BMC Genomics

RESEARCH

Open Access

Precise exome analysis of blastocyst biopsy scale samples using primary template-directed amplification

Alina Samitova[1*], Vera Belova[1], Iuliia Vasiliadis[1], Zhanna Repinskaia[1], Tatiana Gorodnicheva[1], Evgeny Romanov[2], Mariam Pogosyan[2], Emil Gaysin[2], Tatyana Nazarenko[2], Denis Rebrikov[1,2] and Dmitriy Korostin[1]

Abstract

This study evaluates primary template-directed amplification (PTA) for whole exome sequencing (WES) of small fibroblast cell groups, which mimics the limited cell quantities typical of trophectoderm embryo biopsies. PTA's consistent amplification reduces allelic dropout (ADO) and improves uniform coverage, overcoming challenges associated with conventional methods such as multiple displacement amplification (MDA). Using fibroblast samples alongside well-characterized genomic references (E701, NA12878), we benchmarked PTA-WES, achieving 97.5% target region coverage at 10x, meeting American College of Medical Genetics and Genomics (ACMG) standards. The completed filtering and variant calling provide a foundation for further optimization and analysis aimed at evaluating the reliability of PTA for routine clinical use. Preliminary results from embryo biopsies sequenced with PTA-WES revealed a median coverage of 102x, significantly improving upon the variability and coverage gaps observed with MDA-WES. These findings support the potential of PTA to increase the clinical applicability of WES for preimplantation genetic testing for monogenic disorders (PGT-M), expanding its ability to detect inherited and *de novo* mutations in embryos.

Keywords PTA, Exome sequencing, Preimplantation genetic testing, WGA

Background

Preimplantation genetic testing (PGT) is a crucial tool in assisted reproductive technologies, allowing for the selection of embryos free from genetic abnormalities before implantation. Previously, many analytical approaches were not feasible for trophectoderm biopsies due to the extremely limited amount of DNA. In the absence of whole genome amplification (WGA), molecular analyses were restricted to targeted methods, such as fluorescence in situ hybridization (FISH) and PCR assays. In contrast, whole genome and whole exome sequencing, as well as reliable detection of *de novo* variants, were technically unattainable. However, with improvements in WGA protocols, it has become possible to obtain sufficient starting material from as little as 6–7 picograms of DNA from a single cell, overcoming the limitations of existing methods and expanding the potential for embryo genome analysis [1–5]. Among current techniques, multiple displacement amplification (MDA) and Multiple annealing and looping-based amplification cycles (MALBAC) have shown significant promise, particularly in PGT-A (testing

*Correspondence:
Alina Samitova
alinasamitova16@gmail.com
[1]Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Pirogov Russian National Research Medical University, Ostrovityanova str. 1, Moscow 117997, Russia
[2]V.I. Kulakov National Medical Research Center for Obstetrics Gynecology and Perinatology, Ministry of Healthcare of the Russian Federation, Moscow 117997, Russia

for aneuploidies), owing to their ability to amplify sufficient DNA from very small inputs [6–8]. Nevertheless whole-exome sequencing (WES), a powerful tool for detecting a wide range of genetic variants associated with inherited disorders, remains largely unavailable in PGT. Despite its advantages, uneven amplification of DNA by MDA leads to coverage gaps, limiting the effectiveness of WES for preimplantation genetic testing for monogenic disorders (PGT-M) [9–12].

Currently, PGT-M involves the collection of maternal, paternal, and control samples, which are analyzed alongside amplified genomic DNA from trophectoderm biopsies to determine the mutation status of each embryo, via methods such as PCR with NGS or Sanger sequencing, karyomapping, and haplarithmisis [12–14]. Several studies highlight the critical role of PGT-M in the genetic analysis of embryos. For example, one study specifically highlighted the use of PGT-M to prevent the transmission of genetic conditions such as Marfan syndrome through PGT, where a PGT-M protocol was developed and tested via multiplex fluorescent PCR and mini-sequencing [15]. There is also a known case of PGT for Meckel syndrome, where the WGA products of each embryo were subjected to Sanger sequencing for direct identification of variant sites, and haplotyping analysis was conducted via SNP markers [16]. Although these approaches are effective for detecting target variants, they do not provide the same resolution as WES and do not allow for reliable identification of *de novo* variants. Examining preimplantation embryos for such mutations is increasingly recognized as an important, albeit technically challenging, task [17]. Several studies have already reported data on the sequencing of whole genomes and exomes of single cells, providing a foundation for these analyses [18, 19]. However, limitations in current WGA techniques such as allele drop-out (ADO), locus drop-out (LDO), chimeric DNA molecules, base replication errors and unevenness in amplification often prevent these techniques from achieving the high coverage required by the standard of the American College of Medical Genetics and Genomics (ACMG) for accurate variant detection [20]. It is also necessary to conduct not only mutation site detection but also SNP linkage analysis to ensure accuracy. Therefore, an improved ability to sequence high-quality exomes directly from trophectoderm biopsies could significantly increase the speed and accessibility of PGT-M, providing a broader range of genetic insights.

A potential solution to this challenge is the introduction of primary template-directed amplification (PTA), a novel method that offers more uniform WGA from small amounts of material, including preimplantation biopsy samples. PTA more evenly amplifies both alleles in the same cell, resulting in significantly diminished ADO and skewing. While PTA offers distinct advantages over other methods, it can also lead to specific artifacts due to the amplification process [21–23]. Therefore, it is essential to consider these artifacts when interpreting the results. Recent research has led to the development of programs such as SCAN2 and PTATO with the objective of minimizing the number of false positives in PTA data analysis [24, 25]. These programs filter the artifact variants both with machine learning algorithms and with Variant Call Format (vcf) thresholds outside the usage of models.

In this study, we utilized fibroblast cultures from the well-characterized in-lab reference genome E701, genomic DNA from the E701 sample, and the Platinum Genome DNA sample NA12878 [26, 27]. Additionally, we developed an artifact-filtering approach tailored to our specific needs. This method leverages variant-calling results from WES and whole-genome sequencing (WGS) data of the well-characterized reference genome sample E701, along with four PTA-WES samples. This research aims to assess the applicability of PTA for expanding the capabilities of PGT, making WES a feasible and reliable option in clinical settings.

Results

PTA-WGA of genomic DNA and DNA from fibroblast cells

To assess the quality of material obtained after whole genome amplification (WGA) using the PTA method, we used both genomic DNA and fibroblast cells. Fibroblast cells were grouped by count ([4, 10, 16], and [25]) to evaluate whether the method would be effective at low input levels. The 4-cell group was intended to model the typical number of cells obtained from a blastocyst trophectoderm biopsy. The gDNA group included the laboratory standard sample E701 and the NA12878 reference sample, both of which were subjected to PTA at amounts of 64 pg and 256 pg, corresponding to approximately 10 and 40 genome equivalents, respectively. The f-4 group consisted of three replicate fibroblast samples, each containing four cells used for PTA. Similarly, the f-10 group consisted of three replicate fibroblast samples of ten cells each, whereas the f-16 group consisted of three replicate samples of sixteen cells each. The f-25 group consisted of three replicate fibroblast samples of twenty-five cells each for WGA. The PTA-WGA was successful in all tested samples with no observed outliers. On the basis of electrophoresis and concentration measurements all the samples presented similar characteristics. On average, we obtained ~ 1600 ng (median [IQR]: 1618.5 [902.3–2061.8] ng; range 516–3240 ng) of DNA per sample after PTA-WGA and further purification using 2x Kapa Pure Beads (Roche).

