List of Supplementary Figures

False mutation detection caused by WGA artifacts

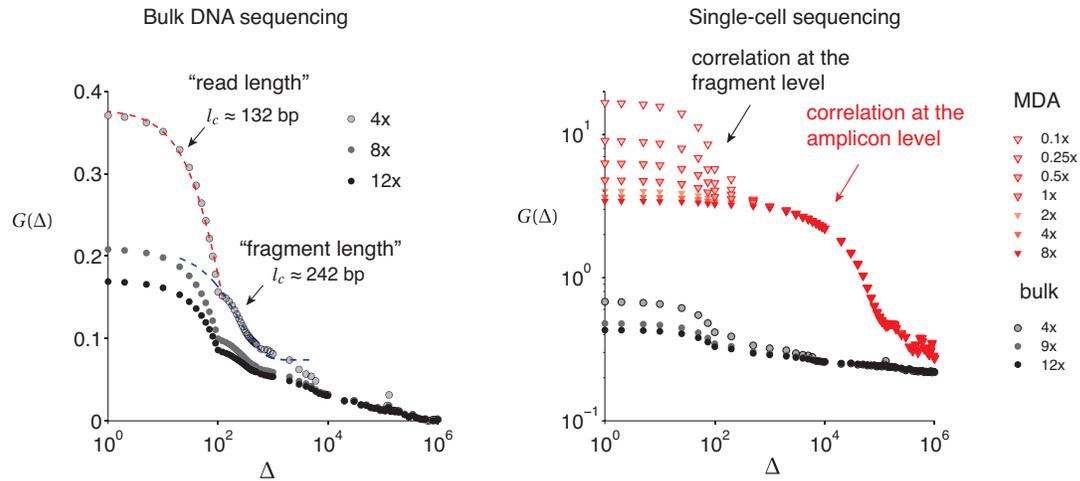**Supplementary Figure 1 | False mutation detection due to whole-genome amplification artifacts.**
From left to right: **false negative SNV detection** due to allelic distortion; **false positive translocation detection** due to artificial chimera; **false positive loss-of-heterozygosity** due to chromosomal segments dropping out of amplification; **false positive SNV detection** due to early amplification errors. There are two major differences between allelic distortion and chromosome segments dropping out of amplification. First, allelic distortion is primarily caused by preferential amplification of one chromosome allele; from the auto-correlation analysis, such preferential amplification extends roughly to the amplicon size (10 ~ 100 kb). But "false LOH" here refers to much longer chromosomal segments (> 1 Mb) that drop out of amplification. Second, it is possible to perform deep targeted sequencing by PCR amplification from the MDA product to uncover both chromosome alleles even under allelic distortion. It is impossible to reveal the chromosome segment that has dropped out of amplification by targeted sequencing of the MDA product; a biological replicate is needed to validate if such dropout is an artifact or true deletion.

1

a Coverage correlations at different scales

Bulk DNA sequencing

"read length"
$l_c \approx 132$ bp

"fragment length"
$l_c \approx 242$ bp

○ 4x
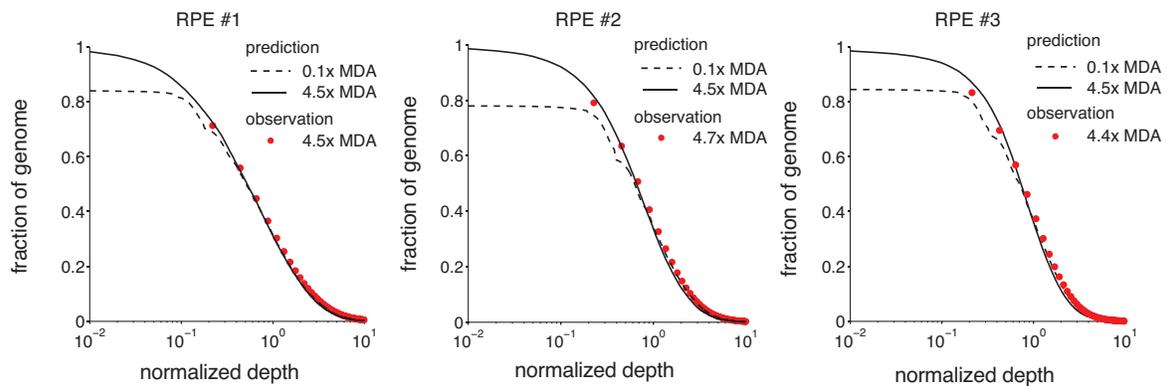● 8x
● 12x

$G(\Delta)$

Single-cell sequencing

correlation at the fragment level

correlation at the amplicon level

$G(\Delta)$

MDA
▽ 0.1x
▽ 0.25x
▽ 0.5x
▽ 1x
▽ 2x
▼ 4x
▼ 8x

bulk
● 4x
● 9x
● 12x

b Amplicon-level correlation fitting in RPE libraries by MDA

$G(\Delta)$

			l_c from fitting	correlation magnitude at 1 kb
RPE #1	▽	0.1x Miseq	32.5 kb	2.45
	▼	4.5x Hiseq	34.2 kb	2.35
RPE #2	▷	0.1x	18.6 kb	1.51
	▶	4.4x	16.8 kb	1.47
RPE #3	◁	0.1x	12.3 kb	1.01
	◀	4.7x	13.1 kb	1.00

c Coverage predictions for single RPE MDA libraries

RPE #1

fraction of genome

prediction
--- 0.1x MDA
— 4.5x MDA
observation
● 4.5x MDA

normalized depth

RPE #2

fraction of genome

prediction
--- 0.1x MDA
— 4.5x MDA
observation
● 4.7x MDA

normalized depth

RPE #3

fraction of genome

prediction
--- 0.1x MDA
— 4.5x MDA
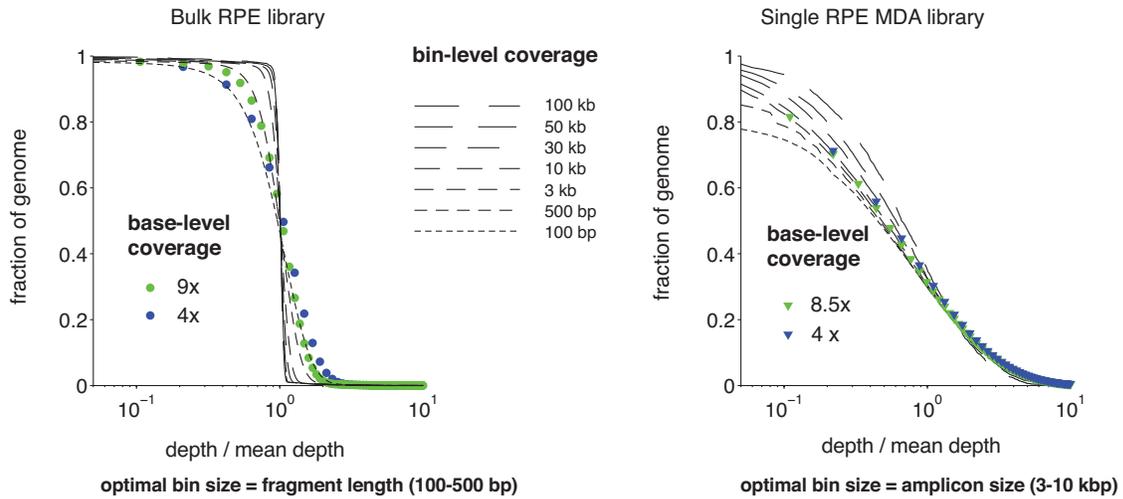observation
● 4.4x MDA

