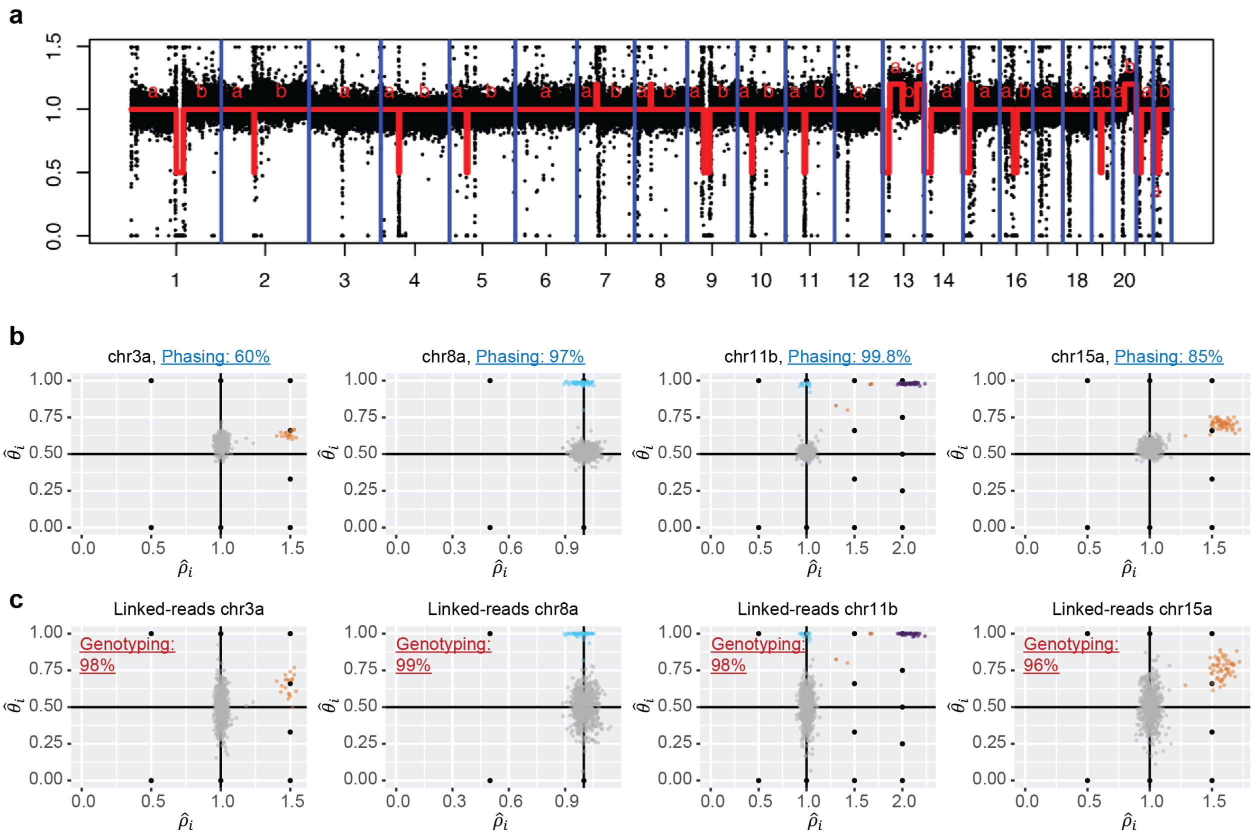
Supplementary Information

Table of Contents

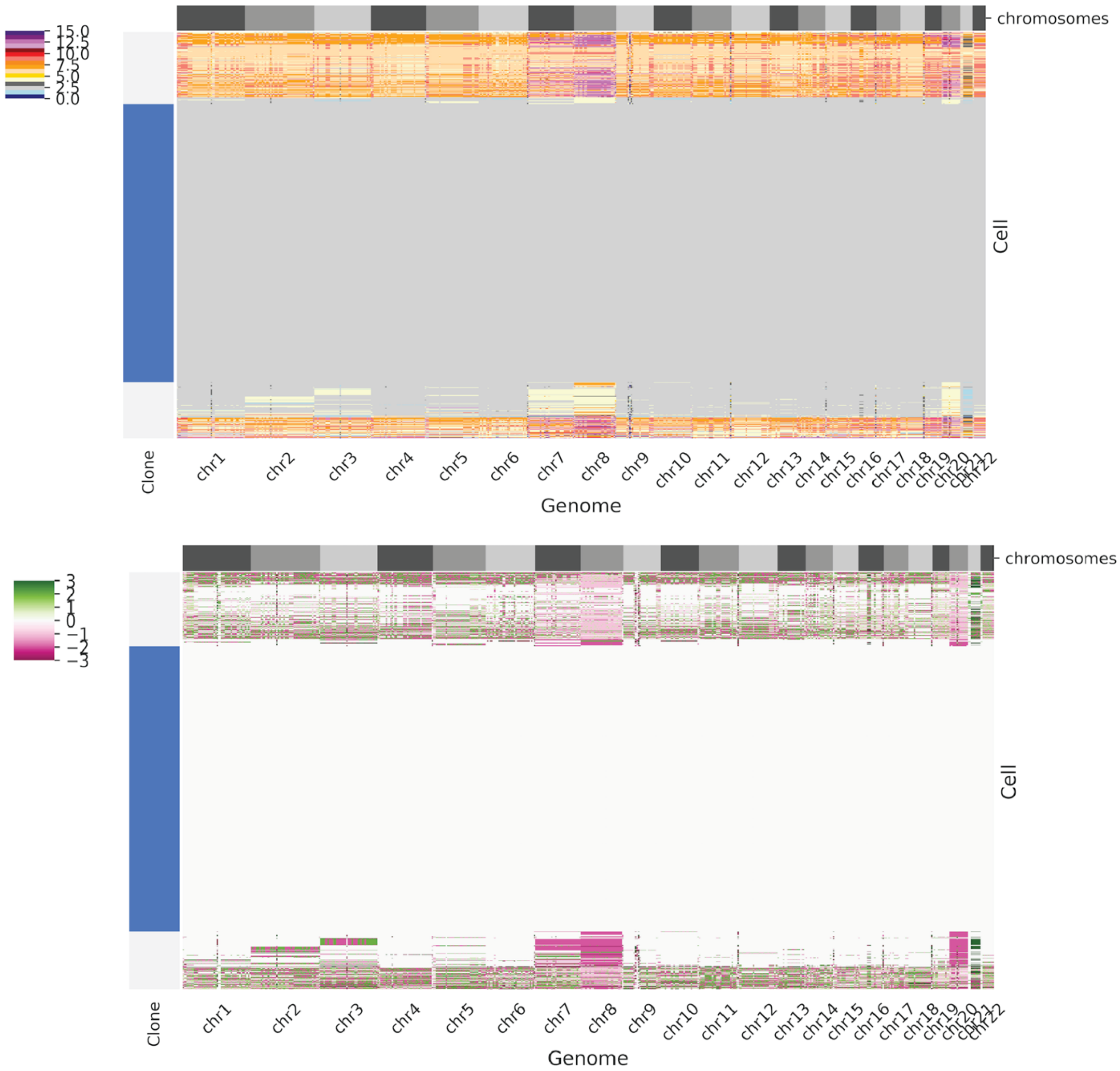
Supplementary Figures

Supplementary Figure 1.



