Accurate CNV identification from only a few cells with low GC bias in a single-molecule sequencing platform

Liang Hu[1,2,3,4#], Qunting Lin[5#], Pingyuan Xie[4,6], Lidong Zeng[5], Lichun Liu[5], Mengxian Huang[5], Qin Yan[5], Meng Zhang[5]*, Ge Lin[1,2,3,4*]

[1]Institute of Reproduction and Stem Cell Engineering, School of Basic Medical Science, Central South University, Changsha, Hunan, China

[2] Reproductive and Genetic Hospital of CITIC-Xiangya, Changsha, Hunan, China

[3] NHC Key Laboratory of Human Stem Cell and Reproductive Engineering（Central South University）, Changsha, Hunan, China

[4]National Engineering Research Center of Human Stem Cells, Changsha, Hunan, China

[5] GeneMind Biosciences Company Limited, Shenzhen, Guangdong, China

[6] Hunan Normal University School of Medicine, Changsha, Hunan, China

[#] Equally contributed

*** Correspondence:** linggf@hotmail.com (G.L.), zhangmeng@genemind.com (M.Z.)

ABSTRACT

A technical problem of characterizing copy number variation of several cells with next-generation sequencing is the whole genome amplification induced bias. The result of CNVs and mosaicism detection is affected by the GC bias. Here, we report a rapid non-WGA sample preparation strategy for a single-molecule sequencing platform GenoCare1600. This approach, combined with a single-molecule sequencing platform that avoids the use of WGA and bridge PCR processes, can provide higher reliability with its lower GC bias. By combining our optimized Tn5-based transposon

insertion approach with GenoCare, we successfully detected CNVs as small as 1.29M and mosaicism as small as 20%, which is consistent with next-generation sequencing (NGS) data. Moreover, our GenoCare-TTI protocol showed less GC bias and less Mad of Diff. These results suggest that the optimized TTI approach, together with the GenoCare1600 sequencing platform, is a promising option for CNV characterization from maybe one single cell.

Keywords: single-molecule sequencing, transposon insertion, CNVs, mosaicism

INTRODUCTION

In vitro fertilization (IVF) is a technology that has been widely used in the treatment of infertility to improve pregnancy rates. However, a high proportion of first-trimester spontaneous miscarriages, which is associated with chromosomal aneuploidy in human pregnancies, greatly affected IVF outcome, especially for patients with advanced maternal age [1]. Preimplantation genetic testing for aneuploidy (PGT-A) can maximize the possibility of euploid embryo transfer, thus are thought to be a gospel to some IVF patients. Different genetic diagnostic technologies have been developed for PGT-A, such as fluorescence in situ hybridization (FISH) [2], quantitative polymerase chain reaction (qPCR) [3], microarray technologies including single nucleotide polymorphism (SNP) microarrays [4], array-based comparative genomic hybridization (aCGH) [5], and next-generation sequencing (NGS) [6]. Among them, FISH was not recommended for PGT-A since it could only screen a limited number of chromosomes, and the error rate of 5-15% usually led to disappointing pregnancy outcomes [7]. Array-based CGH and SNP microarray are reliable, but the expensive chip expenses significantly increase the PGT-A cost. The NGS approach is a widely used technique for PGT-A due to its ability to comprehensively screen all chromosomes at a competitive price. A recent study, using NGS platform, successfully detected CNV close to 1Mb in size [8]. Besides, NGS also can detect the presence of 20%-80% abnormal cells in a blastocyst biopsy [9]. The standard processes for PGT-A NGS library typically include whole genome amplification (WGA) of cells, fragmentation of WGA products, end-polishing, ligation of adaptor sequences, PCR amplification, and size-selection. However, WGA before NGS not only increases the PGT-A duration but also creates noises.

Recently, single-molecule sequencing arises as a new technology for clinical applications. It circumvents many library preparation issues by avoiding DNA amplification. In the past few years, previous work demonstrated sequencing of SNPs, M13 virus genome, and detection of trisomy 21/18/13 by single-molecule sequencing

platform GenoCare [10 11]. Advantages of GenoCare include time-saving and straightforward sample preparation; (ii) absence of PCR amplification and low GC bias; (iii) significantly more sequencing reads than other single-molecule platforms.

WGA is a crucial step to enable comprehensive chromosome analysis. Over the years, there have been significant advances in WGA techniques, and several alternatives are available. The first method is primer extension pre-amplification (PEP), followed by the more widely adopted degenerate oligonucleotide-primed PCR (DOP-PCR). The basic principle of DOP-PCR is to use degenerate primers containing a random six-base sequence and a fixed sequence. However, in the process of PCR, due to uncertain factors such as input DNA amount and GC content, overamplified regions and unamplified regions appear, leading to amplification bias [12]. Multiple displacement amplification (MDA) was developed using isothermal amplification to solve this problem. In 2012, Zong et al. reported a single-cell WGA method called multiple annealing and looping-based amplicon cycles (MALBAC) that employed quasi-linear amplification through looping-based amplicon protection followed by PCR to reduce non-linear amplification [13]. Despite these advances in WGA for NGS-based PGT-A, Several issues still hinder its widespread clinical application: (1) Time: WGA and NGS library constructions require two days from beginning to the end with common workflows: WGA, DNA fragmentation and repair, ligation to specific adapters, PCR enrichment. (2) Cost: reagents and equipment used in this process are expensive. (3) Bias: Amplification biases are generated during the WGA and sequencing library construction [14]. The Tn5 transposase-based sequencing library preparation was then applied to address these problems. The enzyme catalyzes translocation by integrating the ME sequence into the target sites of DNA strands [15 16]. Due to its fast workflow, low DNA input, and limited hands-on time, this method has been used widely by researchers [17].

In this study, we developed a new library preparation method based on Tn5 transposase (**Figure 1**). Workflow and experimental conditions were optimized to

skip the MDA process and give less hands-on time and GC bias. It was validated by comparison with traditional methods through sequencing on single-molecule sequencer GenoCare1600. Samples with copy number variations (CNVs) were also sequenced to demonstrate the ability to identify aneuploidy and mosaicism.

MATERIALS AND METHODS

Cell lines

The cell lines used in this study were established and cultured in human embryonic stem cell (hESC) bank of the National Engineering Research Center of human Stem Cells [18]. In this paper, a 5-cell or 50-cell sample was picked up by microinjection. chHES90 was a normal hESC and was used to demonstrate the impact of MDA, and the other hESC lines were used for the comparison of library preparation methods and sequencing platforms.

