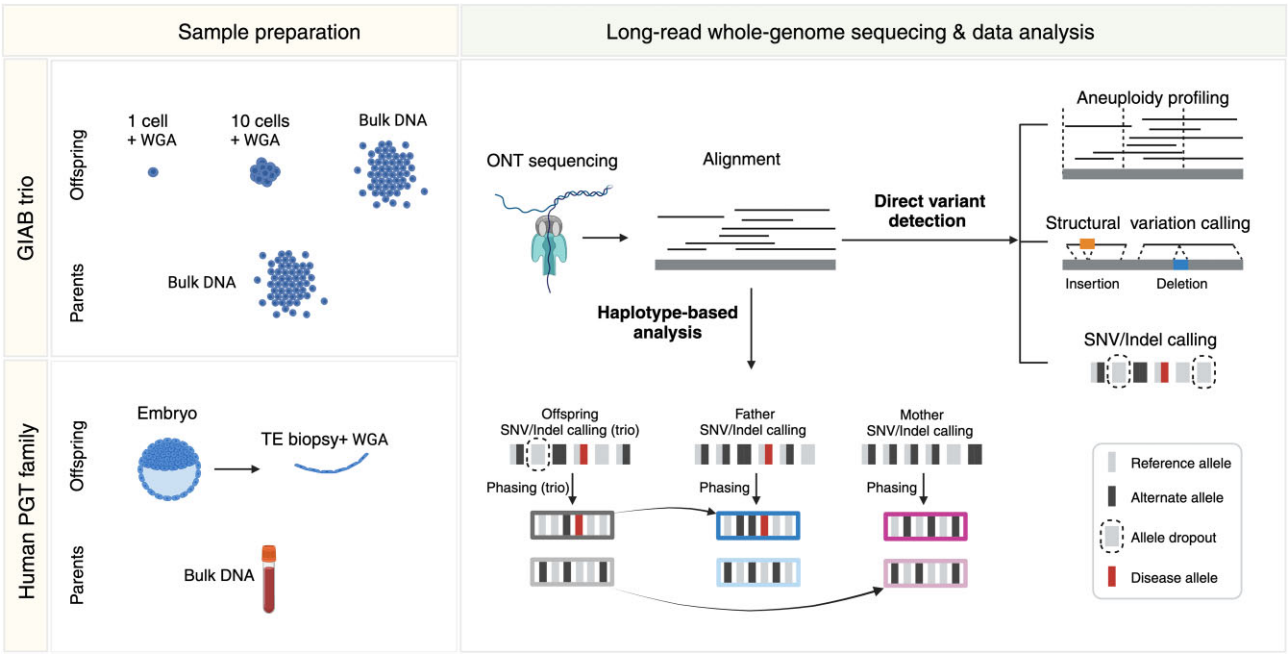


Figure 1. lrWGS-based concurrent direct variant detection and haplotyping workflow. The benchmarking study utilized a GIAB trio consisting of HG002 (offspring), HG003 (father), and HG004 (mother). Single-cell (1 cell), multi-cell (10 cells), and bulk samples were collected from the HG002 cell line. The single-cell and multi-cell samples underwent WGA before lrWGS using nanopore technology. After sequencing, the long reads were aligned to the human reference genome. Direct variant detection for aneuploidies, SVs and small variants (SNVs and Indels) was performed using mapped lrWGS data from the offspring. In addition to direct SNV/Indel calling, the performance of haplotype-based analysis to determine whether a disease allele present in the parents was inherited by the offspring was evaluated. For haplotype-based analysis, publicly available aligned lrWGS data for the parents were used, with individual variant calling and phasing performed. The offspring was analyzed in a trio setting, incorporating parental information for better results. By comparing the offspring's haplotypes with those of the parents, the inherited parental haplotypes were inferred and used for diagnosis. In the shown example, the disease allele inherited by the offspring was identified by both haplotype-based analysis and direct variant detection. The utility of lrWGS-based small variant calling, haplotyping and aneuploidy profiling was further evaluated in families undergoing PGT-M, where DNA from TE biopsies of embryos and bulk DNA from the prospective parents were analyzed.

procedures have been performed according to the standard operating procedures of UZ Leuven. Specifically, for universal PGT, parental bulk DNA was extracted from whole blood. TE biopsy was performed on blastocyst-stage embryos, obtaining an average of five TE cells, which were subjected to WGA using the MDA method with the REPLI-g SC kit (Qiagen, Germany). Residual DNA from both parents and corresponding embryo biopsies (MDA amplified) is available at the hospital, and an aliquot of excess DNA material were utilized for this study. ONT library preparation and sequencing were done using the same procedures as for cell line samples, with the base calling model and base caller applied for each sample listed in Supplementary Table S1.

Read preprocessing and mapping

Read quality was assessed using NanoPlot (v1.39.0) [23] and FastQC (v0.11.7). Reads with an average quality score below 9 or a length shorter than 500 bp were filtered out using NanoFilt (v2.8.0) [23]. The processed reads were then aligned to the human reference genome hg38 using minimap2 (v2.12) [24]. Mapping statistics were generated using samtools flagstat (v1.9) [22]. Chimeric read count was determined by counting the number of unique reads for all alignments with 0 × 800 SAM flag. The lengths of chimeric read-derived segments were determined using information from the mapped BAM file as follows: For primary alignments, segment lengths were calculated as the lengths of the 'seq' field minus the lengths of the soft-clipped bases. For supplementary alignments, segment lengths corresponded to the lengths of the 'seq' field. Depth of coverage for each genomic position was computed using samtools depth with -aa flag (v1.9) [22], and average depth of coverage was determined by dividing the sum of depths across all positions by the size of the genome.

Variant calling

Clair3 (v0.1) [25] was used for small variant calling with BAM file of a single individual. Clair3-Trio (v0.3) [26] was used for trio small variant calling with BAM files of the parents and the offspring. For both strategies, default parameters were used and SNVs and short indels (≤50 bp) were called. GIAB benchmark data (v4.2.1) was used to assess small variant calling performance on high-confidence regions. VCF files containing high confident small variants on autosomes and corresponding BED files containing high confident regions were obtained for each trio member from https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/. RTG Tools (v3.12.1) [27] was used for benchmarking analysis with the vcfeval command. To include supplementary alignments in SNV calling (omitted by default), the SAM flags of the supplementary alignments (2048 and 2064) were first modified as normal flags (0 and 16, respectively). The modified BAM files were then utilized for SNV calling with Clair3 (v0.1).

Sniffles2 (v2.5) [28] was used for SV calling, incorporating tandem repeat annotations to improving SV detection in repetitive regions. SV calling performance was evaluated using the SV draft benchmark set (v1.1) for HG002 and corresponding

benchmark regions (obtained from https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/data/AshkenazimTrio/analysis/NIST_HG002_DraftBenchmark_defrabbV0.019-20241113). Filter-passed SVs in the benchmark set (32 551 deletions and 41 143 insertions), along with filter-passed deletions, insertions, and duplications in the call set, were retained for benchmarking using Truvari (v4.3.1) [29] bench command with the –dup-to-ins flag.

Haplotype phasing of small variants

The default phasing mode of WhatsHap (v1.0) [30], which uses data from a single individual, was employed to phase variants identified by Clair3. The pedigree phasing mode of WhatsHap (v1.0) [30], utilizing data from both parents and the offspring, was used to phase parental variants called by Clair3 and offspring's variants called by Clair3-Trio. For both modes, default parameters were applied to phase only SNVs, while the –indels flag was added to phase both SNVs and indels. Only variants with GQ values higher than 2 in VCF files were retained for phasing. Preliminary conservative paternal|maternal phasing data from GIAB (available at https://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/AshkenazimTrio/HG002_NA24385_son/NISTv4.2.1/GRCh38/SupplementaryFiles/HG002_GRCh38_1_22_v4.2.1_benchmark_phased_MHCassembly_StrandSeqANDTrio.vcf.gz) was used as a benchmark for evaluating phasing performance with WhatsHap Compare (v1.0) [30].

Inference of inherited parental haplotypes

To deduce the parental haplotypes inherited by the offspring, haplotypes of the offspring obtained from the pedigree phasing mode were compared with parental haplotypes from the default phasing mode. Only biallelic loci were retained for comparison, excluding those with unphased genotypes in either parent or identical homozygous genotypes in both parents. Each chromosome was divided into 1 Mb consecutive segments, referred to as comparison units, and haplotype comparisons were performed within each segment. Since in the resulting VCF file from pedigree phasing mode, haplotype alleles of the offspring are given as paternal|maternal, for each comparison unit, we compared the first haplotype of the offspring to the two paternal haplotypes and the second haplotype of the offspring to the two maternal haplotypes. During comparison, loci with unphased genotypes or genotypes that violated Mendelian inheritance rules in the offspring were disregarded. The inherited parental haplotypes were identified as those exhibiting the highest number of matched SNVs and were used to replace the original haplotype information in the offspring. For each locus of interest, the offspring's genotype was determined from these inherited parental haplotypes.

