

# UC Santa Cruz

## UC Santa Cruz Previously Published Works

### Title

Machine learning multi-omics analysis reveals cancer driver dysregulation in pan-cancer cell lines compared to primary tumors

### Permalink

<https://escholarship.org/uc/item/7wk433xs>

### Journal

Communications Biology, 5(1)

### ISSN

2399-3642

### Authors

Sanders, Lauren M

Chandra, Rahul

Zebarjadi, Navid

et al.

### Publication Date

2022

### DOI












10.1038/s42003-022-04075-4

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

## Machine learning multi-omics analysis reveals cancer driver dysregulation in pan-cancer cell lines compared to primary tumors

Lauren M. Sanders <sup>1,2</sup>, Rahul Chandra<sup>3</sup>, Navid Zebarjadi<sup>2,4</sup>, Holly C. Beale <sup>2,4</sup>, A. Geoffrey Lyle <sup>2,4</sup>, Analiz Rodriguez<sup>5</sup>, Ellen Towle Kephart <sup>2</sup>, Jacob Pfeil<sup>1,2</sup>, Allison Cheney <sup>2,4</sup>, Katrina Learned<sup>1,2</sup>, Rob Currie <sup>1,2</sup>, Leonid Gitlin<sup>6</sup>, David Vengerov<sup>7</sup>, David Haussler <sup>1,2,9</sup>, Sofie R. Salama <sup>1,8,9</sup> & Olena M. Vaske <sup>2,4,9</sup>

Cancer cell lines have been widely used for decades to study biological processes driving cancer development, and to identify biomarkers of response to therapeutic agents. Advances in genomic sequencing have made possible large-scale genomic characterizations of collections of cancer cell lines and primary tumors, such as the Cancer Cell Line Encyclopedia (CCLE) and The Cancer Genome Atlas (TCGA). These studies allow for the first time a comprehensive evaluation of the comparability of cancer cell lines and primary tumors on the genomic and proteomic level. Here we employ bulk mRNA and micro-RNA sequencing data from thousands of samples in CCLE and TCGA, and proteomic data from partner studies in the MD Anderson Cell Line Project (MCLP) and The Cancer Proteome Atlas (TCPA), to characterize the extent to which cancer cell lines recapitulate tumors. We identify dysregulation of a long non-coding RNA and microRNA regulatory network in cancer cell lines, associated with differential expression between cell lines and primary tumors in four key cancer driver pathways: KRAS signaling, NFκB signaling, IL2/STAT5 signaling and TP53 signaling. Our results emphasize the necessity for careful interpretation of cancer cell line experiments, particularly with respect to therapeutic treatments targeting these important cancer pathways.

<sup>1</sup> Department of Biomolecular Engineering, UC Santa Cruz, Santa Cruz, CA, USA. <sup>2</sup> UC Santa Cruz Genomics Institute, Santa Cruz, CA, USA. <sup>3</sup> Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA. <sup>4</sup> Department of Molecular, Cell and Developmental Biology, UC Santa Cruz, Santa Cruz, CA, USA. <sup>5</sup> Department of Neurosurgery, University of Arkansas for Medical Sciences, Little Rock, AR, USA. <sup>6</sup> Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, California, USA. <sup>7</sup> Oracle Labs, Oracle Corporation, Pleasanton, CA, USA. <sup>8</sup> Howard Hughes Medical Institute, UC Santa Cruz, Santa Cruz, CA, USA. <sup>9</sup> These authors jointly supervised this work: David Haussler, Sofie R. Salama, Olena M. Vaske. email: [lmsh@ucsc.edu](mailto:lmsh@ucsc.edu); [olena@ucsc.edu](mailto:olena@ucsc.edu)

Tumor-derived cell lines provide a robust model environment for testing treatment hypotheses, identifying biomarkers of response to therapies, and studying underlying cancer biology. Cancer cell line models grow quickly, are comparatively cost-effective, and are readily available. Their integration into preclinical research has led to remarkable advances in cancer characterization and treatment<sup>1</sup>.

However, additional genomic characterization of cancer cell lines has indicated that the transition from *in vivo* to *in vitro* may introduce key genomic alterations. One of the first groups to compare tumor and cell line used microarray gene expression profiles to identify breast cancer cell lines that seemed to be genetically inappropriate models for breast carcinoma<sup>2</sup>. As mutation detection has become more accurate, multiple studies have reported that head and neck cancer cell lines tend to harbor more mutations than their tumors of origin<sup>3,4</sup>. A recent study showed that colorectal cancer cell lines recapitulate colorectal tumor subtypes, but that cell lines have more mutations than tumors<sup>5</sup>. In general, the current consensus concerning cancer cell lines as primary tumor models is that cell lines share many of the original tumor characteristics, but can harbor genetic changes of poorly characterized significance; and that some cancer cell lines may not even be molecularly appropriate or representative models for their tumor of origin<sup>6</sup>.