Tn5 transposon

To accommodate with single-molecule sequencer GenoCare, the adapter sequences for Tn5were designed as follows:

ME-SEQ: 5'-[phos]-CTGTCTCTTATACACATCT-[NH_2]-3';

Adapter 1:

5'-TCCTTGATACCTGCGACCATCCAGTTCCACTCAGATGTGTATAAGAGACAG-3';

Adapter 2:

5'-CTCAGATCCTACAACGACGCTCTACCGATGAAGATGTGTATAAGAGACAG-3'.

Library construction via transposon insertion

Three library preparation methods were developed. In each case, 50ng gDNA was used as DNA input. Fragmentation and transposon insertion reactions were done as follows: 4μL 5×buffer L, 1μL target DNA at 50 ng/ μL, 1.5 μL Tn5 transposase, and 13.5 μL H_2O were mixed and incubated at 55˚C for 10min in a preheated thermocycler. After this mutual treatment, in route 1, DNA was purified by 1 volume VAHTS DNA clean bead (Vazyme), and 20μL solution was then transferred into a

PCR tube, following by a five-cycle PCR process with 10 μL 5× GM PCR buffer, 2 μL forward primer, 2 μL reverse primer, 1 μL GM DNA polymerase, and 15 μL H_2O. Afterward, the solution was purified by 1.2 volume VAHTS DNA clean beads, and the final solution was 20μL.

Route 2 is different from route 1 by adopting asymmetrical PCR amplification. The solution mixture contained 10 μL 5×GM PCR buffer, 2 μL forward primer, 0.2 μL reverse primer, 1 μL GM DNA polymerase, and 16.8 μL H_2O. Ten cycles of PCR were performed to produce a sufficient yield. Eventually, a 20μL solution was obtained after beads purification.

Route 3, after fragmentation and transposon insertion, the reaction was terminated by adding 5 μL 5× stop buffer and staying at room temperature for 5 min. No further purification was needed. Asymmetric PCR amplification was carried out for 10 cycles under the condition of 10 μL 5× GM PCR buffer, 2 μL forward primer, 0.2 μL reverse primer, 1 μL GM DNA polymerase, and 11.8 μL H_2O. Eventually, a 20μL solution was obtained after beads purification.

From 50 ng DNA, we routinely got more than 25 ng/uL dsDNA from route 1 and 2, and ~200 ng/uL ssDNA from route 3. Products have a size distribution of 150~1000 bp and can be directly used on GenoCare.

Also, we develop a rapid library preparation kit base on Tn5 for cell lysis. In this paper, the GenoCare library was prepared according to the manufacturer's protocol: gDNA was extracted from 5 or 50 cells. Six microliters of lysis buffer were added, spun down, and the DNA was incubated at 55℃for 1 h in a preheated thermocycler. 0.5 μL lysis stop buffer was then added, spun down, and the tube stayed at room temperature for 35 min to stop lysis reaction. No further purification was needed.

The primers oligonucleotide sequences were as follows:

forward primer 5'-TTCCTCAGATCCTACAACGACGCTCTACCGAT-3',

reverse primer 5'-TTCTCCTTGATACCTGCGACCATCCAGTT-3'.

MDA

MDA was performed on five human cells, as described in Vazyme Discover-sc single-cell kit (Vazyme). Briefly, 3μL cell lysis buffer was freshly prepared and added into each tube. After heating at 65°C for 10 min, 3 μL buffer N was added to stop the lysis reaction. 40 μL of amplification buffer was then added to start the MDA process. PCR steps were as follows: 30°C for 6 hours, 65°C for 3 min. In order to understand the relationship between amplification time and GC bias, 0.5h, 1h, 2h, 4h, and 6h reaction times were studied. The final products were characterized by Qubit (Invitrogen) and agarose gel electrophoresis.

MDA product was according to the manufacturer's protocol for GenoCare library preparation, and the library construction is about 1.5 hours. Meanwhile, 6-hour MDA amplification products were performed using TruePrep DNA Library Prep Kit V2 for Illumina (Vazyme) and sequenced on Illumina Hiseq X10 as a comparison. The library preparation workflow of this NGS method is about 12 hours.

Sequencing and data analysis

The samples were sequenced on Genemind Bioscience's single-molecule sequencer GenoCare and Illumina's HiseqX10 sequencer, yielding more than 4% genome coverage for each sample. GenoCare sequencing was performed according to the previously disclosed protocol [11]. For each sample, 25% of the area of each flowcell channel was imaged, and 72 cycles of sequencing data were collected, yielding more than 4 million reads. NGS Sequencing was done on Illumina Hiseq X10 with standard PE-150 protocol according to the operating instruction manual, and the total sequence time is about three days.

In a basic data analysis process, Illumina X10 sequencing data were mapped to the reference human genome (hg19) by bwa, and antigenocide data were mapped to the same reference using home-written software called DirectAlign. Raw reads with low quality and non-unique alignments were removed. To reduce the influence of GC and mappability differences, we split the reference into 150 Kbp windows and kept the

bins with GC content 32%~60% and mappability bigger than 0.6. To compare GC bias between samples, we calculated the relative bin density (RBD) by $R_{i,j}=r_{i,j}/M$. Where r denotes reads number in each bin, i and j represent different bins and samples, and M is the average number of sequencing reads in bins on autosomes. GC bias ($\Delta {R_{GC}}^2$) was defined as

$$\Delta {R_{GC}}^2 = 1 - \frac{{\sigma_L}^2}{\sigma^2}, \sigma^2 = \frac{1}{n}\sum_{i-1}^{n}(R_i - M)^2 , {\sigma_L}^2 = \frac{1}{n}\sum_{i=1}^{n}(R_i - R_L)^2.$$

Moreover, R_L represents the optimal prediction, which was obtained via a loess regression fit of the RBD against the GC content. Then we developed a weighted correction strategy to correct GC content and mappability in the scale of every 0.1% bin. Weight index *w* was calculated from an R function loss; the corrected bin density (CBD) was calculated in the following formula: *CBD=RBD/w*. This normalized bin density is used for further analysis. CNV identification was made by R packages (DNA copy).

We calculated the median absolute deviation of difference (Mad of Diff) to evaluate the reproducibility instead of the coefficient of variation (CV) because Mad of Diff represents the difference of corrected and normalized adjacent bin copy number, which is more accurate when assessing CNV samples.

Statistics

The differences in coverage, GC bias, and Mad of Diff were compared using the student's t-test. P<0.05 was considered statistically significant. Analyses were performed using the statistical package SPSS, version 18.0 (SPSS).

