Uniform and accurate single-cell sequencing based on emulsion whole-genome amplification

Yusi Fu[a], Chunmei Li[a], Sijia Lu[b,1], Wenxiong Zhou[a], Fuchou Tang[a,c], X. Sunney Xie[a,b,2], and Yanyi Huang[a,c,d,2]

[a]Biodynamic Optical Imaging Center, School of Life Sciences, Peking University, Beijing 100871, China; [b]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138; [c]Peking–Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China; and [d]College of Engineering, Peking University, Beijing 100871, China

Contributed by X. Sunney Xie, July 28, 2015 (sent for review June 7, 2015; reviewed by Luke P. Lee)

Whole-genome amplification (WGA) for next-generation sequencing has seen wide applications in biology and medicine when characterization of the genome of a single cell is required. High uniformity and fidelity of WGA is needed to accurately determine genomic variations, such as copy number variations (CNVs) and single-nucleotide variations (SNVs). Prevailing WGA methods have been limited by fluctuation of the amplification yield along the genome, as well as false-positive and -negative errors for SNV identification. Here, we report emulsion WGA (eWGA) to overcome these problems. We divide single-cell genomic DNA into a large number (10^5) of picoliter aqueous droplets in oil. Containing only a few DNA fragments, each droplet is led to reach saturation of DNA amplification before demulsification such that the differences in amplification gain among the fragments are minimized. We demonstrate the proof-of-principle of eWGA with multiple displacement amplification (MDA), a popular WGA method. This easy-to-operate approach enables simultaneous detection of CNVs and SNVs in an individual human cell, exhibiting significantly improved amplification evenness and accuracy.

single cell | whole-genome amplification | sequencing | microfluidics | emulsion

Single-cell sequencing, characterization the genome of individual cells, is highly needed for studying scarce and/or precious cells, which are inaccessible for conventional bulk genome characterization, and for probing genomic variations of a heterogeneous population of cells (1–3). Recently single-cell genomics has unveiled unprecedented details of various biological processes, such as tumor evolution (4–6), embryonic development (7), and neural somatic mosaicism (8). Single-cell whole-genome amplification (WGA) is required to generate enough replicates of genomic DNAs for library preparation in conjunction with current sequencing protocols. Single-cell WGA has been increasingly used in cutting-edge clinical diagnostic applications such as molecular subtyping of single tumor cells (4, 9) and preimplantation genetic screening of in vitro fertilized embryos (10).

An ideal single-cell WGA method should have high uniformity and accuracy across the whole genome. The WGA uniformity is critical for copy number variation (CNV) detection, whereas the WGA accuracy is essential for avoiding single-nucleotide variation (SNV) detection errors, either false positives or false negatives. The false positives arise from misincorporation of wrong bases in the first few cycles of WGA. In a diploid human cell, the false negatives primarily arise from the allelic dropout (ADO), i.e., heterozygous mutations are mistaken as homozygous ones because of the lack of amplification in one of the two alleles (11).

Existing WGA chemistry includes degenerate oligonucleotide-primed PCR (DOP-PCR) (12), multiple displacement amplification (MDA) (13–17), and multiple annealing and looping-based amplification cycles (MALBAC) (4, 18, 19), which have successively achieved genome analysis at the single-cell level. DOP-PCR is based on PCR amplification of the fragments flanked by universal priming sites, and provides high accuracy for detecting CNVs in single cells but has low coverage and high false-positive and false-negative rates for calling SNVs (5). MDA has a much improved coverage but tends to have lower precision/sensitivity in CNV determination due to its variation of the amplification gain along the genome, not reproducible from cell to cell (20). By virtue of quasilinear amplification, MALBAC suppresses the random bias of amplification and exhibits reduced ADO rates, yielding low false negatives for SNV detection (2, 11, 18, 19). Notwithstanding its drawbacks, MDA still offers comparable or higher genome coverage than MALBAC, at least for single diploid cells, possibly taking advantage of the randomness (2). In fact, even higher coverage has been obtained for cells with aneuploidy, such as dividing cells (21), and cancer cells (22). MDA's main advantage is its lower false-positive rate for SNV detection on account of the use of Phi-29, a highly processive polymerase with high fidelity.

Microfluidic devices have been carried out for single-cell WGA (16, 20, 23, 24), allowing avoidance of contaminations and high-throughput analyses of multiple single cells in parallel. The small total reaction volumes (microliters to nanoliters- or picoliters) of the microfluidic devices not only facilitate the efficiency of reactions but also allow significant cost reduction for enzymes and regents used. It was reported that the nanoliter volume of a microfluidic device improved uniformity of the amplification compared with microliter devices in the WGA of single bacterial cells (20).

Here, we report a method, emulsion whole-genome amplification (eWGA), to use the small volume of aqueous droplets in oil to better the WGA chemistry for uniform amplification of a single

Significance

Uniform and accurate single-cell whole-genome amplification is important when starting material is limited and precious. We develop an emulsion-based amplification method that can suppress the amplification bias to detect high-resolution copy number variations of a single cell, and to simultaneously detect the single-nucleotide variations with high accuracy. This approach is compatible with various amplification protocols including the widely used multiple displacement amplification, which has been demonstrated in this paper.

Author contributions: Y.H. designed research; C.L. prepared sequencing libraries; Y.F., C.L., and W.Z. performed research; Y.F., S.L., F.T., X.S.X., and Y.H. analyzed data; and Y.F., S.L., F.T., X.S.X., and Y.H. wrote the paper.

Reviewers included: L.P.L., University of California, Berkeley.

Conflict of interest statement: S.L. and X.S.X. are cofounders and shareholders of Yikon Genomics.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this paper has been deposited in the NCBI Sequence Read Archive database (accession no. SRP052908).

[1]Present address: Yikon Genomics, Co., Ltd., Taizhou, Jiangsu 225300, China.

[2]To whom correspondence may be addressed. Email: xie@chemistry.harvard.edu or yanyi@pku.edu.cn.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1513988112/-/DCSupplemental.

cell's genome. By distributing single-cell genomic DNA fragments into a large number (10^5) of picoliter droplets, a few DNA fragments in each droplet is allowed to reach saturation of DNA amplification. After merging the droplets by demulsification, the differences in amplification gain among the DNA fragments are significantly minimized.

Although this approach can be used for any chemistry of WGA, we take MDA as an example to greatly reduce the random bias of amplification by separating the reactions into a large amount of emulsion droplets. We carried out detailed comparison with MDA, MALBAC, and DOP-PCR performed in tube using single cells from normal diploid human cells and a monoclonal human cancer cell line with inherited CNVs. Our results indicate that eWGA not only offers higher coverage but also enables simultaneous detection of SNVs and CNVs with higher accuracy and finer resolution, outperforming the prevailing single-cell amplification methods in many aspects.

Results and Discussion

eMDA Sequencing Library Preparation. MDA, an easy-operating and widely used single-cell WGA protocol, is used for the proof-of-concept of eWGA. We lysed individual cells to release the genomic DNA (gDNA) fragments and dehybridized them to single strands by heating. After adding the MDA reaction buffer, the solution (10 μL) was distributed into ~7×10^5 droplets, 14 pL each, using a microfluidic chip (Fig. 1 *A* and *B*, and *SI Appendix*, Fig. S1). This process is carried out at 4 °C to keep the amplification from starting. Under this lysing condition, the estimated mean size of DNA fragments is ~10 kb (18). Thus, for a single diploid cell, each droplet contains one fragment on average. We have tested different dilutions of DNA and observed the decline of mapping rate with further dilution, especially when the average fragment is far less than one per droplet. This is because a large number of empty droplets increases the ratio of nonspecific product of amplification. On the other hand, more DNA fragments in one droplet (>10 per droplet) impairs the evenness of WGA. The aneuploidy of the cell will affect the actual number of fragments per droplet. We found eMDA performance is stable when each droplet has one to two fragments. We collected all of the droplets in a microcentrifuge tube (Fig. 1 *C* and *D*). In contrast to the conventional single-tube MDA reaction, which exhibits more serious amplification bias with longer reaction time, in the emulsion MDA (eMDA) reaction each droplet produces similar amount of amplification products due to the eventual saturation of the polymerization reaction in each droplet (*SI Appendix*, Fig. S2). After heat inactivation of the enzyme and demulsification, the amplification uniformity is accomplished in the aqueous solution, and the amplification products are used to construct sequencing libraries.

eMDA Amplify Normal Diploid Single Cells Evenly and Completely. We chose human umbilical vein endothelial cell (HUVEC), a normal human diploid cell line, to validate the amplification evenness of eMDA using bulk (200 ng) genomic DNA from HUVECs as a reference. We carried out 10 single-cell eMDA experiments and compared the sequencing results with those of single-cell MALBAC or conventional MDA reactions. We divided the human genome into bins with mean size of 52.4 kb using dynamic binning method (5) and applied shallowly sequenced data, 3M uniquely mapped reads for each single cell, to calculate the copy number in each bin (Fig. 2*A* and *SI Appendix*, Figs. S3 and S4). eMDA showed the most uniform amplification across the whole genome, with coefficient of variation (CV) of 0.36, which is significantly lower than the conventional MDA (CV = 2.23) (Fig. 2*A*, Table 1, and *SI Appendix*, Fig. S5). From the reads covering autosomes and sex chromosomes of these single HUVECs (Fig. 2*B*), we found eMDA providing the smallest deviation from a priori expectation using bulk DNA as a reference.

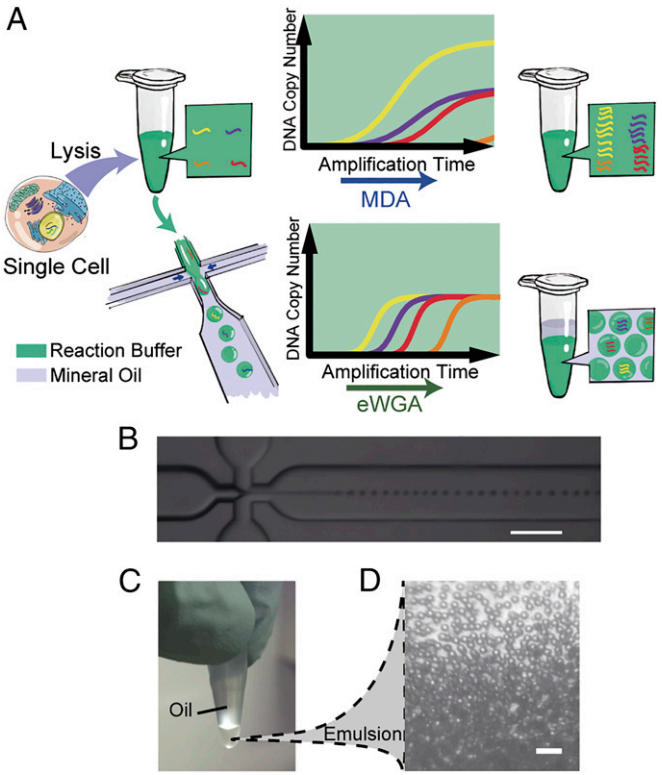


Fig. 1. The experimental process of eWGA-seq and emulsion generation. (*A*) A single cell is lysed and then mixed with MDA reaction buffer in a tube. The solution was either directly used for conventional MDA, generating unevenly amplified DNA fragments, or used for emulsion generation in a microfluidics cross-junction device, resulting in uniformly distributed aqueous reaction droplets and evenly amplified DNA fragments. (*B*) The microfluidics cross-junction. Reaction buffer and mineral oil are driven by compressed air with proper pressure to achieve uniform water-in-oil emulsion. The cross-section of the channel is 105 × 100 μm. The speed of emulsion generation is ~35,000 per min. (Scale bar: 300 μm.) (*C*) All droplets are collected into a 200-μL microcentrifuge tube and incubated at 30 °C to perform eWGA. (*D*) The emulsion is stable during the reaction. (Scale bar: 100 μm.)

We sequenced a few single HUVECs to a greater depth (>14×) using eMDA, MALBAC, and conventional MDA, and plotted the Lorenz curves of coverage to further validate the evenness of eMDA (Fig. 2*C*). As perfectly uniform coverage would result in a diagonal line, eMDA shows the best uniformity across the whole genome, compared with MALBAC and conventional MDA-amplified single cells, and is closest to the unamplified bulk sample. In contrast to the previously reported nanoliter MDA reaction in which the amplification gain is reduced (20, 23), our eMDA yields a similar gain as the conventional MDA to ensure a high coverage breadth of the genome. We showed that emulsion would not result in losing fragments of DNA as eMDA exhibits slightly higher coverage breadth (72.3% at 10×, for a human diploid cell) than MALBAC (67.5%) or conventional MDA (68.5%) at the same sequencing depth (*SI Appendix*, Figs. S6 and S7).

We also plotted the power spectra of read density as a function of the spatial frequency (Fig. 2*D*) based on the sequencing result using different protocols. The analysis confirmed that, for single-cell sequencing, eMDA provides the best uniformity among the three methods by offering smaller copy number fluctuation at all frequencies due to the effectively suppressed amplification bias through compartmentation. Because the intrinsic amplification randomness still exists within each droplet, the uniformity improvement is more significant in the lower frequency (large bin size) region than in the higher frequency domain.

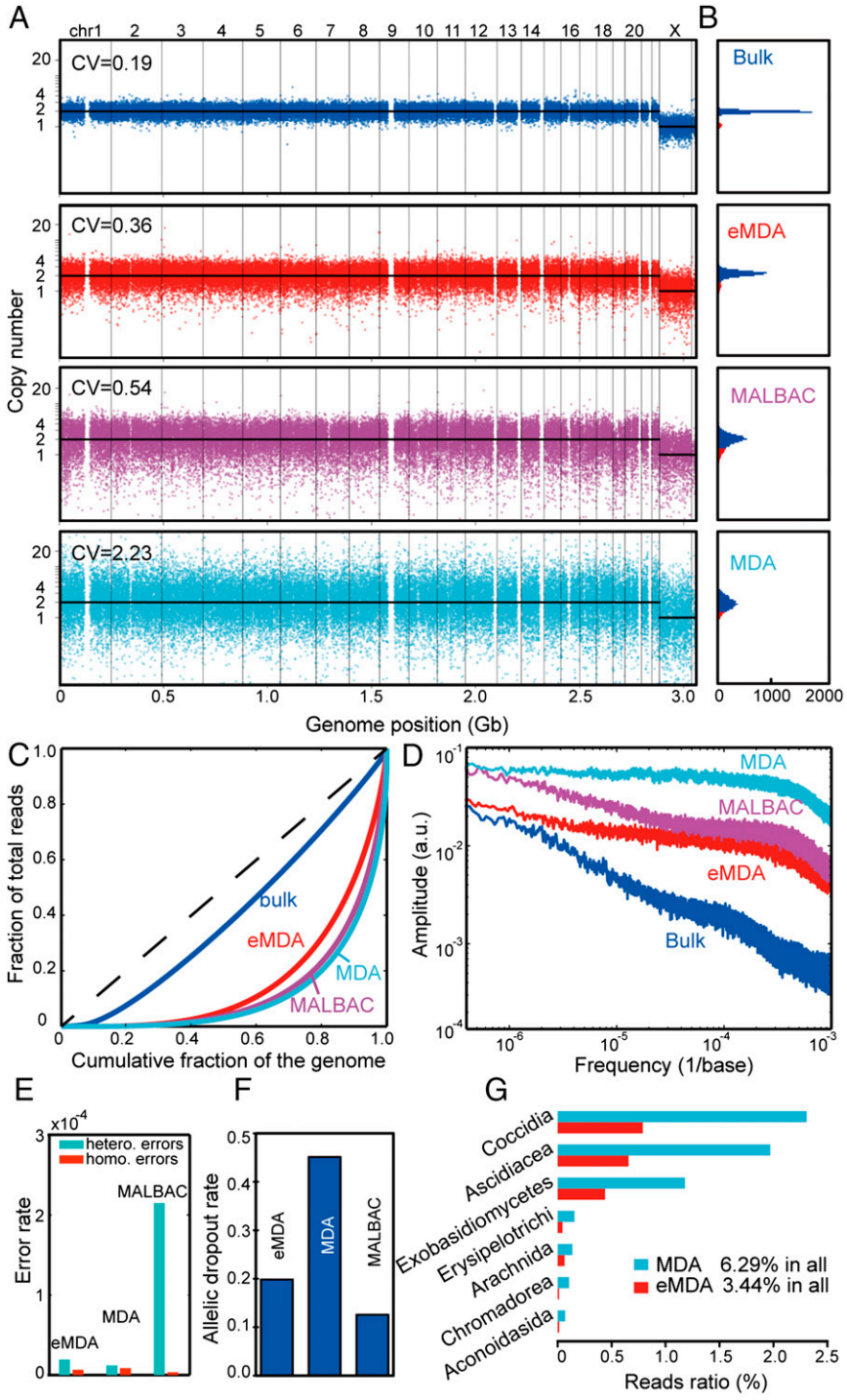


Fig. 2. The comparison of WGA methods for sequencing single HUVECs. (*A*) The copy number across the whole genome with a mean bin size of 52.4 kb; black line shows the expected value. (*B*) The density histogram of copy number distribution (bin size, 502 kb). (*C*) The Lorenz curves of coverage uniformity for single cells amplified by eMDA, MALBAC, conventional MDA, and unamplified genomic DNA. (*D*) The power spectrum of read density as a function of spatial frequency. (*E*) Copying-error rate of single-cell WGA methods. (*F*) ADO rate of single-cell WGA methods. (*G*) The ratio of the sequencing read originated from major pollutes in single-cell eMDA and conventional MDA experiments.