normalized depth
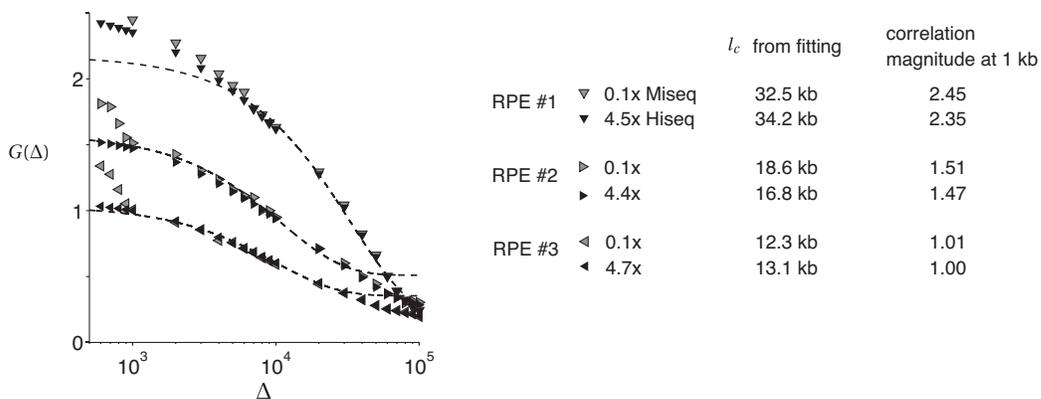
Supplementary Figure 2 | Analysis of amplification bias and genome coverage in MDA-generated libraries of RPE-1 cells.
(**a**) Correlation at the scale of sequencing read length and fragment length is present in both bulk and single-cell libraries (200-300 bp), but correlation at the scale of amplicons (∼ 10 kb) only appears in MDA-derived libraries. (**b**) Amplicon-level correlation independent of the sequencing depth is demonstrated in disomic Chr.1 in three MDA-generated libraries of two-cell RPE samples. (**c**) Independent sequencing of three MDA-generated RPE libraries to 0.1x and to ∼ 4.5x confirms that the bin-level depth-of-coverage predicted from low-pass sequencing agrees with the observed single-base depth-of-coverage in higher-depth sequencing, reflecting the dominant non-uniformity generated at the amplicon level. All correlation functions are calculated for disomic Chr.1
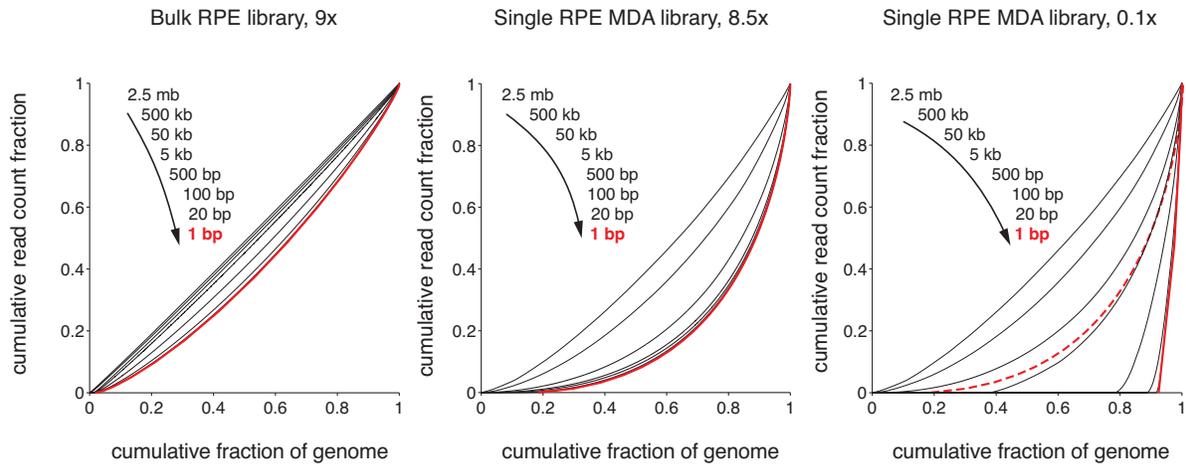
a Depth-of-coverage yield curve for bulk and MDA libraries

Bulk RPE library
- 12x
- 9x
- 4x

Single RPE MDA library
- 8.5x
- 4 x
- 1 x

Normalized DoC yield
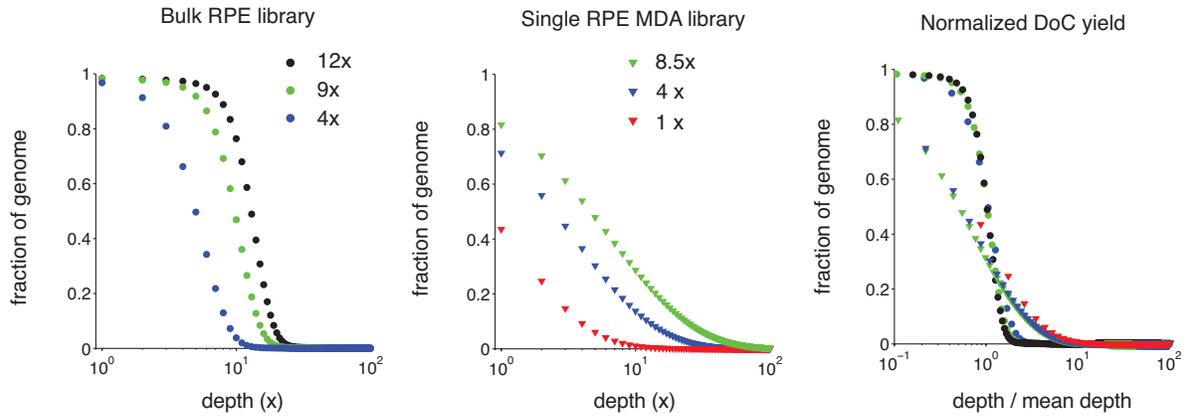
b Normalized depth-of-coverage yield curve at different bin sizes

Bulk RPE library

bin-level coverage
- 100 kb
- 50 kb
- 30 kb
- 10 kb
- 3 kb
- 500 bp
- 100 bp

base-level coverage
- 9x
- 4x

optimal bin size = fragment length (100-500 bp)

Single RPE MDA library

base-level coverage
- 8.5x
- 4 x

optimal bin size = amplicon size (3-10 kbp)

c Lorenz distribution curves for bulk and MDA libraries evaluated at different bin sizes

Bulk RPE library, 9x
- 2.5 mb
- 500 kb
- 50 kb
- 5 kb
- 500 bp
- 100 bp
- 20 bp
- 1 bp

Single RPE MDA library, 8.5x
- 2.5 mb
- 500 kb
- 50 kb
- 5 kb
- 500 bp
- 100 bp
- 20 bp
- 1 bp

Single RPE MDA library, 0.1x
- 2.5 mb
- 500 kb
- 50 kb
- 5 kb
- 500 bp
- 100 bp
- 20 bp
- 1 bp

3

Supplementary Figure 3 | Coverage non-uniformity in bulk and MDA-derived libraries of RPE-1 cells evaluated at different bin sizes.

