nature research

https://doi.org/10.1038/s41588-020-00726-6

Supplementary information

Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity

In the format provided by the authors and unedited

Supplementary Note

Sequencing

Final 3' scRNA-Seq libraries were diluted to 4 nM, denatured, and further diluted to a final concentration of 2.8 pM for sequencing with the following sequencing parameters: Read 1: 26 cycles, i7 index: 8 cycles, i5 index: 0 cycles, Read 2: 58 cycles. Pooled bulk MARS-Seq libraries were diluted to 4 nM, denatured, further diluted to 2 pM and sequenced with the following parameters: Read 1: 75 bp, Read 2: 15 bp, no indices. Sequencing was performed with the NextSeq500 (Illumina) using the NextSeq 75bp High Output Kit (Illumina) at the Weizmann Institute Life Science Core Facility.

Cell cycle analysis by propidium iodide (PI) DNA staining

The EpiSen-high (AXL^-CLDN4^+) and EpiSen-low (AXL^+CLDN4^-) subpopulations were isolated from the JHU006 and UM-SCC47 cell lines as described above. Cells were sorted into ice cold PBS and fixed by adding the cell suspension dropwise to 70% ethanol while vortexing. Fixed cells were stored at 4°C. Following 2x PBS washes, cells were resuspended and incubated in PI/Triton-X-100 staining solution consisting of 0.1% Triton-X-100 (Sigma), 0.2mg/ml DNAse-free RNAse A (Sigma), and 0.04 mg/ml of 500ug/ml PI (Sigma) in PBS at 20°C for 30 min. Singlets were gated by FSC-W vs. FSC-A (PI) and cell cycle analysis was performed based on the PI signal histogram. The experiment was performed two times independently.

Senescence induction by etoposide and SA-β-gal staining

Primary bronchial cells were seeded at 50,000 cells/well in 24 well plates and treated with 5-7.5 μM etoposide (Sigma) 24 hours later to induce senescence. After 48 hours, media was replaced and on day 9 etoposide-treated cells and untreated controls were fixed with 0.5% glutaraldehyde solution in PBS pH 7.4 and incubated with X-gal staining solution (0.2M $K_3Fe(CN)_6$, 0.2M $K_4Fe(CN)_6$ $3H_2O$, and 40X X-Gal stock diluted in PBS/$MgCl_2$) for 6 hours protected from light. X-Gal stock consists of 40 mg/ml X Gal (Roche #745740) in N,N-dimethylformamide (Sigma D-4254). Following PBS washes, stained cells were covered with 80% glycerol prior to imaging.

Bulk RNA-seq by SMART-Seq2

The SMART-Seq2 protocol[7] was adapted with several modifications, as described below, to generate libraries for bulk RNA-Seq of individual cell lines included in the HNSCC custom pool (see 'Cell line assignment') and for profiling of bronchial primary cells following senescence induction with etoposide[8] (**Fig. 5B**). Between 100-200 cells were resuspended in lysis buffer containing Triton 0.2% and RNAase inhibitor, or pelleted, washed with PBS and resuspended in the lysis buffer. For library generation, the plate was incubated at 72°C for 3 min and reverse transcription was performed as described in the SMART-Seq2 protocol[7] followed by cDNA pre-amplification for 17 cycles. Following 1X Agencourt Ampure XP beads cleanup (Beckman Coulter), 200 pg of amplified DNA underwent tagmentation and final amplification adding unique Illumina barcodes for 12 cycles (Nextera XT Library Prep

kit, Illumina). Libraries were quantified by Qubit and TapeStation (Agilent) and pooled prior to sequencing. Sequencing was performed using the Illumina Nextseq 75 cycle high output kit (single read, dual index). Bulk Smart-Seq2 libraries were diluted to 1.6 pm and sequenced with parameters: Read 1: 76bp, i7: 8 cycles, i5:8 cycles. Sequencing was performed with the NextSeq500 (Illumina) using the NextSeq 75bp High Output Kit (Illumina) at the Weizmann Institute Life Science Core Facility.