To estimate the accuracy of CNV identification of these methods, we carried out a simulation by calling the artificial CNVs with both copy number gain (2 to 3) and loss (2 to 1), in silico generated within diploid autosomes (*SI Appendix*, Fig. S8). The accuracy is the ratio of simulated CNVs that could be detected at the 52.4-kb resolution. eMDA shows much higher accuracy to identify the CNVs at the range from 300 kb to 2 Mb. We also performed an intersample correction (24) for MALBAC to eliminate the sequence-dependent bias (*SI Appendix*, Fig. S9), whereas for eMDA such normalization is unnecessary. This feature is very important in various medical applications such as in vitro fertilization preimplantation screening because a standard normalization sample and the expertise of performing complicated cross-sample normalization are often not available. eMDA was

Table 1. Summary of the comparison between different methods for single human cell amplification

Parameter	Amplification method			
	eMDA	MDA	MALBAC*	DOP-PCR
HT-29 single cells				
CNV resolution, kb[†]	619	9,669	5,847	538
CNV accuracy, %[‡]	66.5	50.9	41.3	78.8
False-positive rate, %[§,¶]	0.01	0.02	0.04	0.30
Coverage breadth, %[§]	90.3	74.4	78.8	43.7
SNV false-negative rate, %[§,#]	11.7	39.3	27.2	70.3
ADO rate, %[§,‖,**]	39.3	52.9	23.8	88.5
HUVEC single cells				
CV, 52.4k bin	0.45	2.23	0.55	—
CNV resolution, kb, gain/loss[†,††]	1,150/350	>2,000/1,200	>2,000/700	—
False-positive rate, %[¶,‡‡]	0.08	0.09	0.12	—
Coverage breadth, %[‡‡]	72.3	68.5	67.5	—
SNV false-negative rate, %[#]	29.5	30.6	34.6	—
ADO rate, %[‖,**]	19.8	45.1	12.6	—
Error rate, %[§§]	0.0019	0.0012	0.021	—

*The MALBAC data have been corrected through intersample normalization.
[†]The sensitivity cutoff is set to be 90%.
[‡]CNVs that are larger than 500 kb.
[§]Calculation based on exome-enriched sequencing data.
[¶]False-positive rate (FPR) is higher for the whole-genome sequencing data compared with exome sequencing data due to the relative lower coverage depth of some loci in whole-genome sequencing data.
[#]SNV false-negative rate is lower for HT-29 single cells than for HUVECs because of the hyperploidy of HT-29 cancer cells.
[‖]Calculation was based on the diploid region with coverage depth larger than 30×.
**Allelic dropout rate (ADO) calculated from exome sequencing data is larger than that from whole-genome sequencing data due to the material loss during capture.
[††]Calculation was based on simulation of copy number gains and losses ranging from 250 kb to 2 Mb.
[‡‡]Calculation was based on 30-Gb sequencing data of each method.
[§§]Calculation was based on reads mapped on chromosome X.

superior to both MALBAC and conventional MDA by offering finer smallest detectable CNV events (350 kb for copy number loss and 1.2 Mb for copy number gain, at 90% sensitivity in the diploid genomic region).

eMDA Amplifies Normal Diploid Single Cells with Higher Accuracy. From the deeply sequenced single-cell data, we detected more homozygous and heterozygous SNVs by eMDA than by MALBAC or conventional MDA (*SI Appendix*, Table S1), in accordance with the higher coverage breadth. As the HUVEC cells we used were from a male, we then deduced the error rates of these methods by calculating the ratio between high-confidence heterozygous SNVs and homozygous SNVs on the X chromosome from each dataset. The error rate of eMDA (1.9×10^{-5}) was comparable with that of conventional MDA (1.2×10^{-5}), but one order of magnitude less than that of MALBAC (2.1×10^{-4}) (Fig. 2*E*). These values, which matched well with previous reports (20), faithfully reflected the difference between the high fidelity Phi-29 polymerase used in eMDA and MDA, and the error-prone enzyme used in MALBAC which lacks proofreading capability.

We then examined the ADO rate of these methods by identifying the loss-of-heterozygosity events in the high-confidence heterogeneous SNVs (>20× coverage depth and >20% for each allele) found in autosomes from the bulk. For a normal diploid HUVEC, the ADO rate of eMDA is 19.8% (Fig. 2*F*). This performance is close to MALBAC, with which the ADO rate is ∼12%, making eMDA a great choice for those single-cell applications that could not be implemented by conventional MDA due to its notoriously high ADO rate (45.1%).

MDA is prone to environmental contamination including the trace amount of DNA pollution in reagents. The contaminant DNA could be reduced by applying small reaction volumes (20, 25). With eMDA, the reaction buffer is distributed to a large number of separated droplets, and the contaminant DNA will only exist in a small portion of droplets and not be overamplified. In addition, because the single human cells are carefully picked through micromanipulation under a microscope, and washed multiple times before lysing, the contamination from other mammalian cells is minimized. Metagenomic analysis (Fig. 2*G*) verified that eMDA produced much cleaner (3.4% nonhuman reads) data than MDA did (6.3% nonhuman reads) for single HUVEC sequencing.

High-Resolution Inherited CNV Detection in Single Cancer Cells. We next applied eMDA to sequence nine single HT-29 cancer cells expanded from a single clone. HT-29 is a colon adenocarcinoma cell line with multiple chromosomal aberrations, making its nuclear DNA close to triploid (26). We validated the aneuploidy through flow cytometry (*SI Appendix*, Fig. S10) and observed that the coverage depth pattern (27) of each single cell is similar to that of bulk (200 ng) gDNA (Fig. 3*A*). We called the CNVs from eMDA-amplified single cells at different resolutions, and found that the CNV pattern of each single cell is almost identical to that of the monoclonal expanded bulk sample, with correlation $r = 0.90 \pm 0.03$, 0.95 ± 0.02, and 0.96 ± 0.02 at 52.4-kb, 502-kb, and 5-Mb resolution, respectively. At the 52.4-kb resolution, we were able to identify CNVs with smallest size of ∼250 kb, which was the 5-bin cutoff we applied to the analysis (Fig. 3*B*). We also profiled CNV patterns of single cells amplified from MALBAC, DOP-PCR, or conventional MDA at 52.4-kb resolution (*SI Appendix*, Fig. S11) and found that, compared with MALBAC and conventional MDA, the improved amplification uniformity of eMDA allowed us to obtain more reliable genomewide CNV pattern (Fig. 3*C*) as well as the higher specificity and higher sensitivity of CNV identification in single cells, with performance close to DOP-PCR (Fig. 3*D* and *SI Appendix*, Fig. S12*A*).

Exome Coverage Breadth and SNV Detection in Single Cancer Cells. We then investigated the accuracy of SNV identification from single HT-29 cells using eMDA. We performed exome enrichment and sequencing for all samples and used bulk HT-29 exome as a reference. eMDA shows highest coverage (≥1× depth, 90 ± 5%), followed by MALBAC (79 ± 4%), conventional MDA (74 ± 11%), and DOP-PCR (44 ± 4%) (Fig. 3*E*). eMDA also exhibits high accuracy to identify homozygous SNVs of single cells, with highest true-positive ratio and lowest false-positive rate among all methods we tested (Fig. 3*F*). As expected, eMDA also noticeably reduce the ADO to 24% from 43% of conventional MDA for these nondiploid single cells (*SI Appendix*, Fig. S12*B*).

Conclusion

Our method, eWGA, applies emulsion to divide the DNA fragments from a single cell to a large number of aqueous droplets in oil and drives the amplification to saturation in each droplet. Using MDA protocol as a demonstration, this approach can dramatically reduce the amplification bias while retaining the high accuracy of replication. Unlike other microfluidics-based WGA methods (20, 23), which improved the uniformity by reducing

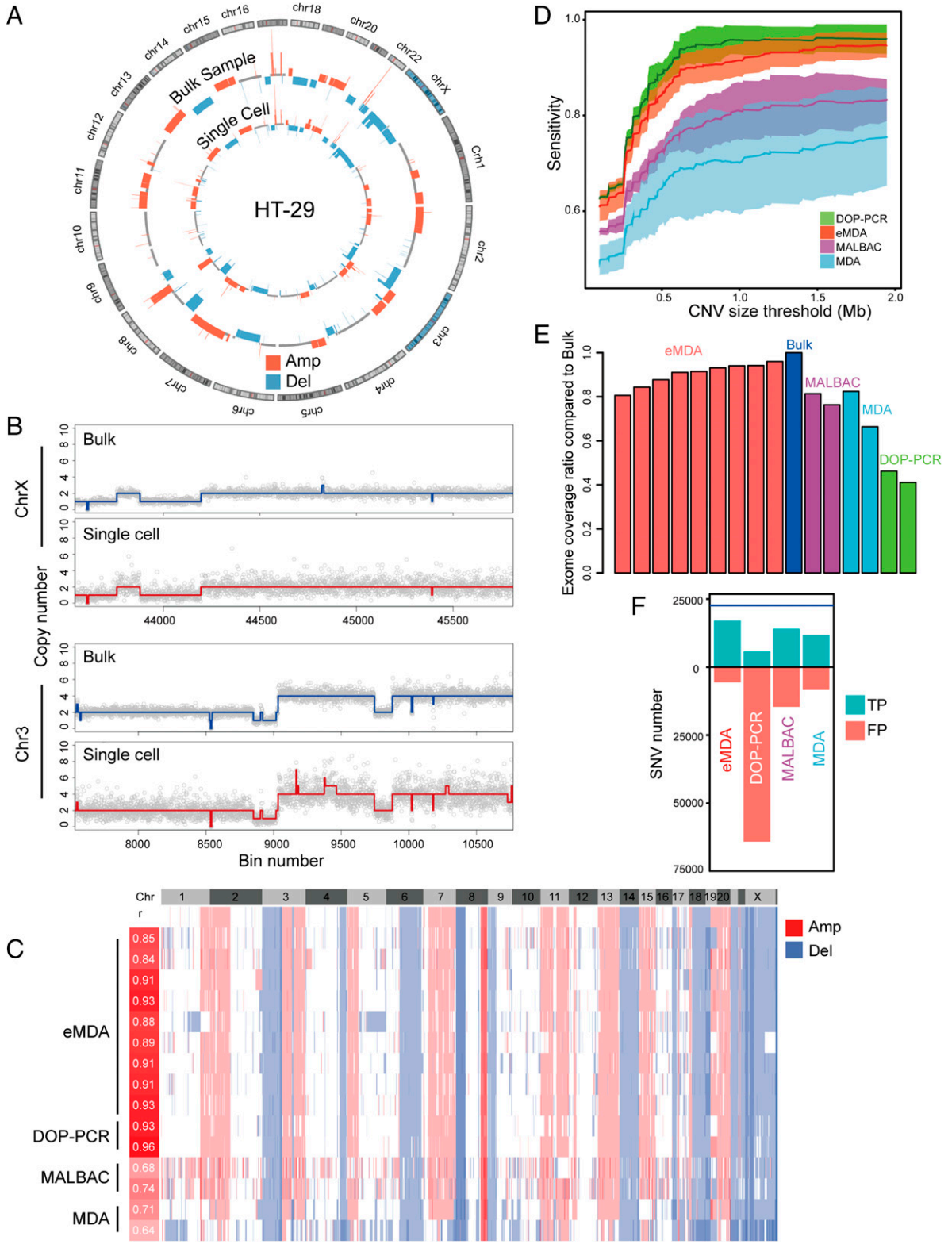


Fig. 3. The comparison of WGA methods for sequencing single HT-29 cells. (*A*) The circos plot (27) showing the copy number profiles from unamplified genomic DNA and from a single cell amplified by eMDA. (*B*) The zoomed-in copy number distribution of chr3 and chrX with a binning size of 52.4 kb. The smallest CNV detected is 5 bins. (*C*) Heat map showing copy number gains and losses of single cells with different amplification methods, with unamplified genomic DNA as reference. The correlation efficiencies between single-cell WGA methods and bulk reference are also listed. (*D*) The CNV detection sensitivity under different bin size threshold of single-cell WGA methods. The filled area represents the SD of each method. (*E*) The coverage ratio of exome captured single-cell WGA samples using unamplified sample as reference. (*F*) The homozygous SNVs detected in single cells using different WGA methods. The blue line shows the number of homozygous SNVs identified in the unamplified sample. The blue bars show the SNVs that matched bulk reference, whereas the red bars show the discordant SNVs.

the gain compared with conventional MDA, eMDA has the gain of $\sim 2 \times 10^6$, which is comparable to the conventional MDA in tube with single human cells as starting material. With the high coverage breadth across the whole genome, eMDA also enables us to detect more SNVs than existing methods and the pollution rate is alleviated with the use of emulsion. eMDA is compatible with targeted enrichment methods such as exome capture, which is useful when only certain regions are of interest in genetic analyses. By using eMDA, the first method (to our knowledge) that enables simultaneous identification of both small CNVs and high-confidence SNVs from a single human cell, we are able to detect CNVs at 250-kb size with 50-kb resolution, and SNVs with error rate $<2 \times 10^{-5}$. We envision that such emulsion approach will also improve the amplification performance of other WGA methods, for example MALBAC, for single-cell genomic studies.

Materials and Methods

Device Fabrication. Microfluidic emulsion-generating chips were made of polydimethylsiloxane (PDMS). The mold used to cast the chips was made by etching photoresist on a silicon wafer using photolithography. In brief, SU-8 2025 (MicroChem) was spin coated onto the wafer at 1700 rpm for 60 s on a spin coater (KW-4A, SETCAS Electronics Co., Ltd), resulting in a thickness of 50 μm of photoresist. Then the wafer was baked at 95 °C for 5 min. The wafer was exposed to UV light for 30 s through a mask defining the channel geometry and then the wafer was baked again at 95 °C for 10 min. The unexposed photoresist was removed with solvent and the wafer was hard-baked at 150 °C for 3 h. The mold was treated with trimethyl chlorosilane vapor for 10 min before use. Then 30 g degased and well-mixed 5:1 (base:curing agent) PDMS (Sylgard 184, Dow Corning) was poured on the wafer, and baked together at 80 °C for 15 min before peeled off. Then we punched the holes for the inlets of reagent/oil and the outlet for connecting a micro-tubing that transferred the emulsion droplets to a 200 μL micro-centrifuge tube. Then the patterned PDMS slab was bonded with a piece of cover glass precoated with 20:1 (base:curing agent) PDMS through baking at 80 °C for 3 h. The resulting chip is shown as *SI Appendix*, Fig. S1.