Raw sequencing data and pooling balance

We obtained between 190 and 249 million (M) reads per sample across 16 samples (Fig. 1). As shown in the

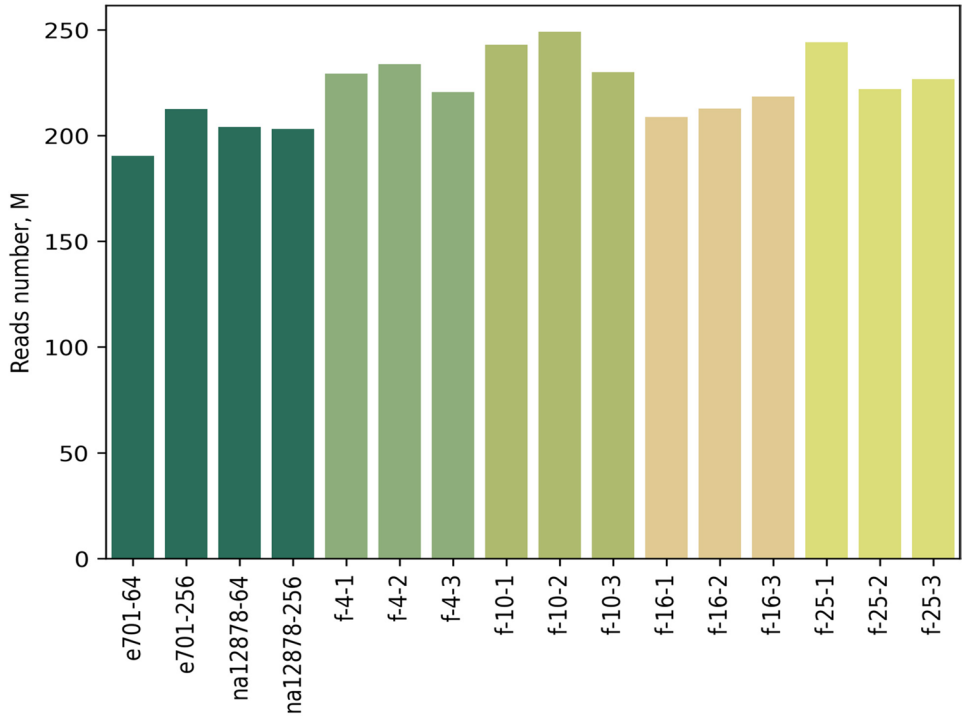

Fig. 1 The barplot shows the number of reads per sample

graph, the amount of data for samples with a low number of cells was comparable to that of samples from other groups, indicating well-balanced pooling. Precapture pooling can be challenging when different sample types (e.g., genomic DNA or FFPE DNA) are used, as variations in initial DNA quality can lead to inconsistent data yields post-enrichment. Therefore, this initial test of pooling different cell groups within a single pool was successful in achieving comparable data output across samples.

Exome enrichment quality after PTA-WGA for fibroblast and genomic DNA

To assess exome sequencing quality after PTA-WGA, we grouped the samples as follows: samples with 4, 10, 16 and 25 fibroblast cells, and genomic DNA (E701 and NA12878). Coverage statistics were calculated via Picard, and metrics were averaged across the sample groups. The results of Picard key metrics for all the samples are presented in Additional file 1: Table 1. The percentages of on-target reads, off-target reads and duplicates for each sample group were calculated (Fig. 2A). The distribution of these metrics did not reveal any particular dependence between groups of samples. However, it should be noted that under practical conditions it is not always possible to obtain ~ 200 M reads per sample. We therefore downsampled each sample to 100 M reads. This process allows us to simulate conditions closer to basic exome data yield for clinical purposes. In this way, we can better estimate sequencing metrics for samples with less data. The graph (Fig. 2B) shows that the use of 100 M reads per sample increased the average on-target percentage by 0.48% compared with the data shown in the graph (Fig. 2A) for samples with 200 million reads. This increase may be due to the reduction in duplicates observed with fewer reads; however, the difference was not statistically significant (Wilcoxon test, $p = 0.06$).

The mean target coverage for all samples was 247x, with a median target coverage of 226x (Fig. 3).

Analysis of the coverage data showed that all samples, regardless of the type of starting material or the number of cells used for PTA, received a comparable amount of sequencing reads, with no noticeable variability within each group. This suggests that the sequencing workflow performed consistently across samples, providing uniform coverage. It should be noted, however, that this observation refers solely to coverage-based metrics and does not account for allele-level properties, which may still vary depending on the input material or WGA performance.

The mean breadth, defined as the percentage of target regions covered at least x times per sample, for all samples in the pool was 97.5% (±0.26%SD) at 10x coverage depth, 96.25% (±0.59%SD) at 20x and 94.84% (±0.90%SD) at 30x coverage depth. These parameters are characterized by a low standard deviation, indicating high homogeneity of the data (Fig. 4).

In this study, we observed an average of 1.28% (± 0.03%SD) regions with zero coverage, which is fully

Fig. 2 Barplots showing the average on-targets, duplicates, and off-targets reads in each group for: (**A**) raw data, and (**B**) down-sampled data

consistent with the exome data typically obtained for samples isolated from blood [28].

ACMG gene coverage analysis

We calculated the depth of coverage for 81 genes recommended by the ACMG for the identification of pathogenic variants in clinical reporting (ACMG SF v3.2) [29]. The average percentage of target regions that covered at least 10x was greater than 98.4%. Further examination of the coverage distribution confirmed that the vast majority of target regions achieved uniform coverage in all samples (Additional file 2: Table 2). Among the 81 clinically important genes analyzed, 49 were fully covered at 100% across all the samples. A further 16 genes achieved average coverage of 95–99% in all the samples. However, coverage issues were observed for 6 genes, with an average coverage of approximately 80%. (Fig. 5). These genes included *BTD*, *MSH2*, *PRKAG2*, *SDHC*, *SDHD*, and *STK11.* Similar coverage patterns are observed in data from standard exome sequencing, which may be attributed to the design of the capture probes used (Agilent SureSelect V8).

These results were obtained without considering the difficult-to-sequence regions reported by Hijikata et al. [30]. However, when these regions were examined separately, it was found that only a limited proportion of the ACMG genes were affected. Of the 81 ACMG genes, three (*TTN*, *PMS2*, and *SDHD*) showed partial overlap with difficult regions. For PMS2 and SDHD, only a single CDS interval intersected these regions, whereas TTN exhibited overlap across 31 CDS segments (8.5% of the gene's 363 CDS intervals). To evaluate the potential

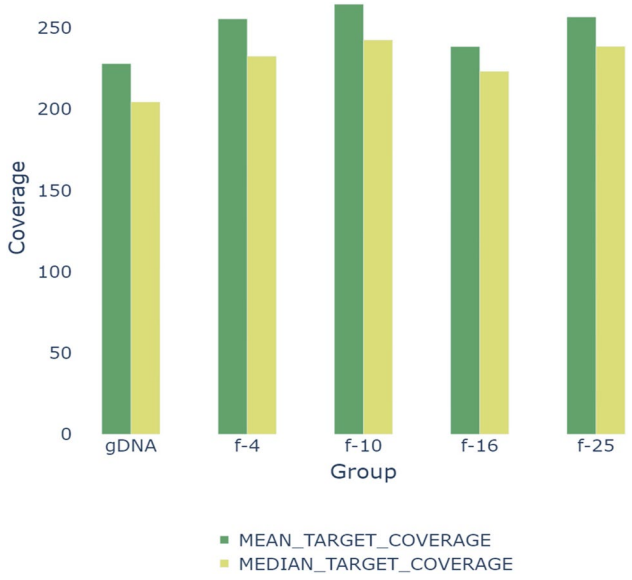

Fig. 3 Average of the mean and median coverage target coverage of the samples that were analyzed without down-sampling

clinical impact of these technically challenging regions, we compared the intersecting CDS intervals with ClinVar (v20240611) entries classified as likely pathogenic or pathogenic (LP/P).A total of 40 LP/P variants were mapped to CDS regions overlapping difficult-to-sequence intervals. This subset represents 1.03% of all LP/P variants reported in the CDS of TTN, PMS2, and SDHD (40/3,901) and only 0.09% of all LP/P variants annotated across CDS of all ACMG SF genes (40/44,005).

Variant filtering criteria

Variant calling results in VCF format were obtained for WES E701 and WGS E701, as well as for the four PTA-WES samples. After filtering by read depth and difficult-to-sequence regions, the WES E701 and WGS E701 variants intersected, resulting in three sets (Fig. 6).

Variant parameter distributions

In the next stage, we focused on variants unique to either WES E701 or WGS E701 (which were absent from their intersection). Such variants are likely false positives, and our goal was to identify their common characteristics that might explain the reasons for their inclusion.

To achieve this, we plotted the variant allele frequency (VAF) and quality (QUAL) distributions for two unique variant sets (WES and WGS) and compared them with the corresponding distributions of common WES/WGS variants.

The analysis of VAF distributions revealed that the majority of common WES/WGS variants presented VAF values in the range of 0.4 to 0.6 (heterozygotes) or equal to 1 (homozygotes) (Fig. 7A and B). In contrast, over 37% of the unique WES and WGS variants had VAF values less than 0.4. As shown in Fig. 7C and D and a significant proportion of common WES/WGS variants were characterized by QUAL values between 50 and 75, whereas more than 79% of the unique WES and WGS variants presented QUAL values less than 50. This may indicate the low confidence of such variants.

