Supplementary Information Guide (Navin et al.)

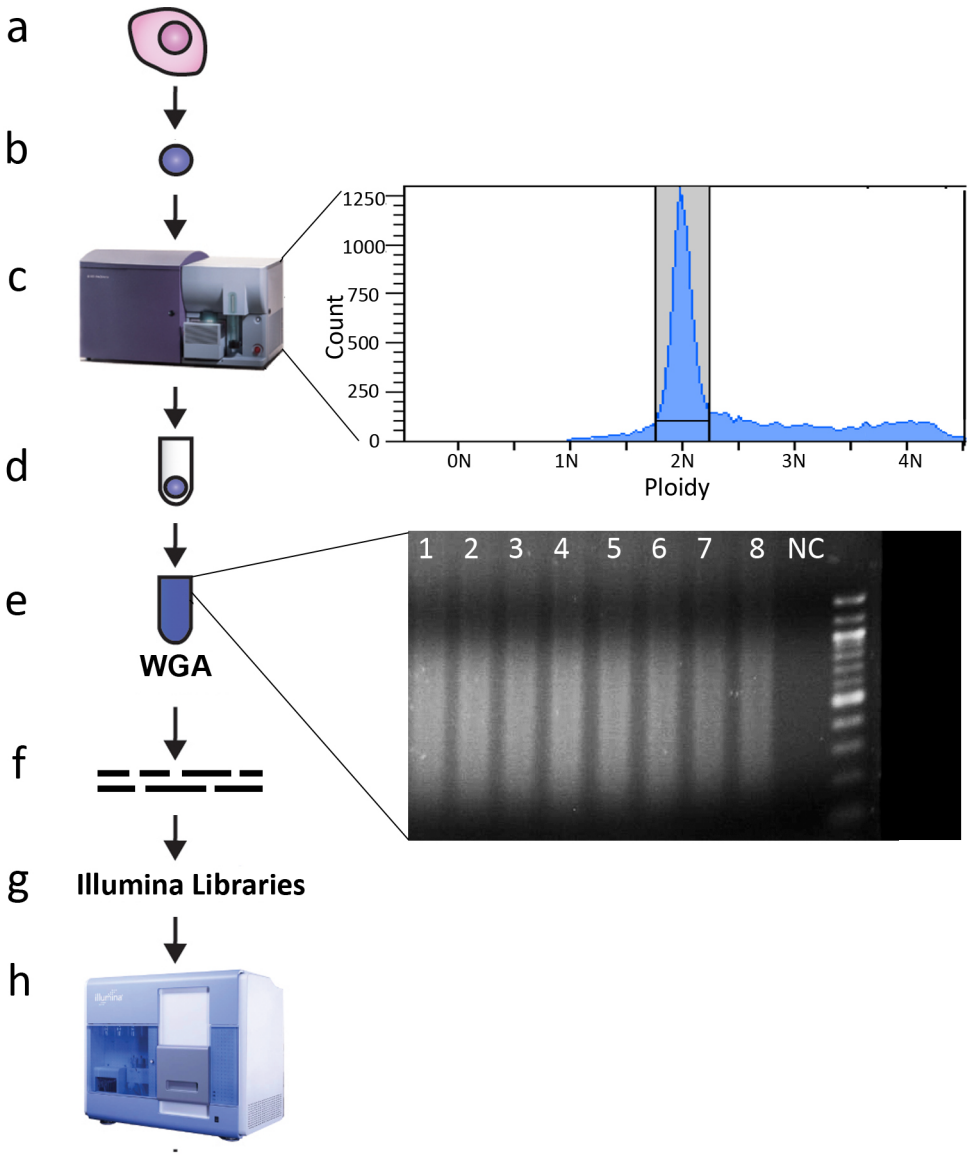
Supplementary Figures

Supplementary Tables

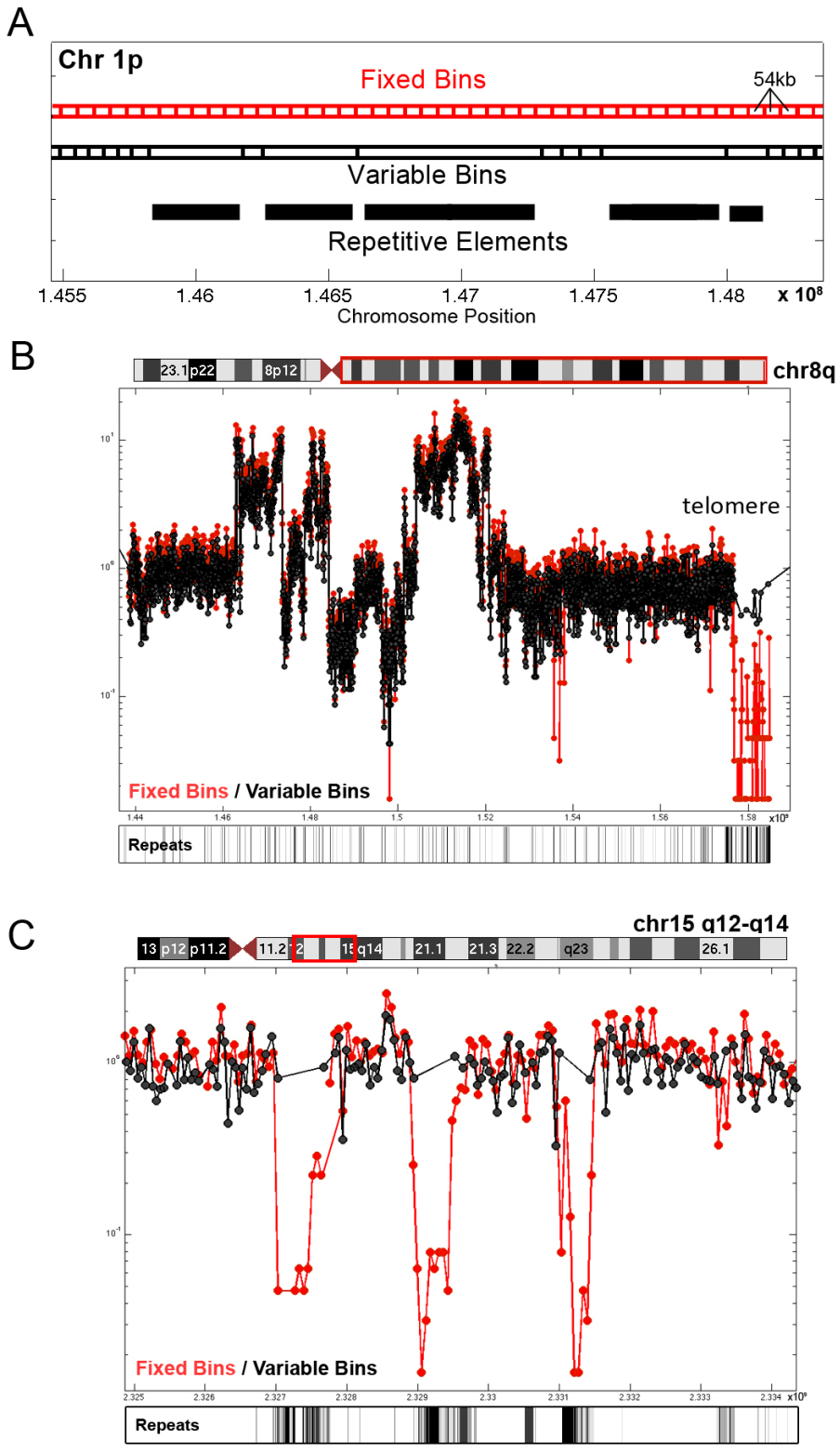
