



# DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

## Genome-Wide Detection of Single-Nucleotide and Copy-Number Variations of a Single Human Cell

The Harvard community has made this article openly available. [Please share](#) how this access benefits you. Your story matters.

<b>Citation</b>	Zong, Chenghang, Alec R. Chapman, and X. Sunney Xie. 2012. Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. <i>Science</i> 338(6114): 1622-1626.
<b>Published Version</b>	<a href="https://doi.org/10.1126/science.1229164">doi:10.1126/science.1229164</a>
<b>Accessed</b>	February 19, 2015 11:49:04 AM EST
<b>Citable Link</b>	<a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:10386977">http://nrs.harvard.edu/urn-3:HUL.InstRepos:10386977</a>
<b>Terms of Use</b>	This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a>

*(Article begins on next page)*

# Genome-Wide Detection of Single Nucleotide and Copy Number Variations of a Single Human Cell

Chenghang Zong<sup>1,†</sup>, Sijia Lu<sup>1,†#</sup>, Alec R. Chapman<sup>1,2,†</sup>, X. Sunney Xie<sup>1,\*</sup>

<sup>1</sup>Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA

<sup>2</sup>Program in Biophysics, Harvard University, Cambridge, MA 02138, USA

<sup>†</sup>These authors contributed equally to the work.

<sup>#</sup>Current address: Yikon Genomics Inc., 1 China Medical City Ave, TQB building 5th floor, Taizhou, Jiangsu, China

<sup>\*</sup>To whom correspondence should be addressed. E-mail: [xie@chemistry.harvard.edu](mailto:xie@chemistry.harvard.edu)

## ABSTRACT

Kindred cells can have different genomes because of dynamic changes in DNA. Single cell sequencing is needed to characterize these genomic differences but has been hindered by whole-genome amplification bias, resulting in low genome coverage. Here we report a new amplification method: Multiple Annealing and Looping Based Amplification Cycles (MALBAC) that offer high uniformity across the genome. Sequencing MALBAC amplified DNA achieves 93% genome coverage  $\geq 1x$  for a single human cell at 25x mean sequencing depth. We detected digitized copy number variations (CNVs) of a single cancer cell. By sequencing three kindred cells, we were able to call individual single nucleotide variations (SNVs) with no false positives observed. We directly measured the genome-wide mutation rate of a cancer cell line and found that purine-pyrimidine exchanges occurred unusually frequently among the newly acquired SNVs.

### Summary:

A new whole genome amplification method with significantly reduced bias allows simultaneous accurate detection of point mutations and copy number variations in single mammalian cells and the direct measurement of mutation rates.

Single molecule and single cell studies reveal behaviors that are hidden in bulk measurements (1, 2). In a human cell, the genetic information is encoded in 46 chromosomes. The variations occurring in these chromosomes, such as single nucleotide variations (SNVs) and copy number variations (CNVs) (3), are the driving forces in biological processes such as evolution and cancer. Such dynamic variations are reflected in the genomic heterogeneity among a population of cells, which demands characterization of genomes at the single cell level (4-6). Single cell genomics analysis is also necessary when the number of cells available is limited to few or one, such as prenatal testing samples (7, 8), circulating tumor cells (9), and forensic specimens (10).

Prompted by rapid progress in next generation sequencing techniques (11), there have been several reports on whole genome sequencing of single cells (12-16). These methods have relied on whole genome amplification (WGA) of an individual cell to generate enough DNA for sequencing (17-21). However, WGA methods in general are prone to amplification bias, which results in low genome coverage. PCR-based WGA introduces sequence-dependent bias because of the exponential amplification with random primers (17, 18, 22). Multiple Displacement Amplification (MDA), which uses random priming and the strand-displacing phi29 polymerase under isothermal condition (19), has provided improvements over PCR-based methods but still exhibits considerable bias, again due to nonlinear amplification.

