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

nature genetics

Check for updates

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

Gabriela S. Kinker[1,2,10], Alissa C. Greenwald[1,10], Rotem Tal[1], Zhanna Orlova[1], Michael S. Cuoco[3], James M. McFarland[4], Allison Warren[4], Christopher Rodman[3], Jennifer A. Roth[4], Samantha A. Bender[4], Bhavna Kumar[5], James W. Rocco[5], Pedro A. C. M. Fernandes[2], Christopher C. Mader[4], Hadas Keren-Shaul[6,7], Alexander Plotnikov[6], Haim Barr[6], Aviad Tsherniak[4], Orit Rozenblatt-Rosen[3], Valery Krizhanovsky[1], Sidharth V. Puram[8], Aviv Regev[3,9] and Itay Tirosh[1]✉

Cultured cell lines are the workhorse of cancer research, but the extent to which they recapitulate the heterogeneity observed among malignant cells in tumors is unclear. Here we used multiplexed single-cell RNA-seq to profile 198 cancer cell lines from 22 cancer types. We identified 12 expression programs that are recurrently heterogeneous within multiple cancer cell lines. These programs are associated with diverse biological processes, including cell cycle, senescence, stress and interferon responses, epithelial–mesenchymal transition and protein metabolism. Most of these programs recapitulate those recently identified as heterogeneous within human tumors. We prioritized specific cell lines as models of cellular heterogeneity and used them to study subpopulations of senescence-related cells, demonstrating their dynamics, regulation and unique drug sensitivities, which were predictive of clinical response. Our work describes the landscape of heterogeneity within diverse cancer cell lines and identifies recurrent patterns of heterogeneity that are shared between tumors and specific cell lines.

Cellular plasticity and heterogeneity are fundamental features of human tumors that play a major role in disease progression and treatment failure[1,2]. For example, rare subpopulations of tumor cells may be an underlying cause of resistance or may facilitate metastasis. Single-cell RNA sequencing (scRNA-seq) has emerged as a valuable tool to study intratumor heterogeneity (ITH) directly in patient samples[3–12]. Yet the mechanisms and functional implications of ITH patterns have been difficult to resolve, thus calling for extensive follow-up studies in model systems. Genetic diversity, epigenetic plasticity and interactions within the tumor microenvironment all contribute to ITH. However, we hypothesize that a considerable fraction of ITH in expression patterns of malignant cells reflects intrinsic cellular plasticity, which exists even in the absence of a native microenvironment. To examine the heterogeneity within cancer cell lines and their ability to recapitulate ITH programs, we sought to define the landscape of cellular diversity within a large number of cell lines from the Cancer Cell Line Encyclopedia (CCLE) collection[13,14].

Results

Pan-cancer scRNA-seq of human cell lines. We developed and applied a multiplexing strategy in which cells from different cell lines were profiled in pools by scRNA-seq and then computationally assigned to the corresponding cell line (Fig. 1a). We used existing pools that were generated from the CCLE collection[13,15]. Each pool consisted of 24–27 cell lines that were from diverse lineages but had comparable proliferation rates and was profiled by scRNA-seq with the 10x Genomics Chromium system, for an average of 280 cells per cell line (Methods). We profiled eight CCLE pools, along with one smaller custom pool that included head and neck squamous cell carcinoma (HNSCC) cell lines.

We assigned profiled cells to cell lines based on the consensus between two complementary approaches, which used genetic and expression profiles (Fig. 1a). First, cells were clustered by their global expression profiles, and each cluster was mapped to the cell line with the most similar bulk RNA-seq profile[14]. Second, by detection of SNPs in the scRNA-seq reads, we assigned cells to the cell line with the highest similarity by SNP profiles derived from bulk RNA-seq[14,16]. Cell line assignments based on gene expression and SNPs were consistent for 98% of the cells, which were retained for analysis (for example, Fig. 1b). The few inconsistent assignments were observed primarily in cells with low data quality, resulting in low SNP coverage, and were therefore excluded. Cell lines with fewer than 50 assigned cells were also excluded from further analyses, as were cells with low-quality data and suspected doublets. Overall, following assignment and quality control filters, we studied the expression profiles of 53,513 cells from 198 cell lines, reflecting 22 cancer types (Fig. 1c and Supplementary Table 1). We detected an average of 19,264 unique molecular identifiers (UMIs) and 3,802 genes per cell, underscoring the high quality of our dataset (Supplementary Fig. 1a).

A potential caveat of our multiplexing approach is that in the previously generated CCLE pools, but not in the custom HNSCC pool,

[1]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [2]Institute of Bioscience, University of Sao Paulo, Sao Paulo, Brazil. [3]Klarman Cell Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Cancer Program, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Department of Otolaryngology–Head and Neck Surgery, The Ohio State University Wexner Medical Center, Columbus, OH, USA. [6]The Nancy & Stephen Grand Israel National Center for Personalized Medicine, Weizmann Institute of Science, Rehovot, Israel. [7]Life Science Core Facility, Weizmann Institute of Science, Rehovot, Israel. [8]Department of Otolaryngology–Head and Neck Surgery, Washington University School of Medicine, St. Louis, MO, USA. [9]Present address: Genentech, South San Francisco, CA, USA. [10]These authors contributed equally: Gabriela S. Kinker, Alissa C. Greenwald. ✉e-mail: itay.tirosh@weizmann.ac.il

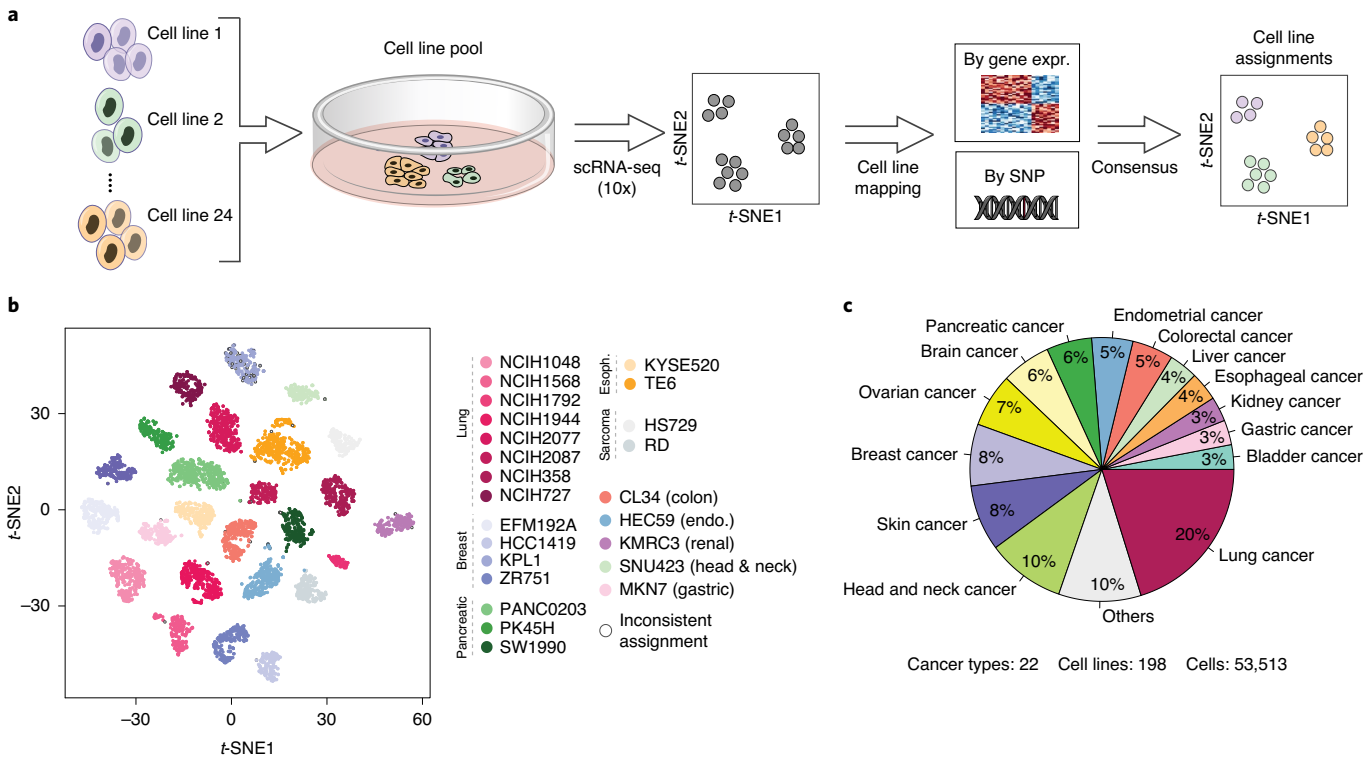


Fig. 1 | Characterizing expression heterogeneity within cell lines by multiplexed scRNA-seq. a, Workflow of the multiplexing strategy used to profile multiple cell lines simultaneously. Cell lines were pooled and profiled by droplet-based scRNA-seq. We used reference CCLE data to assign cells to the most similar cell line based on their overall gene expression and SNP pattern. Expr., expression; endo., endometrial; esoph., eshophageal. **b**, *t*-SNE plot of a representative pool demonstrating the robustness of cell line assignment. Cells with inconsistent assignments (by gene expression and SNPs) are denoted and were excluded from further analyses. **c**, Distribution of the cancer types that were profiled.

cell lines were co-cultured for 3 d before profiling by scRNA-seq, and hence their expression patterns may have been affected. However, our analyses suggest that the effect of co-culturing was limited, particularly when considering the heterogeneity within each cell line, which is the focus of this work. First, the patterns of heterogeneity were as similar between cell lines from the same pool as they were between cell lines from different pools (Supplementary Fig. 1b). Second, we performed a control experiment in which six cell lines were profiled with and without co-culturing for 3 d. Co-culturing had a modest effect on the average gene expression in each cell line, while patterns of heterogeneity within cell lines were highly consistent between the two conditions (Supplementary Fig. 1c–f).

Patterns of expression heterogeneity within cell lines. Extensive variability in gene expression was identified across cells within individual cell lines, including discrete subpopulations of cells, as well as continuous patterns that reflect spectra of cellular states (Fig. 2a). To identify discrete subpopulations, we used dimensionality reduction with *t*-distributed stochastic neighbor embedding (*t*-SNE) followed by density-based clustering (DBSCAN; Extended Data Fig. 1a and Methods). Discrete clusters were found only within a minority (11%) of the cell lines (Fig. 2b and Extended Data Fig. 1b,c), and the expression programs of discrete clusters showed limited similarities between cell lines, indicating that discrete subpopulations were typically cell line specific (Fig. 2c and Extended Data Fig. 1d).

Next, to identify both continuous and discrete variability of cellular states, we applied non-negative matrix factorization (NMF) to each cell line[5] (for example, Fig. 2d and Methods). Overall, we detected 1,445 robust expression programs across all cell lines, with four to nine such programs in individual cell lines (Extended Data Fig. 1e and Supplementary Table 3). To identify common expression programs that varied within multiple cell lines, we first excluded those with limited similarity to all other programs, as well as those associated with technical confounders (Extended Data Fig. 1e,f). Of the remaining 800 programs, only 4.75% corresponded to the discrete subpopulations described above (Fig. 2e).

Comparison of the NMF programs, based on their shared genes, emphasized clusters of similar programs, allowing us to define the consensus of each cluster, which we termed recurrent heterogeneous programs (RHPs) of gene expression, as they were heterogeneous within multiple cell lines. As expected, the two most prominent RHPs reflected the cell cycle (Fig. 2e and Supplementary Table 4). The cell cycle RHPs corresponded to the G1/S and G2/M phases (Fig. 2e), as was also observed in clinical tumor samples (Extended Data Fig. 2a). G2/M programs were highly similar across cell lines and tumors, defining a generic mitotic program (Extended Data Fig. 2b). In contrast, G1/S programs associated with genome replication were more context dependent. A central difference in G1/S programs involved the MCM complex genes (*MCM2–MCM7*) and the linker histone H1 family genes (*HIST1H1B–HIST1H1E*), which were robustly upregulated in only tumors and cell lines, respectively (Extended Data Fig. 2b,d). This may reflect an adaptation to rapid growth and loss of the G1 checkpoint in cell lines in vitro. Consistent with this possibility, cell lines had a much lower percentage of apparent G0 cells (that is, cells lacking both the G1/S and G2/M expression programs; Extended Data Fig. 2e).

RHPs mirror diverse biological processes and in vivo states. The ten additional RHPs reflected diverse biological processes (Fig. 3 and Supplementary Table 4). These RHPs were either

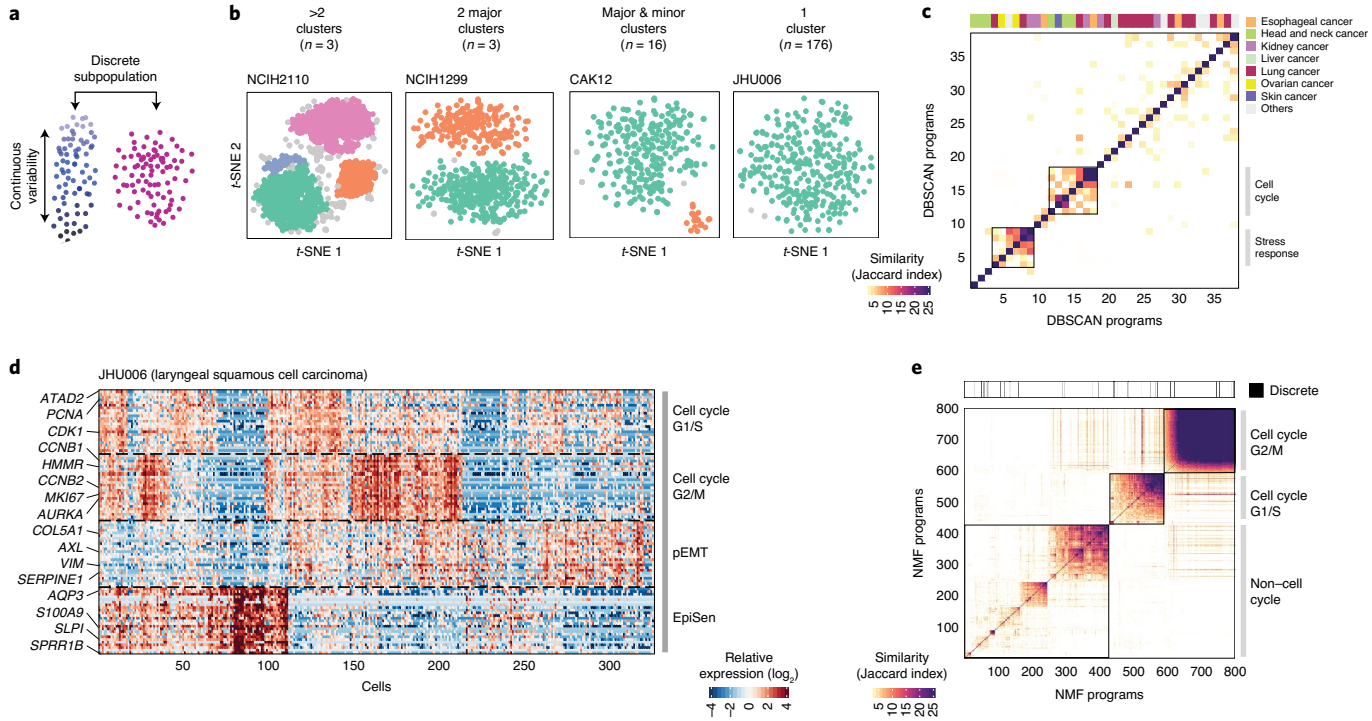


Fig. 2 | Discrete and continuous patterns of heterogeneity within cell lines. a, Illustration of the two types of expression variability that were investigated. **b**, *t*-SNE plots showing exemplary cell lines for the four classes defined by the presence and number of discrete subpopulations that were identified using DBSCAN. The description of each class and number of cell lines is indicated above the *t*-SNE plots. **c**, Heatmap depicting pairwise similarities between the gene expression programs (quantified by Jaccard Index over the programs' genes and expresed as percentages) defined for each of the cell clusters derived from the 22 cell lines that were identified as having one or more discrete subpopulations. Hierarchical clustering identified only two groups of similar programs (metaprograms). The top panel shows assignment of cancer types. **d**, Continuous programs of heterogeneity identified using NMF in a representative cell line that lacks discrete subpopulations (JHU006; see **b**). The heatmap shows relative gene expression from four NMF programs across all JHU006 cells, ordered by hierarchical clustering. NMF programs are annotated (right) and the selected genes are indicated (left). **e**, Hierarchical clustering of pairwise similarities between NMF programs that were identified across all the analyzed cell lines. Programs with limited similarity to all other programs, as well as those associated with technical confounders, were excluded. The top panel indicates the 4% of NMF programs that were consistent with discrete subpopulations identified by DBSCAN ($P < 0.001$, two-sided Fisher's exact test).

largely independent of cell cycle status or preferentially expressed by non-cycling cells (Extended Data Fig. 3a,b). Importantly, each of the RHPs was detected across at least eight different cell lines and from at least four different pools, highlighting their robustness (Extended Data Fig. 3c). We characterized these ten RHPs by functional enrichment of their signature genes (Fig. 3d), by the cell lines in which they were observed (Extended Data Fig. 4d,e) and by their potential regulators[17] (Supplementary Tables 5 and 6).

In addition, we examined the similarity of these in vitro RHPs to recurrent in vivo expression programs that vary across cells within patient tumor samples. In vivo RHPs were defined previously in HNSCC[5], melanoma[6], glioblastoma[18] and ovarian cancer[19], and we defined additional RHPs by NMF analysis of scRNA-seq datasets from HNSCC[5], melanoma[6], breast cancer[9] and lung cancer[12] samples (Extended Data Fig. 2a and Supplementary Table 7). Strikingly, seven of the ten cell line RHPs were highly similar to the in vivo RHPs, as shown by a highly significant overlap of signature genes (Fig. 4a, false discovery rate (FDR)-adjusted $P < 10^{-9}$ by hypergeometric test), as well as by high correlation of cell scores (Fig. 4b). The in vivo relevance of cell line RHPs was further demonstrated in melanoma and in HNSCC by a combined analysis of cells from cell lines and tumors, revealing common patterns of variation as described below (Fig. 4c–f and Extended Data Fig. 4).

RHPs are associated with multiple types of stress responses. One of the RHPs (8) reflected a stress response, including DNA-damage-induced and immediate early genes (for example, *DDIT3*, *DDIT4* and *ATF3*), resembling programs identified in melanoma and HNSCC tumors[5,6] (Fig. 4a,b). Another RHP (4) contained interferon (IFN)-response genes (for example, *IFIT1–IFIT3*), strongly resembling a program of heterogeneity observed in ovarian cancer[19]. The IFN response may be triggered by genomic instability through the cGAS–STING pathway[20]. Accordingly, the IFN-response program was reduced in cell lines with mutations in *MRE11A*, which encodes a protein that recognizes cytosolic double-stranded DNA and activates STING[21] (Extended Data Fig. 5a).