Supplemental Methods

Data Access

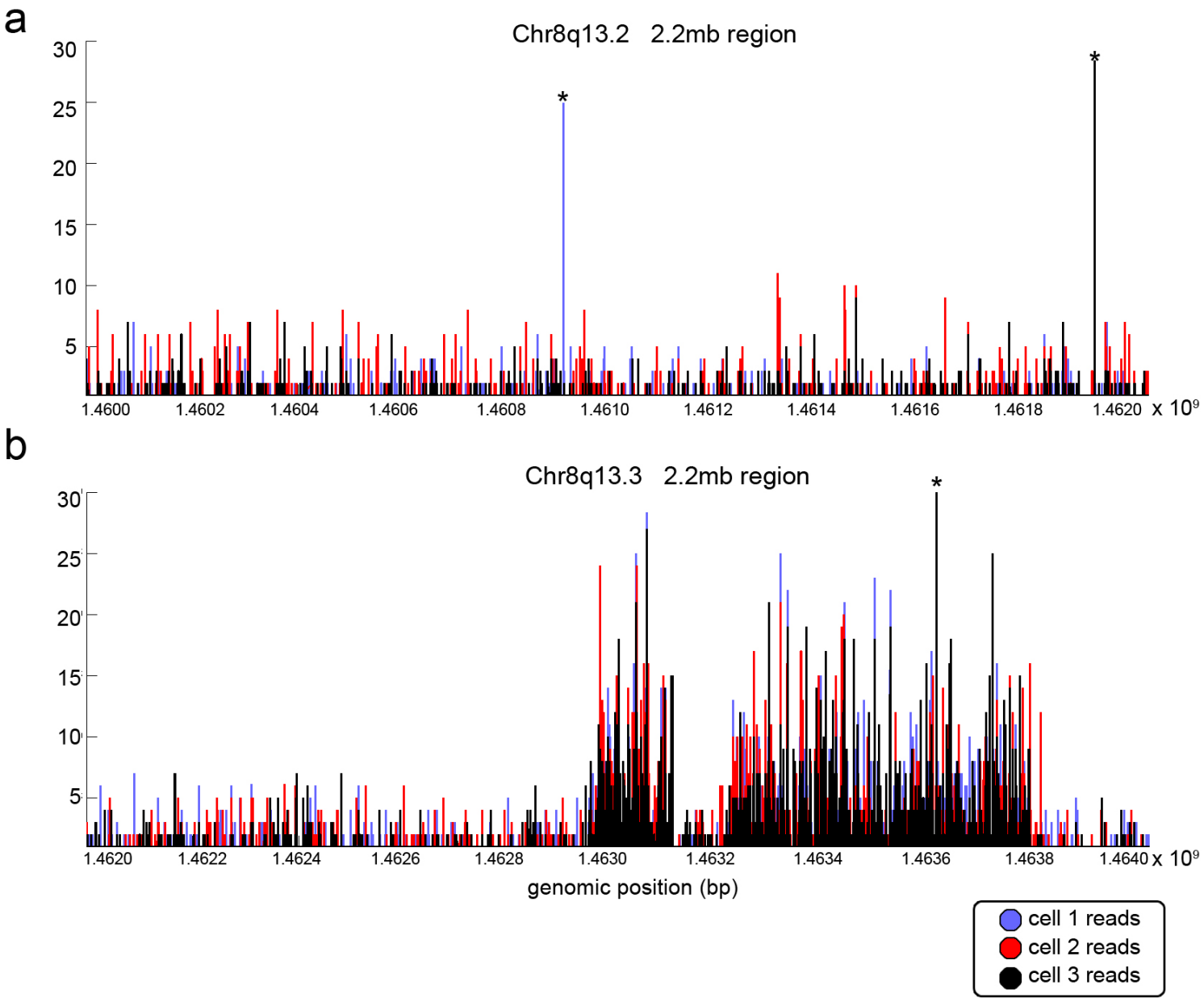
All data has been deposited into the NCBI Sequence Read Archive (SRA018951.105)