Here we report a new WGA method, Multiple Annealing and Looping Based Amplification Cycles (MALBAC), which introduces quasi-linear preamplification to reduce the bias associated with nonlinear amplification. Picograms of DNA fragments (~10 to 100kb) from a single human cell serve as templates for amplification with MALBAC (Fig. 1). The amplification is initiated with a pool of random primers, each having a common 27-nucleotide sequence and 8 variable nucleotides that can evenly hybridize to the templates at 0°C. At an elevated temperature of 65°C, DNA polymerases with strand displacement activity are used to generate semiamplicons with variable lengths (0.5-1.5kb), which are then melted off from the template at 94°C. Amplification of the semiamplicons give full amplicons which have complementary ends. The temperature is cycled to 58°C to allow the looping of full amplicons, which prevents further amplification and cross hybridizations. Five cycles of preamplification are followed by exponential amplification of the full amplicons by PCR in order to generate micrograms of DNA required for next generation sequencing (Fig. 1). In the PCR, oligos with the common 27-nucleotide sequence are used as the primers.

We used MALBAC to amplify the DNA of single SW480 cancer cells. With  $\sim 25x$  mean sequencing depth, we consistently achieved  $\sim 85\%$  and up to  $93\%$  genome coverage at  $\geq 1x$  depth on either strand (Fig. 2A). As a comparison, we performed MDA on a single cell from the same cancer cell line. At  $25x$  mean sequencing depth, MDA covered  $72\%$  of the genome at  $\geq 1x$  coverage. While significant variations of the coverage have been reported for MDA (15, 16, 20, 23), MALBAC coverage is reproducible.

We use Lorenz curves to evaluate coverage uniformity along the genome. Here, we plotted the cumulative fraction of the total reads that cover a given cumulative fraction of genome (Fig. 2B). The diagonal line indicates a perfectly uniform distribution of reads, and deviation from the diagonal line indicates an uneven distribution of reads. We compared the Lorenz curves for bulk sequencing, MALBAC, and MDA at  $\sim 25x$  mean sequencing depth (Fig. 2B). It is evident that MALBAC outperforms MDA in uniformity of genome coverage. We also plotted the power spectrum of read density variations to show the spatial scale at which the variations take place. For MDA, large amplitudes at low frequencies ( $1/\text{genome distance}$ ) were observed, indicating that large contiguous regions of the genome are over- or under-amplified. In contrast, MALBAC has a power spectrum similar to that of the unamplified bulk.

CNVs due to insertions, deletions, or multiplications of genome segments are frequently observed in almost all categories of human tumors (13, 24, 25). MALBAC's lack of large-scale bias makes it amenable to probing CNVs in single cells. We determined the digitized CNVs across the whole genomes of three individual cells from the SW480 cancer cell line (Fig. 3A-C). CNVs of five cells are included in the SOM (Supplemental Online Material). The chromosomes exhibit distinct CNV differences among the three individual cancer cells and in the bulk result (Fig. 3D), which are difficult to resolve by MDA (Fig. 3E). For the MALBAC data, we used a hidden Markov model to quantify CNVs (SOM). We confirmed the gross features of CNVs detected by MALBAC with a previously published karyotyping study (26). For example, both MALBAC-based quantification of CNVs and spectral karyotyping show one copy of chromosome 18 and three copies of chromosome 17 in the SW480 cancer cell line. Although the majority of copy numbers are consistent between single cells, we also observe cell-to-cell variations as labeled by the dashed box in Fig. 3.

Attempts have been made recently to identify SNVs from a single cell by MDA (15, 16, 23). The first challenge in accurate SNV calling from a single cell is substantial human contamination from the environment and the operators, given picograms of DNA from a single human cell. The second challenge is low detection yield (high false negative), particularly where alleles drop out due to amplification bias. The third challenge is false positives associated with amplification and sequencing errors, either random or systematic (27).

To meet the first challenge, we took special precautions to decontaminate with UV radiation before each experiment was conducted in a restricted clean room. An alternative approach to reduce contamination is microfluidics (28).

With regard to the second challenge, MALBAC allowed us to call  $2.2 \times 10^6$  single cell SNVs compared with  $2.8 \times 10^6$  detected SNVs in bulk, yielding a 76% detection efficiency, in contrast to 41% with MDA (Table 1). This improvement resulted from improved uniformity by MALBAC (SOM, Fig S6). Listed separately in Table 1 are heterozygous and homozygous SNVs. Next we calculate the allele dropout rate. Comparison of single-cell and bulk SNVs showed that 7,288 of the SNVs genotyped as homozygous mutations by MALBAC are actually heterozygous in bulk, which corresponds to a ~1% allele dropout rate in MALBAC (SOM). In contrast, with MDA we found 172,563 incorrect homozygous calls, corresponding to an allele dropout rate of ~65% (SOM).