Gap closure migration assay

Following sorting by AXL and ITGA6 to isolate the EMT-II-high and EMT-II-low subpopulations as described above, 75,000 cells of each population (EMT-II-high cells, EMT-II-low cells, unsorted cells, and unsorted cells treated with 10 ng/µl TGF-β3 (Peprotech)) were resuspended in 70 µL RPMI and loaded per well containing a wound healing assay culture insert (Ibidi). After 24 hour incubation, the inserts were removed to generate the gap and cells were imaged at 0, 6, 12, 24, and 48 hours. The experiment was performed independently three times.

Cell line assignment

We used both expression-based and SNP-based methods to assign cells to cell lines. In each of these methods, we compared the single cells to external bulk data of the corresponding cell lines and then either assigned the cells to the most similar cell line or excluded them as potential doublets or low-quality cells. Only cells whose assignments were also consistent between both methods were retained for further analysis. Bulk RNA-seq data was obtained from the DepMap portal (https://depmap.org/; 18q3 data release)[1] for cell lines in the eight CCLE-based pools and was generated as described above for the eight HNSCC cell lines in the custom pool.

For SNP based classification, for each cell we determined the cell line from the pool whose SNP profile (based on bulk RNA-seq data) best matched the observed reference and alternate allele reads across a panel of SNP sites, similar to the Demuxlet method[2]. Specifically, we used a logistic regression model for each cell, where the probability of a read at SNP site i being from the alternate allele is modeled as $P_i = \sigma(\beta_0+\beta_j*X_{i,j})$, where σ is the logistic function, $X_{i,j}$ is the allelic fraction of cell line j at site i (estimated from bulk RNA-seq data), and β_j are parameters estimated for each single cell and reference cell line by maximizing the data likelihood under a binomial model. Models were fit using the R package glmnet [3], and the cell line whose SNP profile produced the highest likelihood under this model was selected. Goodness-of-fit was quantified by the model deviance relative to the null-model deviance. We used a reference panel of 100,000 SNPs that were frequently detected across a panel of 200 cancer cell lines (based on bulk RNA-seq data), and that were detected in the scRNA-seq data from the same cell lines.

For expression-based classification, we used a broadly similar approach. First, we subsetted the gene expression data to genes that were expressed in at least half of all cells or had a maximal expression (measured by the 98th percentile of that gene's expression across all cells) greater than 3 $\log_2$(CPM). Next, we used the Rtsne R package to estimate, for each pool, 3 t-

SNE embedding dimensions for each cell, computed based on 50 principle components, using a perplexity parameter of 30, and a theta value of 0.2 (using the first three principle components for initialization). We applied a local 'smoothing' to the normalized and centered single-cell expression profiles (*ER*), using a Gaussian kernel applied to the cell-cell distances in the t-SNE embedding space. The Gaussian kernel bandwidth was set using the method 'sigest' from the R package kernlab. Finally, for each cell, we identified the reference cell line from the pool with the most similar bulk RNA-seq gene expression profile (*ER*, and Pearson correlation similarity).

Detection of putative 'doublets', where two or more cells are labeled with the same barcode during droplet-based library preparation, was performed based on the SNP data, using the same generalized linear modeling approach to identify a mixture of two reference cell lines whose combined SNP profiles best explained the SNP data from a given cell. To efficiently estimate the best-fitting reference cell line pair we used a Lasso-regularized generalized linear model. After determining the best-fitting 'singlet' and 'doublet' models for each putative cell, we then determined whether each putative cell was a singlet, doublet, or a 'low-quality' cell based on several statistics. To identify low quality cells we took the maximum of the deviance explained by the singlet model and the deviance explained by the doublet model. We observed that the maximum deviances formed a bimodal distribution. We thus used the local minimum between the two distributions as a threshold and classified all cells with a maximum deviance below this threshold as 'low quality'. To separate putative doublets from singlets, we then fit a two-component Gaussian mixture model using three variables: (1) the amount of deviance explained by the singlet model; (2) the (log-transformed) deviance-improvement of the doublet model over the singlet model; and (3) the fraction of genes detected in that cell. Cells with a probability greater than 0.75 of belonging to the cluster with higher average doublet improvement were classified as doublets.

Analysis of co-culture (pool) effects

To evaluate pooling effects, we first examined if cell lines co-cultured in the same pool tend to have higher similarity in the patterns of heterogeneity than cell lines cultured in different pools. We calculated pairwise correlations between NMF programs (**Supplementary Table 3**) gene scores and assessed the proportion of the total variance (η^2, one-way ANOVA) explained by whether or not the cell lines were in the same pool. This was performed for all cell lines, as well as separately for each of the cancer types with a sufficient number of cell lines. In all cases, we observed minimal or no effect of co-culturing on the patterns of heterogeneity within cell lines.