Two other RHPs (9 and 10) consisted of genes related to protein folding and maturation and to proteasomal degradation, respectively. These were the only RHPs that did not resemble any of the in vivo heterogeneity programs. However, it is possible that such programs exist in vivo and have not been detected yet.

hence, very little to no red in the last two columns of 4a

RHPs recapitulate in vivo EMT programs. Three distinct RHPs were related to the epithelial–mesenchymal transition (EMT). The first (RHP 2; EMT-I) was specific to melanoma (Extended Data Fig. 3d) and correlated negatively with another melanoma-specific RHP (1), which was associated with skin pigmentation (SkinPig) genes (for example, *MITF* and *PMEL*). Both of these melanoma-specific RHPs and their negative correlation recapitulated patterns observed in melanoma tumors (Fig. 4a,b), in which the RHPs were linked to drug resistance[6,22]. Accordingly, in a combined analysis of in vitro

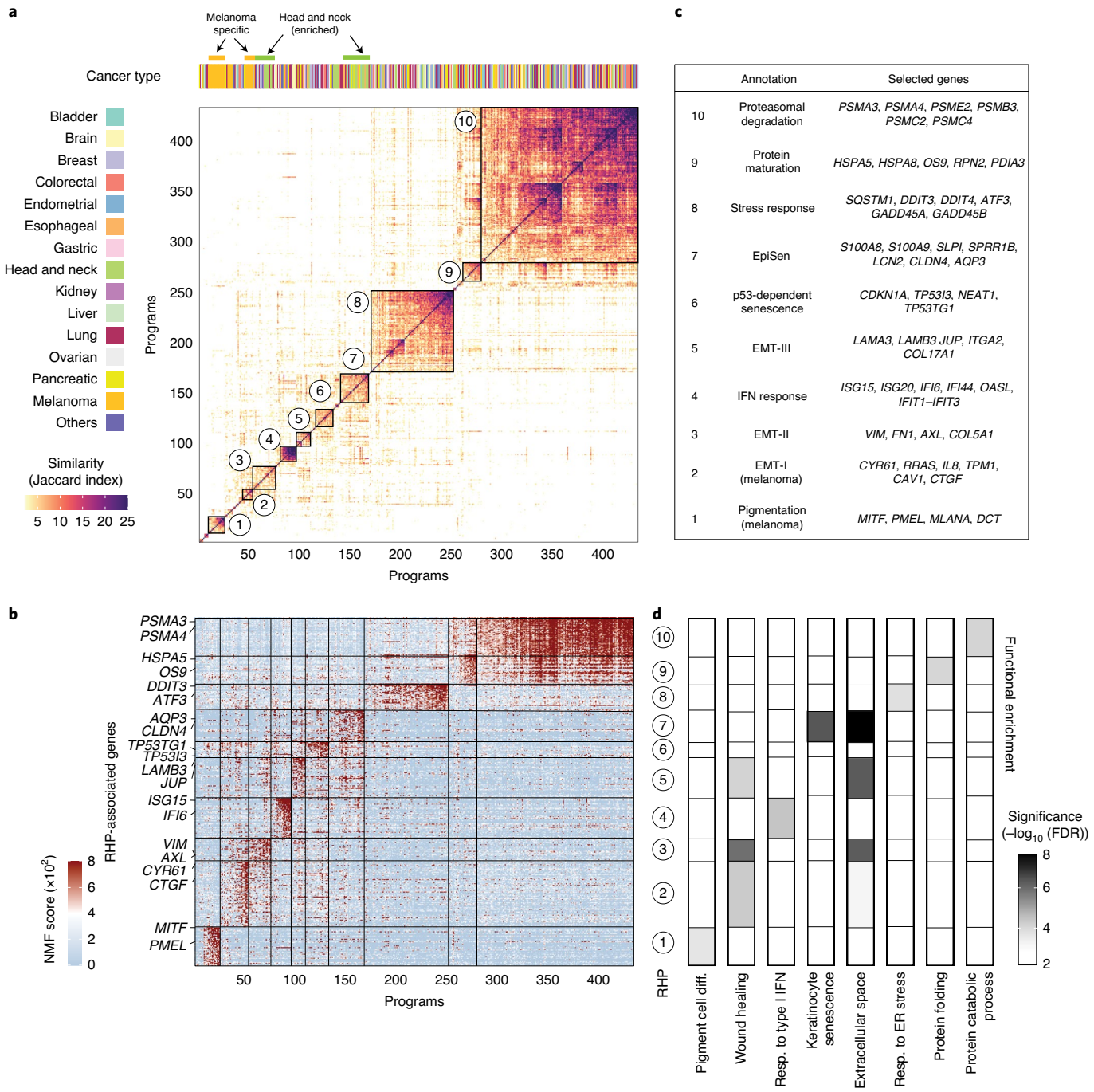
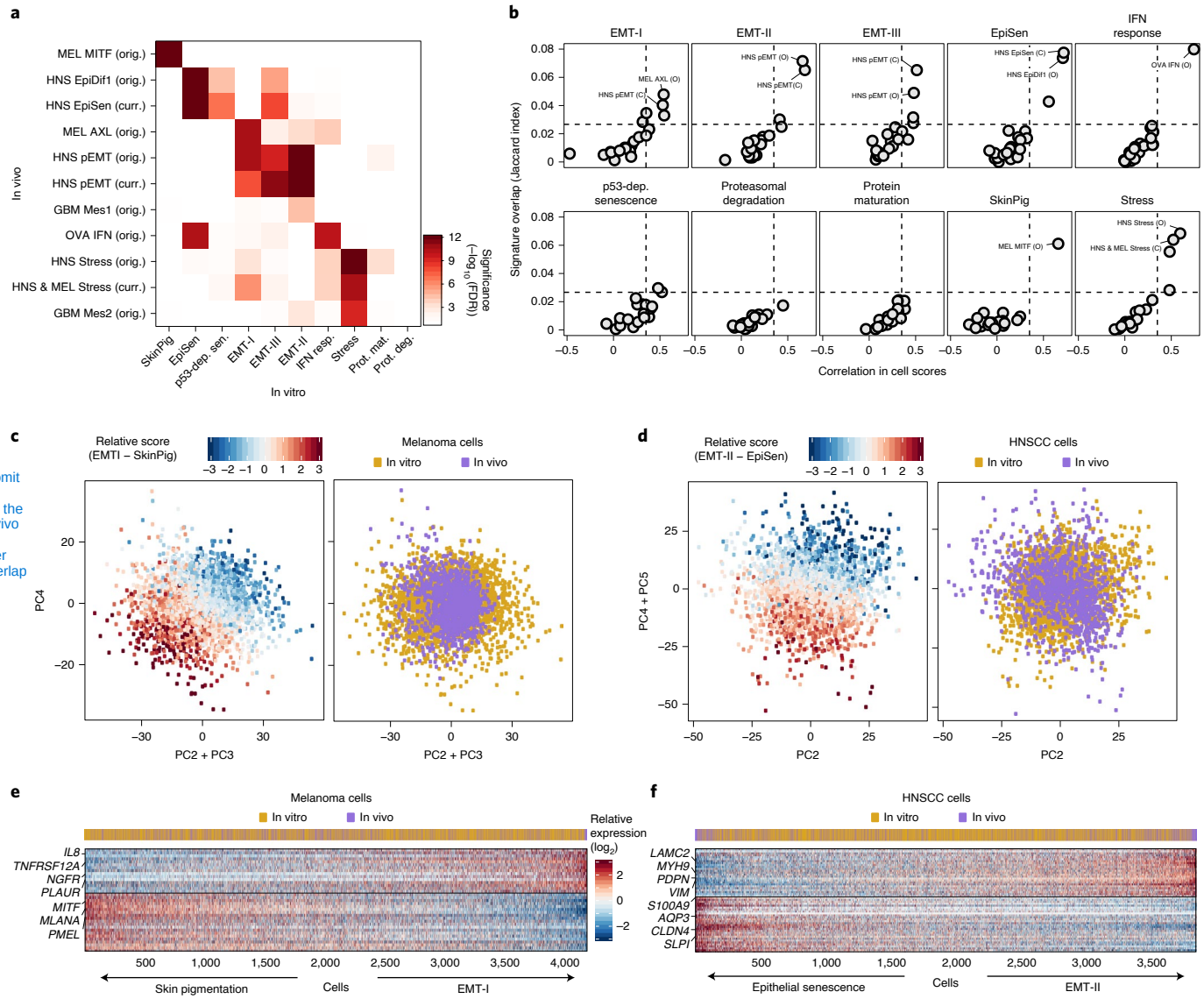


Fig. 3 | Functional annotation of RHPs. a, The main heatmap depicts pairwise similarities between all NMF programs (except for those linked to the cell cycle, Fig. 2e), ordered by hierarchical clustering. Ten clusters (RHPs) are indicated by squares and numbers. The top panel shows assignment of cancer types, highlighting significant enrichment (hypergeometric test, $P < 0.001$) of melanoma and HNSCC cell lines (Extended Data Fig. 3d). **b**, NMF scores of signature genes for each RHP (rows), with selected genes labeled. Programs (columns) are ordered as in **a**. **c**, Annotation and selected top genes for each of the ten RHPs. **d**, Functional enrichment ($-\log_{10}$ of FDR-adjusted P values, hypergeometric test) of RHP genes with eight annotated gene sets.

and in vivo melanoma cells, these two RHPs corresponded to three of the top five principal components (PCs) and defined a range of cellular states (Fig. 4c,e and Extended Data Fig. 4a,b,e). Notably, our data highlight certain melanoma cell lines as faithful model systems for these in vivo-related RHPs (Extended Data Fig. 4e).

Two other RHPs, EMT-II (3) and EMT-III (5), also reflected EMT-like processes in distinct cell lines. EMT-II was mainly observed in HNSCC cell lines, although across seven distinct pools (Extended Data Fig. 3c,d). It included *VIM* (encoding vimentin), *FN1* (fibronectin) and other genes, closely mirroring the partial EMT state we previously observed in HNSCC tumors (Fig. 4a,b,d,f and Extended Data Fig. 4), in which it was linked to metastasis[5]. Cell lines harboring EMT-II had a reduced frequency of *NOTCH4* mutations and were more sensitive to inhibitors of NOTCH signaling (Extended Data Fig. 5a,b), suggesting a potential role of the NOTCH pathway in enabling EMT-II



why did they omit PC1?
did PC1 make the clusters of in vivo and in vitro separate rather than blend/overlap

Fig. 4 | In vitro RHPs recapitulate in vivo programs of heterogeneity. a, Significance of the overlap ($-\log_{10}$ of FDR-adjusted *P* values, hypergeometric test) between RHP gene sets defined in cell lines (in vitro, *x* axis) and in tumors (in vivo, *y* axis). In vivo RHPs are named by a cancer type abbreviation (MEL, melanoma; HNS, HNSCC; GBM, glioblastoma; OVA, ovarian cancer) followed by an associated functional annotation and are labeled to indicate whether they were defined by the original study (orig., O) or the current study (curr., C; Extended Data Fig. 2a). Sen., senescence; resp., response; mat., maturation; deg., degradation; dep, dependent; prot., protein. **b**, Each panel shows the mean Jaccard index (*y* axis) and mean correlation of single-cell scores (*x* axis) between the NMF programs constituting a specific in vitro RHP (as noted at the top) and all in vivo RHPs. The most similar in vivo RHPs are labeled as in **a**. Dashed lines indicate a 99.9% confidence threshold determined by permutations of NMF programs. **c**, Scatterplots of melanoma cells (3,033 cells from cell lines and 1,169 cells from tumors) based on PC2 + PC3 (*x* axis) and PC4 (*y* axis) coordinates. Cells are colored by the relative score for the EMT-I and SkinPig genes shared between cell line and tumor RHPs (left) and by whether the cells were from tumors or cell lines (right). **d**, Scatterplots of HNSCC cells (2,780 cells from cell lines and 1,078 cells from tumors) based on PC2 (*x* axis) and PC4 + PC5 (*y* axis) coordinates. Cells are colored by relative score for the EMT-II and EpiSen genes shared between cell line and tumor RHPs (left) and by whether the cells were from tumors or cell lines (right). **e**, Heatmap showing the relative expression of shared EMT-I and SkinPig RHP genes (rows) across melanoma cells (columns), sorted by relative RHP scores. The origin of cells from tumors or cell lines is shown in the top panel. **f**, Heatmap showing the relative expression of shared EMT-II and EpiSen RHP genes (rows) across HNSCC cells (columns), sorted by relative RHP scores. The origin of cells from tumors or cell lines is shown in the top panel.

variability. In contrast, EMT-III was enriched among non-cycling cells (Extended Data Fig. 3a,b) and contained genes involved in cell-junction organization, such as laminin-encoding genes and *JUP* (plakoglobin)[23]. The identification of three distinct EMT programs, two of which were enriched in specific cancer types, highlights EMT as a common, yet context-specific, pattern of cellular heterogeneity, which may have important implications for metastasis and drug responses.

RHPs related to senescence programs. RHPs 6 and 7 were preferentially observed in G0 cells (Extended Data Fig. 3a,b) and seem to reflect different expression programs related to cellular senescence. RHP 6 was enriched in *TP53*-wild-type cell lines and in those sensitive to the activation of p53 by nutlin-3a (Extended Data Fig. 5). Moreover, this program included the senescence mediator p21 (encoded by *CDKN1A*) and other p53-target genes. Thus, we annotated RHP 6 as a 'classical' p53-dependent senescence program.

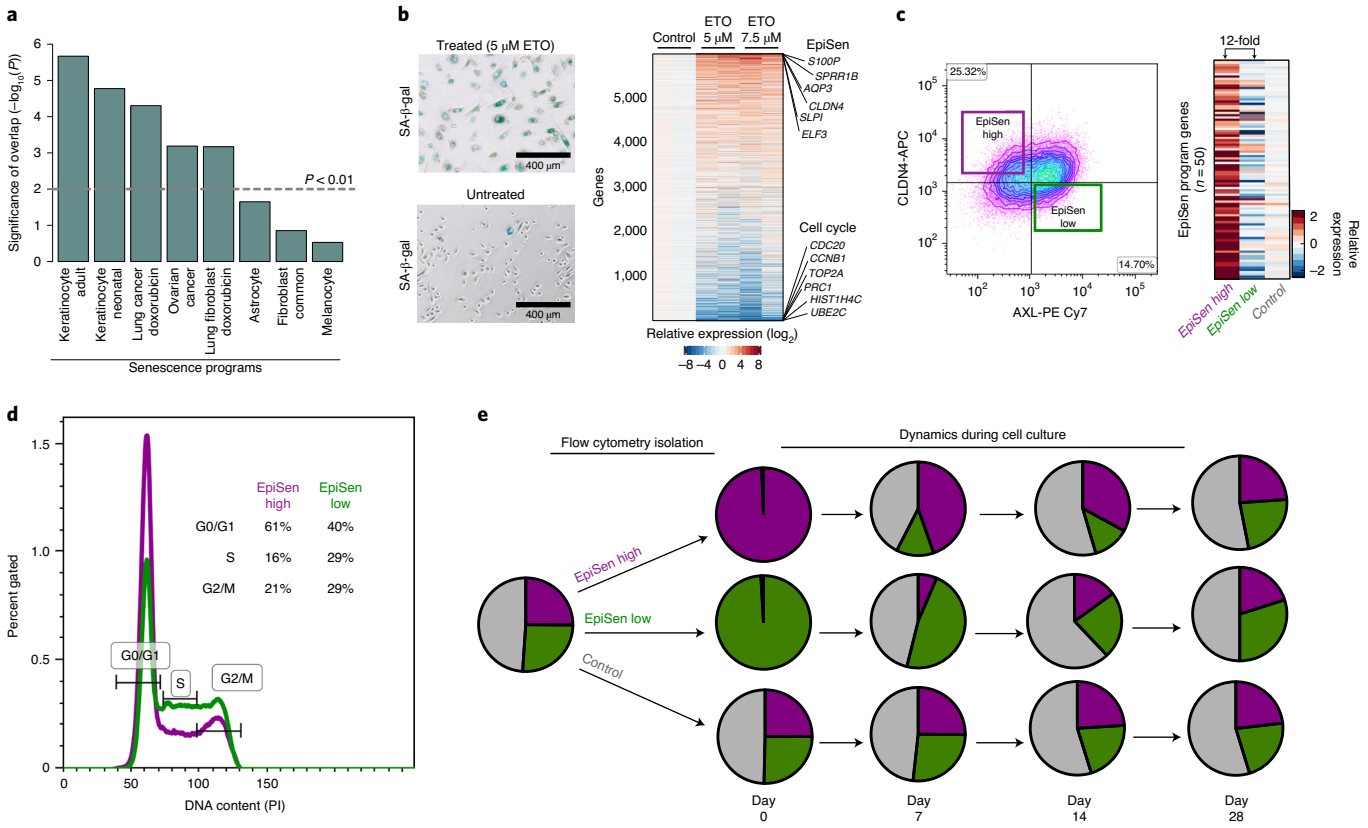


Fig. 5 | Interrogating the EpiSen RHP in primary cells and model cell lines. a, Significance of the overlap ($-\log_{10}(P)$, hypergeometric test) between the EpiSen RHP ($n=38$ genes) and eight previously reported senescence programs, keratinocyte adult ($n=72$), keratinocyte neonatal ($n=365$), lung cancer doxorubicin ($n=414$), ovarian cancer ($n=200$), lung fibroblast doxorubicin ($n=200$), astrocyte ($n=200$), fibroblast common ($n=120$) and melanocyte ($n=200$). **b**, Left, induction of senescence with etoposide in primary lung bronchial cells, confirmed by SA-β-gal staining. Right, heatmap depicting the relative expression of the top 6,000 expressed genes (rows) in primary lung bronchial cells, 9 d after induction of senescence by etoposide treatment for 48 h at two concentrations with two biological replicates. EpiSen and cell cycle programs were the most upregulated and downregulated programs, respectively (Extended Data Fig. 6c); select genes from these programs are labeled. ETO, etoposide. **c**, Left, isolation of the EpiSen-high ($AXL^{-}CLDN4^{+}$) and EpiSen-low ($AXL^{+}CLDN4^{-}$) populations by flow cytometry in JHU006 cells. Right, heatmap showing the relative expression of EpiSen program genes in three sorted subpopulations, including a control population. **d**, Flow cytometry analysis of the cell cycle was performed by staining sorted EpiSen-high and EpiSen-low JHU006 cells with the DNA-binding dye propidium iodide (PI). **e**, Pie charts depicting relative proportions of EpiSen-high (purple), EpiSen-low (green) and intermediate (gray) subpopulations in SCC47 cells, for an unsorted sample (left, initial distribution) and for sorted subpopulations (right) that were analyzed immediately after sorting (day 0) and at three additional time points (days 7, 14 and 28 in culture).

