ACS APPLIED MATERIALS & INTERFACES

Research Article

www.acsami.org

Single-Cell-Based Platform for Copy Number Variation Profiling through Digital Counting of Amplified Genomic DNA Fragments

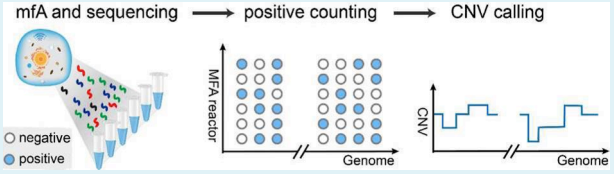
Chunmei Li,[†] Zhilong Yu,[†] Yusi Fu,[†] Yuhong Pang,[†] and Yanyi Huang*[,†,‡,§]

[†]Beijing Advanced Innovation Center for Genomics (ICG), Biodynamic Optical Imaging Center (BIOPIC), School of Life Sciences, Peking University, Beijing 100871, China

[‡]College of Engineering and [§]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China

Supporting Information

ABSTRACT: We develop a novel single-cell-based platform through digital counting of amplified genomic DNA fragments, named multifraction amplification (mfA), to detect the copy number variations (CNVs) in a single cell. Amplification is required to acquire genomic information from a single cell, while introducing unavoidable bias. Unlike prevalent methods that directly infer CNV profiles from the pattern of sequencing depth, our mfA platform denatures and separates the DNA molecules from a single cell into multiple fractions of a reaction mix before amplification. By examining the sequencing result of each fraction for a specific fragment and applying a segment-merge maximum likelihood algorithm to the calculation of copy number, we digitize the sequencing-depth-based CNV identification and thus provide a method that is less sensitive to the amplification bias. In this paper, we demonstrate a mfA platform through multiple displacement amplification (MDA) chemistry. When performing the mfA platform, the noise of MDA is reduced; therefore, the resolution of single-cell CNV identification can be improved to 100 kb. We can also determine the genomic region free of allelic drop-out with mfA platform, which is impossible for conventional single-cell amplification methods.



KEYWORDS: *single-cell, digital counting, DNA amplification, copy number variation, maximum likelihood, genomic sequencing*

INTRODUCTION

The genomic copy number variation (CNV) is a type of duplication or deletion event that affects a considerable number of bases of DNA.[1−3] The size of CNV varies from a few bases to even a complete chromosome.[4−6] CNV plays an important role in generating necessary variations in the population as well as disease phenotypes.[2,7,8] Most reported studies focused on shared CNVs or average CNV patterns among a bulk amount of cells.[4,6,9] However, recent studies revealed that each cell may have harbored its characteristic CNV profile, and such profiles might help reconstruct the full spectrum of the cellular complexity and the evolutionary connections between cells.[10−12] Highly accurate and precise identification of CNVs in single cells is extremely important and necessary to fundamental biology studies or medical researches since many CNVs are associated with critical biofunctions and diseases.[7,8,13−16] Large fragment CNVs are easy to detect through conventional molecular biology methods,[17,18] microarray-based analyses,[19−23] or karyotyping.[7,22] However, identifying the small-size CNVs from a single cell with high confidence is challenging, thus becoming one of the highly demanded features in single-cell whole genome sequencing.[10,14]

Various approaches have been developed to obtain genomic information in single cells using next-generation sequencing platforms.[9,10] The key to these methods is to amplify the genomic DNA of a single cell with high coverage breadth, low replication error rate, and low amplification bias. We have compared a few commonly practiced methods, including degenerate oligonucleotide-primed PCR (DOP-PCR),[24] multiple displacement amplification (MDA),[25,26] multiple annealing and looping-based amplification cycles (MALBAC),[27,28] and our microfluidic-based emulsion amplification (eWGA),[29] and found that eWGA shows a superb performance in balancing the identification capability between CNVs and single nucleotide variations (SNVs). The limited identification resolution of CNV for eWGA is around 0.25−1 Mb, which may cause false negative calling of small CNVs. In previously reported methods, the CNV pattern is directly reflected by the sequencing depth, i.e., the number of reads that covered the certain positions in the genome, and the actual values of the copy numbers can be deduced from ratiometric analysis. Hence the CNV identification is highly affected by the amplification bias.[10]

In this article, we present a new method to detect small CNVs in single cells through a multifraction amplification (mfA) approach coupled with high-throughput sequencing. The key concept of mfA is to avoid the bias-sensitive ratiometric assessment and simply count the number of fractions that contain an amplified product of specific fragment. This mfA approach is independent of amplification bias, making this strategy universally applicable to other single-cell WGA methods.

Received: March 4, 2017
Accepted: March 24, 2017
Published: March 24, 2017

DOI: 10.1021/acsami.7b03146
ACS Appl. Mater. Interfaces 2017, 9, 13958−13964

RESULTS AND DISCUSSION

Experimental Process. This strategic design can be applied to any amplification chemistry, but in this work we chose MDA since it has four major advantages to amplify the single cell's whole genome: (a) the reaction is isothermal, which is probably the most favorable protocol for experimental operators; (b) amplification is highly efficient, producing microgram-level amplified product which is sufficient for all available experimental protocols for sequencing library construction; (c) amplification is highly accurate with overall error rate around 10^{-5} owing to the high fidelity of phi 29 polymerase; (d) amplification has high coverage breadth across the whole genome, typically 70−80% for a normal diploid human cell.[10] The mfA with MDA method was named as mfMDA, and the general experimental design of mfMDA was schematically illustrated in Figure 1. Each single cell was manually picked and

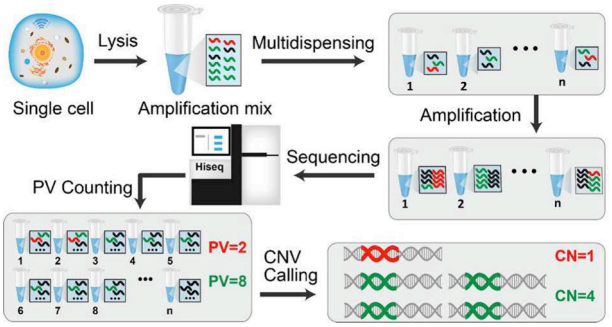


Figure 1. Experimental procedure of multifraction amplification (mfA). An intact single cell is lysed, and the DNA fragments are denatured to single-strand fragments. Then the lysate is mixed with an amplification reaction reagent and is equally dispensed into n (n = 7−20) reaction microtubes. The different single-stranded DNA fragments for the same DNA sequence will be likely separated into different tubes. The amplification products from each tube are purified and sequenced separately. For any target genomic sequence, the number of tubes that performs positive signals of this locus is defined as positive value (PV), from which the copy number can be deduced.

lysed to release DNA, which was denatured into single strands, and then was separated into multiple (n = 7−20, according to the maximum copy number of the single cell) fractions, each of which was amplified separately in a small-volume reaction tube. The MDA product of each fraction was uniquely indexed through sequencing library construction and then sequenced using an Illumina platform.

After sequencing, we analyzed the sequencing reads of all the fractions, and for each fraction we mapped the sequenced fragments to the reference genome. For a genomic region that contains more than one copy of DNA, we should have large chances to observe the signal in more than one fraction (reaction tubes). A straightforward way to deduce the possible copy number of a specific location of the genome is to count the number of fractions in which this locus has been sequenced. However, simply counting the positive tubes may underestimate the real copy numbers since the fraction number is not infinite; hence, there is a certain probability that multiple copies of a single genomic locus coparticipate in the same tube. The probability of such a distribution can be accurately predicted by Poisson statistics,[30,31] similar to the consideration of digital PCR but not identical. In digital PCR applications, we commonly desire a large dynamic range for highly accurate detecting target DNA or RNA fragments from a few copies to a few thousand copies.[32−35] However, single-cell CNV studies may only require a small number of compartmentations since genomic DNA copy number is typically no more than 10.

Amplification Probability of ssDNA Fragments. Single-cell whole genome amplification started with cell lysis and DNA denature, providing fragmentized single-strand DNA (ssDNA) as starting material. It is important to point out that unlike digital PCR WGA does not guarantee successful amplification of every DNA fragment in a reaction system. To experimentally obtain the average probability of an ssDNA fragment that can be amplified by MDA reaction, we specifically used two human haploid single cells (sperms) as starting material in a single tube. The exome was enriched after amplification and library preparation and then sequenced and analyzed. The result (see Supporting Information, Section 2) indicates that the probability of successful amplification of an ssDNA fragment is about 0.4. Thus, for a haploid cell, the theoretical maximum coverage breadth of mapped reads across the whole genome is $1 - (1 - 0.4)^2 = 0.64$, and for diploid cells with 4 single-strand copies, the coverage limit is $1 - (1 - 0.4)^4 = 0.87$. Similarly, the limit is 0.95 for a triploid cell and 0.98 for tetraploid.