To further evaluate the impact of the pooling procedure on cell lines' expression profiles, we designed a control experiment in which six cancer cell lines were profiled by scRNA-seq, as described above, after 3 days of being co-cultured, or individually cultured as control. The cancer cell lines included JHU006, SCC47 and CAL27 (HNSCC), RKO (colon cancer), HCC1954 (breast cancer), and SKMEL2 (melanoma) cultured in RPMI medium supplemented with 10% fetal bovine serum. Data processing and cell line assignment were performed as described above. For each cell line, we then compared the expression profiles between the co-

cultured and individually-cultured experiments. We used t-test to identify differential expression associated with the co-culture, and genes with an absolute fold change larger than 2 and FDR-adjusted p-value smaller than 0.05 were considered significant. We assessed the enrichment of significantly up and down regulated genes with Gene Ontology terms (C5:BP MSigDB[4]) using hypergeometric test (FDR-adjusted $p < 0.05$ was considered significant). Finally, we applied NMF (with factor equals to 6) to each cell line, cultured as a pool or individually, and compared the 12 programs obtained by hierarchical clustering, using one minus Pearson correlation coefficient across NMF gene scores as a distance metric.

Identifying candidate regulators of RHPs

First, any transcription factor (based on Gene Ontology annotation) that is part of a RHP was considered as a candidate regulator (**Supplementary Table 5**). Second, we performed a binding site enrichment analysis for transcription factors and chromatin regulators within RHPs, using the GTRD database[5] (gtrd.biouml.org) and FDR-adjusted $P<0.05$ by hypergeometric test (**Supplementary Table 5**). Third, in order to gain greater insight specifically into potential regulators of the EpiSen and EMT-II programs in the SCC47 and JHU006 cell lines, we ran SCENIC v1.1.1[6] (**Supplementary Table 6**). We used the GENIE3, RcisTarget and AUCell R packages, and the human motif collection version 9 (mc9nr); we filtered-out "extended" regulons if there were matching high-confidence regulons (i.e. "onlyNonDuplicatedExtended"). The input matrix represented $\log_2(CPM/10+1)$ expression values and only included genes expressed at high or intermediate levels (above 3.5) in at least 2% of cells. Enrichment of regulons with EMT-II and EpiSen genes were calculated using hypergeometric test, and those with an FDR-adjusted $p < 0.05$ in both cell lines were selected.

Predictive value of the EpiSen program for Cetuximab response in HNSCC patients

We analyzed microarray data of 40 pre-treatment samples from HNSCC recurrent-metastatic patients treated with platinum-based chemotherapy plus Cetuximab[49]. Patients were stratified into short (n = 30) and long (n = 14) progression-free survival (PFS < 5.6 and PFS > 12, respectively). Processed data and clinicopathological information were downloaded from the Gene expression Omnibus (GEO) repository (GSE65021). Differential expression analysis comparing the two groups of patients was performed using t-test. For each patient we calculated three different EpiSen scores using the following gene sets: (i) *in vitro*, defined as the top 100 most common genes among EpiSen programs of HNSCC cell lines, (ii) *in vivo*, previously defined from scRNA-seq data of HSNCC tumors[5], and (iii) shared, reflecting genes found in both the *in vitro* an *in vivo* signatures. Program scores were calculated based on $\log_2$-transformed data using the approach described above in "**Defining program scores in each cell".** We evaluated the ability of each of these signatures to discriminate patient outcomes using multivariate logistic regression, adjusting for age, gender, sample type, tumor stage (T), tumor grade, and site of primary tumor, and by calculating the area under receiver operator characteristic (ROC) curves.

Computational cell cycle analysis

Scoring cells for the G1/S and G2/M RHPs reveals a circle-like structure, reflecting different phases of the cell cycle. This pattern recurs across cell lines, and was also previously described

for different human cancers and mouse hematopoietic stem cells [6,7,63]. Since these patterns are continuous, borders between cell cycle phases were defined conservatively by manual inspection into four patterns: non-cycling ($SC_{G1/S}$ < -0.75 and $SC_{G2/M}$ < -0.5), G1 ($SC_{G1/S}$ > -0.5 and $SC_{G2/M}$ < 0), S ($SC_{G1/S}$ > 0.25 and $SC_{G2/M}$ > 0), and G2/M cycling ($SC_{G1/S}$ < 0.25 and $SC_{G2/M}$ > 0.5).