On the basis of the observed differences in VAF and QUAL distributions, we aimed to establish threshold values to distinguish unique WES and WGS variants from common WES/WGS variants. To achieve this, we

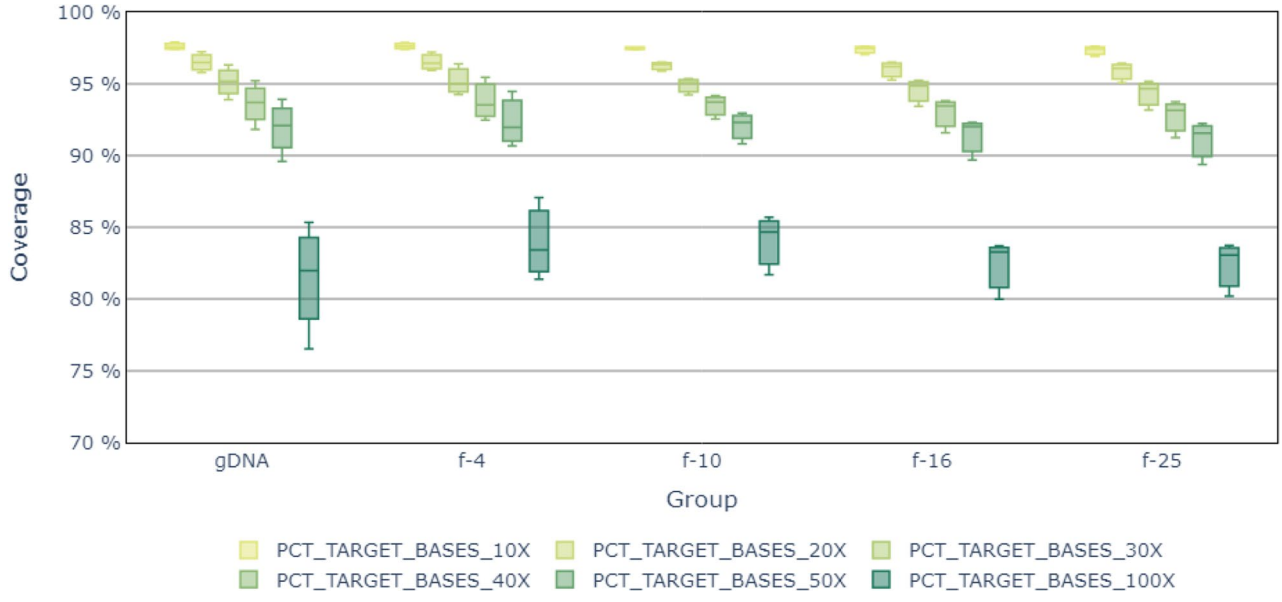

Fig. 4 The average percentage of regions with 10x, 20x, 30x, 40x, 50x and 100x coverage depths for different groups of samples

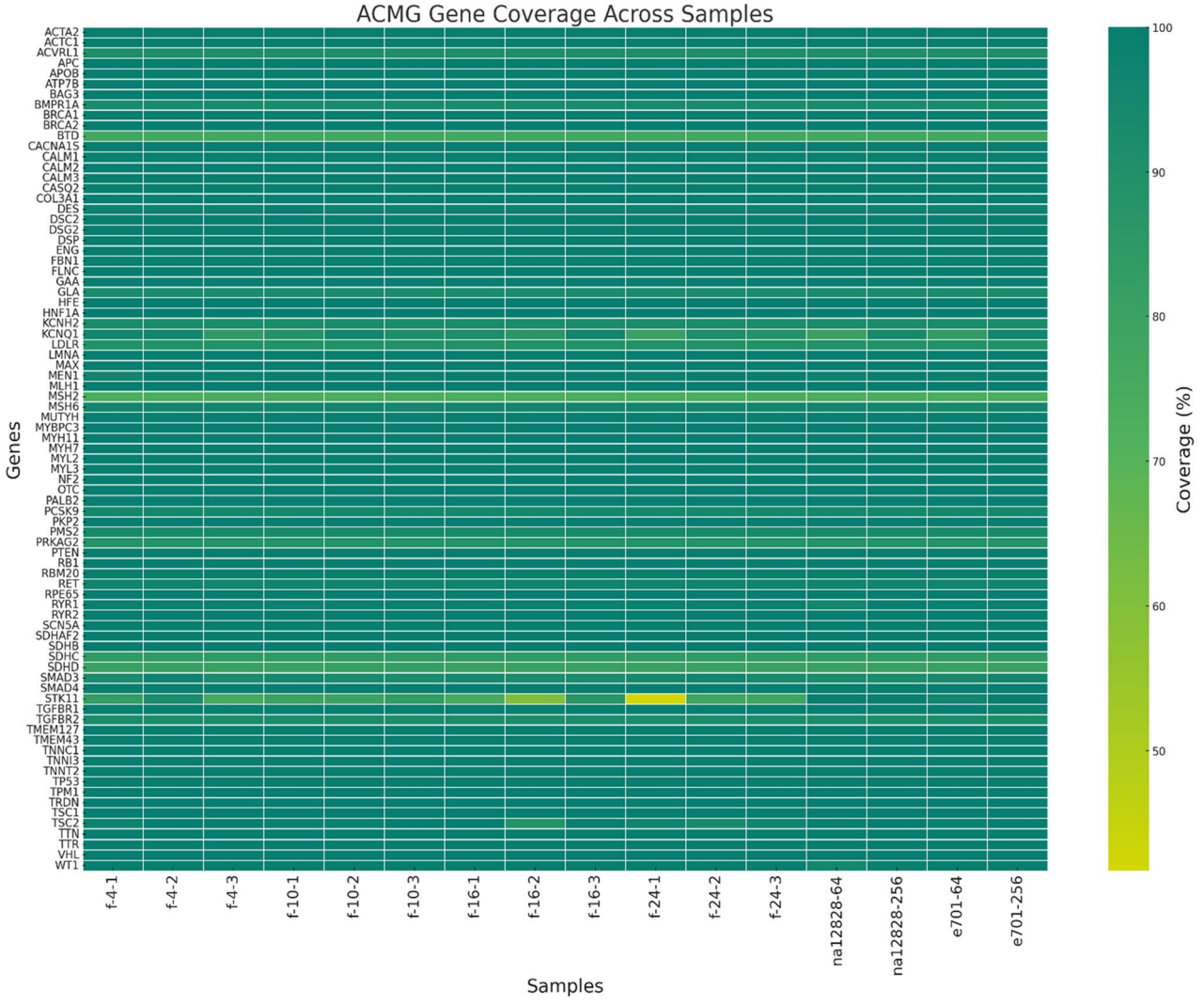

Fig. 5 The percentage of the breadth of coverage with per-site read depth ≥ 10x is shown for each of the 81 genes recommended by the American College of Medical Genetics and Genomics (ACMG) among the samples

calculated the 5th percentile of VAF and QUAL distributions for common WES/WGS variants. The values obtained were as follows: VAF: 0.371429 (WES) and 0.380952 (WGS), QUAL: 30.4 (WES) and 32.8 (WGS). On the basis of these data, we propose the use of thresholds of VAF < 0.37 and QUAL < 30 for variants filtering. We suggest that applying these thresholds in sample processing ensures minimal loss of true positive variants while effectively excluding false positives.

Variant type distributions

The subsequent stage of the analysis involved a pairwise intersection of variants identified in four PTA-WES samples with common WES/WGS variants. Unique PTA-WES variants constituted approximately 5–6% of total variants identified in each sample (Fig. 8, Additional file 3:S1). For unique PTA-WES and common WES/WGS variants, we plotted the distributions of mutation types, expressed as percentages, to assess the relative frequency of each mutation type (Fig. 9A and B, Additional file 3: Fig. S2). Among the common WES/WGS variants, the most prevalent mutations were G-> A (16.53%), C-> T (16.20%), A-> G (14.89%) and T-> C (14.84%) transitions. The proportion of indels was 11.94%. In contrast, unique PTA-WES variants were characterized by a high proportion of indels (> 45%), as well as high frequency of C-> T and G-> A transitions compared with other mutation types.

