SUPPLEMENTARY INFORMATION APPENDIX

Joint single cell DNA-Seq and RNA-Seq of gastric cancer cell lines reveals rules of *in vitro* evolution

AUTHORS

Noemi Andor[1], Billy T. Lau[3], Claudia Catalanotti[4], Anuja Sathe[2], Matthew Kubit[2], Jiamin Chen[2], Cristina Blaj[5], Athena Cherry[6], Charles D. Bangs[6], Susan M. Grimes[3], Carlos Jose Suarez[6], Hanlee P. Ji[2,3]

INSTITUTIONS

[1]Integrated Mathematical Oncology, Moffitt Cancer Center, Tampa, 33612, FL, United States

[2]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Stanford, CA, United States

[3]Stanford Genome Technology Center, Stanford University, Palo Alto, CA, United States

[4]10X Genomics, Pleasanton CA, United States

[5]Department of Molecular and Cell Biology, University of California, Berkeley, United States

[6]Department of Pathology, Stanford University School of Medicine, Stanford, CA, United States

TO WHOM CORREPSONDENCE MAY BE ADDRESSED

Hanlee P. Ji

Email: genomics_ji@stanford.edu

Noemi Andor

Email: Noemi.Andor@moffitt.org

TABLE OF CONTENTS

SUPPLEMENTARY METHODS – single cell DNA-Seq

Generation of microfluidic chip, cell beads (CBs)

In the first microfluidic chip, CBs were generated by partitioning approximately 10,000 cells of each sample in a hydrogel matrix. A cell suspension is combined with an activation reagent, hydrogel precursors, paramagnetic particles, and loaded into one inlet well. In the other two inlet wells, CB polymer reagent and partitioning oil were added. To ensure a low multiplet rate, cells were delivered at a dilution such that the majority of CBs contain either a single cell or no cell. Once generated, the emulsion was immediately transferred into a PCR strip tube and incubated with orbital shaking at 1000 rpm overnight. The incubation yields polymerized magnetic CBs for subsequent steps.

Encapsulated cells were processed by the addition of lysis and protein digestion reagents to yield accessible DNA for whole-genome amplification. The presence of magnetic particles in the cell bead matrix enabled CB retention and streamlined washing and buffer exchange steps. After lysis, CBs were washed by magnetic capture, concentrated by reduction of liquid volume, and buffer exchanged with the addition of 1X PBS buffer. CBs were then denatured by NaOH, neutralized with Tris, and diluted in storage buffer. Finally, aggregates of cell beads were removed by filtration through a Flowmi strainer before a volume normalization procedure to set the CB concentration.

Generation of cell bead-gel beads (CBGBs)

CBGBs were generated by loading CBs, barcoded gel beads, enzymatic reaction mix, and partitioning oil in a second microfluidic chip. A majority of the **CBGBs** (~80%) contained a single CB and a single gel bead, which once encapsulated then dissolved to release their contents. To amplify and barcode gDNA, the emulsion was then incubated at 30°C for 3 hours, 16°C for 5 hours, and finally heat inactivated at 65°C for 10 minutes before a 4°C hold step. This two-step isothermal incubation yielded genomic DNA fragments tagged with an Illumina read 1 adapter followed by a partition-identifying 16bp barcode sequence. Conventional end-repair and a-tailing of the amplified library was performed, after which a single-end sequencing adapter containing the Illumina read 2 priming site was ligated.

ScDNA-Seq data processing and CNV calling

The computational pipeline includes **preprocessing** and **single cell copy number calling.** The outputs of this pipeline are CNV calls *and* read counts in 20kb bins across the genome as genomic bin-by-cell matrices. In the preprocessing stage, the first 16 base pairs of read 1 are compared to a whitelist of all possible droplet barcodes (totaling ~737,000). All observed droplet barcodes were tested for the presence of a cell by using mapped read abundances to the human genome. Reads were aligned to GRCh38 using bwa-mem version 0.7.12-r1039. Each read in the bam file was annotated with a cellular barcode tag ‘CB’. Confidently mapped reads were counted across the genome in 20kb non-overlapping windows. GC bias correction, modelled as a polynomial of degree 2 with fixed intercept, was applied.

Copy number calls are determined by modeling binned read abundances to a Poisson distribution with the copy number, GC bias, and a scaling factor as parameters. Candidate breakpoints were estimated by applying a log-likelihood ratio statistic against fluctuations in read coverage over neighboring genomic bins. These breakpoints were refined and reported as a set of non-overlapping segments across the genome. The copy numbers were scaled to integer-level ploidies. Copy number calls for non-mappable regions were imputed with neighboring copy number calls in confidently mapped regions, provided that the copy number on both sides of a non-mappable region were the same and the region was < 500 kb.

SUPPLEMENTARY METHODS – Cell cycle analysis

Assigning cell cycle state to scDNA-sequenced cells: For a given sample, we classified the genome of each sequenced cell $i \in I$ to one of three states (G0/G1, S, apoptotic) as follows. Under the assumption that the G0/G1 population is larger than any of the other populations, we defined the G0/G1 ploidy, p_{g0g1}, as the median ploidy across all sequenced cells of a given sample. Then, we calculate three features for each cell, *x*: i) its distance, d_x, to p_{g0g1}; ii) its total number of breakpoints, b_x, and iii) the Pearson correlation coefficient, r_x, between the number of rare breakpoints observed in the cell per each chromosome and the number of replication origins per chromosome. Rare breakpoints were defined as breakpoints that were shared among less than 1% of cells.

We distinguished G0/G1 cells from cells with higher genome fragmentation. We divided cells into two groups – $P_A := \{x \in I \mid p_x \geq p_{g0g1}\}$ and $P_B := \{x \in I \mid p_x < p_{g0g1}\}$ – containing cells above and below the sample's G0/G1 ploidy respectively. We fitted two sigmoid functions, one for each subgroup, to model cell ploidy as function of the number of breakpoints per cell: $p_i \sim \begin{cases} f_A(b_i), if\ i \in P_A \\ f_B(b_i), if\ i \in P_B \end{cases}$

We then calculated $B := argmin_b |f_A(b) - f_B(b)|$, as the threshold distinguishing G0/G1 cells from apoptotic cells:

$$apoptotic := \{x \in P_B \mid b_x \geq B\}$$

and from replicating cells:

$$S := \{x \in P_A \mid b_x \geq B\}$$

I.e. the yet unclassified cells were assigned to the G0/G1 state:

$$G0G1 := \{x \in I \mid b_x < B\} - \{S \cup apoptotic\}$$

The highest correlation to replication origins was observed for replicating cells (***SI Appendix, Fig. S2*B-J**), supporting the accuracy of above cell cycle phase assignment strategy. We removed cell cycle

specific breakpoints from further analysis, keeping only those breakpoints present among at least 1% of G0/G1 cells and encompassing segments of at least 5 Mb. Population-average copy number per segment per sample was calculated as the mean copy number across G0/G1 cells of that sample.

Inferring clonal dynamics from distribution of replicating cells among clones: For each detected and confirmed clone of a given sample, we calculated whether its % replicating cell assignment was different than expected by chance from its G0/G1 representation. Hereby we excluded clones below 4% size, because their absolute cell count was too small to reliably perform these calculations. Let N_{g0g1} and N_s be the total number of G0G1 and S cells detected in a cell line respectively. Further let $G0G1_i$ and S_i be the number of G0G1 and S cells assigned to clone *i* respectively. Then we calculated the % G0/G1 cells and the % replicating cells as $G0G1_i / N_{g0g1}$ and S_i / N_s respectively (X- and Y-axes in Fig. 2D). To infer positive selection, we used the hypergeometric distribution and calculated the p-value of sampling at least the observed number of replicating clone members, $\boldsymbol{S_i}$, as:

P = phyper(S_i, ***m***, $G0G1_i$, ***k***) where ***k*** is the total number of cells sampled from the clone and ***m*** is the expected number of clone members that are replicating (assuming proportionality to G0/G1 clone size). The p-value of sampling maximum S_i replicating cells was calculated by subtracting above value from 1, and was used to infer negative selection. P-values were adjusted for multiple hypotheses testing using the FDR method (R function "p.adjust").