Performance evaluation for haplotype linkage analysis and direct variant detection

To evaluate the performance of both haplotype linkage analysis and direct variant detection, familial high-confidence regions were first obtained by intersecting the high-confidence regions of each trio member using bedtools multiinter (v2.27.1) [31]. The familial high-confidence regions were then intersected with protein-coding gene regions (extracted from the genome annotation file downloaded from https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_42/gencode.v42.annotation.gtf.gz) using bedtools intersect (v2.27.1) [31]. Within the resulting intersected regions, loci with parental genotype combinations of heterozygous (0/1) and homozygous reference (0/0) were selected. The performances of the two methods were then assessed by evaluating whether the genotypes of the offspring at these loci could be accurately identified from the inferred parental haplotypes (haplotype linkage analysis) or from default variant calling results (direct variant detection).

DNM screening

Genome-wide DNM screening was restricted to biallelic SNV loci within familial high confidence gene regions as detailed above. DNMs were identified by comparing the genotypes of the trio members. A locus was classified as DNM if both parental genotypes were homozygous reference while the offspring showed a different genotype. Further detailed analysis of the identified DNMs were done with a customized R script.

Aneuploidy profiling

Aneuploidy profiling was performed using NanoGLADIATOR (v1.0) [32] with a window size of 1000 000 bp.

To determine parental haplotype contributions across the genome, we selected SNV loci where the parents exhibited differing homozygous genotypes (homozygous reference (0/0) for one parent and homozygous alternate (1/1) for the other). Only loci with GQ > 2 and depth between 5 and 50 in the offspring were retained. For each locus, we computed the paternal and maternal allele fractions for the offspring. If a parent's genotype was 1/1, the allele frequency (AF) value for the offspring was used as the allele fraction for that parent. Conversely, 1 minus AF was used as the allele fraction if the parent's genotype was 0/0. Subsequently, we grouped loci within fixed bin sizes of 1 MB and calculated the average paternal and maternal allele fractions within each bin. Bins containing fewer than 30 000 loci were excluded. The calculated average paternal and maternal allele fraction values then underwent Circular Binary Segmentation using the R package PSCBS. Finally, we plotted the mean parental allele fraction values for each bin along with the segmentations across the genome. The parental haplotype fraction for each chromosome was determined through visual inspection of the plot, enabling the identification of the parental origin of any aneuploidies.

To determine the mitotic or meiotic origin of an aneuploidy with identified parental origin, we analyzed SNV loci that were heterozygous in the parent causing the aneuploidy and, in the offspring, but homozygous in the other parent. For each locus in the offspring, the unique allele fraction (UAF) was inferred from the AF value. The unique allele refers to the allele in the heterozygous parent that differs from the allele in the homozygous parent. The procedures used for segmentation and visualization of parental allele fraction values, as described above, were applied to visualize UAF values across the genome. The mitotic or meiotic origin of the aneuploidy was inferred through visual inspection of the plot.

Results

Small variant (SNV and indel) and SV calling with lrWGS data from single cells

To evaluate the potential of lrWGS-based haplotyping for single cells, we took one single-cell and one multi-cell (10 cells)

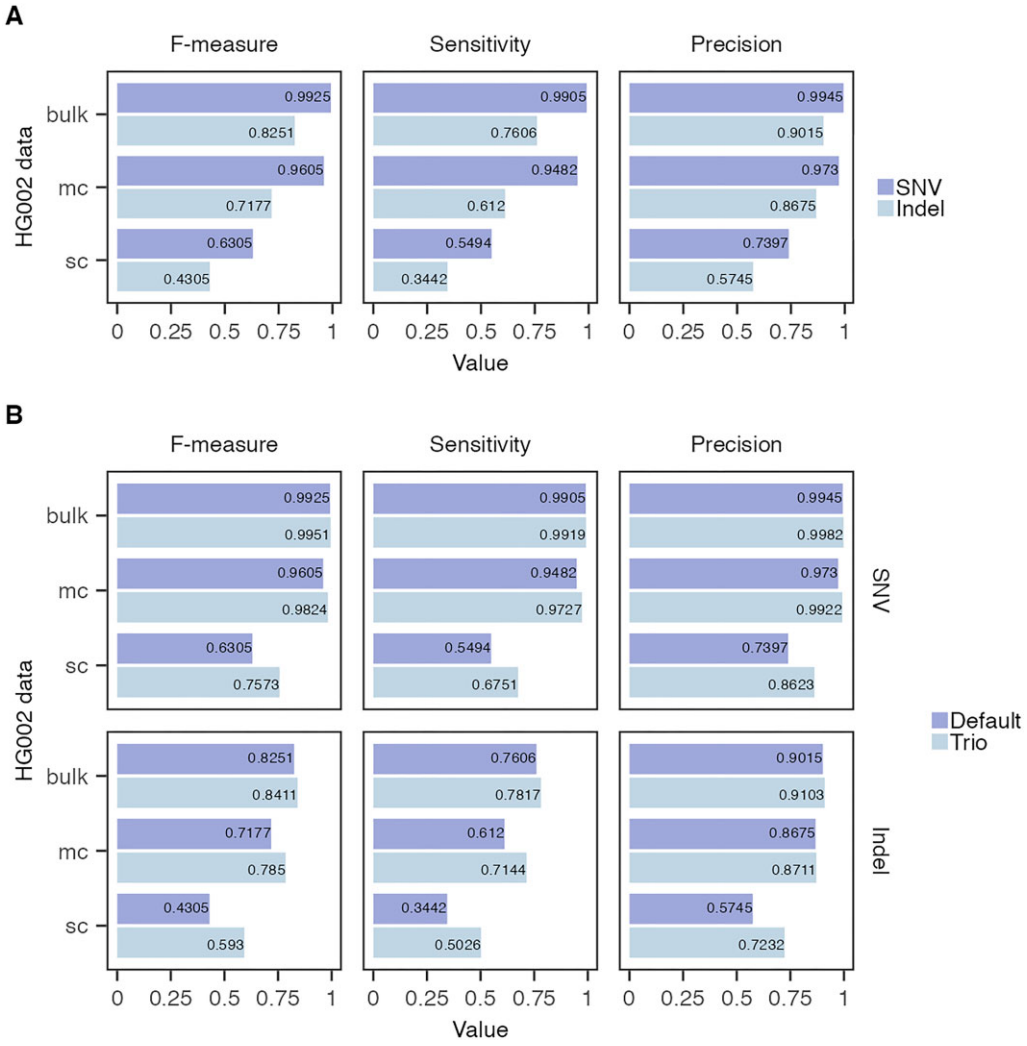


Figure 2. Small variant (SNV and indel) calling performance for lrWGS data from single-cell (sc), multi-cell (mc), and bulk samples of the offspring. (**A**) Variant calling performance in individual samples. (**B**) Trio variant calling yields varying degrees of improvements in variant calling performance for both SNVs and indels. F-measure represents the harmonic mean of precision and sensitivity.

sample from the offspring of the GIAB trio, mimicking single blastomere and TE biopsy, respectively. Additionally, a bulk sample was included for comparison (Fig. 1). Basic statistics of the reads generated for each sample are presented in Supplementary Fig. S1 and Supplementary Table S2. We obtained 22–24× lrWGS data for the bulk, multi-cell, and single-cell samples, covering 95%, 94%, and 88% of the human genome, respectively (Supplementary Table S3).

Given that SNVs and indels are the primary focus of most PGT-M cases and serve as genetic markers for haplotype construction, we first performed SNV and indel calling for lrWGS data using Clair3 [25]. Variant calling performance was assessed by comparing with GIAB benchmark data. Across all sample types, we observed better variant calling performance for SNVs than for indels (Fig. 2A). Among different sample types, single-cell data showed the lowest performance, while multi-cell data was more similar to bulk data. Specifically, multi-cell data exhibited an SNV F-measure of 0.9605, comparable to bulk data at 0.9925, whereas the F-measure for single-cell data decreased significantly to 0.6305 (Fig. 2A). This underscores that MDA-induced amplification errors, in addition to sequencing errors, negatively impact the genotyping accuracy of single-cell and multi-cell samples. Not surprisingly, for single-cell data, the sensitivity for heterozygous SNVs was notably lower (0.3844) compared to homozygous SNVs (0.8422) (Supplementary Fig. S2), suggesting a high rate of ADO.

Both single- and multi-cell data contain a high percentage of chimeric reads (55% and 48% for single- and multi-cell data, respectively) (Supplementary Fig. S3A). This is expected due to the nature of MDA amplification which is characterized by the formation of chimeric DNA rearrangements. During mapping, each chimeric read was fragmented into multiple smaller segments and mapped to their original positions within the genome. Among these alignments, one was selected as the representative alignment, while all others were classified as supplementary alignments (Supplementary Fig. S3B). For both multi- and single-cell data, the segments derived from chimeric reads during mapping were significantly shorter than the original chimeric reads, with an N50 (3570 bp for multi-cell and 3284 bp for single-cell data) approximately half the N50 of the original chimeric reads (Supplementary Fig. S3C

and D). By default, supplementary alignments were not utilized for small variant calling. We hypothesized that incorporating supplementary alignments might enhance coverage in specific genomic regions and improve SNV calling performance. Hence, we conducted tests that deliberately included supplementary alignments for variant calling. In contrast to expectation, we noted slightly reduced SNV F-measures compared to results obtained without including supplementary alignments (Supplementary Fig. S3E). Supplementary alignments were thus not utilized for small variant calling throughout this study.