Conventional Single-Cell MDA Reaction. The gDNA was fragmented by heating (4 min at 98 °C, and 2 min at 95 °C) in 4 μL lysis buffer [30 mM Tris-HCl (pH = 8.0), 10 mM NaCl, 1 mg/mL proteinase (Qiagen), 5 mM EDTA and 0.5% Triton X-100]. Then 6 μL MDA reaction buffer was added to reach 10 μL total volume with a final concentration of 1x Phi-29 buffer (NEB), 50 μM N6 primer with two phosphorothioate bonds at the 3′-side (Invitrogen), 1 mM dNTP (NEB), 0.2 mg/mL BSA (NEB). We heated the tube at 95 °C for 5 min, and then immediately put it on ice for at least 20 min to anneal the random hexmers to fragmented gDNA. We then added 8 units of Phi-29 polymerase (NEB) and briefly centrifuged. Then MDA reactions were carried out at 30 °C. Reactions were terminated at 65 °C for 10 min after 10 h amplification.

Single-Cell eMDA Reaction. The reaction buffer preparation is identical to MDA reactions. However, to prevent the reaction from initiating prior to droplet generation, the Phi-29 polymerase was added to the reaction mix immediately before emulsion generation. The reaction buffer was kept at 4 °C to prevent the amplification from starting before being dispersed into droplets. The emulsion droplets were collected into a tube and then incubated at 30 °C for 8∼10 h before termination at 65 °C for 10 min.

A detailed description of remaining material and methods can be found in *SI Appendix*.

ACKNOWLEDGMENTS. We thank Dr. Ruiqiang Li, Dr. Zhilong Yu, Xiannian Zhang, Wei Fan, Ang Li, Zitian Chen, Tao Chen, and Zhe Su for their help on experiments and data analysis, Dr. Yun Zhang and the High-Throughput Sequencing Center of Peking University for assistance with Illumina sequencing, and the National Center for Protein Sciences Beijing (Peking University) for assistance with FACS analysis. This work was supported by National Natural Science Foundation of China Grants 21327808, 91313302, and 21222501 (to Y.H.).

1. Kalisky T, Quake SR (2011) Single-cell genomics. *Nat Methods* 8(4):311–314.
2. Huang L, Ma F, Chapman A, Lu S, Xie XS (2015) Single-cell whole-genome amplification and sequencing: Methodology and applications. *Annu Rev Genom Human Genet*, 10.1146/annurev-genom-090413-025352.
3. Wang Y, Navin NE (2015) Advances and applications of single-cell sequencing technologies. *Mol Cell* 58(4):598–609.
4. Ni X, et al. (2013) Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc Natl Acad Sci USA* 110(52):21083–21088.
5. Navin N, et al. (2011) Tumour evolution inferred by single-cell sequencing. *Nature* 472(7341):90–94.
6. Vogelstein B, et al. (2013) Cancer genome landscapes. *Science* 339(6127):1546–1558.
7. Yan L, et al. (2013) Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat Struct Mol Biol* 20(9):1131–1139.
8. McConnell MJ, et al. (2013) Mosaic copy number variation in human neurons. *Science* 342(6158):632–637.
9. Klein CA, et al. (2002) Combined transcriptome and genome analysis of single micrometastatic cells. *Nat Biotechnol* 20(4):387–392.
10. Hou Y, et al. (2013) Genome analyses of single human oocytes. *Cell* 155(7):1492–1506.
11. Navin NE (2014) Cancer genomics: One cell at a time. *Genome Biol* 15(8):452.
12. Telenius H, et al. (1992) Degenerate oligonucleotide-primed PCR: General amplification of target DNA by a single degenerate primer. *Genomics* 13(3):718–725.
13. Dean FB, Nelson JR, Giesler TL, Lasken RS (2001) Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* 11(6):1095–1099.
14. Zhang K, et al. (2006) Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* 24(6):680–686.
15. Chitsaz H, et al. (2011) Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat Biotechnol* 29(10):915–921.
16. Wang J, Fan HC, Behr B, Quake SR (2012) Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* 150(2):402–412.
17. Peters BA, et al. (2012) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 487(7406):190–195.
18. Zong C, Lu S, Chapman AR, Xie XS (2012) Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science* 338(6114):1622–1626.
19. Lu S, et al. (2012) Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science* 338(6114):1627–1630.
20. de Bourcy CFA, et al. (2014) A quantitative comparison of single-cell whole genome amplification methods. *PLoS One* 9(8):e105585.
21. Wang Y, et al. (2014) Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 512(7513):155–160.
22. Xu X, et al. (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* 148(5):886–895.
23. Gole J, et al. (2013) Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat Biotechnol* 31(12):1126–1132.
24. Yu Z, Lu S, Huang Y (2014) Microfluidic whole genome amplification device for single cell sequencing. *Anal Chem* 86(19):9386–9390.
25. Blainey PC, Quake SR (2011) Digital MDA for enumeration of total nucleic acid contamination. *Nucleic Acids Res* 39(4):e19.
26. Kawai K, et al. (2002) Comprehensive karyotyping of the HT-29 colon adenocarcinoma cell line. *Genes Chromosomes Cancer* 34(1):1–8.
27. Krzywinski M, et al. (2009) Circos: An information aesthetic for comparative genomics. *Genome Res* 19(9):1639–1645.

SI Appendix

Uniform and Accurate Single Cell Sequencing Based on Emulsion Whole Genome Amplification

Yusi Fu, Chunmei Li, Sijia Lu, Wenxiong Zhou, Fuchou Tang,

Xiaoliang Sunney Xie, and Yanyi Huang

Supporting Materials:

SI Materials and Methods

Figs. S1-S14

Tables S1-S4

SI Materials and methods

Cell culture

The primary human umbilical vein endothelial cells (HUVECs), kindly provide by Haidian Maternal & Child Health Hospital of Beijing, were originally isolated from umbilical cord. The HUVECs were cultured with Medium 199 (M199, Invitrogen) supplemented with 5% heat-inactivated fetal bovine serum (FBS, Invitrogen), 1% penicillin–streptomycin (PS, Invitrogen), 1% endothelial cell growth supplement (ECGS, ScienCell). HUVECs were cultured at 37 ^{o}C in a humidified incubator containing 5% CO_2. When HUVECs became confluent, they were washed 5 times with PBS to the flush out the dead cells, then detached by 0.25% trypsin with 0.1% EDTA (Invitrogen), and centrifuged at 1200 rpm for 3 min. Then the supernatant was discarded, and cells were resuspended. These cells can be used to continue culturing or to isolate single cells for downstream experiment. HUVECs with more than eight passages were discarded.

The HT-29 cells, expanded from a monoclone, were kindly provided by Professor Wensheng Wei in the School of Life Sciences at Peking University.

Preparation of single cells

We first diluted the cell suspension using PBS and pipetted gently to make cells disperse into single cells. Then we used a clean petri dish and dipped with pipet tips to form some ~10 μL clean PBS buffer drops to further dilute the cell suspensions. With only a few cells were in one drop, we used mouth pipet to pick a single cell from this PBS drop and to release to another clean PBS drop. Then we changed the mouth pipet tip and transferred the single cell to another clean PBS drop to wash the cell. We repeated this washing process for at least four times and each time we aspirated as few buffer with the single cell as possible. After washing, we used stereoscope to confirm that only one cell is in the PBS drop (***SI Appendix*, Fig. S13**). We changed the mouth pipet tip as often as possible to make sure only one cell was picked. We used a new mouth pipet to move this cell to 4 μL lysis buffer (30mM Tris-HCl (pH=8.0), 10mM NaCl, 1mg/mL proteinase (Qiagen), 5mM EDTA and 0.5% Triton X-100.). The success of transfer was monitored under stereoscope to ensure that there was no cell left in the PBS drop. Lysis was carried out at 50 ^{o}C for 3 h and followed by denaturing proteinase for 30 min at 70 ^{o}C. The single-cell lysate was stored at -80 ^{o}C.

MALBAC and DOP-PCR

MALBAC was performed using the Single Cell Whole Genome Amplification (WGA) kit (Yikon Genomics, Taizhou, China), and DOP-PCR was performed using GenomePlex Single Cell Whole Genome Amplification kit (WGA4, Sigma), following the protocol provided by the manufacturers.

Emulsion generation

Emulsion generation is in general very robust. We've tried both house-made devices and commercially available devices, with different recipes to generate stable emulsion droplets. For the experiments done with house-made PDMS microfluidic chips, compressed air was used to push the reaction buffer as well as the mineral oil (supplemented with 3% ABIL-EM90 and 0.1% Triton-X100) through a focus-flowing chip to generate w/o droplets ~30 μm diameter (**Fig. 1*B***). To keep the air pressure stable, two buffering bottles were placed before the inlet (***SI Appendix*, Fig. S1**). The reaction buffer was followed by 2 μL mineral oil

to ensure the complete recover of all the DNA-containing aqueous droplets into a 200-μL collecting tube. The reaction buffer in the inlet tube was placed between ice bags when generating droplets to keep the temperature low before gDNA fragments were dispersed into droplets. The whole process was carried out at 4˚C to prevent the amplification from starting and was completed within 20 mins .

For those experiments done with commercially available droplet generation devices (Dolomite, UK) we used different composition of oil (Mineral Oil supplemented with 4.5% Span80, 0.4% Tween80 and 0.05% Triton X-100). We did not see any noticeable difference between the experiments done with these two emulsification methods. PDMS chips were disposable and used only once. For commercial devices, we cleaned them with 200 μL ethanol and 400 μL water and repeated once to prevent carry over pollution between experiments.

Demulsification and DNA purification

After heat inactivation, the stability of emulsion was checked under microscope to make sure the droplets did not merge. Demulsification was done by adding 700 μL isobutanol with ~30 s fierce vortexing until the droplets disappeared and the solution was clear. Then 70 μL Binding Buffer (Zymo Research) was added, followed with centrifugation at 17,000 g for 3 min. The upper layer of the solution was discarded. The remaining solution was purified by DNA (PCR) Clean-up & Concentration kit (Zymo Research) following the manufacture's recommended protocol and was finally eluted with 10 μL water.

Quality control of single-cell whole genome amplification

After purification the amplified DNA was first quantified by Qubit dsDNA HS Assay (Invitrogen) and then the amplification bias was briefly examined through quantitate PCR (***SI Appendix*, Fig. S14**). qPCR was set up by adding 5 μl PCR SsoAdvanced SYBR Green Supermix (Biorad), 0.5 μL of 10 μM forward- and reverse-primer (***SI Appendix*, Table S2**), 1 μL template and 3 μL water. Then qPCR was carried out at Illumina Eco thermocycler with melting curve analysis. The melting curves were examined to make sure the result was not due to the formation of primer dimers. Blank experiment was added to make sure the Ct

value was not from the polluted reagents. We also used 1 ng purified genomic DNA as positive control.

Preparation of sequencing libraries

For each amplified sample, 100 ng DNA was used to build the sequencing library for Illumina platform using NEBNext Ultra DNA Library Prep Kit (NEB). The libraries were sequenced by Illumina Hiseq 2500 or MiSeq sequencers. The summary of the sequencing data of all samples and the corresponding average depth and coverage are listed in ***SI Appendix*, Table S3**.

Exome capture

Exome capture was performed on all the single cell experiments and unamplified bulk sample of HT-29 cells. We used SureSelectXT Human All Exon V5 (Agilent Technologies) to capture ~50M coding regions by pooling 4 sequencing libraries with different index together in each run. The summary of sequencing result is listed in ***SI Appendix*, Table S4**.

FACS to determinate the ploidy of HT-29

We use the fluorescence-activated cell sorter (FACS) to determine the ploidy number of HT-29 cell. The human fibroblast, which is a normal diploid cell line, was used as a reference. We placed ~$2x10^6$ cells in a tube and centrifuge at 2,000 rpm for 5 min, then the supernatant was removed and 1 mL HT-29 cell culture medium (Dulbecco's Modified Eagle Medium (DMEM, Invitrogen) supplemented with 10% heat-inactivated fetal bovine serum (FBS, Invitrogen), 1% penicillin–streptomycin (PS, Invitrogen)) was added. Then we prepared three tubes of cells with 10^6 fibroblasts, 10^6 HT-29 cells, and $5x10^5$ fibroblasts mixed with $5x10^5$ HT-29 cells. We span down the cells to replace the supernatant with 500μL pre-warmed medium containing 1:1000 Hoechest 33342 (10mg/ml, Invitrogen) in each tube, and incubated at 37 ^oC for 30 min to stain the cells, followed by PBS washing twice. Finally the cells were suspended in 500μL 1% FBS-PBS and loaded into a FACS machine (BD Aria SORP) to measure the fluorescent intensity, which is corresponding to the DNA content of each cell for ploidy calculation.

Quality control and alignment of sequencing reads

The raw images were converted to sequences in fastq format using the RTA v1.9 and CASAVA v1.8.2. We kept the high-quality reads and discarded the reads with adaptors, or with too many undermined bases, or with too low quality (quality value $\leqslant$5). Then the filtered reads were mapped to the human assembly US National Center for Biotechnology Information (NCBI) bulild 37(hg19) using Bowtie2 (1). The alignment SAM format results were converted to BAM format via samtools (2) and sorted by chromosome coordinates.

CNV identification

The CNV identification was mainly based on the protocol published by Baslan et al (3). In brief, a fastq file was generated throughout the genome, with each read was 1bp base apart and with a length of 100bp. Then all the reads were remapped to the reference genome using Bowtie2 with the default setting. If the read was mapped to the same position it was generated and without the XS tag reported by Bowtie2, the position was regarded as a unique base. Continuous unique bases were converted to regions and dynamic binning was used to calculate the bin boundaries. Then the uniquely mapped reads number in each bin was calculated for each sample. CBS segmentor was used to find non-overlapping regions of differing copy number, then the copy number of each region was calculated.

CNV simulation for normal diploid cell

We used the data from the autosomes of diploid cell and randomly picked 100 non-overlapping bins and simulate CNV for these bins. Then we used this data to identify CNV for each sample and calculated the ratio of the simulated CNV that could be detected. Different bin-sizes from 300 kb to 2 Mb were used. The whole process was repeated several times to give a more accurate assessment for each method. For MALBAC, we used the GC corrected and cross sample normalized data.

Genome coverage

The relationship between genome coverage breadth and the sequencing depth

was calculated from unsorted SAM file. We firstly calculated the bases that have been covered by each uniquely mapped reads. Then the genome coverage and the number of read are calculated. Since the result could be affected by the read length, when comparing with published data (***SI Appendix*, Fig. S7**) the read length of our data was trimmed to be the same with data published by others. Otherwise, 100bp paired-end reads were used.

Identification of SNV, ADO rate, detection efficiency, error rate, true positive SNVs and false positive SNVs

Samtools mpileup was used to identify SNVs for sorted bam, and bcftools was used to generate the VCF file with the default setting. We used the bulk sample as reference and standard to calculate the detection efficiency.

Heterozygous SNVs were picked from bulk sample, filtered with at least 20X sequencing depth, and the smaller allele fraction should cover at least 20% of all the reads. Then for each single-cell sample we compared these SNVs. If at one site, enough reads covered that position, both kind of allele must exist and occupy more than 5% of all the reads covered that position. If not, a loss of heterozygous happened. The number of loss of heterozygous/all the sites detected with enough depth was computed as ADO rate.

Error rate was calculated using male HUVEC with the SNVs at the single copy of X chromosome. The heterozygous SNVs identified on X chromosome were regarded as errors. To accurately detect the error rate, we first filtered out all the sites that have an insertion or deletion within ±100 bp range. The error heterozygous site was defined as position that both the reference base and alteration base were supported by forward and reverse reads cover this site. Also the base with smaller fraction should be supported by at least 20% of all the reads aligned to this site. The high confidence homozygous position was counted if more than 10 reads covered both in the sample calculated and the unamplified sample. Then the heterozygous position/homozygous position is defined as error rate.

For HT-29, homozygous SNVs across the HT-29 genome are called from single cells amplified from different methods. The results are compared to the unamplified sample. If the SNV exists in the unamplified sample, it is regarded

as true positive SNVs, if not, it is regarded as false positives.

The detailed comparison between different single-cell WGA methods are listed in **Table 1**.