SUPPLEMENTARY METHODS – single cell RNA-Seq

ScRNA-Seq library preparation and sequencing: We used the Chromium Controller instrument (10X Genomics Inc., Pleasanton, CA) and the Single Cell 3' Reagent kit (v2) to prepare individually barcoded single cell RNA-Seq libraries following the manufacturer's standard protocol. Briefly, single cell suspensions were loaded on a Chromium Controller instrument and were partitioned in droplets. Reverse transcription is performed, followed by droplet breaking, and cDNA amplification. Each cDNA molecule thus contained the read 1 sequencing primer, a 16bp cell-identifying barcode, and a 10bp UMI sequence[1]. We performed enzymatic fragmentation, end-repair, and a-tailing followed by ligation of a single-end adapter containing the read 2 priming site. PCR was performed using the Illumina P5 sequence and a sample barcode as described earlier. Libraries were purified with SPRIselect beads (Beckman Coulter, Brea, CA) and size-selected to ~450bp. Finally, sequencing libraries were quantified by qPCR before sequencing on the Illumina platform using 26x98 paired-end reads. The Cellranger software suite 1.2.1 was used to process scRNA data, sample demultiplexing, barcode processing, and single cell 3' gene counting. The cDNA insert, which is contained in the read 2, was aligned to the GRCh38 human reference genome. The reference GTF contained 33,694 entries, including 20,237 genes, 2,337

pseudogenes and 5,560 Antisense (non-coding DNA). Cellranger provided a gene-by-cell matrix, which contains the read count distribution of each gene for each cell. Cellranger's preprocessing pipeline is described in more detail at: https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/algorithms/overview.

ScRNA-Seq data preprocessing: We used a curated set of seven biological and technical features to detect and remove low-quality cells[1,2]. Biological features included: 1) transcriptome variance and expression of 2) cytoplasm localized genes, 3) mitochondrially localized genes, 4) mtDNA encoded genes. Technical features included: 5) % mapped reads, 6) % multi-mapped reads and 7) %non-exonic reads (intergenic & intronic). These features robustly identify low quality cells independently of cell type and of the experimental setting. The analysis was performed using Celloline version 0.9[2] and the R-package Cellity version 1.4[2]. For additional processing, we used the software suite Seurat (v2.3.2)[3]. Briefly, UMI counts were capped at the 99% quantile and only cells expressing at least 1,000 genes were included in subsequent analysis. Also, most cells classified by Cellity as low quality had a high percentage expressed mitochondrial genes as quantified by Seurat (data not shown).

Assigning cell cycle state to scRNA-sequenced cells: Leveraging prior knowledge in form of cell-cycle annotated genes and deploying a rank-based comparison across single cells, has been shown to robustly capture the transcriptional cell-cycle signature across different cell types and experimental protocols[4]. We employed such pathway-centric approach to classify the transcriptome of each sequenced cell to a cell cycle state as follows below.

Pathway quantification: The gene membership of 1,417 pathways was downloaded from the Reactome database[5] (v63). First the transcriptome profiles of high-quality cells detected within a given sample were scaled to the number of UMIs per cell (Seurat function "ScaleData"). We used the GSVA function[6] to model variation in pathway activity across cells of the sample (R function "gsva", mx.diff=TRUE). GSVA starts by evaluating the expression magnitude of a given gene in a given cell, in the context of the sample population distribution. To reduce gene specific biases (i.e. caused by GC content and gene length), an expression-level statistic was calculated for each gene from a kernel estimation of its cumulative density function. GSVA then calculated a rank-based, cell specific enrichment scores using the Kolmogorov-Smirnov like random walk statistic. For any given sample, pathways for which less than ten gene members were expressed in the scRNA-Seq data were not quantified.

Quantification of cell cycle pathways activity**:* The gene membership of 39 cell cycle pathways was downloaded from the Reactome database[5] (v63), whereby each pathway consisted of at least ten genes (SI Appendix,* Supplementary Table 4)**. We used the GSVA method[6] to model variation in pathway activity across cells of a given cell line, as described later.

Pathway and cell classification: Pathways were classified into three groups depending on their main activation timing during: i) G0/G1 (10 pathways, further referred to as P_{G0G1}); ii) S (5 pathways, further referred to as P_S) and iii) G2M (26 pathways, further referred to as P_{G2M}). Each class was normalized by its maximum activity across cells. As previously described, the 39 pathways were used as features to perform hierarchical clustering of cells (Euclidean distance metric and ward.D2 agglomeration method) into four clusters C:={C_1, C_2, C_3, C_4}. To classify each cluster $x \in C$ as either an G0/G1, S or G2M representative, we tested 39 null-hypotheses, one for each pathway p, namely that the activity of p in cells from x exceeds the activity of p in cells from {C – x}. We tested our hypotheses using the Wilcox rank-sum test and p-values were adjusted for multiple testing. For each pathway class $\delta \in \{G0G1, S, G2M\}$ we calculated the average effect size as:

$P(x|\delta) := \frac{1}{|P_\delta|}\sum_{p \in P_\delta} e_{p,x}$, where:

$$e_{p,x} = \begin{cases} effect\ size, if\ Wilcox\ p \leq 0.05 \\ 0, otherwise \end{cases}$$

Finally, we assigned cell cycle phase $argmax_{\delta \in \{G0G1,S,G2M\}} P(x|\delta)$ to each cluster $x \in C$.

SUPPLEMENTARY METHODS – LIAYSON: Calling CNVs from scRNA-Seq

To infer a cell's copy number state at any given locus, the LIAYSON algorithm uses a cell's read counts across the entire genome, thereby mitigating the influence of non-genetic factors on mRNA expression. In addition to raw UMI counts, LIAYSON requires as input the population-average (bulk) segmentation profile of the sample and the classification of cells into G0/G1, S and G2M subsets. It consists of two steps – aggregating expression across copy number segments, and calling copy number from segmental expression. These two steps are detailed as follows.

Calculating cell-by-segment expression matrix. Let $S := \{S_1, S_2, \ldots S_n\}$ be the set of *n* genomic segments that have been obtained from DNA-sequencing $i \in I$ cells of given sample (e.g. from bulk exome-sequencing, scDNA-sequencing, etc.). To prepare each cell's RNA-seq profile for copy number analysis

we first grouped genes by their segment membership, such that $\boldsymbol{E_{ij}}$ and $\boldsymbol{G_{ij}}$ are the average number of UMIs and the number of expressed genes per segment S_j per cell *i*.

To reduce data sparsity and the effect of non-genetic factors on gene expression we excluded genomic segments shorter than 10 Mb. For each cell cycle phase (**G0/G1, S, G2/M**), we also excluded genomic segments j for which $\overline{\boldsymbol{G_{*j}}}$ – the average number of expressed genes per cell – was below 20 (setting this threshold too high would also exclude single copy losses).

Calling copy numbers from G0/G1 cell-by-segment expression matrix. We first normalize the cell-by-segment expression matrix to gene coverage, by fitting a linear regression model for each j ∈ S:

$E_{*j} \sim Z_*$, where $Z_i \coloneqq \sum_{j \in S} G_{ij}$ – is the overall gene coverage of a given cell.

The model's residuals R_{ij} reflect inter-cell differences in expression per segment that cannot be explained by differential gene coverage per cell. A first approximation of the cell-by-segment copy number matrix C is then given by: $\mathbf{C_{ij}} := R_{ij} * (cn_j / \mu_j)$, where $\mu_j \coloneqq \frac{1}{|I|}\sum_{i \in I} R_{ij}$, is the mean residual per segment across cells and cn_j is the G0/G1 population-average copy number of segment j derived from DNA-seq. Above transformation of $\mathbf{E_{ij}}$ into $\mathbf{C_{ij}}$ is in essence a numerical optimization, shifting the distribution of each segment to the average value expected from bulk DNA sequencing.
Let x' ∈ C be the measured copy number of a given cell-segment pair, and x its corresponding true copy number state. The probability of assigning copy number x to a cell i at locus j depends on:

A. **Cell i's read count at locus j**, calculated conditional on the measurement x'.
   We fit a Gaussian kernel on the read counts at locus j across cells to identify the major (M) and the minor (m) copy number states of j as the highest and second highest peak of the fit respectively. Then we calculate the proportion of cells expected at state m as: $f \coloneqq \frac{cn_j - M}{m - M}$. The probability of assigning copy number x to a cell i at locus j is calculated as:

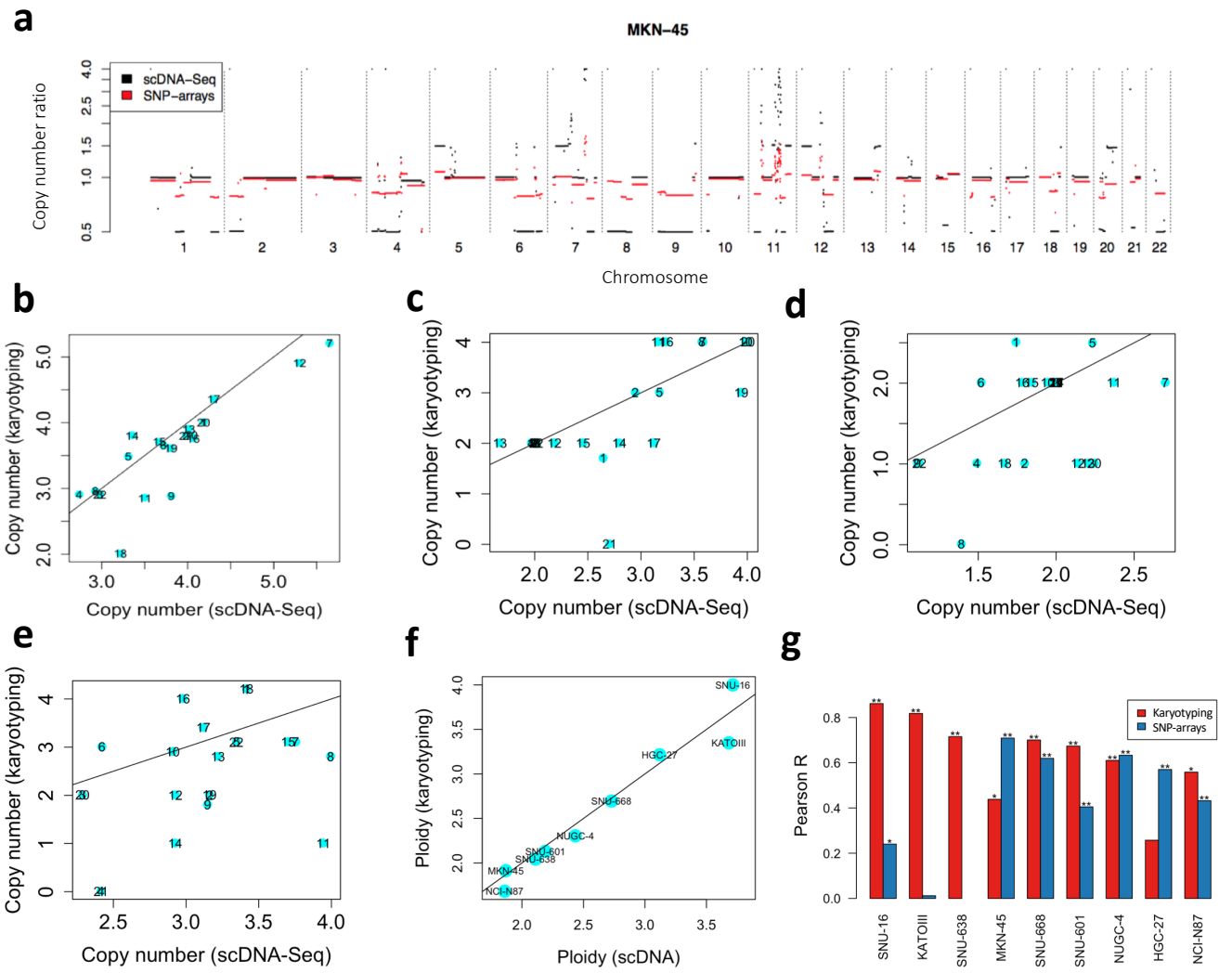
$$P_A(x|x') \coloneqq \begin{cases} 0, if\ x \notin \{m, M\} \\ P_{ij}(x'|N(m, sd = f)), if\ x == m \\ P_{ij}(x'|N(M, sd = 1 - f)), if\ x == M \end{cases}$$

B. **Cell i's read count at other loci**, i.e. how similar the cell is to other cells that have copy number x at locus j. We use Apriori – an algorithm for association rule mining – to find groups of loci that tend to have correlated copy number states across cells. Let $R^j_{K \to x}$ be the set of rules concluding copy number

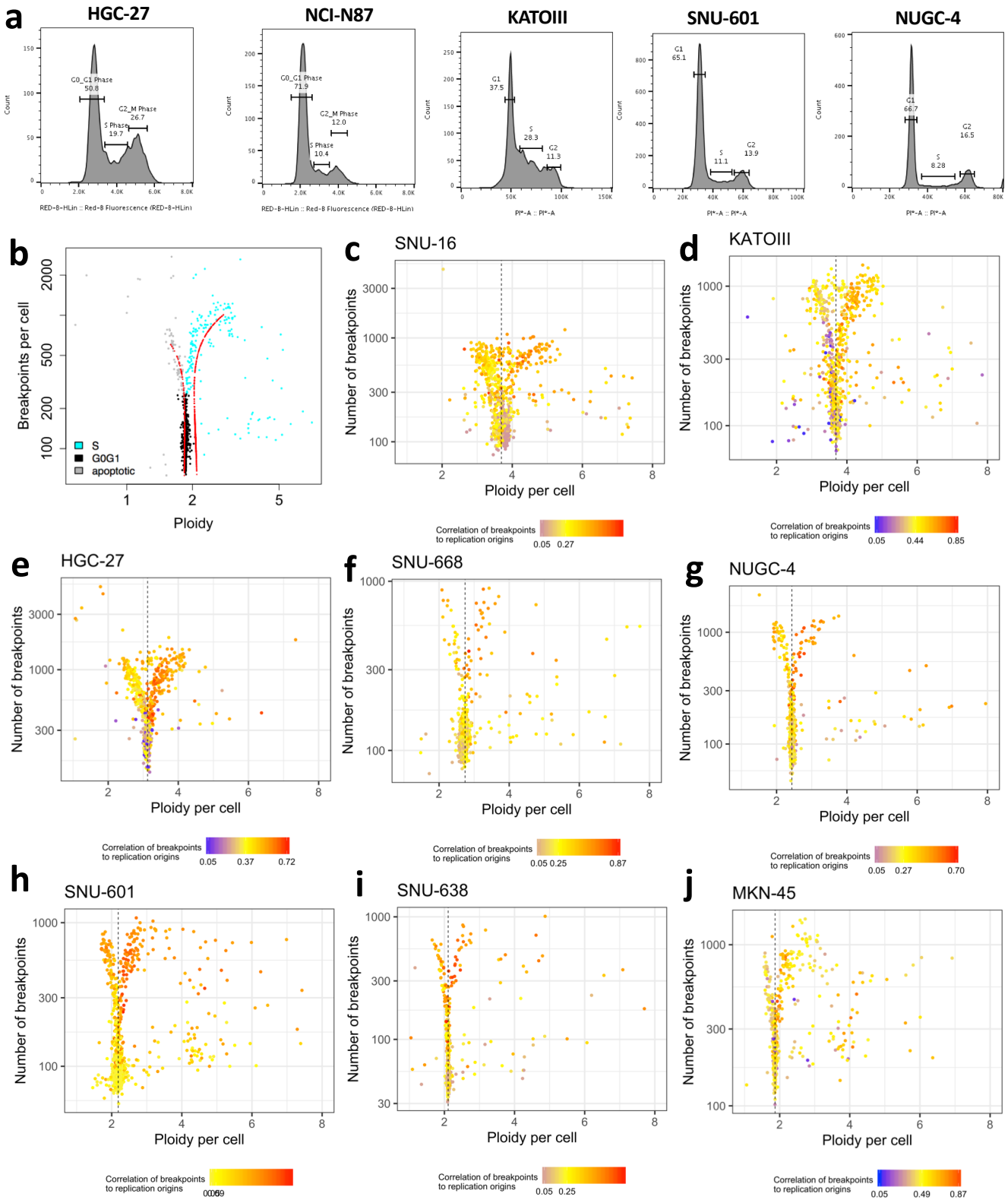
x for locus j, where k ∈ K are copy number profiles of up to n=4 loci in the form {S_1=x_1, S_2=x_2, … S_n=x_n}. For each cell $i \in I$ corresponding to any of the copy number profiles in K, we calculate:

$\mathrm{P_B}(\mathrm{x}) \sim \sum_{r \in \mathrm{R}^{j}_{K \to x}} C_r$, the cumulative confidence of the rules in support of x at j.