High-quality SNVs and indels are required to enable haplotype phasing. Since both single- and multi-cell data showed lower SNV and indel calling performance compared to bulk data (Fig. 2A), we aimed to improve variant calling performance by incorporating parental data (publicly available lrWGS ONT data with ∼24× coverage, see the 'Materials and methods' section) to enable trio information-aware variant calling and reduce Mendelian inheritance violation variants. Trio variant calling was performed using Clair3-Trio [26], resulting in varying degrees of improvements in variant calling performance for the offspring, with the most substantial increase observed for single-cell data (Fig. 2B). Parental variant detection performance did not yield obvious benefits from trio variant calling (Supplementary Fig. S4).

We then performed SV calling for bulk, multi-cell, and single-cell lrWGS data of the offspring using Sniffles2 [28]. The F-measures were 0.736 for bulk data, 0.6057 for multi-cell data, and 0.4347 for single-cell data, relative to the SV benchmark set (Supplementary Fig. S5A,B). The relatively low F-measures, even for bulk data, could be attributed to the draft nature of the benchmark set, which may contain false SVs (see the 'Materials and methods' section). The lower SV calling performance of multi-cell and single-cell data compared to bulk data indicated an adverse consequence of the high abundance of chimeric reads in these datasets, which hinders SV calling. Additionally, a smaller fraction of SVs within tandem repeat regions were correctly identified compared to those outside these regions across all data types (bulk, multi-cell, and single-cell), underscoring the challenges of SV calling in tandem repeat regions (Supplementary Fig. S5C).

To evaluate the consistency of SNV, indel and SV calling performance for lrWGS data from the same sample type at similar depth of coverage, we generated both technical and biological replicates for different sample type. Technical replicates were created by downsampling the bulk, single-cell, and multi-cell data to 6×, 12×, and 18× coverage. Each downsampling was performed twice to produce pseudo-technical replicates. Analysis of these technical replicates showed high consistency in variant calling performance, with F-measure values increasing as coverage depth rose from 6× to 18× (Supplementary Fig. S6A). For biological replicates, we used bulk lrWGS data for HG002 from the GIAB project as a biological replicate. Additionally, two datasets derived from distinct single-cell samples (sc_2 and sc_3) served as biological replicates for the original single-cell sample (Supplementary Fig. S1 and Supplementary Tables S2 and S3). Consistent variant calling performance was observed across these biological replicates, with F-measure values improving as coverage increased from 6× to 24× (Supplementary Fig. S6B). These findings confirmed that lrWGS data from the same sample type, with similar read depths, resulted in comparable variant calling performance.

lrWGS of single cells enables accurate phasing of the human genome

The next step is to achieve accurate phasing. We used the pedigree phasing mode of WhatsHap [30] to achieve high phasing performance by incorporating parental data, combining both read-based and genetic phasing. Meanwhile, we also tested the default phasing mode of WhatsHap [30], which relies solely on read-based phasing. We then compared the trio-based variant calling and phasing results with those obtained using the default settings. We tested phasing with only SNVs and with both SNVs and indels, as these two variant types demonstrated different variant calling performances. When phasing only SNVs, 95%–98% of heterozygous SNVs were phased into 2350, 59 763, and 121 116 blocks, with corresponding block NG50 of 2964 293, 29 851, and 0 bp, covering 90%, 64%, and 40% of the autosomal regions for bulk, multi-cell, and single-cell data respectively. Adding indels for phasing achieved similar statistics (Fig. 3A–E). When performing pedigree phasing the outcome was spectacularly improved, especially for multi- and single-cell data (Fig. 3A–E). The most significant improvement was the generation of chromosome-long haplotypes with one block per autosome, covering 98% of the autosomal region for all data types (Fig. 3C and E). We assessed the accuracy of the phased blocks by comparing them with the phased GIAB benchmark data and obtained switch error rates as an indication of phasing accuracy. A lower switch error rate indicates higher accuracy. We observed higher switch error rates in the phased blocks when phasing both SNVs and indels compared to phasing only SNVs, likely due to lower indel calling performance. Compared to default phasing, pedigree phasing resulted in decreased accuracy for bulk data but improved accuracy for multi- and single-cell data. For bulk data, the switch error rates increased from 0.09% to 0.11% when phasing only SNVs and from 0.22% to 0.49% when phasing both SNVs and indels. In contrast, for multi-cell data, the switch error rates decreased from 0.40% to 0.11% when phasing only SNVs and from 0.54% to 0.44% when phasing both SNVs and indels. For single-cell data, the rates decreased from 1.31% to 0.19% when phasing only SNVs and from 1.46% to 0.50% when phasing both SNVs and indels (Fig. 3F).

Performance comparison of direct variant detection and haplotype-based variant inference

Given the promising quality of trio-based variant calling and phasing from single- and multi-cell data, we hypothesized that the resulting haplotypes could be used to infer parental haplotypes inherited by the offspring, enabling genotype extraction based on these inferred parental haplotypes. To infer the transmitted parental haplotypes, we compared parental haplotypes from default phasing mode with haplotypes of the offspring from pedigree phasing mode. Haplotype comparison was conducted within 1Mb segments across the genome (the 'Materials and methods' section; Fig. 1). During haplotype comparison, we observed that Mendelian inconsistency and ADO rates were significantly higher when using single-cell data from the offspring, compared to using bulk data. In contrast, the rates were only slightly higher when using multi-cell data (Supplementary Table S4). This observation further indicates MDA-induced amplification errors in amplified DNA. Compared to the transmitted parental haplotypes inferred from the offspring's bulk data phasing results, those

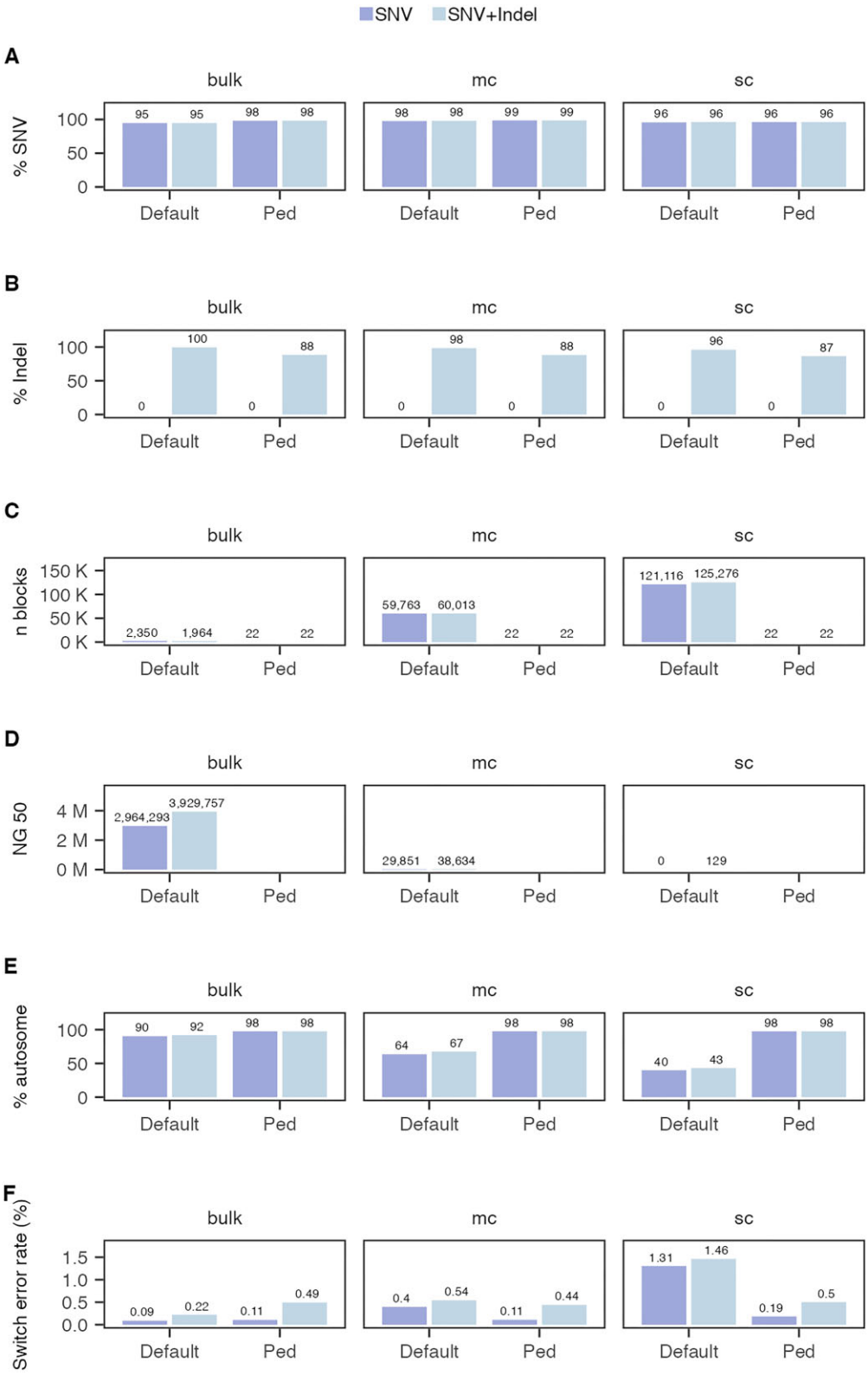


Figure 3. Phasing performance is superior with the pedigree phasing mode (ped) compared to the default phasing mode (default) for single-cell (sc) and multi-cell (mc) data of the offspring. (**A**) Percentage of phased heterozygous SNVs. (**B**) Percentage of phased heterozygous indels. (**C**) Total number of phased blocks. (**D**) NG50 of phased blocks. (**E**) Percentage of autosomal region covered by phased blocks. (**F**) Switch error rate. Shown are statistics for autosomes. For NG50, values are calculated per chromosome and averaged across autosomes. NG50 values are not applicable for pedigree phasing results because each chromosome has only one phased block.

inferred from the offspring's multi- and single-cell data phasing results demonstrated consistencies of 95% and 82% when phasing only SNVs, and 93% and 82% when phasing both SNVs and indels, respectively (Supplementary Fig. S7).