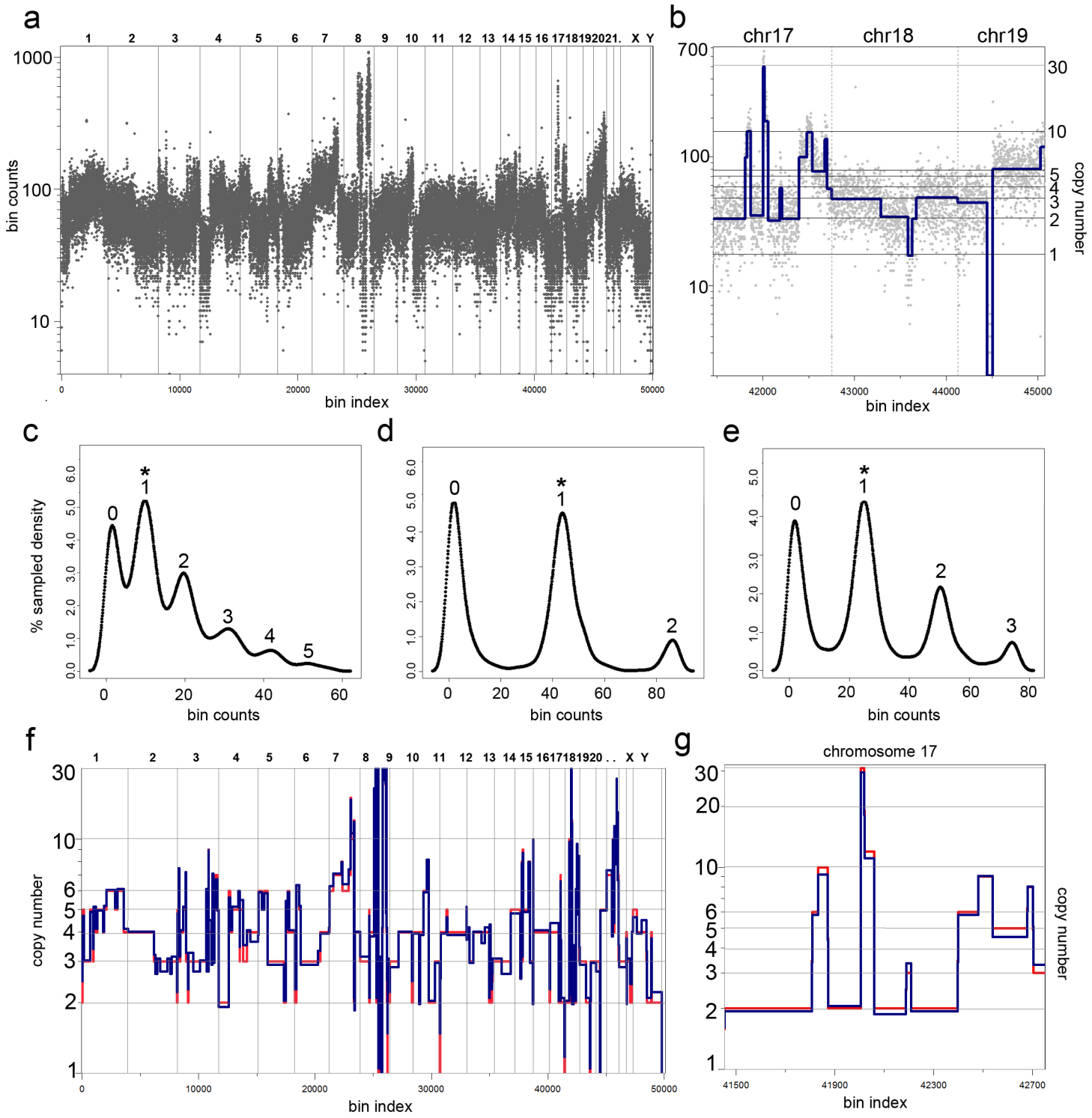
Supplementary Fig.1 | Overview of the SNS Approach
(a) Cells are lysed and nuclei are isolated. (b) DAPI stained nuclear suspensions are loaded into the FACS machine. (c) Cells are run to measure ploidy, and then subpopulations are gated from which individual nuclei are deposited into single wells on a 96-well plate containing lysis buffer. (d) DNA is fragmented in alkali solution and amplified by WGA resulting in a distribution of fragments from 100-1000bp that can be visualized on an agarose gel. Negative control (NC) reactions are also prepared without any nuclei deposited which show no DNA fragments on agarose gels. (e) WGA amplified DNA is sonicated to remove specific 28bp adapters (f). Illumina single-end libraries are prepared and each library is sequenced on individual flow-cell lanes.