A comparison of the variant type distributions across the samples revealed notable differences in the overall data structure. Furthermore, these mutation types were integrated into the variant filtering process.

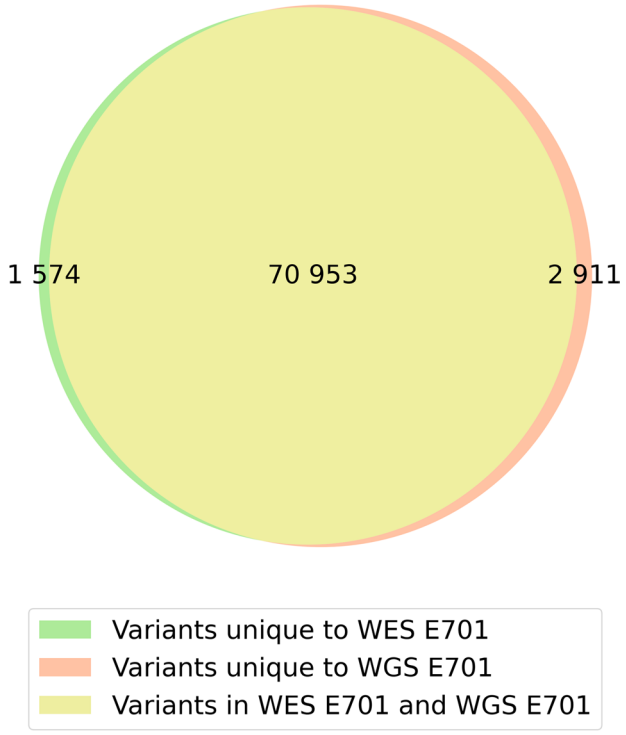

Fig. 6 Venn diagram showing the intersection of the WES E701 and WGS E701 variants

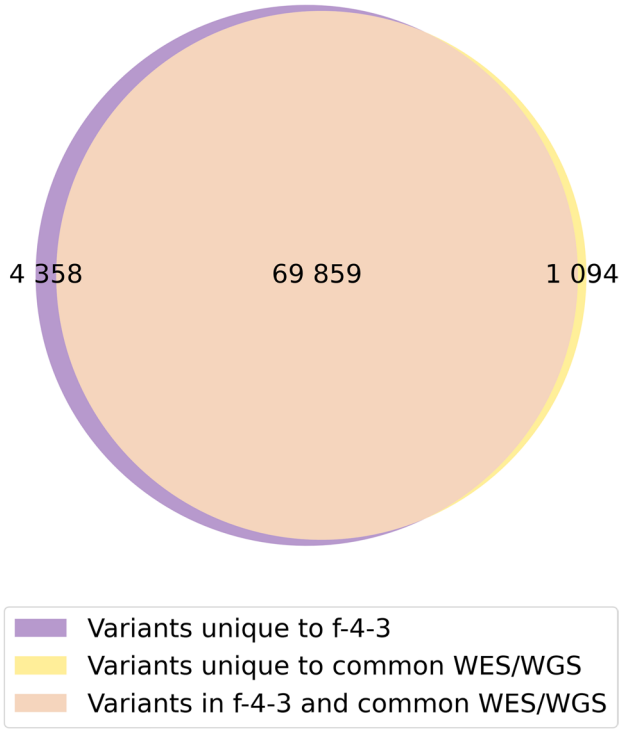

Fig. 8 Venn diagram showing the intersection of PTA-WES and common WES/WGS variants

Indel length distributions

Given the high proportion of indels among the unique PTA-WES variants, our objective was to identify any patterns that can be considered typical for these variants. To achieve this, we plotted the indel length distributions for two variant sets: common WES/WGS and unique PTA-WES. Indel length was defined as the difference between the lengths of the reference and alternative alleles. The analysis revealed that single-nucleotide insertions and deletions were the most prevalent among the common

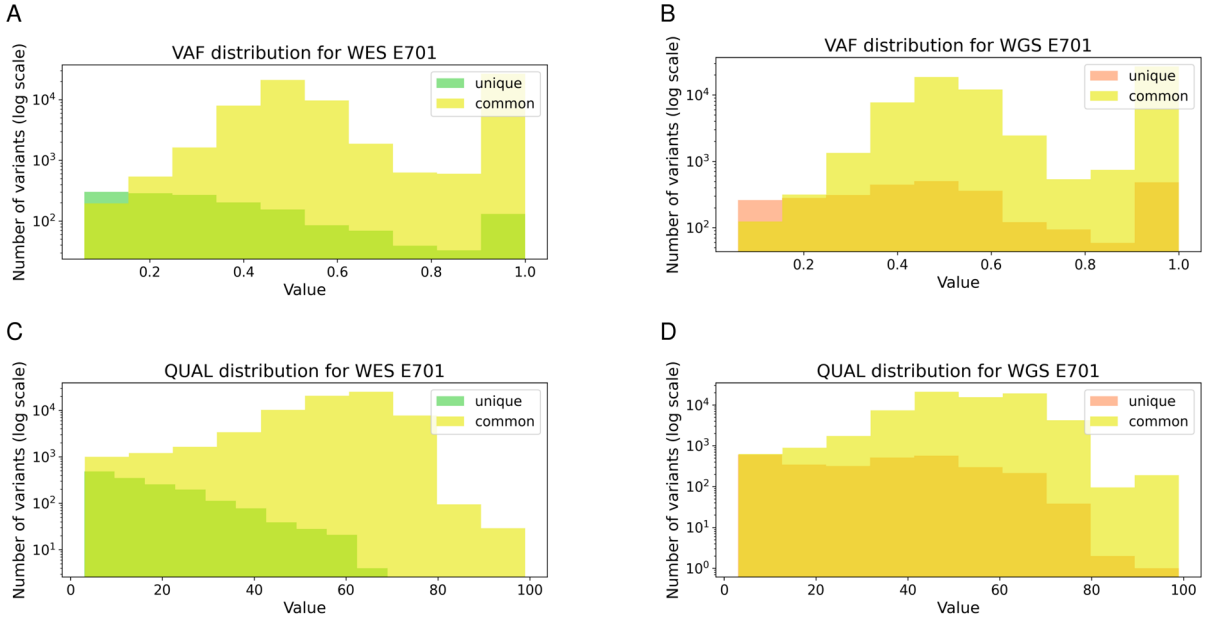

Fig. 7 Histograms of distributions: **A**) VAF for WES E701 variants, **B**) VAF for WGS E701 variants, **C**) QUAL for WES E701 variants, and **D**) QUAL for WGS E701 variants

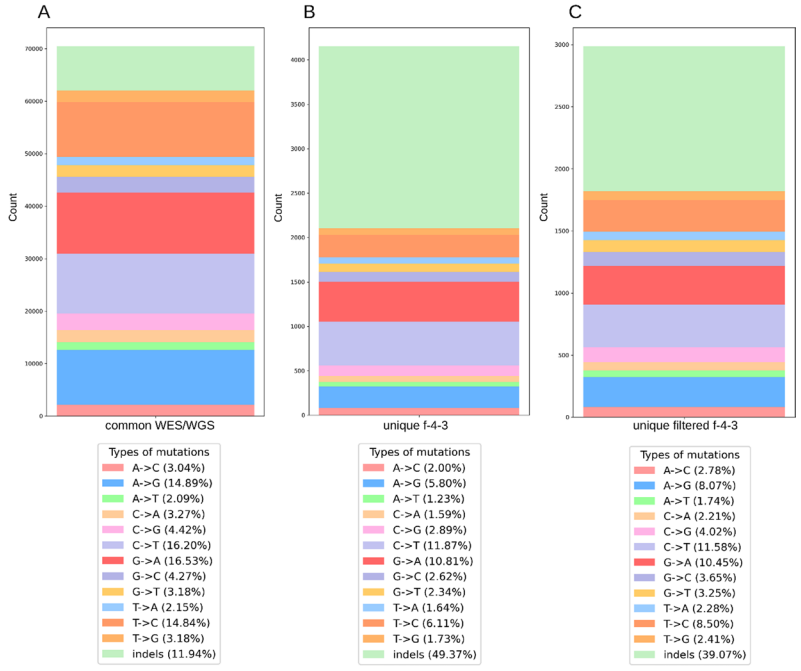

Fig. 9 Distributions of variant types among **A**) common WES/WGS variants, **B**) unique PTA-WES variants, and **C**) unique PTA-WES variants after filtering