Association between program variability and mutations or drug responses

Mutation calls (coding region, germline filtered) were downloaded from the CCLE portal (https://portals.broadinstitute.org/ccle), and drug response data (CTRP v2, area under the curve, AUC) were downloaded from the CTD^2 portal (https://ocg.cancer.gov/programs/ctd2/data-portal) [64]. We restricted the analysis to non-silent mutations and compounds tested in at least 160 out of the 198 cell lines analyzed. We compared program variability scores of mutated and wild-type cell lines using two-sided t-test. The association between drug sensitivity (1-AUC) and program variability scores was assessed using multiple linear regression, with cancer type as a covariate: $Y_d \sim cancer\ type + PV_j$, for drug *d* and program *j*.

Association between programs of variability and CNA subclones

In cell lines presenting discrete programs of variability and CNA subclones, we evaluated the association between the expression-based classification of cells into subpopulations, as defined by DBSCAN, and the subclone-based classification, as defined by GMM, using Chi-square test. In cell lines presenting continuous programs of variability and CNA subclones, we compared NMF cell scores of each program between clones using t-test, for cases with 2 subclone modes, and one-way ANOVA, for cases with >2 subclone modes. P<0.001 were considered statistically significant.

Drug screen analysis

In order to define potential hits from the primary screen for follow-up in the secondary screen, we considered the differential viability between the EpiSen-high and EpiSen-low states for each compound in each cell line. In order to define hits that were differentially sensitive for only one cell line, we used 2.5 standard deviations from the mean of the difference in viability of the vehicle (DMSO-treated) controls between states as the threshold. In order to define shared hits that were differentially sensitive between the EpiSen-high and EpiSen-low states in both cell lines we used a less strict threshold of 2 standard deviations from the mean of the difference in viability of the vehicle controls between states. A third category of compounds that killed cells in both the EpiSen-high and EpiSen-low states at 10 μM (defined as ≤10% viability) was also selected for follow-up in the secondary screen at lower concentration (1 μM).

The secondary screen was performed in duplicates. In order to determine the differential viability between the EpiSen-high and EpiSen-low states, we compared the mean of each pair of duplicate measurements. To avoid an impact from outlier measurements, in each case where the difference between duplicates was larger than 20%, we calculated three potential values for differential viability between EpiSen-high and EpiSen-low populations: one value based on the

mean of the two duplicate measurements and two additional values based on each measurement alone. We then conservatively used the minimal value of differential viability to ensure that individual outlier measurements will not lead to the appearance of differential viability. In order to define hits in the secondary screen, the threshold was defined by the upper and lower bounds of the adjusted control values over replicates between states.