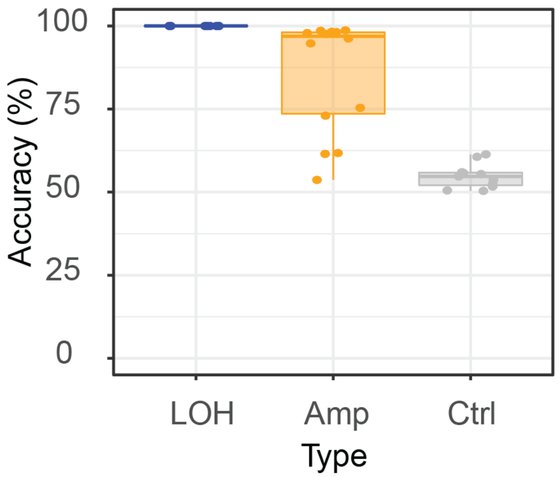
Supplementary Fig. 1: Validation of Alleloscope genotyping results for the P6335 colorectal cancer sample with linked-reads sequencing data. (a) Segmentation of the pooled scDNA-seq data using the HMM algorithm. (b) $\hat{\theta}_i$ values recapitulate CNV carriers that are detected using only coverage for four chromosomal regions on the P6335 tumor sample. Different colors represent different genotype clusters. Phasing accuracy for each region is shown in the title. (c) $\hat{\theta}_i$ calculation using known SNP phases from paired linked-reads sequencing data. Genotyping accuracy is labeled in the plots. The colors follow the clustering results from b. The color scheme is the same as that in Fig. 2 and Supplementary Fig. 5.

Supplementary Figure 2.



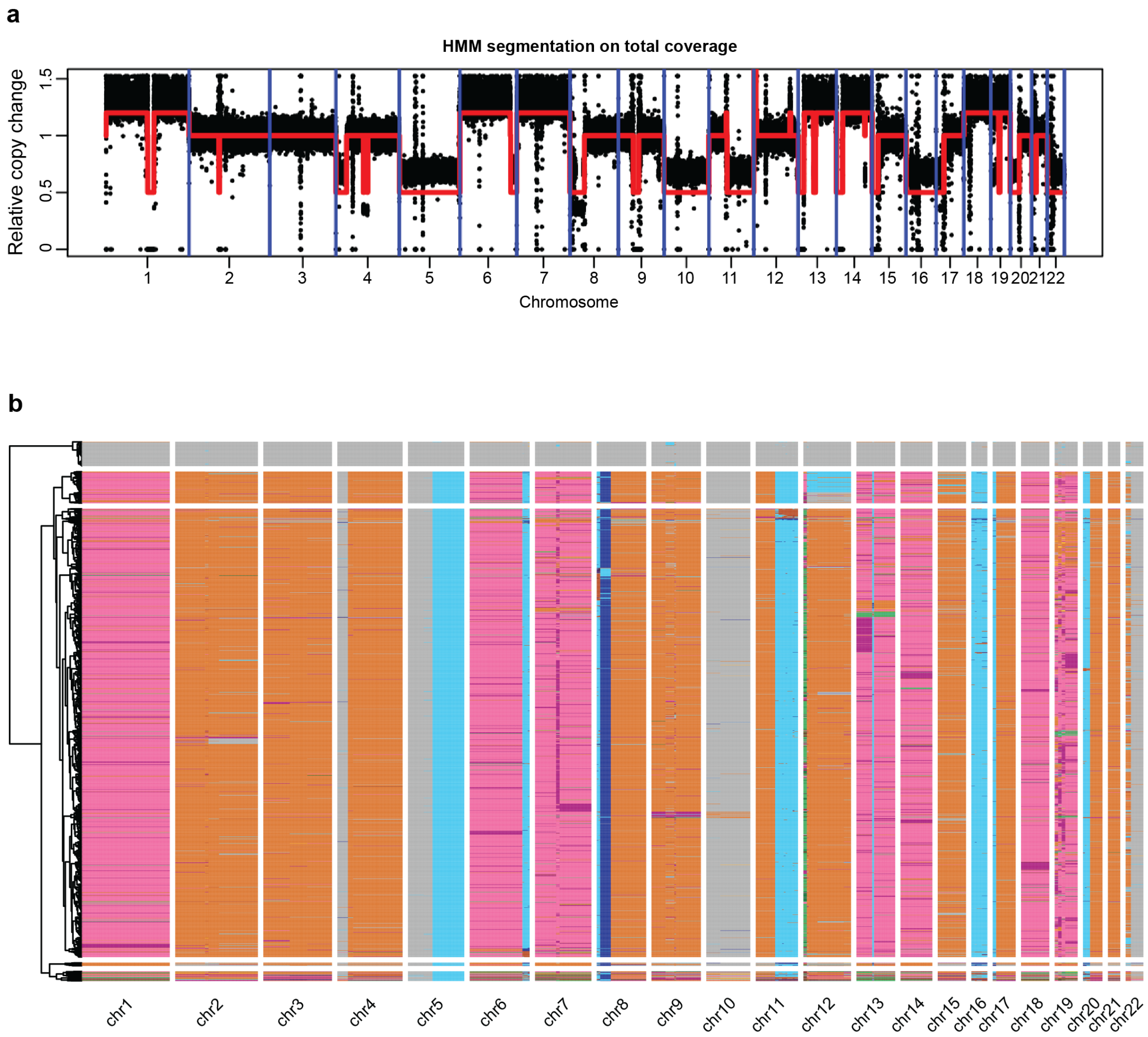
Supplementary Fig. 2: Heatmaps of allele-specific genotypes and haplotype-specific genotypes from CHISEL for the P5931 sample.

Supplementary Figure 3.