In contrast, RHP 7 was not enriched in *TP53*-wild-type cell lines but was enriched in HNSCC cell lines (Extended Data Fig. 3d,e).

RHP 7 was highly similar to the senescence program of keratinocytes and was also similar to other published senescence programs[24–30] (Fig. 5a, Extended Data Fig. 6a,b and Supplementary Table 8). To further examine senescence-related programs of epithelial cells, we profiled primary lung bronchial cells by bulk RNA-seq after inducing senescence with etoposide. The senescence phenotype was validated by senescence-associated β-galactosidase (SA-β-gal) staining and RNA-seq showing the downregulation of cell cycle genes (Fig. 5b). Genes in both of the senescence-associated RHPs (6 and 7) were upregulated, although the effect was much stronger for genes in RHP 7 (Fig. 5b and Extended Data Fig. 6c). RHP 7 also contained many genes encoding secreted proteins, consistent with a senescence-associated secretory phenotype (SASP). These included *S100A8*, *S100A9*, *SAA1*, *SAA2*, *LCN2*, *CXCL1* and *SLPI*, which are involved in inflammatory responses and may influence cancer, stromal and immune cells in the tumor microenvironment. While most of these factors are not traditionally considered to be classical SASP genes[31], we found a significant overlap ($P<0.01$, hypergeometric test) with secreted factors from multiple other senescence-related programs, including those from the in vivo counterpart in HNSCC tumors (Extended Data Fig. 6d,e).

In sum, RHP 7 was associated with low levels of proliferation and a secretory phenotype and strongly resembles the senescence response of keratinocytes, lung bronchial cells and other epithelial cells. This program lacked classical senescence markers (for example, p16 and p21) and differed from published senescence signatures of fibroblasts and melanocytes[24], underscoring the context specificity of senescence expression programs. We therefore denoted RHP 7 as an epithelial senescence-associated (EpiSen) program. We note that, although the EpiSen program was induced in senescent cells, its expression does not necessarily imply a complete senescence phenotype.

Notably, EpiSen recapitulated a program we previously observed in HNSCC tumors, 'EpiDif1' (Figs. 4a,b and 5a), which was negatively associated with the cell cycle and spatially restricted to the hypoxic tumor core[5]. This program was negatively correlated with the EMT-II program, comprising a spectrum of cellular states that were shared by multiple HNSCC cell lines and tumors. Accordingly, in a combined analysis of HNSCC cells in vitro and in vivo, the two RHPs were associated with three of the top five PCs and defined a range of cellular states (Fig. 4d,f and Extended Data Fig. 4c–e).

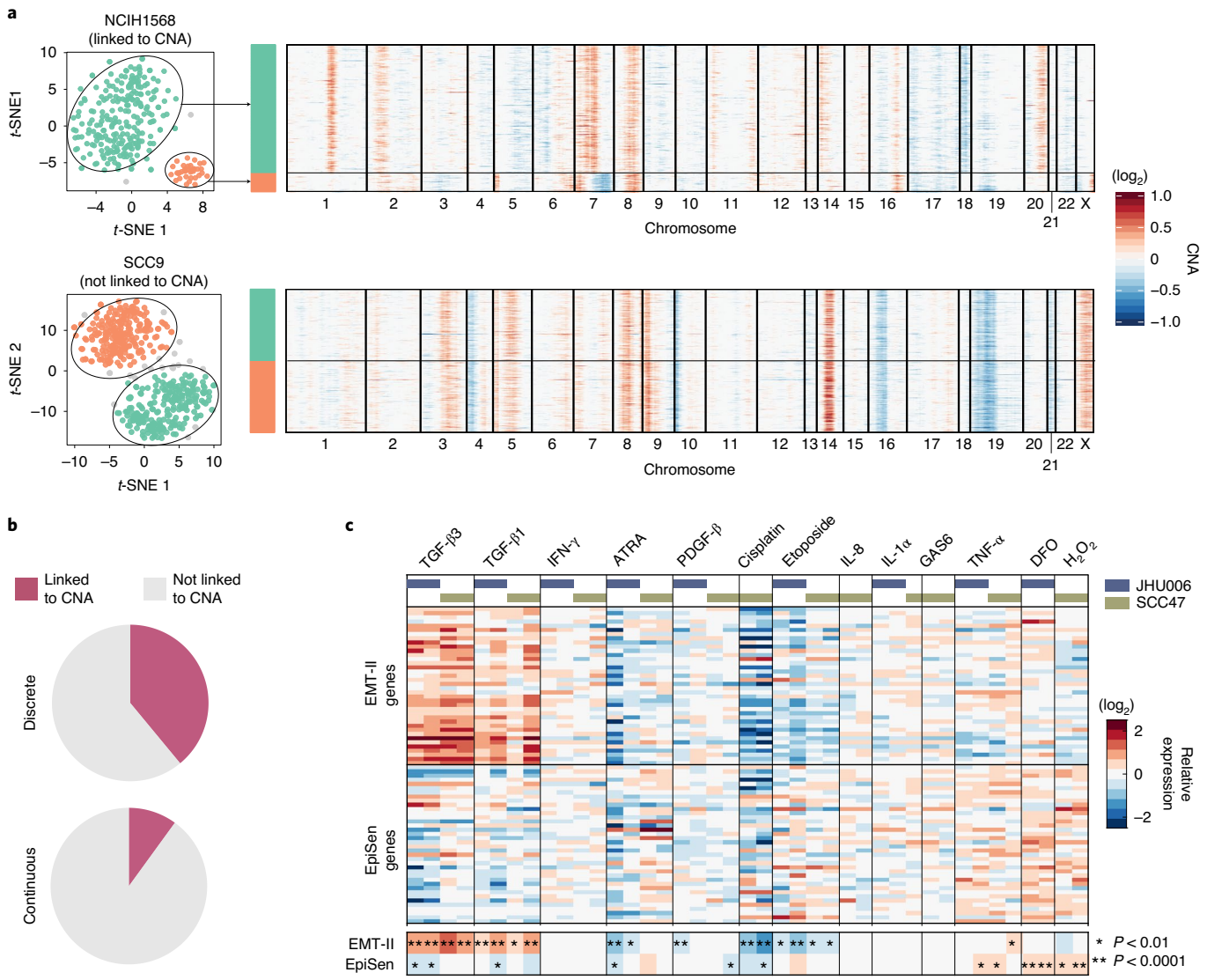


Fig. 6 | Genetic and microenvironmental factors partially explain expression heterogeneity. a, Representative cell lines showing the association (top) or lack thereof (bottom) between discrete subpopulations and CNA-based subclones. Left, *t*-SNE plots showing discrete subpopulations identified using DBSCAN (as in Fig. 2b and Extended Data Fig. 1b). Right, heatmaps depicting inferred CNAs ordered according to the expression-based clusters. **b**, Percentage of discrete (top) and continuous (bottom) heterogeneity programs that were associated with genetic subclones. For discrete programs, associations were assessed by comparing the assignment of cells to CNA subclones and to expression-based subpopulations ($P < 0.001$, two-sided Fisher's exact test); for continuous programs, we compared NMF cell scores between different clones ($P < 0.001$, two-sided *t*-test or one-way ANOVA). **c**, The main heatmap depicts relative expression of EpiSen and EMT-II program genes following multiple perturbations in SCC47 and JHU006 cells. The smaller heatmap at the bottom shows the average expression values for the EMT-II and EpiSen genes, and asterisks denote significant up- or downregulation (by two-sided *t*-test; *P* values are indicated in the figure).

Proliferation and dynamics of EpiSen subpopulations. We selected two HNSCC cell lines (JHU006 and SCC47) with high variability of the EpiSen and EMT-II RHPs for further analysis. EpiSen-high and EpiSen-low subpopulations of cells could be prospectively isolated (as $AXL^{-}CLDN4^{+}$ and $AXL^{+}CLDN4^{-}$ cells, respectively; Extended Data Fig. 7a), with a difference in the expression of the EpiSen program of ~12-fold (Fig. 5c). We noted that EpiSen-low cells were only minimally enriched for the EMT-II RHP; they were used as a negative control for the EpiSen RHP. The EpiSen-high subpopulation was enriched for the G0/G1 phases, consistent with lower levels of proliferation (Fig. 5d and Extended Data Fig. 7b). Nevertheless, this subpopulation still contained cells in the S and G2/M phases, similar to its in vivo counterpart (Extended Data Fig. 7c), and did not stain for the classical senescence marker SA-β-gal (data not shown). These results suggest that the EpiSen program represents an incomplete or reversible cell cycle arrest, consistent with previous studies in cancer cells[32].

The proportions of EpiSen-high and EpiSen-low sorted subpopulations began to shift by 1 week of culture, and each subpopulation returned to the presorting distribution of cellular states by 4 weeks, suggestive of cellular transitions (Fig. 5e and Extended Data Fig. 7d). This distribution of cellular states was stably maintained in culture, suggesting the existence of a steady state, which was maintained by balancing proliferation (favoring EpiSen-low cells) with cellular transitions (favoring EpiSen-high cells). These results indicate that the EpiSen program is dynamically regulated, although we cannot determine whether cellular transitions occur only from EpiSen-low to EpiSen-high cells or in both directions.

RHP regulation by genetics and tumor microenvironment. Expression heterogeneity could be driven by genetic or non-genetic mechanisms. To search for the contribution of genetic heterogeneity, we identified large-scale copy number aberrations (CNAs) in each cell, based on average expression levels in windows of 100 genes around each locus[3–8] (Supplementary Fig. 2). CNA patterns allowed for the robust identification of multiple genetic subclones in 26% (58 of 198) of the cell lines, based on the gain or loss of chromosomes (or chromosome arms) that was restricted to subsets of cells (Methods)[33]. Next, we compared the assignment of cells to CNA-based genetic subclones with their patterns of expression heterogeneity. Among the discrete expression-based clusters, 39% were significantly associated with specific genetic subclones, suggesting a genetic basis for these cases of expression heterogeneity (Fig. 6a,b; $P<0.001$, two-sided Fisher's exact test). In contrast, only 8% of the continuous NMF programs were significantly associated with genetic subclones (Fig. 6b; $P<0.001$, two-sided t-test). This analysis likely underestimates the contribution of genetic heterogeneity, as it relies on CNAs for subclone identification. However, it suggests that genetic heterogeneity contributes primarily to discrete clusters, while the continuous programs of heterogeneity may primarily reflect cellular plasticity, consistent with the established plasticity of EMT and the dynamics of the EpiSen program.

Next, we examined the induction of these programs by soluble factors of the tumor microenvironment and related perturbations. The most dynamic programs were EMT-II and EpiSen, which responded in opposite ways to several of the perturbations (Fig. 6c and Extended Data Fig. 8a). As expected, transforming growth factor (TGF)-β1 and TGF-β3 caused upregulation of the expression of EMT-II genes and increased migration in a wound-healing assay (Extended Data Fig. 8b,c). Interestingly, treatment with TGF-β also caused downregulation of the expression of EpiSen genes, underscoring the potential interplay between EpiSen and EMT, consistent with other analyses (Fig. 4d,f and Extended Data Fig. 4e) and with our prior findings that EMT-high cells were enriched at the invasive edge, while EpiSen-high cells were enriched at the core of tumors[5]. Tumor cores are often associated with increased hypoxia, which suggests a potential mechanism for the spatial enrichment of senescent cells. Accordingly, the hypoxia mimetic desferrioxamine (DFO) induced the expression of the EpiSen program. A similar effect was observed upon hydrogen peroxide treatment, consistent with oxidative stress being a potent inducer of senescence[34] (Fig. 6c). In sum, the EpiSen and EMT-II programs reflect cellular plasticity that exists in certain cell lines, even in the absence of perturbations, and the native tumor microenvironment, but these programs are further induced by stresses and secreted factors.

Coexisting subpopulations differ in drug sensitivity. An important implication of cellular diversity in cancer is the possibility that distinct subpopulations of cells respond differently to treatments, facilitating treatment failure and recurrence. Thus, we compared the sensitivities of EpiSen-high and EpiSen-low subpopulations sorted from each of the two selected model cell lines (Fig. 7a). We initially screened 2,198 bioactive compounds using a CellTiter-Glo-based viability assay (Extended Data Fig. 9a–c). Compounds ($n=248$) with differential killing or the ability to kill both subpopulations (≤10% viability) were selected for a secondary screen performed in duplicate for each cell line (Fig. 7b). The secondary screen identified 113 compounds with differential killing of the subpopulations in at least one cell line (Supplementary Table 9).

Of the hits resulting in preferential sensitivity of EpiSen-high cells, >40% were shared by both cell lines. This fraction of shared hits further increased to 71.4% when considering the targets of compounds rather than the exact compounds, highlighting consistent vulnerabilities of EpiSen-high cells. Fourteen compounds causing differential sensitivities, including five shared hits and nine that were specific to one cell line, were analyzed by full dose–response experiments (Fig. 7c, Extended Data Fig. 9d and Supplementary Table 10). All five of the shared compounds and five of the nine cell-line-specific compounds (56%) caused significant differential sensitivity, as in the secondary screen ($P<0.05$, paired t-test).

EpiSen-high cells were more sensitive to the senolytic compound ABT-737 (ref. [35]), as expected, and to multiple inhibitors of epidermal growth factor receptor (EGFR), AKT, phosphatidylinositol-3-OH kinase (PI(3)K), DNA-dependent protein kinase (DNA-PK), insulin-like growth factor 1 receptor (IGF1R) and Janus kinase (JAK) (Fig. 7b). Several of these targets (DNA-PK, IGF1R and AKT) converge on the repair of double-strand breaks, forming part of the DNA repair machinery[36,37]. The PI(3)K–AKT axis is hyperactivated in HNSCC, and resistance to PI(3)K inhibition in HNSCC is AXL dependent[38]. Accordingly, EpiSen-high cells (which are defined by low AXL expression) were more sensitive to inhibitors of PI(3)K and AKT, as well as those of EGFR and IGF1R that signal via the PI(3)K–AKT axis.

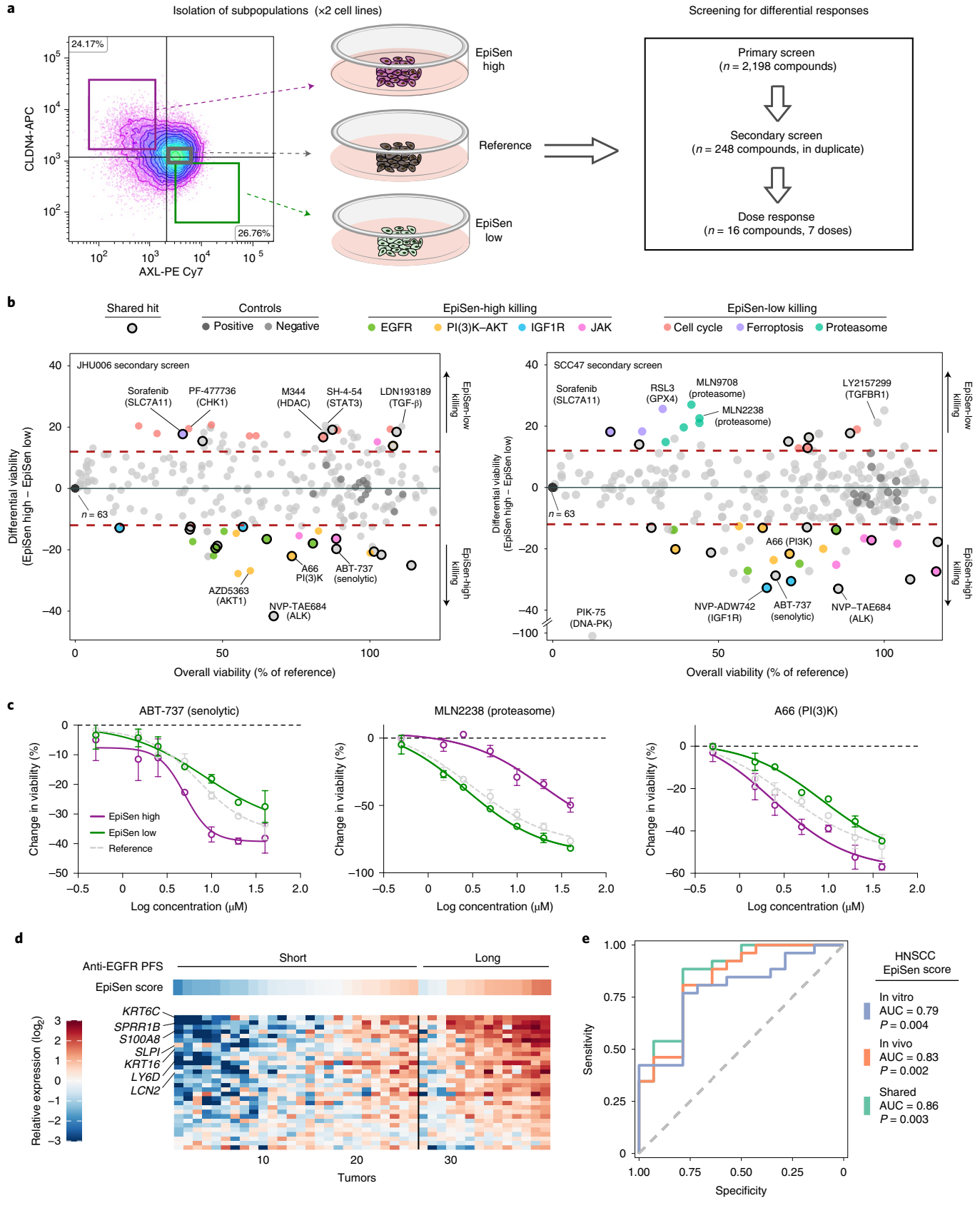
EpiSen-low cells were more sensitive to inhibitors of cell cycle regulators (cyclin-dependent kinases (CDKs), checkpoint kinase 1 (CHK1) and topoisomerase), consistent with their increased levels of proliferation (Fig. 7b). In SCC47 but not in JHU006 cells, EpiSen-low cells were also more sensitive to multiple proteasome inhibitors and to multiple drugs that induce ferroptosis.

In sum, EpiSen-high and EpiSen-low cells are associated with different vulnerabilities that are largely consistent across two model cell lines. In addition to these different sensitivities, nine compounds killed both subpopulations (viability ≤10%) in both of the cell lines (Supplementary Table 9), including disulfiram (Antabuse), which was recently proposed as a potential HNSCC therapy[39,40].