We then assessed the specific added value of using long reads for phasing. With short-read or SNP arrays data, genetic phasing is typically applied within a pedigree. This method can phase variant loci that are heterozygous in one parent and homozygous in the other, as well as loci exhibiting different homozygous genotypes in the parents. However, loci that are heterozygous across all trio members remain unresolved. In contrast, long-read data have the potential to phase these loci. We evaluated the proportions of these loci among all loci used for haplotype comparison (Supplementary Fig. S8). The results showed that 15%–19% of informative SNVs and 14%–17% of informative indels were heterozygous across trio members and could only be phased with long reads (Supplementary Fig. S8). This indicates that lrWGS data enable more loci to be phased and utilized for subsequent analyses.

Next, we compared the performance of two variant detection approaches: direct variant detection and haplotype-based analysis. In direct variant detection, variants are identified from default variant calling results, whereas in haplotype-based analysis, variants are inferred from the transmitted parental haplotypes (Fig. 1). Using GIAB benchmark data, we selected loci within high-confidence protein-coding gene regions where one parent is homozygous reference (0/0) and the other is heterozygous (0/1). The offspring can be either heterozygous (0/1) or homozygous reference (0/0), allowing us to evaluate the performance of both approaches in detecting transmitted coding variants. In total, we selected 812 973 SNV loci, with 407 232 heterozygous and 405 741 homozygous reference in the offspring, and 102 465 indel loci, with 50 650 heterozygous and 51 815 homozygous reference in the offspring. Using direct variant detection, we correctly identified 94% of heterozygous SNVs in multi-cell data and 43% in single-cell data, with no false positives for homozygous reference loci (Fig. 4A). Performance for indel loci was lower, with 68% of heterozygous indels detected in multi-cell data and 26% in single-cell data, along with 1% false positives for homozygous reference loci (Fig. 4B). In contrast, haplotype-based analysis demonstrated superior performance over direct variant detection, accurately inferring 98% of heterozygous SNVs in multi-cell data and 92% in single-cell data, with 2%–3% false positives for homozygous reference loci (Fig. 4A). The performance for indel loci was still lower, identifying 78% of heterozygous indels in multi-cell data and 74% in single-cell data, with 4%–5% false positives for homozygous reference loci (Fig. 4B). Multi-cell data demonstrated direct variant detection performance comparable to bulk data and also benefited from haplotype-based analysis, identifying more heterozygous loci despite a small rise in false positives for homozygous reference loci (Fig. 4).

Exploration of DNM screening

An essential aspect of single-cell genomics is exploring new mutations that arise in individual cells during processes such as cell division, cancer genesis, and tissue differentiation. Understanding DNM rates at the single-cell level will deepen our knowledge of the heterogeneity of both cancerous and normal cells, as well as the mechanisms driving cancer progression. Additionally, a large proportion of common and rare genetic disorders are a consequence of DNMs [33]. Therefore, we assessed the efficacy of lrWGS data for genome-wide DNM screening, focusing on SNVs. We identified 452 heterozygous SNVs in the offspring as "true DNMs" within high-confidence protein-coding regions of the GIAB benchmark data, with both parents exhibiting homozygous reference genotypes at the corresponding loci. This is significantly higher than expected, since the number of new point mutations present in human offspring is on average 60 (30–90 depending on parental age at conception), with around one DNM per exome [33–36]. Hence, most are likely false positives or mosaic variants introduced by cell culture. Of the 452 DNMs detected in the benchmark data, 407 (90%) were identified using multi-cell data, while 167 (37%) were detected with single-cell data from the offspring. However, the false positive rates were 87 434 out of 87 841 (99.5%) for multi-cell data and 321 201 out of 321 368 (99.9%) for single-cell data (Supplementary Table S5). Most of the false candidate DNMs in bulk data should be due to genotyping errors caused by sequencing errors. In contrast, the significantly larger number of candidate DNMs observed in single-cell and multi-cell data should be due to both sequencing errors and errors introduced during the WGA process. The latter is the primary cause, as evidenced by the substantially higher number of false positive DNM candidates observed with single-cell and multi-cell data compared to bulk data—~61 times more with single-cell data and 17 times more with multi-cell data (Supplementary Table S5). In summary, the abundance of false positives complicates DNM screening when using single-cell or multi-cell lrWGS data of the offspring for DNM screening.

Direct variant detection, haplotype-based variant inference, and haplotype-aware aneuploidy profiling in human embryos using lrWGS data

Since the proof-of-concept study with GIAB cell lines demonstrated promising performance, we proceeded to test lrWGS-based PGT on TE biopsies from five human embryos derived from two different couples. Each TE biopsy contains 5–10 cells, corresponding to the multi-cell sample in above benchmark study. The DNA was previously analyzed using clinically accredited SNP array-based comprehensive PGT, and the results were used as a reference (Supplementary Table S6 and Supplementary Fig. S9). For family ONT1, the father carried a pathogenic SNV (c.1384C > T) in the *MSH2* gene causing Lynch syndrome, an autosomal dominant cancer predisposition syndrome. For family ONT2, the mother carried an indel (c.2955delG), and the father carried a pathogenic SNV (c.1133–708A > G) in the *LAMB3* gene, responsible for autosomal recessive junctional epidermolysis bullosa. Two embryos from family ONT1 (ONT1-E02, ONT1-E03) and three from family ONT2 (ONT2-E04, ONT2-E06, ONT2-E20) were tested using lrWGS-based PGT. Basic statistics of the reads generated for each sample are presented in Supplementary Fig. S1 and Supplementary Table S2. We obtained 21–31× coverage following lrWGS of the parents and embryos, covering 93%–95% of the human genome (Supplementary Table S7). With direct mutation detection, the carrier status of the variant alleles was accurately determined, except for one indel that was missed in ONT2-E20. However, we identified three out of five reads supporting

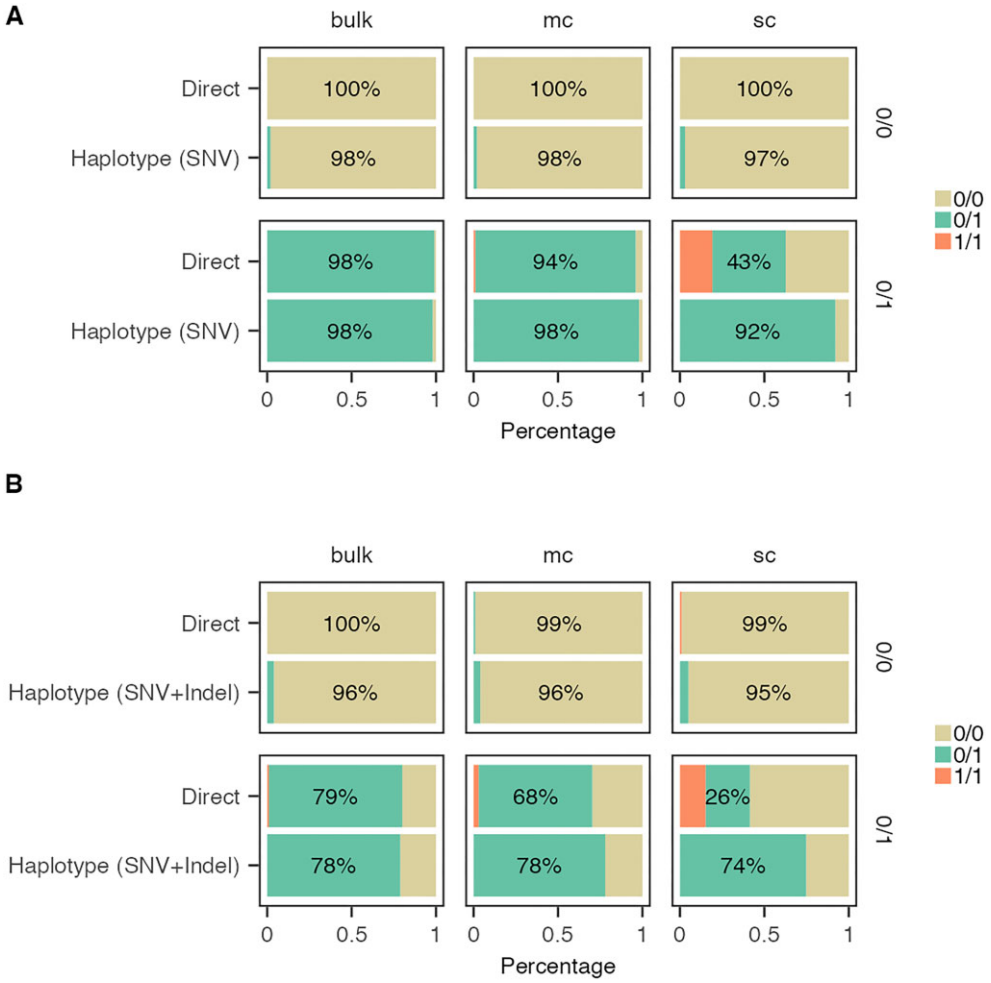


Figure 4. Performance comparison of direct variant detection and haplotype-based variant inference on selected (**A**) SNV and (**B**) indel loci using single-cell (sc), multi-cell (mc), and bulk data of the offspring. Direct: direct variant detection. Haplotype (SNV): haplotype-based analysis with only SNVs phased in the haplotypes. Haplotype (SNV + Indel): haplotype-based analysis with both SNVs and indels phased in the haplotypes. The true genotypes from GIAB benchmark data are indicated on the right. The proportions of inferred genotypes are color-coded. 0/0 for homozygous reference, 0/1 for heterozygous, and 1/1 for homozygous alternate.