(**a**) The depth-of-coverage (DoC) curves at different sequencing depths reflect the same underlying density distribution of the coverage at different loci in the sequencing library when the coverage depth is normalized by the mean sequencing depth. (**b**) The depth-of-coverage yield evaluated at the single-base level is almost identical to the depth-of-coverage evaluated at the level where the dominant variation is generated. For bulk DNA libraries, the dominant variation occurs at the level of sequencing fragments (presumably in the PCR step of library preparation). Therefore, the DoC curve evaluated at bin size ~200-300 bp is indistinguishable from the DoC curve evaluated at the single-base level, but at larger bin sizes, the coverage bias is further attenuated. For MDA-generated libraries, the DoC curve is almost invariant until the bin size reaches the level of single amplicons ~10 kb. Therefore, the DoC curve at the single-base level can be reliably estimated from the bin-level (~10 kb) coverage in low-pass sequencing data (~0.1x). (**c**) Lorenz curves evaluated at different bin sizes also confirm the dominant variation at ~100 bp scale for the bulk DNA library and at ~5 kb scale for the MDA library. In the bulk DNA library (left panel), the Lorenz curve for the bin-level coverage (black curves) is similar to the Lorenz curve at the single base level (red curve) but deviates towards more uniformity as the bin size exceeds 100 bp. In the MDA library sequenced to 8.5x (middle panel), the Lorenz curve is essentially unchanged until the bin size reaches 5 kb and deviates towards more uniformity as the bin size exceeds 50 kb, consistent with an amplicon size ~10 kb. In the same MDA library sequenced to 0.1x (right panel), the majority of the chromosome is uncovered. In the uncovered region, amplification non-uniformity cannot be evaluated from the single-base level coverage (red curve); however, the bin-level coverage (5 kb bins) agrees well with the single-base level coverage observed in 8.5 x sequencing (dashed red curve), suggesting that this distribution is intrinsic to the MDA but not dependent on the sequencing depth. All Lorenz curves are evaluated for disomic Chr.1

**a** Correlation analysis of single glioblastoma MDA libraries
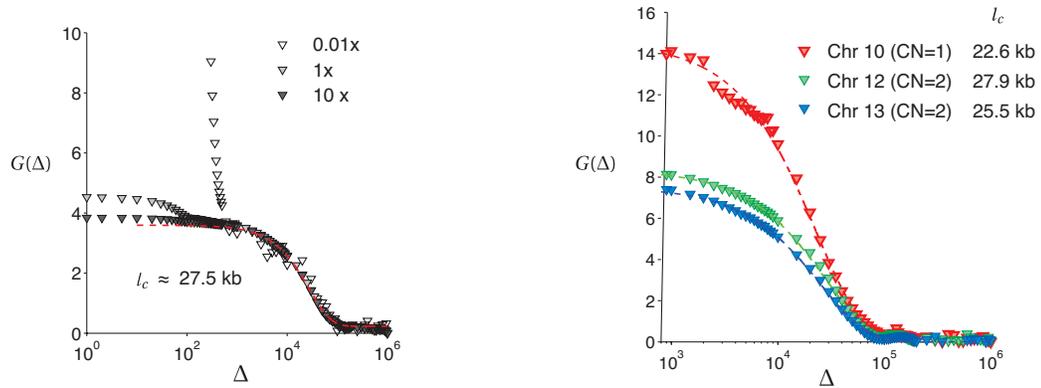
**b** Coverage predictions for single glioblastoma MDA libraries
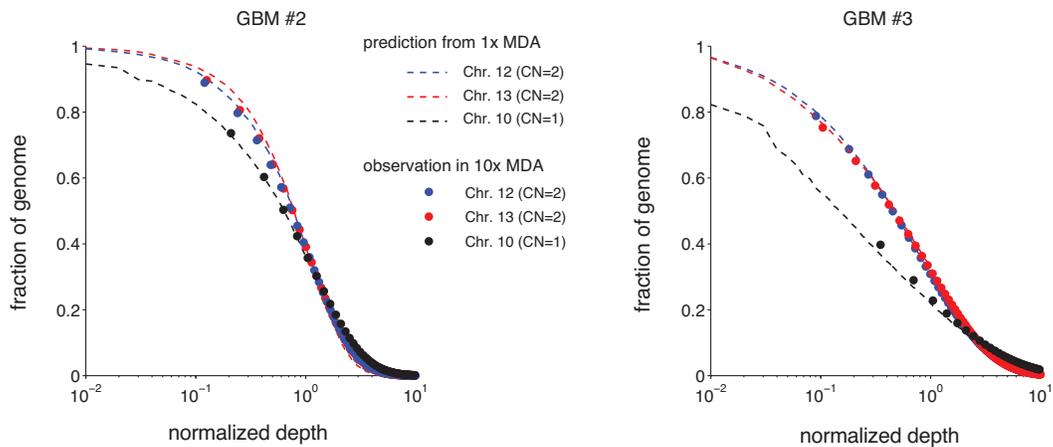
Supplementary Figure 4 | Analysis of amplification bias and genome coverage in MDA-generated libraries of frozen glioblastoma nuclei (BT325) in Francis *et al.* (2014).
(**a**) (Left) The same amplicon-level correlation (red dashed curve, $l_c = 27.5$ kb) is revealed from independent sequencing of a single glioblastoma library (GBM #4) to 0.01x, 1x, and 10x. (Right) Chromosomes with different copy number show the same amplicon size but different magnitude of amplification bias, reflecting differential priming efficiencies (from one copy of monosomic chromosome vs. two copies of disomic chromosomes) during the early stage of amplification. (**b**) The cumulative coverage curve at higher sequencing depths can be accurately predicted from the bin-level coverage at lower depths for chromosomes with different copy number.
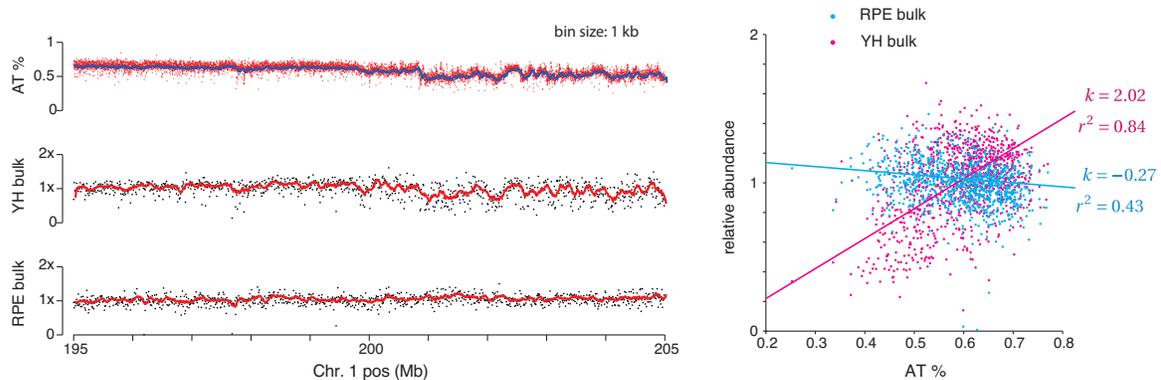
a Auto-correlation analysis of single and bulk DNA libraries in Hou et al. (2012)
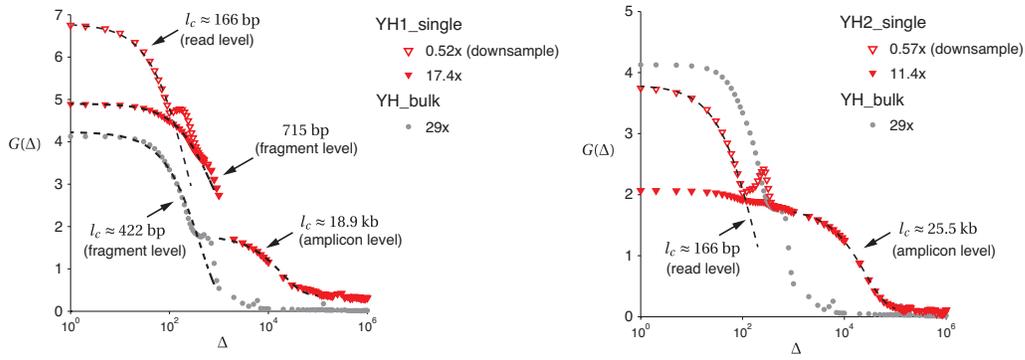
b Prediction of allelic coverages of single-cell DNA libraries in Hou et al. (2012)
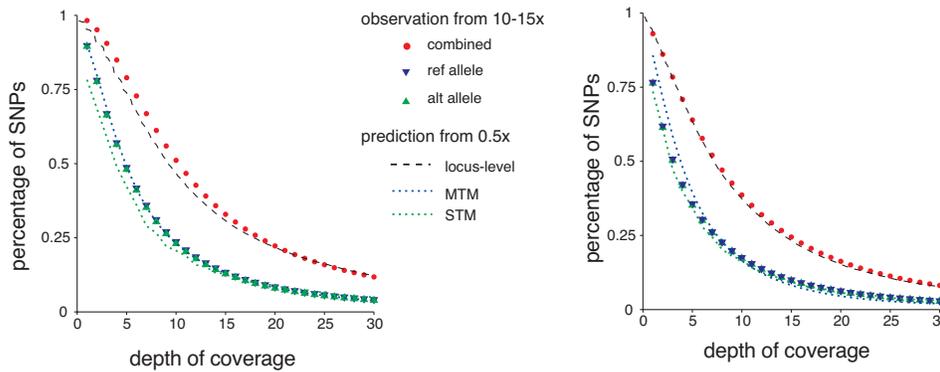