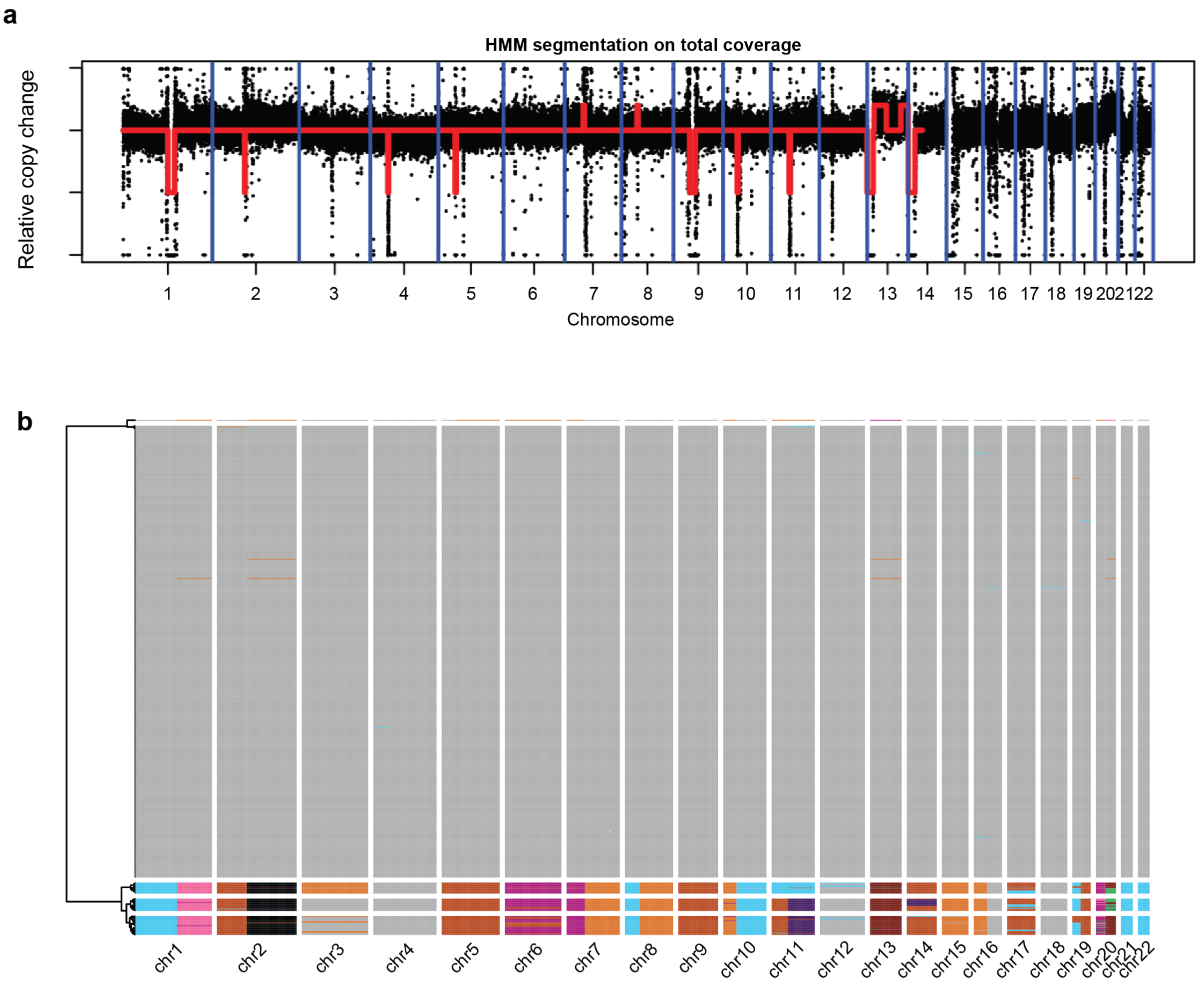
Supplementary Fig. 3: Phasing accuracy for the CNA regions in the P6198 sample by comparing to the matched linked-reads sequencing data. LOH: segments with any LOH events. Amp: segments with amplifications that lead to allelic imbalance. Ctrl: control segments without allelic imbalance.

Supplementary Figure 4.



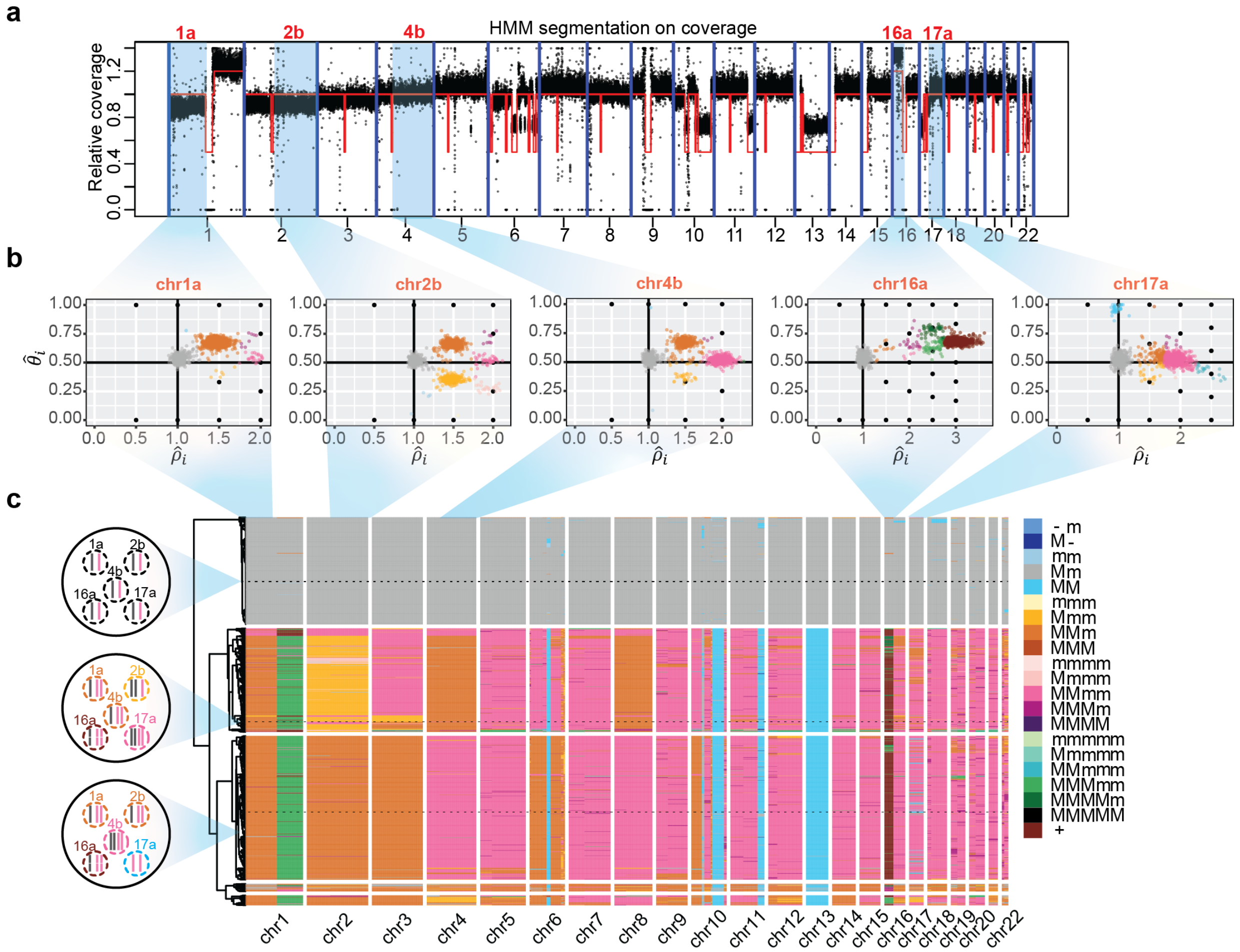
Supplementary Fig. 4: Segmentation plot and genotype heatmap for the P6198 sample. The color scheme is the same as that in Fig. 2 and Supplementary Fig. 5.