Nevertheless, cancer cell lines continue to be widely used in cancer research and therapeutic discovery. As the focus on molecularly targeted therapeutics grows, so does the need to thoroughly characterize how cancer cell lines diverge phenotypically from tumors due to their *in vitro* growth environment<sup>6</sup>. It is essential that preclinical researchers know which biological pathways behave similarly *in vivo* and *in vitro*, and even more importantly, which pathways demonstrate altered activity as a result of alterations in environmental signals and stressors in the *in vitro* growth setting. If a key pathway behaves differently in cell lines as compared to primary tumors, preclinical testing of a drug targeting that pathway will not accurately predict patient tumor response.

In order to characterize the specific pathway alterations that occur between primary tumors and tumor-derived cell lines, we analyze three types of high-throughput molecular data from The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE). We perform transcriptomic analysis of bulk RNA sequencing of TCGA and CCLE samples and identify a set of differentially expressed genes. We integrate micro-RNA sequencing from the same projects and identify an interaction network of micro-RNA (miRNA), long non-coding (lncRNA) and protein-coding genes that is aberrantly expressed in cell lines compared to tumors. This network implicates four key cancer driver pathways that are often the subject of preclinical drug evaluation in cell lines, but whose activity in cell lines is not representative of original tumors. We use proteomic quantification data from the same studies to demonstrate that the aberrant cancer driver pathway expression observed in cell lines extends to the proteomic level. We also demonstrate similar findings in separate datasets of single-cell RNA sequencing from tumors and cancer cell cultures.

## Results

**Support vector machine classifier identifies a set of genes differentially expressed between primary tumors and tumor-derived cell lines.** We hypothesized that genes with differential expression between the TCGA and CCLE datasets would represent differences in biological pathway activity. In order to identify novel sources of variation within these datasets, we eliminated immune-related genes because it is already known that cancer cell

lines are unable to recapitulate the immune signatures of the primary tumors<sup>7</sup> (see “Methods”, Supplementary Fig. 1 and Supplementary Data 1, Tab 2).

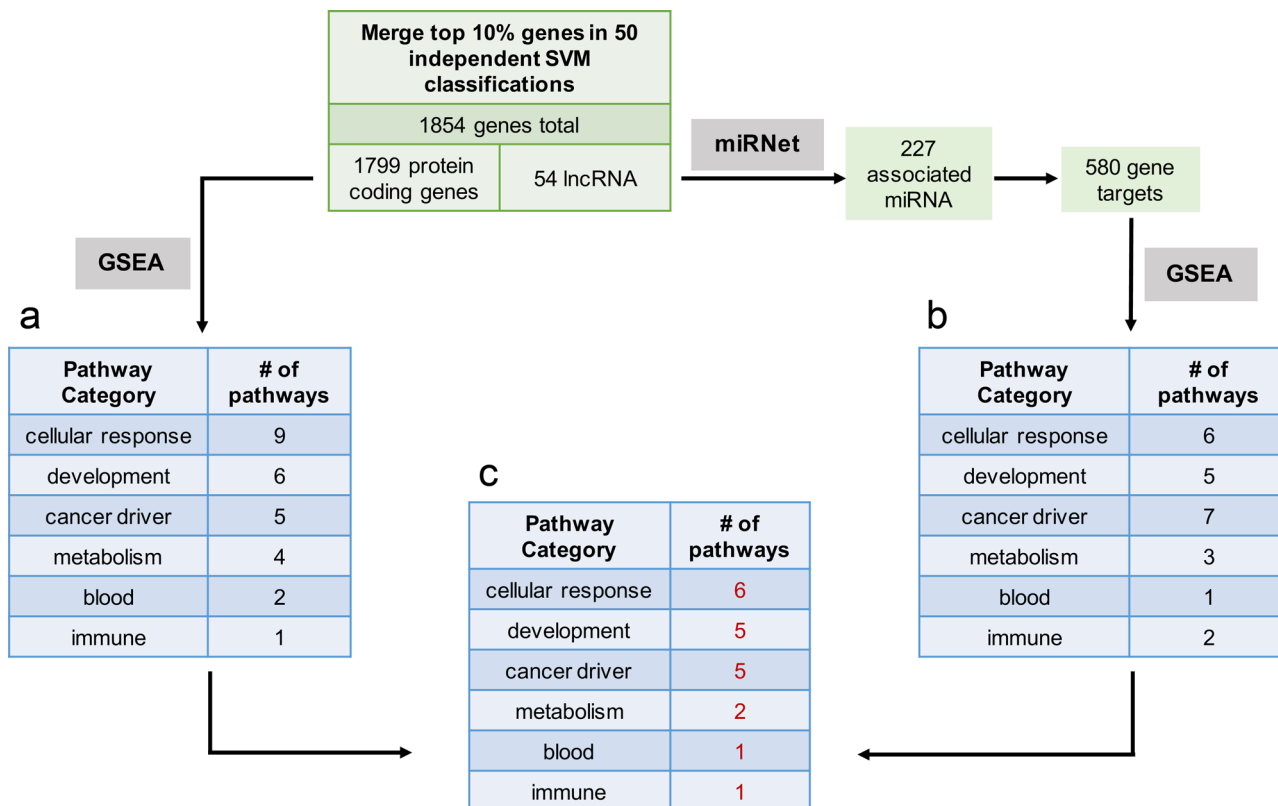
We then used a support vector machine (SVM) linear classifier within the Python *sklearn* module to identify genes (features) which are the most useful and important for classifying a new tumor or cell line based on a trained model<sup>8</sup>. Feature identification using SVM allowed us to identify the set of genes that best differentiate between tumor and cell line regardless of whether each gene is more highly expressed in the tumor group or the cell line group. The SVM approach has been shown to eliminate noisy results in a high-dimensional gene expression comparison, by drawing out the greatest sources of variation across all samples<sup>9,10</sup>. In order to ensure robust results, we repeated the classification on fifty different random 80/20 test/training splits of the data. After sorting the genes by their SVM-assigned feature importance coefficients, we merged the top 10% of genes from all fifty classifications, resulting in 1854 genes that were in the top 10% of most important genes for each classification (Supplementary Data 1, Tab 1). These genes included 54 lncRNA and 1799 protein-coding genes.