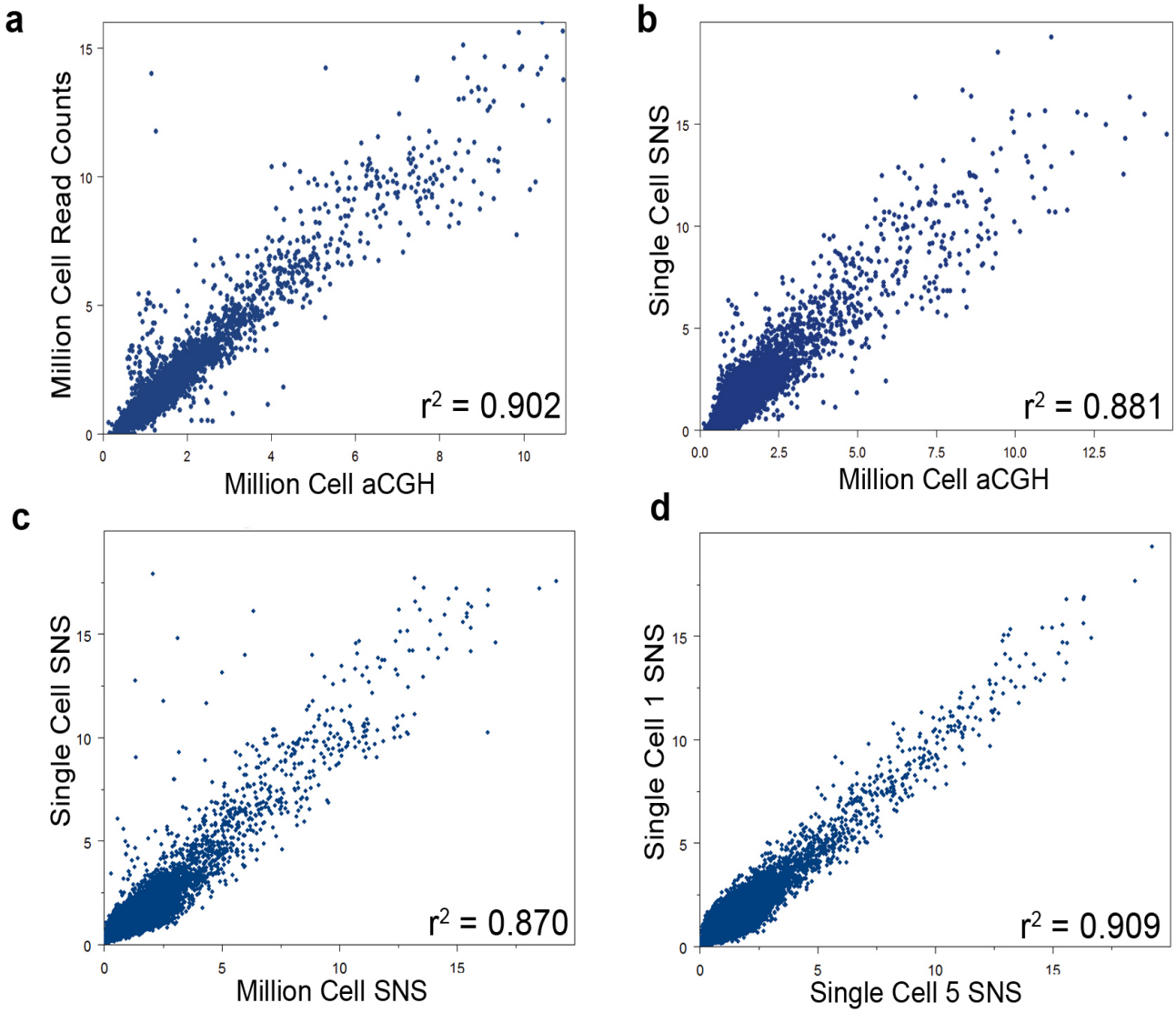
Supplementary Fig.2 | Fixed vs. Variable Binning of Sequence Read Depth
(a) 54kb uniform length and variable (uniform expected count) bin intervals are plotted on a repetitive region of chromosome 1p, illustrating that variable bin size increases to maintain a constant mean read count. (b-c) Uniform length and variable bins were calculated for a single SK-BR-3 cell and plotted along with an annotation track for repetitive elements below. (b) Chromosome 8q shows similar copy number in both uniform and variable bin profiles, but the uniform bin method erroneously shows the region near the telomere as a large homozygous deletion. (c) Chromosome 15q12-14 shows three erroneous homozygous deletions in a repetitive region of the human genome using the uniform length bin profile, while the variable bin method maintains a constant ground state copy number.