Theoretical Simulation. A fragment from a region with a larger copy number will have more ssDNA copies. In the mfMDA approach, these ssDNA copies are separated into different fractions. More original ssDNA copies mean more fractions may contain reads mapped to this fragment after sequencing. If a fragment is detected in n fractions, we define that this fragment has a positive value of n (PV = n). If we consider all the fragments in a given region, each fragment may have a specific PV (PV = 0, 1, 2, ..., n_{fraction}). For each specific region, we can calculate the frequency of different PV.

We performed a simulation to quantitatively assess the PV frequency distribution of ssDNA fragments, assuming each of which has average amplification probability of 0.4, and each region contains 100 fragments (Figure 2). Different original

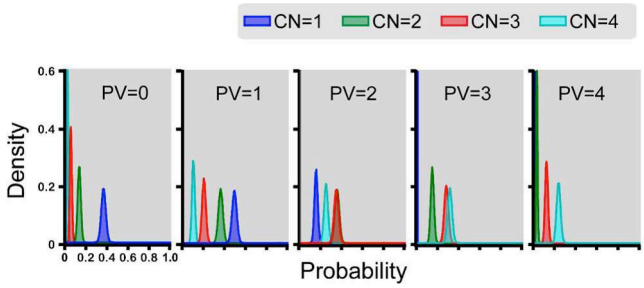


Figure 2. Probability distribution of positive values (PVs) for different fragments with different copy number (CN). Considering the proportion of fragments with PV of 0, it is easy to distinguish CN = 1 from CN > 1 cases. Considering the proportion of PV of 0, 1, and 3, it is possible to distinguish CN = 2 and CN = 3 cases, and considering the proportion of PV of 1, 2, and 4, it can be better to distinguish CN = 3 and CN = 4 cases.

copy number will lead to different PV frequency distributions. For example, CN = 1 exhibits a high frequency for PV = 0. While CN = 2 and CN = 3 show quite similar frequency for PV = 2, they have quite different frequencies for PV = 1 and PV = 3. This suggests that after we obtain the experimental PV data from mfMDA we can determine the copy number of each region through a maximum likelihood estimation according to the PV frequency distribution statistics. For each region, we performed the maximum likelihood estimation calculation for

DOI: 10.1021/acsami.7b03146
ACS Appl. Mater. Interfaces 2017, 9, 13958−13964

all possible copy numbers. We then obtained the likelihood of each possible copy number by calculating the product of probability of all the PV values in the region. The copy number containing the maximum product of probability was then assigned as the copy number of the region.

Amplification of Single HT-29 Cells. We then tested mfMDA for single human cells with intrinsic CNV patterns. We used the HT-29 cell line (a human colon cancer cell) as the model system. HT-29, a cell line with average copy number around three,[36] has been thoroughly studied by other single-cell WGA approaches, making it a good candidate for assessing technical performance of our new method. We applied 15 fractions in this mfMDA experiment, and the sequencing libraries were constructed using the Illumina Nextera kit and sequenced by Hiseq 2500 sequencer. The fraction number should be larger than the number of ssDNA strands, or there will be a chance that a target sequence appears in all the fractions. We will not be able to tell if this phenomenon is due to the contamination. Therefore, we choose $2 \times (CN_{max}) + 1$ as the fraction number. A larger fraction number is acceptable, but it will make the experiment and library preparation more complicated. Fifteen fractions for the HT-29 cell will be able to handle a maximum CN of 7, which we believe is enough for the HT-29 cell with average CN of about 3. In contrast, in our previous technical verification experiments using single sperms, each of which has only two ssDNAs after denature, a fraction number of 3 will be enough. However, in the experiment we need to make sure that we get only one sperm cell but not two. In addition we need to make sure the experiment is not contaminated by diploid cells. Hence the experiment needs to be designed to distinguish a copy number of 2, and we choose at least 5 fractions for single sperm experiments. The sequencing reads were mapped to human genome reference hg19 using bowtie2.[37] Duplicates were removed by samtools.[38] The insert size was 200−300 bp during the library construction, and the read length is 2 × 150 bp using paired-end sequencing. To make the best use of the data, we choose 200 bp as the size of a bin. The whole genome was divided into 200-bp bins, and each bin was considered as a fragment. Contamination was also removed during processing. For each fraction we calculated the sequencing depth of each bin and obtained the PV of all bins. The whole genome is segmented into regions. In a given region, we calculated the frequency of different PVs, which were the observed values in our maximum likelihood estimation.

Determining the copy numbers from PV frequency requires the probability density function, which can be generated from simulations based on other parallel mfMDA experiments or, preferably, from the same experiment based on a landscape of the CNV distribution (Figure 3). Such a landscape can be easily depicted from bulk sequencing, but for single cell sequencing, especially for those systems in which intercellular heterogeneity exists, we have to generate an approximated profile from the sum of read depth of all fractions in an mfMDA experiment. It is also critical to determine a proper block size to provide such a low-resolution CNV landscape. If the block size is too small, the small bias-related variations will result in extra noise, while if the block size is too large, the resolution of the landscape will be too low. After analyzing the relationship between the normalized coefficient of variance (CV) of the read depth inside the block and the block size (Figure S3), in our experiment we chose the block size of 0.7 Mb. When the bin size is small, the random noise of MDA affects the accuracy of

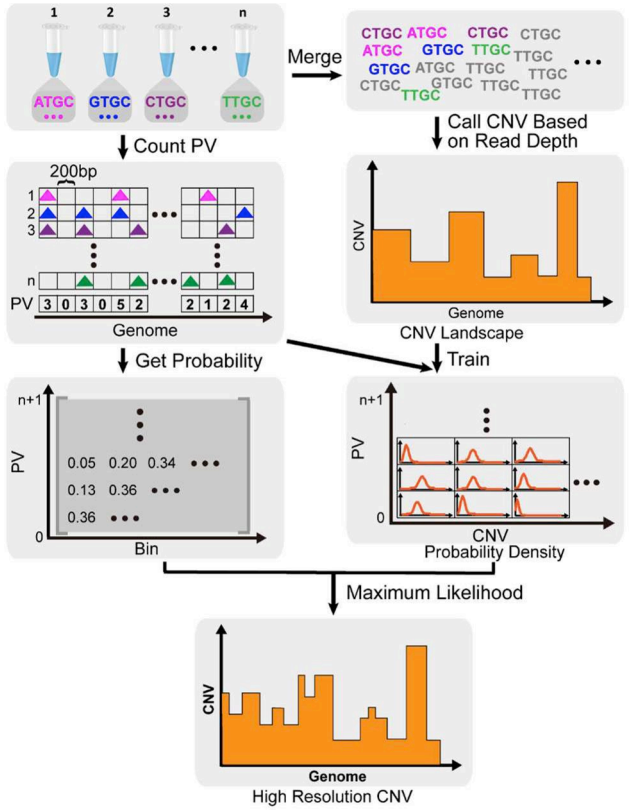


Figure 3. CNV identification through maximum likelihood estimation. The whole genome is divided into 200-bp fragments, and for each fragment the PV is counted. Then the sequencing data are merged together to call a low-resolution CNV landscape of the genome. A probability density is obtained by calculating the probability distribution of each PV for bins sampled across low-resolution CNV regions. At the same time, the whole genome is divided into small bins (>200 bp), and we can get the proportion of different PV for each bin. With the maximum likelihood estimation, a high-resolution CNV landscape of the genome can be obtained.

CNV segmentation. The noise gets smaller as the bin size grows. However, when the bin size is too large, regions with different copy number may be merged into a single bin and then make CNV calling inaccurate, leading to an increase of CV. We find that the CV result has a local minimal value when the bin size is around 0.7 M. The total read depth was binning to the block size and processed using DNACopy.[39] The CNV regions with a size larger than 20 Mb were picked to generate the probability density function (Table S4).

A naïve approach is to segment whole genome into regions with equal size and apply maximum likelihood estimation to all the regions to get the most probable copy numbers. A possible problem of this method is that it will be influenced significantly by contamination or amplification noise. Higher resolution requires smaller size of the regions, but smaller regions are more susceptible to noise. In addition, the number of fragments in a single region is limited in small-size regions; hence, the proportion distribution of the fragments of a certain PV becomes more dispersed, leading to an increase in discrimination difficulty (Figure S4).