RESULTS

Cell line

To develop an effective library method strategy for CNVs and mosaicism detection at the cell level through a single-molecule sequencing platform, we chose five human embryonic stem cell lines with different sizes of CNV range from 1.29 Mbp to 56.26 Mbp (**Table 1)**. NGS confirmed all the CNVs of the cells lines with their total

genomic DNA **(Figure 1A)**. The range of copy number gain (CNG), whose copy ratios were 3/2, were from 2.47 Mbp to 56.26 Mbp; And the range of copy number loss (CNL), whose copy ratios were 1/2, were from 1.29 Mbp to 6.28 Mbp **(Figure 1B)**.

Whole-genome amplification induced significant bias

It is known that different WGA approaches induce error and template bias. To evaluate the influence of short-time MDA, we sequenced the short-time MDA products (from 0.5 hours to 6 hours) of five chESC90 cells and compared with those of unamplified 5-cell DNA, 50-cell DNA, and DNA of bulk cells (**Figure 2A**). To minimize the effect of amplification, we used Tn5-based transposon insertion (TTI) to construct the library for next-generation sequencing. As expected, the coverage of unamplified 5-cell DNA and 50-cell DNA was lower than that of bulk DNA and MDA-amplified products **(Figure 2B)**. Surprisingly, even MDA for as short as 30 minutes induced statistically significant bias on GC bias and Mad of Diff. As shown in Figure 2B, after an only 30-minute amplification, the GC bias increased to 0.2. After 6-hour MDA, GC bias even reached as high as 0.62, which was 62 fold compare to bulk DNA without amplification (0.01).

Importantly, although the GC bias of unamplified 5 cells and 50 cells also increased to 0.03 and 0.05 respectively, they were only about 1/5 of that of 0.5-hour MDA, indicating that cell number does not significantly affect GC bias **(Figure 2C)**. Similarly, Mad of Diff, a criterion which reflects the standard deviation of the sequencing data, increased to about 3 fold of that of bulk DNA (0.09) after amplification. Moreover, the Mad of Diff only increased to 0.15 for unamplified 5 cells and 50 cells **(Figure 2D)**. The results were summarized in **Table S1**.

To get further insight into correlations between those library preparation methods, we calculated Pearson's cross-correlation coefficients of relative bin density (RBD) between each sequencing result (**Figure 2E**). The hierarchical clustering of the

correlation coefficient matrix showed that the MDA methods were well separated from each other and were also very different than the PCR-free bulk cells. Results from unamplified 5 cells and 50 cells and bulk cells are highly correlated, indicating that cell lysis following by unamplified Tn5-based library preparation had the least bias.

Optimization of Tn5-based transposon insertion protocol

Our above data showed that Tn5-based transposon insertion (TTI) is a promising library prep strategy since the DNA was unamplified. However, amplification of the trace amount of DNA template is necessary during PGT-A, especially when the patients choose both PGT-A and PGT for monogenic diseases (PGT-M). Thus, we developed three routes of different pretreatment protocols and PCR strategies to optimize the amplification-based TTI (**Figure 3A**). Route 1 is a frequently used approach that purifies the DNA by beads and then amplified the purified DNA by symmetrical PCR. Since symmetrical PCR might lose some of single-strand DNA (ssDNA), we amplified the beads-purified DNA by asymmetrical PCR (Route 2). To minimize the loss of DNA during purification, we just use stop buffer to stop the TTI reaction and then amplified the unpurified DNA by asymmetrical PCR (Route 3). To test the performance of three routes, we sequenced six hESC cell line MDA product samples on the GenoCare platform and analyzed sequencing results of unique data, coverage, GC bias, and Mad of Diff (**Table S2**).

We first calculated the correlation coefficient matrix among the three routes. As shown in **Figure 3B**, the hierarchical clustering of the correlation coefficient matrix showed that different library preparation routes and conditions clustered, respectively. They also displayed a weak correlation with each other, which indicated that each route had its built-in pattern of the library.

As shown in **Figure 3C and 3D**, compared with other routes, route 3 can obtain the most unique data (route 3 vs. route 1, 4.49 vs. 2.94 M reads, $p<0.01$; route 3 vs. route 2, 4.49 vs. 3.94 M reads, $p<0.05$) and the highest coverage (route 3 vs. route 1, 6.27% vs. 3.74%, $p<0.01$; route 3 vs. route 2, 6.27 vs. 4.62%, $p<0.05$), indicating that the

stop buffer and asymmetric PCR amplification can avoid the DNA loss. Besides, the average GC bias of route 3 is much less than the other two routes (**Figure 3E**, route 3 vs. route 1, 0.15 vs. 0.33, $p<0.01$; route 3 vs. route 2, 0.15 vs. 0.35, $p<0.01$).

Besides, since route 3 does not need the beads purification, its library preparation time is only 1.5 hours, which saves 0.5-1 hour compare to route 1 and route 2 (**Table S2**). The above data indicated that route 3 has a clear advantage over the other two routes, which makes it an ideal method for CNV detection for trace amount cells.

Optimized TTI protocol has less GC bias and can identify small CNVs

To test the performance of the optimized TTI protocol in identifying CNVs, we sequenced six hESC cell line gDNA samples with small CNVs on the GenoCare platform. The results were compared with those amplified by MDA and sequenced on Illumina HiseqX10 with 150bp paired-end reads. As shown in **Table 2**, the average coverage of GenoCare-TTI protocol was less than that of X10-MDA protocol (7.61% vs. 14.90), because the average read length of GenoCare1600 was less than that of Illumina HiseqX10 (42.06 vs. 150.00). Notably, the GC bias of GenoCare-TTI protocol was much less than that of the X10-MDA protocol (0.03 vs. 0.23, $p<0.01$), indicating that GenoCare-TTI protocol might have better coverage with fewer reads. Moreover, the average Mad of Diff of GenoCare-TTI protocol was less than that of X10-MDA protocol (0.22 vs. 0.31, $p<0.05$), indicating that GenoCare-TTI protocol might be more powerful to detect small CNVs.

Besides, we also evaluated whether GenoCare-TTI protocol can identify the small CNVs. As shown in **Figure 4A**, all the CNVs, including a 1.29M small CNV, can be successfully identified both with GenoCare-TTI protocol and with X10-MDA protocol. It should be noted that since GenoCare-TTI protocol has less Mad of Diff, the small CNV was more easily to be characterized than X10-MDA protocol (**Figure 4B**).

Optimized TTI protocol can identify mosaicism

To investigate the accuracy of PGT-A in mosaicism detection, we made a series of

mosaicism samples by mixing CNV-cell line chHES488 with the normal hESC cell line 90-P51 (**Figure 5A**). Those mosaic samples were characterized by either GenoCare-TTI protocol or with X10-MDA protocol. As shown in **Figure 5B**, even 20% of mosaicism can be identified by both protocols. We also found that GenoCare-TTI protocol has much less GC bias (**Table S3**) in this experiment, the average bin copy number remained stable as the GC content varied (**Figure 5C**).