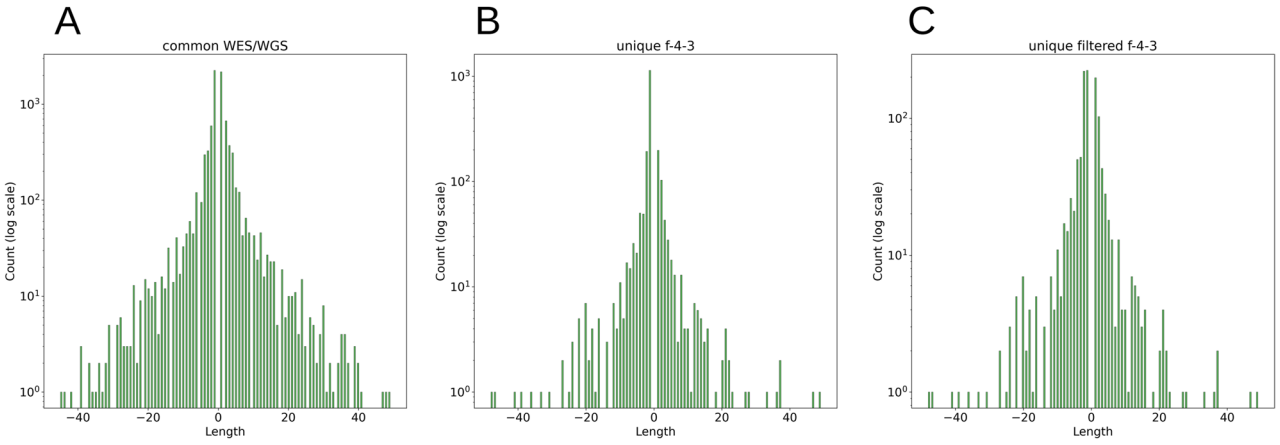

Fig. 10 Indel length distributions among **A**) common WES/WGS variants, **B**) unique PTA-WES variants, and **C**) unique PTA-WES variants after filtering. Indels with a length greater than 50 were excluded

WES/WGS variants. In contrast, single-nucleotide insertions constituted the majority of unique PTA-WES variants (Fig. 10A and B, Additional file 3). The observed differences in indel length distributions suggest that single-nucleotide insertions may significantly contribute to the total number of artifacts observed.

Therefore, on the basis of the results obtained, we propose the following filters to exclude low-quality and potentially artifactual variants: VAF < 0.37 and QUAL < 30, and additionally fulfill at least one of the following conditions (a C-> T, a G-> A, or a 1-bp insertion).

Applying filters to PTA-WES samples

We applied the aforementioned filters to 16 PTA-WES samples (Table 1). The highest proportions of filtered variants were observed in samples e701-64 (3.86%) and na12828-64 (4.12%). For the other samples, the proportion of filtered variants was approximately 2–3% of the total identified variants.

To evaluate the effectiveness of filtering, we intersected filtered PTA-WES variants with common WES/WGS and compared the number of variants not falling into the intersection before and after filtering. On average, this proportion decreased by 1.38% across all samples (Fig. 11, Additional file 3: S4). According to the Wilcoxon test, this reduction is statistically significant ($p = 3.05 \times 10^{-5}$).

Table 1 Results of PTA-WES variants filtering

Sample	Total variants (count)	Filtered variants (count)	Filtered variants (%)
f-4-1	73967	2078	2.81
f-4-2	74427	1775	2.38
f-4-3	74217	1786	2.41
f-10-1	74051	1606	2.17
f-10-2	74021	1845	2.49
f-10-3	74191	1628	2.19
f-16-1	74024	1514	2.05
f-16-2	73804	1587	2.15
f-16-3	74072	1591	2.15
f-24-1	73540	1588	2.16
f-24-2	73785	1452	1.97
f-24-3	73993	1442	1.95
na12828-64	72571	2992	4.12
na12828-256	74252	1993	2.68
e701-64	72513	2802	3.86
e701-256	73842	1895	2.57

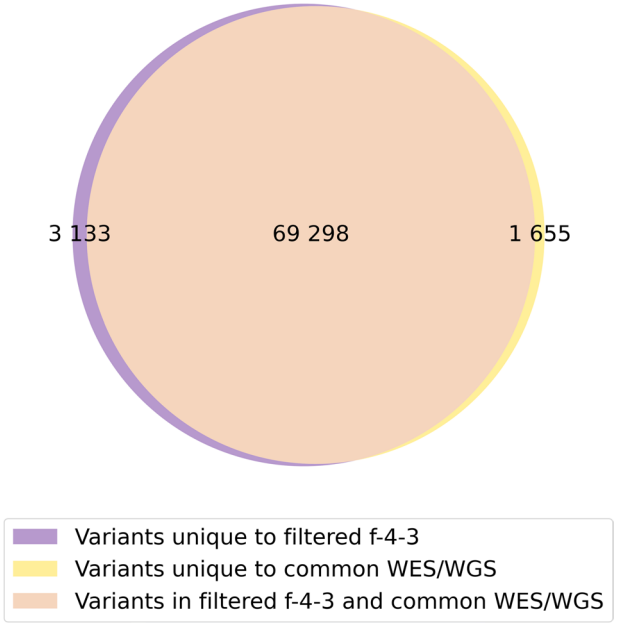

Fig. 11 Venn diagram showing the intersection of filtered PTA-WES and common WES/WGS variants

We further analyzed the distributions of variant types and indel lengths for unique PTA-WES variants after filtering. The results revealed a decrease in the relative proportion of indels among all mutation types and a significant decrease in the histogram peak corresponding to single-nucleotide insertions (Figs. 9C and 10C).

These findings indicate that the proposed filtering thresholds are expected to effectively eliminate low-quality and artifactual variants from the PTA-WES samples, thereby substantially enhancing the accuracy and robustness of the analysis.

Analytical sensitivity and precision of PTA-WES

To evaluate the analytical accuracy of PTA-WES, we benchmarked variant calls obtained from all PTA samples against their corresponding non-PTA reference samples (e701, NA12878). Metrics for SNPs and indels (recall, precision, and F1) were calculated using the haplotype-comparison tool hap.py v0.3.15 (https://github.com/Illumina/hap.py). Full metric results are provided in Additional file 4: Table 3, and a summary is presented in Fig. 12. Across all input groups, SNP calling exhibited consistently high performance, with a mean precision of approximately 0.97 (median = 0.9825 [0.9801–0.9846]), a mean recall of 0.94 (median = 0.9382 [0.9344–0.9405]), and an average F1 score of 0.96 (median = 0.9602 [0.9548–0.9621]). Indel detection showed a mean precision of 0.80 (median = 0.8241 [0.8186–0.8262]) and a mean recall of 0.78 (median = 0.7716 [0.7679–0.7754]), corresponding to an average F1 score of 0.79 (median = 0.7969 [0.7931–0.7996]). On the Fig. 12, six outlier points – mostly corresponding to genomic DNA samples after PTA –are labeled, while the remaining points cluster tightly according to variant type (SNP or indel) (Additional file 4: Table 3) .

Evaluation of carrier variant detection in fibroblast-derived and standard exomes

After applying the defined filtering criteria to remove low-quality and potentially artifactual variants, carrier screening reports were generated for all analyzed samples. Tables 2 and 3 show the clinically relevant variants identified in each sample, including the chromosomal location and the corresponding gene. Any variants that were also detected in the matched control DNA sample and were therefore considered to be expected findings are highlighted in bold.