We first obtain a seed of cell-segment pairs by assigning a-priori copy number states only when $\mathrm{argmax}_{x \in [1,8]}\ \mathrm{P_A}(x|x'') > t$. We use this seed as input to B. Finally, a-posteriori copy number for segment j in cell i is calculated as: $argmax_{x \in [1,8]}\ P_A(x|x'') + P_B(x)$.

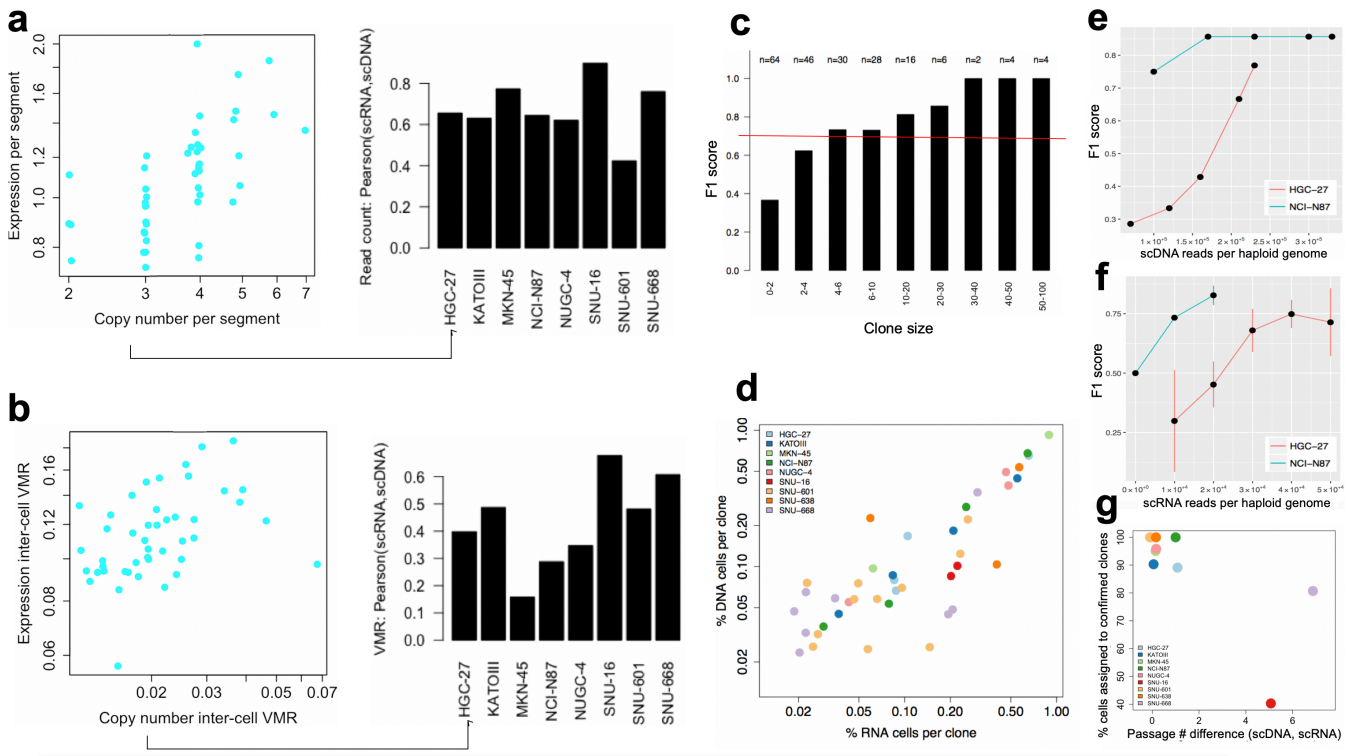


Supplementary Figure 1: Karyotyping and SNP-array analysis confirm scDNA-Seq derived aneuploidy. (**a**) ScDNA-Seq derived aneuploidy of MKN-45 (black) is confirmed by SNP-array data (red). (**b-e**) Correlation between average scDNA-Seq derived (x-axis) and karyotyping derived (y-axis) copy number per chromosome is shown for SNU-16 (**b**; Pearson r=0.86; P<1E-5), SNU-668 (**c**; Pearson r=0.7; P=3E-4), MKN-45 (**d**; Pearson r=0.44; P=0.04) and HGC-27 (**e**; Pearson r=0.26; P=0.32). (**f**) Average ploidy per cell line inferred by cellranger-dna from scDNA-Seq (x-axis) correlates to cell line's Karyotype (y-axis) (Pearson r=0.98; P<1E-5). (**g**) ScDNA-Seq derived copy number per segment correlates with karyotyping and with SNP-arrays, as is shown for MKN-45 and the other cell lines (**: P<=0.005; *: P<=0.05). In general, SNP-array data was obtained from older passages of the respective cell lines and therefore had weaker correlation to the scDNA-Seq data than did karyotyping. SNP-array data was not available for SNU-638.

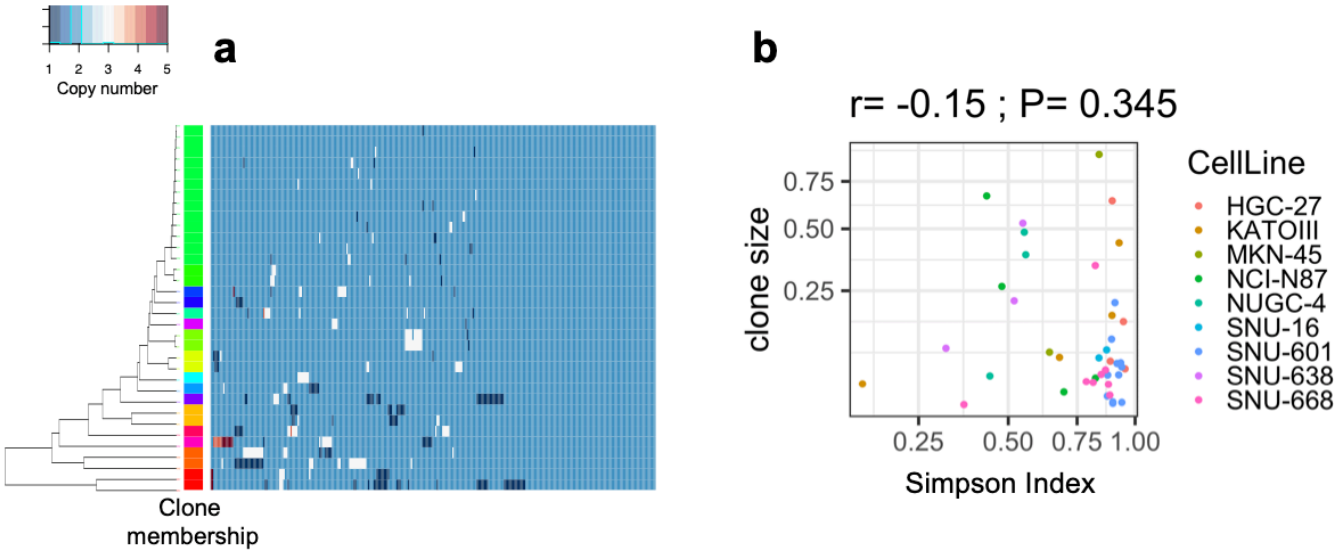


Supplementary Figure 2: ScDNA-Seq cell cycle phase assignment strategy and validation. (**a**) Cell cycle analysis of five gastric cancer cell lines by flow cytometry. (**b**) Demonstration of cell cycle phase assignment strategy with scDNA-Seq using NCI-N87 as an example. Two sigmoid functions (dotted red lines) are fitted to model cell ploidy (x-axis) as a function of the number of breakpoints per cell (y-axis), for cells above- (right) and below (left) median sample ploidy. Cell cycle phase assignment according to