EpiSen is predictive of clinical drug response. The increased sensitivity of EpiSen-high cells to multiple EGFR inhibitors captured our interest, as cetuximab is routinely used for the treatment of patients with HNSCC[41]. Most patients with recurrent or metastatic

Fig. 7 | Coexisting cellular states differ in drug sensitivity. a, Experimental scheme for drug screening. Three subpopulations were isolated by flow cytometry and subjected to a primary screen, a secondary screen and a dose–response analysis of selected hits. **b**, Viability of the reference population (x axis) and differential viability of the EpiSen-high versus EpiSen-low populations (y axis) upon treatment with 248 compounds tested in the secondary screen, in JHU006 (left) and SCC47 (right) cells, averaged over two replicates. Dashed lines represent thresholds for differential sensitivity (as described in the Methods). Selected hits and controls are colored by target as specified in the top legends. **c**, Dose–response curves of three selected compounds in SCC47 cells measured in duplicate at seven concentrations, presented as the change in viability relative to vehicle controls. Error bars represent the s.d.; data points represent the means of replicates. **d**, A heatmap showing the expression of EpiSen genes shared between the HNSCC cell lines (in vitro) and tumors (in vivo) (Extended Data Fig. 10) in bulk pretreatment samples of 40 patients with recurrent or metastatic HNSCC, stratified into those with short and long PFS following treatment with cetuximab and platinum-based chemotherapy. The top panel shows the corresponding EpiSen scores. Genes are ordered by differential expression ($\log_2$(fold change)), comparing patients with short and long PFS, and tumors are ordered within each group according to the EpiSen score. Selected genes are labeled. **e**, Receiver operating characteristic (ROC) curves for predicting patients with long versus short PFS following cetuximab treatment. Curves depict the predictive power of three potential HNSCC EpiSen signatures (in vitro, in vivo and shared). P values were calculated for each signature separately using multivariate logistic regression, correcting for relevant clinicopathological features.

HNSCC progress shortly after cetuximab treatment, combined with platinum-based chemotherapy, but a minority of patients have long progression-free survival (PFS). To examine the potential relevance of EpiSen in clinical response to cetuximab, we examined bulk pretreatment transcriptome data of 40 patients with recurrent or metastatic HNSCC, stratified by PFS following cetuximab treatment[41]. Twenty-six patients had short PFS (PFS < 5.6 months), while 14 patients had long PFS (PFS > 12 months).



Consistent with our in vitro observations, bulk EpiSen scores, a proxy for the abundance of EpiSen cells, were significantly higher in patients with long PFS than in those with short PFS and predicted patient responses with an area under the curve (AUC) of 0.86 (Fig. 7d,e and Extended Data Fig. 10). Notably, the predictive power of EpiSen was comparable between the RHP for HNSCC cells in vitro defined in this work and the in vivo program defined previously[5] and was slightly higher for the genes shared by the two programs (Fig. 7e).

Discussion

We identified 12 RHPs (2 for cell cycle and 10 others), 9 of which were highly similar to programs of heterogeneity observed within tumors, highlighting their importance and indicating that they were retained even in the absence of a native microenvironment. We suggest that the relevance of RHPs is directly derived from their recurrence. In the same way that cancer genetics has focused on recurrent mutations, based on the premise that they are drivers of tumorigenesis, we propose that cancer transcriptomics should focus on recurrent programs, as they may drive important cancer phenotypes, such as drug resistance and metastasis. Extending this analogy to therapeutics, we envisage that, while current targeted therapies attempt to reverse the effects of recurrent oncogenic mutations, future therapies may also be targeted at recurrent programs associated with proliferation, drug resistance or metastasis.

The continuous pattern of such RHPs contrasts with the discrete nature of genetic heterogeneity. Accordingly, we observed dynamic plasticity of the EpiSen program and found only limited associations with genetic subclones (albeit identified only by inferred CNAs). Thus, cancer cells may harbor variability through two largely distinct processes, that is, genetic and non-genetic mechanisms, both of which may contribute to drug resistance and tumor progression. We speculate that, by focusing on recurrent patterns of heterogeneity, our analysis highlights non-genetic plasticity, as this form of variability tends to be shared across cell lines, while genetic forms of variability tend to be unique to each cell line.

Careful examination of the recurrent in vitro programs highlights their consistency with in vivo tumor programs, but also the divergence from their developmental 'normal' counterparts. During development and wound healing, both EMT and senescence are associated with precise phenotypes and well-defined regulators. Yet in the context of tumors and cancer cell lines, we observe only partial phenotypes and limited dependence on these regulators. The EMT-like profiles that we observe include many EMT-related genes and are associated with increased migration, but do not involve other EMT hallmarks, such as the loss of epithelial markers, a change in morphology or high expression of core EMT transcription factors. Similarly, EpiSen-high cells resemble the senescence response of keratinocytes and lung bronchial cells, are associated with reduced proliferation and possess markers of SASP, yet they retain some proliferative capacity, do not express high levels of p16 and p21 and do not stain with SA-β-gal. This is consistent with findings from previous studies that showed evidence for incomplete and reversible senescence programs in cancer[42,43]. We hypothesize that cancer cells often activate partial or distorted programs, possibly not through canonical developmental mechanisms, in a context-dependent manner. This could contribute to the difficulties in resolving long-standing debates in the cancer field about the role of EMT and senescence, which is often evaluated by assessing the activity of developmental regulators and markers that may fail to detect certain partial programs. Comprehensive single-cell profiling helps to detect such partial programs that vary in their magnitude across cells.

Multiple RHPs hold potential clinical relevance. The EpiSen program is associated with distinct responses to several drugs and, most notably, with patient responses to cetuximab. EMT-II is highly consistent with an HNSCC program[5] that is localized to the invasive edge, is predictive of nodal metastasis and is evaluated for clinical decision-making[44]. EMT-I is similar to multiple melanoma programs, variably designated as 'invasive' (refs. [45,46]), 'AXL-high' or 'MITF-low' (refs. [47,48]) or 'resistance' (ref. [22]), but invariably involves the upregulation of EMT-related genes, the downregulation of skin pigmentation genes (for example, *MITF*) and resistance to targeted therapies. The p53-dependent senescence program (RHP 6) is significantly correlated with the response to the p53-activating drug nutlin-3a. Notably, in nutlin-sensitive cell lines, only a minority of cells express this program before treatment, while most or all cells appear to express it after treatment[49], suggesting that rare senescent cells may serve as a biomarker for p53 activity that implies sensitivity to nutlin-3a or similar treatments. Finally, expression of the IFN-response program (RHP 4) by subpopulations of cancer cells[19] may influence immune cells in the tumor microenvironment and the response to immunotherapies. For example, recent work demonstrated opposing functions of the IFN response by cancer and immune cells through complex cancer–immune cross-talk[50].

With the advent of single-cell genomics, cellular heterogeneity is now being characterized in various clinical contexts. However, the ability to model ITH is a prerequisite for deeper understanding of the mechanisms that govern such heterogeneity. Here we described the landscape of cellular diversity across 198 cell lines, highlighting particular models that recapitulate programs of heterogeneity observed in human tumors. Further studies of these programs and model systems will provide a better understanding of ITH and may help to transform this understanding into novel treatment strategies that exploit ITH.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-00726-6.

Received: 10 February 2020; Accepted: 25 September 2020;
Published online: 30 October 2020

References

1. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* **27**, 15–26 (2015).
2. Chaffer, C. L., San Juan, B. P., Lim, E. & Weinberg, R. A. EMT, cell plasticity and metastasis. *Cancer Metastasis Rev.* **35**, 645–654 (2016).
3. Filbin, M. G. et al. Developmental and oncogenic programs in H3K27M gliomas dissected by single-cell RNA-seq. *Science* **360**, 331–335 (2018).
4. Patel, A. P. et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science* **344**, 1396–1401 (2014).
5. Puram, S. et al. Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell* **171**, 1611–1624 (2017).
6. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
7. Tirosh, I. et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature* **539**, 309–313 (2016).
8. Venteicher, A. S. et al. Decoupling genetics, lineages, and microenvironment in IDH-mutant gliomas by single-cell RNA-seq. *Science* **355**, eaai8478 (2017).
9. Chung, W. et al. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nat. Commun.* **8**, 15081 (2017).
10. Kim, K. T. et al. Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. *Genome Biol.* **17**, 80 (2016).
11. Li, H. et al. Reference component analysis of single-cell transcriptomes elucidates cellular heterogeneity in human colorectal tumors. *Nat. Genet.* **49**, 708–718 (2017).
12. Lambrechts, D. et al. Phenotype molding of stromal cells in the lung tumor microenvironment. *Nat. Med.* **24**, 1277–1289 (2018).
13. Barretina, J. et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
14. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).

15. Yu, C. et al. High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat. Biotechnol.* **34**, 419–423 (2016).
16. Kang, H. M. et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
17. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
18. Neftel, C. et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell* **178**, 835–849 (2019).
19. Izar, B. et al. A single-cell landscape of high-grade serous ovarian cancer. *Nat. Med.* **26**, 1271–1279 (2020).
20. Chen, Q., Sun, L. & Chen, Z. J. Regulation and function of the cGAS–STING pathway of cytosolic DNA sensing. *Nat. Immunol.* **17**, 1142–1149 (2016).
21. Kondo, T. et al. DNA damage sensor MRE11 recognizes cytosolic double-stranded DNA and induces type I interferon by regulating STING trafficking. *Proc. Natl Acad. Sci. USA* **110**, 2969–2974 (2013).
22. Shaffer, S. M. et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. *Nature* **546**, 431–435 (2017).
23. Aceto, N. et al. Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* **158**, 1110–1122 (2014).
24. Hernandez-Segura, A. et al. Unmasking transcriptional heterogeneity in senescent cells. *Curr. Biol.* **27**, 2652–2660 (2017).
25. Jang, D. H. et al. A transcriptional roadmap to the senescence and differentiation of human oral keratinocytes. *J. Gerontol. A Biol. Sci. Med. Sci.* **70**, 20–32 (2015).
26. Musiani, D. et al. PRMT1 is recruited via DNA-PK to chromatin where it sustains the senescence-associated secretory phenotype in response to cisplatin. *Cell Rep.* **30**, 1208–1222 (2020).
27. Yang, L., Fang, J. & Chen, J. Tumor cell senescence response produces aggressive variants. *Cell Death Discov.* **3**, 17049 (2017).
28. Pawlikowski, J. S. et al. Wnt signaling potentiates nevogenesis. *Proc. Natl Acad. Sci. USA* **110**, 16009–16014 (2013).
29. Hanzelmann, S. et al. Replicative senescence is associated with nuclear reorganization and with DNA methylation at specific transcription factor binding sites. *Clin. Epigenetics* **7**, 19 (2015).
30. Basisty, N. et al. A proteomic atlas of senescence-associated secretomes for aging biomarker development. *PLoS Biol.* **18**, e3000599 (2020).
31. Coppe, J. P., Desprez, P. Y., Krtolica, A. & Campisi, J. The senescence-associated secretory phenotype: the dark side of tumor suppression. *Annu. Rev. Pathol.* **5**, 99–118 (2010).
32. Lee, S. & Schmitt, C. A. The dynamic nature of senescence in cancer. *Nat. Cell Biol.* **21**, 94–101 (2019).
33. Ben-David, U. et al. Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325–330 (2018).
34. te Poele, R. H., Okorokov, A. L., Jardine, L., Cummings, J. & Joel, S. P. DNA damage is able to induce senescence in tumor cells in vitro and in vivo. *Cancer Res.* **62**, 1876–1883 (2002).
35. Yosef, R. et al. Directed elimination of senescent cells by inhibition of BCL-W and BCL-X_L. *Nat. Commun.* **7**, 11190 (2016).
36. Bozulic, L., Surucu, B., Hynx, D. & Hemmings, B. A. PKBα/Akt1 acts downstream of DNA-PK in the DNA double-strand break response and promotes survival. *Mol. Cell* **30**, 203–213 (2008).
37. Wong, R. H. et al. A role of DNA-PK for the metabolic gene regulation in response to insulin. *Cell* **136**, 1056–1072 (2009).
38. Elkabets, M. et al. AXL mediates resistance to PI3Kα inhibition by activating the EGFR/PKC/mTOR axis in head and neck and esophageal squamous cell carcinomas. *Cancer Cell* **27**, 533–546 (2015).
39. Park, Y. M. et al. Anti-cancer effects of disulfiram in head and neck squamous cell carcinoma via autophagic cell death. *PLoS ONE* **13**, e0203069 (2018).
40. Shah O'Brien, P. et al. Disulfiram (Antabuse) activates ROS-dependent ER stress and apoptosis in oral cavity squamous cell carcinoma. *J. Clin. Med.* **8**, 611 (2019).
41. Bossi, P. et al. Functional genomics uncover the biology behind the responsiveness of head and neck squamous cell cancer patients to cetuximab. *Clin. Cancer Res.* **22**, 3961–3970 (2016).
42. Beausejour, C. M. et al. Reversal of human cellular senescence: roles of the p53 and p16 pathways. *EMBO J.* **22**, 4212–4222 (2003).
43. Sage, J., Miller, A. L., Perez-Mancera, P. A., Wysocki, J. M. & Jacks, T. Acute mutation of retinoblastoma gene function is sufficient for cell cycle re-entry. *Nature* **424**, 223–228 (2003).
44. Parikh, A. S. et al. Immunohistochemical quantification of partial-EMT in oral cavity squamous cell carcinoma primary tumors is associated with nodal metastasis. *Oral Oncol.* **99**, 104458 (2019).
45. Hoek, K. S. et al. Metastatic potential of melanomas defined by specific gene expression profiles with no BRAF signature. *Pigment Cell Res.* **19**, 290–302 (2006).
46. Verfaillie, A. et al. Decoding the regulatory landscape of melanoma reveals TEADS as regulators of the invasive cell state. *Nat. Commun.* **6**, 6683 (2015).
47. Konieczkowski, D. J. et al. A melanoma cell state distinction influences sensitivity to MAPK pathway inhibitors. *Cancer Discov.* **4**, 816–827 (2014).
48. Muller, J. et al. Low MITF/AXL ratio predicts early resistance to multiple targeted drugs in melanoma. *Nat. Commun.* **5**, 5712 (2014).
49. McFarland, J. M. et al. Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nat. Commun.* **11**, 4296 (2020).
50. Benci, J. L. et al. Opposing functions of interferon coordinate adaptive and innate immune responses to cancer immune checkpoint blockade. *Cell* **178**, 933–948 (2019).

Methods

Cell line pools. We obtained eight previously generated pools of cell lines[15], each containing 24–27 cell lines from diverse cancer types. Cell lines were combined in pools based on growth rates (doubling time). To ensure comparable representation over short-term culturing, each pool was thawed and cultured in RPMI supplemented with 10% FBS for 3 d before scRNA-seq. A ninth custom pool was generated by freshly pooling eight additional head and neck cell lines immediately before scRNA-seq (JHU006, JHU011, JHU029, SCC9, SCC90, SCC25, UM-SCC47, 93-VU-147T). Additionally, two pools were generated for the co-culture control experiment, each consisting of CAL27 (HNSCC), UM-SCC47 (HNSCC), JHU006 (HNSCC), RKO (colon), HCC1954 (breast) and SKMEL2 (melanoma). The first pool was generated by co-culturing all cells for 72 h before profiling; the second pool was generated by combining cells immediately before profiling.

Individual cell line cultures. Human HNSCC cell lines (laryngeal: JHU006, JHU011, JHU029; oropharyngeal: UM-SCC47, SCC90, 93VU147T, oral cavity: SCC9, SCC25) were provided by J.W.R. after confirmation by short tandem repeat analysis. The laryngeal cell lines were grown in RPMI 1640 (Biological Industries). The oropharyngeal and oral cavity cell lines were grown in a 3:1 mixture of Ham's F12:DMEM (Biological Industries). All growth media for HNSCC cell lines were supplemented with 10% FBS (Biological Industries), 1× penicillin–streptomycin and 1× L-glutamine (Biological Industries). Additional human cancer cell lines used for the co-culture control experiment were cultured in either RPMI 1640 (SKMEL2, melanoma; HCC1954, breast) or DMEM (RKO, colon) supplemented with 10% FBS, 1× penicillin–streptomycin and 1× L-glutamine.

Human primary bronchial epithelial cells (PCS-300-010) were purchased from the American Type Culture Collection (ATCC) and grown in Airway Epithelial Cell Basal Medium (ATCC, PCS-300-030) supplemented with the bronchial epithelial cell growth kit (ATCC, PCS-300-040). All cell lines tested negative for mycoplasma by the EZ-PCR Mycoplasma Detection kit (Biological Industries).

Droplet-based scRNA-seq. scRNA-seq libraries were generated using the 10x Genomics Chromium Single Cell 3′ kit version 2 (eight CCLE pools and a custom HNSCC pool) or the 10x Genomics Chromium Single Cell 3′ kit version 3.1 (co-culture control experiment) and the 10x Chromium Controller (10x Genomics), according to the 10x Single Cell 3′ version 2 protocol. Briefly, for each pool, a single-cell suspension was generated in 0.04% BSA in PBS (≥95% viability) and approximately 10,500 single cells were loaded in the Chromium Controller with a targeted recovery of 6,000 cells. cDNA was purified with Dynabeads and amplified by 12 cycles of PCR (98 °C for 45 s, (98 °C for 20 s, 67 °C for 30 s, 72 °C for 1 min) × 12, 72 °C for 1 min). The amplified cDNA was fragmented, end repaired, ligated with index adaptors and size selected with cleanups between each step using the SPRIselect Reagent kit (Beckman Coulter). Quality control of the resulting barcoded libraries was performed with the Agilent TapeStation and by PCR with primers specific to the P5 and P7 sequence (NEBNext Library Quant Kit for Illumina, New England Biolabs).

Bulk RNA-seq by MARS-seq. A bulk adaptation of the massively parallel RNA-seq (MARS-seq) protocol[51] was used to generate RNA-seq libraries for validating the isolation of selected subpopulations by flow cytometry (Fig. 5c) and for expression profiling of HNSCC cell lines following perturbations (Fig. 6c). RNA was isolated from sorted and unsorted cells using the Quick-RNA Microprep kit (Zymo Research). Briefly, 50 ng of input RNA from each sample was barcoded during reverse transcription and pooled. Following cleanup with Agencourt AMPure XP beads (Beckman Coulter), the pooled samples underwent second-strand synthesis and were linearly amplified by T7 in vitro transcription. The resulting RNA was fragmented and converted into a sequencing-ready library by tagging the samples with Illumina sequences during ligation, reverse transcription and PCR. Libraries were quantified by Qubit and TapeStation, as well as by qPCR for *GAPDH* as previously described[51]. Sequencing was done with the Illumina NextSeq 75 Cycle High-Output kit (paired-end sequencing).