DISCUSSION

Our present study demonstrated the validity of a new 1.5-hour PGT-A sample preparation method for as little as five cells, like the blastocyst biopsy samples during PGT-A. The technical accuracy was measured in MDA amplification products CNVs and mosaic ratios. We compared the influence of the gain achieved in MDA on the characteristics of interest. Overall, the amplification bias in MDA is a direct function of the overall reaction gain, with more significant gain leading to more bias. This observation underlines the importance of tailoring the gain of amplification to yield a higher quality of DNA products for the subsequent sequencing workflow. The overall gain can be set through by reducing the reaction time.

For single-cell sequencing, WGA is required before sequencing library construction during NGS based PGT-A. It has been demonstrated that significant GC bias will arise during the WGA process, which would affect the sequencing profile accuracy, leading to false positive or false negative [19]. In our study, we found that the cell genome amplification by cell lysis Tn5-based library preparation method is highly efficient. It rendered better reproducibility and uniformity than MDA, especially concerning GC content.

We also compared the efficiency of CNVs detection by different Tn5-based library preparation methods. All three optimized Tn5-based library preparation methods generated sequence data on the SMTS GenoCare platform and are concordant with the date from the Illumina X10 platform. Among the three routes, Route 1 and 2

showed high GC bias, which may be caused by two reasons. Previous research revealed that half of the genomic DNA fragments were lost due to transposon symmetry in the conventional method [20]. As a result, PCR amplification lost uniformity due to the lower template amount.

Moreover, since the same adaptor sequences were attached to genomic DNA fragments on both ends, self-looping of the template strand may happen, which could lead to PCR amplification failure in the GC-rich regions. To solve this problem, we designed an asymmetric PCR amplification to enrich ssDNA [21]. Meanwhile, we replaced beads purification with stop buffer to stop fragmentation reaction to reduce the operation time and DNA loss. Interestingly, we found that a combination of asymmetric PCR and stop buffer treatment (route 3) could not only yield reliable results for CNVs diagnosis but also was faster than the other two Tn5-based library preparation methods (route 1 and 2). Meanwhile, NGS results of route 3 showed much less GC bias with compare to those of route 1 and 2. Therefore, route 3 may be an ideal method for PGT-A.

Mosaicism in embryos as a result of postzygotic mitotic aneuploidy could contribute to biologic variation in blastocysts. Chromosomal mosaicism frequently occurs in human blastocysts as detected during PGT-A [22]. A mosaic embryo or cell line is that one has cells with different CNVs in at least one chromosome. Levels of mosaicism at the cleavage stage were estimated to range from 15% to 90% [23]. According to the Preimplantation Genetic Diagnosis International Society (PGDIS), less than 20% mosaicism is deemed as euploidy. When the mosaic ratio is more than 80%, the embryo is considered to be aneuploidy. When the mosaic ratio is between 20% and 80%, PGT-A results will help doctors make embryo transfer decisions. Empirically, >50% mosaic embryos are easy to detect with the use of multiple PGT-A platforms. To fully validate route 3 for PGT-A, we also applied it on the mosaic curve detection. As our data showed, the six different proportions of mosaic samples of this curve have a noticeable trend of gradually decreasing. Analysis of these

well-controlled samples by route 3 demonstrated perfect consistency with the expected mixing proportions. The standard curve of the mosaic ratio can provide a relatively accurate reference method for the assessment of the chimeric ratio during PGT-A. The approach may help reduce the impact of mosaicism and biologic variation on evaluating the technical accuracy of new methods. Route 3, combining with SMTS GenoCare 1600 sequencer, can accurately and reproducibly measure 20% mosaicism in a known sample.

To our knowledge, there were very few reports that a single molecule sequencer was applied in PGT-A. In this study, we compared the sequencing performance of GenoCare1600 and Illumina Hiseq X10 for as little as five cells in the same 11 hESC lines to mimic the real PGT-A circumstances. At least 4 million sequencing reads were collected for each sample, which meets the requirement of PGDIS. With the PE-150 sequencing protocol, X10 delivered higher RL and genome coverage, while GenoCare 1600 gave lower GC bias and shorter sequencing time with less sequencing cycles. GenoCare 1600 could produce more than 10 million reads per flowcell channel, 160 million reads per flowcell, comfortably accommodating 16 PGT-A samples for each run. Only one-fourth of each flowcell channel was imaged in this study to reduce sequencing time further. The reads number is one to two orders of magnitude higher than other single-molecule sequencers, namely Pacific biosciences Sequeland Oxford Nanopore's Minion/GridIon. Although GenoCare's read length is short, its large amount of reads number makes it more suitable for CNV detection.

In this study, we characterized an alternative method of library construction, which combines with Genocare sequencing platform produced fast and accurate PGT-A data. We have successfully used the Tn5-based library preparation method to investigate the different CNVs of the cell-line samples, which was consistent with NGS results. Besides, for samples with a 20% to 80% mosaic ratio, Genocare sequencing results are very consistent with NGS results. We believe that this new platform delivers a

competitive solution to PGT-A by reducing the test time and generating data with lower GC bias and comparable CNV sensitivity.

ACKNOWLEDGMENTS

Supported by the National Key R&D Program of China (grant 2018YFC1003100, to L.H.), the National Natural Science Foundation of China (grant 81873478, to L.H.), and the science and technology major project of the ministry of science and technology of Hunan Province, China (grant 2017SK1030, to G.L.).

CONFLICT OF INTEREST

Q.L, L.Z, L.L, M.H, Q.Y, and M.Z are employees of GeneMind Biosciences Company Limited.

REFERENCES

1. Brezina PR, Kutteh WH. Clinical applications of preimplantation genetic testing. BMJ 2015;**350**:g7611 doi: 10.1136/bmj.g7611[published Online First: Epub Date]|.
2. Fernandez SF, Toro E, Colomar A, et al. A 24-chromosome FISH technique in preimplantation genetic diagnosis: validation of the method. Syst Biol Reprod Med 2015;**61**(3):171-7 doi: 10.3109/19396368.2014.1002869[published Online First: Epub Date]|.
3. Treff NR, Tao X, Ferry KM, et al. Development and validation of an accurate quantitative real-time polymerase chain reaction-based assay for human blastocyst comprehensive chromosomal aneuploidy screening. Fertil Steril 2012;**97**(4):819-24 doi: 10.1016/j.fertnstert.2012.01.115[published Online First: Epub Date]|.
4. Tan YQ, Tan K, Zhang SP, et al. Single-nucleotide polymorphism microarray-based preimplantation genetic diagnosis is likely to improve the clinical outcome for translocation carriers. Hum Reprod 2013;**28**(9):2581-92 doi: 10.1093/humrep/det271[published Online First: Epub Date]|.
5. Gutierrez-Mateo C, Colls P, Sanchez-Garcia J, et al. Validation of microarray

comparative genomic hybridization for comprehensive chromosome analysis of embryos. Fertil Steril 2011;**95**(3):953-8 doi: 10.1016/j.fertnstert.2010.09.010[published Online First: Epub Date]|.