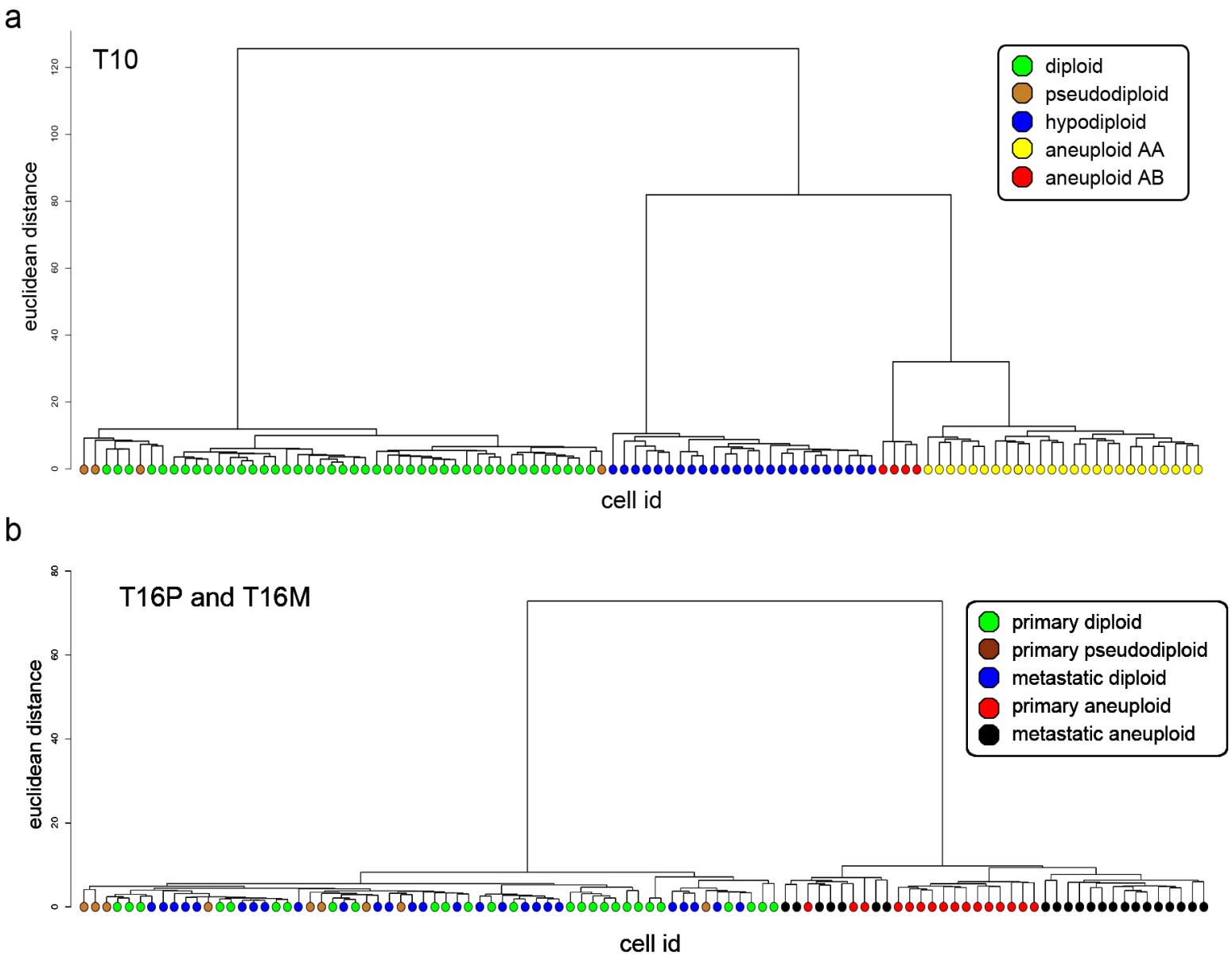
Supplementary Fig.3 | Pileups and Distributions of WGA Sequencing Reads in Single Cells
(a) Sequence read depth from three single SK-BR-3 cells are plotted for a 2.2mb region on chromosome 8q13.2 illustarting that reads are fairly randomly distributed and non-overlapping between three different single cells. Two regions of over-replicated pileups are shown and marked with asterisks. (b) Read depth from three different single SK-BR-3 cells are also plotted on chromosome 8q13.3 that contains an amplification of the MYC locus, showing an increase in sequence read density.



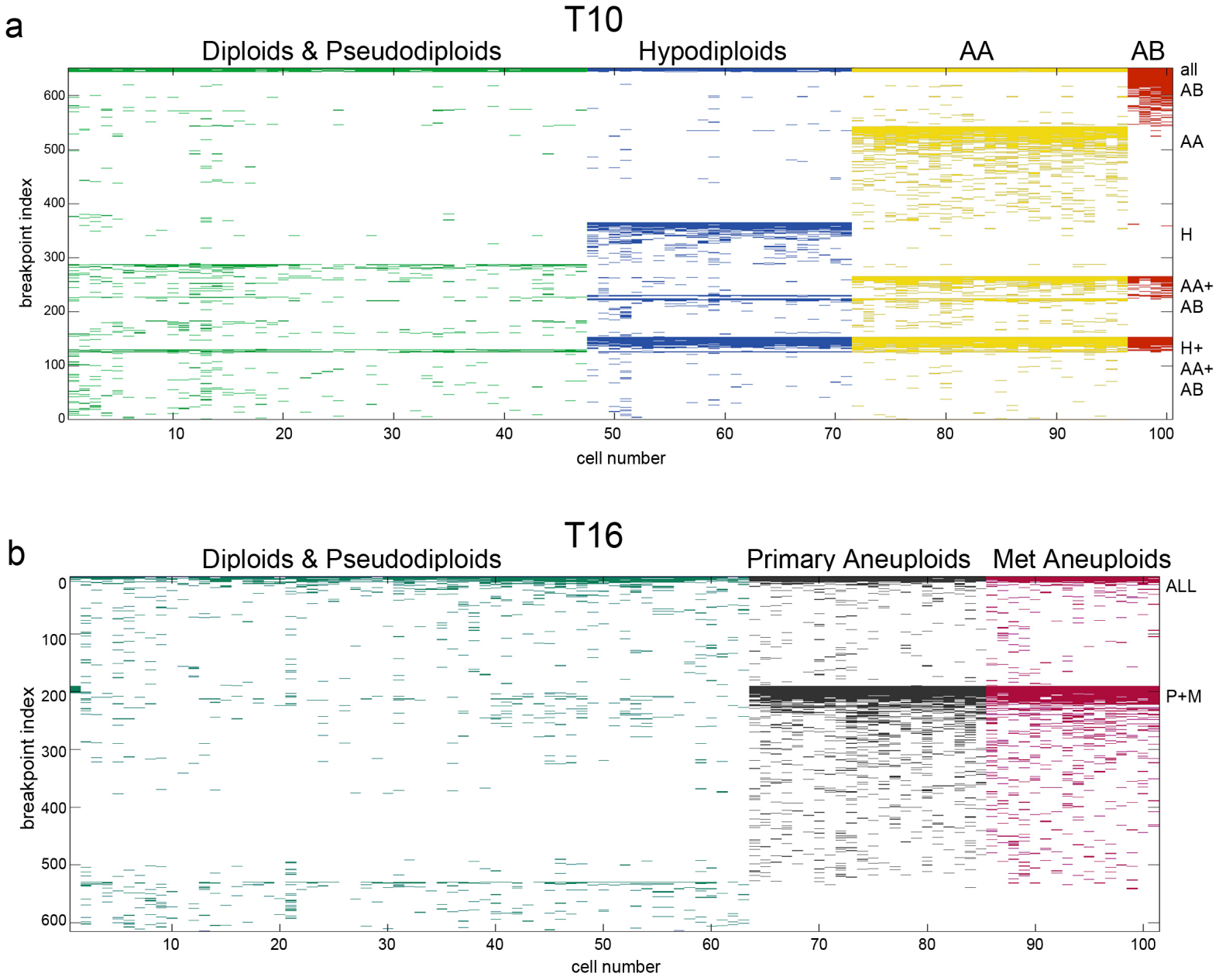
Supplementary Fig.4 | Integer Copy Number Quantification from Read Density
(a-c, f-g) Integer copy number calculations are shown for a single SK-BR-3 cell. (a) Mapped sequence reads are counted in variable bins of uniform expected read density. (b) Variable bin counts in grey are plotted on a log scale and KS-segments are plotted in blue for a region from chromosome 17-19. (c-e) Gaussian kernel smoothed density plots are shown with asterisks denoting the first increment peak, used for normalization, for (c) an SK-BR-3 cell, (d) a hypodiploid tumor cell from T10, and (e) an aneuploid tumor cell from T10. (f) A normalized KS-segmented profile in blue is compared to the integer copy number profile in red for an SK-BR-3 cell, and (g) this region is shown for chromosome 17 with integer copy number on the ordinate.