c YH bulk DNA library showed considerable GC bias at the fragment level

Supplementary Figure 5 | Analysis of amplification bias and genome/allele coverage in bulk and MDA-derived libraries of diploid lymphoblastoid YH cells.
(**a**) Comparison of coverage correlation in single-cell and bulk libraries confirms the dominant bias occurring at the amplicon level ~20 kb. There is considerable fragment-level correlation in both the bulk and one of the two single-cell libraries that is likely introduced during the PCR step of library preparation. (**b**) The coverage at heterozygous sites (reference, alternate, or both alleles combined) evaluated at high sequencing depths agrees well with predictions from low-pass sequencing. (**c**) The fragment-level bias in the bulk DNA library shows a strong correlation with the local sequence content (AT%) that is absent in the bulk library of RPE-1 cells shown in **Supplementary Figure 2**.

a Auto-correlation analysis of single and bulk MDA libraries in Evrony et al. (2012)

(i) MDA of single neurons

(ii) MDA of 100 neurons

(iii) MDA of 50,000 neurons
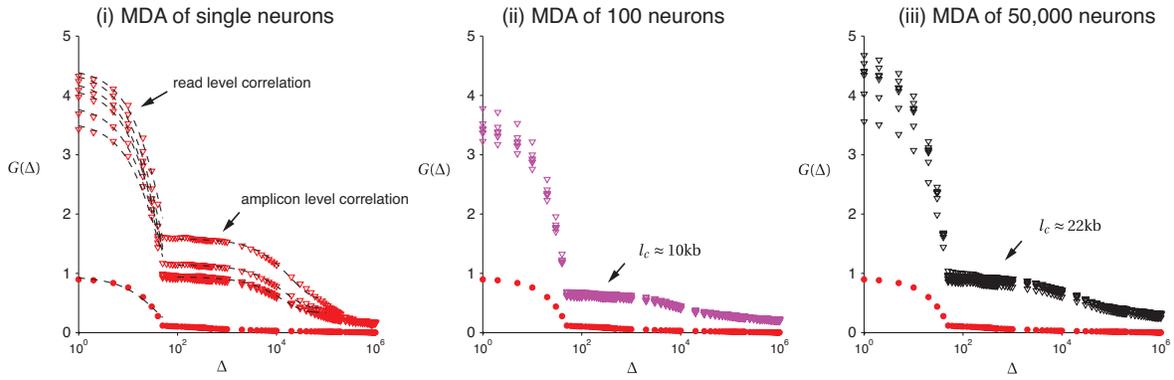
Sample	read level correlation	amplicon correlation
1465-neuron-1	50 bp	16.5 kb
1465-neuron-2	49 bp	14.8 kb
1465-neuron-3	48 bp	15.3 kb
1465-neuron-4	60 bp	23.3 kb
1465-neuron-5	46 bp	15.6 kb
1465-neuron-6	46 bp	14.3 kb
1465-bulkDNA	38 bp	NA

b Correlation between GC content and library representation at the amplicon level (bin size: 20 kb)

single neuron
(1465-cortex-MDA-1)

100 neurons
(1465-cortex-100-MDA-1)

50,000 neurons
(1465-cortex-50k-MDA-1)

bulk
(1465-cortex-bulkDNA)