We then applied a segment-merge algorithm before maximum likelihood estimation to enable the detectability of smaller CNVs. This algorithm uses variable size of regions which contain different numbers of fragments. The data were

DOI: 10.1021/acsami.7b03146
ACS Appl. Mater. Interfaces 2017, 9, 13958−13964

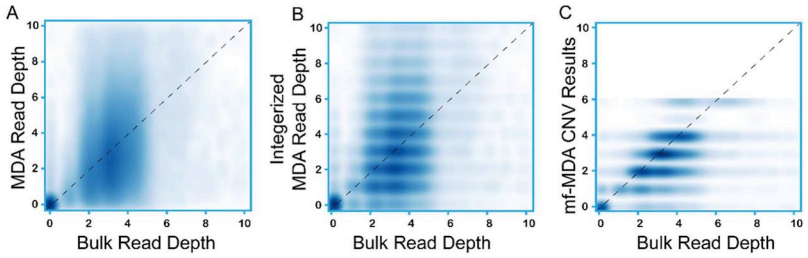


Figure 4. Consistency of CNV identification between single-cell WGA methods and bulk sequencing data. (A) The correlation between bulk read depth and the MDA read depth. The low correlation implies that conventional single cell MDA results cannot accurately present the CNV. (B) The correlation between bulk read depth and the integerized read depth of single-cell MDA result. The correlation is still low. (C) The correlation between bulk read depth and the copy number identified through mfMDA approach.

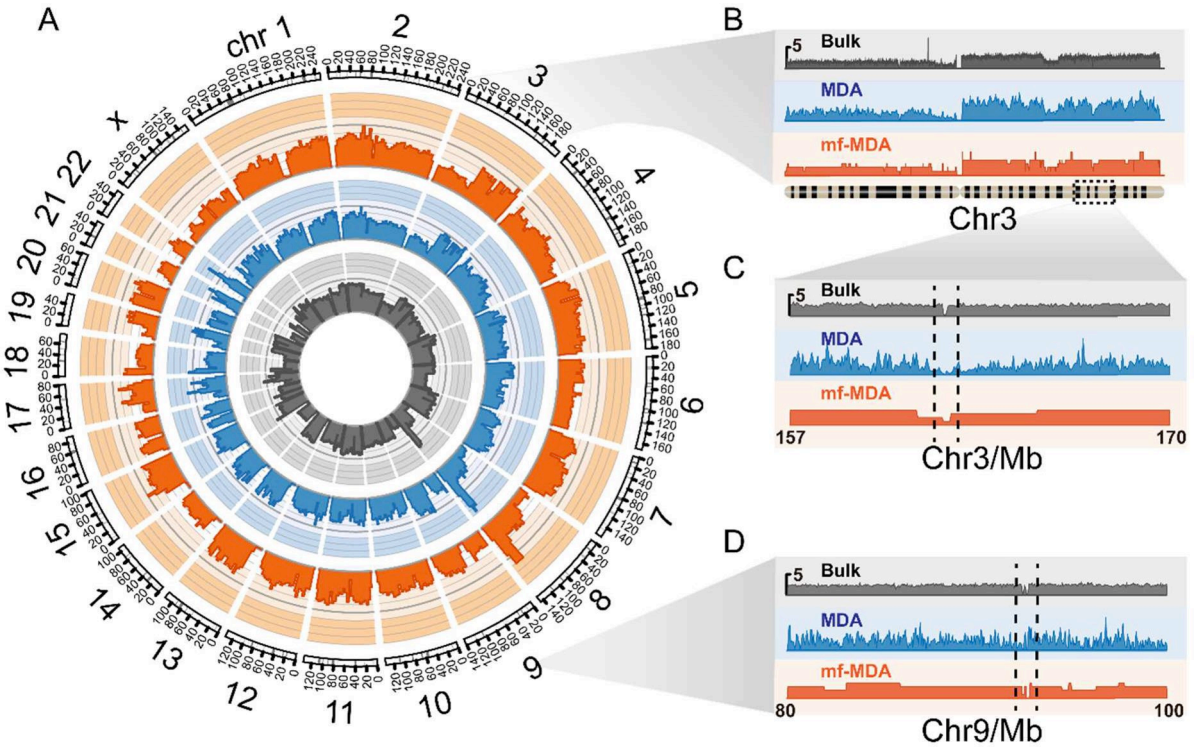


Figure 5. CNV results of HT-29 cells. (A) The genome-wide CNV pattern of HT-29 cells. Gray: CNV pattern deduced from sequencing depth of bulk sample. Blue: CNV pattern reflected by conventional MDA of single HT-29 cell. Orange: CNV pattern reflected by the single-cell mfMDA approach. (B) The CNV patterns of chromosome 3. (C),(D) The detailed CNV patterns.

segmented into small regions, and the counts of PV values for each region were obtained. These series of counts, containing information on genome-wide copy number changes, were segmented to form change points that defined the boundaries between genome segments with different copy numbers. We can simply use these change points to produce the genomic CNV pattern; however, these change points contained many false positives which might be further filtered. During the filtering process, we combined the most significant change points in the series. The small regions were merged into large regions with variable sizes divided by the filtered change points. These regions were then processed using the maximum likelihood method to determine their copy numbers.

We found that this segment-merge algorithm could highly reproduce the bulk cell CNV results compared to conventional MDA (Figure 4) and achieved higher resolution to identify CNVs from single-cell WGA product through mfMDA (Figure 5A). The result showed that noise across all the regions was much lower than the raw read depth of MDA, with spikes and channels removed. With high noise, CNV was not accurately detected for conventional single-tube MDA but clearly revealed through the mfMDA approach. Particularly, some small CNVs, at the size smaller than 500 kb, which have been difficult to identify in previous studies, can be successfully identified by mfMDA. For example, bulk sequencing showed that in chromosome 3 there was a 150-kb copy number loss. This 2-to-1 copy number loss can be accurately detected by mfMDA but obscured by the bias-induced noise in conventional MDA (Figure 5B). Another similar example was a 100-kb copy number deletion in chromosome 9 (Figure 5C).

In addition to facilitating the accurate assessment of small CNVs for single-cell sequencing, this mfMDA approach also provides a unique advantage to identify the high confidence regions that are free of allelic drop-out (ADO) in the single-cell sequencing data. None of the previously developed single-cell amplification and sequencing methods can provide such capability. Despite its high coverage and accuracy, MDA suffers from a notoriously high rate of allelic drop out (ADO) due to

DOI: 10.1021/acsami.7b03146
ACS Appl. Mater. Interfaces 2017, 9, 13958−13964

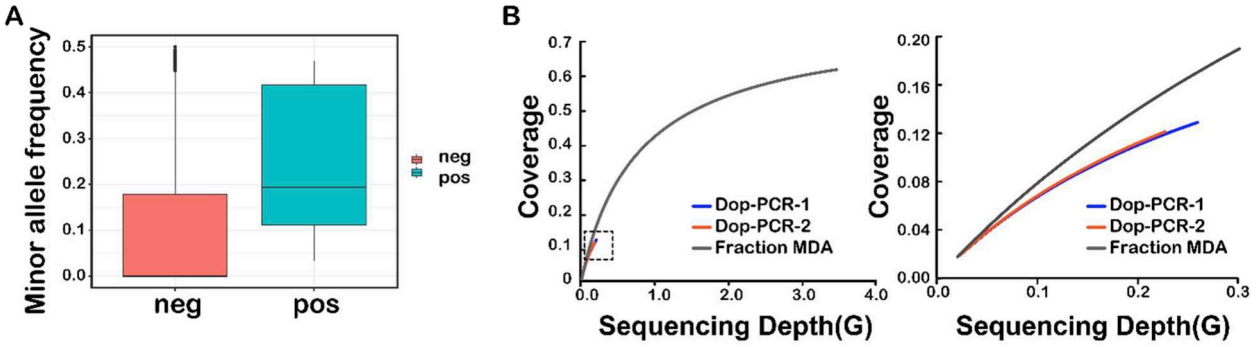


Figure 6. WGA performance of mfMDA. (A) The minor allele frequency in ADO-free regions (marked as "pos") and in uncertain regions (marked as "neg"). (B) mfMDA exhibits higher genomic coverage than DOP−PCR.

the random amplification nature, which significantly limited the application of MDA. However, with our mfMDA data, we can determine the ADO-free region of single-cell WGA result. We analyzed the PV of each segment in the diploid region and screened out those regions that have PV larger than 2. In these regions, at least three ssDNA fragments of the same genomic location have been successfully captured in mfMDA, ensuring that both alleles have been sequenced. Since every single cell can only be amplified once, it is important to know which region is ADO-free, and only in such regions we can confidently identify and differentiate homozygous and heterozygous mutations. We have analyzed all the sequencing data from mfMDA and identified the 360 Mb (12% of the genome) ADO-free region from the WGA sequencing data of a single HT-29 cell. The experimental result showed that the minor allele frequency is obviously higher than the other uncertain region (Figure 6A).