6. Tan Y, Yin X, Zhang S, et al. Clinical outcome of preimplantation genetic diagnosis and screening using next generation sequencing. Gigascience 2014;**3**(1):30 doi: 10.1186/2047-217X-3-30[published Online First: Epub Date]|.
7. Debrock S, Melotte C, Spiessens C, et al. Preimplantation genetic screening for aneuploidy of embryos after in vitro fertilization in women aged at least 35 years: a prospective randomized trial. Fertil Steril 2010;**93**(2):364-73 doi: 10.1016/j.fertnstert.2008.10.072[published Online First: Epub Date]|.
8. Wang L, Cram DS, Shen J, et al. Validation of copy number variation sequencing for detecting chromosome imbalances in human preimplantation embryos. Biol Reprod 2014;**91**(2):37 doi: 10.1095/biolreprod.114.120576[published Online First: Epub Date]|.
9. Fiorentino F, Biricik A, Bono S, et al. Development and validation of a next-generation sequencing-based protocol for 24-chromosome aneuploidy screening of embryos. Fertil Steril 2014;**101**(5):1375-82 doi: 10.1016/j.fertnstert.2014.01.051[published Online First: Epub Date]|.
10. Gao Y, Deng L, Yan Q, et al. Single molecule targeted sequencing for cancer gene mutation detection. Sci Rep 2016;**6**:26110 doi: 10.1038/srep26110[published Online First: Epub Date]|.
11. Zhao L, Deng L, Li G, et al. Single molecule sequencing of the M13 virus genome without amplification. PLoS One 2017;**12**(12):e0188181 doi: 10.1371/journal.pone.0188181[published Online First: Epub Date]|.
12. Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. Nature 2011;**472**(7341):90-4 doi: 10.1038/nature09807[published Online First: Epub Date]|.
13. Zong C, Lu S, Chapman AR, et al. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. Science 2012;**338**(6114):1622-6 doi: 10.1126/science.1229164[published Online First:

Epub Date]|.

14. Huang L, Ma F, Chapman A, et al. Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. Annu Rev Genomics Hum Genet 2015;**16**:79-102 doi: 10.1146/annurev-genom-090413-025352[published Online First: Epub Date]|.
15. Adey A, Morrison HG, Asan, et al. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. Genome Biol 2010;**11**(12):R119 doi: 10.1186/gb-2010-11-12-r119[published Online First: Epub Date]|.
16. Reznikoff WS. Tn5 as a model for understanding DNA transposition. Mol Microbiol 2003;**47**(5):1199-206 doi: 10.1046/j.1365-2958.2003.03382.x[published Online First: Epub Date]|.
17. Marine R, Polson SW, Ravel J, et al. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. Appl Environ Microbiol 2011;**77**(22):8071-9 doi: 10.1128/AEM.05610-11[published Online First: Epub Date]|.
18. Lin G, Xie Y, Ouyang Q, et al. HLA-matching potential of an established human embryonic stem cell bank in China. Cell Stem Cell 2009;**5**(5):461-5 doi: 10.1016/j.stem.2009.10.009[published Online First: Epub Date]|.
19. Sabina J, Leamon JH. Bias in Whole Genome Amplification: Causes and Considerations. Methods Mol Biol 2015;**1347**:15-41 doi: 10.1007/978-1-4939-2990-0_2[published Online First: Epub Date]|.
20. Buenrostro JD, Giresi PG, Zaba LC, et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 2013;**10**(12):1213-8 doi: 10.1038/nmeth.2688[published Online First: Epub Date]|.
21. Tolnai Z, Harkai A, Szeitner Z, et al. A simple modification increases specificity and efficiency of asymmetric PCR. Anal Chim Acta 2019;**1047**:225-30 doi: 10.1016/j.aca.2018.10.017[published Online First: Epub Date]|.
22. Goodrich D, Tao X, Bohrer C, et al. A randomized and blinded comparison of

qPCR and NGS-based detection of aneuploidy in a cell line mixture model of blastocyst biopsy mosaicism. J Assist Reprod Genet 2016;**33**(11):1473-80 doi: 10.1007/s10815-016-0784-3[published Online First: Epub Date]|.

23. Harton GL, Cinnioglu C, Fiorentino F. Current experience concerning mosaic embryos diagnosed during preimplantation genetic screening. Fertil Steril 2017;**107**(5):1113-19 doi: 10.1016/j.fertnstert.2017.03.016[published Online First: Epub Date]|.

FIGURE LEGENDS

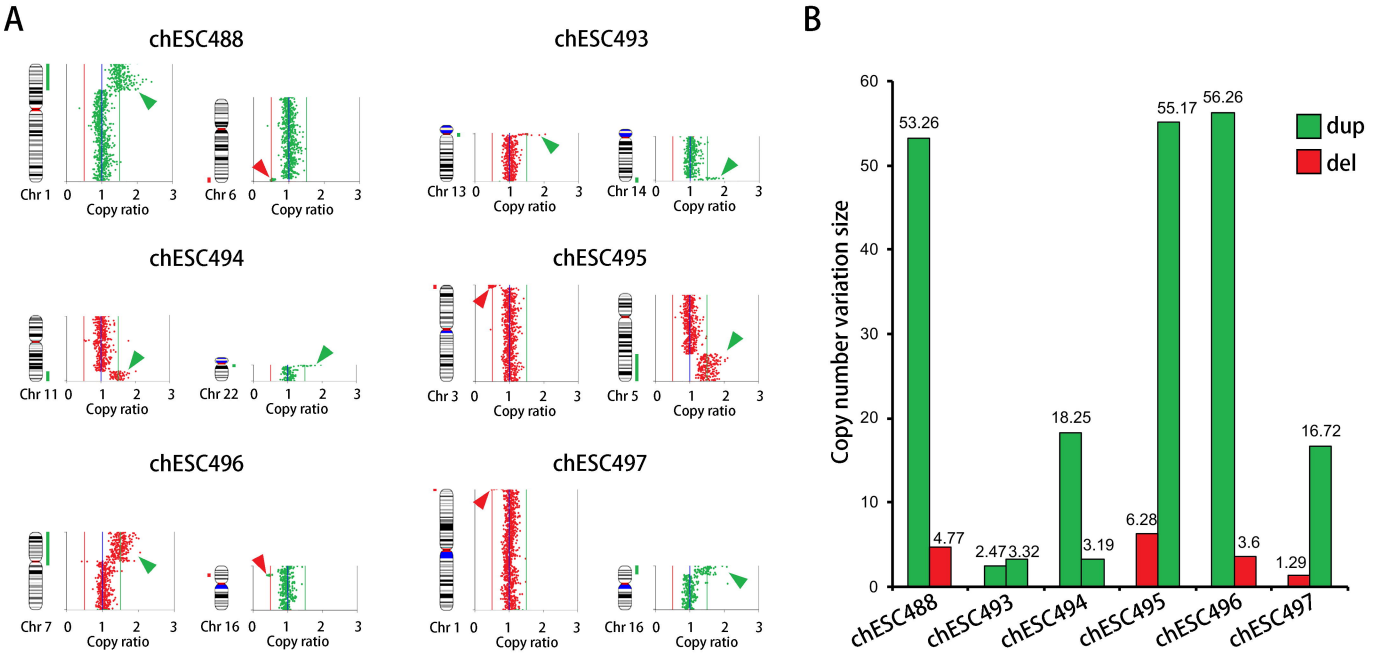
Figure 1. Next-generation sequencing (NGS) results of the six abnormal human embryonic stem cell lines. (A) The copy number variation (CNV) of affected chromosomes; (B) The CNV size of the six abnormal human embryonic stem cell lines.