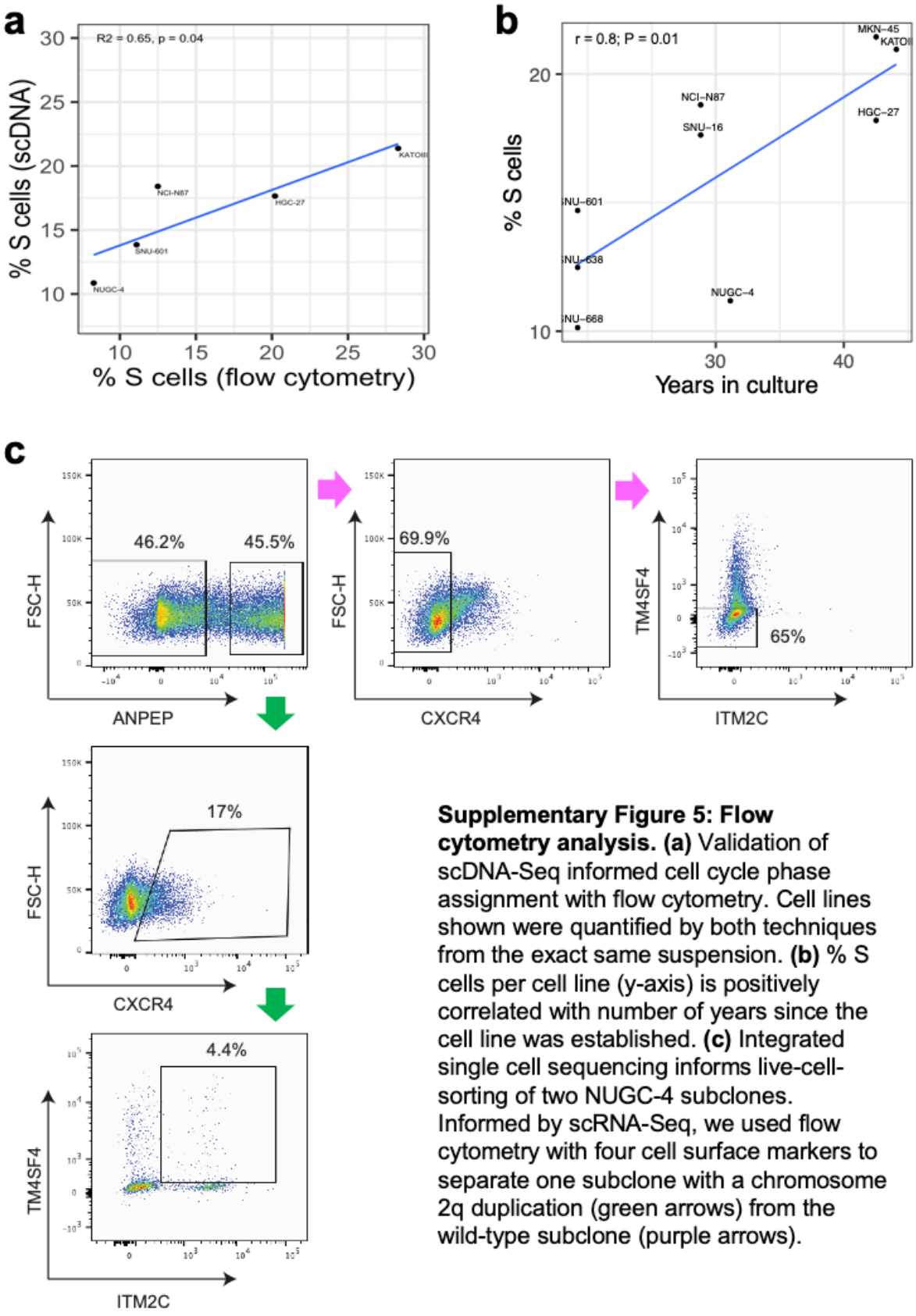
the fitted functions is color-coded. (**c-j**) Validation of assignment strategy using a third, independent feature: correlation to replication origins**.** The highest correlation was observed for S cells, whereas G0/G1 cells had the lowest values as would be expected from cells that are not actively replicating.



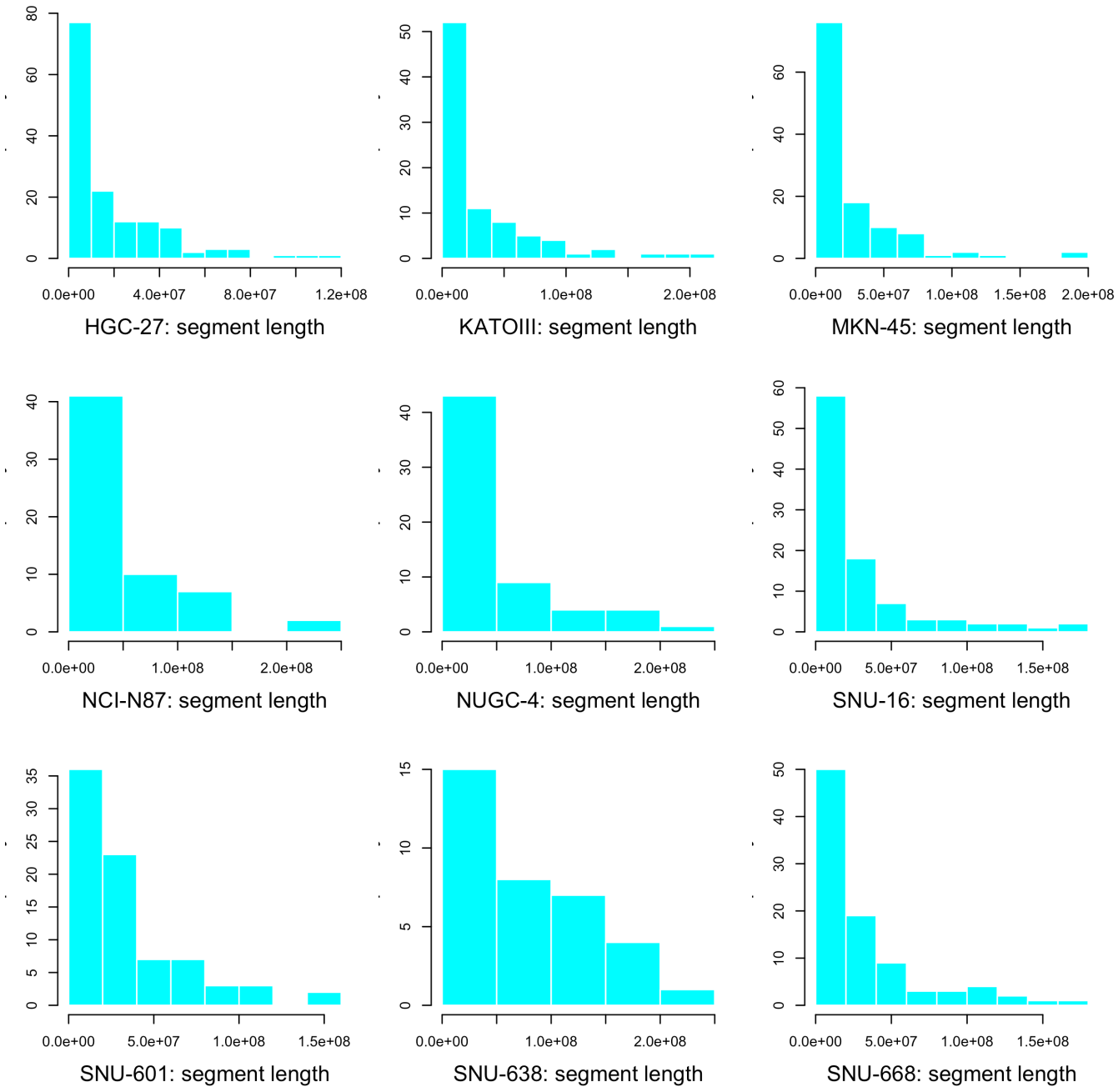
Supplementary Figure 3: ScDNA- and RNA-Seq mutual validation. (**a-b**) **Validation at meta-population resolution**. (**a**) Average expression per segment across KATOIII cells (left) reflects the cell population's average copy number (Pearson r=0.63; P=8.6E-6). Correlation coefficients between scDNA- and scRNA-Seq derived average read counts per segment are shown for KATOIII along with the other seven gastric cancer cell lines (right). (**b**) Variance-to-mean ratio (VMR) of expression per segment across KATOIII cells (left) correlates with the cells' variability in copy number states (Pearson r=0.49; P=8.0E-4). Correlation coefficients between scDNA- and scRNA-Seq derived VMR are shown for all gastric cancer cell lines (right). (**c-f**) **Validation at subpopulation resolution**. (**c**) Clone detection F1 score increases with increasing clone size across all nine cell lines. (**d**) The relative proportions of scDNA-Seq (y-axis) and scRNA-Seq (x-axis) cells per clone correlate within and across cell lines. (**e-f**) Dependence of clone detection F1 score on sequencing depth in NCI-N87 and HGC-27. F1 score increases with increasing number of reads sequenced per each cell's haploid genome. F1 score of scDNA-Seq (**e**) and scRNA-Seq (**f**) clone detection was calculated using the respectively other technique as control. For (**e,f**), between 20% and 95% of reads were sampled randomly for each of the two CLs and used to estimate performance at different sequencing depths. (**g**) Differences in passage number between scDNA- and scRNA-Seq experiments accompany differences in clonal composition observed between the two techniques for SNU-16 and SNU-668 (Pearson r=-0.71; P=0.032).