Compared to the bulk data, the MALBAC data contains  $1.1 \times 10^5$  false positives (Table 1) out of  $3 \times 10^9$  bases in the genome. This corresponds to a  $\sim 4 \times 10^{-5}$  false positive rate, which is due to the errors made by the polymerases in the semi-amplicons generated in first MALBAC cycle and propagated in the later amplification. Although improving the polymerase's error rate is possible, our strategy to reduce the false positive rate was to sequence two or three kindred cells derived from the same cell. The simultaneous appearance of an SNV in the kindred cells would indicate a true SNV. The false positive rate due to uncorrelated random errors can be reduced to  $\sim 10^{-8}$  with two kindred cells and  $\sim 10^{-12}$  with three kindred cells.

However, there are false positives due to correlated errors i.e. systematic sequencing and amplification errors. We filtered out these errors by comparing two unrelated single cells

that are not from the same lineage (SOM, Figure S5). After this procedure, we can call true SNVs of a single cell with no false positives observed (Table 2).

To gain insight into the mutation process in the cancer cells, we clonally expanded a single ancestor cell picked from a heterogeneous population of the SW480 cancer cell line for 20 generations (Fig. 4A). We extracted DNA from this single cell clonal expansion for bulk sequencing, which reflects the genome of the ancestor cell. We then picked a single cell from this clone. To detect SNVs acquired by the cell during expansion, we grew another four generations to obtain the kindred cells denoted C1 to C16. We individually sequenced three kindred cells, C1, C2, and C3 after MALBAC amplification. After filtering correlated and uncorrelated errors (Fig. 4B), we detected 35 unique SNVs shown in Fig. 4C.

We randomly chose 8 out of a total of 35 unique SNVs and confirmed that they are neither false positives by Sanger sequencing C4-C6, nor false negatives by Sanger sequencing the bulk (Please refer to SOM for Sanger sequencing data). As an example, Fig. 4D and 4E shows the MALBAC and Sanger sequencing result of one such SNV.

These 35 unique SNVs are newly acquired during the 20 cell divisions. Adjusting for a detection efficiency of 72% for heterozygous SNVs, we estimate that ~49 mutations occurred in the 20 generations, yielding a mutation rate of ~2.5 nucleotides per cell generation, consistent with our estimation based on the bulk data (SOM). The mutation rate of this cancer cell line is about 10 fold higher than the mutation rate estimated based on germ line studies (29-31).

Mutations can be transitions (purine $\leftrightarrow$ purine exchange i.e. A $\leftrightarrow$ G or pyrimidine $\leftrightarrow$ pyrimidine exchange, i.e. C $\leftrightarrow$ T) or transversions (purine  $\leftrightarrow$  pyrimidine exchanges, i.e. A/G $\leftrightarrow$ C/T). Transitions are more common. Surprisingly, we found that the transition/transversion (tstv) ratio for the 35 newly acquired SNVs detected is only 0.30, whereas the ratio for the total SNVs of this cell line is 2.01, as expected for common human mutations (32). To further confirm that this observation is not due to single cell amplification, we sequenced the bulk DNA of the original heterogeneous culture (SOM). The tstv ratio for SNVs detected in the single cell expanded bulk but not in the original heterogeneous bulk was 0.75. Both significantly low tstv ratios indicate that transitions are not favored over transversion for newly acquired SNVs in this cancer cell line (SOM). While understanding the underlying mechanism of this phenomenon will require

similar measurements in other systems, it is evident that, by allowing precise characterization of CNVs and SNVs, MALBAC can shed light on the individuality, heterogeneity, and dynamics of the genomes of single cells.

**Table 1: Comparison of Single cell SNVs for bulk, MDA and MALBAC**

	Heterozygous SNVs	Homozygous SNVs	Total SNVs
Bulk			
SNVs	911,958	1,930,204	2,842,162
Single cell MDA			
SNVs	93,140 (2,828)*	1,238,286 (1,973)	1,331,426 (4,801)
Detection efficiency	10%	63%	41%
Single cell MALBAC			
SNVs	756,812 (108,481)	1,539,326 (6,821)	2,296,138 (115,302)
Detection efficiency	71%	80%	76%

\*The number in the bracket indicates the number of false positives.