GC correction and MALBAC cross-sample normalization

In single-cell MALBAC samples, we found the amplification bias is highly correlated with GC-content of the genome, while this correlation was not found in DOP-PCR, MDA, or eMDA amplified samples. The MALBAC bias pattern is reproducible between samples. After GC correction by LOWESS smoothing, the MALBAC produced a smaller bias with a reduced CV (***SI Appendix*, Fig. S9**). Also the cross-sample normalization was applied by divided the read number in each bin of one sample to another to further improve the evenness (4). The X and Y chromosomes were excluded for all the CV calculation in this paper.

References

1. Langmead B & Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Meth* (9):357–359.
2. Li H *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* (25):2078–2079.
3. Baslan T *et al.* (2012) Genome-wide copy number analysis of single cells. *Nat Protoc* (7): 1024–1041.
4. Peters BA *et al.* (2013) Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* (487):190–195.

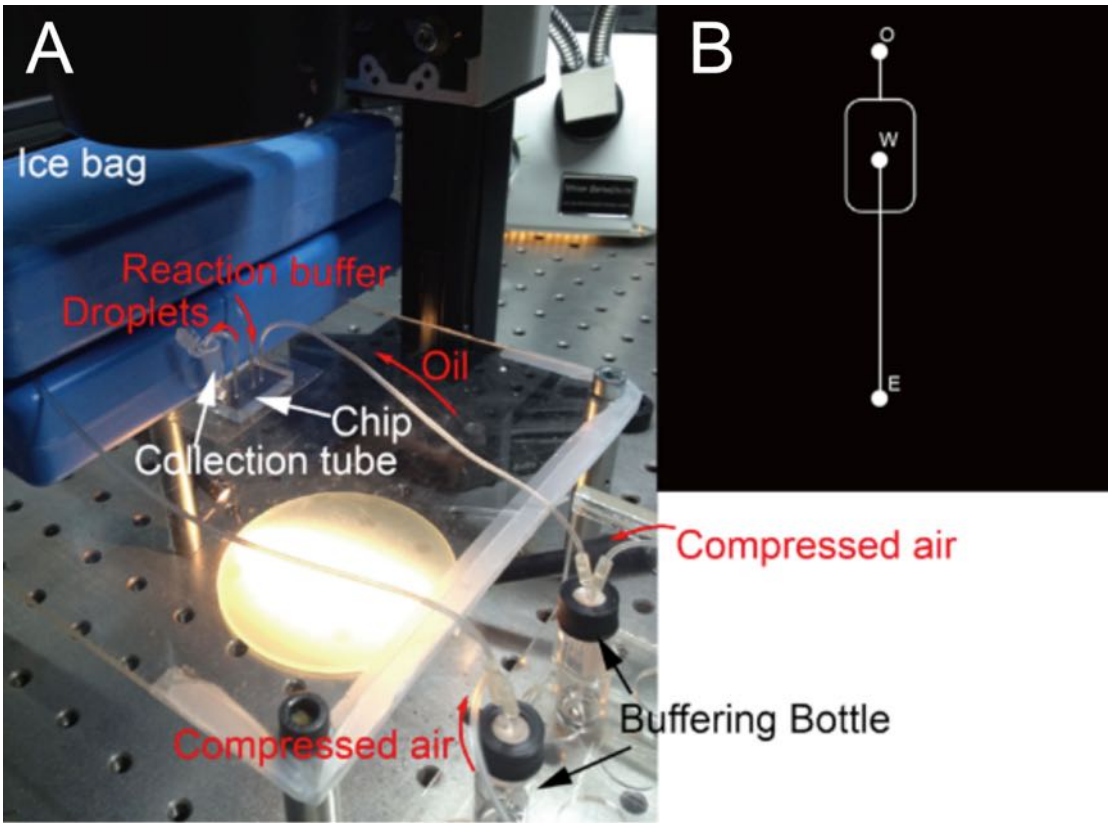


Fig. S1. The experimental setup for droplet generation. (*A*) Compressed air, stabilized by buffering bottles, was used to drive both mineral oil and the reaction buffer into a PDMS chip to generate emulsion droplets. A short segment of the collection tube was connected to the outlet port of the chip to collect the emulsion droplets for downstream experiments. The reaction buffer was sandwiched between two ice bags to prevent the reaction from initiating before droplets formation. The whole process was monitored and recorded under a stereoscope. (*B*) The layout of the microfluidics chip for emulsion generation. O: oil input. W: reaction buffer input. E: outlet port for emulsion collection. The channel width: 50 μm.

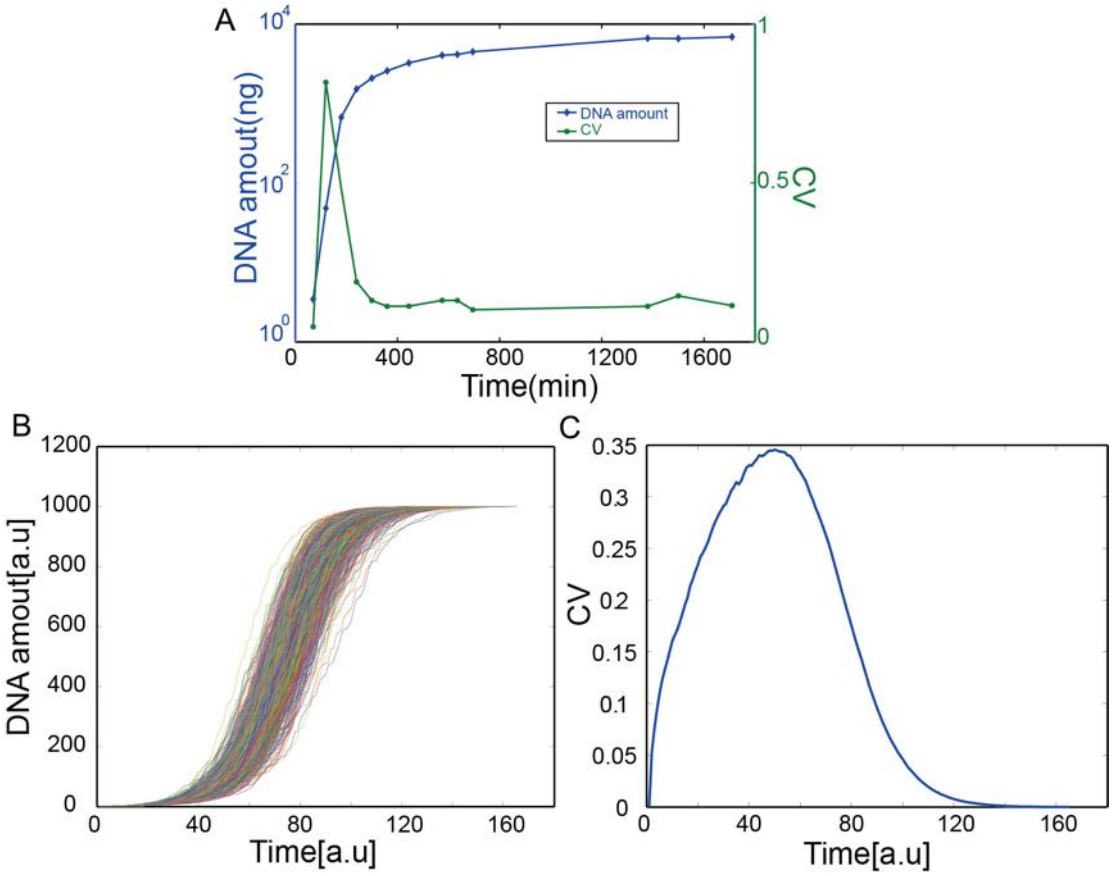
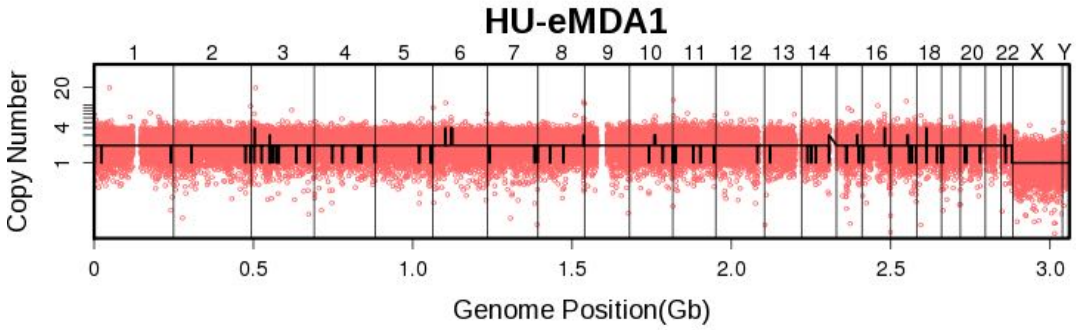
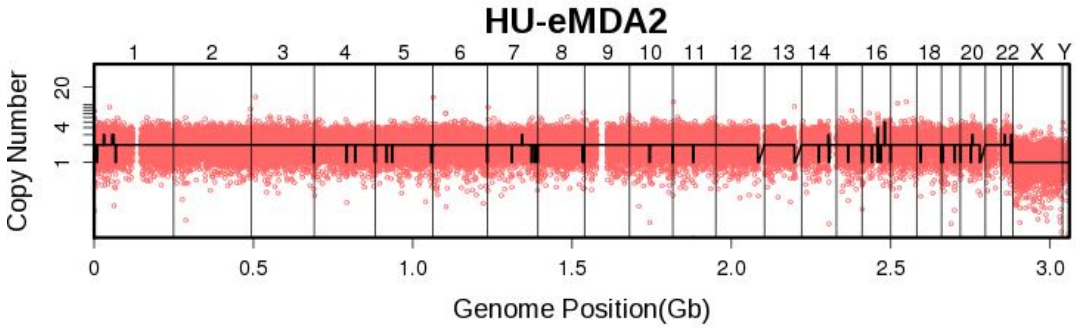
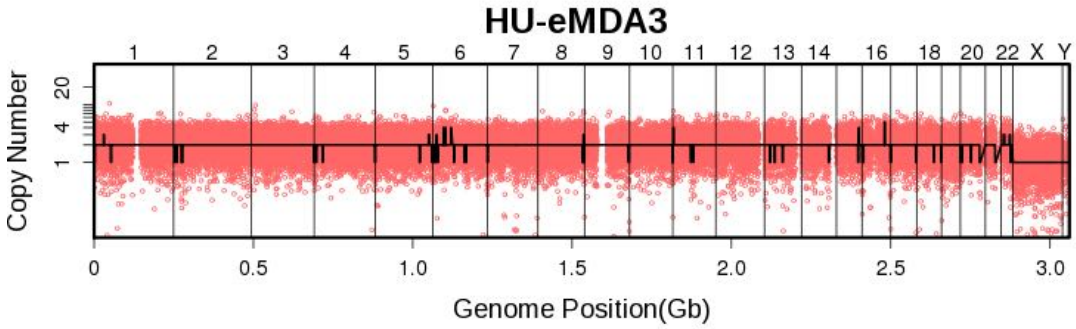
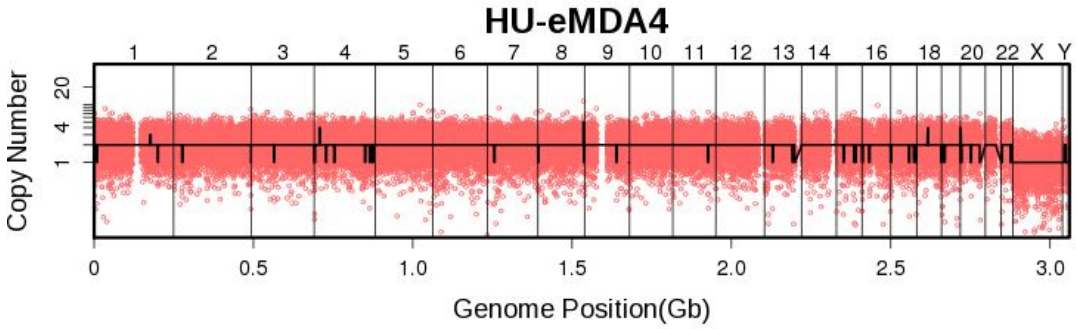


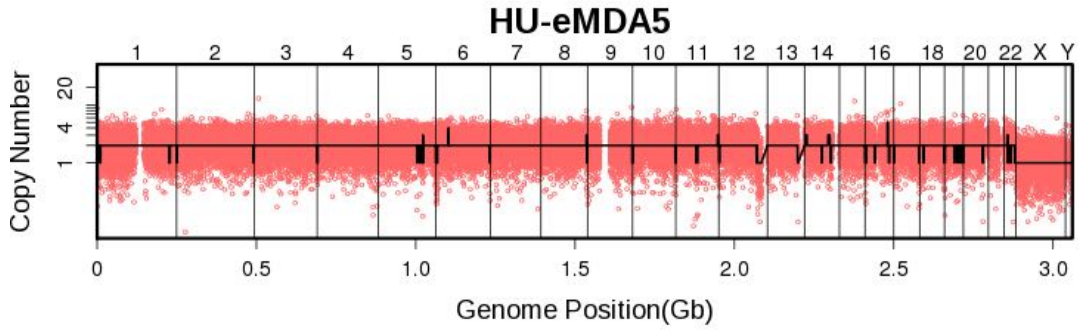
Fig. S2. Amplification bias reduction through reaction saturation. (*A*) The coefficient of variation (CV) of the total amount of the DNA produced by five separate MDA reactions starting with 6 pg DNA throughout the amplification process. As the reagent consumed, the reaction slowed down; the CV was first increased and then reduced to a stable value after the reaction reached near saturation (~8h). (*B*) The simulation result of the amplification yield of independent MDA reactions as function of reaction. We added random amplification velocity to each reaction. (*C*) The CV of the simulated individual MDA reactions. The CV first rises and then falls with the extension of reaction time, in accordance with the experimental observation.