Fig. 12 Benchmarking analytical performance of PTA-WES in variant detection: precision and recall for SNPs and Indels across all input samples

Table 2 Summary of variants detected in carrier status reports from cell groups

Replicate	4 cells	10 cells	16 cells	25 cells
#1	chr13:20189346 AG/A *GJB2* chr15:41815388 C/T *MAPKBP1*	chr13:20189346 AG/A *GJB2* chr15:41815388 C/T *MAPKBP1* chr16:28902318 G/GC *ATP2A1*	chr13:20189346 AG/A *GJB2* chr15:41815388 C/T *MAPKBP1*	chr13:20189346 AG/A *GJB2*
#2	chr13:20189346 AG/A *GJB2*	chr15:41815388 C/T *MAPKBP1*	chr13:20189346 AG/A *GJB2*	chr13:20189346 AG/A *GJB2*
#3	chr13:20189346 AG/A GJB2 chr15:41815388 C/T *MAPKBP1* chr17:65536466 G/GC *AXIN2*	chr13:20189346 AG/A GJB2 chr15:41815388 C/T *MAPKBP1* chr16:30725049 G/GC *SCAP* chr19:17844302 T/TG *JAK3*	chr13:20189346 AG/A *GJB2*	chr13:20189346 AG/A *GJB2*

Table 3 Summary of variants detected in carrier status reports from primary genomic DNA samples, including both PTA-derived and standard (non-WGA) samples

	e701	NA12878
64 pg after PTA-WES	chr13:20189346 AG/A *GJB2* chr16:28984905 T/TG *LAT*	chr11:6617154 C/T *TPP1* chr12:102840474 T/C *PAH*
256 pg after PTA-WES	chr13:20189346 AG/A *GJB2*	chr11:6617154 C/T TPP1 chr12:102840474 T/C *PAH* chr1:25543762 T/TG *LDLRAP1*
Standard WES	chr13:20189346 AG/A *GJB2*	chr11:6617154 C/T *TPP1* chr12:102840474 T/C *PAH*

To ensure the reliability of the reported findings, all variant loci were manually reviewed using the Integrative Genomics Viewer (IGV). This analysis revealed that many variants identified in one technical replicate were also present in others but with lower sequencing depth as verified by manual review in IGV, and thus did not pass variant calling thresholds. Importantly, we observed that samples derived from smaller cell numbers more frequently contained variants not detected in standard exome sequencing from blood-derived DNA of the same samples.

Of the variants identified, those located in the *JAK3* and *LAT* genes displayed low variant allele frequencies (VAF < 0.14 and < 0.21, respectively). However, their

QUAL scores (34.6 and 33.7) prevented exclusion based on our established filtering thresholds. In addition, the read depth at these loci was high (average DP = 224), even in samples where the variants were not detected in the corresponding control. Thus, coverage-based filtering alone would not have eliminated these variants.

Additional variant calling using the Genome Analysis Toolkit (GATK) v4.6.2.0 confirmed the presence of the variant detected in the matched control DNA sample across all samples, including the one in which the variant had not been identified by DeepVariant v1.5.0. This further supports the importance of using multiple tools to ensure accurate detection and minimize false negatives.

PTA-WES Analysis of embryonic trophectoderm samples

Here we also mention our preliminary results from an experiment in which trophectoderm biopsies from blastocyst trophectoderms were subjected to the PTA-WGA method. Seven different exomes from embryo biopsies samples were sequenced, resulting in 70–120 M reads per sample with mean coverage of 136x and a median coverage of 102x with 85.71% on-target reads. Coverage statistics across samples indicated that the percentage of target regions covered 10x ranged from 86.39% to 96.58%, with an average of 92.29% (± 4.25%SD); for coverage at 20x, it ranged from 77.73% to 94.75%, with an average of 86.93% (SD = 6.86%). The proportion of target regions with zero coverage remained low (1.27–2.23%, SD = 0.33%) (Additional file 5: Table 4).

Discussion

In this study, we evaluated the quality of WES on small fibroblast groups that underwent PTA-based WGA, simulating the scale of trophectoderm embryo biopsies. Additionally, we propose a custom artifact-filtering approach tailored to address the specific challenges associated with the PTA method. To ensure a robust comparison, we included in-lab reference genomic DNA E701 and DNA of NA12878 Platinum genome sample.

The results demonstrate that PTA-based amplification generates sufficient amplified DNA material for library preparation and further exome enrichment regardless of the initial amount of fibroblast cells. After the PTA step, the amplified products display a wide range (of ~ 250 to > 1,500 bp) of fragment sizes, as confirmed by gel electrophoresis results, suggesting that library preparation could bypass the fragmentation step. However, we chose to retain this step in the current study to avoid any potential loss of unique fragments.

All fibroblast DNA library samples containing varying initial cell counts (4, 10, 16, and 25 cells) were combined using precapture pooling prior to enrichment. Low cell-count samples produced data yields similar to those of higher cell counts, highlighting that PTA-WGA might support high-throughput sequencing in diverse cell input conditions.

The sequencing analysis results demonstrated that PTA enables consistent and efficient amplification across fibroblast single-cell genomes. Samples amplified by PTA achieved a mean percentage of target regions of 97.5% at 10x depth, meeting the ACMG recommendations of 95% at 10x depth. The mean and median target coverage values were comparable to GIAB and E701 DNA exomes after PTA. Additionally, we observed that PTA-amplified samples contained only approximately 1.28% uncovered regions, aligning with the expected coverage gaps often encountered in exome sequencing owing to low mappability genomic regions [28, 30].

The primary quality metrics of PTA-amplified fibroblast samples at the embryo biopsy scale meet established ACMG standards, showing their potential for clinical genetic analysis. For the 81 genes recommended by the ACMG for pathogenic variant identification, the average coverage for more than 65 genes exceeded 95% at a depth of > 10x, demonstrating the effectiveness of the PTA amplification method.

To ensure the reliability of variant calling from PTA-based WGA data, accurate filtering of amplification artefacts is critical. Our initial aim was to implement existing bioinformatic tools, such as PTATO and SCAN2, for artefact filtering in our WGA-derived samples. However, despite consulting with the developers and receiving their guidance and support, we were unable to successfully integrate these tools into our data analysis workflow. As a result, we developed and applied our own filtering strategy tailored to the characteristics of PTA-derived data. The proposed approach was successfully implemented across all samples and was specifically optimized to eliminate false-positive variants while preserving clinically relevant findings. These results indicate that PTA-based WGA meets initial requirements for further accurate variant detection for screening for monogenic disorders.

Our study is the first to achieve successful WES from trophectoderm human embryo biopsies. The PTA-WGA approach for embryos demonstrated superior performance providing more uniform amplification and thus higher median exome coverage (102x), low zero-coverage regions (mean 1.51%), and stable target region coverage at 10x and 20x depths (mean 92.29% and 86.93%, respectively). Although the data obtained from embryo biopsies appear less uniform than that derived from fibroblasts or genomic DNA, this variability is likely attributable to the intrinsic biological heterogeneity of the samples. Factors such as biopsy size, embryo quality, and sample handling conditions including storage and freeze-thaw cycles may further contribute to the observed differences in sequencing performance. Based on our current results, we suggest that increasing the sequencing depth per

sample beyond standard WES levels may further improve target region coverage in embryo biopsy samples, as several samples already exceeded 95% of target regions at 10x depth. These findings underscore the potential of PTA-based WGA to support robust WES in PGT-M, facilitating the detection of genetic variants. Our future work will include variant calling analyses to enhance our evaluation of PTA-WES.

During the initial stages of developing a method for whole-exome sequencing (WES) of low-input samples, we also evaluated the performance of MDA. Although these results are not included in the primary dataset of this study, we consider it important to report them here for the benefit of researchers exploring similar approaches. MDA-based WES presented significant challenges, including a high proportion of uncovered regions and pronounced coverage unevenness. Despite a comparable range of sequencing output per exome (70–160 million reads), the mean target coverage reached 124× (± 40.56% SD), whereas the median was only 6× (± 9.97% SD), indicating substantial variability in regional coverage. The percentage of target regions covered at ≥ 10× did not exceed 95% in any sample and ranged from 20.79% to 63.30% (SD = 16%). Moreover, the proportion of target regions with zero coverage ranged from 11.78% to 55.13% (SD = 18.5%), highlighting the limitations of MDA-WES in achieving uniform and reliable exome coverage. These observations are consistent with the limitations of MDA previously reported in whole-genome and exome applications [9].

In our study, PTA is a promising WGA tool for exome sequencing in PGT-M. It should be noted that WES is not suitable for standard PGT-A as it does not provide uniform genome coverage, which is essential for reliable aneuploidy detection. However, within the context of PGT-M, PTA's improved performance could expand the utility of WES in clinical applications by enabling the detection of both inherited and de novo variants. Detecting de novo pathogenic variants is particularly important in PGT-M, as these mutations arise spontaneously, are not inherited from either parent, and may lead to serious developmental disorders. Despite the progress achieved, certain gaps remain in the methodology, primarily the detection of variants that were not expected based on the corresponding control samples. Although the exact cause of this phenomenon is not yet clear, such findings may be related to PCR artifacts, inherent properties of the biological material, or other unidentified factors. As a result, samples processed via WGA remain difficult to interpret with full confidence, complicating the integration of PTA-WES into standard PGT-M workflows. Additionally, variability in the quality of starting material and sample preparation significantly affects the reliability of the sequencing output.