Supplementary Figure 5.



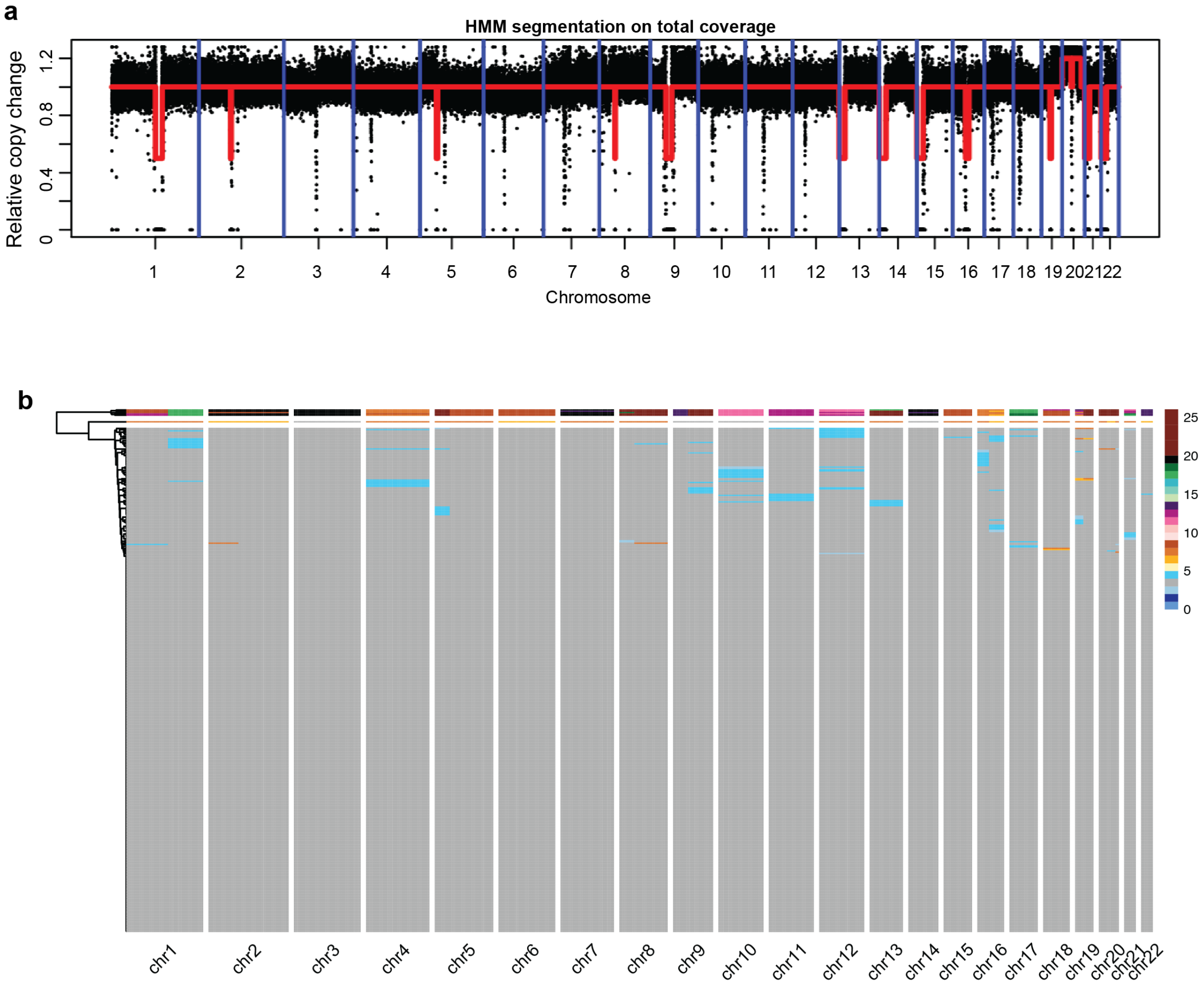
Supplementary Fig. 5: Segmentation plot and genotype heatmap for the P6335 sample. The color scheme is the same as that in Fig. 2 and Supplementary Fig. 5.

Supplementary Figure 6.