the presence of this deletion (Supplementary Table S8). During haplotype comparison, we observed Mendelian inconsistency and ADO rates comparable to those observed in the proof-of-concept study when using multi-cell data from the offspring (Supplementary Table S9). Haplotype-based analysis of disease allele inheritance showed 100% concordance with SNP array-based PGT-M results (Fig. 5). Specifically, SNVs were phased to identify the risk parental haplotypes carrying the disease alleles and to infer the parental haplotypes inherited by the embryo, from which the embryo's carrier status can be determined (Fig. 5). For the indel in the mother of family ONT2, the risk haplotype carrying the indel was identified by phasing both the indel and the SNVs. The indel locus was then incorporated into the SNV phasing results, with each allele assigned to the corresponding maternal haplotype (Fig. 5B). We also assessed the genome-wide consistency of embryonic SNV genotypes inferred from SNP array versus lrWGS-based haplotyping. The analysis showed 98–99% concordance between the two technologies across the five embryos (Supplementary Table S10), demonstrating high consistency in haplotyping results for both technologies.

Since SNP array-based comprehensive PGT allows for the concurrent analysis of aneuploidy (PGT-A) and enables the identification of mosaic aneuploidy and the origin of aneuploidy, we explored whether these can also be detected by lrWGS data. We identified all aneuploidies and their mechanistic origins using lrWGS data, achieving 100% consistency with SNP array-based PGT (Fig. 6 and Supplementary Figs S9 and S10). Specifically, in ONT1-E02 we observed complex segmental aberration on chromosome 2, with mosaic duplication of short arm and mosaic deletion of the long arm. In addition, the same embryo exhibited complex segmental aberrations on chromosome 14 (Supplementary Fig. S10). Additionally, we detected trisomy on chromosomes 9 and 16 in ONT2-E20 (Fig. 6A), with parental allele fractions indicating the presence of an extra maternal chromosome (Fig. 6B). To determine the mitotic or meiotic origin of the extra maternal chromosomes, we analyzed loci with homozygous genotypes in the father and heterozygous genotypes in the mother and the embryo. For each locus, we calculated the UAF in the embryo, with the unique allele being the distinct allele among the four parental alleles. The distribution of the UAF across the chromosome helps identify the mechanistic origin of the

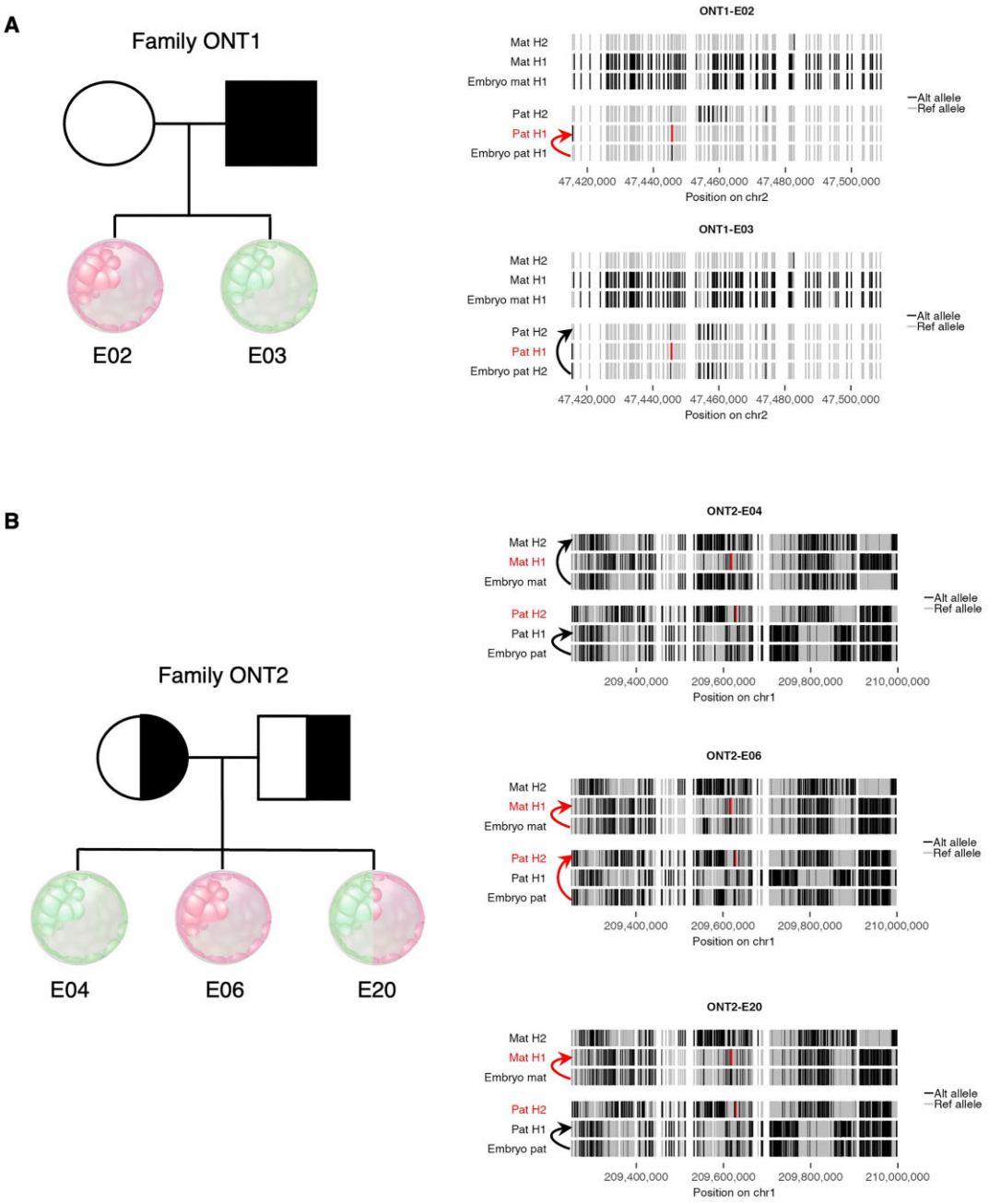


Figure 5. Visualization of haplotype-based PGT-M in human embryos. Pedigree plots for the two PGT families are shown on the left. Embryos are color-coded based on their carrier status as determined by haplotype-based PGT-M: green indicates that the disease allele was not inherited, red indicates that the dominant mutation was inherited or both recessive mutations were inherited, and green/red mosaic indicates the inheritance of a single recessive mutation. SNV haplotyping results for the chromosomal region linked to the disease loci are shown on the right. In the maternal (Mat H1, Mat H2) and paternal (Pat H1, Pat H2) haplotypes, each vertical line represents an allele from an informative locus: gray indicates the reference allele, black indicates the alternate allele, and disease alleles are shown in red. Parental haplotypes that carry disease alleles are labeled in red text. Arrows link the embryonic haplotypes to their corresponding maternal and paternal haplotypes, with red arrows used when the corresponding parental haplotypes carry the disease alleles. For the indel in the mother of family ONT2, both SNVs and indels were first phased to obtain phasing information for the indel locus. This locus was then added to the SNV phasing results, with each allele assigned to its corresponding haplotype. (**A**) In Family ONT1, ONT1-E02 inherited the paternal haplotype carrying the disease allele, while ONT1-E03 inherited the normal paternal haplotype. (**B**) In Family ONT2, ONT2-E04 inherited normal haplotypes from both parents, ONT2-E06 inherited pathogenic allele-carrying haplotypes from both parents, and ONT2-E20 inherited the normal haplotype from the father and the pathogenic allele-carrying haplotype from the mother.