**Table 2: MALBAC calling of total SNVs and newly acquired SNVs using two and three kindred cells**

	Heterozygous SNVs	Homozygous SNVs	Total SNVs
Two kindred cells			
SNVs	615,387	1,322,555	1,937,942
Detection efficiency	67%	68%	68%
Newly acquired SNVs	145 (~100)*	3 (~0)	148 (~100)
Three kindred cells			
SNVs	660,246	1,577,798	2,238,044
Detection efficiency	72%	81%	80%
Newly acquired SNVs	30 (~0)	5 (~0)	35 (~0)

\*The number in the bracket indicates the number of false positives.



## Figure Captions:

**Figure 1:** MALBAC single cell whole genome amplification. A single cell is picked and lysed. First, genomic DNA of the single cell is melted into single-stranded DNA molecules at 94°C. MALBAC primers then anneal randomly to single-stranded DNA molecules at 0°C and are extended by a polymerase with displacement activity at elevated temperatures, creating semi-amplicons. In the following five temperature cycles, after the step of looping the full amplicons, single stranded amplicons and the genomic DNA are used as template to produce full amplicons and additional semi-amplicons, respectively. For full amplicons, the 3' end is complementary to the sequence on the 5' end. The two ends hybridize will form the looped DNA, which can efficiently prevents the full amplicon from being used as template, therefore warrant a close-to-linear amplification. After the five cycles of linear preamplification, only the full amplicons can be exponentially amplified in the following PCR using the common 27-nucleotide sequence as the primer. PCR reaction will generate microgram level of DNA material for sequencing experiments.

**Figure 2.** Characterization of amplification uniformity. **(A)** Histograms of reads over the entirety of Chromosome 1 of a single cell from the SW480 cancer cell line and the zoom-in of a ~8 million base region (chr1: 62,023,147-70,084,845). **(B)** Lorenz curves of MALBAC, MDA and bulk sample. A Lorenz curve gives the cumulated fraction of reads as a function of the cumulated fraction of genome. Perfectly uniform coverage would result in a diagonal line and a large deviation from the diagonal is indicative of a biased coverage. The green and blue arrows indicate the uncovered fractions of the genome for MALBAC and MDA respectively. All samples are sequenced at 25x depth. **(C)** Power spectrum of read density throughout the genome (as a function of spatial frequency). MALBAC performs similarly to bulk, while the MDA spectrum shows high amplitude at low frequency, demonstrating that regions of several megabases suffer from under- and over- amplification. This observation is consistent with the variations of read depth in Fig. S3 (SOM).

**Figure 3.** CNVs of single cancer cells. Digitized copy numbers across the genome are plotted for three single cells (Panel **A** to **C**) as well as the bulk sample (Panel **D**) from the SW480 cancer cell line. The bottom panel shows the result based on MDA amplification (Panel **E**). Green lines are fitted CNV numbers obtained from the hidden Markov model (SOM). The single cells are sequenced at only 0.8x depth, while the bulk and MDA are done at 25x. More single cells' CNV analyses are included in the SOM (Fig. S4). The regions within the dashed box exhibit the CNV differences among single cells and the bulk, which cannot be resolved by MDA. The binning window is 200kb.

**Figure 4.** Calling newly acquired SNVs and estimation of mutation rate of a cancer cell line (SW480). (**A**) Experiment design. A single ancestor cell is chosen and cultured for ~20 generations. The vast majority of cells are used to extract DNA for bulk sequencing to represent the ancestor cell's genome. A single cell from this culture is chosen for another expansion of four generations. The kindred cells are isolated for single cell whole genome amplification. Single cell sample C1, C2, and C3 are used for high-throughput sequencing. Sample C4, C5, and C6 are used for varying SNVs with Sanger sequencing. (**B**) 3D p-value plot of a one-sided binomial test for SNV candidates from the three kindred cells. The black dots are the false positives due to uncorrelated amplification errors; all of them are on the x-y-z axis and x-y, y-z, x-z planes. Outside of the three planes, the 166 green dots are the residual false positives due to correlated errors from homopolymers, tandem repeats, high-GC content and high density SNV regions, and the 35 red dots are the newly acquired SNVs during the 20 generations of clonal expansion (SOM). We note that the homozygous SNVs are located at the (1,1,1) position. (**C**) Locations of the 35 newly acquired SNVs on the chromosomes of a single cell (SOM). (**D**) Next-generation sequencing data of a newly acquired SNV. The SNV (C→G) exists in the high throughput data of all three kindred cells but not in the bulk data. (**E**) Sanger sequencing data of single cells C4, C5, and C6 confirms that this SNV is not a false positive, while the Sanger sequencing of the bulk confirms that this SNV is not a false negative of next generation sequencing of the bulk (i.e. this SNV is indeed absent in the bulk).