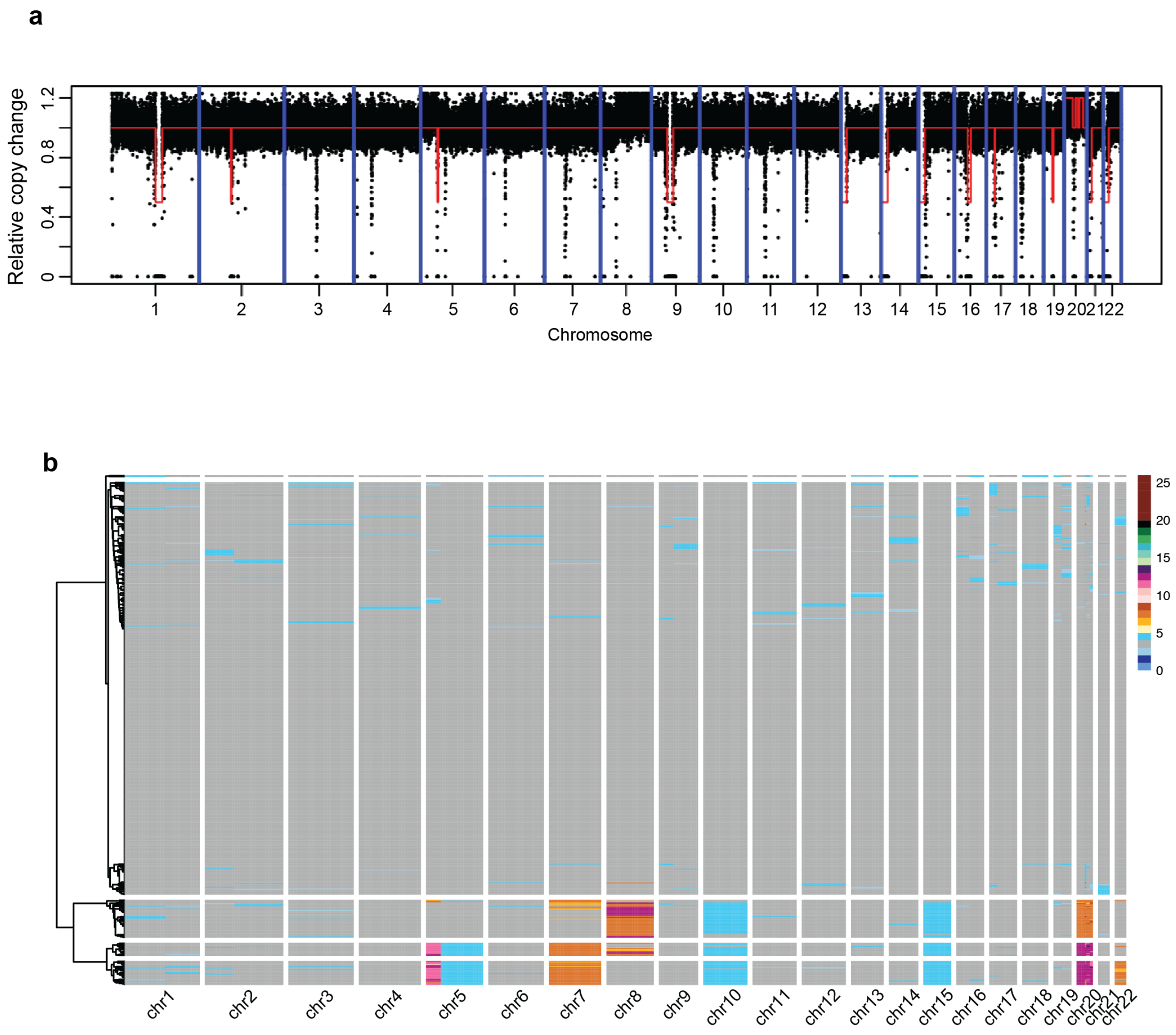
Supplementary Fig. 6: Segmentation plot and genotype heatmap for the BC10X sample. (a) Genome segmentation using HMM on the pooled coverage signals across the cells. (b) Genotype profiles of five example regions. The coloring scheme is same as that in part (c). (c) Hierarchical clustering of single-cell ASCN genotypes reveals complex subclone structure. Genotypes of the five regions in three example cells from the three major subclones are shown in the left. Different colors represent different genotypes. In the color panel, M and m represent the “Major haplotype” and “minor haplotype” respectively.

Supplementary Figure 7.



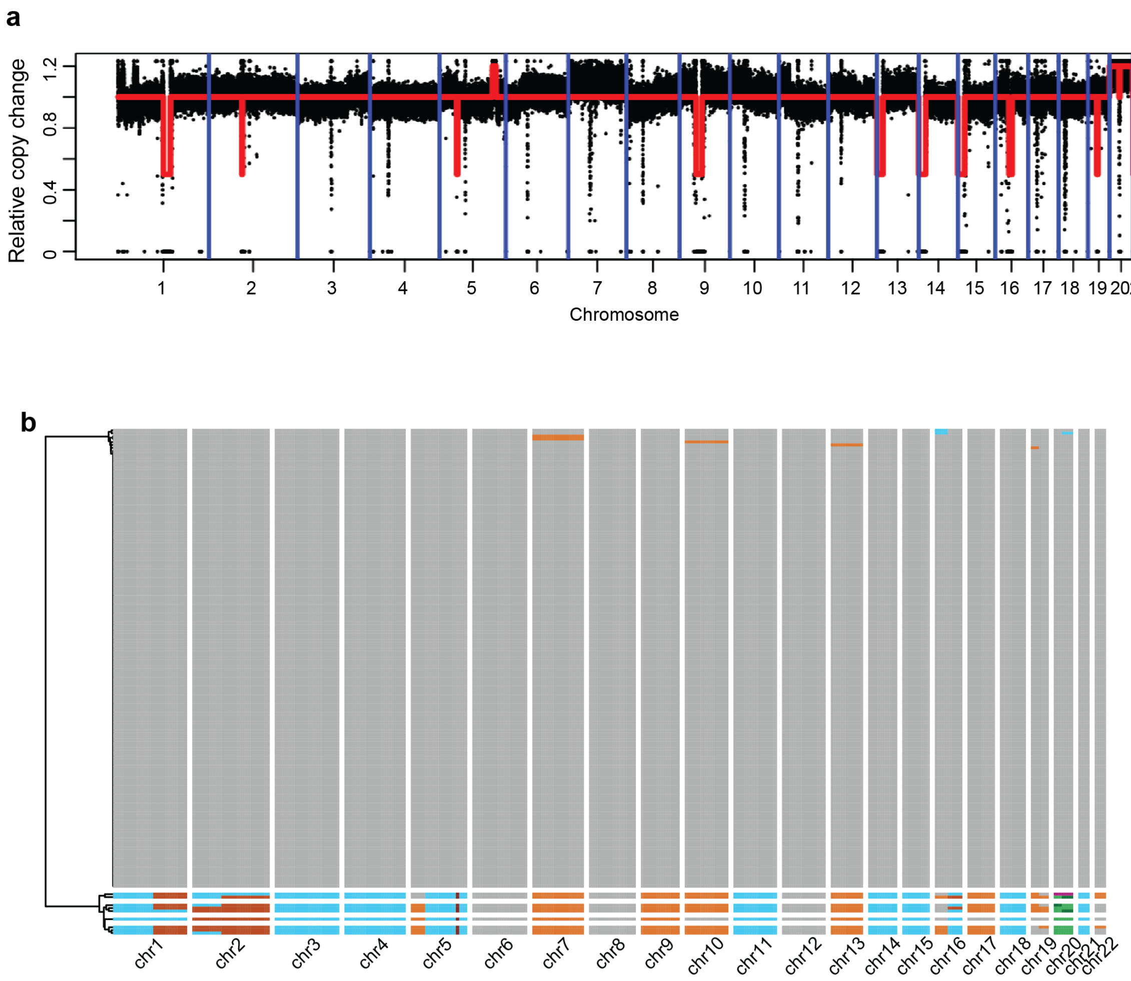
Supplementary Fig. 7: Segmentation plot and genotype heatmap for the P5846 sample. The color scheme is the same as that in Fig. 2 and Supplementary Fig. 5.