In order to characterize the functional significance of our SVM-derived gene set, we performed gene set enrichment analysis (GSEA) on the 1799 protein-coding genes using the Hallmark cancer pathway set from the Molecular Signatures Database (mSigDB v7.0)<sup>11</sup>. We found 27 gene sets with significant enrichment in the SVM-derived differentially expressed protein-coding genes (Fig. 1a and Supplementary Data 1, Tab 3).

Since lncRNA play known regulatory roles in normal tissue and in cancer, we hypothesized that the 54 differentially expressed lncRNA may be involved in regulating the differentially expressed coding genes, and may compose an interaction network with aberrant expression in cell culture. In order to characterize the functional interactions of these lncRNA, we employed miRNet, a tool that integrates multiple interaction databases for identification of lncRNA-miRNA and miRNA-gene regulatory networks<sup>12</sup>. miRNet identified 227 miRNA with known interactions to the 54 lncRNA. In turn, these 227 miRNA had 580 known gene targets among the 1799 differentially expressed coding genes ( $P$  value  $< 7.8 \times 10^{-27}$ , hypergeometric test). GSEA of the 580 coding genes revealed 24 Hallmark gene sets with significant enrichment ( $P$  value  $< 0.05$ , Fig. 1b and Supplementary Data 1, Tab 3). Strikingly, 20 of these Hallmark pathways overlapped with the enriched pathways from the coding genes GSEA (Fig. 1c and Supplementary Data 1, Tab 3), indicating that the set of SVM-derived important lncRNA is closely involved in many of the same pathways as the set of SVM-derived important coding genes.

We categorized the pathways into 6 categories (cellular response, development, cancer driver, metabolism, blood, and immune). Notably, our results (Supplementary Data 1, Tab 3) recapitulate the findings of a recent study by Yu and colleagues which found key differences in developmental, cell cycle, and immune pathways between cell lines and tumors<sup>7</sup>. In this study, we were particularly interested in the five cancer driver pathways which overlapped between the coding genes GSEA and the lncRNA-derived GSEA, as these molecular pathways are most likely to be the focus of preclinical trials in cancer cell lines.

**Four main types of cancer driver pathways exhibit differential expression and protein levels in cancer cell lines compared to primary tumors.** The five cancer driver pathways with significant enrichment in both SVM-derived coding genes and lncRNA-related genes are KRAS Signaling Up and Down, P53 pathway,



**Fig. 1 Workflow for support vector machine (SVM) classification of samples from TCGA and CCLE to assign feature importance to all genes.** **a** GSEA pathway enrichment results of 1799 protein-coding genes in the overlap of the top 10% most important genes in 50 independent SVM classifications. **b** GSEA pathway enrichment results of 580 protein-coding genes linked by miRNet databases to the 54 lncRNAs found in the overlap of the top 10% most important genes in 50 independent SVM classifications. **c** Overlap of pathways from the protein-coding gene GSEA and the lncRNA-based GSEA.