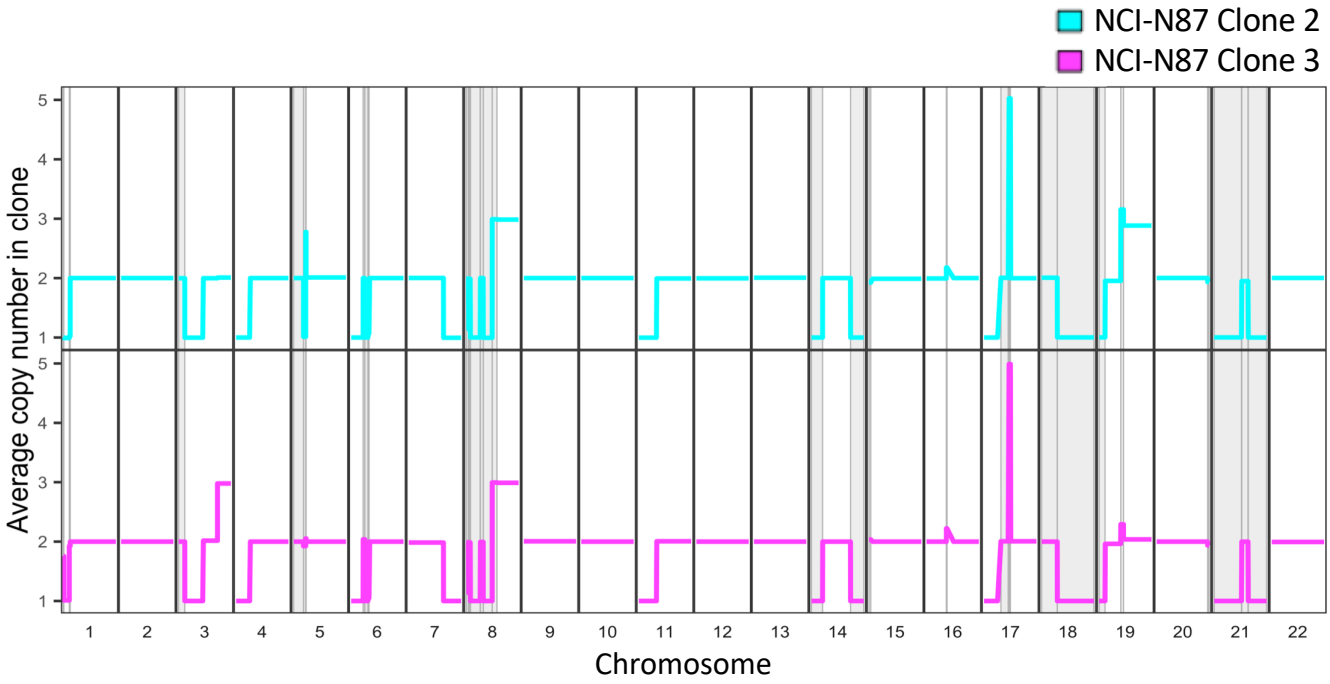
Supplementary Figure 4: Quantification of intra-clone diversity as surrogate of CNV accumulation rate. (**a**) As a surrogate of CNV accumulation rate per clone, we clustered cells according to rare CNVs and calculated the Simpson diversity index of cell-clusters found within a given clone. (**b**) Intra-clone diversity was not confounded by clone size.



Supplementary Figure 5: Flow cytometry analysis. (a) Validation of scDNA-Seq informed cell cycle phase assignment with flow cytometry. Cell lines shown were quantified by both techniques from the exact same suspension. **(b)** % S cells per cell line (y-axis) is positively correlated with number of years since the cell line was established. **(c)** Integrated single cell sequencing informs live-cell-sorting of two NUGC-4 subclones. Informed by scRNA-Seq, we used flow cytometry with four cell surface markers to separate one subclone with a chromosome 2q duplication (green arrows) from the wild-type subclone (purple arrows).



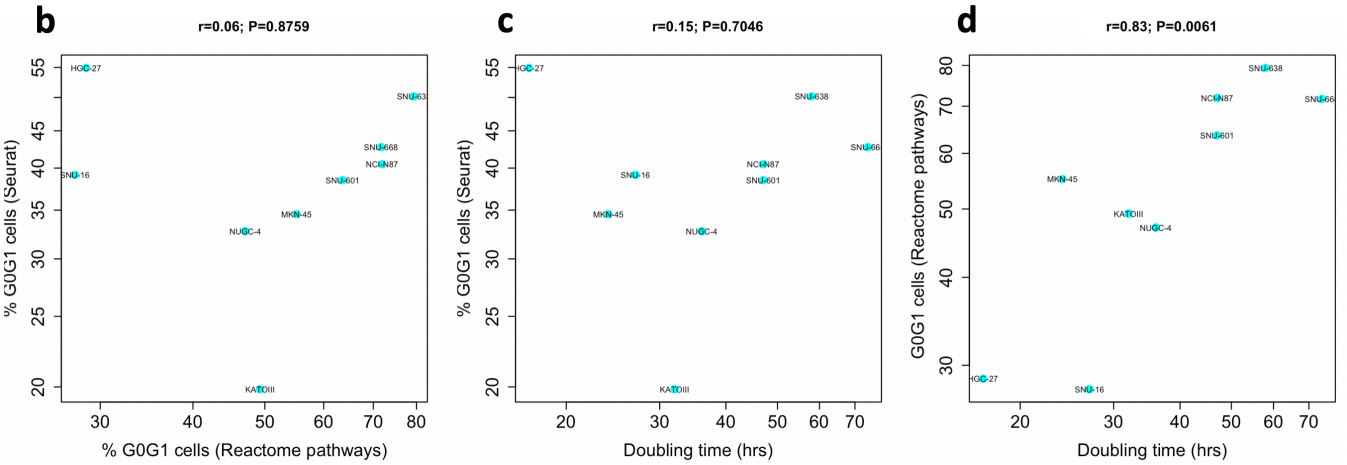
Supplementary Figure 6: Segment length distribution. scDNA-Seq derived segment length distribution is shown for the nine cell lines.



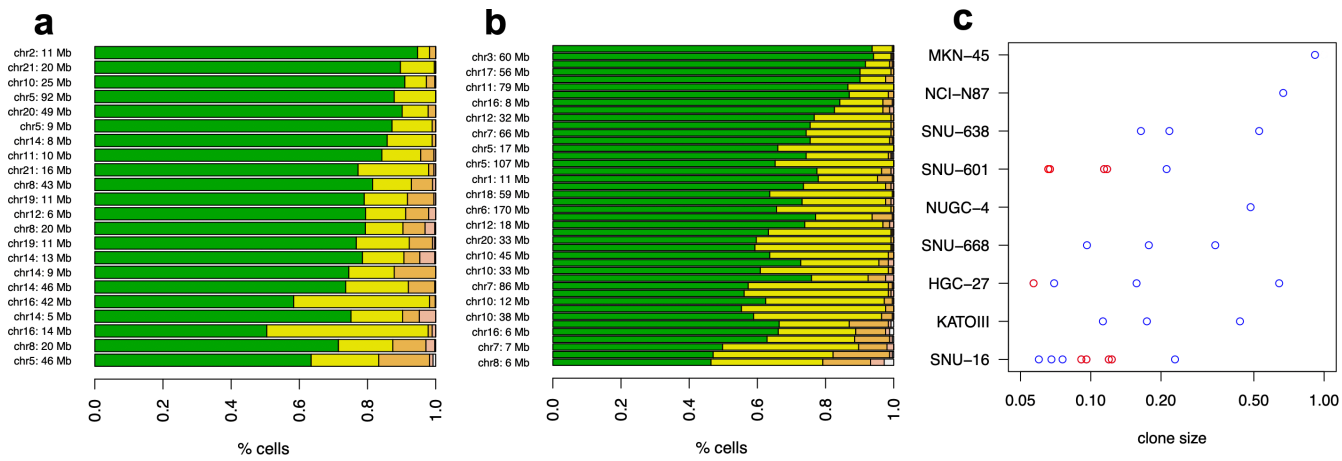
Supplementary Figure 7: Resolution of scDNA- and scRNA-Seq on clone-specific CNVs. Clone specific differences in CNVs are shown for the two largest NCI-N87 clones (pink and cyan). Highlighted as gray bands are the genomic regions too small to be assigned clone-specific CNVs by scRNA-Seq and thus detectable only with scDNA-Seq. See **Fig. 3 B-D** in main text for details on the two NCI-N87 clones.