Supplementary Figure 8.



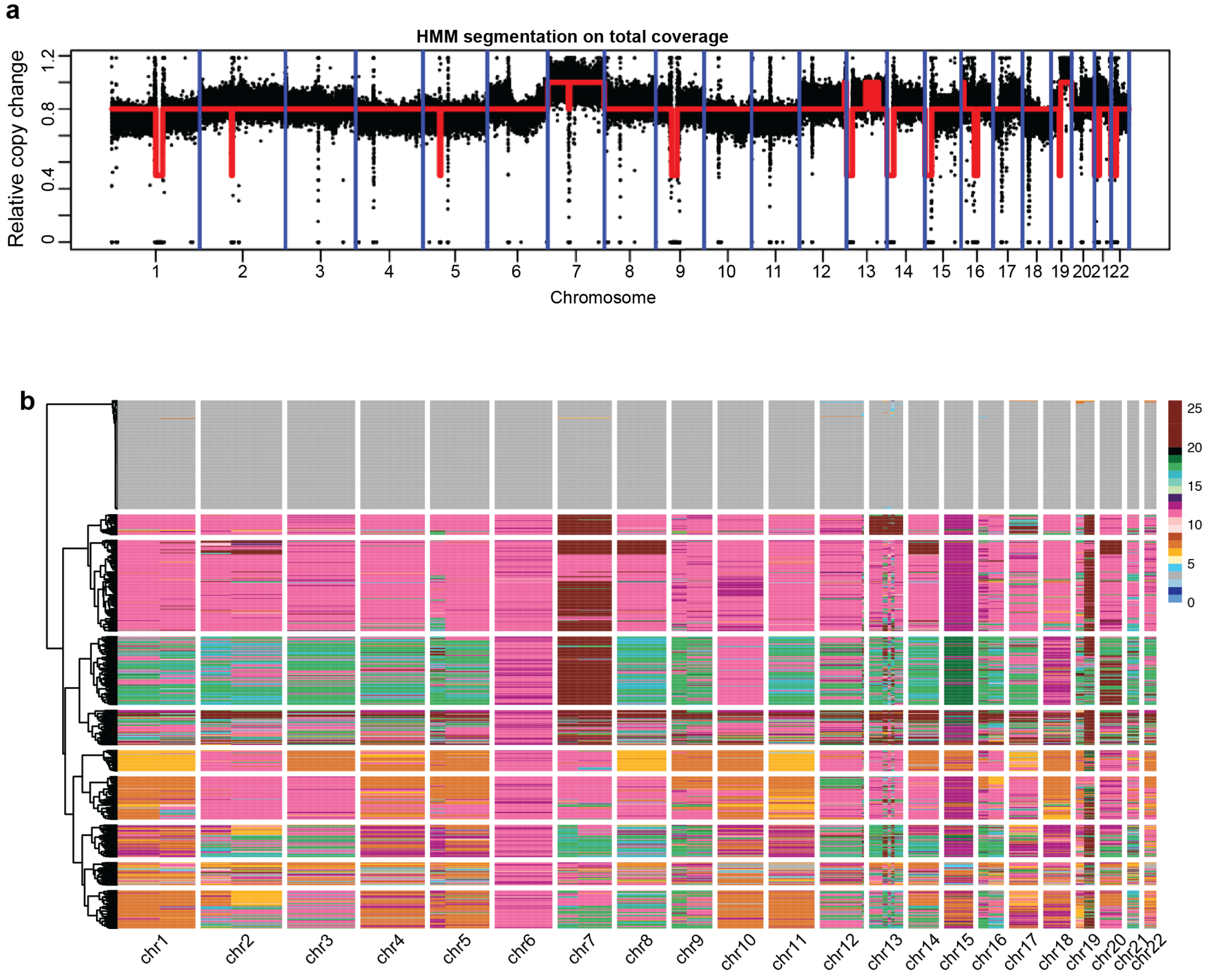
Supplementary Fig. 8: Segmentation plot and genotype heatmap for the P5847 sample. The color scheme is the same as that in Fig. 2 and Supplementary Fig. 5.

Supplementary Figure 9.



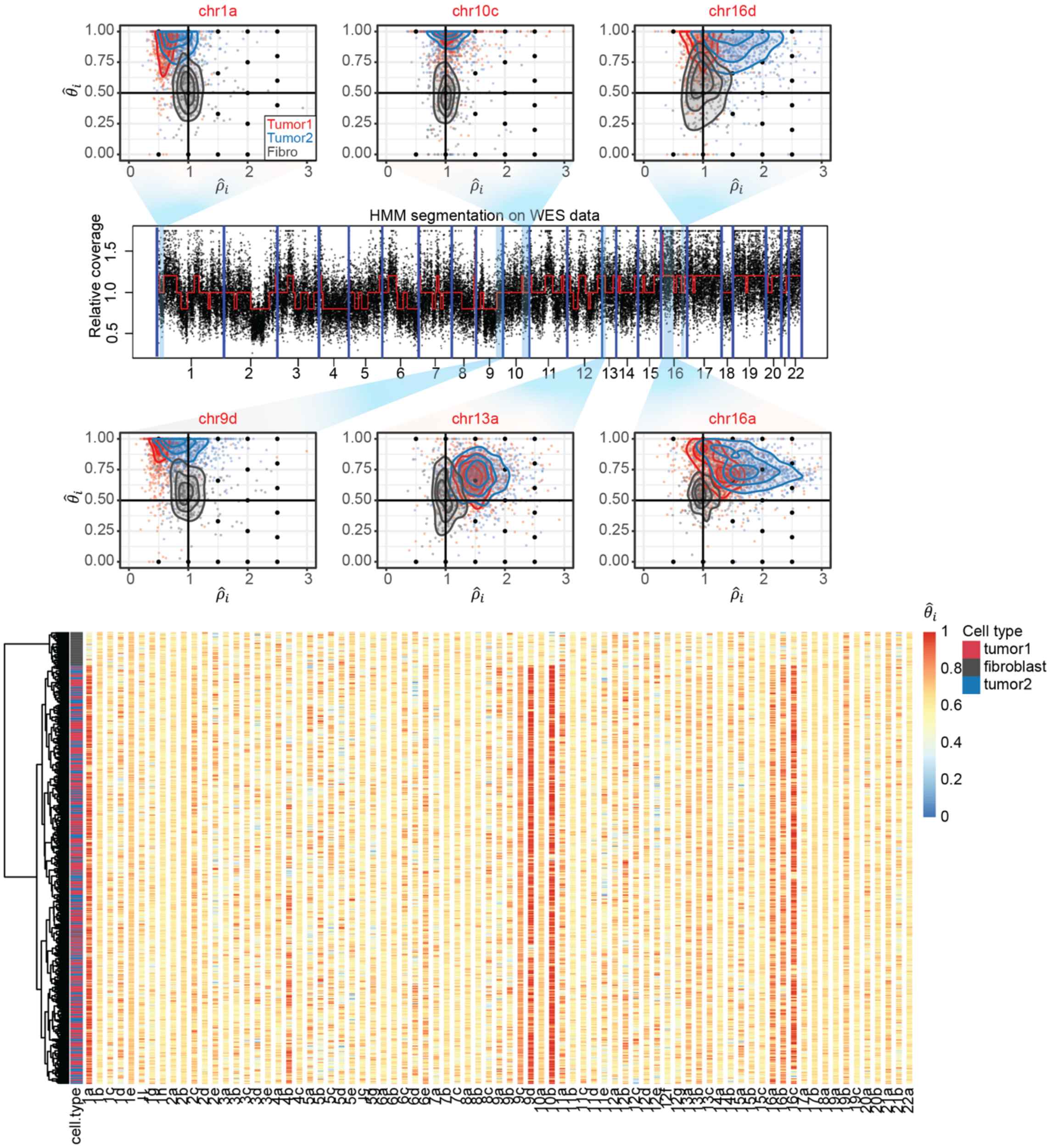
Supplementary Fig. 9: Segmentation plot and genotype heatmap for the P5915 sample. The color scheme is the same as that in Fig. 2 and Supplementary Fig. 5.