## References

1. M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic gene expression in a single cell. *Science* **297**, 1183 (Aug 16, 2002).
2. G. W. Li, X. S. Xie, Central dogma at the single-molecule level in living cells. *Nature* **475**, 308 (Jul 21, 2011).
3. S. Negrini, V. G. Gorgoulis, T. D. Halazonetis, Genomic instability--an evolving hallmark of cancer. *Nature reviews. Molecular cell biology* **11**, 220 (Mar, 2010).
4. C. Lengauer, K. W. Kinzler, B. Vogelstein, Genetic instabilities in human cancers. *Nature* **396**, 643 (Dec 17, 1998).
5. S. Yachida *et al.*, Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114 (Oct 28, 2010).
6. P. J. Campbell *et al.*, The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109 (Oct 28, 2010).
7. Y. M. Lo *et al.*, Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus. *Science translational medicine* **2**, 61ra91 (Dec 8, 2010).
8. J. O. Kitzman *et al.*, Noninvasive whole-genome sequencing of a human fetus. *Science translational medicine* **4**, 137ra76 (Jun 6, 2012).
9. S. Nagrath *et al.*, Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature* **450**, 1235 (Dec 20, 2007).
10. E. K. Hanson, J. Ballantyne, Whole genome amplification strategy for forensic genetic analysis using single or few cell equivalents of genomic DNA. *Analytical biochemistry* **346**, 246 (Nov 15, 2005).
11. M. L. Metzker, Sequencing technologies - the next generation. *Nature reviews. Genetics* **11**, 31 (Jan, 2010).

12. H. C. Fan, J. Wang, A. Potanina, S. R. Quake, Whole-genome molecular haplotyping of single cells. *Nature biotechnology* **29**, 51 (Jan, 2011).
13. N. Navin *et al.*, Tumour evolution inferred by single-cell sequencing. *Nature* **472**, 90 (Apr 7, 2011).
14. M. Gundry, W. G. Li, S. B. Maqbool, J. Vijg, Direct, genome-wide assessment of DNA mutations in single cells. *Nucleic acids research* **40**, 2032 (Mar, 2012).
15. Y. Hou *et al.*, Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. *Cell* **148**, 873 (Mar 2, 2012).
16. J. Wang, H. C. Fan, B. Behr, S. R. Quake, Genome-wide Single-Cell Analysis of Recombination Activity and De Novo Mutation Rates in Human Sperm. *Cell* **150**, 402 (Jul 20, 2012).
17. L. Zhang *et al.*, Whole genome amplification from a single cell: implications for genetic analysis. *Proceedings of the National Academy of Sciences of the United States of America* **89**, 5847 (Jul 1, 1992).
18. H. Telenius *et al.*, Degenerate oligonucleotide-primed PCR: general amplification of target DNA by a single degenerate primer. *Genomics* **13**, 718 (Jul, 1992).
19. F. B. Dean, Nelson, J.R. , Giesler, T.L. & Lasken, R.S. , Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res.* **11**, 1095 (2001).
20. K. Zhang *et al.*, Sequencing genomes from single cells by polymerase cloning. *Nature biotechnology* **24**, 680 (Jun, 2006).
21. K. Lao, N. L. Xu, N. A. Straus, Whole genome amplification using single-primer PCR. *Biotechnology journal* **3**, 378 (Mar, 2008).
22. W. Dietmaier *et al.*, Multiple mutation analyses in single tumor cells with improved whole genome amplification. *The American journal of pathology* **154**, 83 (Jan, 1999).
23. X. Xu *et al.*, Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell* **148**, 886 (Mar 2, 2012).
24. R. Beroukhim *et al.*, The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899 (Feb 18, 2010).
25. P. J. Stephens *et al.*, Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27 (Jan 7, 2011).
26. P. J. Rochette, N. Bastien, J. Lavoie, S. L. Guerin, R. Drouin, SW480, a p53 double-mutant cell line retains proficiency for some p53 functions. *J Mol Biol* **352**, 44 (Sep 9, 2005).