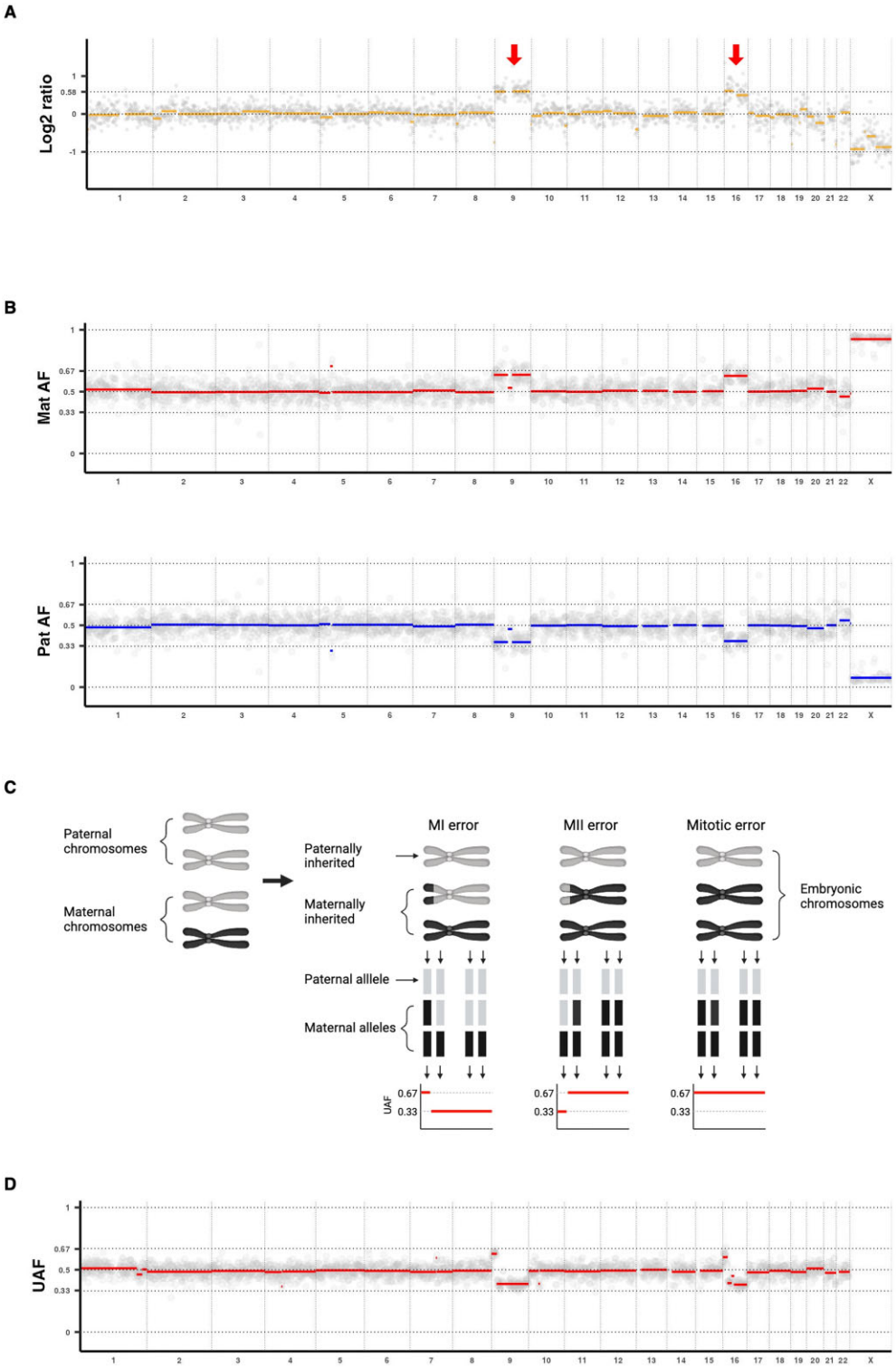


Figure 6. PGT-A analysis using lrWGS data identified aneuploidies and their mechanistic origins in ONT2-E20. (**A**) The copy number plot shows trisomy on chromosomes 9 and 16. (**B**) Maternal allele fractions (Mat AF) and paternal allele fractions (Pat AF) across the genome for SNV loci where the parents exhibited differing homozygous genotypes [homozygous reference (0/0) for one parent and homozygous alternate (1/1) for the other], reflecting parental genome contributions across the embryonic genome. In the trisomic regions, Mat AF = 0.67 and Pat AF = 0.33, indicating one additional chromosome from the maternal side for chromosomes 9 and 16. (**C**) A schematic illustrates the reasoning used to infer the mechanistic origin of a maternally derived trisomy. Loci with homozygous genotypes in the father and heterozygous genotypes in both the mother and the embryo are selected. For each shown heterozygous locus of the embryo, gray bar(s) represent the allele matching the homozygous paternal allele, while black bar(s) represent the unique allele from the heterozygous maternal locus that differs from the homozygous paternal allele. The UAF—the fraction of this unique allele in the embryo—is inferred from the AF value in variant calling results, which is expected to be 0.33 if two different maternal alleles were inherited and 0.67 if two same unique maternal alleles were inherited. The pattern of UAF values across the trisomic chromosome indicates the mechanistic origin of the trisomy. MI: meiotic I; MII: meiotic II. (**D**) UAF values across the genome. For trisomic chromosomes 9 and 16, the UAF is 0.67 at the beginning and decreases to 0.33 across the remaining chromosomal regions. This pattern indicates that the origin of these two trisomies is maternal meiotic I nondisjunction.

trisomy (Fig. 6C). Both chromosomes 9 and 16 were inferred to originate from maternal meiotic I nondisjunction (Fig. 6D), consistent with SNP array results (Supplementary Fig. S9E). To explore the minimum depth of coverage required for accurate aneuploidy detection from TE biopsies, we downsampled BAM files for ONT1-E02 and ONT2-E20 to depth of 1×, 0.75×, 0.5×, 0.25×, 0.1×, 0.05×, and 0.01×. Aneuploidies were clearly detectable at 0.25× and remained detectable at depths as low as 0.1× and 0.05×, though the plots became increasingly noisy. At 0.01×, the aneuploidy signal was no longer discernible (Supplementary Fig. S11). These findings are consistent with those of Liu *et al.* [37], who reported that 0.25×–0.75× ONT data are sufficient for PGT-A.

In summary, we confirmed in human embryos that lrWGS can be used for concurrent PGT-M and PGT-A analysis. For PGT-M, while direct variant detection missed an inherited indel in one of the embryos, inclusion of haplotype-based analysis mitigated this drawback, resulting in 100% concordance with SNP array-based PGT results. For PGT-A, lrWGS data enables not only the detection of aneuploidies but also their parental and mechanistic origins.

Discussion

Here, we explored the characteristics of lrWGS data from single cells and evaluated its performance for variant calling and reference-free haplotyping. Our benchmark analysis revealed that with lrWGS data from TE biopsies (10 cells), direct mutation analysis has a 94% probability of identifying an inherited SNV. For haplotype-based analysis, the probabilities are 98% for TE biopsies (10 cells) and 92% for single blastomere biopsies (1 cell). Considering that TE biopsy is becoming the golden standard, lrWGS-based PGT can thus enable direct variant detection coupled with haplotype-based analysis for increased diagnostic accuracy. Using human embryos, we validated the high performance of lrWGS-based PGT-M and concurrent PGT-A, with the mechanistic origins of aneuploidies correctly identified.

Correctly matching the haplotypes of the offspring to those of their parents is essential for effective haplotype linkage analysis. Key factors contributing to successful matching include: (i) Attainment of high-quality variant calling and phasing outcomes for the parents, which played a pivotal role in identifying high-risk and low-risk haplotypes. (ii) Trio variant calling for the offspring predicted significantly fewer Mendelian inheritance violation loci. (iii) Pedigree-based phasing for the offspring generated chromosome-spanning haplotypes with increased phasing accuracy by combining read based phasing and genetic phasing. (iv) The increased density of informative loci in our study, primarily due to loci that could only be phased using long reads, enabled successful haplotype matching over relatively short genome regions. (v) Exclusion of loci violating Mendelian rules from haplotype comparison. (vi) Flexibility in the haplotype matching process, allowing for discrepancies between the offspring's haplotype and the inferred parental haplotype.

Compared to targeted long-read sequencing-based PGT-M, which requires family- and disease-specific workups and is limited by short amplicons with few informative SNVs [13, 14], lrWGS-based method enables genome-wide haplotype-based PGT-M without these constraints. Additionally, lrWGS-based PGT offers several advantages over short-read WGS-based PGT. For example, while short-read WGS-based haplarithmisis enables all forms of PGT in a single assay, it requires additional family members for phasing [38], which often causes delays, increases costs, and may not always be available. In contrast, lrWGS-based approach phases both parental and embryonic genomes without requiring additional relatives. Both approaches allow direct variant detection in embryos, but long-read sequencing excels in resolving challenging regions, such as repetitive sequences, which short-read sequencing struggles with [39]. Moreover, unlike short-read WGS-based PGT methods that use mutant embryos for phasing [40, 41], lrWGS-based method requires only parental genomic DNA, which is widely accessible and involves less effort for sample preparation. Another advantages of lrWGS-based PGT is its ability to directly phase DNMs present in prospective parents, eliminating the need for additional steps such as analyzing single sperm, polar bodies [42–44] or affected sibling embryos [40]. Additionally, lrWGS-based PGT allows for the detection of both the parental origin and the mitotic or meiotic nature of chromosomal anomalies, providing valuable insights into the etiology of aneuploidies. This feature is also enabled by SNP array-based APCAD [45] and short-read lrWGS-based haplarithmisis [38]. Such information is crucial in clinical practice, as aneuploidies resulting from meiotic chromosome segregation errors rarely survive to term and often lead to adverse pregnancy outcomes; thus, selecting against these embryos could improve IVF success rates [46]. The potential applications of lrWGS-based single-cell haplotyping and aneuploidy profiling go beyond human PGT and can be adapted for other species, such as equine and bovine, to improve reproductive outcomes. It also holds promise for cell-based noninvasive prenatal diagnosis by analyzing single fetal cells present in maternal blood [47].