It is reasonable to couple this fractionalization strategy with other single-cell amplification chemistries besides MDA, using the same algorithm to deduce the CNV pattern. However, unlike high-coverage and random-bias MDA, some of those amplification chemistries have limited coverage breadth (Figure 6B)[24] and will cause inaccurate assessment of the CNV when PV counting is applied, while some of the others exhibit sequence-dependent bias[20,28] which need normalization before to obtain the probability density matrix. Furthermore, MDA is an isothermal process, making the whole experimental process convenient to operate. Fractionalization is technically easy to be realized by either robotic automation or by microfluidic devices if needed and will facilitate the robustness of this approach by reducing the possible errors during experimental operations. In addition, fractionalization is technically easy to be realized by either robotic automation or by microfluidic devices if needed. Multiplex sequencing can be accomplished using dual-indexing library preparation or adding individual barcodes to each sample. With these techniques, we believe tens to hundreds of cells can be handled in parallel during the experimental part. For data analysis, the number of cells being analyzed in parallel is only limited by computing capabilities and resources. Usually four to eight cells can be analyzed in parallel with a mainstream personal computer.

CONCLUSION

In summary, we have developed a new approach to identify single-cell copy number variations based on multifraction amplification (mfA) and sequencing of the single-cell whole genome. This mfA strategy is fundamentally different from previously reported methods that deduced copy number from sequencing depth. By examining each fraction if a specific fragment has been sequenced, sequencing-depth-based CNV identification can be transformed into digital counting and thus become insensitive to the amplification bias. Since not every DNA fragment can be successfully amplified, we then applied the segment-merge maximum likelihood algorithm to calculate the copy number, taking the amplification probability into consideration. We have demonstrated the method with MDA. The new method has lower noise, facilitating the identification of single-cell CNVs with resolution at 100 kb. Combined with the high fidelity of MDA, the new method allows for further analysis of single nucleotide variations. With our method we could also determine the genomic region without allelic drop-out, which is impossible for traditional MDA or other single-cell amplification methods.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acsami.7b03146.

Detailed experimental process, bioinformatics analysis, Figures S1−S4, Tables S1−S4 (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: yanyi@pku.edu.cn.

ORCID

Yanyi Huang: 0000-0002-7297-1266

Author Contributions

Y.H. and Z.Y. conceived the project. C.L. and Z.Y. conducted the experiment. Z.Y., Y.F., and C.L. performed the data analysis. All authors discussed the data and wrote the paper.

Funding

This work was supported by National Natural Science Foundation of China (21327808 and 21525521) and Ministry of Science and Technology of China (2015AA0200601 and 2016YFC0900100).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank He Chen and Dr. Xiannian Zhang for constructive comments on data analysis and Dr. Yun Zhang and the Peking University High-throughput Sequencing Center at BIOPIC for the experimental assistance.

DOI: 10.1021/acsami.7b03146
ACS Appl. Mater. Interfaces 2017, 9, 13958−13964

REFERENCES

(1) Alkan, C.; Coe, B. P.; Eichler, E. E. Applications of Next-Generation Sequencing Genome Structural Variation Discovery and Genotyping. *Nat. Rev. Genet.* **2011**, *12*, 363−375.

(2) Freeman, J. L.; Perry, G. H.; Feuk, L.; Redon, R.; McCarroll, S. A.; Altshuler, D. M.; Aburatani, H.; Jones, K. W.; Tyler-Smith, C.; Hurles, M. E.; Carter, N. P.; Scherer, S. W.; Lee, C. Copy Number Variation: New Insights in Genome Diversity. *Genome Res.* **2006**, *16*, 949−961.

(3) Sharp, A. J.; Locke, D. P.; McGrath, S. D.; Cheng, Z.; Bailey, J. A.; Vallente, R. U.; Pertz, L. M.; Clark, R. A.; Schwartz, S.; Segraves, R.; Oseroff, V. V.; Albertson, D. G.; Pinkel, D.; Eichler, E. E. Segmental Duplications and Copy-Number Variation in the Human Genome. *Am. J. Hum. Genet.* **2005**, *77*, 78−88.

(4) Redon, R.; Ishikawa, S.; Fitch, K. R.; Feuk, L.; Perry, G. H.; Andrews, T. D.; Fiegler, H.; Shapero, M. H.; Carson, A. R.; Chen, W.; Cho, E. K.; Dallaire, S.; Freeman, J. L.; Gonzalez, J. R.; Gratacos, M.; Huang, J.; Kalaitzopoulos, D.; Komura, D.; MacDonald, J. R.; Marshall, C. R.; Mei, R.; Montgomery, L.; Nishimura, K.; Okamura, K.; Shen, F.; Somerville, M. J.; Tchinda, J.; Valsesia, A.; Woodwark, C.; Yang, F.; Zhang, J.; Zerjal, T.; Zhang, J.; Armengol, L.; Conrad, D. F.; Estivill, X.; Tyler-Smith, C.; Carter, N. P.; Aburatani, H.; Lee, C.; Jones, K. W.; Scherer, S. W.; Hurles, M. E. Global Variation in Copy Number in the Human Genome. *Nature* **2006**, *444*, 444−454.

(5) Tuzun, E.; Sharp, A. J.; Bailey, J. A.; Kaul, R.; Morrison, V. A.; Pertz, L. M.; Haugen, E.; Hayden, H.; Albertson, D.; Pinkel, D.; Olson, M. V.; Eichler, E. E. Fine-Scale Structural Variation of the Human Genome. *Nat. Genet.* **2005**, *37*, 727−732.

(6) Zarrei, M.; MacDonald, J. R.; Merico, D.; Scherer, S. W. A Copy Number Variation Map of the Human Genome. *Nat. Rev. Genet.* **2015**, *16*, 172−183.

(7) Buysse, K.; Delle Chiaie, B.; Van Coster, R.; Loeys, B.; De Paepe, A.; Mortier, G.; Speleman, F.; Menten, B. Challenges for CNV Interpretation in Clinical Molecular Karyotyping: Lessons Learned From a 1001 Sample Experience. *Eur. J. Med. Genet.* **2009**, *52*, 398−403.

(8) McCarroll, S. A.; Altshuler, D. M. Copy-Number Variation and Association Studies of Human Disease. *Nat. Genet.* **2007**, *39*, S37−S42.

(9) Xie, C.; Tammi, M. T. CNV-Seq, a New Method to Detect Copy Number Variation Using High-Throughput Sequencing. *BMC Bioinf.* **2009**, *10*, 1−9.

(10) Huang, L.; Ma, F.; Chapman, A.; Lu, S.; Xie, X. S. Single-Cell Whole-Genome Amplification and Sequencing: Methodology and Applications. *Annu. Rev. Genomics Hum. Genet.* **2015**, *16*, 79−102.

(11) Kalisky, T.; Quake, S. R. Single-Cell Genomics. *Nat. Methods* **2011**, *8*, 311−314.

(12) Wang, Y.; Navin, N. E. Advances and Applications of Single-Cell Sequencing Technologies. *Mol. Cell* **2015**, *58*, 598−609.

(13) Vogelstein, B.; Papadopoulos, N.; Velculescu, V. E.; Zhou, S., Jr.; Diaz, L. A.; Kinzler, K. W. Cancer Genome Landscapes. *Science* **2013**, *339*, 1546−1558.

(14) Gawad, C.; Koh, W.; Quake, S. R. Single-Cell Genome Sequencing: Current State of the Science. *Nat. Rev. Genet.* **2016**, *17*, 175−188.

(15) Yan, L.; Yang, M.; Guo, H.; Yang, L.; Wu, J.; Li, R.; Liu, P.; Lian, Y.; Zheng, X.; Yan, J.; Huang, J.; Li, M.; Wu, X.; Wen, L.; Lao, K.; Li, R.; Qiao, J.; Tang, F. Single-Cell Rna-Seq Profiling of Human Preimplantation Embryos and Embryonic Stem Cells. *Nat. Struct. Mol. Biol.* **2013**, *20*, 1131−1139.

