SUPPORTING INFORMATION

Accurate and sensitive single-cell-level detection of copy number variations by micro-channel multiple displacement amplification (µcMDA)

Supporting Information:

Supplementary MethodsSupplementary Table S1, S2SupplementaryFigureS1-S6

Supplementary Methods

Cell culture

The K562 cell line was maintained suspending in 90% (v/v) RPMI 1640 medium (GIBCO, Life Technologies) supplemented with 10% (v/v) fetal bovine serum (Sigma Aldrich) and 100 μ g/ml Penicillin-Streptomycin (GIBCO, Life Technologies) in incubator at 37°C with 5% CO₂. The YH-1 cell line was incubated in a similar system including 80% (v/v) RPMI 1640 medium supplemented with 20% (v/v) fetal bovine serum, 100 μ g/ml Penicillin-Streptomycin, and additional 2mM/ml L-Glutamine (Sigma Aldrich) at 37°C with 5% CO₂. Before each culturing cycle or DNA extraction, cell suspension was centrifuged at 1200 rpm for 3 min.

Estimation of DNA content in K562 single cell

We estimated the DNA content of K562 cell using the sequencing data of bulk sample. First, we used the sorted bam file to achieve the mapping ratio of reads Mr, and the overall coverage Co. Then we respectively calculated the total coverage of each ploidy to obtain an average ploidy:

$$P_{ave} = \frac{\sum_{i=1}^{n} iP(i)}{\sum_{i=1}^{n} P(i)}$$
(1)

, where *i* stands for a certain ploidy and P(i) stands for the total read coverage belongs to this ploidy. As illustrated in the coverage breadth figure (Figure 4C), the coverage gain sharply decreased when using further enlarged sequencing data. As a result, *Co* is approximate to Co_{max} when applying a data large enough. Suppose that the unmapped reads would constitute the rest genome with the same ploidy-assignment like those mapped read, the DNA content of K562,

$$Dc_{K562} = \frac{Dc_{hg-19} \cdot Co \cdot P_{ave}}{Mr}$$
(2)

Finally, we substitute the values into the formula with $Dc_{hg-19} = 3.3$ pg, Co = 91.33%, $P_{ave} = 2.66$ and Mr = 97.35% and achieved an estimated DNA content of 8.24 pg for a single K562 cell.

CNV recognition

The CNV detection was based on the protocol provided by Baslan et al (1,2). Firstly, variable bins were generated by aligning simulated reads constructed from reference back to the reference. Simulated reads were yielded by fetching all 150-bp regions with 1-bp dislocation from reference genome. Only those reads uniquely mapped to its generated place were reserved for the subsequent good-zone calculation. Then good zones were successively gathered with the total length reaching the setup value of bin size. After variable bins were constructed, uniquely mapped reads of each sample were assigned in to these bins. Then the read counts of each bin were preserved as an original input for the follow-up CNV calling steps. Some bad bins were removed

according to the acquiescent parameter of the Ginkgo software. GC correction was then applied by regressing the normalized bin counts using their LOWESS fit curve. Afterwards, all bins were segmented by their respective corrected count value using circular binary segmentation (CBS) algorithm. Finally, the copy number states of the segments were determined according to a presupposed ploidy, which was predicted with our bulk sample.

For the YH-1 experiments, we only used 4 million reads, identical to $0.2 \times$ genome, for CNV analysis because at this down sampling rate, correlation coefficients between down sampled bulk data were around 0.01. Under this condition, the internal non-uniform bias of data as well as its influence on down sampling process could be simultaneously exhibited (Supplementary Figure S5). As for the K562 CNV detection, more abundant reads of 60 M (3 ×) were utilized for more accurate comparison, also due to the significantly lower DNA input.

GC-bias evaluation

In the YH-1 experiments, we measured the coverage depths of each 100-kb fixed bin in chromosome 1, and normalized them with their median value (3). It is noticed that coverage depths were generated only by those uniquely mapped reads with acceptable read and mapping qualities. Because of the relatively high coverage, uncovered regions in all bins remained untreated. Differently, N regions were removed before the depth calculation.

With respect to K562 GC-content analysis, we calculated the copy-number-state normalized bin counts in 100-kb variable bins instead of calculating the depths in fixed-sized bins. Accordingly, we filtered out reads with bad read quality or mapping quality and N regions and repeats were directly excluded in the calculation due to the benefit of variable binning algorithm.

For each sample, a smoothed regression curve using the loess function with the *span* parameter identical to 0.16 was fitted to the scatters representing GC-related depths in each bin.

Uniformity characterization in K562 experiments

Uniformity for amplifying aneuploid K562 genomic DNA was demonstrated in the following four aspects: (1) Lorenz curves showed the cumulative fraction of total reads versus the cumulative fraction of covered genomic region. In this analysis, only those uniquely mapped reads were involved in generating the normalized bin counts, then the Lorenz curves. Hence these curves reflect more on the covering uniformity by the successfully aligned reads at the certain variable-bin size of 50 kb. (2) The coefficients of variation plot, in which the CV values were determined by the copy-number-state normalized depths within the N-region removed bins of different sizes across the chromosome 1. Hence, the CV plot comprehensively presented the uniformity at all scales in spite of the copy number states. (3) In the coverage breadth plot, coverages were calculated directly using the duplication-reserved reads, as this plot mainly exhibited the efficiency of the transformation from sequencing efforts provided by the amplification products to sequencing coverage. (4) Autocorrelations (4) were calculated with 160/80/40/20/10/5 -Mb respectively down sampled data. Still, only unique reads were expected to be adopted here. Auto-correlations of base-level coverage in chromosome 1 at various depths were

then calculated to verify the coverage correlations at full length scales. Then, the exponential function:

$$G(\Delta) = a + be^{-\Delta/l_c}$$
(3)

was fitted to the range from 0.5 kb to 17 kb, where Δ stands for the distance between a pair of bases, $G(\Delta)$ represents the Auto-correlation value and the characteristic lengths (l_c) suggests the amplicon-originated bias of non-uniformity.