Defining mode of differential sensitivity in drug screen

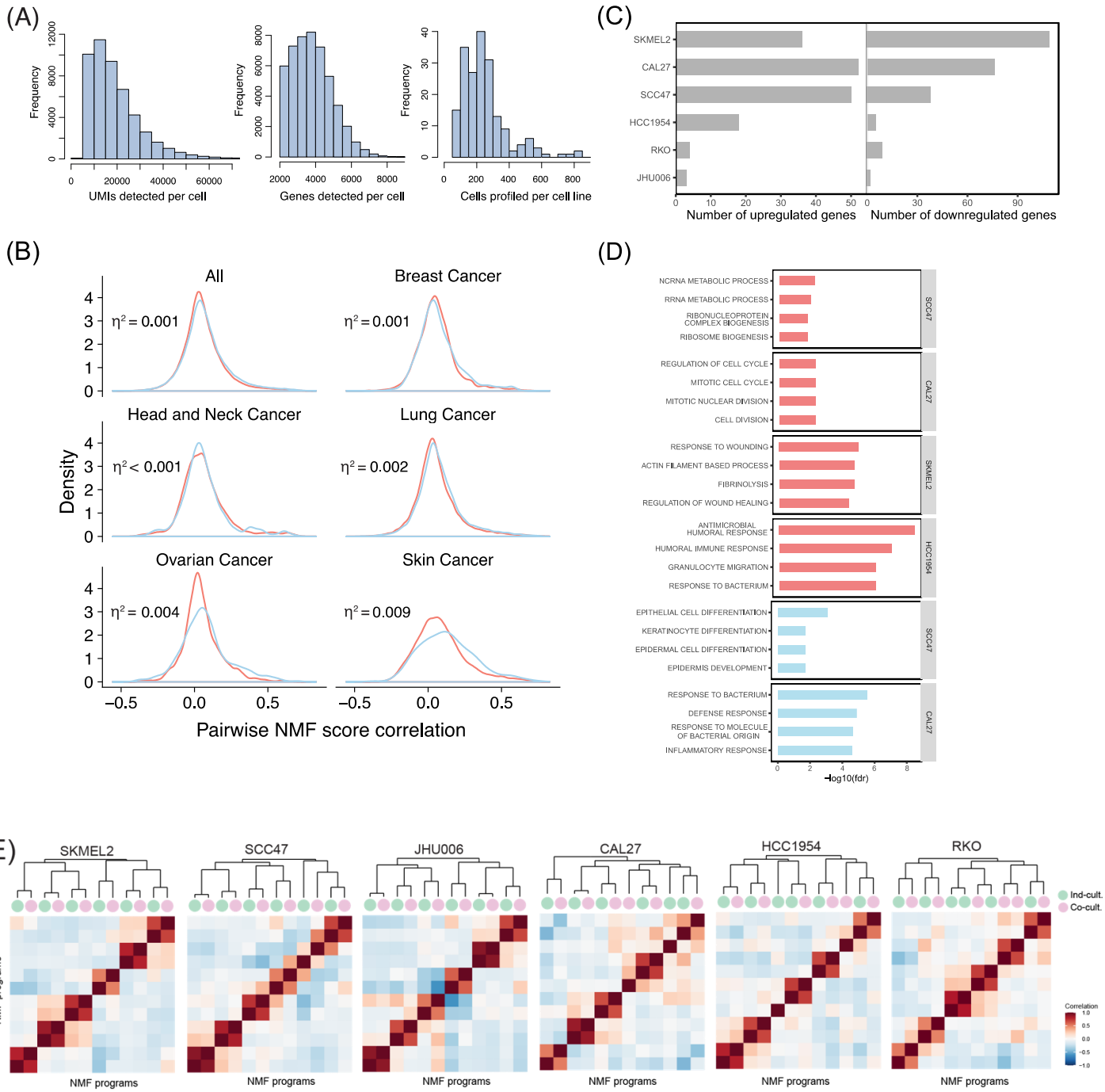
In order to define the mode of differential sensitivity among secondary screen hits (i.e. whether one subpopulation is killed more than most cells, one population is protected more than most cells, or both), we considered the differential viability of each subpopulation compared to the reference population, i.e. $\Delta v(i)=v(i)-v(ref.)$, where v reflects percent viability. Any hit in which one of the two subpopulations had $\Delta v<-10$ or $\Delta v>10$ was defined as having a protection or killing effect, respectively. If both subpopulations passed this threshold, then we deemed the hit as reflecting primarily one effect (if the Δv of one subpopulation was higher, in absolute value, than the other by at least 10), or alternatively as reflecting a combined effect. The mode of differential sensitivity is noted in **Supplementary Table 9**.