(16) Navin, N.; Kendall, J.; Troge, J.; Andrews, P.; Rodgers, L.; McIndoo, J.; Cook, K.; Stepansky, A.; Levy, D.; Esposito, D.; Muthuswamy, L.; Krasnitz, A.; McCombie, W. R.; Hicks, J.; Wigler, M. Tumour Evolution Inferred by Single-Cell Sequencing. *Nature* **2011**, *472*, 90−119.

(17) Fiegler, H.; Redon, R.; Andrews, D.; Scott, C.; Andrews, R.; Carder, C.; Clark, R.; Dovey, O.; Ellis, P.; Feuk, L.; French, L.; Hunt, P.; Kalaitzopoulos, D.; Larkin, J.; Montgomery, L.; Perry, G. H.; Plumb, B. W.; Porter, K.; Rigby, R. E.; Rigler, D.; Valsesia, A.; Langford, C.; Humphray, S. J.; Scherer, S. W.; Lee, C.; Hurles, M. E.; Carter, N. P. Accurate and Reliable High-Throughput Detection of Copy Number Variation in the Human Genome. *Genome Res.* **2006**, *16*, 1566−1574.

(18) Newman, T. L.; Tuzun, E.; Morrison, V. A.; Hayden, K. E.; Ventura, M.; McGrath, S. D.; Rocchi, M.; Eichler, E. E. A Genome-Wide Survey of Structural Variation Between Human and Chimpanzee. *Genome Res.* **2005**, *15*, 1344−1356.

(19) Carter, N. P. Methods and Strategies for Analyzing Copy Number Variation Using DNA Microarrays. *Nat. Genet.* **2007**, *39*, S16−S21.

(20) Yu, Z.; Lu, S.; Huang, Y. Microfluidic Whole Genome Amplification Device for Single Cell Sequencing. *Anal. Chem.* **2014**, *86*, 9386−9390.

(21) Yang, F.; Zuo, X.; Li, Z.; Deng, W.; Shi, J.; Zhang, G.; Huang, Q.; Song, S.; Fan, C. A Bubble-Mediated Intelligent Microscale Electrochemical Device for Single-Step Quantitative Bioassays. *Adv. Mater.* **2014**, *26*, 4671−4676.

(22) Chen, P.; Pan, D.; Fan, C.; Chen, J.; Huang, K.; Wang, D.; Zhang, H.; Li, Y.; Feng, G.; Liang, P.; He, L.; Shi, Y. Gold Nanoparticles for High-Throughput Genotyping of Long-Range Haplotypes. *Nat. Nanotechnol.* **2011**, *6*, 639−644.

(23) Chen, Z.; Fu, Y.; Zhang, F.; Liu, L.; Zhang, N.; Zhou, D.; Yang, J.; Pang, Y.; Huang, Y. Spinning Micropipette Liquid Emulsion Generator for Single Cell Whole Genome Amplification. *Lab Chip* **2016**, *16*, 4512−4516.

(24) Telenius, H. K.; Carter, N. P.; Bebb, C. E.; Nordenskjold, M.; Ponder, B. A.; Tunnacliffe, A. Degenerate Oligonucleotide-Primed PCR: General Amplification of Target DNA by a Single Degenerate Primer. *Genomics* **1992**, *13*, 718−725.

(25) Dean, F. B.; Nelson, J. R.; Giesler, T. L.; Lasken, R. S. Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. *Genome Res.* **2001**, *11*, 1095−1099.

(26) Wang, J.; Fan, H. C.; Behr, B.; Quake, S. R. Genome-Wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. *Cell* **2012**, *150*, 402−412.

(27) Lu, S.; Zong, C.; Fan, W.; Yang, M.; Li, J.; Chapman, A. R.; Zhu, P.; Hu, X.; Xu, L.; Yan, L.; Bai, F.; Qiao, J.; Tang, F.; Li, R.; Xie, X. S. Probing Meiotic Recombination and Aneuploidy of Single Sperm Cells by Whole-Genome Sequencing. *Science* **2012**, *338*, 1627−1630.

(28) Zong, C.; Lu, S.; Chapman, A. R.; Xie, X. S. Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell. *Science* **2012**, *338*, 1622−1626.

(29) Fu, Y.; Li, C.; Lu, S.; Zhou, W.; Tang, F.; Xie, X. S.; Huang, Y. Uniform and Accurate Single-Cell Sequencing Based On Emulsion Whole-Genome Amplification. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 11923−11928.

(30) Shoukri, M. M. On a Generalization for the Double Poisson Distribution. *Communications in Statistics -Theory and Methods* **1982**, *11*, 151−164.

(31) Srivastava, R. C. A Note On the Rao-Rubin Characterization of the Poisson-Distribution. *SIAM J. Appl. Math.* **1982**, *42*, 261−265.

(32) Hindson, B. J.; Ness, K. D.; Masquelier, D. A.; Belgrader, P.; Heredia, N. J.; Makarewicz, A. J.; Bright, I. J.; Lucero, M. Y.; Hiddessen, A. L.; Legler, T. C.; Kitano, T. K.; Hodel, M. R.; Petersen, J. F.; Wyatt, P. W.; Steenblock, E. R.; Shah, P. H.; Bousse, L. J.; Troup, C. B.; Mellen, J. C.; Wittmann, D. K.; Erndt, N. G.; Cauley, T. H.; Koehler, R. T.; So, A. P.; Dube, S.; Rose, K. A.; Montesclaros, L.; Wang, S.; Stumbo, D. P.; Hodges, S. P.; Romine, S.; Milanovich, F. P.; White, H. E.; Regan, J. F.; Karlin-Neumann, G. A.; Hindson, C. M.; Saxonov, S.; Colston, B. W. High-Throughput Droplet Digital PCR System for Absolute Quantitation of DNA Copy Number. *Anal. Chem.* **2011**, *83*, 8604−8610.

(33) Belgrader, P.; Tanner, S. C.; Regan, J. F.; Koehler, R.; Hindson, B. J.; Brown, A. S. Droplet Digital PCR Measurement of Her2 Copy Number Alteration in Formalin-Fixed Paraffin-Embedded Breast Carcinoma Tissue. *Clin. Chem.* **2013**, *59*, 991−994.

(34) Heredia, N. J.; Belgrader, P.; Wang, S.; Koehler, R.; Regan, J.; Cosman, A. M.; Saxonov, S.; Hindson, B.; Tanner, S. C.; Brown, A. S.;

DOI: 10.1021/acsami.7b03146
ACS Appl. Mater. Interfaces 2017, 9, 13958−13964

Karlin-Neumann, G. Droplet Digital (Tm) PCR Quantitation of Her2 Expression in FFPE Breast Cancer Samples. *Methods* **2013**, *59*, S20−S23.
(35) Vogelstein, B.; Kinzler, K. W. Digital PCR. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96*, 9236−9241.
(36) Kawai, K.; Viars, C.; Arden, K.; Tarin, D.; Urquidi, V.; Goodison, S. Comprehensive Karyotyping of the HT-29 Colon Adenocarcinoma Cell Line. *Genes, Chromosomes Cancer* **2002**, *34*, 1−8.
(37) Langmead, B.; Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357−359.
(38) Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map Format and Samtools. *Bioinformatics* **2009**, *25*, 2078−2079.
(39) Baslan, T.; Kendall, J.; Rodgers, L.; Cox, H.; Riggs, M.; Stepansky, A.; Troge, J.; Ravi, K.; Esposito, D.; Lakshmi, B.; Wigler, M.; Navin, N.; Hicks, J. Genome-Wide Copy Number Analysis of Single Cells. *Nat. Protoc.* **2012**, *7*, 1024−1041.

DOI: 10.1021/acsami.7b03146
ACS Appl. Mater. Interfaces 2017, 9, 13958−13964

Supporting Information

A single-cell-based platform for copy number variation profiling through digital counting of amplified genomic DNA fragments

Chunmei Li[†], *Zhilong Yu*[†], *Yusi Fu*[†], *Yuhong Pang*[†], *and Yanyi Huang**[†‡ǁ]

[†]Beijing Advanced Innovation Center for Genomics (ICG), Biodynamic Optical Imaging Center (BIOPIC), School of Life Sciences, Peking University, Beijing 100871, China.

[‡]College of Engineering, Peking University, Beijing 100871, China

[ǁ]Peking-Tsinghua Center for Life Sciences, Peking University, Beijing 100871, China.

* To whom correspondence should be addressed. E-mail: yanyi@pku.edu.cn

I. Materials and Methods

Cells and reagents

The HT-29 cells, expanded from a monoclone, were kindly provided by Professor Wensheng Wei in the School of Life Sciences at Peking University. Semen sample was collected from 50-year-old Asian male volunteers. Random primers (N_6, 33μg dissolved in 80μL water), bovine serum albumin (BSA, 20mg/mL) and deoxyribonucleoside triphosphate (dNTP, solution, 10mM each) were purchased from New England Biolabs China. High purity water was purchased from Ambion. Protease comes from Qiagen. The remaining biochemical reagents used in the amplification reaction were purchased from Sigma Aldrich.