a

	G2M.Reactome	S.Reactome	G0G1.Reactome
G2M.Seurat	0.133	0.110	0.086
S.Seurat	0.026	0.095	0.173
G0G1.Seurat	0.003	0.063	0.312



Supplementary Figure 8: Comparing pathway-centric and gene-centric cell cycle scoring. (**a**) Confusion matrix resulting from comparing Seurat's gene-centric assignment to that of our classification using pathways from the REACTOME database[7]. Confusion matrix was calculated across all nine cell lines. Seurat's `CellCycleScoring` function was used with a list of cell cycle genes from Tirosh *et al*.[8] to denote g2m and s markers. (**b**) The assignments were well correlated for most cell lines, except for SNU-16, HGC-27 and KATOIII. (**c,d**) Comparing both assignments to the doubling times of the cell lines, we observed that the same three cell lines show divergent proportions for Seurat's assignment, but not for the REACTOME-based assignment.



Supplementary Figure 9: Single-cell sequencing informs when assumptions of bulk sequencing deconvolution algorithms are not met. (**a-b**) For each genomic segment affected by subclonal CNVs (y-axis) colors distinguish the five most common copy number states measured for that segment (x-axis) in SNU-668 (a) and SNU-16 (b). Segments are sorted according to the Simpson-entropy index of their copy number states. (**c**) Subclones at or above 5% cellular frequency are displayed for each of the nine cell lines. Size-adjacent clone pairs whose cell frequency ratio is below 1.1 are highlighted red.

Supplementary Table 1: ScDNA-Seq metrics.

	SNU-16	KATOIII	HGC-27	SNU-668	NUGC-4	SNU-601	SNU-638	MKN-45	NCI-N87
Sequenced cells	825	973	912	1,238	829	1,531	724	787	1,005
Mean mapped, deduplicated reads per cell	706,285	945,622	998,063	463,375	2,164,914	565,648	507,254	720,892	885,548
Median effective reads per 1Mbp	227	313	327	144	653	180	154.5	220	290
Median cnv resolution (mb)	1.42	1.16	1.14	1.84	0.59	1.55	1.77	1.41	1.11
Ploidy (scDNA-Seq)	3.71	3.68	3.12	2.73	2.43	2.19	2.11	1.87	1.86
Breakpoint count (all cells)	2,037	2,545	5,497	1,621	1,477	1,198	425	3,677	1,306
Breakpoint count (G0G1 cells)	96	86	144	92	61	81	35	118	60
G0G1	58%	62%	63%	82%	79%	77%	74%	68%	74%
S	17%	21%	18%	10%	11%	14%	12%	22%	18%
Apoptotic	25%	17%	19%	8%	10%	9%	14%	10%	8%

Supplementary Table 2: Passage number, confluence and general information about nine gastric cancer cell lines. Confluence was typically similar between scDNA- and scRNA-Seq experiments (80-90%), but diverged considerably for NUGC-4, explaining the discrepancy in the % cycling cells between the two techniques for this cell line (**Fig. 1C**).

		SNU-16	KATOIII❖	HGC-27	SNU-668	NUGC-4	SNU-601❖	SNU-638❖	MKN-45	NCI-N87
Passage # | % confluence	> Karyotyping	P8	P2	P7(o)	P2	P24(o)	P2	P2	P24(o)	P2
	> ScDNA-Seq	P7	P4 | 65	P7(o) | NA	P9 | 80-90	P23(o) | 20-30	P3 | 80	P1 | 50-60	P23(o) | NA	P3 | NA
	> ScRNA-Seq	P2	P4 | 65	P7(o) | 80-90	P2 | 80-90	P23(o) | 80-90	P3 | 80	P1 | 50-60	P23(o) | 80-90	P2 | 80-90
Doubling time (h)		27	32	17	74	36	47	58	24	47
Ploidy (Karyotyping)		3.76	3.71	3.29	2.83	2.60	2.29	2.25	1.95	1.94
Age		33	55	NA	68	35	34	48	62	NA
Sex		Female	Male	Unspecified	Male	Female	Male	Male	Female	Male
Tumor Differentiation		Carcinoma	Carcinoma	Carcinoma	Carcinoma	Adenocarcinoma	Carcinoma	Carcinoma	Adenocarcinoma	Carcinoma
MSI status		MSS	MSS	MSS	MSS	MSS	MSS	MSI	MSS	MSS
Growth type		suspension	adherent + suspension	adherent	adherent	suspension + adherent	adherent	adherent	adherent + suspension	adherent
Year of 1st report		1990	1974	1976	1997	1988	1997	1997	1976	1990

(o) passage number includes that from vendor

❖ scDNA & scRNA sequenced from the same suspension

Supplementary Table 3: ScRNA-Seq statistics. (*: Low quality cells excluded).

	SNU-16	KATOIII	HGC-27	SNU-668	NUGC-4	SNU-601	SNU-638	MKN-45	NCI-N87
Replicates	2	1	2	1	1	1	1	1	1
Sequenced cells*	2,088	3,084	1,808	5,319	3,797	5,186	867	2,814	3,246
Mean reads per cell	138,104	76,707	150,155	51,949	53,046	84,420	71,714	54,031	40,555
Mean genes per cell	5,683	4,851	5,584	3,920	3,686	4,464	3,841	3,611	3,299
Median genes per cell	5,661	4,781	5,545	3,821	3,590	4,355	3,904	3,538	3,135
G0G1	28%	49%	29%	72%	47%	64%	79%	55%	72%
G2M	37%	18%	21%	9%	25%	13%	9%	12%	11%
S	35%	33%	51%	19%	28%	24%	11%	33%	17%
Cycling	72%	51%	71%	28%	53%	36%	21%	45%	28%

Supplementary Table 4: Cell cycle pathways. A total of 39 cell cycle pathways from the REACTOME database are listed along with their activation timing during S, G2M or G0/G1 phases of the cell cycle.

	Pathway	Class
1	Establishment of Sister Chromatid Cohesion	S
2	Cyclin A:Cdk2-associated events at S phase entry	S
3	S Phase	S
4	Synthesis of DNA	S
5	Ubiquitin-dependent degradation of Cyclin D	S
6	Condensation of Prometaphase Chromosomes	G2M
7	Condensation of Prophase Chromosomes	G2M
8	Conversion from APC/C:Cdc20 to APC/C:Cdh1 in late anaphase	G2M
9	Cyclin A/B1 associated events during G2/M transition	G2M
10	FBXL7 down-regulates AURKA during mitotic entry and in early mitosis	G2M
11	G2/M Checkpoints	G2M
12	G2/M DNA damage checkpoint	G2M
13	G2/M Transition	G2M
14	Inhibition of the proteolytic activity of APC/C required for the onset of anaphase by mitotic spindle	G2M
15	Loss of Nlp from mitotic centrosomes	G2M
16	M Phase	G2M
17	MASTL Facilitates Mitotic Progression	G2M
18	Mitotic Anaphase	G2M
19	Mitotic G2-G2/M phases	G2M
20	Mitotic Metaphase and Anaphase	G2M
21	Mitotic Prometaphase	G2M
22	Mitotic Prophase	G2M
23	Mitotic Spindle Checkpoint	G2M
24	Recruitment of NuMA to mitotic centrosomes	G2M
25	Recruitment of mitotic centrosome proteins and complexes	G2M
26	Regulation of PLK1 Activity at G2/M Transition	G2M
27	Regulation of mitotic cell cycle	G2M
28	TP53 Regulates Transcription of Genes Involved in G2 Cell Cycle Arrest	G2M
29	The role of GTSE1 in G2/M progression after G2 checkpoint	G2M
30	Cyclin D associated events in G1	G0G1
31	G0 and Early G1	G0G1
32	G1 Phase	G0G1
33	G1/S DNA Damage Checkpoints	G0G1
34	G1/S Transition	G0G1
35	G1/S-Specific Transcription	G0G1
36	TP53 Regulates Transcription of Genes Involved in G1 Cell Cycle Arrest	G0G1
37	p53-Dependent G1 DNA Damage Response	G0G1
38	p53-Dependent G1/S DNA damage checkpoint	G0G1
39	p53-Independent G1/S DNA damage checkpoint	G0G1

Supplementary Table 5: Mutual validation of scDNA- and scRNA-Seq derived clone identification.