Conclusions

Our findings underscore the potential of the PTA-WGA to provide more uniform amplification, improved coverage uniformity, and reduced zero-coverage regions compared with alternative methods such as the MDA-WGA. These advancements are crucial for detecting both inherited and *de novo* pathogenic variants, which are vital for addressing genetic disorders. The application of a tailored artifact-filtering approach further improved data accuracy, demonstrating the adaptability of the PTA-WGA for diverse clinical genetic analyses.

However, the study also highlights ongoing challenges, including variability in data output influenced by sample preparation and starting material quality—particularly in trophectoderm embryo biopsies—as well as difficulties in confidently interpreting PTA-processed samples due to the presence of variants that currently remain unresolved through existing filtering approaches. Future work should therefore focus not only on optimizing laboratory protocols and computational methods to improve PTA's overall reliability but also on developing more effective strategies for filtering false-positive variants. Overall, while the results are promising, PTA-based WGA is not yet a fully mature technology and requires further refinement before it can be implemented in clinical practice.

Materials and methods

Sample collection

Human fibroblasts from a reference sample E701 from our laboratory were used for this study. They were carefully thawed from cryostorage and cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS), penicillin-streptomycin (50 U/ml) and L-glutamine (2 mM) to promote cell growth and maintain optimal conditions for cell proliferation. Once the number of cells was reached, the fibroblasts were detached from the culture flask surface via 0.25% trypsin-EDTA solution according to a standard trypsinisation protocol to ensure cell viability and maintain consistent cell morphology for downstream applications. The fibroblasts were divided into groups of 4, 10, 16, and 25 cells, which were then placed in 0.2 ml tubes. The number of cells in each group was determined using an automatic cell counter and viability analyser (Logos Biosystems, South Korea). Each group was prepared in triplicate. In addition, reference genomic DNA NA12878 and our in-lab reference DNA sample E701 was used at concentrations equivalent to 10 and 40 genomes.

In addition to fibroblast-derived material and genomic DNA, the study also included embryonic samples. The study included both large trophectoderm samples from arrested embryos (approximately 50 cells) and standard biopsies consisting of 5–7 cells. Additionally, we analyzed a sample obtained from a day-3 arrested embryo

Table 4 Characteristics of embryo samples: quality and cell count

Sample ID	Number of cells	Embryo quality
PTA-1	8	8 A
PTA-2	≈50	4BC
PTA-3	≈50	4BC
PTA-4	≈50	4BC
PTA-5	5	Unknown
PTA-6	5	Unknown
PTA-7	5	Unknown

at the 8-cell stage. This range of input material allowed us to evaluate the performance of the PTA-WES protocol across different sample types and input amounts. Detailed information on the type and quality of biological material used is provided in Table 4.

All embryo samples were donated for research purposes and provided by the V.I. Kulakov National Medical Research Center for Obstetrics, Gynecology, and Perinatology under the category of not suitable for implantation. Embryos were thawed using a standard vitrification warming protocol with Kitazato media. Approximately two hours after thawing, trophectoderm biopsies were performed following standard operating procedures (SOP) in Kulakov Center, using laser-assisted dissection. Biopsied cells were transferred into 0.2 mL PCR tubes containing 3 µL of buffer (PBS + 3% PVP) and stored at −40 °C. The WGA was performed directly in the same tubes to minimize sample loss and contamination.

WGA

PTA was performed according to the manufacturer's instructions (BioSkryb Genomics). After PTA, the DNA was purified via 2X Kapa Pure Beads (Roche). The DNA yield was quantified with the Qubit dsDNA HS Assay system (Life Technologies) and its quality was assessed by 1.5% agarose gel electrophoresis. MDA was measured using QIAGEN REPLI-g kits according to the manufacturer's instructions (Qiagen).

Sample Preparation and exome sequencing

For each library, 500 ng of PTA product or genomic DNA (for NA12878 and E701) was sheared via a Covaris LE220 according to the manufacturer's protocol, followed by size selection with Kapa Pure Beads (Roche) to achieve a fragment distribution peak at ~ 250 bp. DNA libraries were prepared using the MGI Universal DNA Library Prep Set, followed by final amplification with 8 PCR cycles. The concentration of the prepared libraries was measured using the Qubit Flex system (ThermoFisher) and the dsDNA HS Assay Kit. Quality control of the DNA libraries was performed via high-sensitivity analysis on a Bioanalyzer 2100 system (Agilent Technologies). Then, we pooled 900 ng of each of the 16 libraries into a single pool which was concentrated using a SpeedVac concentrator (ThermoFisher) at 60 °C. Exome enrichment of the pool with the Agilent SureSelect Human All Exon V8 probes was performed according to the RSMU_exome protocol [31]. The pool was then circularized, downloaded into 4 flow cell lanes and sequenced paired-end mode on the DNBSEQ-G400 platform, using the DNBSEQ-G400RS high-throughput sequencing set PE100 kit, following the manufacturer's instructions (MGI Tech).

Genomic data analyses

The quality of the obtained fastq files was analyzed using FastQC v0.11.9 [32]. On the basis of the quality metrics, the fastq files were trimmed via BBDuk by BBMap v38.96 [33]. The reads were aligned to the indexed reference genome GRCh38.p14 using bwa-mem2 v2.2.1 [34]. SAM files were converted into BAM files and sorted using SAMtools v1.9 to check the percentage of the aligned reads [35]. On the basis of the obtained BAM files, the quality metrics of exome enrichment and sequencing were calculated using Picard v2.22.4, and the number of duplicates was calculated using Picard MarkDuplicates v2.22.4 [36]. To evaluate the genomic regions of interest within the ACMG gene set, we used the panel described in reference [29]. Using the hg38 genome assembly (NCBI RefSeq), we generated BED files representing both the full genomic intervals and the coding sequences (CDS) of the target genes. We then intersected these intervals with a benchmark dataset of difficult-to-sequence regions in order to exclude potentially problematic loci prior to conducting any further analyses [30].

Variant calling and filtering

Variant calling was performed using DeepVariant v1.5.0 [37]. The multiallelic variants in VCF files were decomposed into biallelic variants using vt decompose v0.5772 and then normalized using vt normalize v0.5772 [38]. A depth coverage filter (DP > = 3) and FILTER = PASS was applied to the variants obtained. The variants were subsequently filtered by the target exome panel (Agilent SureSelect v8). Additionally, difficult-to-sequence regions in exome data, which are affected mainly by low-mappability regions, such as pseudogenes, tandem repeats, homopolymers, and other low-complexity regions (1.19 Mb in sum), were excluded [30].

Carrier screening analysis

The search for variants in genes associated with monogenic diseases was performed using a script written in Python3 and conducted using our own database, incorporating data from major publicly accessible resources such as ClinVar, PanelApp, OMIM and OrphaNet and filtered for variant and disease significance.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-025-12511-y.

Additional file 1.

Additional file 2.

Additional file 3.

Additional file 4.

Additional File 5.

Author contributions

AS: designed, performed research and analyzed data; writing - original draft; writing - review & editing; VB: designed, performed research and analyzed data; supervision; writing - review & editing; IV: software and visualization; ZR: software and visualization; TG: methodology; provided experimental samples; ER: methodology; provided experimental samples; MP: provided experimental samples; EG: provided experimental samples; TN: provided experimental samples; DR: funding acquisition; DK: designed research; resources and funding acquisition; conceptualization; project administration.

Funding

The study was been carried out within the framework of state assignment № 124020400004-9 on the topic Development of a virally delivered gene therapy drug for the treatment of Crigler-Najjar syndrome types I and II.

Data availability

Exome sequences (24 fastq pairs) from the E701 reference DNA were deposited into the NCBI open-access sequence read archive (SRA) in fastq.gz format under BioProject ID PRJNA1137605. Additionally, whole genome sequencing data have been deposited under BioProject ID PRJNA1083205. The exome sequences of the samples used in this study are available upon request.