Basis for determining the fraction number

The fraction number should be larger than the number of single-strand DNA, or there will be a chance that a target sequence appears in all the fractions. We will not be able to tell it from contamination. So we choose 2 × (maximum CNV number) + 1 as the fraction number. A larger fraction number is OK but it will make the experiment and library preparation more complicated. The average copy number of HT-29 cell is 3, 15 fractions for HT-29 cell will be able to handle a maximum CNV number of 7, which we think is enough for HT-29 cell. For a single sperm, which has only 2 single-strand DNA, a fraction number of 3 is enough. However, in the experiment we need to make sure that we have get only one sperm cell but not two. Also we need to make sure the experiment is not contaminated by diploid cells. So the experiment is designed to be distinguish a copy number of 2, which means 4 single-strand DNA. So we choose at least 5 fractions for single sperm.

HT-29 single cell multiple fractions MDA

HT-29 cells for MDA were washed in PBS (Invitrogen), and then dispersed into single cell suspension by gently pipetting. Each morphological good single cell was picked by mouth pipette into 1μL cell lysis buffer (30 mM Tris-HCl, 10 mM KCl, 5 mM EDTA, 0.5% Triton-X100, and 2 mg/mL protease, pH = 8.0) followed by incubation at 50 °C for 180 min. Then, the protease was thermally inactivated by raising the temperature to 70°C for 30 minutes. After that, 1μL of Phi29 polymerase reaction buffer (50 mM Tris-HCl, 10 mM

$MgCl_2$, 10mM $(NH_4)_2SO_4$, 4 mM DTT, pH 7.5), 5 μL of N_6 random primer and high purity water were added to the lysis mix. The tube was heated to 95 °C for 5 min to denature and fragment double-strand gDNA, then quickly chilled on ice for at least 5 min for keeping DNA as single-strand state and annealing N_6 primers. Then we added 0.8 μL of Phi29 polymerase, 1.0 μL of dNTP mix, and 0.2 μL of BSA to the mix on ice. Immediately, the total reaction solution of 10 μL was equally divided into 15 tubes at 4 °C, and MDA reactions were carried out at 30 °C. After 10-hour amplification, reactions were terminated at 65°C for 10 min.

Sperm single cell multiple fractions MDA

Sperm single cell multiple fractions MDA were carried out under the same procedure as HT-29 single cell above except for two major differences. One was cell lysis buffer and time. Each morphological good individual spermatozoa was picked using mouth pipette into 0.5 μL of cell lysis buffer (30 mM Tris-HCl, 10 mM KCl, 5m M EDTA, 0.5% Triton-X100, 40 mM DTT and 2 mg/mL protease, pH=8.0) followed by incubation at 50 ° C for 720 minutes. The other was the number of portioning and amplification time. The total 10 μL reaction mix was equally divided into 5-7 tubes quickly at 4 °C, and incubated at 30 °C for 12 hours.

Purification & Quality control

Each tube of MDA product was purified separately using DNA Clean-up & Concentration kit (Zymo Research) after addition of 50 μL water to the reaction mix. Purified DNA was firstly quantified by Qubit dsDNA HS Assay (Invitrogen). Then, same amount of DNA from each tube belonging to same single cell were pooled together as qPCR template to examine the amplification bias. In total, 5 primers which targeted to different chromosomes were applied for quantification with PCR SsoAdvanced SYBR Green Supermix (Biorad) on Illumina Eco thermocycler. Amplification is considered successful when at least 3 primers can result in correct products (confirmed by melting curve analysis) and Ct value lower than 30. These single cell products were chosen for next step of library preparation.

Library preparation

For bulk cells, 1μg gDNA was used to construct the sequencing library according to standard procedure of NEBNext Ultra DNA Library Prep Kit (NEB). For multiple fractions MDA of

single cells, 1ng DNA of each tube was used to perform library construction with Nextera DNA Library Preparation Kit (Illumina). All qualified libraries were pooled together and sequenced on Illumina Hiseq2500 platform.

II. Theoretical Simulation

Distribution of original single-strand copies in multiple fractions

In order to know the probability of a number of ssDNA fragments divided into N sub-systems, we firstly consider the probability that 2 ssDNA fragments are divided into N fractions. The probability of 2 ssDNA copies distributed into the same fraction is:

$$p_{N,2,1} = 1/N \quad (1)$$

The probability of 2 ssDNA copies distributed into different fractions is:

$$p_{N,2,2} = 1 - 1/N \quad (2)$$

Similarly, the probability of 3 ssDNA copies distributed into 1 fraction is:

$$p_{N,3,1} = 1/N^2 \quad (3)$$

The probability of 3 ssDNA copies distributed into 2 fractions is:

$$p_{N,3,2} = 3(N\text{-}1)/N^2, \quad (4)$$

The probability of 3 ssDNA copies distributed into 3 different fractions is:

$$p_{N,3,3} = (N\text{-}1)(N\text{-}2)/N^2 \quad (5)$$

If the number of ssDNA copies is i, the probability that they are distributed into j fractions is $p_{N,i,j}$, we have:

$$p_{N,i,1} = 1/N^{(i\text{-}1)} \quad (6)$$

$$p_{N,i,i} = \left(\prod_{k=1}^{i-1}(N\text{-}k)\right)/N^{(i\text{-}1)} \quad (7)$$

$$p_{N,i,j} = p_{N,i\text{-}1,j}{}^{*}j/N + p_{N,i\text{-}1,j\text{-}1}{}^{*}(N\text{-}j+1)/N \quad (8)$$

For a typical diploid cell which has 4 single-strand copies, the result of distribution into multiple fractions was shown in **Table S1**. When the number of total fractions is fixed, a different number of ssDNA copies will result in different distributions. **Table S2** shows the probability distributions of 2, 4, 6, and 8 ssDNA copies (corresponding to copy numbers 1, 2, 3, and 4) dispersed into 20 fractions.

The probability of an ssDNA copy been amplified successfully

To simplify the model, we used haploid sperm cells as samples. We distributed the single sperm reaction mix into five or seven fractions. After amplification, the exons were enriched and sequenced. In order to obtain accurate and reliable results, we selected unique SNP sites of the individual in the exon group for analysis to get rid of potential contaminations. Sperm

1 was distributed into 5 fractions. 20790 SNP sites were detected in only 1 of the 5 fractions, and 4599 sites in 2 of the 5 fractions. Sperm 2 was distributed into 7 fractions. 19054 SNP sites were detected in only 1 of the 7 fractions, and 3216 sites in 2 of the 7 fractions.

When the probability of an ssDNA copy been amplified successfully is r, for sperm 1, the probability of 2 single-strand copies separated into 2 fractions is:

$$p_{5,2,2} = 1 - 1/5 = 0.8 \tag{9}$$

If the ssDNA copy number is l, the probability of m ssDNA copies being amplified $q_{l,m}$ can be calculated using binomial distribution:

$$q_{l,m} = \mathrm{C}_l^m r^m (1-r)^{l-m} \tag{10}$$

For sperm cells, the probability of 2 ssDNA copies both being amplified is

$$q_{2,2} = \mathrm{C}_2^2 r^2 (1-r)^0 = r^2 \tag{11}$$

Combined with the probability that 2 ssDNA copies are separated into 2 fractions, the probability that 2 ssDNA copies are detected in 2 fractions is

$$t_{2,2} = p_{5,2,2} * q_{2,2} = 0.8\, r^2 \tag{12}$$

If both of the 2 ssDNA copies fail to be amplified, the copy will be detected in none of the 5 fractions. The probability is t2,0=(1-r)2. So the probability that a certain copy is detected in only 1 fraction is

$$t_{2,1} = 1 - t_{2,2} - t_{2,0} = 1 - (1-r)^2 - 0.8\, r^2 \tag{13}$$

Combined with experimental value:

$$\frac{t_{2,1}}{t_{2,2}} = \frac{1-(1-r)^2 - 0.8\, r^2}{0.8\, r^2} = \frac{20790}{4599} \tag{14}$$

We will have

$$r = 0.37 \tag{15}$$

For sperm 2:

$$\frac{t_{2,1}}{t_{2,2}} = \frac{1-(1-r)^2 - 0.85\, r^2}{0.85\, r^2} = \frac{19054}{3216} \tag{16}$$

We have

$$r = 0.30 \tag{17}$$

According to the result and analysis above, we find that in our experiments, the probability for an ssDNA copy being amplified is about 0.3-0.4. When the probability is 0.4, for a haploid cell, the theoretical maximum coverage of MDA is 0.64, and for diploid cells, the maximum coverage is 0.87. Similarly, for triploid and tetraploid cells, the maximum

coverage is 0.95 and 0.98.