DNMs arise during various biological and pathological processes, such as cell division and cancer development. Additionally, DNMs are a major cause of rare human disorders. Genome-wide DNM screening would be valuable for identifying these mutations. Using multi- and single-cell lrWGS data of the offspring, we identified 90% and 37% of DNMs in benchmark data, respectively. However, true DNMs represented only 0.5% and 0.1% of all identified DNM candidates, highlighting the abundance of false positives. These results demonstrate that whole-genome DNM screening with ONT lrWGS data remains a challenging task at present However, with increasing sequencing accuracy and methodological improvements, this is likely to be possible in the future. A previous pilot study attempted to use variant annotation databases and functional prediction algorithms to identify real pathogenic DNMs among numerous DNM candidates [48]. Such strategies and additional quality metrics could be integrated into DNM screening to enhance detection accuracy. It is worth noting that we used cell line samples for DNM detection. Since DNMs arise during each cell division and increase with each passage of culturing (Londin *et al.*, 2011), the culture process may have influenced the observed high incidence of DNMs.

To explore the impact of advancements in sequencing technology on phasing, we compared our study with a previous study by Sakamoto *et al.* [49], which phased human cancer genomes by first calling high-quality SNVs using NGS data, followed by phasing with nanopore lrWGS data. In their study, DNA was processed using the SQK-LSK109 and SQK-LSK110 kits and sequenced on R9.4.1 flow cells. With a read length N50 of 20 092 bp at 35× coverage, they achieved a

phased block N50 of 1336 363 bp [49]. In contrast, our study used nanopore lrWGS data for both SNV calling and phasing, with DNA processed using the SQK-LSK114 kit and sequenced on R10.4.1 flow cells. Despite a shorter read length N50 of 17 860 bp and lower coverage of 24× for our HG002 bulk sample, we achieved a phased block NG50 of 2964 293 bp. The corresponding N50 is expected to be longer, as not all genomic regions are fully phased. The nearly doubled phased block N50 in our study, compared to that of Sakamoto *et al.*, despite lower read length N50 and coverage, highlights the advancements in sequencing technology for improved phasing.

We observed a read length N50 ranging from ∼5000 to 25 000 across different bulk samples. Similarly, Sakamoto *et al.* [49] reported significant variation in ONT read length N50, ranging from ∼5000 to 30 000 for various bulk samples, and found that longer read lengths led to an overall increase in phased block length. These findings highlight that read length is a variable factor influencing the overall performance of ONT-based phasing. For single-cell and multi-cell data, the amplification step resulted in smaller read length N50 values, ranging from ∼2 kb to 8 kb. Additionally, the splitting of chimeric reads during mapping further reduced the effective segment length, halving the N50 compared to the original chimeric reads. Compared to bulk samples, the shorter reads from single-cell and multi-cell samples resulted in smaller genome regions being phased into shorter blocks. Despite negatively impacting phasing performance, the alignments generated by chimeric reads also complicated SV calling. These observations emphasize the distinct challenges in phasing and SV calling when using ONT data from amplified DNA, underscoring the need for further exploration.

A constant recombination rate of 1.26 cM/Mb was used during phasing, which is a suitable assumption for the human genome. However, the inherent errors in lrWGS data, especially those derived from single-cell and multi-cell samples, may hinder the accurate detection of actual recombination sites. Consequently, unidentified recombination events may affect haplotype comparison. Moreover, imperfect phasing results contain switch errors that may influence haplotype matching. To mitigate the impact of recombination events and switch errors on the inference of inherited parental haplotypes, we manually constrained the maximum comparison block length to 1 Mb in this study. Additionally, visual inspection of haplotype blocks can help identify recombination events and switch errors, further reducing the risk of misdiagnosis. With ongoing improvements in sequencing accuracy, read length, and bioinformatics algorithms, this constraint on block length may eventually become unnecessary, making accurate phasing of the entire genome feasible.

This study has limitations and areas for potential improvement. First, we utilized ONT sequencing, a cost-effective long-read sequencing technology that is rapidly advancing in read length and accuracy. These improvements will enhance variation discovery and phasing performance, making it essential to conduct updated benchmark studies regularly. Second, due to the limited number of benchmark samples and human PGT samples included in this study, further validation with larger cohorts is necessary to evaluate the practical utility of lrWGS-based haplotyping and aneuploidy profiling for PGT before its adoption in clinical diagnostic settings. Third, the bioinformatics analysis of lrWGS-based PGT, utilizing the pipeline established in this study, can be completed within a few days and is relatively straightforward. However, sequencing cost remains a key consideration. Achieving 20–30× coverage for lrWGS, as demonstrated in this study, currently costs ∼850–1000 EUR per sample, which is higher than Illumina short-read sequencing at similar coverage (∼530–750 EUR per sample). Nevertheless, the total cost is comparable because lrWGS-based PGT does not require a phasing reference. We anticipate that the cost of ONT sequencing will continue to decrease over time, as has been observed with other sequencing technologies, enabling the broader application of lrWGS-based PGT. Forth, we observed an MIC rate averaging 4.68% in our validation study with human PGT families, which is higher than the 2.42% observed at lower coverage (10×) in a recent study using short-read whole-genome sequencing-based comprehensive PGT [38]. This difference underscores the still lower read accuracy of ONT reads compared to Illumina short reads, even with the latest flow cell. However, continued advancements in ONT technology are expected to further enhance accuracy and improve the overall performance of lrWGS-based PGT.

To summarize, we developed a bioinformatics pipeline that enables genome-wide concurrent haplotyping and aneuploidy profiling of single cells using lrWGS data, and validated its effectiveness for genome-wide, reference-free comprehensive PGT. Additionally, we evaluated the performance of lrWGS data from single cells for direct variant detection and DNM screening. Beyond PGT for human embryos, our bioinformatics pipeline has potential applications in other areas of single-cell genomics. For instance, it can be adapted for PGT in animal species like bovine and equine to improve reproductive outcomes, and for cell-based noninvasive prenatal testing by analyzing single circulating trophoblast cells in maternal blood [47].

Acknowledgements

The authors would like to thank the couples who participated in the study.

Author contributions: Conceptualization: J.R.V.; Formal Analysis: Y.Z.; Funding Acquisition: J.R.V.; Investigation: G.P., T.J., O.T., A.V., K.P., H.V.E., E.D., S.D., and M.G.; Methodology: Y.Z., J.R.V., and O.T; Supervision: J.R.V.; Visualization: Y.Z.; Writing—original draft: Y.Z.; Writing and Editing: Y.Z., J.R.V., E.S., O.T., and T.J.

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

Funding was received from the Marie Skłodowska-Curie grant agreement No 813707 (MATER) and from the KU Leuven, C1-C14/22/125 to J.R.V. Y.Z. was partially supported by the Marie Skłodowska-Curie grant agreement No 813707 (MATER).

Data availability

Raw data has been deposited at the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under accession number EGAD50000000787. It is available to academic users upon request to the Data Access Committee (DAC) of KU Leuven via the corresponding author (JRV). We have provided the bioinformatical scripts via the following link: https://doi.org/10.5281/zenodo.14617878