Figure 2. WGA induced significant bias in NGS. (A) The experimental design scheme to characterize the WGA-induced bias; (B) Unamplified gDNA had relatively low coverage; (C) Unamplified gDNA had less GC bias; (D) Unamplified gDNA had less Mad of Diff; (E) Amplified gDNA and unamplified gDNA correlated together, respectively.

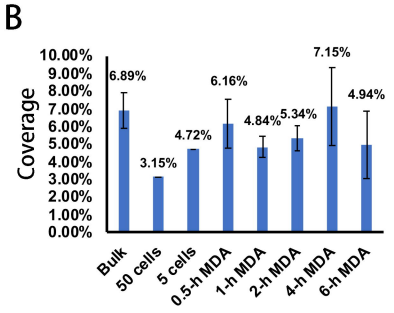
Figure 3. Stop buffer, and asymmetrical PCR optimized the Tn5-based transposon insertion (TTI) protocol. (A) The experimental design scheme to optimize the TTI protocol; (B) The sequencing results of three routes had a weak correlation with each other; (C, D, E) Route 3 had more unique data, higher coverage, and less GC bias.

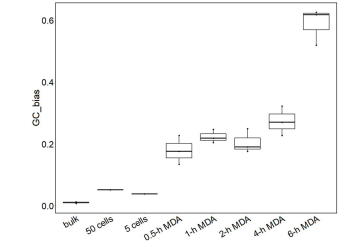
Figure 4. Optimized TTI protocol can identify small CNVs. (A) The small CNVs can be identified by both GenoCare-TTI protocol and X10-MDA protocol; (B) Both protocols could detect a CNV as small as 1.29M. Moreover, X10-MDA protocol had less Mad of Diff.

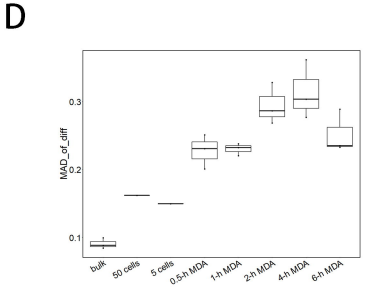
Figure 5. Optimized TTI protocol can identify as low as 20% mosaicism. (A) The experimental design scheme of mosaicism characterization; (B) Both protocols could detect as low as 20% mosaicism; (C) The average bin copy number of GenoCare-TTI protocol did not vary with GC content. The bin coverage rate distribution across the genome from different libraries preparation and a bulk-cell sample. Each bar represents the coverage rate in a 150 kb bin, and all 17019 bins are plotted.

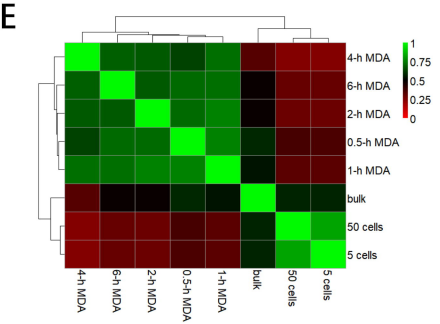

A
chESC488
Chr 1
Chr 6
Copy ratio
chESC493
Chr 13
Chr 14
chESC494
Chr 11
Chr 22
chESC495
Chr 3
Chr 5
chESC496
Chr 7
Chr 16
chESC497
Chr 1
Chr 16
B
Copy number variation size
dup
del
53.26
4.77
2.47
3.32
18.25
3.19
6.28
55.17
56.26
3.6
1.29
16.72
chESC488
chESC493
chESC494
chESC495
chESC496
chESC497