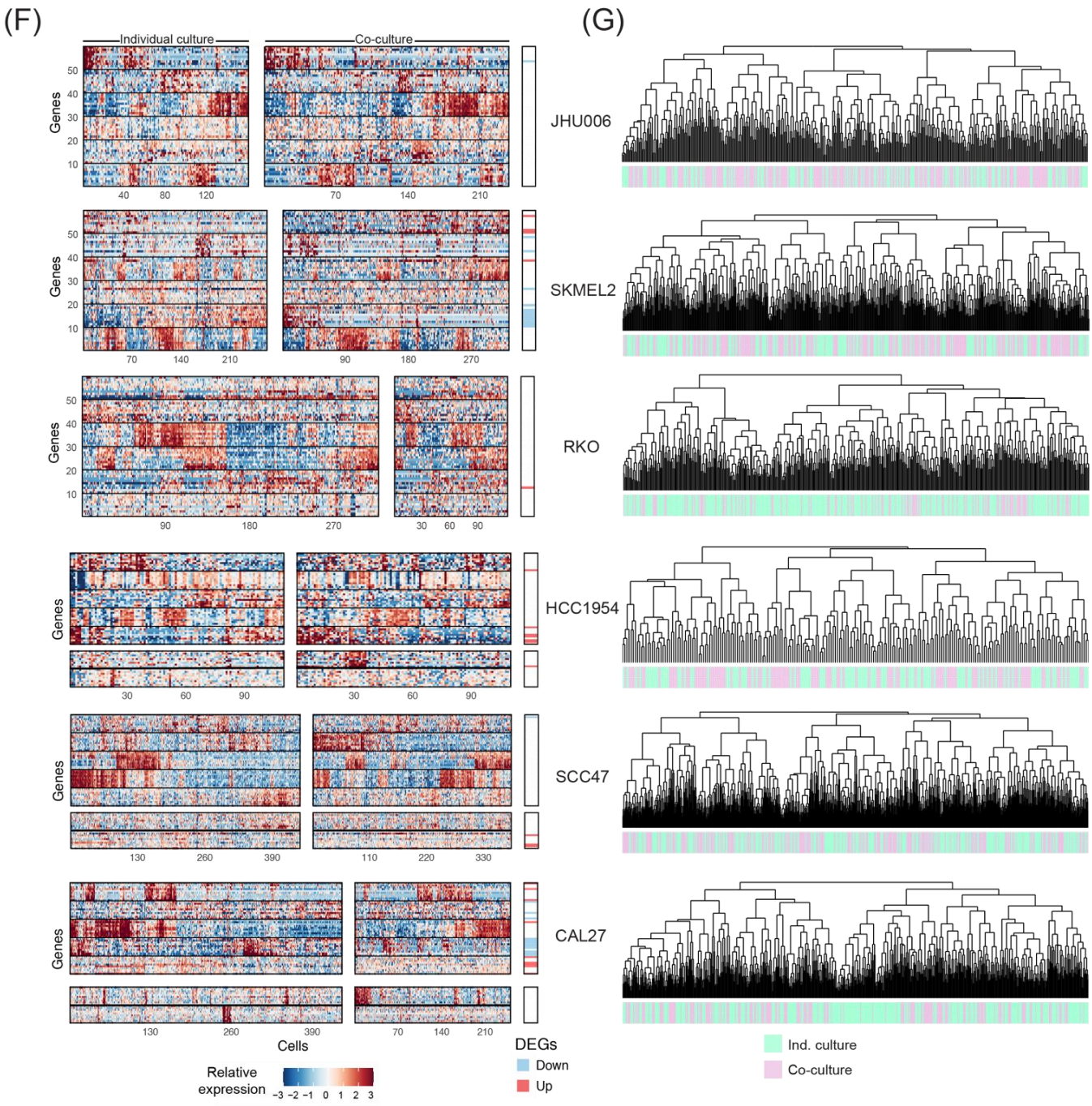
Analysis of dose response curves

A subset of compounds that were differentially sensitive between the EpiSen-high and EpiSen-low cell lines were selected for follow-up dose response studies in SCC47 in each sorted subpopulation (EpiSen-low, EpiSen-high, reference). To generate dose response curves, viability at each concentration of the seven-point dose response series was averaged over replicates and normalized to the viability of vehicle (DMSO) controls. Percent change in viability was calculated at each concentration using the normalized viability and curves were fit using these values with a three-parameter nonlinear regression model in GraphPad Prism 8 (GraphPad Software, La Jolla CA, USA) where:

Y=max%change in viability+(min%change in viability-max%change in viability)/(1+10^(X-LogEC50)) with Hill Slope=-1.0. EC50 values as determined by the three-parameter model were defined as the concentration at which the fitted curve crossed the value corresponding to half of the maximal inhibition (i.e. maximal percent change in viability). For visualization of dose response curves, normalization was performed to lowest concentration of drug. Two compounds for which EC50 could not be accurately calculated were omitted from further analysis. A paired t-test was performed using the aggregated differences in viability at each concentration to determine statistical significance ($P \leq 0.05$) of differential viability between curves (**Supplementary Table 10**).