Flow cytometry and sorting of cell lines. Sorting of JHU006 and UM-SCC47 cells was performed on a BD FACSMelody running BD FACSChorus version 1.0 using the following antibodies: anti-human AXL PE-Cy7 (eBioscience) at 1:300, anti-human CLDN4 APC (Miltenyi) at 1:200 and anti-human ITGA6/CD49f APC (eBioscience) at 1:200. Gating of positive and negative cells was defined by the unstained control. For sorting, the top 10% of the high population and bottom 10% of the low population were taken.

For EpiSen program dynamics experiments, 200,000 cells of each subpopulation (EpiSen-high, AXL$^-$CLDN4$^+$; EpiSen-low, AXL$^+$CLDN4$^-$; control sort, all single cells) were sorted by a two-step sequential sort (in which the isolated EpiSen-high and EpiSen-low populations were sorted and then subsequently resorted for the same markers to improve purity) and reanalyzed by flow cytometry immediately following the second sort and throughout culturing at weekly intervals for 4 weeks. Analysis was performed using Kaluza Analysis Software version 2.1 (Beckman Coulter). The experiment was performed three times independently. Successful isolation of the EpiSen subpopulation with anti-AXL and anti-CLDN4 antibodies was validated by bulk RNA-seq of sorted cells as described in the 'Bulk RNA-seq by MARS-seq' section above. Similarly, the EMT-II subpopulation was isolated by purifying AXL$^+$ITGA6$^+$ (EMT-II-high) and AXL$^-$ITGA6$^-$ (EMT-II-low) cells for a migration assay.

Cytokine treatment and perturbation of HNSCC cell lines. JHU006 and SCC47 cells were seeded at 50,000 cells per well in 24-well plates in their standard media and treated in duplicate with drug, cytokine or vehicle (0.1% DMSO with 1 μg ml^{-1} BSA) 24 h after seeding. Cells were harvested 24 h after treatment, and RNA was isolated with the Quick-RNA Microprep kit (Zymo Research) for bulk expression profiling using the MARS-seq protocol (above). Treatments included 10 μM all-*trans* retinoic acid (ATRA; Sigma), 25 ng ml^{-1} IFN-γ (Peprotech), 25 ng ml^{-1} TNF-α (Peprotech), 25 ng ml^{-1} PDGF-BB (Miltenyi), 10 nM etoposide (Sigma), 10 ng ml^{-1} TGF-β1 (Peprotech), 10 ng ml^{-1} TGF-β3 (Peprotech), 10 μM cisplatin (Sigma), 200 μM hydrogen peroxide (Sigma), 25 ng ml^{-1} IL-8 (CXCL8; Peprotech), 25 ng ml^{-1} GAS6 (Sino Biological), 50 ng ml^{-1} S100A8–A9 (Sino Biological) and 500 μM DFO (Sigma).

Drug screening and viability assay. The Selleck Bioactive Compound Library (Selleck Chemicals) as well as DMSO-only controls and staurosporine-positive (killing) controls was dispensed into 384-well plates with an Echo 550 Liquid Handler (Labcyte). The drug concentration was 10 μM for the primary screen and 1 μM or 10 μM for the secondary screen (performed in duplicate), depending on the hit category. The purity of the compounds selected for follow-up by dose–response analysis was confirmed by LC–MS (data not shown). For the dose–response curves, a seven-point twofold dilution series with an upper limit of 40 μM was tested in duplicate. EpiSen-high (AXL$^-$CLDN4$^+$) cells, EpiSen-low (AXL$^+$CLDN4$^-$) cells and a third neutral reference population were sorted from JHU006 and SCC47 cell lines, as described above, and sorted subpopulations were seeded into the compound-treated plates in their standard media at a concentration of 15,000 cells per ml (750 cells per well) with a Combi Multidrop (Thermo Fisher). Plates were incubated at 37 °C for 48 h following sorting and compound treatment, and cell viability was determined based on luminescence readings following the addition of CellTiter-Glo (Promega) according to the manufacturer's instructions. Luminescence was measured on a BMG PHERAstar plate reader. Data were normalized in Genedata Screener, in which samples treated with DMSO (vehicle) were defined as the neutral control (that is, 100% viability) and samples without cells were defined as the inhibitor control (that is, 0% viability). Compound-centric data were visualized in CDD Vault from Collaborative Drug Discovery.

Processing of scRNA-seq data. Cell barcode filtering, alignment of reads and UMI counting were performed using Cell Ranger 3.0.1 (10x Genomics). Expression levels were quantified as $E_{i,j} = \log_2(1 + \mathrm{CPM}_{i,j}/10)$, in which counts per million $(\mathrm{CPM})_{i,j}$ refers to $10^6 \times \mathrm{UMI}_{i,j}/\mathrm{sum}(\mathrm{UMI}_{1...n,j})$, for gene i in sample j, with n being the total number of analyzed genes. The average number of UMIs detected per cell was less than 100,000; thus, CPM values were divided by 10 to avoid inflating the differences between detected ($E_{i,j} > 0$) and undetected ($E_{i,j} = 0$) genes as previously described[7]. For each cell, we quantified the number of detected genes as a proxy for sample quality. We conservatively retained cells with a number of detected genes ranging from 2,000 to 9,000. When analyzing cell lines individually, we only considered genes expressed at high or intermediate levels ($E_{i,j} > 3.5$) in at least 2% of cells, yielding an average of 6,758 genes analyzed per cell line. Values were then centered by cell line to define relative expression values, by subtracting the average expression of each gene i across all k cells: $Er_{i,j} = E_{i,j} - \mathrm{average}(E_{i,1...k})$, where Er represents relative expression values. When analyzing cell lines collectively, we selected the 7,000 most highly expressed genes across all cell lines, resulting in a minimum average expression of 12 CPM. Values were centered by subtracting the average expression across all 53,513 cells analyzed: $ER_{i,j} = E_{i,j} - \mathrm{average}(E_{i,1...\ 53,513})$, for gene i in sample j.

Processing of bulk RNA-seq data. Reads were aligned to the GRCh38/hg38 human genome using Bowtie, and expression values were quantified using RSEM. Data are presented as $E_{i,j} = \log_2(\mathrm{TPM}_{i,j} + 1)$, where $\mathrm{TPM}_{i,j}$ refers to transcripts per million for gene i in sample j, as calculated by RSEM[52].

Systematic characterization of transcriptional heterogeneity. For each of the 198 cell lines that passed quality control, we applied two distinct approaches to identify discrete and continuous patterns of expression heterogeneity. First, to identify discrete (highly distinct) subpopulations within cell lines, we used nonlinear dimensionality reduction (*t*-SNE) followed by DBSCAN, which assumes that clusters are contiguous regions with high cell density. *t*-SNE was applied to each cell line individually using relative expression values (*Er*) and a perplexity of 30. To identify dense regions, DBSCAN classifies each point according to a minimum points (minPts) threshold, defined as the minimum number of neighbors within a user-defined radius (eps) around core points. To optimize this parameter selection, we tested the ability of DBSCAN to correctly distinguish cells from two distinct cell lines. We combined cells from two different cell lines and tested the classification accuracy of DBSCAN using different eps values (0.6–3) and minPts thresholds (5 and 10). DBSCAN classification was evaluated using two-sided Fisher's exact

test and considered correct if $P<0.001$. This procedure was repeated 1,000 times, and in each iteration, we randomly selected the cell lines, the total number of cells (56–1,990) and the proportion of cells selected from each cell line (2–98% of total). The parameter combination yielding the highest rate of correct classification (eps = 1.8, minPts = 5) was used for further analyses. We also applied DBSCAN with additional, less stringent eps values (1.2 and 1.5) to show the robustness of the results. To define gene signatures that characterized the discrete subpopulations that were identified, gene expression of cells in a given cluster was compared to that of all other cells within the same cell line using a two-sided *t*-test. Genes with fold change ≥2 and $P<0.001$ were selected, and the top 50 (by fold change) were defined as the gene signature. Clusters containing more than 90% of the cells of a given cell line were excluded from this analysis.

Second, each cell line was analyzed separately using NMF to identify both discrete and continuous programs of expression heterogeneity. NMF was applied to *Er*, by transforming all negative values to zero, as previously described[5]. We performed NMF with the number of factors *k* ranging from 6 to 9 and initially defined expression programs as the top 50 genes (by NMF score) for each *k*. For each cell line, we sought robust expression programs by selecting those with an overlap of at least 70% (35 of 50 genes) with a program obtained using a different *k* value. To avoid redundancies, only one program was selected from each set of overlapping programs, based on having the highest overlap with an NMF program identified in another cell line. To determine which programs reflected discrete patterns of expression variability for each cell line, we compared the classification of cells by DBSCAN and the assignment of cells to the most highly expressed NMF program using two-sided Fisher's exact test ($P<0.001$ was considered significant). The association between programs and technical artifacts was inspected for each cell line, by calculating the Pearson correlation coefficient between the number of genes detected in a cell (that is, complexity) and the respective NMF scores. This approach identified a cluster of NMF programs (based on 50 minus the number of overlapping genes across expression programs) that share negative correlations with complexity. This cluster appeared to reflect technical artifacts (also based on manual inspection and the inclusion of many mitochondrial genes and pseudogenes) and was excluded from further analysis.

To identify RHPs across cell lines, we compared expression programs derived from DBSCAN and NMF, separately, by hierarchical clustering, using 50 minus the number of overlapping genes as a distance metric. Given the high number of NMF programs, clustering was restricted to programs with at least a minimum overlap of 20% (10 of 50 genes) with a program observed in another cell line. Twelve clusters (that is, metaprograms) were defined by manual inspection of the hierarchical clustering results. For each cluster of programs, an RHP was then defined as all genes included in at least 25% of the constituent programs. We assessed the enrichment of RHP signatures with Gene Ontology terms (C5:BP/CC MSigDB[53]) using a hypergeometric test (FDR-adjusted $P<0.05$ was considered significant). We used the same approach to compare the EpiSen RHP with senescence-like and SASP programs previously described in the literature ($P<0.01$ was considered significant; Supplementary Table 8). For SASP analysis, only secreted factors were considered.

Defining program scores in each cell. Program scores were calculated for each cell individually to evaluate the degree to which they expressed a given RHP. Cells with higher complexity (defined as having a larger number of genes detected) would be expected to have higher cell scores for any gene set. To account for this effect, for each gene set analyzed, we created a control gene set to be used to calculate a normalization factor, as previously described[7]. Control gene sets were selected in a way that ensured a similar distribution of expression levels as that in the input gene set. First, all analyzed genes were ordered by mean expression across all cell lines and partitioned into 75 bins. Next, for each gene in a given gene set, 100 genes were randomly selected from the same expression bin. Finally, given an input set of genes (G_j), we defined a score, $SC_j(i)$, for each cell *i*, as the mean relative expression of the genes in G_j. A similar cell score was then calculated for the respective control gene set and subtracted from the initial cell scores: $SC_j(i) = \text{average}(ER(G_j,i)) - \text{average}(ER(G_j^{cont},i))$.

Comparison of in vitro and in vivo expression heterogeneity programs. In vivo RHPs were defined previously[5,6,18] or generated using published scRNA-seq data[5,6,9,12] with the NMF-based strategy described above applied to the malignant cells in each dataset separately. Initial comparisons with in vitro RHPs were performed using hypergeometric tests. Next, two additional approaches were used. First, we calculated the average similarity (Jaccard index) and single-cell score (that is, SC) correlation between each in vivo RHP and the in vitro programs composing each RHP. In vitro programs were defined as the top 50 NMF-scoring genes as previously described. We assessed statistical significance by permuting in vitro programs 100 times and considered a confidence threshold of 99.9% as significant. To generate permuted gene sets, we first ordered all genes by mean expression across all cell lines and partitioned them into 75 bins. Each gene in a given gene set was then replaced by a randomly selected gene from the same expression bin. Second, shared patterns of expression heterogeneity were highlighted by analyzing tumors and cell lines simultaneously. To this end, we selected melanoma tumors or cell lines harboring the EMT-I and SkinPig RHPs and HNSCC tumors or cell lines harboring the EMT-II and EpiSen RHPs and combined the tumors and cell lines of each cancer type into joint datasets. Expression levels ($\log_2(CPM_{i,j}/10+1)$) of each individual dataset were mean centered per gene before the cells were combined. PC analysis was performed using the 4,500 most highly expressed genes in each joint dataset, and we identified which PCs correlated strongly ($r>0.35$ or $r<-0.35$) with single-cell scores for the respective RHPs of interest.

Defining program variability in each cell line. To evaluate the degree of heterogeneity of RHPs in each cell line, the variability of cell scores was examined. First, given a program *j* and cell line *i*, we defined program variability, $PV_j(i)$, by first ranking cells according to the program score (SC_j) and comparing the average signal between the top and bottom 10% of cells: $PV(j) = \text{average}(SC_j(\text{top } 10\%)) - \text{average}((SC_j(\text{bottom } 10\%))$. Next, to control for the potential association between the mean and variability of program scores, we applied a local polynomial regression with a smoothing span of 0.8 to infer the relationship between program variability and the mean in each cell line and used the residuals of the model as the corrected program variability score: $PV_c(j) = PV(j) - RG(\text{mean}(SC_j))$, where RG represents the local regression model for PV based on the average program scores, SC.

CNA estimation. Initial values (CNA_0) were estimated by first sorting the analyzed genes by their chromosomal location and calculating a moving average of *ER* with a sliding window of 100 genes as previously described[7]. To avoid considerable impact of any particular gene on the moving average, in this analysis, we limited relative expression values to (−3, 3). To define proper CNA reference values for use as the baseline, we downloaded gene-level copy number data (Affymetrix SNP6.0 arrays, $\log_2(\text{copy number}/2)$) from the CCLE portal (https://portals.broadinstitute.org/ccle) and calculated for each cell line the average copy number signal by chromosome arm. Next, for each chromosome arm, we selected a set of reference cell lines, defined as those presenting an average copy number signal ranging from −0.2 to 0.2. For a given CNA window, in a given chromosome arm, we then calculated the average CNA estimates of the respective reference cell lines and defined the minimum (BaseMin) and maximum (BaseMax) values obtained as the lower and upper baseline limits. The final CNA estimate of cell *i* at position *j* was defined as:

$$CNA_f(i,j) = \begin{cases} CNA_0(i,j) - BaseMax(j), & \text{if } CNA_0(i,j) > BaseMax(j) + 0.1 \\ CNV_0(i,j) - BaseMin(j), & \text{if } CNA_0(i,j) < BaseMin(j) - 0.1 \\ 0, & \text{if } BaseMin(j) - 0.1 < CNA_0(i,j) < BaseMax(j) + 0.1 \end{cases}$$

Detection of CNA subclones within cell lines. To confidently identify CNA-based subclones, we focused on CNAs encompassing whole chromosome arms, as these are more reliably inferred than focal CNAs. We reasoned that the presence of multiple subclones in a single cell line would be reflected in a multimodal distribution of CNA signal for at least one chromosome arm across cells. Thus, we first calculated for each cell and each chromosome arm the average CNA_0 estimate across all loci in the chromosome arm. Next, we fitted arm-level CNA values to a bimodal Gaussian mixture and calculated the probability of each cell belonging to each component. Models were fitted using expectation–maximization, as implemented by the R function mclust. A cell line was then defined as having subclones if, for at least one chromosome arm, a minimum of 20 cells were classified into a second mode with at least 99% confidence. For cell lines harboring one chromosome arm with a bimodal distribution, we defined two clones corresponding to the two modes. For cell lines harboring multiple chromosome arms with a bimodal distribution, we considered all combinations of modes with at least five cells. In both cases, we only considered cells assigned to modes with at least 90% confidence.

Software packages. Data analysis was performed in R (versions 3.5.3 and 3.6.3) with the following packages: Rtsne (version 0.15), dbscan (version 1.1-4), NMF (version 0.21.0), mclust (version 5.4), glmnet (version 2.0-16), kernlab (version 0.9-26), SCENIC (version 1.1.1), ggplot2 (version 3.1.0), GENIE3 (version 1.9) and RcisTarget (version 1.9). Additional software used included Cell Ranger (version 3.0.1, 10x Genomics), Bowtie (version 1.2.2), RSEM (version 1.3.1), GraphPad Prism (version 8), Kaluza Flow Cytometry Analysis software (version 2.1, Beckman Coulter) and Genedata Screener (version 12).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw and processed scRNA-seq data are available through the Broad Institute's single-cell portal (SCP542) and at the Gene Expression Omnibus (GEO) (accession number GSE157220). Publicly available databases used in our analysis included the DepMap portal (18q3 data release; https://depmap.org/), the CCLE portal (https://portals.broadinstitute.org/ccle), the CTD2 portal (https://ocg.cancer.gov/programs/), GTRD database version 20.06 (http://gtrd.biouml.org) and MSigDB version 7.0 (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp).

Code availability

R code for reproducing the analyses shown in the main figures is available at https://github.com/gabrielakinker/CCLE_heterogeneity. Additional code related to extended data and supplementary figures is available upon request from the corresponding author. Code used for the assignment of cells to reference expression profiles is available at https://github.com/broadinstitute/single_cell_classification.

References

51. Keren-Shaul, H. et al. MARS-seq2.0: an experimental and analytical pipeline for indexed sorting combined with single-cell RNA sequencing. *Nat. Protoc.* **14**, 1841–1862 (2019).
52. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
53. Liberzon, A. A description of the Molecular Signatures Database (MSigDB) web site. *Methods Mol. Biol.* **1150**, 153–160 (2014).

Acknowledgements

This work was supported by funding from the Israel Science Foundation (I.T.), the Zuckerman STEM leadership program (I.T.), a Mexican Friends New Generation grant (I.T.), the Rising Tide Foundation (I.T.), the AMN Fund for the Promotion of Science, Culture and Arts in Israel (I.T.), the Estate of Dr. David Levinson, the Dr. Celia Zwillenberg-Fridman and Dr. Lutz Zwillenberg Career Development Chair (I.T.), the Sao Paulo Research Foundation (FAPESP) (fellowships 2014/27287-0 and 2017/24287-8 (G.S.K.)), the Clore Foundation Postdoctoral Fellowship (A.C.G.), the Klarman Cell Observatory (A.R.), the Howard Hughes Medical Institute (A.R.), the National Cancer Institute (K08CA237732; S.V.P.), a V Foundation V Scholars Award (S.V.P.), a Cancer Research Foundation Young Investigator Award (S.V.P.) and the Dorris Duke Fund to Retain Clinical Scientists (S.V.P.).