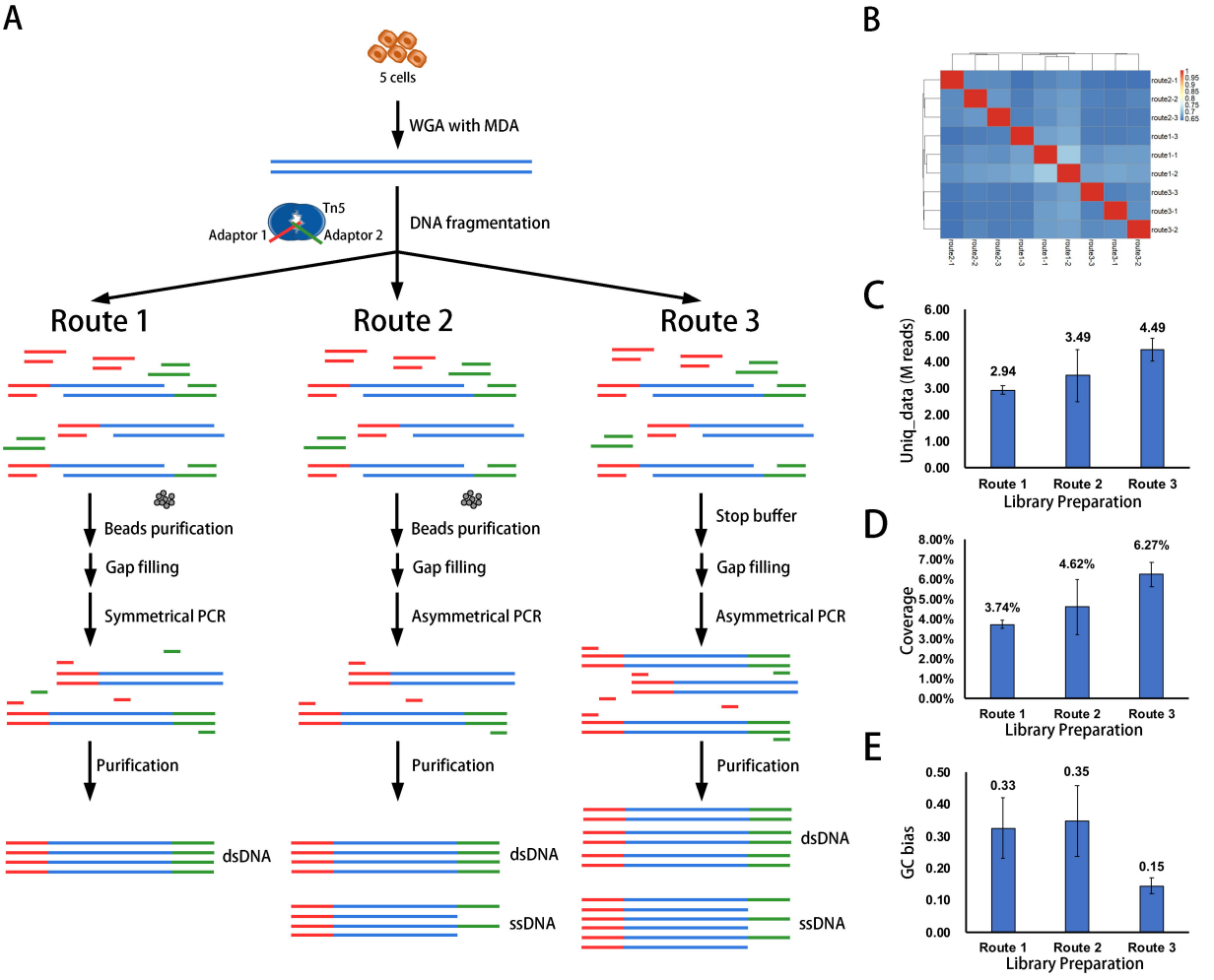

A
bulk gDNA
50 cells
5 cells
hESC cell line
No whole genome amplification
5 cells
Whole genome amplification
(MDA for 0.5hr, 1hr, 2hrs, 4hr, 6hrs)
GenoCare 1600


B
Coverage
10.00%
9.00%
8.00%
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
6.89%
3.15%
4.72%
6.16%
4.84%
5.34%
7.15%
4.94%
Bulk
50 cells
5 cells
0.5-h MDA
1-h MDA
2-h MDA
4-h MDA
6-h MDA

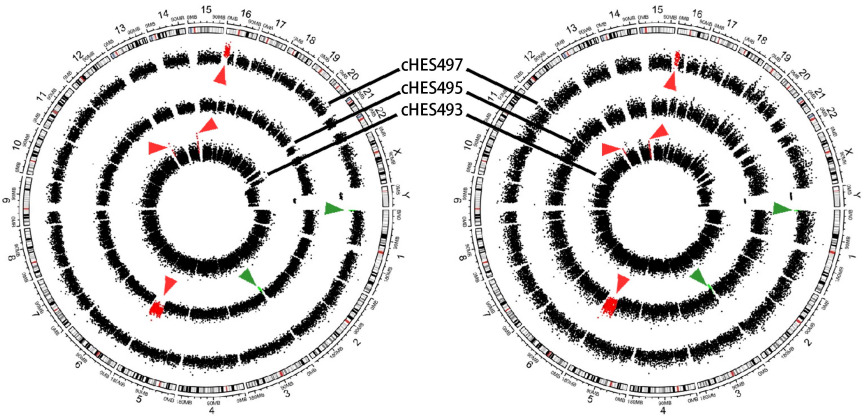

C
GC_bias
0.6
0.4
0.2
0.0
bulk
50 cells
5 cells
0.5-h MDA
1-h MDA
2-h MDA
4-h MDA
6-h MDA


D
MAD_of_diff
0.3
0.2
0.1
bulk
50 cells
5 cells
0.5-h MDA
1-h MDA
2-h MDA
4-h MDA
6-h MDA

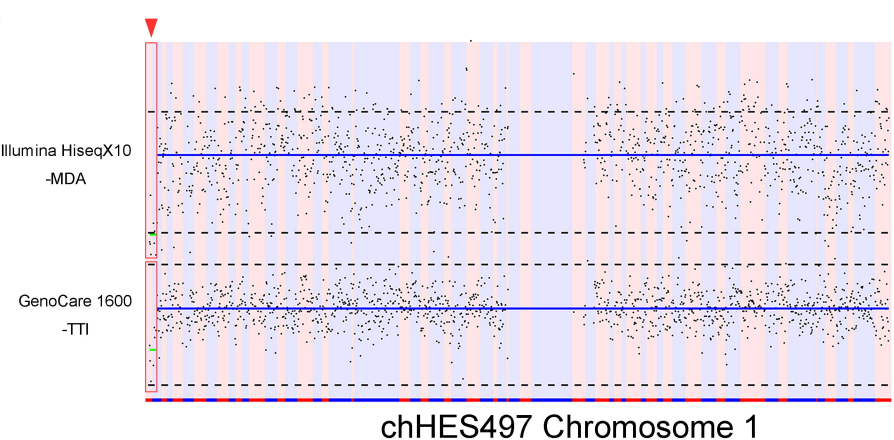

E
4-h MDA
6-h MDA
2-h MDA
0.5-h MDA
1-h MDA
bulk
50 cells
5 cells
1
0.75
0.5
0.25
0


A
5 cells
WGA with MDA
Tn5
Adaptor 1
Adaptor 2
DNA fragmentation
Route 1
Route 2
Route 3
Beads purification
Gap filling
Symmetrical PCR
Purification
dsDNA
Beads purification
Gap filling
Asymmetrical PCR
Purification
dsDNA
ssDNA
Stop buffer
Gap filling
Asymmetrical PCR
Purification
dsDNA
ssDNA
B
route2-1
route2-2
route2-3
route1-3
route1-1
route1-2
route3-3
route3-1
route3-2
0.95
0.9
0.85
0.8
0.75
0.7
0.65
C
Uniq_data (M reads)
6.00
5.00
4.00
3.00
2.00
1.00
0.00
2.94
3.49
4.49
Route 1
Route 2
Route 3
Library Preparation
D
Coverage
8.00%
7.00%
6.00%
5.00%
4.00%
3.00%
2.00%
1.00%
0.00%
3.74%
4.62%
6.27%
Route 1
Route 2
Route 3
Library Preparation
E
GC bias
0.50
0.40
0.30
0.20
0.10
0.00
0.33
0.35
0.15
Route 1
Route 2
Route 3
Library Preparation