References

1. Tewhey R, Bansal V, Torkamani A *et al.* The importance of phase information for human genomics. *Nat Rev Genet* 2011;**12**:215–23. https://doi.org/10.1038/nrg2950
2. Logsdon GA, Vollger MR, Eichler EE. Long-read human genome sequencing and its applications. *Nat Rev Genet* 2020;**21**:597–614. https://doi.org/10.1038/s41576-020-0236-x
3. Eid J, Fehr A, Gray J *et al.* Real-time DNA sequencing from single polymerase molecules. *Science* 2009;**323**:133–8. https://doi.org/10.1126/science.1162986.
4. Mikheyev AS, Tin MMY. A first look at the Oxford Nanopore MinION sequencer. *Mol Ecol Resour* 2014;**14**:1097–102. https://doi.org/10.1111/1755-0998.12324
5. Wenger AM, Peluso P, Rowell WJ *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol* 2019;**37**:1155–62. https://doi.org/10.1038/s41587-019-0217-9
6. Nurk S, Koren S, Rhie A *et al.* The complete sequence of a human genome. *Science* 2022;**376**:44–53.
7. Zhang CZ, Adalsteinsson VA, Francis J *et al.* Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat Commun* 2015;**6**:6822.
8. Natesan SA, Bladon AJ, Coskun S *et al.* Genome-wide karyomapping accurately identifies the inheritance of single-gene defects in human preimplantation embryos in vitro. *Genet Med* 2014;**16**:838–45. https://doi.org/10.1038/gim.2014.45
9. Zamani Esteki M, Dimitriadou E, Mateiu L *et al.* Concurrent whole-genome haplotyping and copy-number profiling of single cells. *Am Hum Genet* 2015;**96**:894–912. https://doi.org/10.1016/j.ajhg.2015.04.011
10. Esteki MZ, Melotte C, Coonen E *et al.* Agilent Technologies OnePGT solution: external verification on both blastomere and trophectoderm biopsies. *Reprod Biomed Online* 2019;**38**:e11–2. https://doi.org/10.1016/j.rbmo.2019.03.022
11. Backenroth D, Zahdeh F, Kling Y *et al.* Haploseek: a 24-hour all-in-one method for preimplantation genetic diagnosis (PGD) of monogenic disease and aneuploidy. *Genet Med* 2019;**21**:1390–9. https://doi.org/10.1038/s41436-018-0351-7
12. Masset H, Ding J, Dimitriadou E *et al.* Single-cell genome-wide concurrent haplotyping and copy-number profiling through genotyping-by-sequencing. *Nucleic Acids Res* 2022;**50**:e63. https://doi.org/10.1093/nar/gkac134
13. Wu H, Chen D, Zhao Q *et al.* Long-read sequencing on the SMRT platform enables efficient haplotype linkage analysis in preimplantation genetic testing for β-thalassemia. *J Assist Reprod Genet* 2022;**39**:739–46. https://doi.org/10.1007/s10815-022-02415-1
14. Tsuiko O, El Ayeb Y, Jatsenko T *et al.* Preclinical workup using long-read amplicon sequencing provides families with *de novo* pathogenic variants access to universal preimplantation genetic testing. *Hum Reprod* 2023;**38**:511–9. https://doi.org/10.1093/humrep/deac273
15. Zhang P, Zhao X, Li Q *et al.* Proband-independent haplotyping based on NGS-based long-read sequencing for detecting pathogenic variant carrier status in preimplantation genetic testing for monogenic diseases. *Front Mol Biosci* 2024;**11**:1329580. https://doi.org/10.3389/fmolb.2024.1329580
16. Hård J, Mold JE, Eisfeldt J *et al.* Long-read whole-genome analysis of human single cells. *Nat Commun* 2023;**14**:5164. https://doi.org/10.1038/s41467-023-40898-3
17. Hassold T, Hunt P. To err (meiotically) is human: the genesis of human aneuploidy. *Nat Rev Genet* 2001;**2**:280–91. https://doi.org/10.1038/35066065
18. Vanneste E, Voet T, Le Caignec C *et al.* Chromosome instability is common in human cleavage-stage embryos. *Nat Med* 2009;**15**:577–83. https://doi.org/10.1038/nm.1924
19. Bhatt SJ, Marchetto NM, Roy J *et al.* Pregnancy outcomes following *in vitro* fertilization frozen embryo transfer (IVF-FET) with or without preimplantation genetic testing for aneuploidy (PGT-A) in women with recurrent pregnancy loss (RPL): a SART-CORS study. *Hum Reprod* 2021;**36**:2339–44. https://doi.org/10.1093/humrep/deab117
20. Madjunkova S, Sundaravadanam Y, Antes R *et al.* Detection of structural rearrangements in embryos. *N Engl J Med* 2020;**382**:2472–4. https://doi.org/10.1056/NEJMc1913370
21. Tan VJ, Liu T, Arifin Z *et al.* Third-generation single-molecule sequencing for preimplantation genetic testing of aneuploidy and segmental imbalances. *Clin Chem* 2023;**69**:881–9. https://doi.org/10.1093/clinchem/hvad062
22. Li H, Handsaker B, Wysoker A *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9. https://doi.org/10.1093/bioinformatics/btp352
23. De Coster W, Rademakers R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* 2023;**39**:btad311. https://doi.org/10.1093/bioinformatics/btad311
24. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100. https://doi.org/10.1093/bioinformatics/bty191
25. Zheng Z, Li S, Su J *et al.* Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nat Comput Sci* 2022;**2**:797–803. https://doi.org/10.1038/s43588-022-00387-x
26. Su J, Zheng Z, Ahmed SS *et al.* Clair3-trio: high-performance Nanopore long-read variant calling in family trios with trio-to-trio deep neural networks. *Brief Bioinform* 2022;**23**:bbac301. https://doi.org/10.1093/bib/bbac301
27. Cleary JG, Braithwaite R, Gaastra K *et al.* Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines. bioRxiv, https://doi.org/10.1101/023754, 3 August 2015, preprint: not peer reviewed.
28. Smolka M, Paulin LF, Grochowski CM *et al.* Detection of mosaic and population-level structural variants with Sniffles2. *Nat Biotechnol* 2024;**42**:1571–80.
29. English AC, Menon VK, Gibbs RA *et al.* Truvari: refined structural variant comparison preserves allelic diversity. *Genome Biol* 2022;**23**:271. https://doi.org/10.1186/s13059-022-02840-6
30. Martin M, Patterson M, Garg S *et al.* WhatsHap: fast and accurate read-based phasing. bioRxiv, https://doi.org/10.1101/085050, 14 November 2016, preprint: not peer reviewed.
31. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2. https://doi.org/10.1093/bioinformatics/btq033
32. Magi A, Bolognini D, Bartalucci N *et al.* Nano-GLADIATOR: real-time detection of copy number alterations from nanopore sequencing data. *Bioinformatics* 2019;**35**:4213–21. https://doi.org/10.1093/bioinformatics/btz241
33. Veltman JA, Brunner HG. De novo mutations in human genetic disease. *Nat Rev Genet* 2012;**13**:565–75. https://doi.org/10.1038/nrg3241
34. Jónsson H, Sulem P, Kehr B *et al.* Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland. *Nature* 2017;**549**:519–22. https://doi.org/10.1038/nature24018
35. Goldmann JM, Wong WSW, Pinelli M *et al.* Parent-of-origin-specific signatures of *de novo* mutations. *Nat Genet* 2016;**48**:935–9. https://doi.org/10.1038/ng.3597

36. Sasani TA, Pedersen BS, Gao Z *et al.* Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *eLife* 2019;**8**:e46922. https://doi.org/10.7554/eLife.46922
37. Liu S, Wang H, Leigh D *et al.* Third-generation sequencing: any future opportunities for PGT? *J Assist Reprod Genet* 2021;**38**:357–64. https://doi.org/10.1007/s10815-020-02009-9
38. Janssen AEJ, Koeck RM, Essers R *et al.* Clinical-grade whole genome sequencing-based haplarithmisis enables all forms of preimplantation genetic testing. *Nat Commun* 2024;**15**:7164. https://doi.org/10.1038/s41467-024-51508-1
39. Kosugi S, Terao C. Comparative evaluation of SNVs, indels, and structural variations detected with short- and long-read sequencing data. *Hum Genome Var* 2024;**11**:18. https://doi.org/10.1038/s41439-024-00276-x
40. Yuan P, Xia J, Ou S *et al.* A whole-genome sequencing–based novel preimplantation genetic testing method for *de novo* mutations combined with chromosomal balanced translocations. *J Assist Reprod Genet* 2020;**37**:2525–33. https://doi.org/10.1007/s10815-020-01921-4
41. Chen S, Yin X, Zhang S *et al.* Comprehensive preimplantation genetic testing by massively parallel sequencing. *Hum Reprod* 2021;**36**:236–47.
42. Altarescu G, Eldar-Geva T, Varshower I *et al.* Real-time reverse linkage using polar body analysis for preimplantation genetic diagnosis in female carriers of *de novo* mutations. *Hum Reprod* 2009;**24**:3225–9. https://doi.org/10.1093/humrep/dep293
43. Rechitsky S, Pomerantseva E, Pakhalchuk T *et al.* First systematic experience of preimplantation genetic diagnosis for de-novo mutations. *Reprod Biomed Online* 2011;**22**:350–61. https://doi.org/10.1016/j.rbmo.2011.01.005
44. Crugnola E, Gobbetti A, Fiandanese N *et al.* P–562 PGT-M with *de novo* mutations: how to deal with it? *Hum Reprod* 2021;**36**:i393. https://doi.org/10.1093/humrep/deab130.561
45. Verdyck P, Berckmoes V, Fernandez Gallardo E *et al.* APCAD part 2: a novel method for detection of meiotic aneuploidy in preimplantation embryos. *Genes* 2025;**16**:115. https://doi.org/10.3390/genes16020115
46. Dimitriadou E, Melotte C, Debrock S *et al.* Principles guiding embryo selection following genome-wide haplotyping of preimplantation embryos. *Hum Reprod* 2017;**32**:687–97. https://doi.org/10.1093/humrep/dex011
47. Vossaert L, Wang Q, Salman R *et al.* Validation studies for single circulating trophoblast genetic testing as a form of noninvasive prenatal diagnosis. *Am Hum Genet* 2019;**105**:1262–73. https://doi.org/10.1016/j.ajhg.2019.11.004
48. Murphy NM, Samarasekera TS, Macaskill L *et al.* Genome sequencing of human in vitro fertilisation embryos for pathogenic variation screening. *Sci Rep* 2020;**10**:3795. https://doi.org/10.1038/s41598-020-60704-0
49. Sakamoto Y, Miyake S, Oka M *et al.* Phasing analysis of lung cancer genomes using a long read sequencer. *Nat Commun* 2022;**13**:3464. https://doi.org/10.1038/s41467-022-31133-6

Received: September 23, 2024. **Revised:** March 13, 2025. **Editorial Decision:** March 14, 2025. **Accepted:** March 30, 2025