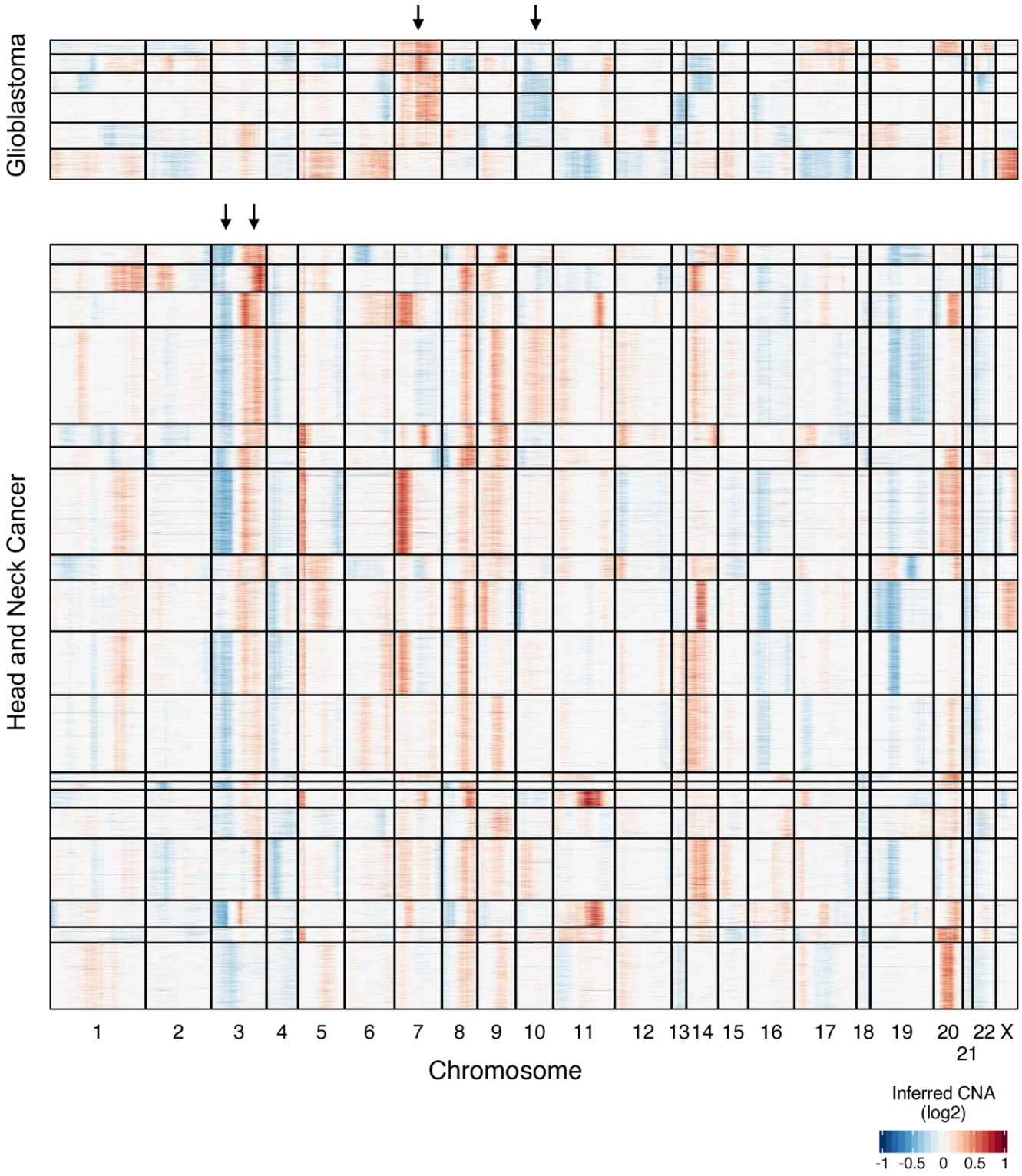
Supplementary Figures




(F)
(G)
Individual culture
Co-culture
Genes
Cells
JHU006
SKMEL2
RKO
HCC1954
SCC47
CAL27
Relative expression
−3 −2 −1 0 1 2 3
DEGs
Down
Up
Ind. culture
Co-culture

Supplementary Figure 1. Quality control for the multiplexed scRNA-seq of cancer cell lines.