A


cHES497
cHES495
cHES493
GenoCare 1600-TTI
Illumina HiseqX10-MDA

B


Illumina HiseqX10
-MDA
GenoCare 1600
-TTI
chHES497 Chromosome 1

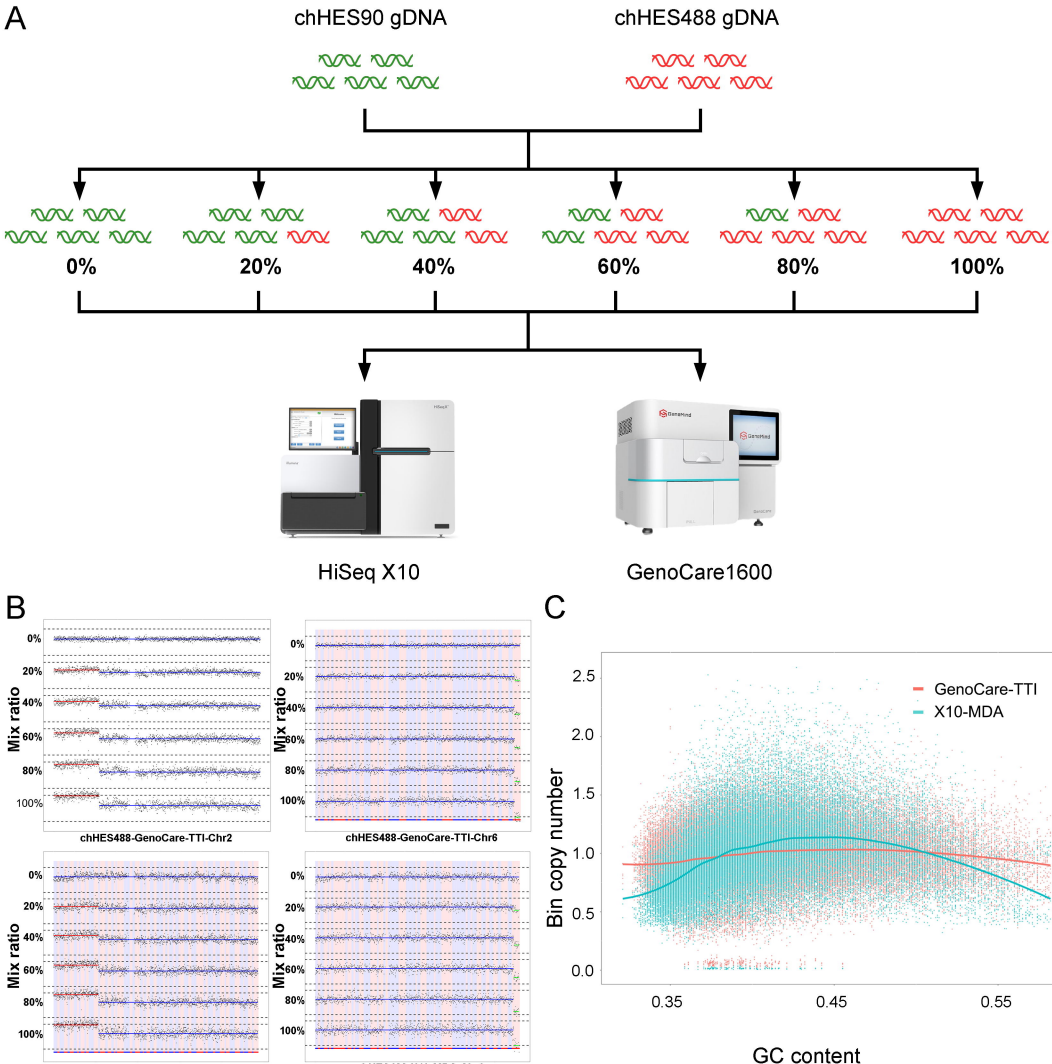

A
chHES90 gDNA
chHES488 gDNA
0%
20%
40%
60%
80%
100%
HiSeq X10
GenoCare1600
B
Mix ratio
chHES488-GenoCare-TTI-Chr2
chHES488-GenoCare-TTI-Chr6
chHES488-X10-MDA-Chr2
chHES488-X10-MDA-Chr6
C
GenoCare-TTI
X10-MDA
Bin copy number
GC content
0.0
0.5
1.0
1.5
2.0
2.5
0.35
0.45
0.55

Table 1. Cell lines information

Cell line name	CNV-Seq result
chHES488	46, XY, dup (2p25.3-2p16.2) 53.26 M, del (6q27) 4.77 M
chHES493	46, XY, dup (13q12.11) 2.47 M, dup (14q32.31-14q32.33) 3.32 M
chHES494	46, XX, dup (11q23.3-11q25) (18.25 M), dup (22q11.1-22q11.21) 3.19 M
chHES495	46, XX, del (3p26.3-3p26.1) 6.28 M, dup (5q23.2-5q35.3) 55.17 M
chHES496	46, XX, dup (7p22.3-7p11.2) 56.26 M, del (16p13.11-16p12.3) 3.60 M
chHES497	46, XX, del (1p36.33-1p36.32) (1.29 M), dup (16p13.3-16.13.11) 16.72 M
chHES90	46, XY

Table 2. Comparison of different library preparation protocols and sequencing platforms

Sample name	Platforms	Average Uniq_ data (M reads)	Average Read length (bp)	Average coverage (%)	Average GC bias	Average Mad of Diff
chHES493-MDA	Illumina Hiseq X10	4.77	150.00	14.62	0.32	0.23
chHES495-MDA	Illumina Hiseq X10	5.72	150.00	16.77	0.21	0.35
chHES497-MDA	Illumina Hiseq X10	4.97	150.00	13.31	0.15	0.34
MDA	**Illumina Hiseq X10**	**5.15**	**150.00**	**14.90**	**0.23**	**0.31**
chHES493-TTI	GenoCare 1600	5.72	42.14	8.03	0.02	0.23
chHES495-TTI	GenoCare 1600	5.22	41.93	7.30	0.03	0.24
chHES497-TTI	GenoCare 1600	5.35	42.10	7.51	0.04	0.20
TTI	**GenoCare 1600**	**5.43**	**42.06**	**7.61**	**0.03**	**0.22**