- Baslan, T., Kendall, J., Rodgers, L., Cox, H., Riggs, M., Stepansky, A., Troge, J., Ravi, K., Esposito, D., Lakshmi, B. *et al.* (2012) Genome-wide copy number analysis of single cells. *Nat. Protoc.*, 7, 1024-1041.
- Garvin, T., Aboukhalil, R., Kendall, J., Baslan, T., Atwal, G.S., Hicks, J., Wigler, M. and Schatz, M.C. (2015) Interactive analysis and assessment of single-cell copy-number variations. *Nat. Methods*, 12, 1058-1060.
- Evrony, G.D., Lee, E., Mehta, B.K., Benjamini, Y., Johnson, R.M., Cai, X., Yang, L., Haseley, P., Lehmann, H.S. and Park, P.J. (2015) Cell lineage analysis in human brain using endogenous retroelements. *Neuron*, 85, 49-59.
- Zhang, C.-Z., Adalsteinsson, V.A., Francis, J., Cornils, H., Jung, J., Maire, C., Ligon, K.L., Meyerson, M. and Love, J.C. (2015) Calibrating genomic and allelic coverage bias in single-cell sequencing. *Nat. Commun.*, 6.

Supplementary Tables

| Sample index | Number of mapped bases | Read mapping ratio (%) | GC content (%) | Mean depth (×) | Genome coverage (%) | CV (10K binsize) |
|-----------------|------------------------------|------------------------------|----------------------|-------------------|---------------------------|---------------------|
| μcMDA-1 | 74,118,624 | 85.70 | 40.51 | 3.3911 | 80.02 | 1.129835536 |
| µcMDA-2 | 74,242,868 | 85.86 | 40.49 | 3.3936 | 79.90 | 1.131428118 |
| μcMDA-3 | 74,172,875 | 85.74 | 40.47 | 3.4011 | 80.26 | 1.121643423 |

Supplementary Table S1 Sequencing statistics of low-coverage whole genome sequencing (LWGS) data from parallel YH-1 μ cMDA assays

Supplementary Table S2 Sequencing statistics of YH-1 and K562 samples amplified through different protocols.

| Sample index | Number of mapped reads | Read mapping ratio (%) | GC content (%) | Mean depth (×) | Genome coverage (%) |
|--------------|------------------------|------------------------------|----------------|-------------------|---------------------------|
| YH-bulk | 724,694,232 | 98.92 | 40.76 | 34.32 | 92.37 |
| YH-MDA | 754,857,816 | 89.80 | 38.86 | 35.25 | 91.15 |
| YH-μcMDA | 757,994,918 | 85.70 | 40.51 | 34.71 | 92.18 |
| K562-bulk1 | 58,368,962 | 97.28 | 40.56 | 2.77 | 71.53 |
| K562-bulk2 | 58,411,610 | 97.35 | 40.86 | 2.78 | 81.40 |
| K562-MDA1 | 59,493,568 | 99.16 | 39.6 | 2.82 | 47.91 |
| K562-MDA2 | 59,581,448 | 99.30 | 40.61 | 2.80 | 42.32 |
| K562-µcMDA1 | 59,139,015 | 98.57 | 40.84 | 2.79 | 53.22 |
| K562-µсMDA2 | 59,330,468 | 98.88 | 41.15 | 2.78 | 48.57 |

Supplementary Figures



4M-downsampled YH-1 Data

Supplementary Figure S1: Cluster dendrogram of the down sampled data in the YH-1 experiment. We kept reducing the sampling ratio until the correlations between each pair of CNV results were around 0.01 in bulk data. Then the certain value of 4 million was determined as the analyzing amount for µcMDA and MDA reads.





Supplementary Figure S2: CNV detection across the K562 genome by grouped contrast assays of μ cMDA and MDA. In this analysis, reads were combined into 50-kb variable bins. The black lines exhibited the estimated copy number states. Scatters representing normalized bin counts gather more in μ cMDA and the CNV patterns were more similar to those of bulk.



Supplementary Figure S3: Correlation analysis of the CNV results of parallel bulk samples. Due to the insufficient sequencing depth, there existed random bias in between the independent bulk data sets. Scatters off the diagonal dash represented those regions with inconsistent copy number states in different bulk samples.



Supplementary Figure S4: two detailed micro-CNVs. (**A**) The region from 17 to 50 Mb on chromosome 22. A brown arrow was pointed at a micro-CNV of ~250 kb. (**B**) a region on chromosome 5 with a brown arrow indicating a micro-CNV of ~300 kb while MDA failed to detect out this CNV.



Supplementary Figure S5: CNV detection across the entire YH-1 genome by bulk, μ cMDA and MDA. Reads were combined into 500-kb variable bins. The black lines exhibited the estimated copy number states. Some small CNVs existed in the YH-1 genome according to the bulk result where μ cMDA data gave much more reasonable results in contrast to the MDA data.



Supplementary Figure S6: Variation of bin size. (A) Variation of 50-kb bin size. (B) Variation of 1-Mb bin size. The pink box indicated values ranging from 25% to 75%. The red line in the box indicated the average value. The whiskers outside the box represented 5% and 95% quantiles respectively.