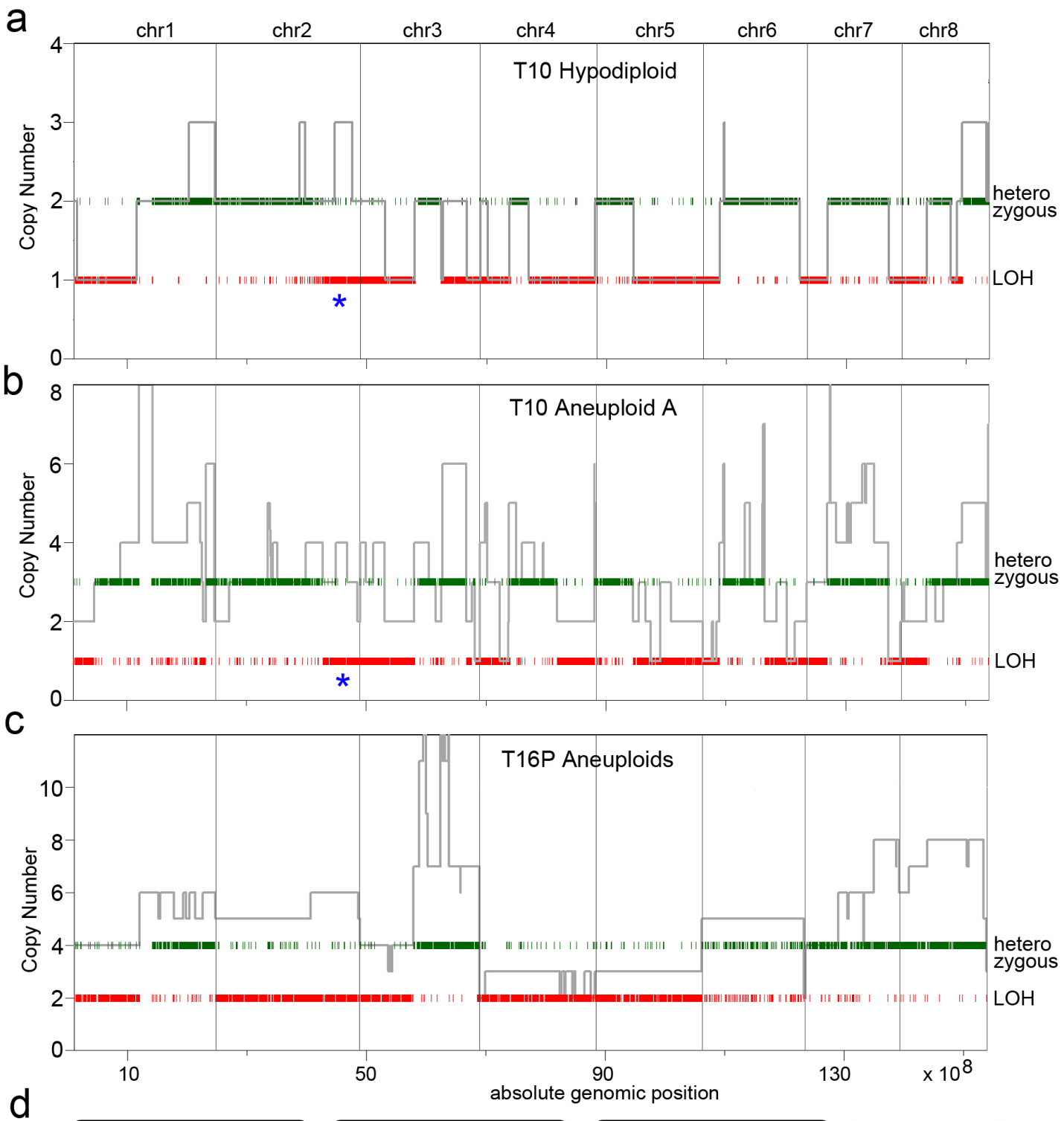
Supplementary Fig.5 | R^2 plots of CGH and SNS Profiles from Single Cells and Bulk DNA
(a-d) R^2 plots of SK-BR-3 copy number profiles using various platforms and number of cells. (a) Sequence read count of a million cell sample compared to an array CGH profile using million of cells (b) Single cell profile measured by SNS versus a million cell array CGH profile, (c) Single cell profile measured by SNS compared to a million cell sequence read count profile, (d) A single cell profile (cell number 1) compared to another single cell profile (cell number 5) measured by SNS.