7
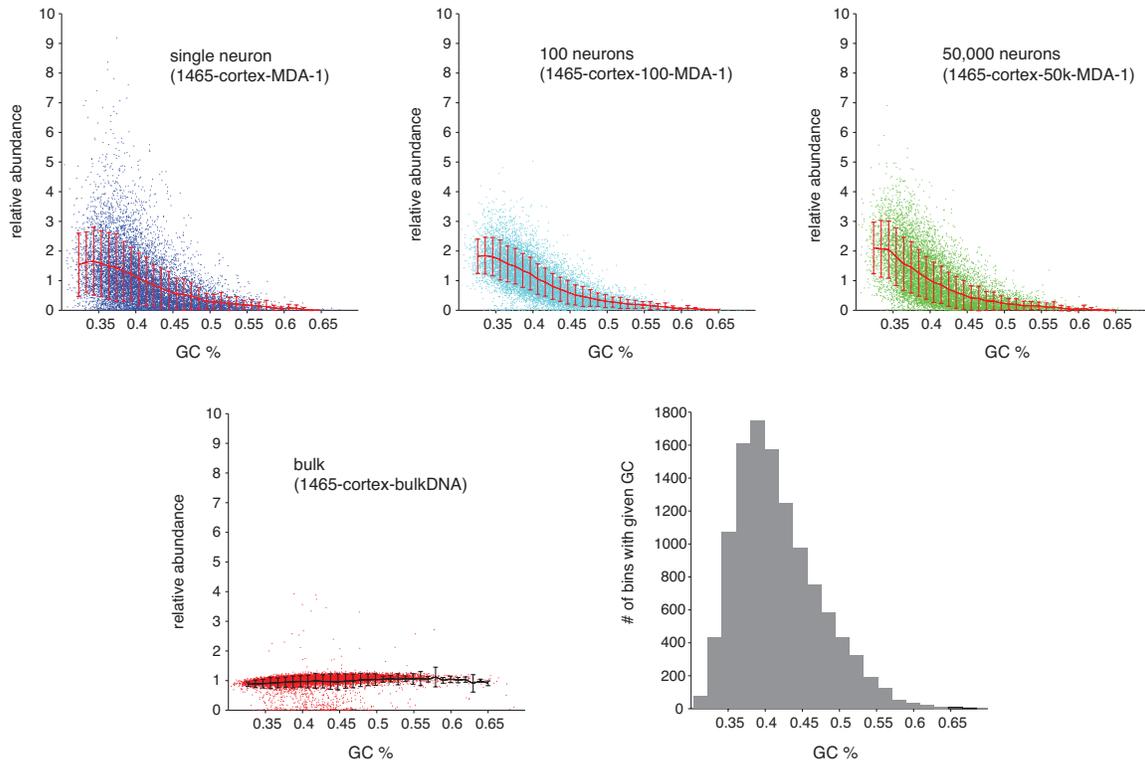
Supplementary Figure 6 | Analysis of coverage correlation in MDA generated libraries of frozen neuron nuclei in Evrony *et al.* (2012).
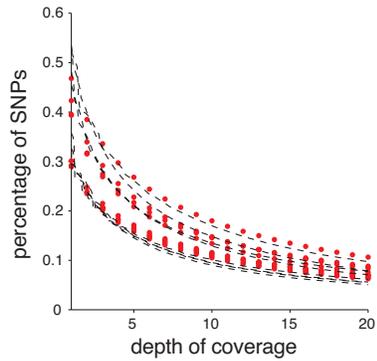(**a**) MDA-generated libraries of single neurons, hundreds of neurons, or thousands of neurons show the same amplicon-level correlation reflecting amplicon size ~20 kb. The short-range correlation reflects the read length (50 bp) at low sequencing depths. (**b**) Comparison of the relative abundance of sequence coverage with local GC content at the amplicon level (evaluated at 20 kb bins) suggests that there is a wide range of coverage variation in regions with similar GC content. By contrast, there is little variation at 20 kb scale in the bulk library. Error bars show 3×standard deviation for each GC stratum. The variation is especially evident for single-neuron libraries, where the range is proportional to the total number of bins with similar GC content, reflecting random, instead of sequence-content generated bias.

a **Auto-correlation analysis of single sperm DNA libraries in Wang et al. (2012)**
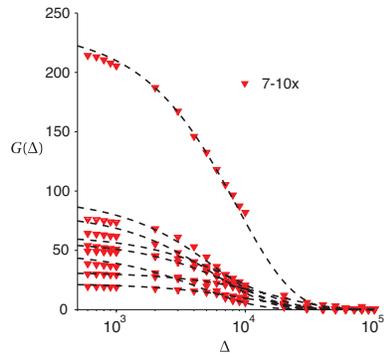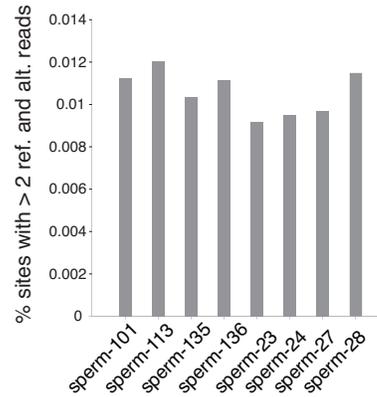

Sample	correlation length	correlation magnitude
sperm-101	16.2 kb	29.2
sperm-113	10.3 kb	48.2
sperm-135	12.8 kb	18.9
sperm-136	5.4 kb	36.1
sperm-23	8.1 kb	50.9
sperm-24	9.1 kb	205.4
sperm-27	6.1 kb	73.6
sperm-28	5.9 kb	61.6

b **Coverage predictions**

c **Error rate estimates**

Supplementary Figure 7 | Analysis of coverage variation in MDA generated libraries of single sperms in Wang *et al.* (2012). (**a**) Different single-sperm libraries exhibit dominant bias at 10-20 kb range consistent with the MDA amplicon size. (**b**) Cumulative coverage at higher sequencing depth (7-10x) can be accurately predicted from the amplicon-level coverage evaluated from low-pass sequencing (<1x). (**c**) About 1% sites showed heterozygosity (with > 2 reads of both reference and alternate alleles) in the haploid sperm libraries, indicating an error frequency of ~1% at allele frequency ~10% or above.

a Coverage correlations in MALBAC, MDA, and bulk SW480 libraries (Zong et al. 2012)

$G(\Delta)$

$l_c \approx 9$ kb

	~20x	~1x
SW480-5 (SRS375673)	◆	◇
SW480-4 (SRS375672)	▼	▽
SW480-3 (SRS375671)	▲	△
SW480-2 (SRS374233)	◀	◁
SW480-1 (SRS373654)	▶	▷
MDA (SRS375060)	■	□
bulk DNA (SRS375016)	●	○

Δ
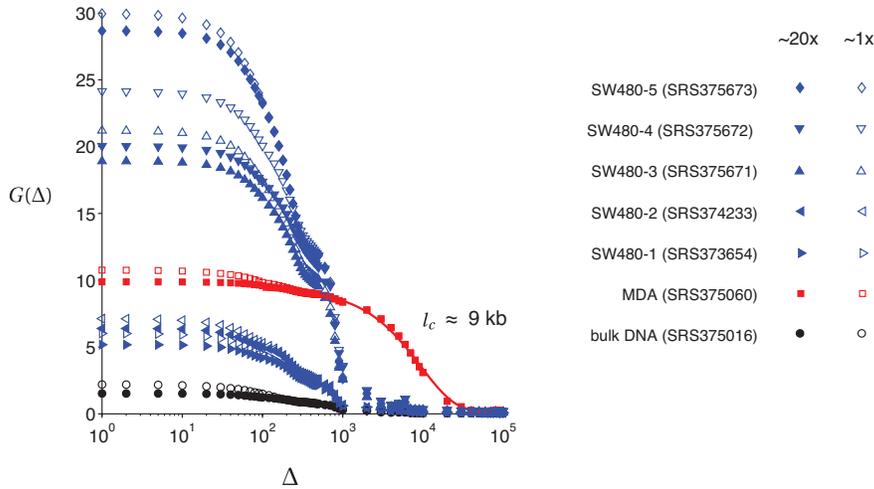
b Snapshots of sequencing coverage in MDA and MALBAC libraries