(**A)** Histograms show distributions of the number of UMIs detected per cell (left), the number of genes detected per cell (middle), and the number of cells detected per cell line (right). **(B)** Distribution of pairwise correlations between expression programs of heterogeneity defined from cell lines in the same (blue) and in different (red) pools. Comparisons included all NMF programs detected in each of the cell lines (**Fig. 2E**). Analyses were performed for all cell lines combined (top left panel) or separately for each cancer type (all other panels, only most abundant cancer types are shown). The proportion of total variance (η^2) explained by whether or not cell lines were in the same pool (calculated using one-way ANOVA) is indicated, suggesting that patterns of expression heterogeneity are largely unaffected by the pooling procedure. **(C)** Number of significantly upregulated or downregulated genes (absolute fold-change>2, FDR-adjusted P<0.05, two-sided t-test) between co-culture and individual-culture conditions, for each of the six cell lines in the control pool (rows). **(D)** Enriched GO annotations among the differentially expressed genes (DEGs) defined in (C). Shown are the hypergeometric test significance values (-log10(P), FDR-adjusted) for the 4 most enriched GO annotations among the upregulated (red bars) or downregulated (blue bars) genes of each cell line with at least one significant enrichment (FDR<0.05). Enriched functions tend to be cell line specific rather than shared across cell lines. **(E)** For each of the cell lines in the control pool, heatmap shows the pairwise correlations between gene scores of 6 NMF programs obtained from the co-culture profiles (indicated by pink circles) and 6 from the individual-culture profiles (indicated by green circles). NMF programs are ordered by hierarchical clustering, as indicated by the top dendrogram. This analysis demonstrates that the vast majority of NMF programs are shared between the two culture conditions, indicating that they are only minimally affected by co-culturing. Specifically, 91.67% (66 of 72) of the NMF programs (12 of 12 in JHU006, SKMEL2 and RKO, and 10 of 12 in SCC47, CAL27 and HCC1954) cluster closely as pairs of NMF programs from the two conditions. **(F)** For each of the cell lines in the control pool, heatmaps show the expression of the top 10 scoring genes in shared and non-shared NMF programs across cells (columns) in the individual-culture (left) and the co-culture (right) experiments. For shared NMF programs, we considered the average gene score in the two matching programs from the two culture conditions, as shown in (E); the six non-matching programs are shown at the bottom parts of the three bottom panels (for SCC47, CAL27 and HCC1954), demonstrating that even though these programs were identified in one condition, most of them have qualitatively similar patterns in the two conditions. The right-most panel indicates if those genes were defined as differentially expressed between co-culture and individual-culture conditions. **(G)** For each of the cell lines in the control pool, dendrogram shows hierarchical clustering of all cells from the two conditions based on the expression of NMF program genes shown in (F). Cells from the individual-culture and the co-culture conditions are shown in green and pink, respectively. This analysis demonstrates that cells from the two conditions do not tend to produce distinct clusters.



Supplementary Figure 2. Inferred CNAs are consistent with expected chromosomal aberrations. Heatmap depicts inferred CNAs for individual cells (rows) from 6 glioblastoma (top) and 19 HNSCC (bottom) cell lines, based on average expression in sliding windows of 100 genes. Arrows highlight expected hallmark alterations - the gain of chromosome 7 and loss of chromosome 10 in glioblastoma, and the loss of chromosome 3p and gain of chromosome 3q in HNSCCs.

Supplementary References

1. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503-508 (2019).
2. Kang, H.M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol* **36**, 89-94 (2018).
3. Friedman, J., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1-22 (2010).
4. Liberzon, A. A description of the Molecular Signatures Database (MSigDB) Web site. *Methods Mol Biol* **1150**, 153-60 (2014).
5. Yevshin, I., Sharipov, R., Valeev, T., Kel, A. & Kolpakov, F. GTRD: a database of transcription factor binding sites identified by ChIP-seq experiments. *Nucleic Acids Res* **45**, D61-D67 (2017).
6. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nat Methods* **14**, 1083-1086 (2017).
7. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc* **9**, 171-81 (2014).
8. Rauner, G., Kudinov, T., Gilad, S., Hornung, G. & Barash, I. High Expression of CD200 and CD200R1 Distinguishes Stem and Progenitor Cell Populations within Mammary Repopulating Units. *Stem Cell Reports* **11**, 288-302 (2018).