Positive values and copy number

In the experiment, there are l original ssDNA copies, and they are distributed into N fractions. After amplification, if the copy is detected in u of N fractions, we call the copy has a positive value of u. The probability that a copy has a positive value of u is

$$t_{l,u} = \sum_{m=u}^{l} \left(q_{l,m} * p_{N,m,u} \right) \quad (18)$$

If the probability for an ssDNA copy being amplified r is 0.4, the probability of different positive values when 2-8 ssDNA copies are distributed into 20 fractions is shown in **Table S3**. According to Table S3, a single fragment with a given number of ssDNA copies may have different positive values. For a number of fragments which share the same copy number, if we count the number of all the positive values, we will find the percentage of each positive values get more and more closed to the probability shown in Table S3 as the number of fragment grows. For example, a region which has M fragments with the same copy number is analyzed. The fragments are distributed and amplified independently, which can be treated as N Bernoulli trials. The result shows that e_0 fragments have a positive value of 0, e_1 fragments have a positive value of 1, e_2 fragments have a positive value of 2… and e_n fragments have a positive value of n. The frequency of a positive value n across the region f_n is e_n / M. For multiple regions the f_n should follow the multinomial distribution. **Figure S1** shows the probability distributions of frequencies for different positive values in a region containing 20 fragments. With different original copy numbers, the frequencies of different positive values will be different. The more fragments the region contains, the more Bernoulli trials there will be, the probability distributions of frequencies for different positive values would be more close to the mean. **Figure S2** shows the probability distributions of frequencies for different positive values in a region containing 50 fragments and **Figure 2** shows the probability distributions of frequencies for different positive values in a region containing 100 fragments.

III. Whole genome amplification of single HT-29 cells

From read depth to positive values

The insert size was 200-300 bp during the library construction and the read length is 2*150 bp using paired-end sequencing. We choose 200 bp as the size of a fragment. Each fragment is treated as a locus. If the starting position of a read falls within the range of a fragment, the read depth of the locus is increased by one.

For each fraction, the read depth of each locus was summed across the entire genome, with an average depth of 27.7. To remove potential contamination, only locus with a reading depth >= 3 will pass the filter. Then for each fraction, we count the positive values of all the fragments.

Based on our calculation described above, when the probability that an ssDNA copy being successfully amplified is 0.4, the probability that at least 10 out of 12 ssDNA copies of a fragment are amplified is:

$$C_{12}^{10}*0.4^{10}(0.6)^2 + C_{12}^{11}*0.4^{11}(0.6)^1 + 0.4^{12} = 0.003 \quad (19)$$

Since this probability is small, and the Ht-29 cell is average triploid (6 ssDNA copies), we set another filter to remove those fragments with positive values larger than 10, which may be contamination or the none-specific regions.

Low resolution Copy Number Variation

Low resolution CNV regions are needed for training the frequency probability distribution of different positive values. The data of multiple fractions are merged and treated as a conventional MDA experiment. We use bin sizes from 0.05 M to 2.50 M and get the CNV result using DNACopy package. Then we calculate the coefficient of variation (CV) of bin read depth in the largest 50% of CNV regions, and get the size-weighted normalization as the final result. As shown in **Figure S3**, as the bin size grows, the CV value drops first, then raises, and finally drops again. That is because when the bin size is small, the random noise of MDA affects the accuracy of CNV segmentation. The noise gets smaller as the bin size grows. However, when the bin size is too large, regions with different copy number may be put into a single bin and makes CNV calling inaccurate, leading to an increase in CV result. We find that the CV result has a local minimal value when the bin size is 0.7-0.8 M. So we

choose 0.7M as the bin size for calling low resolution CNV. We select CNV regions with a size larger than 20M, as shown in **Table S4**, to generate the probability distribution matrix. The probability distribution is generated according to the processing region size *s*. For example, s can be 20, 100 or 500 kb, which means a single region contains 100, 500 or 2500 fragments. For each low resolution CNV region, regions are sample from the beginning to the end with an increment of starting point by 200 bp. For each region sampled, the frequency of 0-8 positive value is obtained. And the probability distribution of different positive values can be obtained after all the low resolution CNV region with different CNVs are processed. Each probability distribution is smoothed using moving average method with a window size of 0.04 to remove noise. The probability distribution is denoted as:

$$f(c, p, e) \tag{20}$$

c is for the copy number, and e is the probability density of positive value p in the region.

Get copy number using maximum likelihood

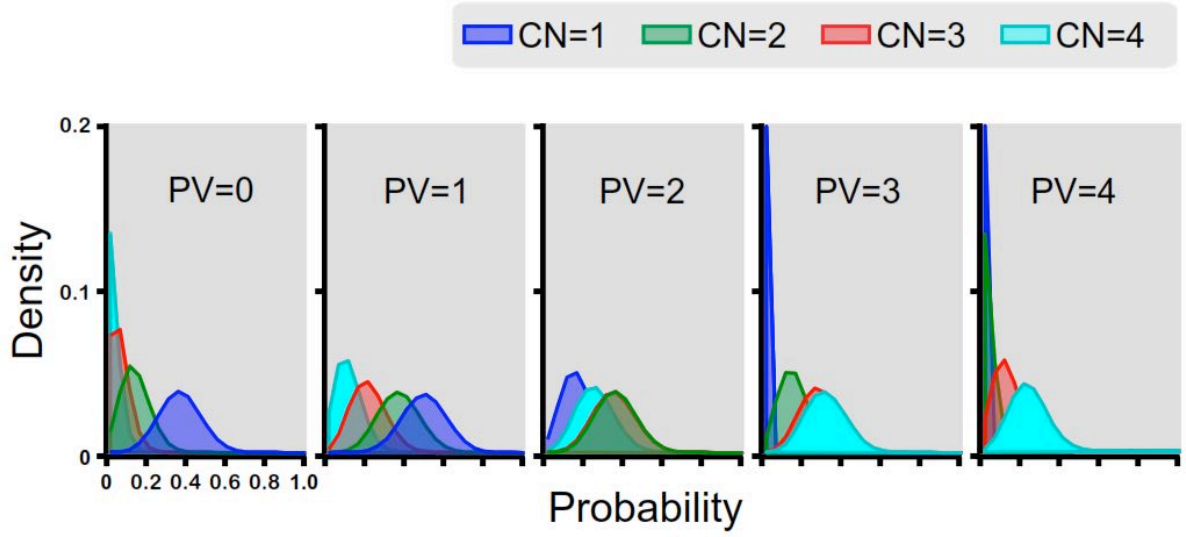
A region is sampled from the genome data in a given size. The frequencies of positive values 0 – 8 are calculated as $e_0, e_1, \ldots e_8$. For each possible copy number c, the probability is:

$$P_c = \prod_{i=0}^{8} f(c, i, e_i) \tag{21}$$

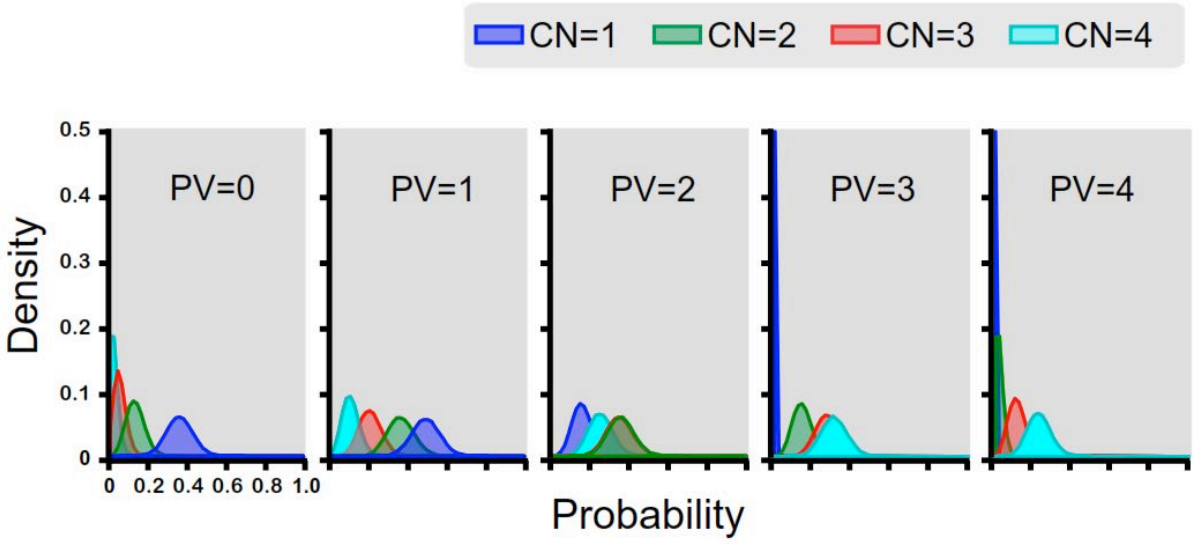
The copy number with max probability P_c is the most possible CNV for the region.