25 kb

MDA
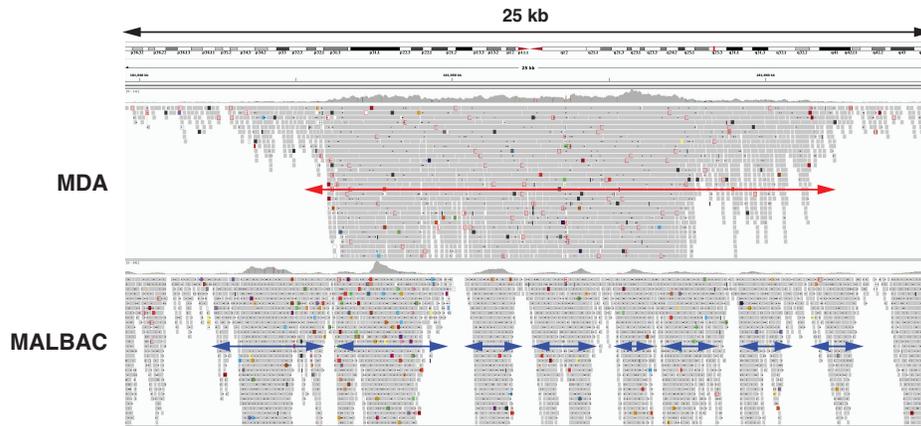
MALBAC

c Coverage predictions for single SW-480 libraries

prediction
- - - from 1x SW480-MDA
(bin size: 4.5 kb)

observation
● 25x SW480-MDA

fraction of genome

depth of coverage

prediction
- - - from 1x SW480-1 (MALBAC)
(bin size: 500 bp)

observation
● 25x SW480-1 MALBAC
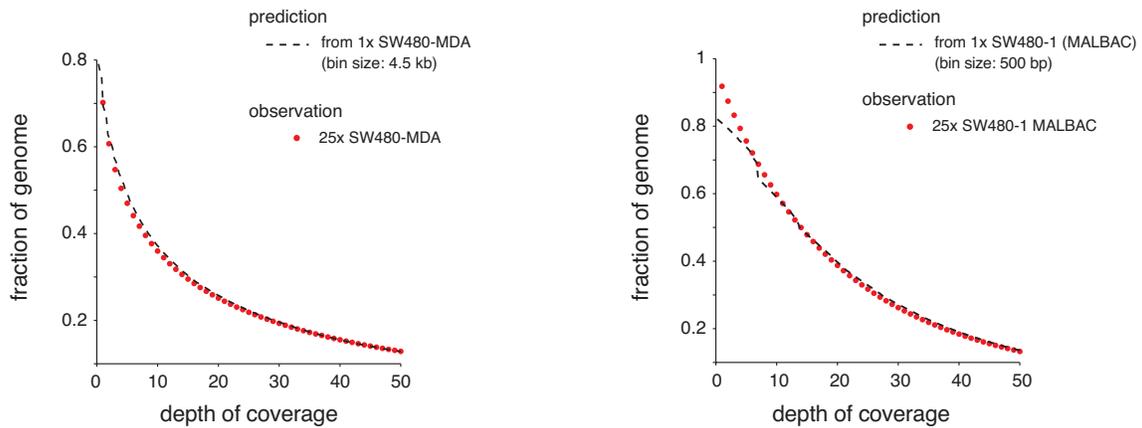
fraction of genome

depth of coverage

Supplementary Figure 8 | Analysis of coverage variation in MDA and MALBAC generated libraries of SW480 tumor cells in Zong *et al.* (2013).
(**a**) All five MALBAC-generated single SW480 libraries (blue symbols) show a much shorter amplicon compared to the MDA-generated library (red squares). (**b**) Snapshots of local sequencing coverage confirm the shorter amplicons in MALBAC in comparison to MDA. (**c**) In both MDA and MALBAC-generated libraries, the dominant source of non-uniformity exists at the amplicon-level and can be accurately predicted from low-pass sequencing.

10

a Coverage correlations in DOP-PCR amplified single-cell libraries (Wang et al. 2014)
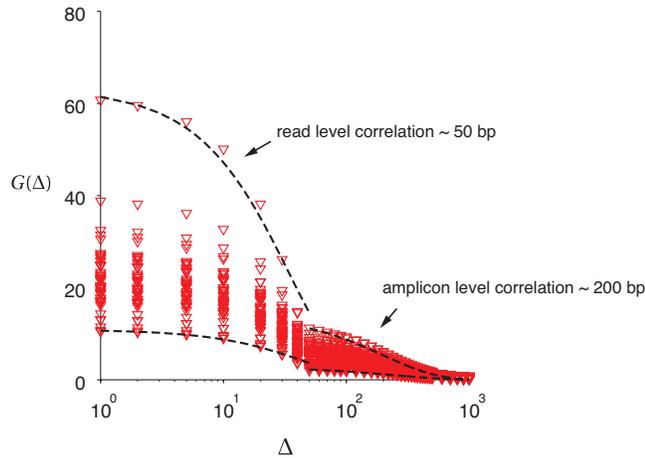
b Correlation magnitude and spread (amplicon size) in DOP-PCR libraries (Wang et al. 2014)
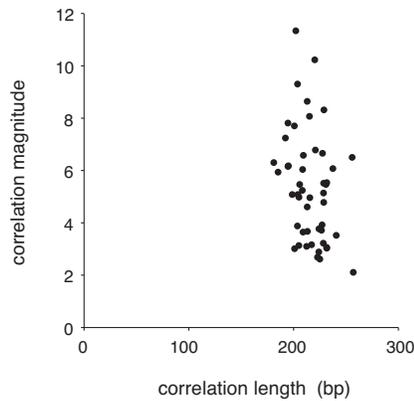
c Correlation magnitude and spread (amplicon size) in MDA and MALBAC libraries
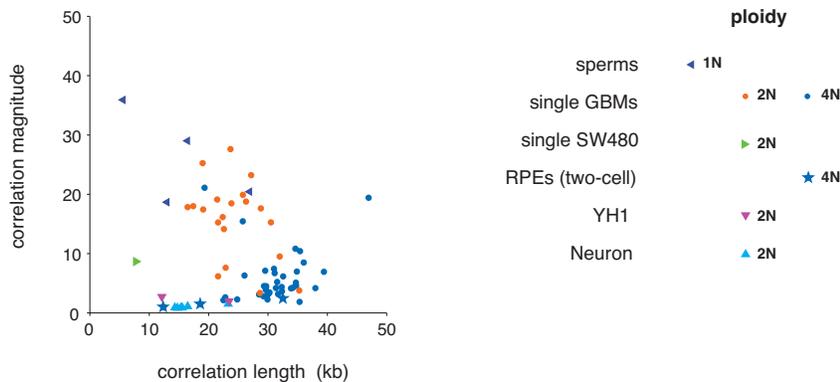

Supplementary Figure 9 | Analysis of coverage variation in DOP-PCR libraries of tumor nuclei in Wang *et al.* (2014)
(**a**) Different libraries (49 total) exhibit similar fragment-level correlations (l_c = 50 bp) and amplicon-level bias (l_c = 200 bp). (**b**) Magnitude and spread of DOP-PCR bias. The amplicon size (estimated by the correlation length) agrees well with the size selection protocol in the original report (Navin *et al.*, 2011). The magnitudes of correlation in DOP-PCR libraries are generally smaller than those observed in MDA-generated libraries as shown in (**c**), reflecting less overall bias in DOP-PCR. (**c**) Magnitude and spread of MDA bias in different studies. More DNA content in 4N samples results in better uniformity than 2N or 1N samples. This is consistent with improved uniformity of MDA from G2 nuclei in Wang *et al.* (2014).