Supplementary Figure 10.



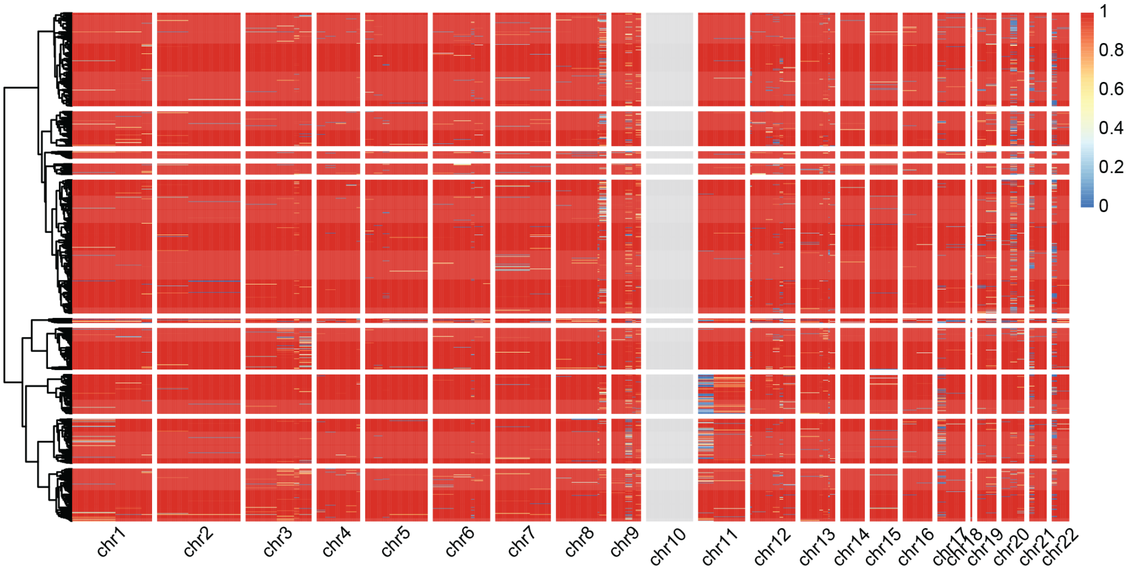
Supplementary Fig. 10: Segmentation plot and genotype heatmap for the P6461 sample. The color scheme is the same as that in Fig. 2 and Supplementary Fig. 5.

Supplementary Figure 11.



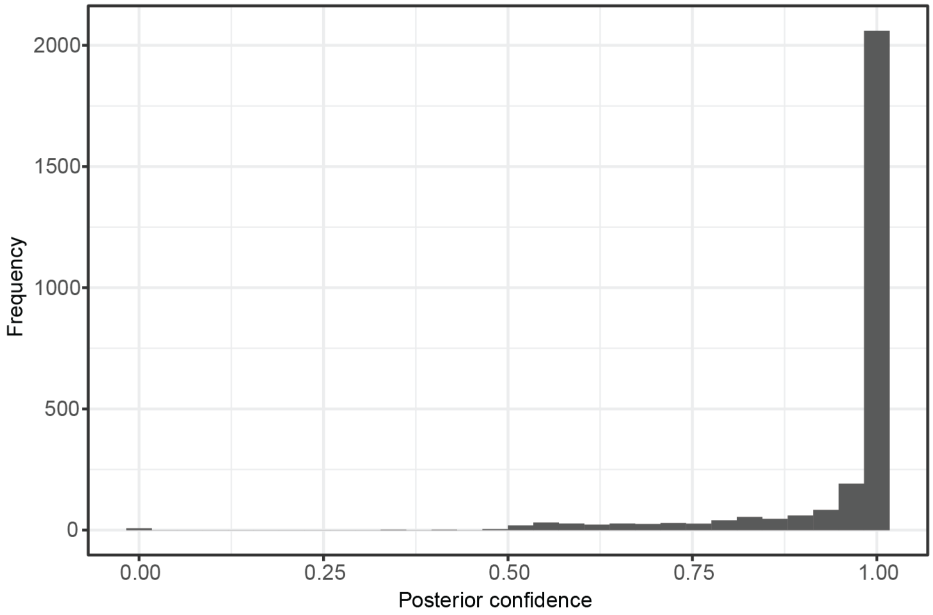
Supplementary Fig. 11: Single cell genotyping of CNV events by Alleloscope for scATAC-seq data of a basal cell carcinoma sample (SU006[1]). (a) Genotype profiles of six example regions. The regions were taken from the segmentation of whole exome sequencing (WES) data. Each dot represents a cell-specific ($\hat{\rho}_i$, $\hat{\theta}_i$) pair. Cells are colored by annotation derived from peak signals[1]. Two tumor cell clusters, identified using ATAC peaks, are labeled by red and blue; fibroblasts (Fibro) are labeled by grey. Density contours of the three cell subpopulations are also shown. (b) Hierarchical clustering of cells in scATAC-seq by $\hat{\theta}_i$ reveals that the two tumor subpopulations are differentiated by peak signals that don't correlate with broad copy number events.