	SNU-16	KATOIII	HGC-27	SNU-668	NUGC-4	SNU-601	SNU-638	MKN-45	NCI-N87
Passage # difference (scDNA, scRNA)	5	0	1	7	0	0	0	0	1
Clone # (scDNA)	11	5	5	10	4	12	4	2	4
Clone # (scRNA)	7	5	5	10	4	11	3	3	4
Confirmed clone membership (%)	40.3	90.3	89.1	80.7	95.8	100.0	100.0	95.0	100.0
TP #	2	4	4	8	3	11	3	2	4
FP #	5	1	1	2	1	0	0	1	0
FN #	6	1	1	2	1	1	1	0	0

TP=true positive; FP=false positive; FN=false negative;

scDNA=single-cell DNA sequencing; scRNA=single-cell RNA sequencing.

Supplementary Table 6: Multiple regression model of the clone count per cell line as a function of ploidy and time in culture. Coefficients were calculated by fitting a least square linear regression on the nine gastric cancer cell lines. Number of clones per cell line increases with ploidy and decreases with the number of years since a cell line was first established.

	Coefficient	Std. Error	t value	Pr(>|t|)
Intercept	7.112	3.814	1.865	0.111
Ploidy	3.098	1.294	2.394	0.054
Years in culture	-0.290	0.098	-2.956	0.025

Multiple R-squared: 0.6461; Adjusted R-squared: 0.5282
F-statistic: 5.478 on 2 and 6 DF, p-value: 0.04431

Supplementary Table 7: We used the experimentally measured clone frequencies to calculate a-posteriori saturation curves of scDNA-Seq library sizes for each cell line as previously described (Ruli Gao et al., Nature Genetics 2016). Minimum number of cells required to keep the risk of observing fewer than five cells per clone below 0.01, was calculated as previously described[9], based on a multinomial distribution (3rd column). The number of G0/G1 cells actually sequenced was greater than that minimum for all cell lines, suggesting that all cell lines were sequenced at or above saturation.

Cell line	Number of clones	Minimum cells	Sequenced G0/G1 cells
SNU-16	11	279	477
KATOIII	5	274	604
HGC-27	5	233	576
SNU-668	10	759	1009
NUGC-4	4	264	656
SNU-601	12	698	1186
SNU-638	4	126	537
NCI-N87	4	451	742
MKN-45	2	149	533

Supplementary Table 8: Clone-specific cell surface marker expression. For each of the 41 confirmed clones (rows), combinations of up to four cell surface markers are listed along with a quantitative estimate of their ability to separate a clone of interest from the remaining clones in a cell line. Cells were clustered based on the expression of corresponding cell surface markers. Across all G0G1 cells we then calculated the Pearson correlation coefficient between clone membership and cluster membership. Horizontal line marks 16 clones for which sorting efficacy was estimated at least as high as for the two isolated NUGC-4 clones presented in main **Fig. 4**.

	cell line	profile	p-value	pearson
1	NCI-N87	CST6 (CST6)-, PSMD2 (PSMD2)-, TM4SF1 (TM4SF1)-	0	0.509
2	NCI-N87	ABCC3 (ABCC3)-, AP2M1 (AP2M1)+, CST6 (CST6)+, TM4S	0	0.524
3	SNU-601	ARPC1A (ARPC1A)-, FLNA (FLNA)+, SEC61G (SEC61G)-	0	0.295
4	SNU-668	COBL (COBL)+, DSP (DSP)+	0	0.297
5	NCI-N87	HLA-C (HLA-C)+	6.91976E-37	0.274
6	SNU-601	PROCR (PROCR)+, RARRES3 (RARRES3)+, SIPA1 (SIPA1)+	2.29189E-35	0.240
7	SNU-601	ARPC1A (ARPC1A)+, CEACAM5 (CEACAM5)+, S100A14 (S10	2.59009E-33	0.233
8	SNU-668	DAB2 (DAB2)+, IL7R (IL7R)+	3.01065E-30	0.220
9	SNU-601	SLC2A3 (SLC2A3)+	6.3601E-30	0.221
10	NUGC-4	ANPEP+, CXCR4-, TM4SF4-, ITM2C-	2.36074E-29	0.273
11	NUGC-4	ANPEP (ANPEP)-, AQP5 (AQP5)+	7.35314E-28	0.266
12	NCI-N87	PCDH7 (PCDH7)+	6.54338E-27	0.233
13	MKN-45	NRP2 (NRP2)-, RAB3B (RAB3B)-	1.93044E-25	0.278
14	MKN-45	NRP2 (NRP2)+, RAB3B (RAB3B)+	1.93044E-25	0.278
15	SNU-601	CST6 (CST6)+, PAFAH1B2 (PAFAH1B2)+	1.67825E-22	0.190
16	NUGC-4	ANPEP-, CXCR4+, TM4SF4+, ITM2C+	2.79415E-13	0.179
17	SNU-601	KDSR (KDSR)+, YTHDC1 (YTHDC1)-	1.50904E-17	0.166
18	SNU-668	DYSF (DYSF)+, LAG3 (LAG3)+, NOTCH4 (NOTCH4)+	6.35257E-17	0.162
19	SNU-668	HERPUD1 (HERPUD1)+, SLC9A3R2 (SLC9A3R2)+	3.04908E-16	0.158
20	SNU-601	TMBIM6 (TMBIM6)-	3.80253E-14	0.148
21	SNU-601	DNAJC15 (DNAJC15)+	7.63287E-14	0.146
22	SNU-668	EPB41L4A (EPB41L4A)+, STAMBP (STAMBP)-	8.88155E-12	0.132
23	SNU-668	ABCG2 (ABCG2)+	1.46816E-11	0.131
24	SNU-638	DNAJC10 (DNAJC10)+, HMGCS1 (HMGCS1)-, LCP1 (LCP1)+	6.57171E-11	0.256
25	SNU-638	BET1 (BET1)-, HMGCS1 (HMGCS1)+, PAIP1 (PAIP1)+	9.75065E-11	0.253
26	HGC-27	EEPD1 (EEPD1)+, FOLH1 (FOLH1)+, PIK3IP1 (PIK3IP1)+	2.26357E-10	0.301
27	KATOIII	MARCH8 (MARCH8)+, PROM2 (PROM2)+	4.4999E-10	0.186
28	KATOIII	FXYD3 (FXYD3)+	8.91705E-09	0.171
29	KATOIII	KCNH2 (KCNH2)+, PDE4A (PDE4A)+	1.74299E-08	0.168
30	SNU-668	ELOVL7 (ELOVL7)+	2.46115E-08	0.108
31	SNU-601	DNAJC6 (DNAJC6)+	3.47837E-07	0.100
32	SNU-601	PEBP1 (PEBP1)-	6.81414E-07	0.097
33	KATOIII	MUC13 (MUC13)-	9.6855E-07	0.146
34	HGC-27	MKRN3 (MKRN3)+, ZNF10 (ZNF10)+	1.13173E-06	0.233
35	HGC-27	BSCL2 (BSCL2)+	1.16961E-06	0.233
36	HGC-27	ODC1 (ODC1)-	1.92832E-06	0.228
37	SNU-601	TRPV6 (TRPV6)+	2.41592E-06	0.092
38	SNU-668	VAMP5 (VAMP5)+	3.38643E-06	0.090
39	SNU-638	SPAG4 (SPAG4)+	2.0822E-05	0.168
40	SNU-16	LRP5 (LRP5)+, PTPRU (PTPRU)-	9.19355E-05	0.328
41	SNU-16	LRP5 (LRP5)-, PTPRU (PTPRU)+	9.19355E-05	0.328

REFERENCES

1. Zheng, G.X.*, et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, 14049 (2017).
2. Ilicic, T.*, et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome Biol* **17**, 29 (2016).
3. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411-420 (2018).
4. Scialdone, A.*, et al.* Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54-61 (2015).
5. Fabregat, A.*, et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46**, D649-d655 (2018).
6. Hanzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
7. Croft, D.*, et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res* **42**, D472-477 (2014).
8. Tirosh, I.*, et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).
9. Gao, R.*, et al.* Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat Genet* **48**, 1119-1130 (2016).