Mathematical models of allelic amplifications

combined coverage from both alleles

locus-level read counts and coverage

sequencing coverage

($\geq$ 1 read): f_1
($\geq$ 2): f_2
($\geq$ 3): f_3
($\geq$ 4): f_4

mixed template model (MTM)
symmetric binomial distribution at each locus

segregated template model (STM)
independent distribution at each locus

allele read counts

allele A

allele B

alellic coverage

allele A ($\geq$ 1 read): λ
allele B ($\geq$ 1 read): λ

$$\langle \lambda \rangle = \sum_{t \geq 1} (1/2)^t \langle f_t \rangle$$

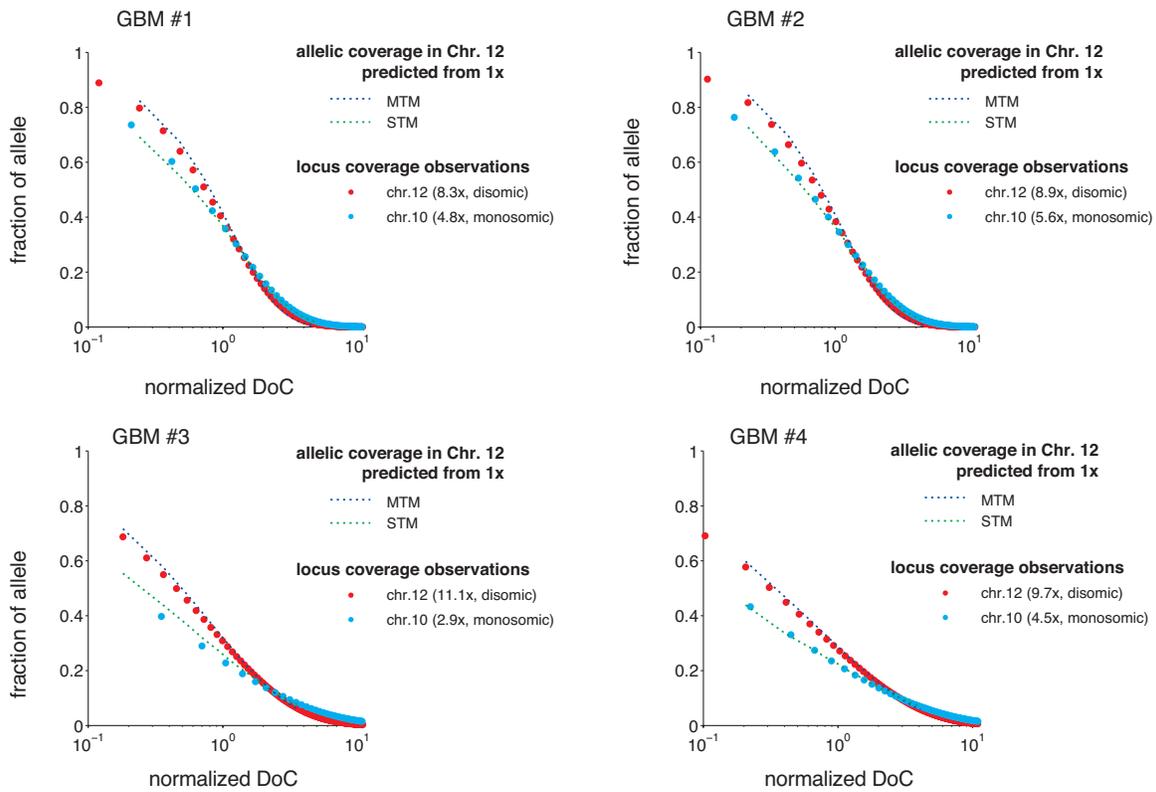$$\langle \lambda \rangle = 1 - (1 - \langle f_1 \rangle)^{1/2}$$
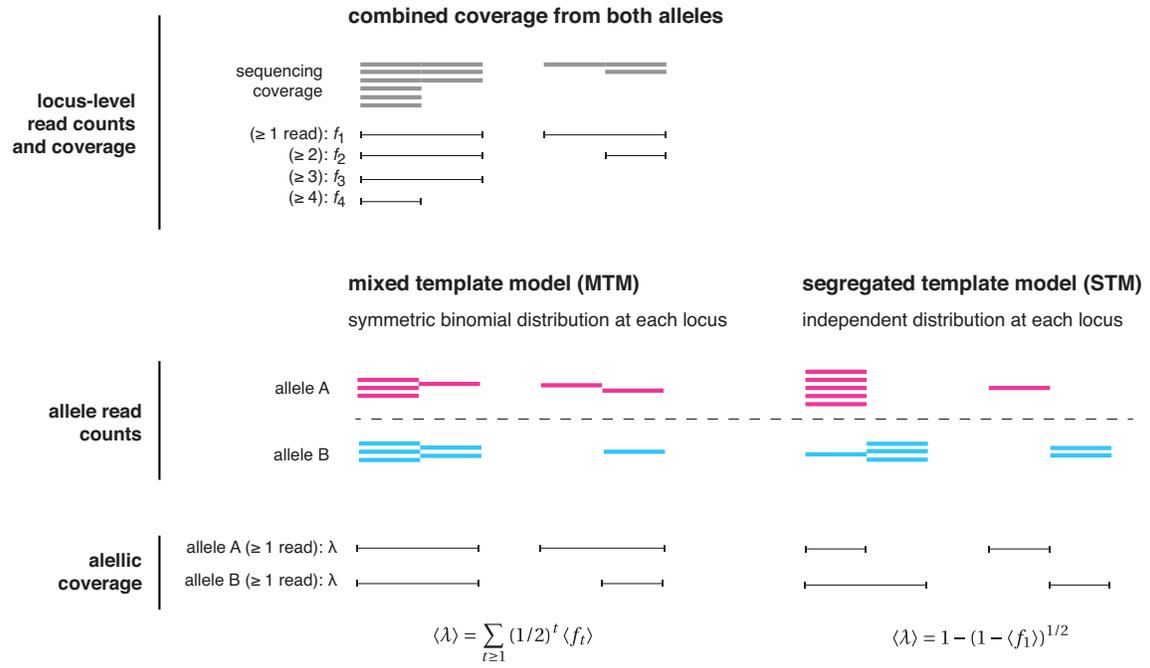
Supplementary Figure 10 | Schematic illustration of two models for allele bias.
Details of the derivations are given in the Methods section.