27. D. MacArthur, Methods: Face up to false positives. *Nature* **487**, 427 (Jul 26, 2012).
28. P. C. Blainey, S. R. Quake, Digital MDA for enumeration of total nucleic acid contamination. *Nucleic acids research* **39**, e19 (Mar, 2011).
29. J. W. Drake, B. Charlesworth, D. Charlesworth, J. F. Crow, Rates of spontaneous mutation. *Genetics* **148**, 1667 (Apr, 1998).
30. J. C. Roach *et al.*, Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* **328**, 636 (Apr 30, 2010).
31. D. F. Conrad *et al.*, Variation in genome-wide mutation rates within and between human families. *Nat Genet* **43**, 712 (Jul, 2011).
32. D. L. Altshuler *et al.*, A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061 (Oct 28, 2010).

### **Acknowledgements**

This work was supported by the United States National Institutes of Health National Human Genome Research Institute Grant (HG005097-1 and HG005613-01) and in part by Bill & Melinda Gates Foundation OPP42867 to X SX. ARC was supported by the NIH Molecular Biophysics Training Grant (NIH/NIGMS T32 GM008313). We thank Paul Choi for his involvement on the early stage of the project and Jenny Lu and Lin Song for their assistance on the experiments. We thank Jun Yong for his help on single cell expansion and isolation and Zhang Yun at Biodynamic and Optical Imaging Center (BIOPIC) at Peking University for assistance on sequencing. The sequencing data is deposited at NCBI with accession number SRA060929.

### **Competing Financial Interests**

CZ, SL and X SX are authors on a patent applied for by Harvard University that covers the MALBAC technology