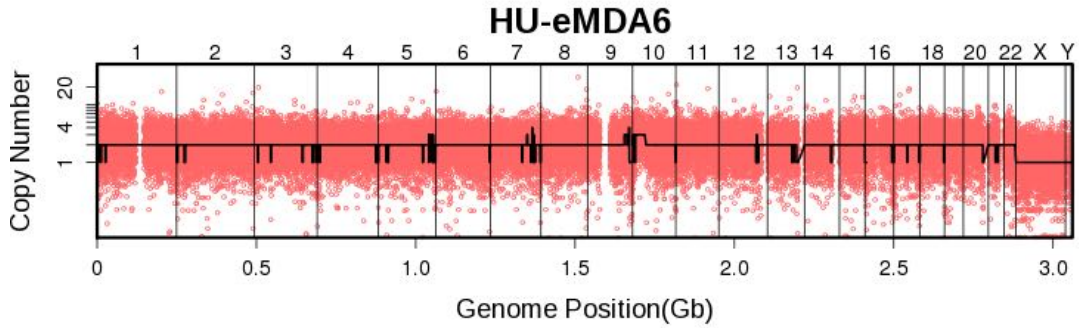

HU-eMDA1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

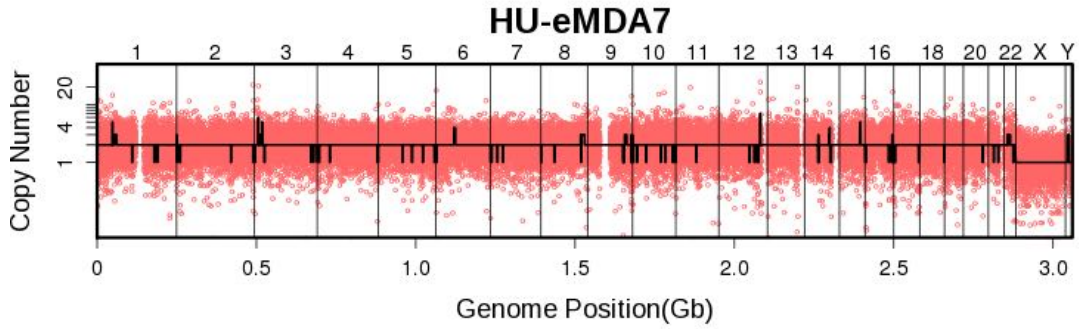

HU-eMDA2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

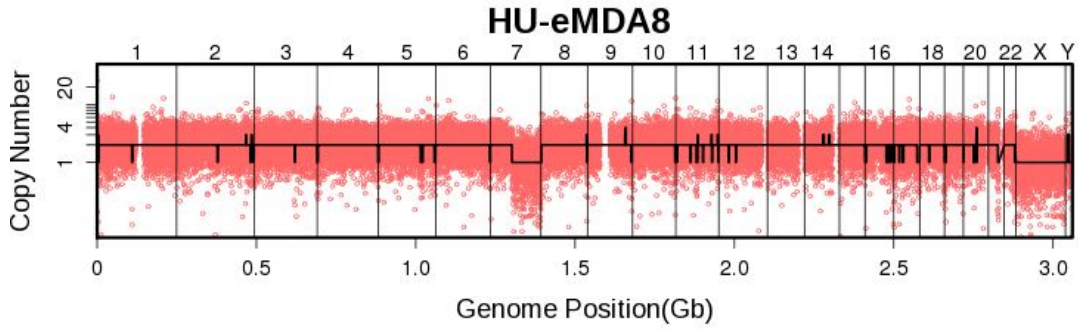

HU-eMDA3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

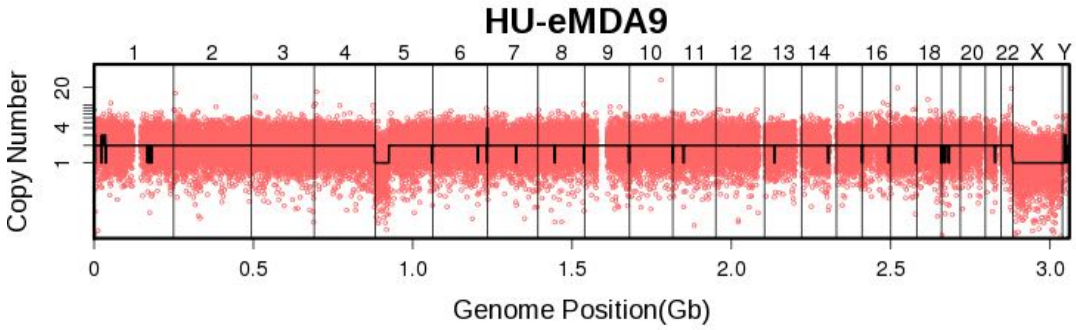

HU-eMDA4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

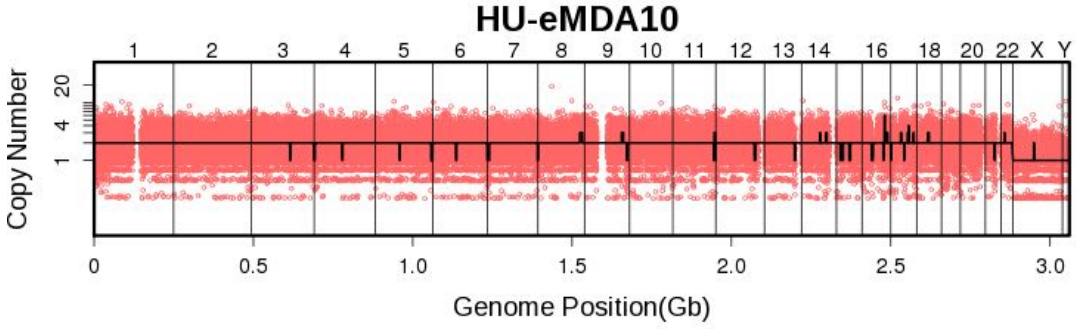

HU-eMDA5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

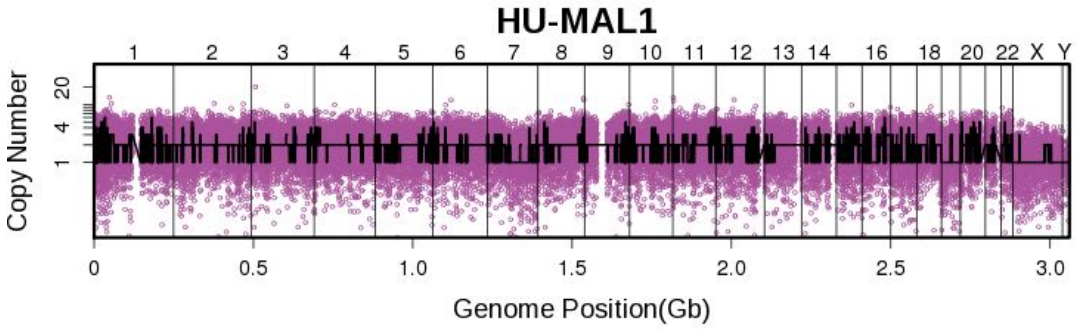

HU-eMDA6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

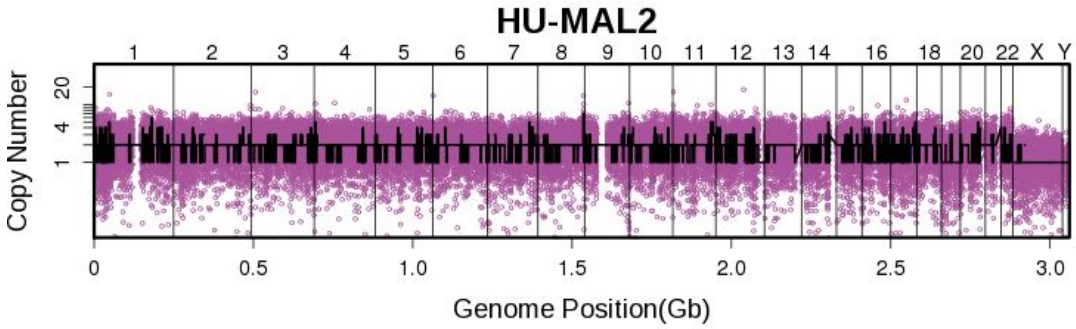

HU-eMDA7
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)


HU-eMDA8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)


HU-eMDA9
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)


HU-eMDA10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)


HU-MAL1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)


HU-MAL2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

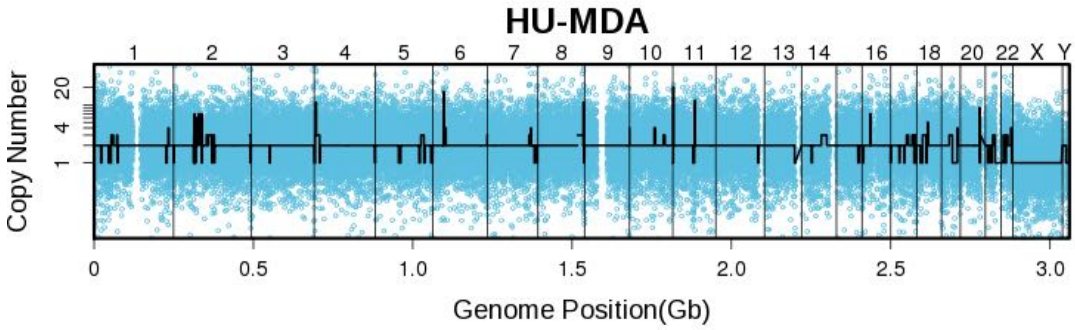


Fig. S3. The copy number distribution of all single HUVEC WGA by MALBAC, eMDA and MDA. The reads are clustered into 52.4-kb (mean size) bins by dynamic binning. The black lines represent the copy number determined by CBS segmentation method.

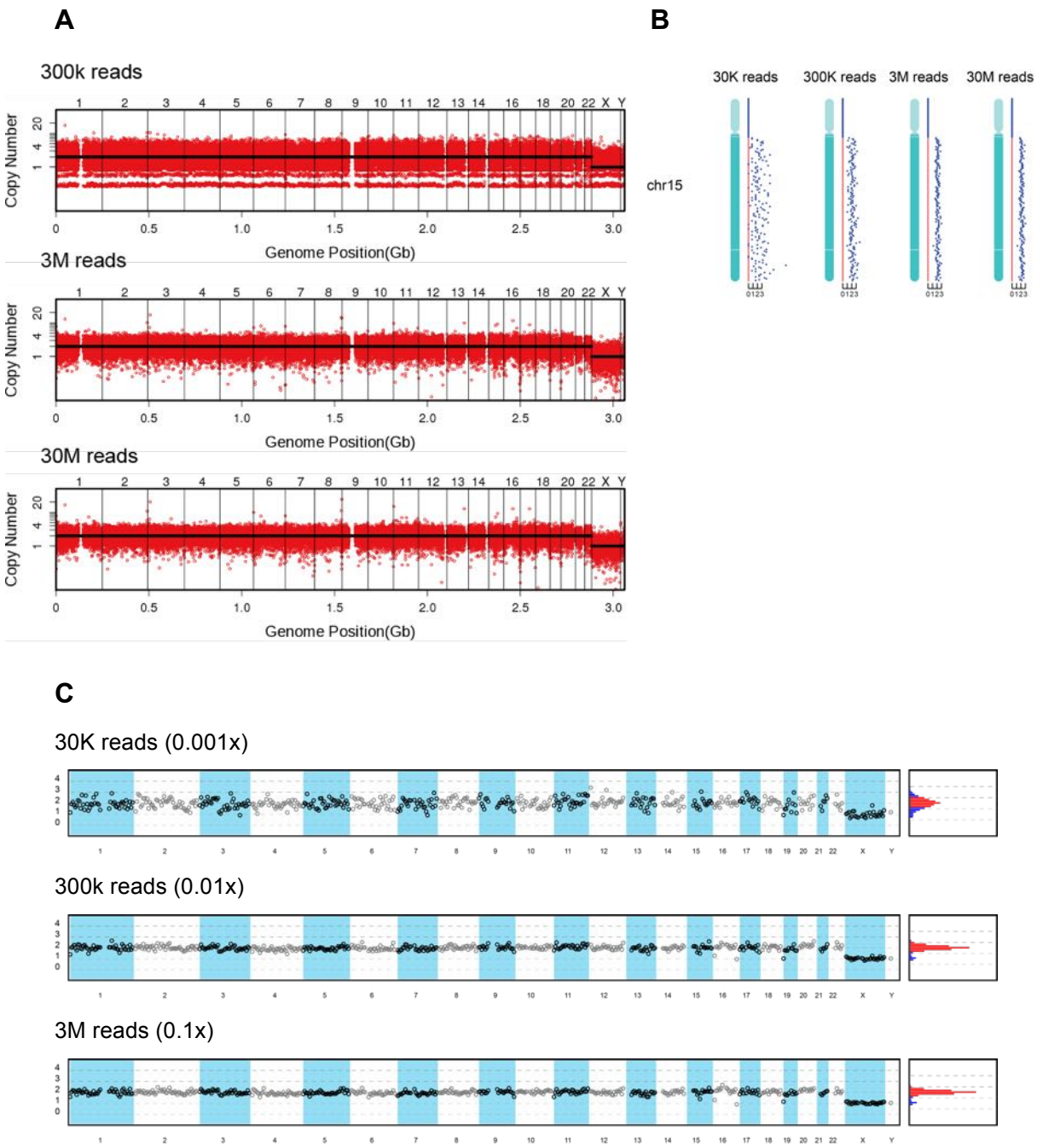


Fig. S4. Influence of eMDA sequencing depth on the coverage distribution. Different sequencing depth was applied to calculate the copy number of each bin for (*A*) 50-kb, (*B*) 500-kb, and (*C*) 5M bin-size using the HU-eMDA1 data. For 50kb-bin and 500kb-bin size analysis at least 3 million reads were needed to accurately identify the CNVs, while for 5M-bin size analysis at least 300,000 reads were needed.

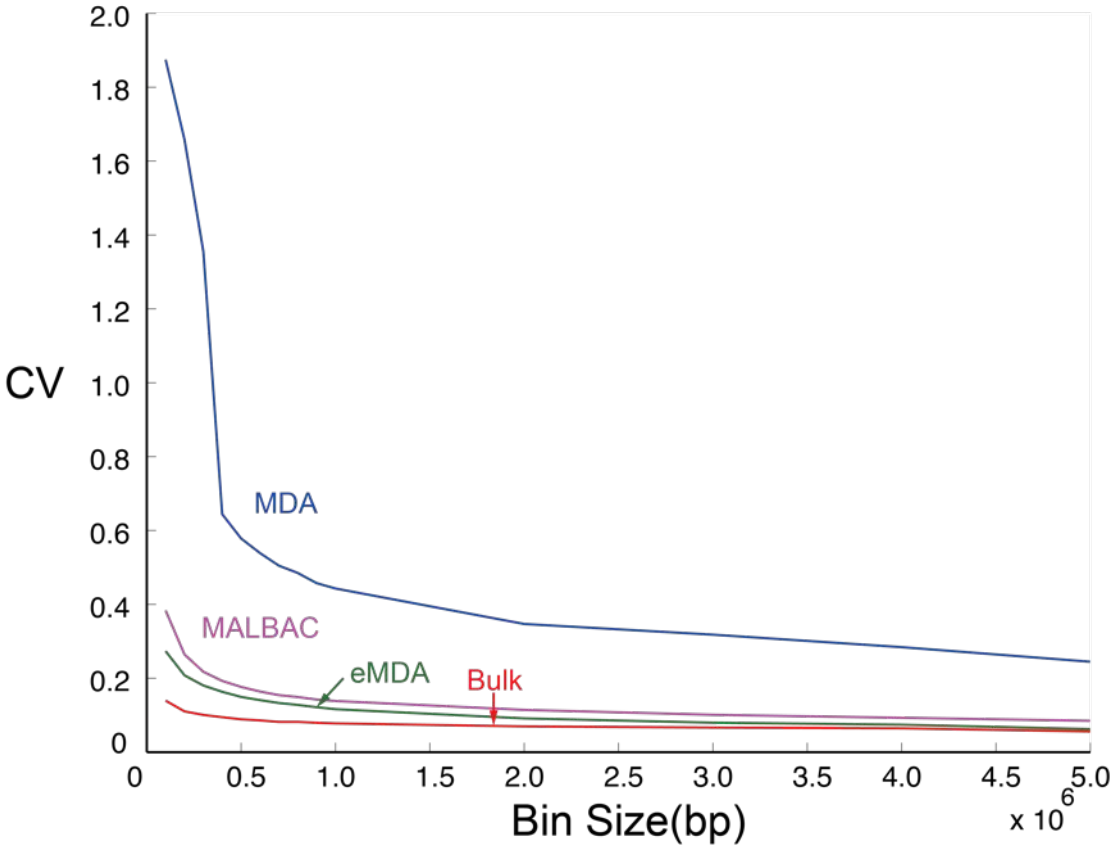


Fig. S5. The relationship between the binning-size and the coefficient of variation (CV) of the sequencing coverage from single cells autosomes. The genome was binning into different sizes, and the CV of copy number distribution of sequencing data calculated for various methods. Fixed bin-size binning method is used here. All ±2 bins around the reference sequence's gaps were deleted. The MALBAC data shown here is the cross-sample normalized data between the two experiments.