Time to analyze

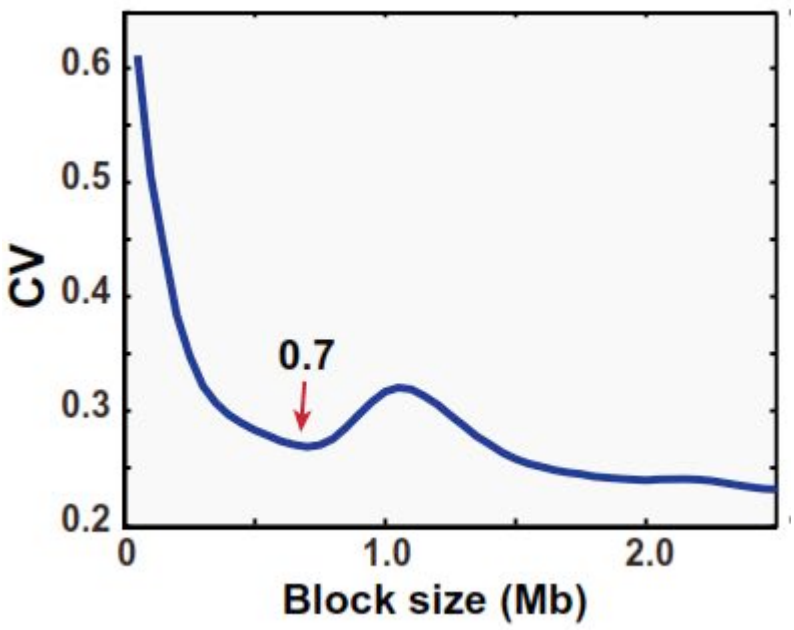
The sequencing data is mapped and BAM files are generated. Starting from BAM files, our naïve MATLAB scripts take about 20 minutes to process a whole human genome with a 3.4GHz i7-3770 CPU computer. The code is rewrite in C and a 5X speedup is obtained. That means the whole analysis can be done in less than 4 minutes. The time can be further reduced with multithreading optimization.



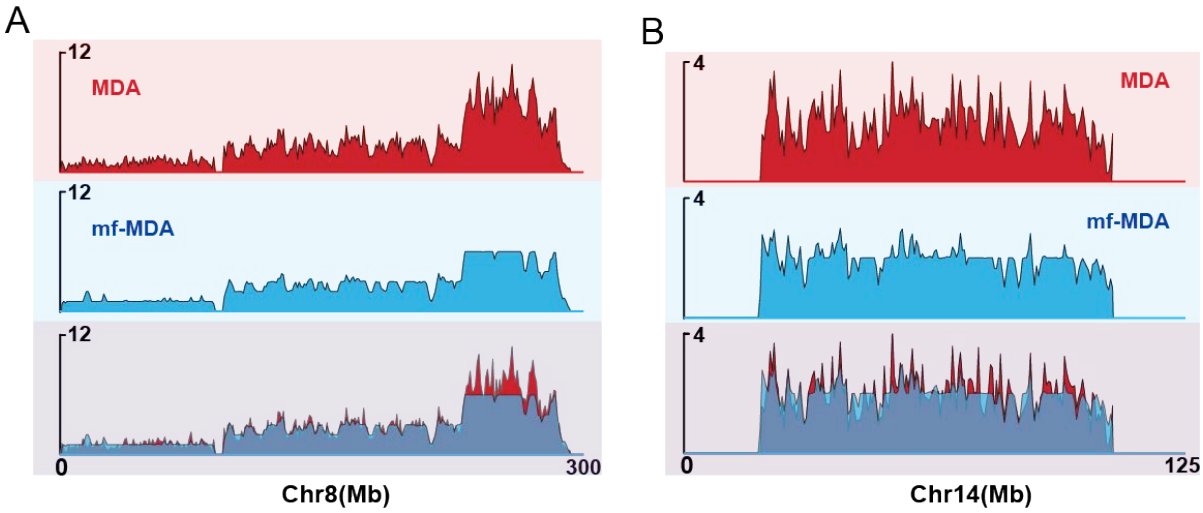
Supplementary Figure S1. The probability distributions of frequencies for different positive values (PV) in a region containing 20 fragments. Positive values are 0, 1, 2, 3, and 4 from left to right. The color blue, green, red, and cyan represent regions with ssDNA copy number of 2, 4, 6, and 8. Different original ssDNA copy numbers will result in different positive value distribution.



Supplementary Figure S2. The probability distributions of block containing 50 fragments with a positive value of 0 to 4.



Supplementary Figure S3. The coefficient of variation (CV) of read depth in each block. The CV value move down quickly when the block size is smaller than 0.7 M, and move up when the block size larger than 0.8M. Then the CV value drops below 0.25, and the noise drop has been slower.



Supplementary Figure S4. The results of maximum likelihood method in chromosome 8 (A) and chromosome 14 (B). There was a significant reduction in the number of copies obtained using multiple copies of MDA compared to a single copy of MDA at a block size of 500 kb.

Supplementary Table S1. The result of the distribution for 4 ssDNA copies distributed into multiple fractions.

Number of positive fractions* / Number of total fractions	1	2	3	4
4	0.09	0.56	0.33	0.02
5	0.19	0.58	0.22	0.01
6	0.28	0.56	0.16	0.00
7	0.35	0.52	0.12	0.00
8	0.41	0.49	0.10	0.00
9	0.46	0.46	0.08	0.00
10	0.50	0.43	0.06	0.00
11	0.54	0.41	0.05	0.00
12	0.57	0.38	0.04	0.00
13	0.60	0.36	0.04	0.00
14	0.63	0.34	0.03	0.00
15	0.65	0.32	0.03	0.00
16	0.67	0.31	0.03	0.00
17	0.68	0.229	0.02	0.00
18	0.70	0.28	0.02	0.00
19	0.71	0.27	0.02	0.00
20	0.73	0.26	0.02	0.00

* Fractions which contain at least 1 original copy.

Supplementary Table S2. The result of the distribution for multiple ssDNA copies distributed into 20 fractions.

Number of positive fractions / Number of ssDNA copies	1	2	3	4	5	6	7	8
2	0.05	0.95	-	-	-	-	-	-
4	0.00	0.02	0.26	0.73	-	-	-	-
6	0.00	0.00	0.01	0.12	0.44	0.44	-	-
8	0.00	0.00	0.00	0.01	0.08	0.29	0.43	0.20

Supplementary Table S3. The probability of positive values for a single fragment with different ssDNA copies in mfMDA.

Number of ssDNA Copies \ Positive Values	0	1	2	3	4	5	6	7	8
2	0.36	0.49	0.15	-	-	-	-	-	-
4	0.13	0.36	0.35	0.14	0.02	-	-	-	-
6	0.05	0.20	0.34	0.27	0.11	0.02	0.00	-	-
8	0.02	0.10	0.24	0.30	0.22	0.09	0.02	0.00	0.00

Supplementary Table S4. The low resolution CNV regions used for training the probability distribution matrix.

Copy number	chromosome	Start	End	Length
1	8	1	42	42
2	3	1	58	58
2	4	161	185	25
2	6	67	156	90
2	14	26	105	80
2	18	22	76	55
2	21	16	47	32
3	1	25	119	95
3	1	152	204	53
3	2	137	240	104
3	4	1	48	48
3	4	55	146	92
3	5	59	179	121
3	6	2	55	54
3	8	52	85	34
3	9	75	138	64
3	10	55	132	78
3	12	44	129	86
4	1	213	246	34
4	2	4	90	87

4	3	95	128	34
4	3	152	196	45
4	7	24	54	31
4	7	66	155	90
4	11	2	47	46
4	11	84	133	50
4	13	24	111	88
4	15	29	98	70
4	19	30	57	28
4	20	33	53	21
6	8	118	142	25