Figure 1

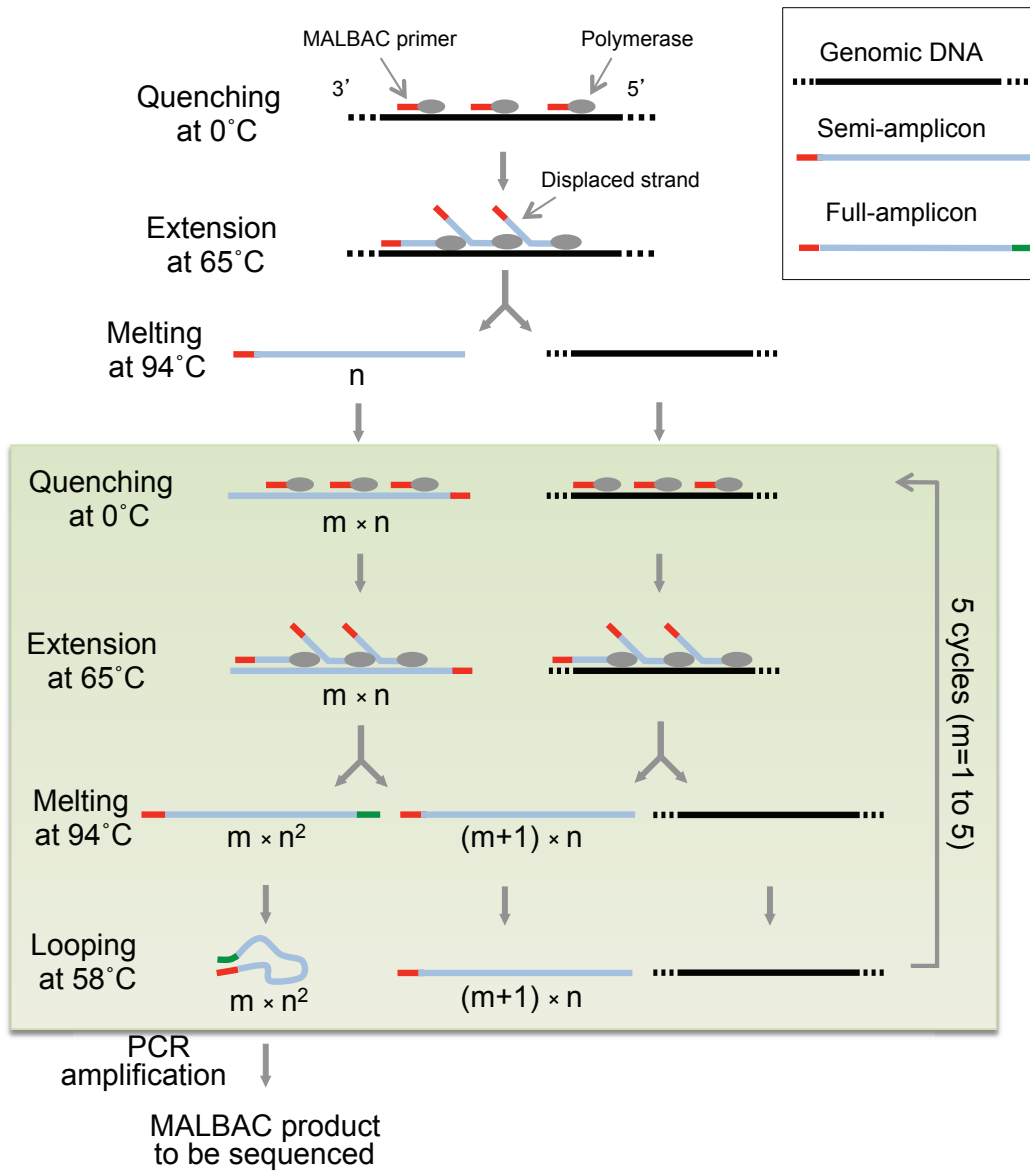


Figure 2

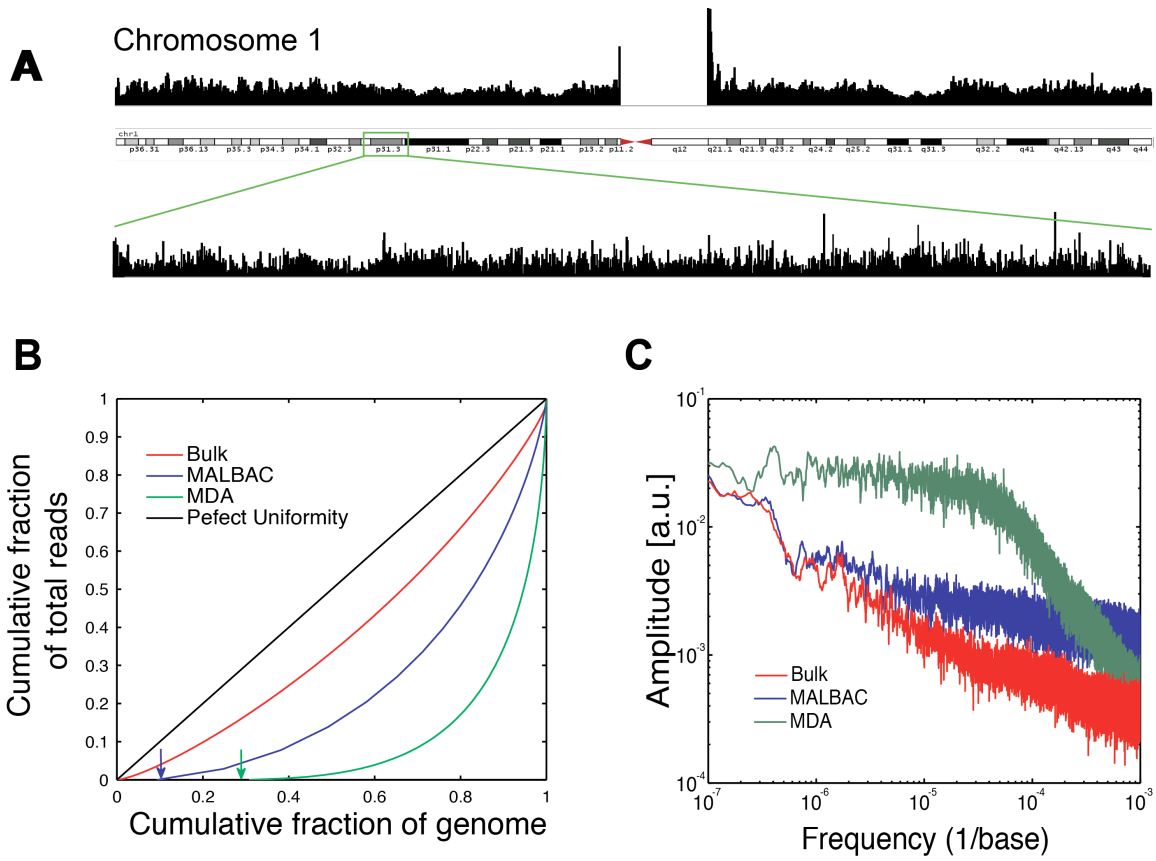