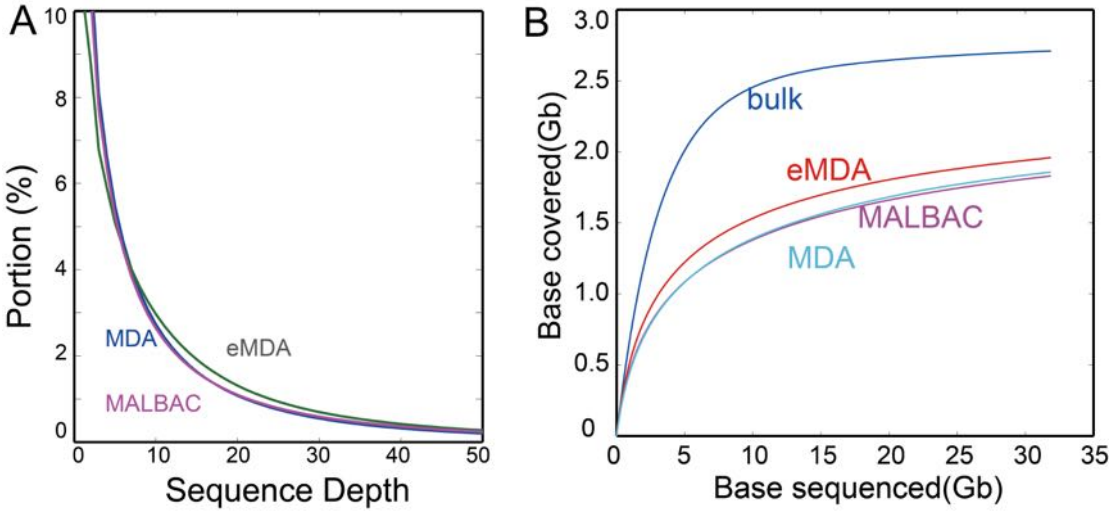


Fig. S6. Distribution of the sequencing coverage depth and the sequencing coverage breadth. (*A*) The distribution of single-base sequencing depth with different single-cell WGA approaches. All methods have obvious bias towards the low-coverage bases while eMDA (data calculated from HU-eMDA1) shows noticeable improvement over the conventional MDA reaction and MALBAC by generating more reads in the deeper-sequenced regions. (*B*) Sequencing coverage breadth as a function of sequencing coverage depth. For a given single haploid cell (HUVEC), eMDA has higher coverage over conventional MDA and MALBAC.

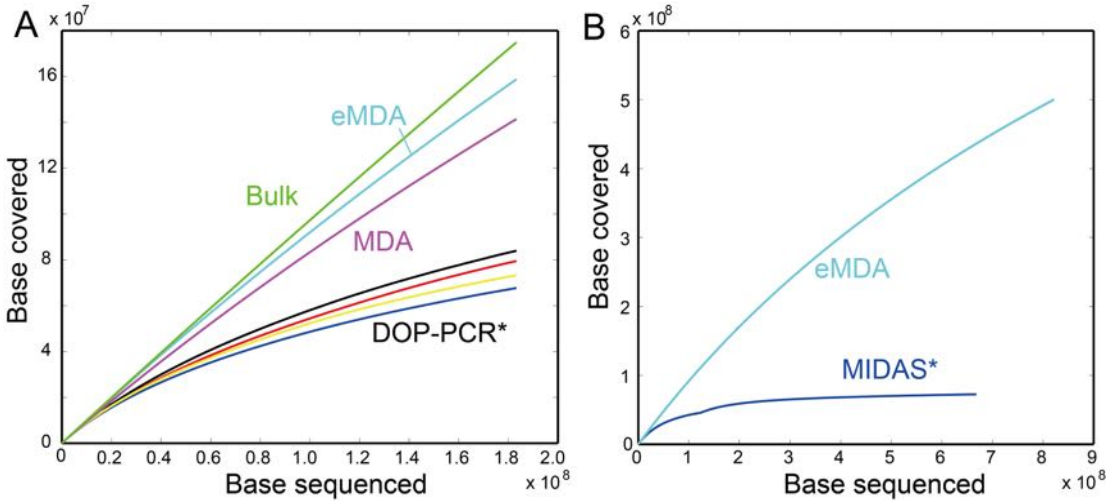


Fig. S7. The relationship between coverage breadth and depth. The read length of eMDA is trimmed to be comparable to previously published data (48bp or 36bp). For all single-cell WGA methods, the coverage breadth is continuously and smoothly increased with deeper sequencing depth. (*A*) The comparison of single cell eMDA, conventional MDA, and the data published using DOP-PCR[1]. (*B*) The comparison between single cell eMDA and MIDAS, a volume-confined MDA method[2]. DOP-PCR has a tendency to saturate on coverage breadth at a relatively shallow depth (0.1x) and covers ~10% of the whole genome. With a small volume and decreased amplification yield, MIDAS did not inherit the relatively high coverage of conventional MDA.

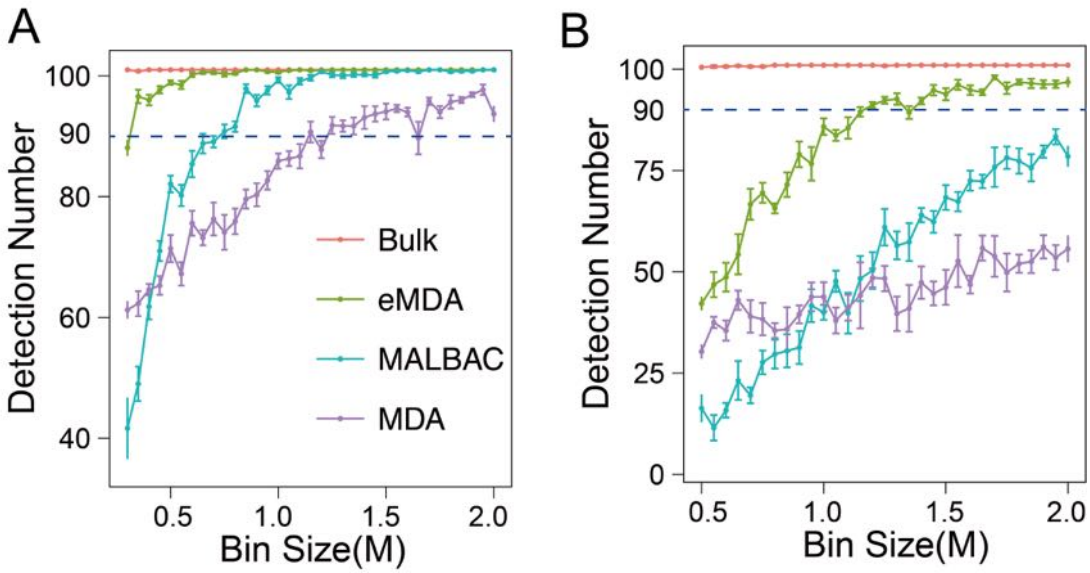


Fig. S8. The detection sensitivity of simulated CNVs. (*A*) The detection sensitivity of simulated copy number losses. We simulated 100 copy number changes from 2 to 1 in autosomes. MALBAC data used here is GC corrected and cross-sample normalized. eMDA shows better detection sensitivity of smaller CNVs than MALBAC and conventional MDA at the 300 kb to 2 Mb bin-size range (smaller bin-size is not reliable for MALBAC and MDA experiment). (*B*) The detection sensitivity of simulated copy number gains. We simulated 100 copy number changes from 2 to 3 within autosomes. The sensitivity for detection copy number gains is lower than for detection of copy number losses. We also observed that GC-correction and cross-sample normalization is less powerful when the bin-size is small. The dash line shows the 90% sensitivity cut-off.

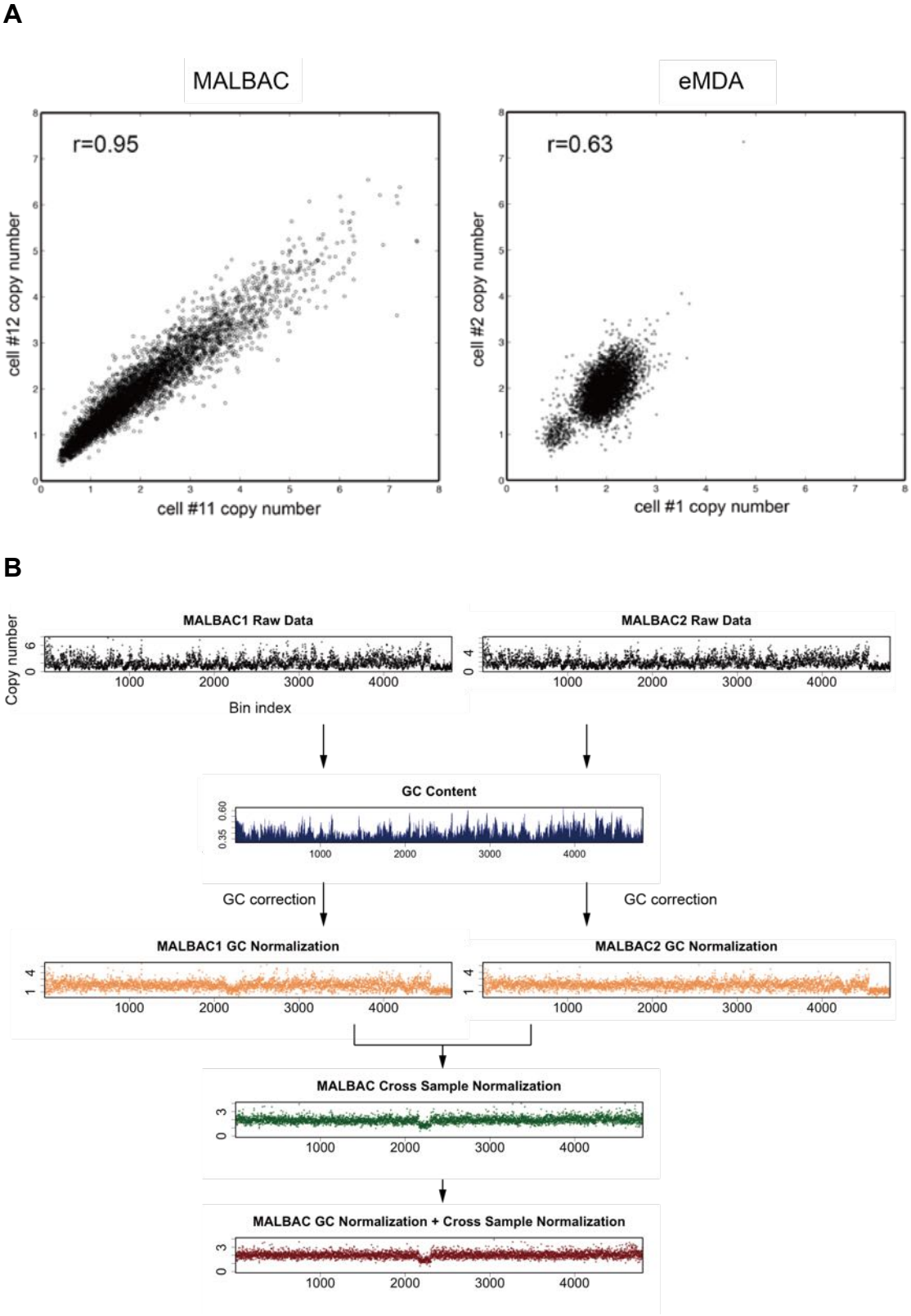


Fig. S9. The GC correction and cross-sample normalization to reduce the systematic bias generated by single-cell MALBAC. (*A*) The coverage depth in

each bin of single cell sequencing data through MALBAC is biased. The copy number of each bin from two duplicate experiments of MALBAC (left) and eMDA (right) are plotted against each other. MALBAC shows a strong correlation between biological duplicates (Pearson r=0.95) while eMDA shows much weaker correlation between samples (Pearson r=0.60), indicating the amplification bias in MALBAC is sequence-dependent while the bias in eMDA is less sequence-dependent. (*B*) The normalization process of the MALBAC data. The sequence depth pattern of single cell MALBAC is correlated with GC content throughout the genome. The amplification bias is corrected by LOWESS GC smoothing. After cross-sample normalization, the unevenness of amplification has been further reduced, allowing us to observe a clear deletion in HU-MAL1. The binning method used here is dynamic binning with a mean bin-size of 502 kb.

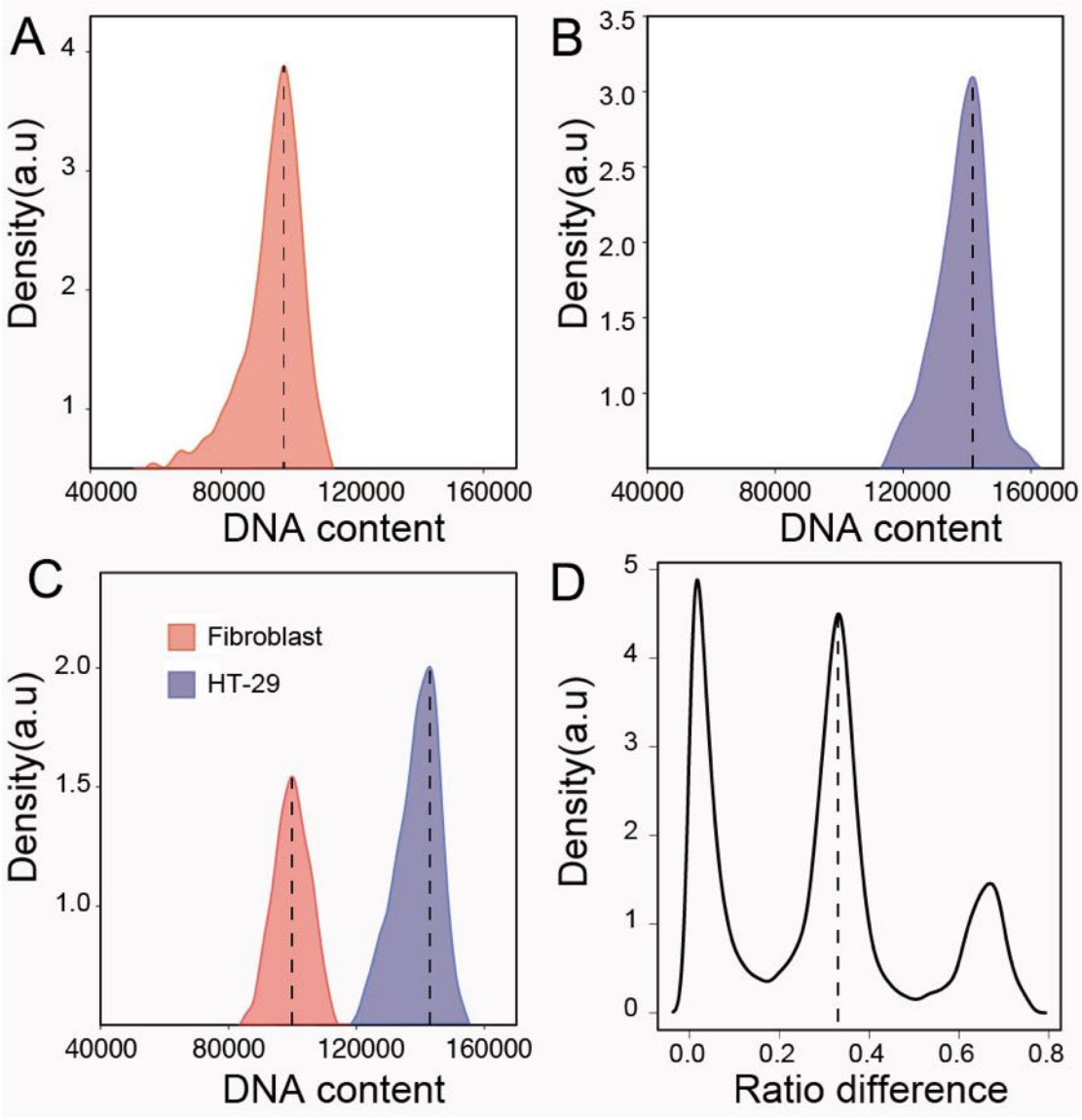
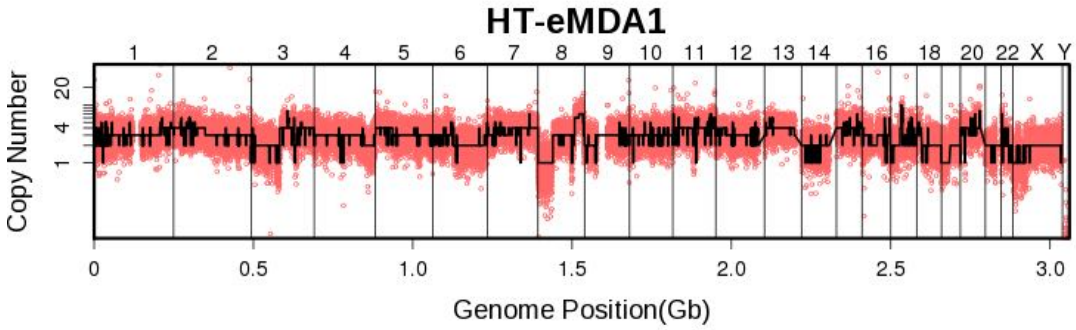
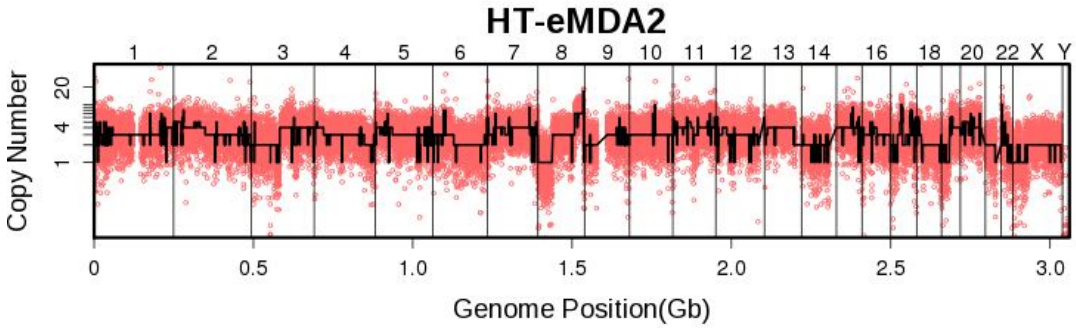
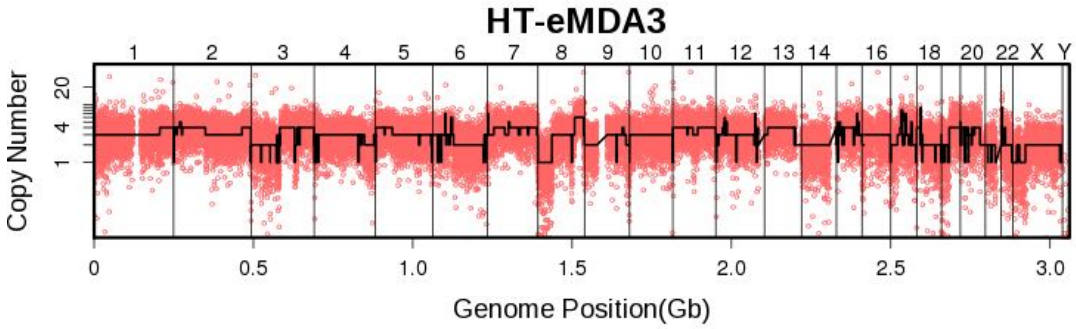