Supplementary Fig.6 | Chromosome Breakpoint trees of T10 and T16
(a-b) Hierarchical clustering was used to construct trees from chromosome breakpoint patterns from single cells (a) hierarchical tree of 100 single cells from T10 (b) hierarchical tree of 100 single cells from T16 combining profiles from both the primary and metastatic tumors.



Supplementary Fig.7 | Biclustered Heatmaps of Chromosome Breakpoints
(a) T10 heatmap of common chromosome breakpoints that have been biclustered and ordered to correspond to the integer copy number tree (b) T16P and T16M heatmap of common chromosome breakpoints that have been biclustered and ordered according to the integer copy number tree.



d

D	H	Events
aa	aa	77648257
ab	ab	7479
ab	a	6986

T10 Hypodiploids

D	A	Events
aa	aa	85351702
ab	ab	8570
ab	a	7974

T10 Aneuploid A

D	A	Events
aa	aa	99751408
ab	ab	6803
ab	a	7609

T16P Aneuploids

copy number
LOH mutations
heterozygous

Supplementary Fig. 8 | Comparison of LOH and Copy Number Profiles

(a-c) Copy number profiles from tumor subpopulations are plotted with LOH at heterozygous SNPs that were identified by comparing to the diploid subpopulation. Asteriks indicates regions where LOH point mutations and copy number are in disagreement. (a) LOH in the T10 hypodiploid subpopulation. (b) LOH in the T10 aneuploid subpopulations. (c) LOH in the T16P aneuploid subpopulation. (d) Summaries of nucleotide classes and LOH detected in each subpopulation. At a given position, "ab" means two nucleotides are seen each in at least 5 cells, and "a" or "b" mean a single nucleotide was observed, always in at least 5 single cells.

SUPPLEMENTARY METHODS

1.1 Samples

The frozen ductal carcinoma T10 (CHTN0173) was obtained from the Cooperative Human Tissue Network, and T16P and T16M were obtained from Asterand (Detroit,MI) Pathology shows that both tumors were poorly differentiated and high grade (III) as determined by the Bloom-Richardson score, and triple-negative (ER-, PR- and Her2/Neu-) as determined by immunohistochemistry. The cell lines used in this study include a normal male immortalized skin fibroblast (SKN1) and a breast cancer cell line (SK-BR-3). Normal breast tissue was obtained from Dr. Hanina Hibshoosh from Columbia University.

1.2 Single Nucleus Sequencing (SNS)

Nuclei were isolated from cell lines and from the frozen tumor using an NST-DAPI buffer (800 mL of NST [146 mM NaCl, 10 mM Tris base at pH 7.8, 1 mM $CaCl_2$, 21 mM $MgCl_2$, 0.05% BSA, 0.2% Nonidet P-40]), 200 mL of 106 mM $MgCl_2$, 10 mg of DAPI, and 0.1% DNase-free RNase A. The frozen tumor was first macro-dissected into 12 sectors of equal size using surgical scalpels and nuclei were isolated from six sectors for FACS by finely mincing a tumor sector in a Petri dish in 1.0–2.0 mL of NST-DAPI buffer using two no. 11 scalpels in a cross-hatching motion. The cell lines were lysed directly in a culture plate using the NST-DAPI buffer, after first removing the cell culture media. All nuclei suspensions were filtered through 37-μm plastic mesh prior to flow-sorting.

Single Nuclei were sorted by FACS using the BD Biosystems Aria II flow cytometer by gating cellular distributions with differences in their total genomic DNA content (or, ploidy) according to DAPI intensity. First a small amount of prepared nuclei from each tumor sample was mixed with a diploid control sample (derived from a lymphoblastoid cell line of a normal person) to accurately determine the diploid peak position within the tumor and establish FACS collection gates. Before sorting single nuclei, a few thousand cells were sorted to determine the DNA content distributions for gating. A 96-well plate was prepared with 10ul of lysis solution in each well from the Sigma-Aldrich GenomePlex© WGA4 kit. Single nuclei were deposited into individual wells in the 96-well plate along with several negative controls in which no nuclei were deposited.

Whole genome amplification was performed on single flow-sorted nuclei as described in the Sigma-Aldrich GenomePlex WGA4 kit (cat # WGA4-50RXN) protocol. WGA fragments from the frozen breast tumor and SK-BR-3 single cells were used directly for Single-read library construction using the Illumina Genomic DNA Sample Prep Kit (cat # FC-102-1001) and following standard protocol with a gel purification size range of 300-250bp. WGA fragments from the fibroblast cell line were first sonicated using the Diagenode Bioruptor© using the following program: 2 times, 7 minutes with 30 seconds high on/off mode in ice cold water. Sonication removes a specific 28bp adapter sequence that is added on during WGA, and improves the total number of sequencing reads per lane.

Single-read libraries from single nuclei were sequenced on individual flow-cell lanes using the Illumina GA2 analyzer for 76 cycles. Data was processed using the Illumina GAPipeline-1.3.2 to 1.6.0 Sequence reads were aligned to the human genome (HG18/NCBI36) using the Bowtie alignment software[44] with the following parameters: 'bowtie –S –t –m 1 –best –strata –p16' to report only top scoring unique mappings for each sequence read. To eliminate PCR duplicates, we removed sequences with identical start coordinates.