Figure 3

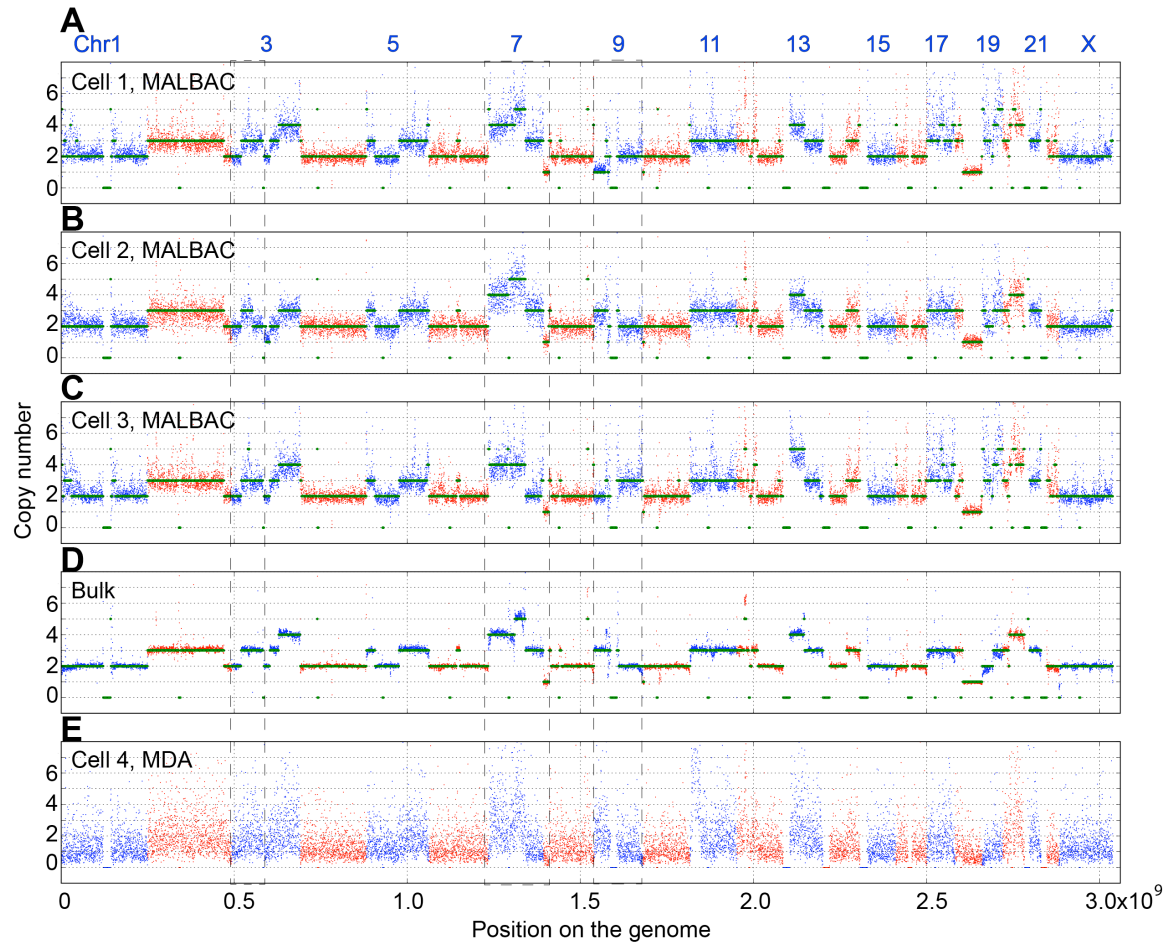




Figure 4