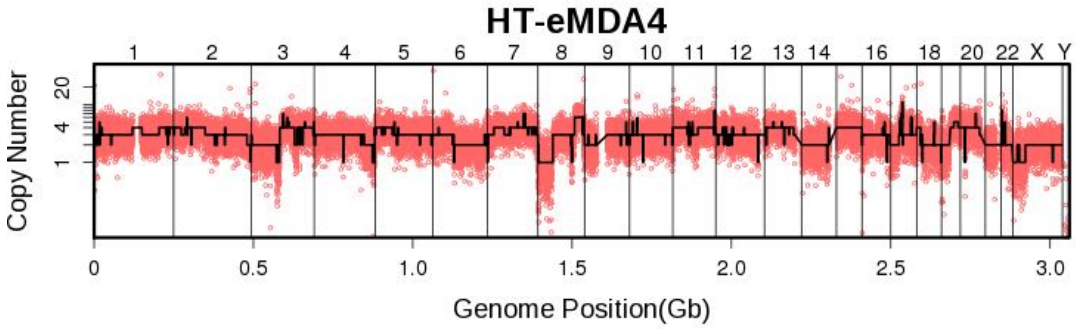
Fig. S10. Ploidy measurement of HT-29 cells using fluorescence-activated cell sorting (FACS) and sequencing. We used the human fibroblast as a reference. We used the Hoechst 33342 (Invitrogen) dye to stain the cells and applied FACS to measure the amount of DNA in each single cell reflected by its fluorescence intensity. (*A*) DNA content distribution of fibroblasts. (*B*) DNA content distribution of HT-29 cells. (*C*) DNA content distribution of the ~1:1 fibroblasts: HT-29 cells. Comparing to the fibroblasts, normal diploid human cells, we can determine the ploidy number (N) of HT-29 cells to be 2.82. (*D*) The distribution of copy number difference between unique dynamic 50kb bins. The peak at 0.325 represents the ratio of 1 copy in HT-29, indicating N = 1/0.325 = 3.08 for all the uniquely mappable regions. The difference between the two

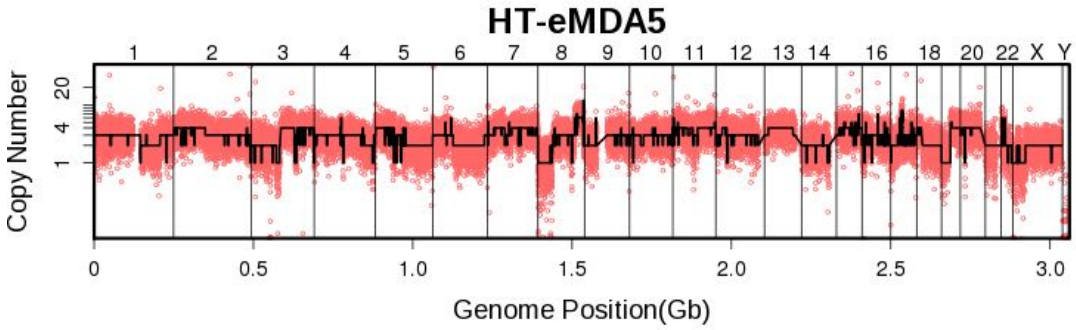
calculated results may come from the copy number difference of the repeat regions or the two cell lines' GC content difference, which affects the binding ability of the Hoechst dye.

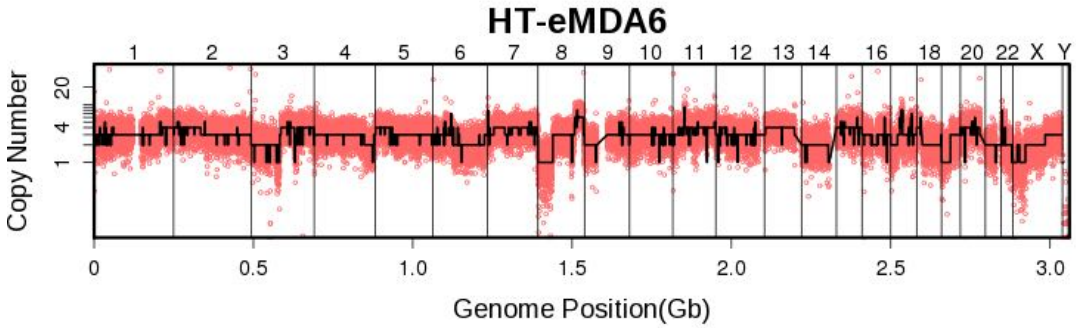

HT-eMDA1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

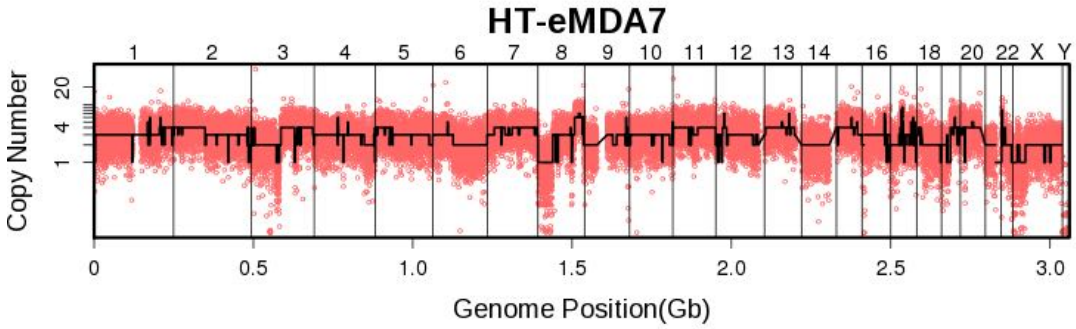

HT-eMDA2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

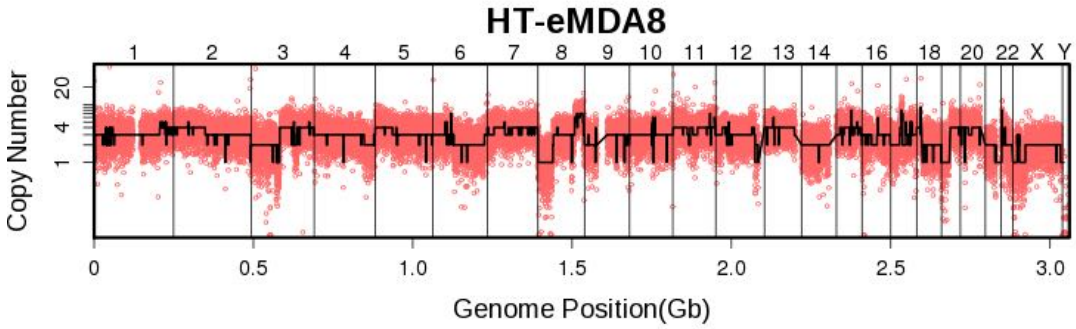

HT-eMDA3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

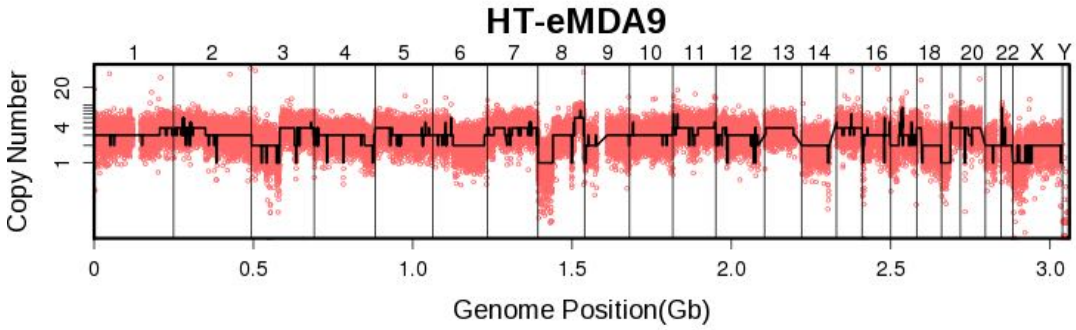

HT-eMDA4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

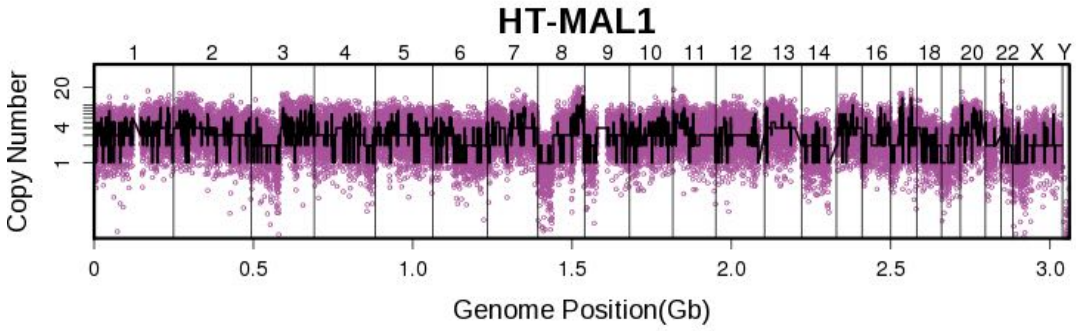

HT-eMDA5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

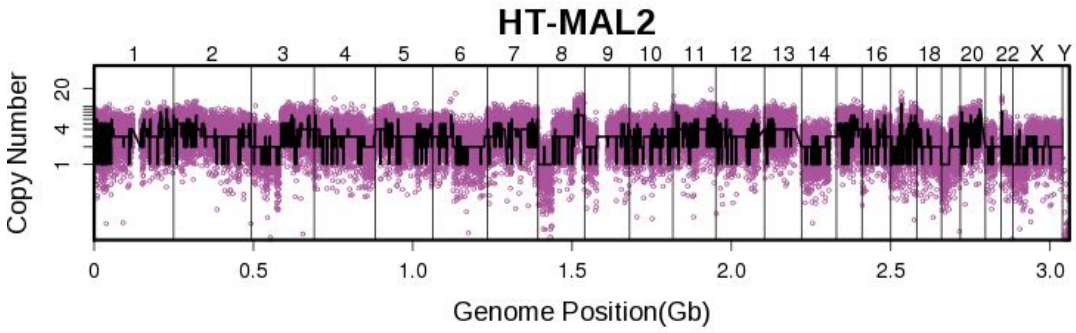

HT-eMDA6
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

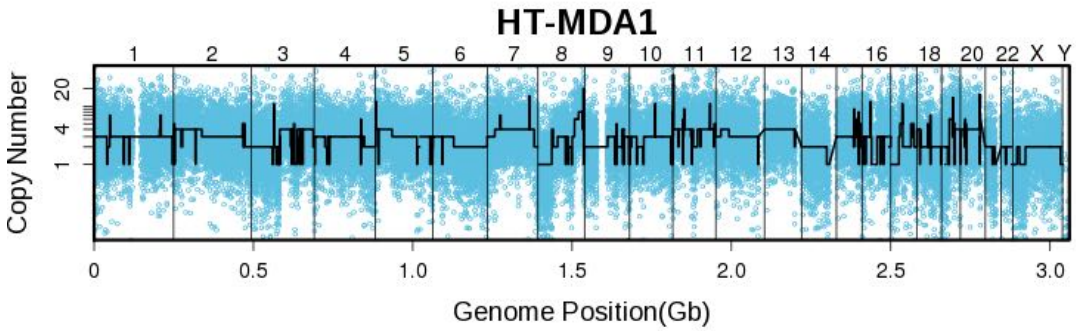

HT-eMDA7
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)


HT-eMDA8
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
20
4
1
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)


HT-eMDA9
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)


HT-MAL1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)


HT-MAL2
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)


HT-MDA1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 18 20 22 X Y
Copy Number
1 4 20
0 0.5 1.0 1.5 2.0 2.5 3.0
Genome Position(Gb)

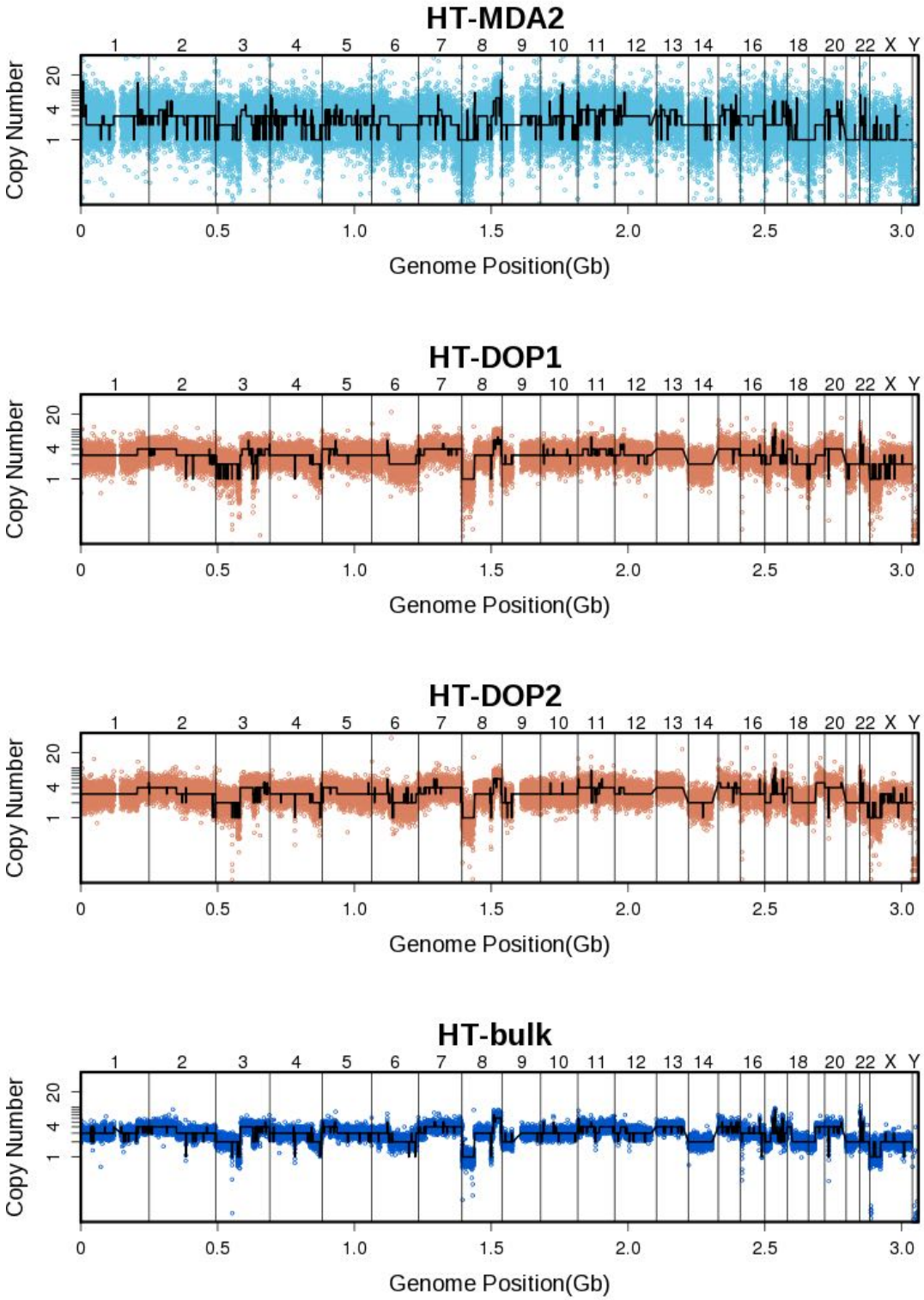


Fig. S11. The copy number distribution of all single HT-29 cell WGA by eMDA, MALBAC, MDA, and DOP-PCR. The reads are clustered into 52.4-kb (mean size) bins by dynamic binning. The black lines represent the copy number determined by CBS segmentation method.