Declarations

Ethics approval and consent to participate

The appropriate institutional review board approval for this study was obtained from the Ethics Committee at the Pirogov Russian National Research Medical University. Consent for publication. All participants (or their legal guardians) provided written informed consent for sample collection, subsequent analyses, and publication thereof.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 25 May 2025 / Accepted: 30 December 2025
Published online: 05 March 2026

References

1. Tšuiko O, Gallardo EF, Voet T, Vermeesch JR. Preimplantation genetic testing: single-cell technologies at the forefront of PGT and embryo research. Reproduction. 2020;160(5):A19–31.
2. Murphy LA, Seidler EA, Vaughan DA, Resetkova N, Penzias AS, Toth TL, Sakkas D. To test or not to test? A framework for counseling patients on preimplantation genetic testing for aneuploidy (PGT-A). Hum Reprod. 2019;34(2):268–75.
3. Polyakov A, Rozen G, Gyngell C, Savulescu J. Novel embryo selection strategies—finding the right balance. Front Reprod Health. 2023;5:1287621.
4. Glotov AS, Kazakov SV, Zhukova EA, Alexandrov AV, Glotov OS, Pakin VS, Baranov VS. Targeted next-generation sequencing (NGS) of nine candidate genes with custom ampliseq in patients and a cardiomyopathy risk group. Clin Chim Acta. 2015;446:132–40.
5. Volozonoka L, Miskova A, Gailite L. Whole genome amplification in preimplantation genetic testing in the era of massively parallel sequencing. Int J Mol Sci. 2022;23(9):4819.
6. Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, et al. Comprehensive human genome amplification using multiple displacement amplification. Proc Natl Acad Sci U S A. 2002;99(8):5261–6.
7. Ren Z, Huang P, Wang Y, Yao Y, Ren J, Xu L, et al. Technically feasible solutions to challenges in preimplantation genetic testing for thalassemia: experiences of multiple centers between 2019 and 2022. J Assist Reprod Genet. 2024(11). https://doi.org/10.1007/s10815-024-03240-4.
8. Niu W, Wang L, Xu J, Li Y, Shi H, Li G, Sun Y. Improved clinical outcomes of preimplantation genetic testing for aneuploidy using MALBAC-NGS compared with MDA-SNP array. BMC Pregnancy Childbirth. 2020;20:1–9.
9. Borgström E, Paterlini M, Mold JE, Frisen J, Lundeberg J. Comparison of whole genome amplification techniques for human single cell exome sequencing. PLoS One. 2017;12(2):e0171566.
10. Daley T, Smith AD. Modeling genome coverage in single-cell sequencing. Bioinformatics. 2014;30(22):3159–65.
11. Zhou X, Xu Y, Zhu L, Su Z, Han X, Zhang Z, Liu Q. Comparison of multiple displacement amplification (MDA) and multiple annealing and looping-based amplification cycles (MALBAC) in limited DNA sequencing based on tube and droplet. Micromachines. 2020;11(7):645.
12. Liu XL, Xu CM, Huang HF. Application and challenge of preimplantation genetic testing in reproductive medicine. Reprod Dev Med. 2019;3(03):129–32.
13. Soloveva EV, Skleimova MM, Minaycheva LI, Garaeva AF, Zhigalina DI, Churkin EO, Stepanov VA. PGT-M for spinocerebellar ataxia type 1: development of a STR panel and a report of two clinical cases. J Assist Reprod Genet. 2024;41(5):1273–83.
14. Unsal E, Ozer L, Polat M, Aktuna S, Baltaci V. HOW EFFECTIVE IS TARGET SEQUENCE ENRICHMENT DURING WHOLE GENOME AMPLIFICATION ON THE IMPROVEMENT OF PGT-M RESULTS? Fertil Steril. 2020;114(3):e434.
15. Piyamongkol S, Makonkawkeyoon K, Shotelersuk V, Sreshthaputra O, Pantasri T, Sittiwangkul R, Piyamongkol W. Pre-implantation genetic testing for Marfan syndrome using mini-sequencing. J Obstet Gynaecol. 2022;42(7):2846–52.
16. Xu H, Pu J, Yang N, Wu Z, Han C, Yao J, et al. First preimplantation genetic testing case of Meckel syndrome with a novel homozygous TXNDC15 variant in a non-consanguineous Chinese family. Mol Genet Genomic Med. 2024;12(1):e2340.
17. Murphy NM, Samarasekera TS, Macaskill L, Mullen J, Rombauts LJ. Genome sequencing of human in vitro fertilisation embryos for pathogenic variation screening. Sci Rep. 2020;10(1):3795.
18. Xu X, Hou Y, Yin X, Bao L, Tang A, Song L, Wang J. Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. Cell. 2012;148(5):886–95.
19. Steuerwald N, Durrett R, Parsons J, Hamilton A, Kontanstinidis M, Licciardi F, et al. Whole exome sequencing of embryo biopsies reveals clinically-significant de novo mutations. Fertil Steril. 2014;102(3):e25.
20. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Lyon E. ACMG clinical laboratory standards for next-generation sequencing. Genet Sci. 2013;15(9):733–47.
21. Xia Y, Katz M, Chandramohan D, Bechor E, Podgursky B, Hoxie M, Siddiqui N. The first clinical validation of whole-genome screening on standard trophectoderm biopsies of preimplantation embryos. F&S Rep. 2024;5(1):63–71.
22. Weier C, Griffith A, McKissock K, Mahmood S, Gordon T, Blazek J, et al. DIRECT MUTATION ANALYSIS FOR PGT-M UTILIZING A NOVEL WHOLE GENOME AMPLIFICATION TECHNOLOGY: AN ALTERNATIVE METHOD FOR RAPID, ACCURATE, AND REFERENCE-FREE RESULTS IN DIFFICULT CASES. Fertil Steril. 2024;122(1):e8–9.
23. Gonzalez-Pena V, Natarajan S, Xia Y, Klein D, Carter R, Pang Y, et al. Accurate genomic variant detection in single cells with primary template-directed amplification. Proc Natl Acad Sci USA. 2021;118(24):e2024176118.
24. Middelkamp, S., Manders, F., Peci, F., van Roosmalen, M. J., González, D. M., Bertrums,E. J., … van Boxtel, R. (2023). Comprehensive single-cell genome analysis at nucleotide resolution using the PTA Analysis Toolbox. Cell genomics, 3(9).
25. Luquette LJ, Miller MB, Zhou Z, Bohrson CL, Zhao Y, Jin H, et al. Single-cell genome sequencing of human neurons identifies somatic point mutation and indel enrichment in regulatory elements. Nat Genet. 2022;54(10):1564–71.

26. Vasiliadis, I., Belova, V., Shmitko, A., Kuznetsova, A., Samitova, A., Suchalko, O.,… Korostin, D. (2024). Experience in developing the human genome standard E701. bioRxiv,2024-09.
27. Genome in a Bottle Consortium. Genome in a Bottle NA12878 vcf/bed file repository [Internet]. 2014. ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/data/NA12878/analysis/GIAB_integration/NIST_RTG_PlatGen_merged_highconfidence_v0.2_Allannotate.vcf.gz
28. Belova, V., Vasiliadis, I., Repinskaia, Z., Samitova, A., Shmitko, A., Ponikarovskaya,N., … Korostin, D. (2024). Comparative evaluation of four exome enrichment solutions in 2024: Agilent, Roche, Vazyme and Nanodigmbio. bioRxiv, 2024-07.
29. Miller DT, Lee K, Abul-Husn NS, Amendola LM, Brothers K, Chung WK, et al. ACMG SF v3.2 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med. 2023;25(8):100866. https://doi.org/10.1016/j.gim.2023.100866.
30. Hijikata, A., Suyama, M., Kikugawa, S., Matoba, R., Naruto, T., Enomoto, Y., … Ohara,O. (2024). Exome-wide benchmark of difficult-to-sequence regions using short-read next-generation DNA sequencing. Nucleic acids research, 52(1), 114–124.
31. Belova V, Pavlova A, Afasizhev R, Moskalenko V, Korzhanova M, Krivoy A, et al. System analysis of the sequencing quality of human whole exome samples on BGI NGS platform. Sci Rep. 2022;12(1):609.
32. Andrews S, FastQC:. A quality control tool for high throughput sequence data. Cambridge, UK: Babraham Institute; 2017.
33. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. 2014. Available online: https://github.com/BioInfoTools/BBMap (accessed on 20 February 2023).
34. Li H, Durbin H. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics. 2009;25:1754–60.
35. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
36. Broad Institute. Picard Toolkit. 2014. Available online: https://broadinstitute.github.io/picard/
37. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983–7.
38. Tan A, Abecasis GR, Kang HM. Unified representation of genetic variants. Bioinformatics. 2015;31(13):2202–4.

Publisher's Note