2.1 Read Depth Counting in Variable Bins

Copy number is calculated from read density, by dividing the genome into an 'bins' and counting the number of unique reads in each bin. In previous copy number studies read density was calculated using bins with uniform fixed length [16-19]. In contrast we use bins of variable length, that adjust size depending on the mappability of sequences to regions of the human genome. In regions of repetitive elements, lower numbers of reads are expected and thus the bin size is increased. To determine interval sizes we simulated sequence reads by sampling 200 million sequences of length 48 from the human reference genome (HG18/NCBI36) and introduced single nucleotide errors with a frequency encountered during Illumina sequencing. These sequences were mapped back to the human reference genome using Bowtie[15] with unique parameters as described above. We assigned a number of bins to each chromosome based on the proportion of simulated reads mapped. We then divided each chromosome into bins with an equal number of simulated reads. This resulted in 50009 genomic bins with no bins crossing chromosome boundaries. The median genomic length spanned by each bin is 54kb. For each cell the number of reads mapped to each variable length bin was counted. This variable binning efficiently reduces false deletion events when compared to uniform length fixed bins as shown in Supplementary Fig. 2b and 2c. For a single cell we typically measure 138 sequence reads per bin.

2.2 Integer Copy Number Quantification

Single cells will have integer copy number states that we can infer from sequence read counts, as follows. Unique sequence reads are counted in variable bins (Supplemental Fig. 4a) and segmented using the Kolmogorov-Smirnov (KS) statistic (Supplemental Fig. 4b). To estimate the integer differences of copy number states, we calculate Gaussian kernel smoothed density plots using Splus (MathSoft, Inc.), showing the difference between median bin counts for all pair-wise combinations of different segments (Supplemental Fig. 4c-e) The uniform steps between groups are very apparent, and is a general property of single cell data. We then convert our KS-segmented data into profiles of integer copy number as follows. We take the differential bin count of the second peak, denoted by an asterisk in Supplemental Fig. 4a, to represent a copy number "increment" of 1. We then divide every bin count in the profile by the increment and round to infer the integer copy number. We show in Supplemental Fig. 4f-g how closely the segmentation profile agrees with the integer copy number profile. However, for diploid or near diploid cells there are few to no steps from which to observe the increment, and we use a different method, taking the increment as the median bin count on the autosomes divided by two.

2.3 Gene Annotations

Amplifications and deletions identified in the single cell copy number profiles were annotated to identify UCSC genes. Cancer genes were identified using a compiled database from the cancer gene consensus and the NCI cancer gene index (Sophic Systems Alliance Inc., Biomax Informatics A.G).

3.1 Neighbor-joining Trees of Copy Number Profiles

Integer copy number profiles of single cells were used to calculate Neighbor-joining trees using a Euclidean distance metric with Matlab (Mathworks). Branches were flipped to orient nodes within subpopulations and trees were rooted using the last common diploid node.

3.2 Common Breakpoint Detection

Breakpoints are defined as bins with a copy number different than the previous bin in genome order. A transition from a lower copy number to a higher copy number (in genome order) is considered to be a different event than the opposite transition. To find breakpoint regions we

count each breakpoint in each cell and the immediately neighboring bins. A contiguous set of bins with counts greater than 1 is designated a breakpoint region. This results in a set of common breakpoint regions. Each cell is then scored for the occurrence of each of these events, a one meaning the cell has a copy number transition of that type (low to high or high to low) in that genomic region and a zero meaning no copy number transition of that type in that region.

3.3 Hierarchical Tree of Chromosome Breakpoints

We used chromosome breakpoints patterns to build a neighbor-joining tree. To eliminate breakpoints events with a high standard deviation, we limited our analysis to breakpoint regions covering no more than seven adjacent bins (N = 657). Using a Euclidean metric, we calculated a distance matrix from the binary chromosome breakpoint patterns identified in the single cells using Matlab (Mathworks). From this distance matrix we constructed a tree using average-linkage.

3.4 Heatmap of Chromosome Breakpoints

The heatmap is based on the same set of breakpoints used to build the neighbor-joining tree. Blue indicates the presence of an event, and white means no event. The columns are ordered as in the tree. The rows are ordered to show clearly which of the subsets of the four main groups in the tree share which events. The groups are ordered by subpopulation. A four dimensional binary vector represents each of the 16 possible subsets of these groups (subset vector). Each breakpoint is represented by a four dimensional vector of the percent of cells in each group having an event at that breakpoint (the "breakpoint vector"). The angle from each breakpoint vector to each subset vector is computed as well as the length of each projection vector. If the length of the projection vector is less than 0.05 the breakpoint vector is assigned to the empty (0,0,0,0) subset, otherwise it is assigned to the subset vector with the smallest angle to the breakpoint vector. The rows are ordered by subset vector in the following order: (1,1,1,1), (0,0,0,1), (0,0,1,0), (0,1,0,0), (1,0,0,0), (0,0,1,1), (0,1,0,1), (1,0,0,1), (0,1,1,0), (1,0,1,0), (1,1,0,0), (0,1,1,1), (1,0,1,1), (1,1,0,1), (1,1,1,0), (0,0,0,0). Within each subset the rows are in descending order by the number of cells in that subset having an event and then in ascending order by the number of cells not in that subset having an event.

4.1 Analysis of LOH Sequence Mutations in Tumor Subpopulations

PCR duplicates were removed from mapped sequence reads and bases with a quality score below 30 were excluded from analysis. We then determined the set of observed nucleotide types for each cell sequenced from the T10 and T16P and T16M tumors and every position in the genome. For each subpopulation we classified a position as the observed nucleotides only if one or two nucleotide types were each observed in five or more cells in the subpopulation. For each grouping of subpopulations DH, DA, if a classification was made in every subpopulation in the group, we translated the classifications into the generic nucleotides (a,b) based upon the order in which they were seen in the group, from left to right. We counted the resulting classifications of positions for each group by class, and determined whether long blocks of identical classifications along a chromosome were expected by chance. To establish the significance of our classification counts we repeated our analysis 100 times with randomly permuted cell labels within each group of subpopulations. We eliminated any effects from differing subpopulation size in a separate set of runs of the same analysis, each with 24 randomly selected cells in every subpopulation.