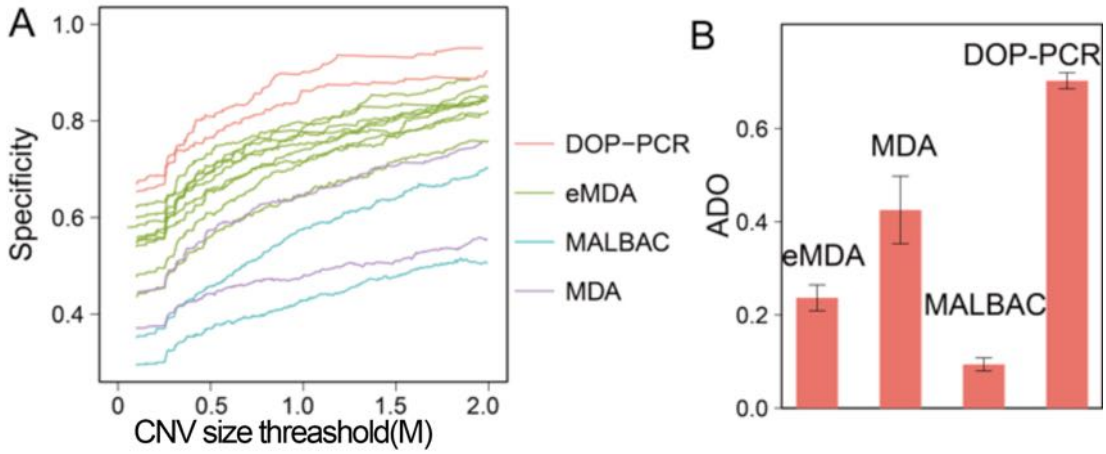


Fig. S12. CNV identification specificity and ADO rates of different single-cell WGA methods using HT-29 cells. (*A*) The specificity of CNV detection. Unamplified gDNA sample of HT-29 cells was used as a reference. DOP-PCR has the highest specificity. eMDA performs very close to DOP-PCR and much better than MALBAC or conventional MDA. (*B*) The ADO rates of deeply sequenced exome data of single HT-29 cells. eMDA improves the coverage of both alleles significantly, compared to conventional MDA.

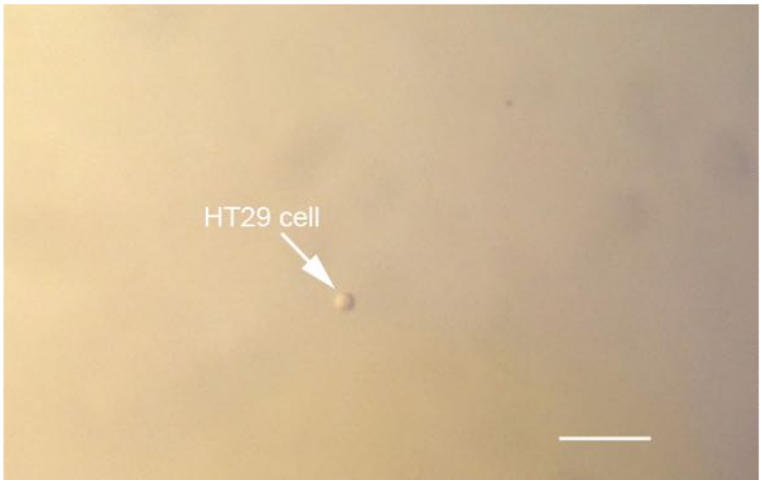


Fig. S13. The single cell observed under a stereomicroscope before lysis. Each single cell was identified under a microscope and then picked using a mouth pipet. After repeated washes (> 4 times) with clean PBS drops, we used a new mouth-pipetting capillary to transfer a single cell to the lysis buffer. By this stringent process, we were able to make sure only one cell had been picked into the lysis buffer. (Scale bar: 50 μm)

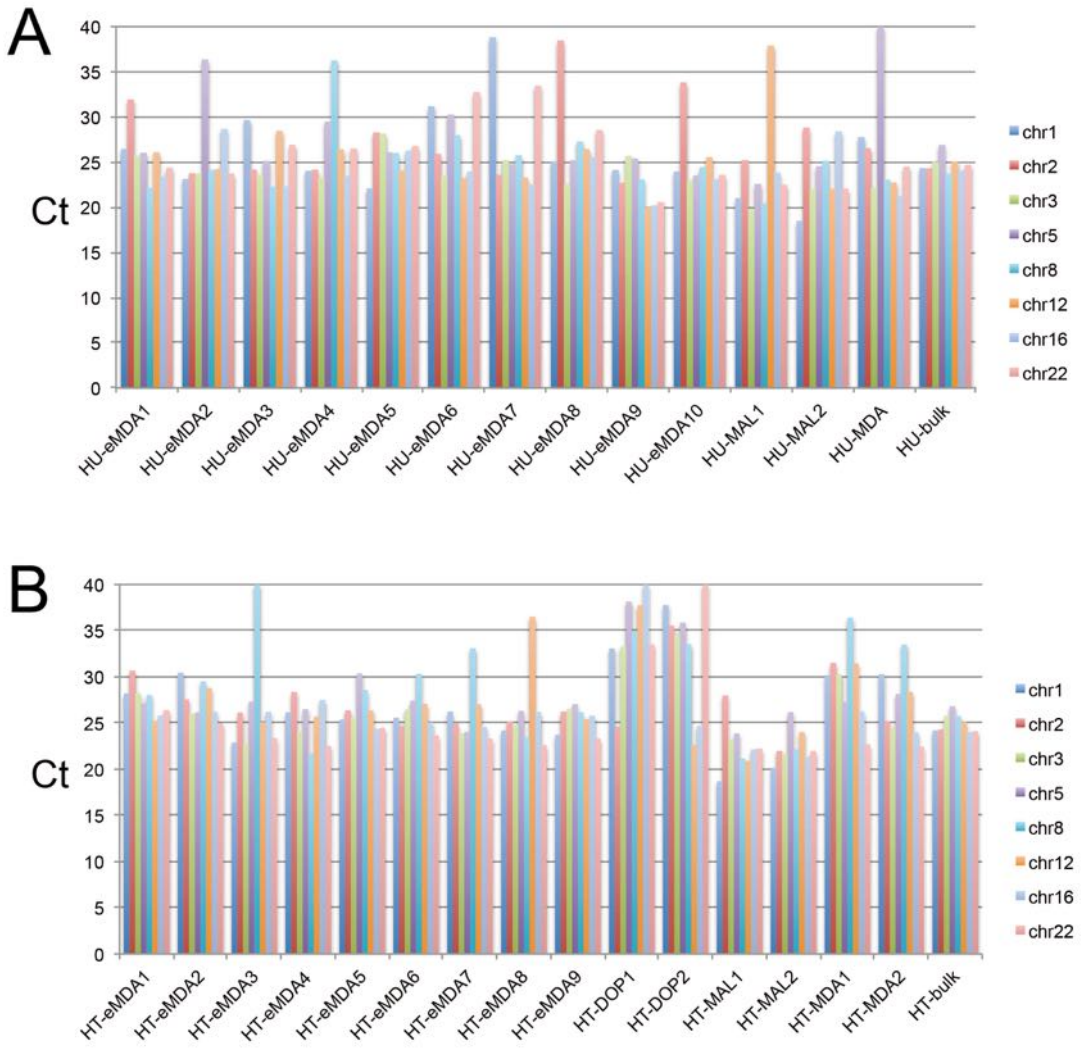


Fig. S14. Quantitative PCR (qPCR) result of selected loci from each single-cell WGA sample. After the amplification and purification, each sample was examined by qPCR with 8 sites on the genome. These sites are picked randomly with different product lengths to fully examine the amplification quality. Primers used for these qPCR are listed in **Table S2.** (*A*) The qPCR Ct values of each locus of the amplification products of single HUVECs. (*B*) The qPCR Ct values of each locus of the amplification products of single HT-29 cells. Single HUVECs show more fluctuation than single HT-29 cells, probably because HUVECs are diploid while HT-29 cells are nearly triploid.

Table S1. The detection efficiency comparison between MALBAC and eMDA

	Heterozygous SNVs	Homozygous SNVs	Total SNVs
	Bulk		
SNVs	2,896,269	1,157,899	4,054,168
	Single-Cell eMDA		
SNVs	2,076,937	782,835	2,859,772
Detection efficiency	72%	68%	70%
	Single-Cell MALBAC		
SNVs	1,931,223	720,860	2,652,083
Detection efficiency	67%	62%	65%

Table S2. qPCR primers used for quality control of each amplification. The amplified regions are randomly picked throughout the genome and the lengths of the PCR products span from ~50bp to ~400bp to give a comprehensive examination of the amplification.

chromsome	primer	sequence	PCR product length
chr1	forward primer	TTTGATGGAGAAATCCGAGG	150
	reverse primer	CTGACTCGGAGAGCAGGAC	
chr2	forward primer	GTGGAGTGGGCCTGGTTTAGAT	372
	reverse primer	AAATTACCAACTGCCCGGAGAC	
chr3	forward primer	AGGCTGCTTGACACTTTGAGGA	64
	reverse primer	TAGCATTGAAGGTGTGCCTTGC	
chr5	forward primer	CTTGCACCAGAATTGCACTGAA	53
	reverse primer	GATGTCAATTCTCCCCAGACTGA	
chr8	forward primer	TAGAGCAGGCGGCATGACTAAT	208
	reverse primer	AGCTCCACTCTTGAACGGGAAT	
chr12	forward primer	CGCTCCTGCCCTTACCTCTATC	59
	reverse primer	AAACCCGGGAGAAGGAGTATCA	
chr16	forward primer	ACGGAGTCGTCTCTGATGTATT	149
	reverse primer	TTTTGTGTTTTTCATGACATTGA	
chr22	forward primer	CTGCCAGCCCAATGTTTGTACT	53
	reverse primer	GGAAGGAAATGAGGCTTCAACC	

Table S3. Summary of data from single HUVECs (HU) and HT-29 cells (HT)

Sample	Unique coverage (Gb) [1]	Mean depth[2]	Reads type	Device[3]
HU-eMDA1	2.03	19.71	100x2	1
HU-eMDA2	1.73	8.38	100x2	1
HU-eMDA3	0.12	1.29	100x2	1
HU-eMDA4	1.22	5.13	100x2	1
HU-eMDA5	0.20	1.35	100x2	1
HU-eMDA6	0.13	1.41	100x2	1
HU-eMDA7	0.15	1.39	100x2	1
HU-eMDA8	0.17	1.42	100x2	1
HU-eMDA9	1.17	4.41	100x2	1
HU-eMDA10	0.02	1.02	50x1	1
HU-MAL1	1.83	17.42	100x2	-
HU-MAL2	0.09	1.11	50x1	-
HU-MDA	1.97	22.32	100x2	-
HU-bulk	2.73	13.87	100x2	-
HT-eMDA1	0.55	1.59	100x2	2
HT-eMDA2	0.36	1.50	100x2	2
HT-eMDA3	0.54	1.71	100x2	2
HT-eMDA4	0.73	1.70	100x2	2
HT-eMDA5	0.16	1.16	75x2	2
HT-eMDA6	0.19	1.20	75x2	2
HT-eMDA7	0.14	1.15	75x2	2
HT-eMDA8	0.15	1.18	75x2	2
HT-eMDA9	0.18	1.18	75x2	2
HT-DOP1	0.32	1.72	100X2	-
HT-DOP2	0.30	1.60	100X2	-
HT-MAL1	0.45	2.12	100X2	-
HT-MAL2	0.46	2.05	100X2	-
HT-MDA1	0.34	1.88	100X2	-
HT-MDA2	0.36	1.71	100X2	-
HT-bulk	0.66	1.41	100x2	-

[1] Bases covered by the uniquely mapped reads.
[2] Calculated as the sequencing depth in the covered region.
[3] The droplet generation device. 1: Home made PDMS chip; 2: Glass chip made by Dolomite.

Table S4. Summary of exome sequencing data from single HT-29 cells (HT) using various WGA methods

Sample	Unique coverage (Gb)[1]	Mean depth[2]	Enrich region coverage(M)[3]	Enrich Region Mean Depth[4]	Reads Type
HT-eMDA1.e	0.44	12.88	43.90	75.92	100x2
HT-eMDA2.e	0.37	10.39	42.23	51.64	100x2
HT-eMDA3.e	0.46	14.35	40.35	94.40	100x2
HT-eMDA4.e	0.65	12.27	45.58	96.84	100x2
HT-eMDA5.e	0.56	10.51	48.05	70.01	100x2
HT-eMDA6.e	0.45	10.98	47.07	61.34	100x2
HT-eMDA7.e	0.53	12.37	45.78	82.37	100x2
HT-eMDA8.e	0.82	13.01	46.60	126.49	100x2
HT-eMDA9.e	0.82	11.60	47.14	111.39	100x2
HT-DOP1.e	0.23	11.84	20.58	77.10	100x2
HT-DOP2.e	0.25	11.09	23.14	68.92	100x2
HT-MAL1.e	0.41	27.13	40.73	175.26	100x2
HT-MAL2.e	0.34	17.96	38.21	103.36	100x2
HT-MDA1.e	0.27	12.07	33.23	53.66	100x2
HT-MDA2.e	0.45	13.60	41.25	83.65	100x2
HT-bulk.e	0.55	8.78	50.05	55.68	100x2

[1] Bases covered by the uniquely mapped reads.
[2] Calculated as the sequencing depth in the covered region.
[3] Coverage of the regions that shall be enriched by exome-capture reagent kit.
[4] Calculated as the sequencing depth in the covered bases in the regions that shall be enriched by exome-capture reagent kit.

Reference:

1. Navin N *et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature* (472):90–94.
2. Gole J *et al.* (2013) Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat Biotechnol* (31):1126–1132.