Author contributions

G.S.K., A.C.G. and I.T. conceived and designed the study. J.A.R., S.A.B., C.C.M., B.K., J.W.R. and S.V.P. provided cell lines and pools. A.C.G., R.T., M.S.C. and C.R. performed scRNA-seq experiments. J.M.M., A.W., A.T. and I.T. assigned cells to cell lines. G.S.K., A.C.G. and I.T. analyzed the scRNA-seq data and interpreted the results. A.C.G., R.T. and Z.O. designed and performed follow-up experiments with selected cell lines. A.C.G., R.T., A.P. and H.B. performed drug screening. A.C.G. and H.K.-S. performed bulk RNA-seq experiments. A.C.G. and I.T. analyzed the drug screening and bulk RNA-seq data. V.K. was consulted on senescence. O.R.-R., A.R. and P.A.C.M.F. provided resources and supervision. I.T. supervised the work. G.S.K., A.C.G. and I.T. wrote the manuscript with input from all other authors.

Competing interests

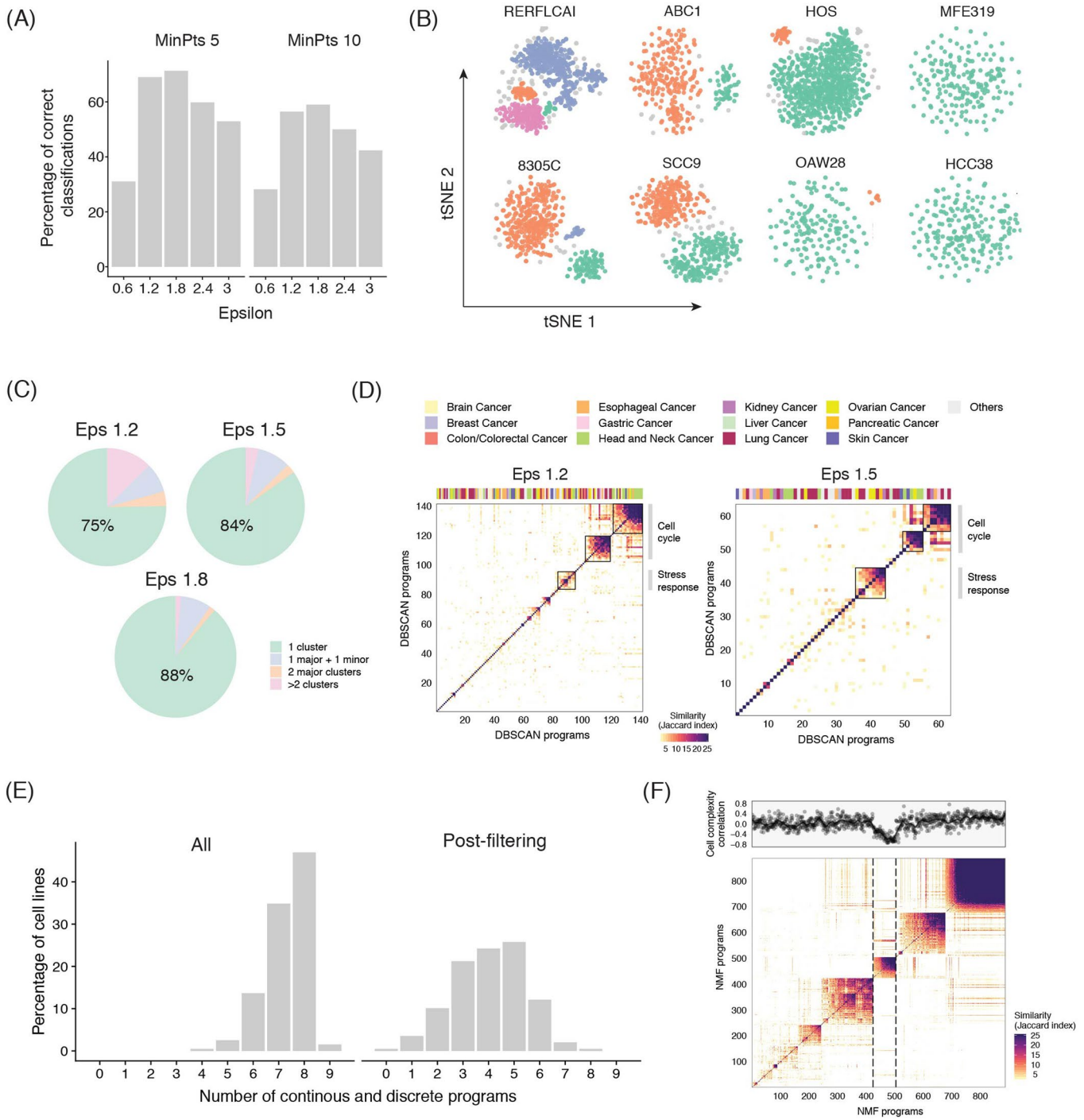
A.R. is a cofounder and equity holder of Celsius Therapeutics, an equity holder of Immunitas and was an SAB member of Neogene Therapeutics, Thermo Fisher Scientific, Asimov and Syros Pharmaceuticals until 31 July 2020. Since 1 August 2020, A.R. is an employee of Genentech, a member of the Roche group. O.R.-R. is a co-inventor on patent applications filed by the Broad Institute for inventions relating to single-cell genomics, such as PCT/US2018/060860 and US provisional application no. 62/745,259. No other authors declare competing interests.

Additional information

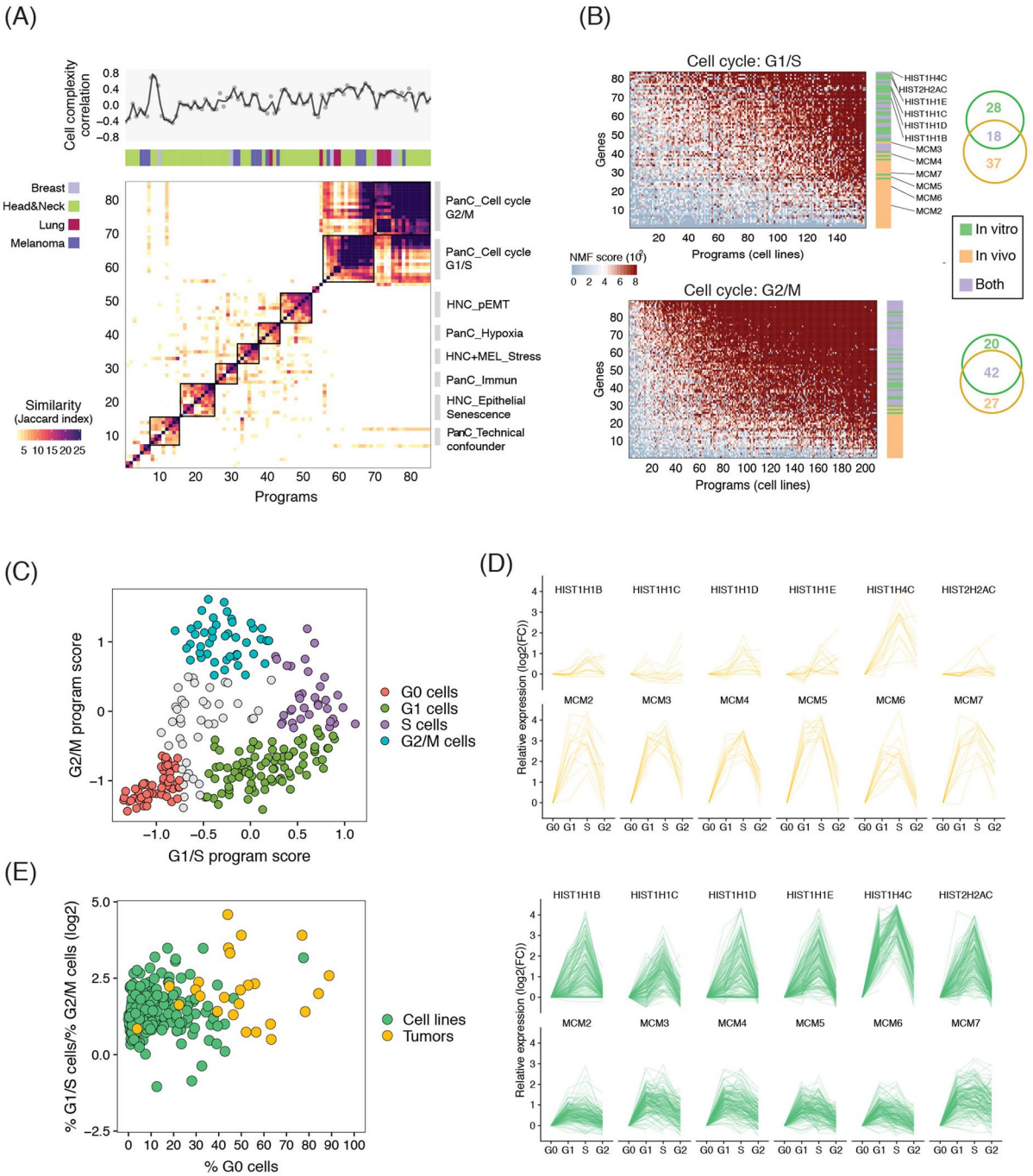
Extended data is available for this paper at https://doi.org/10.1038/s41588-020-00726-6.

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-00726-6.

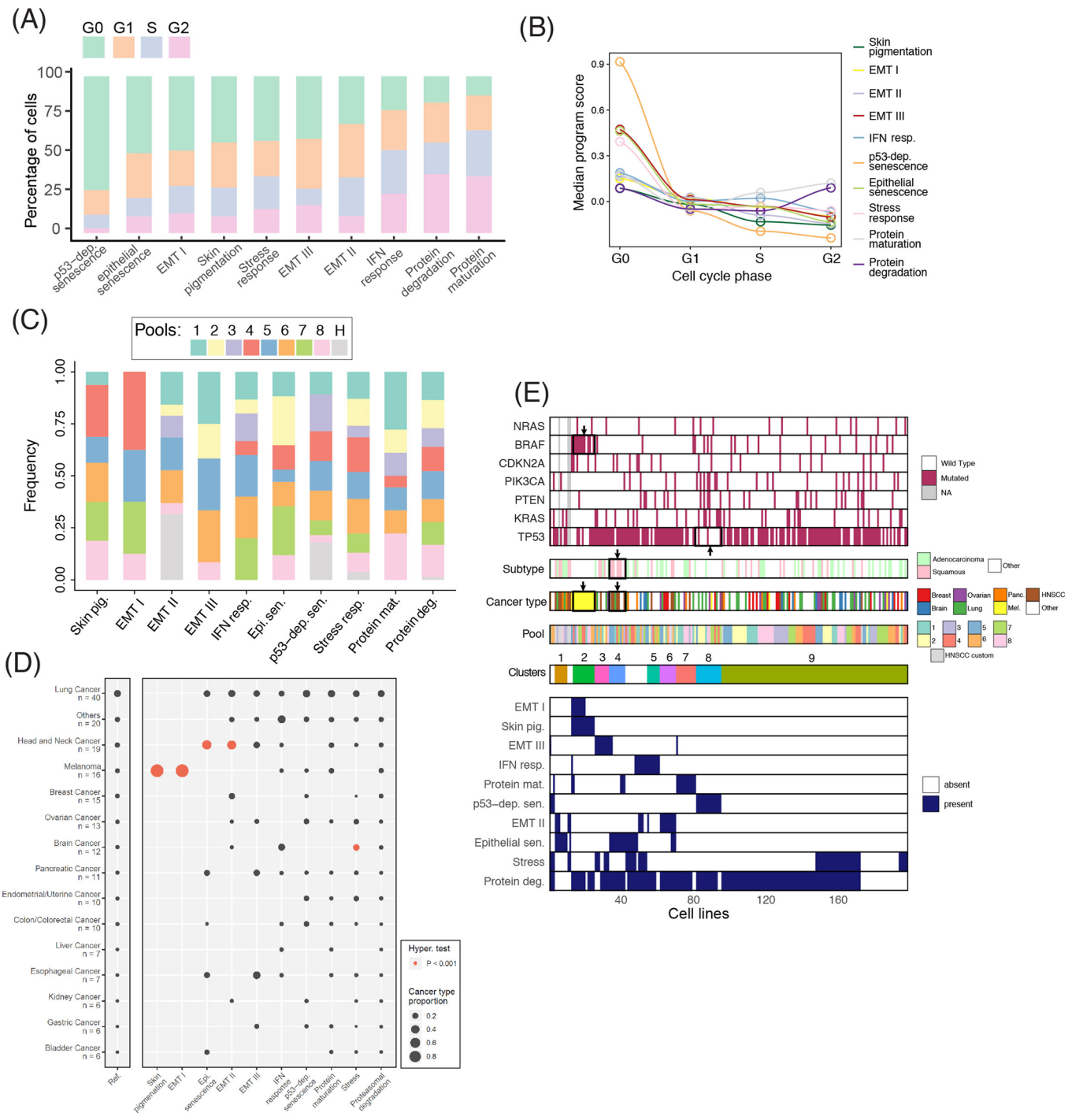
Correspondence and requests for materials should be addressed to I.T.



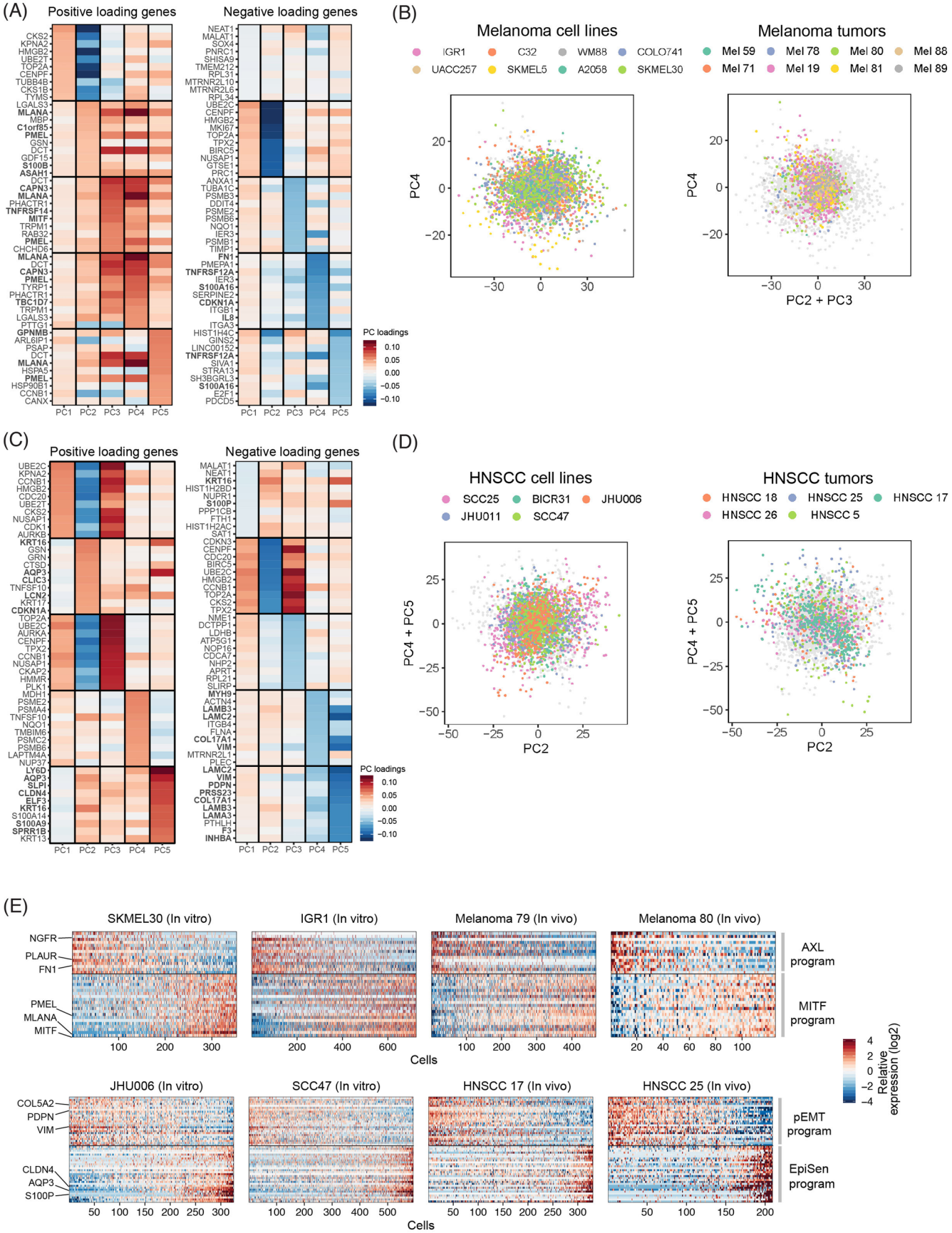
Extended Data Fig. 1 | Identifying discrete clusters and continuous expression programs. a, Performance of DBSCAN using different sizes of epsilon neighborhood (eps) and minimum numbers of points required to form a dense region (MinPts). We randomly selected cells from two different cell lines and tested the ability of DBSCAN to distinguish between them (two-sided Fisher's exact test, $P < 0.001$) using different parameter combinations. The procedure was repeated 1,000 times and the combination yielding the highest rate of correct classification was applied in the subsequent analyses. **b**, t-SNE plots for additional two examples of cell lines from each of the four classes defined by presence and number of discrete subpopulations identified by DBSCAN (as in Fig. 2b). **c**, **d**, Identification of discrete programs of heterogeneity, as in Fig. 2b, using less stringent eps (1.2 and 1.5) highlights common trends. **e**, Number of heterogeneity programs identified per cell line using NMF. NMF was applied to each cell line using k (number of factors) of 6–9, and gene programs identified as variable with 2 or more values of k were retained (left panel, n = 1,445). To identify common expression programs varying within multiple cell lines, we excluded programs with limited similarity to all other programs as well as those associated with technical confounders (right panel, n = 800). **f**, Pairwise similarities between programs identified by NMF across all the cell lines analyzed, with cell lines ordered by hierarchical clustering. Programs with limited similarity to all other programs were excluded. Top panel indicates correlations between program scores and cell complexity (that is number of genes detected per cell). The cluster of programs that correlates with complexity (indicated by dashed lines) was excluded from subsequent analyses.