**a**   **Allelic coverage predictions for single RPE MDA libraries**

observation from 4.5x
- ● combined
- ▼ ref allele
- ▲ alt allele

prediction from 0.1x
- – – – locus-level
- ········ MTM
- ········ STM
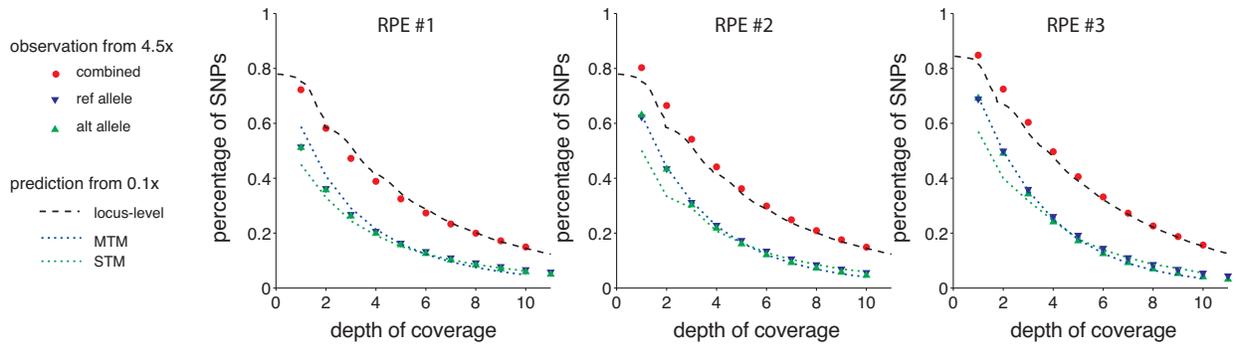
b   **Comparison of allelic coverage in disomic and monosomic chromosomes**

GBM #1

allelic coverage in Chr. 12 predicted from 1x
- ········ MTM
- ········ STM

locus coverage observations
- ● chr.12 (8.3x, disomic)
- ● chr.10 (4.8x, monosomic)

GBM #2

allelic coverage in Chr. 12 predicted from 1x
- ········ MTM
- ········ STM

locus coverage observations
- ● chr.12 (8.9x, disomic)
- ● chr.10 (5.6x, monosomic)

GBM #3

allelic coverage in Chr. 12 predicted from 1x
- ········ MTM
- ········ STM

locus coverage observations
- ● chr.12 (11.1x, disomic)
- ● chr.10 (2.9x, monosomic)

GBM #4

allelic coverage in Chr. 12 predicted from 1x
- ········ MTM
- ········ STM

locus coverage observations
- ● chr.12 (9.7x, disomic)
- ● chr.10 (4.5x, monosomic)

Supplementary Figure 11 | Allele-level amplification bias in fresh RPE-1 and frozen glioblastoma libraries generated by MDA.
(**a**) Comparison of the allele coverage (A or B) predicted from the locus-level coverage (A+B) in two-cell RPE samples with the observation at heterozygous sites in Chr. 1. The observation suggests that amplification of homologous chromosomes more resembles the "mixed template model" (MTM). (**b**) Comparison of the allele coverage prediction from disomic chromosome 12 with the observed coverage in monosomic chromosome 10 in frozen glioblastoma nuclei. The agreement between the "segregated template model" (STM) prediction and the observed coverage in the monosomic chromosome further supports that each chromosome in the single-cell genome is independently amplified.

13

a RPE #3: Positive correlation between ref. and alt. allele counts
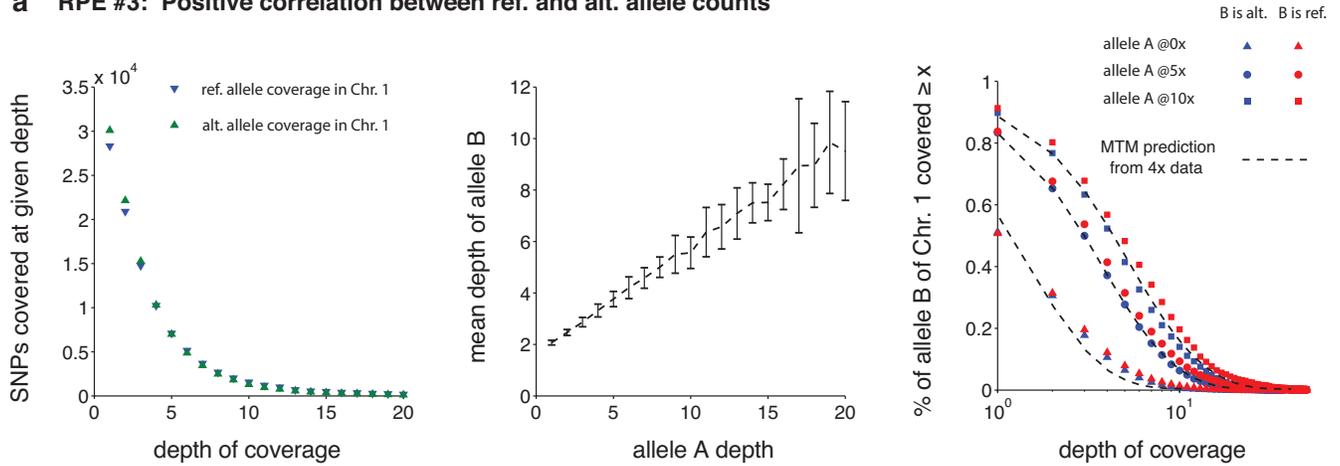
b GBM #4: Independence between ref. and alt. allele counts
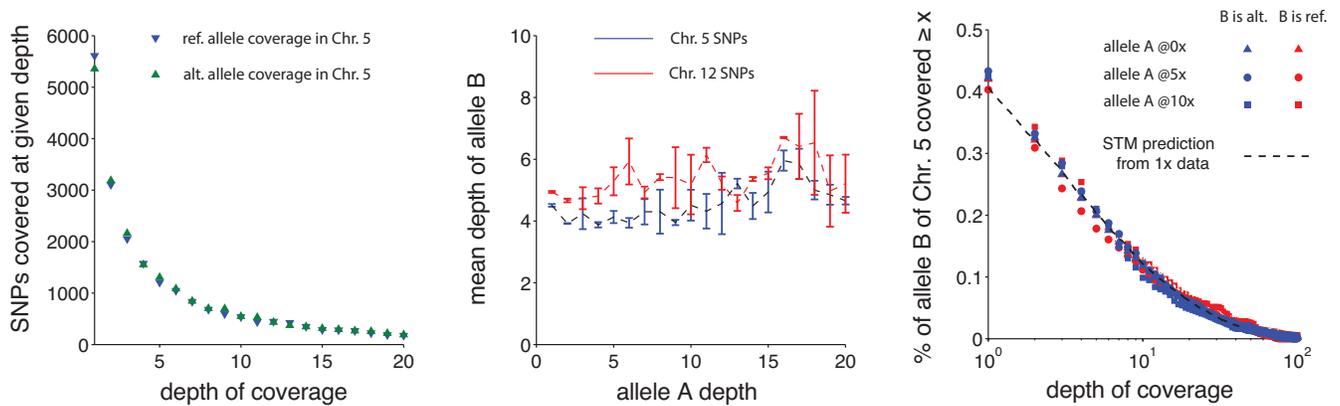
Supplementary Figure 12 | Distinct correlations between the amplification of homologous chromosomes.
(**a**) The depth-of-coverage curves of reference (A) and alternate (B) alleles are identical (left panel) but exhibit a positive correlation (middle panel). The depth-of-coverage curve of allele B at sites where the A allele is covered at different depths (0x, 5x, 10x) also shows a positive correlation with the allele A coverage but can be predicted by the method described above (right panel). The correlation between allele A and allele B coverage indicates that over-amplified regions are not dominated by amplification of one homologue but tend to represent both homologous chromosomes. (**b**) In a representative glioblastoma library, even though the depth-of-coverage curves of reference (A) and alternate (B) alleles are still identical but the correlation between allele A and allele B coverage is absent (middle panel). The depth-of-coverage curves of allele B at sites where the A allele is covered at different depths (0x, 5x, 10x) are almost identical (right panels). These results demonstrate independent amplifications of the two homologous chromosomes.