Supplementary Figure 12.



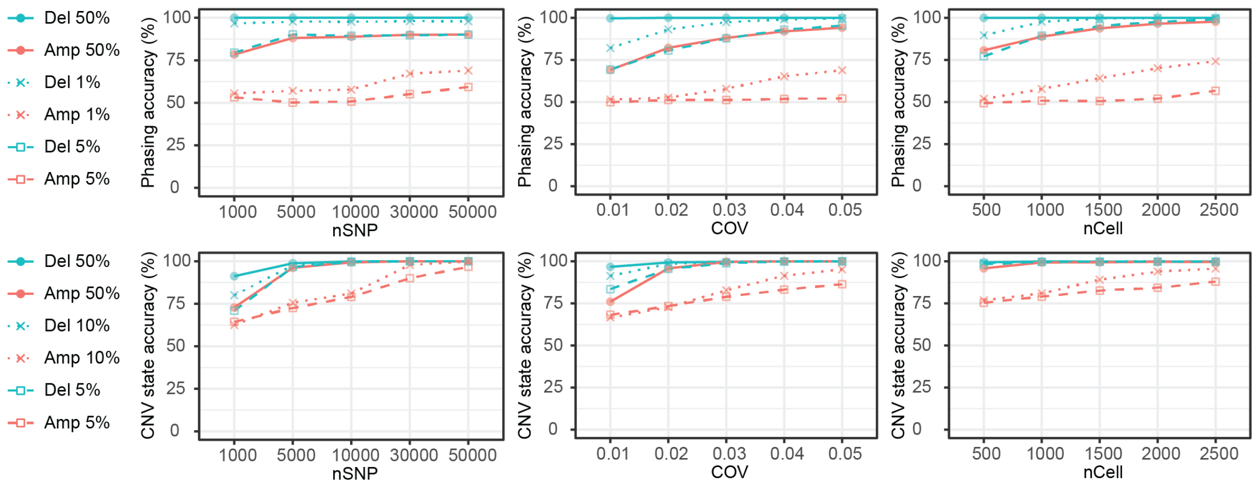
Supplementary Fig. 12: Confidence scores for the genotype assignment of each cell in each region for the SNU601 scDNA-seq dataset.

Supplementary Figure 13.



Supplementary Fig. 13: Distribution of the posterior confidence scores of subclone assignment for the 2,753 cells from SNU601 scATAC-seq.

Supplementary Figure 14.



Supplementary Fig. 14: Power for the detection of 1 copy deletion and 1 copy amplification for data of varying coverage (per base), heterozygous SNP count, and number of cells. The heterozygous SNP count reflects the size of the region: larger regions contain more heterozygous loci. Cells were clustered based on the minimum distance of $\hat{\theta}_i$ to the canonical values. Top: phasing accuracy, defined as the proportion of SNPs with $\hat{I}_j$ correctly estimated; bottom: cell CNV state accuracy, defined as the proportion of cells that are correctly assigned to carrier state. Amp: amplification. Del: deletion. Line types represent different proportions (0.5%, 0.1% and 0.05%) of carrier cells. The number of SNPs, coverage, number of cells and purity were set as 10,000, 0.03, 1000, and 0.5 if not specified.

Supplementary Methods

Simulations and Power Analysis

For a simulated region, let n be the number of cells, m be the number of heterozygous SNPs, θ be the major haplotype proportion, and μ_i be the total coverage of cell i sampled from the cells on chr7 in the P5931 tumor sample. For cell i, we simulated total coverage of SNP j (μ_{ij}) using a Poisson distribution

$$\mu_{ij} \sim Poisson(\mu_i),$$

where $i = 1 \sim n$. Parallelly, phases of SNP j (I_j) were simulated under a Bernoulli distribution

$$I_j \sim Bernoulli(0.5),$$

where I_j indicates whether a reference allele is on the major haplotype for SNP j, and j=1~m. Using μ_{ij} and I_j, simulated read counts of reference alleles of SNP j in cell i (A_{ij}) were simulated under a Binomial distribution

$$A_{ij} \sim Binomial(\mu_{ij}, p_{ij}),$$

where p_{ij} is the proportion of the reference allele at loci j in cell i with the values shown in the following table

p_{ij}	cell i with CNA	cell i without CNA
$I_j = 1$	θ	0.5
$I_j = 0$	$1 - \theta$	0.5

Then simulated read counts of alternative alleles of SNP j in cell i (B_{ij}) were retrieved by

$$B_{ij} = \mu_{ij} - A_{ij}$$

In the first simulation used to illustrate distribution of the estimates from Alleloscope, we fixed the cell number n to be 1,000, the SNP number m to be 10,000 which are typical in real datasets. θ was set to be 1 and 0.66 for cells carrying deletion and one-copy amplification respectively with the purity equal to 0.5. On the simulated A_{ij} and B_{ij} matrices Alleloscope estimated phases for each SNP and CAN states for each cell. Distribution of the estimated values versus the true values are visualized using boxplots.

To know the effects of SNP numbers, cell coverage, cell numbers, and purity, power analysis was performed for one-copy deletion and one-copy amplification scenarios.

We assessed the accuracy for phasing and cell-level CNA state estimation under the following scenarios: SNP numbers from 1,000 to 50,000, mean coverage from 0.01 to 0,05 for each cell, cell number from 500 to 2500. For different scenarios, we assessed the effect of three purity: 0.5, 0.1, and 0.01, reflecting from larger subclones to rare subclones. All parameters remained the same as those in the previous paragraph except for the parameters that were assessed. Phasing accuracy was calculated by comparing true I_j's and estimated $\hat{I}_j$'s in the region. If $\hat{I}_j \geq 0.5$, the values were considered as 1. Otherwise, the values were considered 0. On the other hand, the accuracy of cell CNA state estimation was the clustering accuracy using the estimated $\hat{\theta}_i$ values. Cells with $\hat{\theta}_i$ values smaller than the midpoints between true θ of normal cells ($\theta_o = 0.5$) and true θ of carriers ($\theta_{del} = 1$; $\theta_{amp} = 0.66$) were considered as normal cells; otherwise, cells were considered as carriers. The clustering accuracy was calculated by comparing the clusters to the true cell states.

Reference

1. Satpathy, A.T. et al. Massively parallel single-cell chromatin landscapes of human immune cell development and intratumoral T cell exhaustion. *Nat Biotechnol* **37**, 925-936 (2019).