Extended Data Fig. 2 | *In vivo* programs of variability and comparison to *in vitro* cell cycle programs. a, Heatmap depicts pairwise similarities between programs identified in tumor samples using NMF. Programs with limited similarity to all other programs were excluded. Top panel shows tumor type and correlations between program scores and cell complexity (that is number of genes detected per cell). Hierarchical clustering emphasizes multiple clusters (shown by squares), one of which is correlated with cell complexity and thus excluded as a potential technical artifact. **b**, NMF scores of G1/S genes (top panel) and G2/M genes (bottom panel) across all NMF programs associated with the corresponding cell cycle phase; each program is from a different cell line. Genes are ranked in each panel by average scores, and their assignment to *in vitro* and *in vivo* cell cycle programs is indicated in the right bar, demonstrating that G1/S programs differ both across cell lines and between cell lines and tumors, while G2/M programs are more consistent. Venn diagrams (right) illustrate the overlap of genes between *in vivo* and *in vitro* RHPs. **c**, Single-cell profiles (n = 264 cells from NCIH2126) showing G1/S and G2/M program score thresholds used to assign cells to different cell cycle phases. **d**, Examples of genes with distinct cell cycle upregulation *in vitro* and *in vivo*. Expression of HIST genes (preferentially induced *in vitro*) and MCM genes (preferentially induced *in vivo*) is shown along the cell cycle (relative to cells in G0) in cell lines (C, green lines) and tumors (D, yellow lines). **e**, Comparison of cell cycle phase distribution *in vitro* and *in vivo*. Scatterplot shows the percentage of cells in G0 (x-axis) and the ratio between the percentage of cells in G1/S and G2/M (y-axis) for each cell line (green, n = 198) and tumor (yellow, n = 25) analyzed. Cell lines display a significantly lower percentage of cells in G0 cells ($P=2e^{-10}$, two-sided t-test).

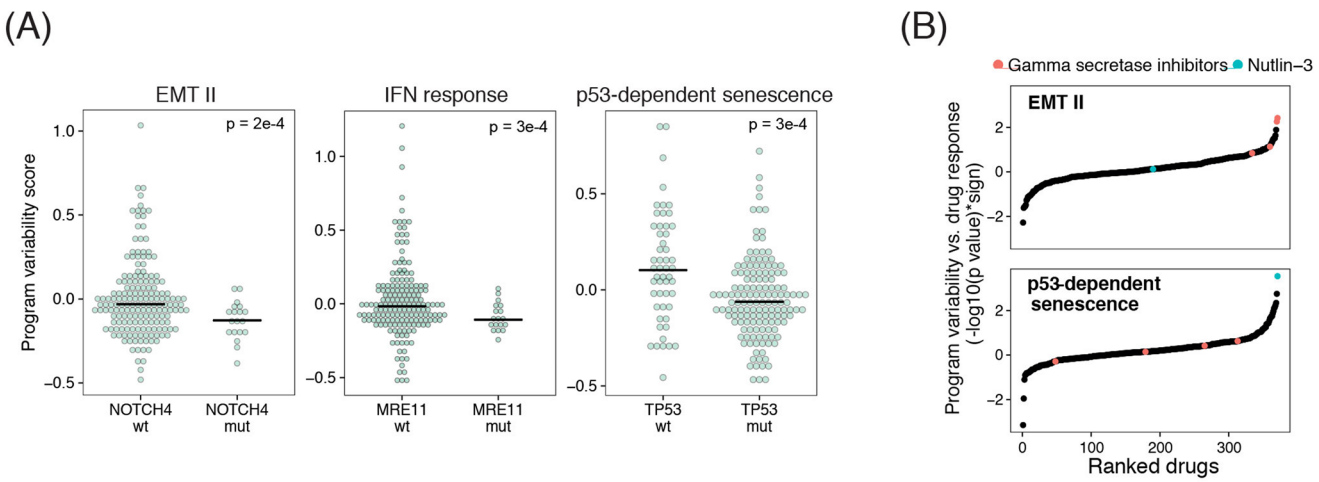


Extended Data Fig. 3 | Association of RHPs with cell cycle, pool identity and cancer types. a, Average distribution of cell cycle phases in cells with high RHP scores (top 5%). **b**, Median RHP scores of cells in each phase of the cell cycle. Cell cycle state was estimated for each individual cell based on the relative expression of the G1/S and G2/M metaprograms (see Extended Data Fig. 2c). For each RHP in (**a**) and (**b**) we only considered the respective model cell lines (see Supplementary Table 3). **c**, Distribution of RHPs (n = 680) across the 9 pools (H corresponds to the custom HNSCC pool). Each RHP was observed in multiple pools, underscoring the lack of pool-specific effects and the robustness of RHPs. **d**, The fraction of cell lines from each cancer type (rows) observed in each RHP (columns) are indicated by circle size. Red circles depict significant enrichments (P<0.001 by hypergeometric test). Left panel shows the fraction of each cancer type in our dataset and the number of cell lines profiled. **e**, Hierarchical clustering of cell lines based on their set of RHPs (bottom panel). Clusters composed of more than 5 cell lines are annotated (1–9). Relevant cell line features are shown on top and their associations with each cell line cluster was tested using hypergeometric test. Significant associations (P < 0.001) are indicated by black squares and arrows. These include associations of cluster 2 with melanoma, cluster 4 with HNSCC and squamous cells, and cluster 8 with wild-type p53.

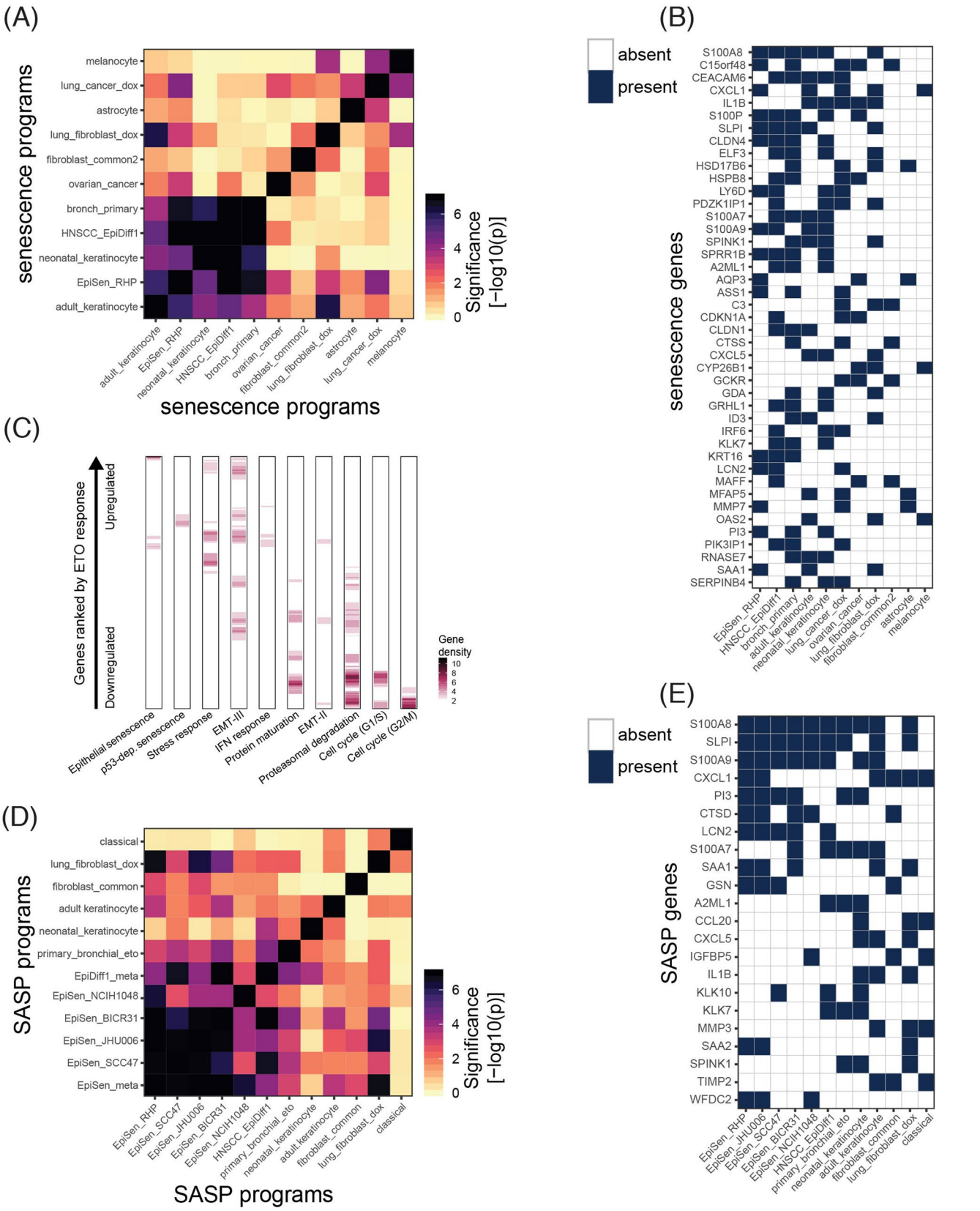


Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Cell lines recapitulate programs of heterogeneity observed in tumor samples. a–**d**, Combined PCA of *in vitro* and *in vivo* cells in melanoma (A-B) and HNSCC (C-D). (**a**, **c**) Top 10 genes by positive (left) and negative (right) loadings of PC1–5; Genes of respective RHPs (SkinPig. and EMT-I in melanoma, EpiSen and EMT-II in HNSCC) are emphasized (bold). (**b**, **d**) Coordinates of the three PCs associated with the respective RHPs across *in vitro* (left, melanoma = 3,033 cells, HNSCC = 2,780 cells) and *in vivo* (right, melanoma = 1,169 cells, HNSCC = 1,078 cells) cells. Cells are colored by their cell line or tumor. Similar results were obtained when PCs were not combined (not shown). (**e**) Heatmap shows relative expression of genes shared by paired *in vivo* and *in vitro* programs in selected melanoma (top panels) and HNSCC (bottom panels) cell lines and tumors, highlighting similar patterns of variability *in vivo* and *in vitro*. Cells are sorted according to the relative average expression of genes in each program, showing the negative correlation between the AXL and MITF programs in melanomas and the pEMT and EpiSen programs in HNSCC. Programs are annotated (right) and selected genes are indicated (left).

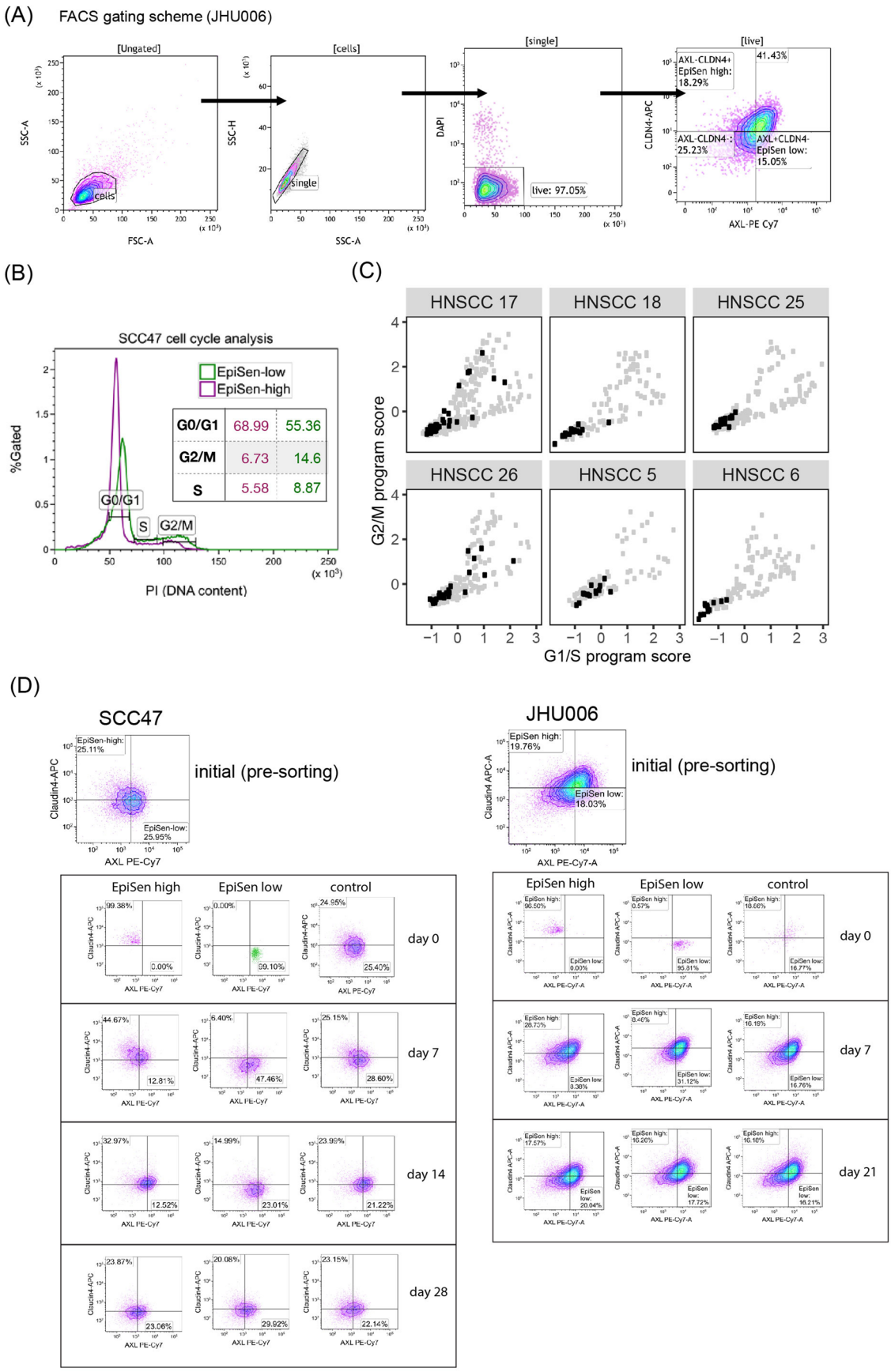


Extended Data Fig. 5 | Determinants and consequences of cellular heterogeneity. a, Association between RHP variability scores and somatic non-silent mutations. We compared the variability of each program in mutated and non-mutated cell lines using two-sided t-test. Model cell lines (high variability) of EMT-II, IFN response and p53-dependent senescence RHPs are depleted of NOTCH4, MRE11 and TP53 mutations, respectively. Horizontal lines indicate the median. **b**, Association between drug response (CTRP database) and program variability calculated using linear regression including tumor type and program variability as independent variables. Increased sensitivity to NOTCH inhibition (gamma secretase inhibitors) and MDM2 inhibition (Nutlin-3) were observed in model cell lines (high variability) of the EMT-II and the p53-dependent senescence respectively.



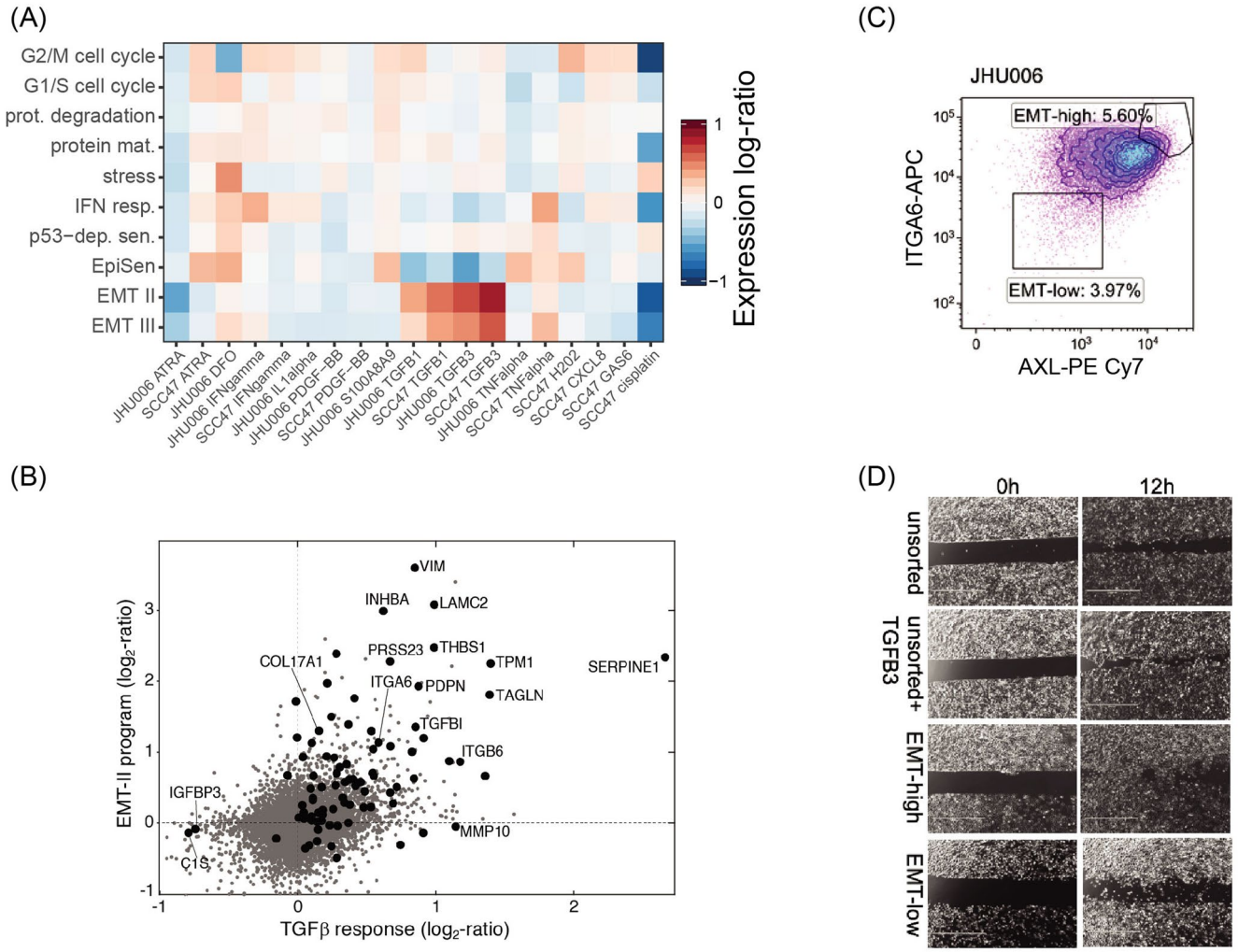
Extended Data Fig. 6 | See next page for caption.

Extended Data Fig 6 | Similarities and differences between senescence-related expression programs. a, Significance (-log10(P), hypergeometric test) of the overlaps between sets of genes upregulated upon senescence, as defined by multiple studies (see Table S8). **b**, Most common genes across the different senescence programs, including all genes appearing in at least three programs, ranked from top to bottom by the number of programs. **c**, Density of RHP signature genes within sliding windows of 300 genes, among the top 6,000 expressed genes, arranged by their expression response to etoposide, from the most upregulated (top) to the most downregulated (bottom), as also shown in Fig. 6b. RHPs are sorted from left to right by their enrichment with upregulated and downregulated genes, respectively. **d**, **e**, Significance of gene-set overlaps (**d**), and the most common genes (**e**), as in (**a**, **b**), when restricting the analysis to secreted genes, in order to define SASP.

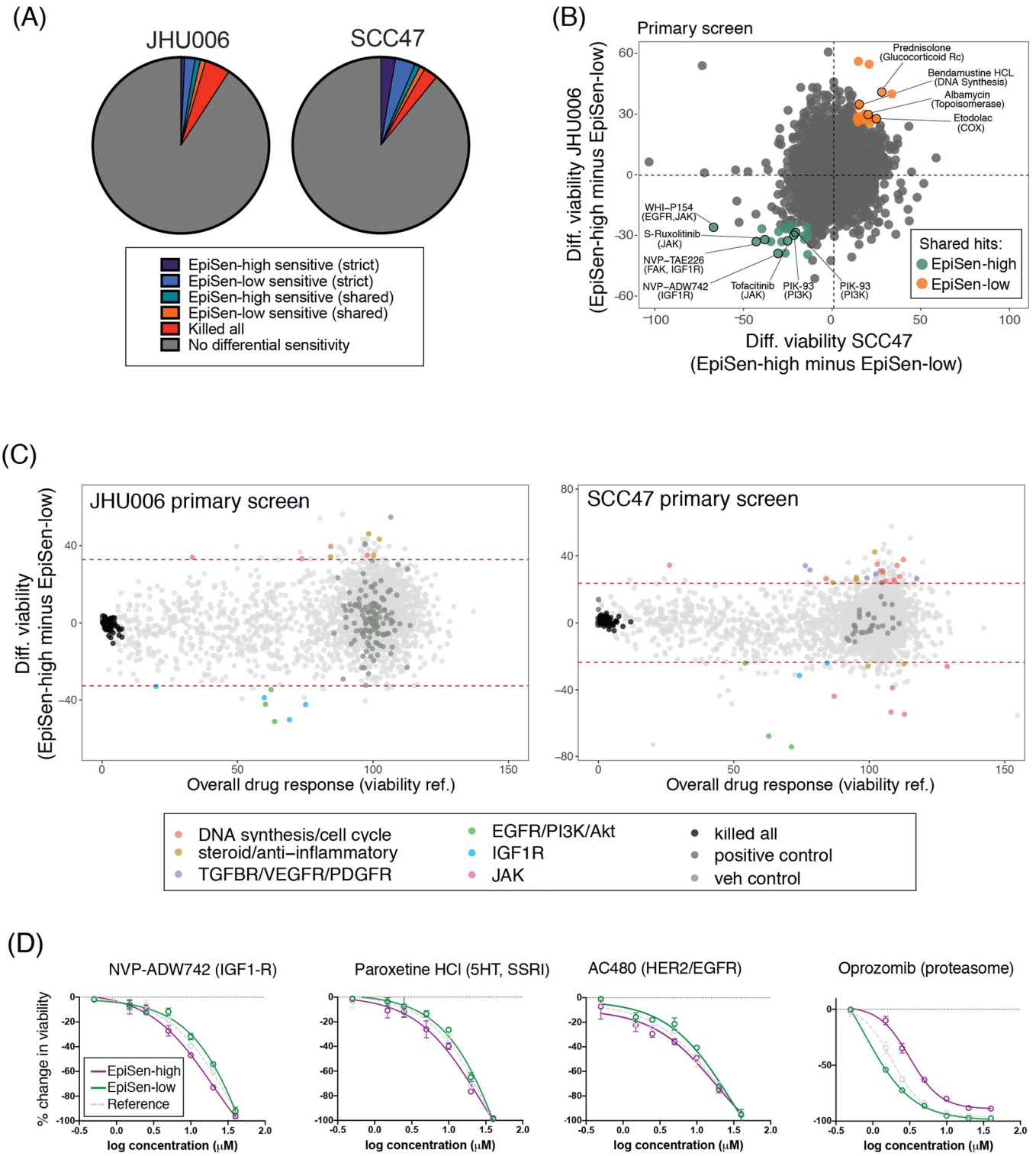


Extended Data Fig. 7 | See next page for caption.

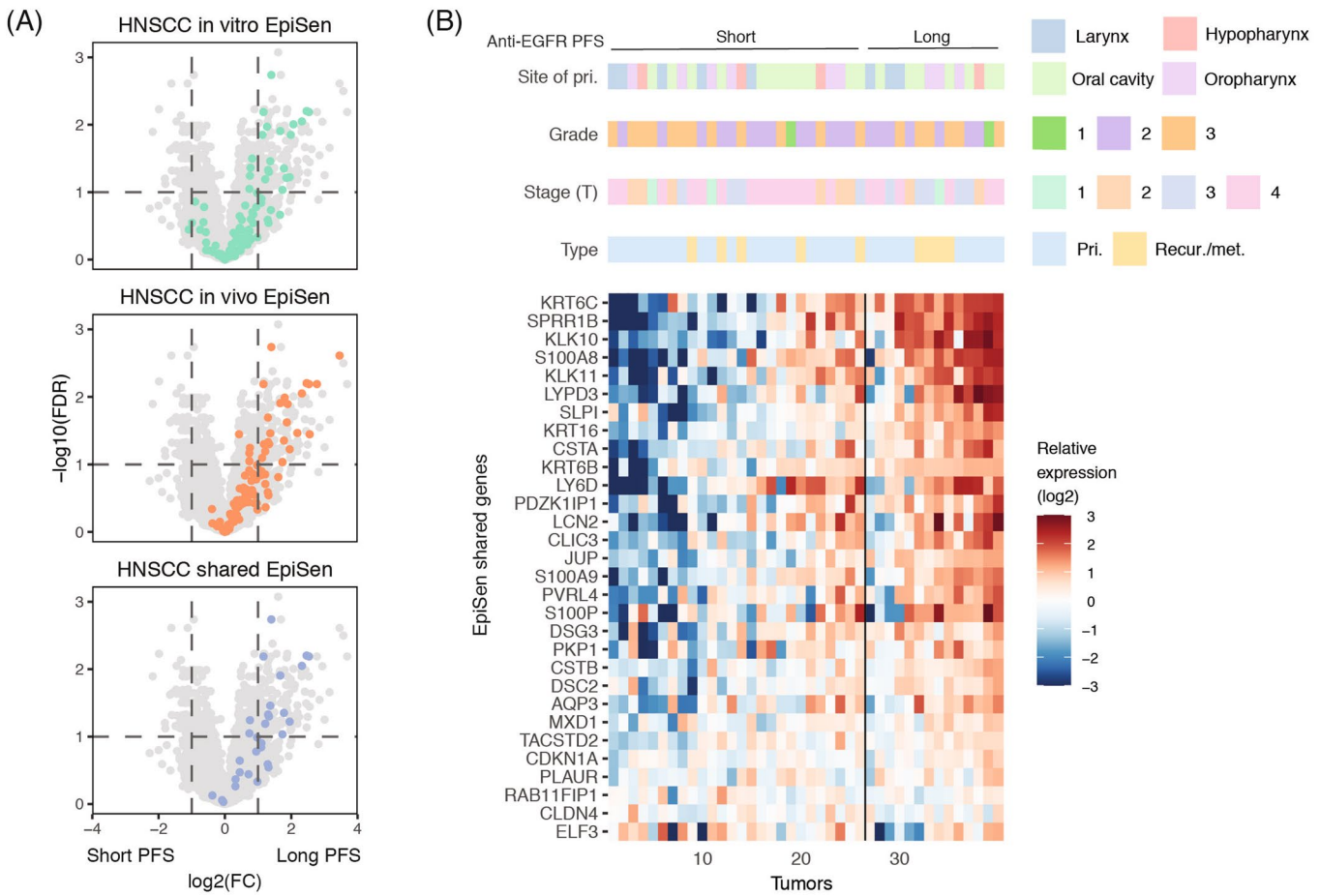
Extended Data Fig. 7 | Isolating EpiSen-high and EpiSen-low cells in model cell lines. a, Gating scheme for isolation of EpiSen-high and EpiSen-low subpopulations in the JHU006 cell line including doublet and dead cell exclusion. For sorting, the top 10% high and bottom 10% low cells were sorted. Validation of AXL+CLDN4- (EpiSen-high)-sorted cells enriching for the full EpiSen RHP was performed by bulk RNA-Seq of sorted cells (Fig. 5c, Methods). **b**, FACS analysis of cell cycle by the DNA binding dye propidium iodide (PI) on sorted EpiSen-high and EpiSen-low cells in SCC47, as shown for JHU006 in Fig. 5d. The table summarizes the results. **c**, Cell cycle scores of the G1/S (X-axis) and G2/M (Y-axis) programs, shown for *in vivo* HNSCC cells from six tumors (panels). Top 10% of cells scoring highly for the EpiSen-related program (EpiDif1) are shown in black, demonstrating their enrichment among non-cycling cells in each of the tumors ($P<0.001$, hypergeometric test). **d**, Three subpopulations were isolated by FACS (EpiSen-high: $AXL^{-}CLDN4^{+}$, EpiSen-low population: $AXL^{+}CLDN4^{-}$, control: unsorted) from SCC47 (left) and JHU006 (right), and analyzed immediately after sorting (day 0, top) and at two additional time points in culture (day 7 and 14, middle and bottom, respectively). Density plots correspond to the pie charts in Fig. 5e.



Extended Data Fig 8 | Activation and isolation of the EMT-II program. a, Average expression log-ratio for the gene-sets representing each RHP (each row represents an RHP), upon treatment of SCC47 or JHU006 with multiple perturbations (each column represents one perturbation in one cell line, as indicated at the bottom averaged over duplicates). **b**, Comparison of the EMT program induced upon TGFβ treatment of unsorted cells (X-axis) vs. the EMT-II RHP gene scores (Y-axis). In both axes, data was averaged over the results for JHU006 and SCC47. **c**, Isolation by FACS of the EMT-II-high population (AXL^+ITGA6^+) and the EMT-II-low population (AXL^-ITGA6^-) in JHU006. **d**, Both TGFβ3-treated cells and EMT-II-high sorted cells are associated with increased migration. Shown is a gap closure (migration) assay, performed on unsorted, unsorted but TGFβ3-treated, EMT-II-high, and EMT-II-low cells, at 0 h and 12 h following gap generation. Scale bar length is 1000 μM.



Extended Data Fig. 9 | Drug sensitivity of EpiSen-high cells and EpiSen-low cells. a, Pie charts depict the proportions of primary screen hits by type. **b**, Shared hits between SCC47 and JHU006 for compounds that preferentially killed the EpiSen-high (green) and EpiSen-low (orange) states. Selected hits are labeled. **c**, Viability of the control population (X-axis) and differential viability of the EpiSen-high vs. EpiSen-low populations (Y-axis) upon treatment with 2198 compounds in JHU006 (left) and SCC47 (right). Dotted lines represent thresholds for differential sensitivity, and hits are colored as defined in the lower legend. **d**, Dose response curves of selected compounds in three SCC47 subpopulations at seven concentrations measured in duplicate (continued from Fig. 6c). Change in viability was calculated relative to vehicle (DMSO-treated) controls. Error bars represent standard deviation, data points represent the mean of replicates.



Extended Data Fig. 10 | Association between the abundance of EpiSen cells and clinical response of HNSCC patients to Cetuximab. a, Volcano plots depict differential expression analysis comparing bulk pretreatment samples of HSNCC patients with short (n = 14) and long (n = 26) PFS following Cetuximab treatment plus platinum-based chemotherapy. Comparisons were performed using two-sided t-test, and P values were adjusted using the FDR procedure. Genes in the HNSCC EpiSen program from cell lines (*in vitro*, top panel) and tumors (*in vivo*, middle panel) are highlighted, as well as those shared between the two programs (bottom panel). **b**, Heatmap showing the expression of shared EpiSen genes in samples stratified into short and long PFS. Genes are order by differential expression (log_2(fold change)), as shown in (**a**), and tumors are ordered within each group according to the EpiSen score. Top panels depicts sample type, tumor stage (T), tumor grade, and site of primary tumors.

nature research

Corresponding author(s): Tirosh, Itay

Last updated by author(s): Sep 1, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	BD FACSChorus (v1.0) was used for acquisition of flow cytometry data.
Data analysis	Data analysis was performed using R (v3.5.3 and v3.6.3) with the following packages: Rtsne (v0.15), dbscan (v1.1-4), NMF (v0.21.0) , mclust (v5.4), glmnet (v2.0-16), SCENIC (v1.1.1), kernlab (v0.9-26), ggplot2 (v3.1.0), GENIE3 (v1.9), and RcisTarget (v1.9). Additional software used included CellRanger (v3.0.1, 10x Genomics), Bowtie (v1.2.2), RSEM (v1.3.1), Graphpad Prism (v8), Kaluza FACS Analysis Software (v2.1, Beckman Coulter) and GeneData Screener (v12). A full description of data analysis is found in 'Methods.' Source code for the main analyses presented in this work can be found at https://github.com/gabrielakinker/CCLE_heterogeneity.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Raw and processed scRNA-seq data is available through the Broad Institute's single-cell portal (SCP542), and at the Gene Expression Omnibus (GEO) with accession GSE157220. Publicly available databases used in our analysis included: DepMap portal (https://depmap.org/; 18q3 data release, CCLE portal (https://

portals.broadinstitute.org/ccle), CTD2 portal (https://ocg.cancer.gov/programs/), and MSigDB V7.0 (https://www.gsea-msigdb.org/gsea/msigdb/index.jsp), the human motif collection version 9 (mc9nr), and GTRD database v20.06 (gtrd.biouml.org).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	As described in the original CCLE publication (Barretina et al. 2012, Nature 483:603-607), cell lines were selected based on commercial availability and unmet need of models. The CCLE pools were supplemented with an additonal custom pool of HNSCC cell lines in order to further enhance our analysis of HNSCC models of gene expression heterogeneity. Given the exploratory nature of the analysis, no calculation was done to determine sample size. Instead, we decided to stop after profiling 9 pools (n= ~200 cell lines) since we already covered 22 cancer types and since we were able to identify 12 recurrent programs of heterogeneity
Data exclusions	In the scRNA-Seq data, we excluded cells with below 2,000 genes detected. When analyzing cell lines individually, we only considered genes expressed at high or intermediate levels (Ei,j > 3.5) in at least 2% of cells, yielding an average of 6,758 genes analyzed per cell line. When analyzing cell lines collectively, we selected the top 7,000 expressed genes across all cell lines, resulting in a minimum average expression of 12 CPM. This exclusion criteria was determined by exploratory data analysis. Additionally, cells identified as doublets, low quality cells, or with inconsistent assignment between the bulk expression and SNP-based identification methods were excluded from further analysis (see 'Methods' for full description) based on pre-established exclusion criteria. In the dose response analysis following the drug screen, two compounds for which EC50 values could not be calculated were omitted from further analysis and this exclusion criteria was pre-established.
Replication	Independent replication was successful and performed for: Fig. 5C&E (three independent replicates), Fig. 5D (two independent replicates), Fig. 6C (where shown, biological replicates were cultured and treated independently, but sequencing libraries were generated at the same time), Extended Data Fig. 7B (two independent replicates), Extended Data Fig. 7D (three independent replicates). Fig. 7B (performed in duplicate) is an independent replicate of the putative hits identified in Extended Data Fig. 9A-C. SCC25 was profiled four times independently by scRNA-Seq as part of CCLE pool ID #19, in the custom HNSCC pool, and in both groups of the co-culture control experiment (Fig. SI1). The HNSCC cell lines JHU006 and UM-SCC47 were profiled three times independently by scRNA-Seq - as part of the custom HNSCC pool and then again in each group of the co-culture control experiment. The experiment described in Fig. SI1E-G consists of independent replicates of cell lines pooled either with or without 72 hours of co-culture.
Randomization	Randomization was generally not relevant for our study, as we had no humans or animal models. Cell cultures subjected to drug treatments were allocated randomly to treatment or control groups.
Blinding	Data collection for scRNA-Seq analysis was performed without investigators' knowledge of cell lines identities. Investigators were not blind to cell line identities during analysis as knowledge of cancer type/cell line identity was necessary in order to perform the analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used: Anti-human AXL PE-Cy7 eBioscience Cat #25-1087-42

Antibodies used	Anti-human Claudin-4 APC Miltenyi Cat #130-114-871 Anti-human ITGA6/CD49f APC eBioscience Cat #17-0495-82
Validation	FACS antibodies were validated by the manufacturer. AXL-PE Cy7: https://www.thermofisher.com/antibody/product/Axl-Antibody-clone-DS7HAXL-Monoclonal/25-1087-42 Claudin4-APC: https://www.miltenyibiotec.com/US-en/products/macs-flow-cytometry/antibodies/primary-antibodies/anticlaudin-4-antibodies-human-rea898-1-50.html ITGA6-APC: https://www.thermofisher.com/antibody/product/CD49f-Integrin-alpha-6-Antibody-clone-eBioGoH3-GoH3-Monoclonal/17-0495-82

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HNSCC cell lines (SCC9, SCC25, JHU006, JHU011, JHU029, UM-SCC47, UM-SCC90, 93VU-147T) were donated by Dr. James Rocco at the Ohio State University. Human primary bronchial cells were purchased from ATCC (Cat #PCS-300-010). A list of the CCLE cell lines and vendors is available on the CCLE portal (www.broadinstitute.org/CCLE).
Authentication	HSNCC cell lines were authenticated by STR analysis. Primary bronchial cells were authenticated by ATCC. Cell lines in the CCLE pools were authenticated by SNP-based DNA fingerprinting.
Mycoplasma contamination	All cell lines tested negative for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	To our knowledge no commonly misidentified cell lines were used in our study.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Following trypisinization and neutralization of trypsin by media, pelleted cells were washed and resusupended in 100ul PBS-0.5% BSA per 1x10^6 cells for staining at the following concentrations: AXL-PECy7: 1:300, Claudin4-APC: 1:200, ITGA6-APC: 1:200. Cells were washed and resuspended in PBS-0.5% BSA for sorting and analysis. Cells were sorted with a 100uM nozzle into 20% FBS-media.
Instrument	BD FACS Melody
Software	BD FACSChorus v1.0 (acquisition); Kaluza Analysis Software v2.1 Beckman Coulter (analysis)
Cell population abundance	Post-sort purity was confirmed by re-analysis of sorted subpopulations of 2000-10000 cells for each population. Post-sort purity is demonstrated in Fig. S8D.
Gating strategy	For all flow experiments, cells were gated based on FSC/SSC, live (DAPI negative based on cells from the unstained control) and single cells (FSC-H/FSC-A). For isolation of gene expression programs, the 'positive' and 'negative' gates were set based on the unstained control and FMO controls. For time course experiments, positive and negative gates were set based on the unstained control at each time point. For sorting, the top 10% of the high and bottom 10% of the low population was taken. Because the cell states isolated by FACS represent continuous (non-discrete) populations, additional validation of the FACS gating scheme was performed by bulk RNA-Seq of sorted cells to confirm isolation of the intended expression program (Fig. 5C). A figure exemplifying the gating strategy is provided in Fig. S8D.

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.