IL2/STAT5 signaling, and TNFA signaling via NFKB (Fig. 2a). With respect to KRAS signaling, since both pathways are a result of activated KRAS signaling, all subsequent analysis focuses on KRAS Signaling Up, which represents genes upregulated as a result of activated KRAS signaling. We focused on upregulated genes since most therapeutic approaches focus on reducing the activity of targeted genes<sup>13</sup>.

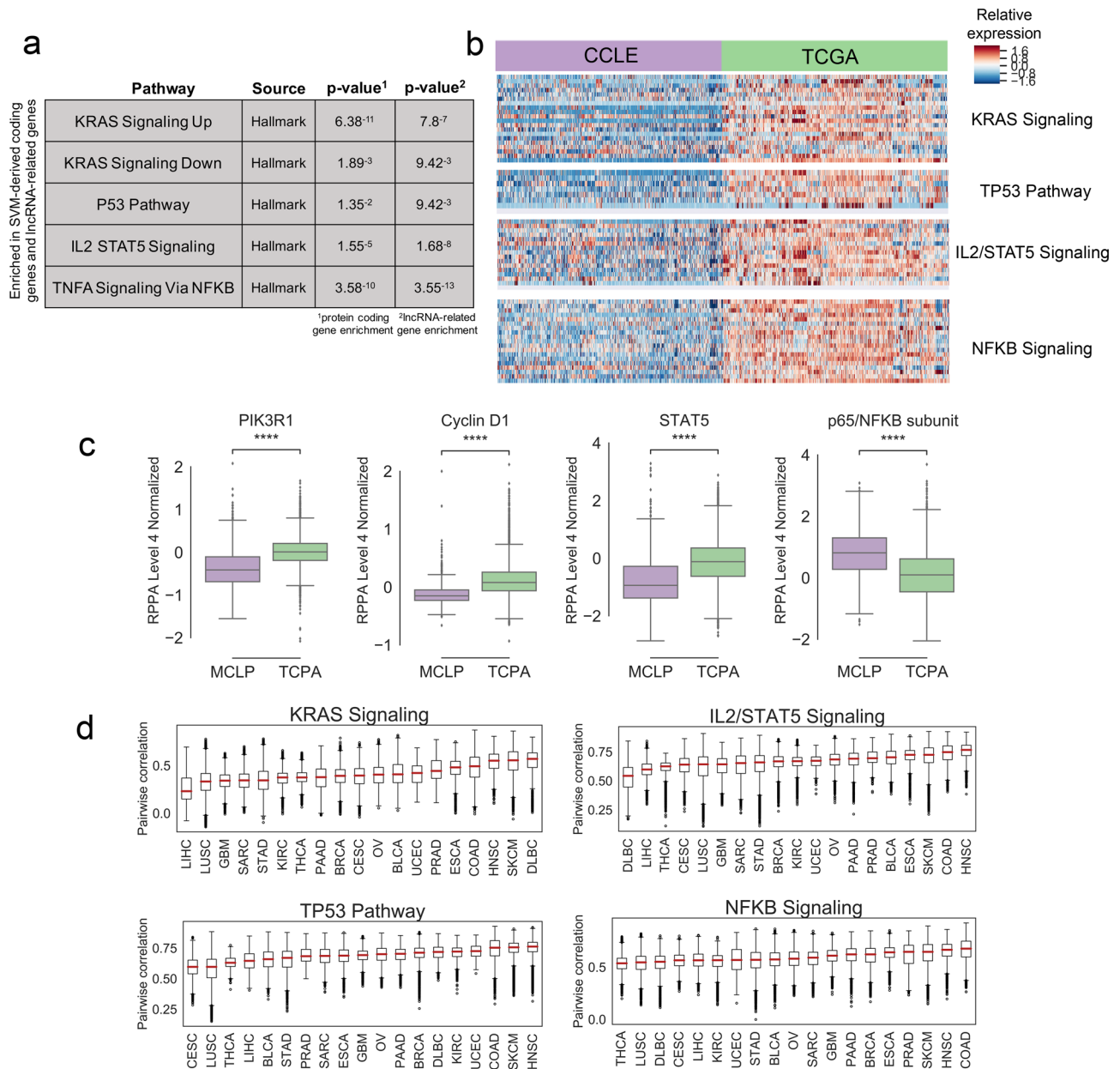
Interestingly, all four pathways show much higher overall expression in tumors than in cell lines (Fig. 2b). As a control, we also examined the gene expression of the Hallmark PI3K-AKT-mTOR pathway, a cancer driver pathway that was not significantly enriched in SVM-derived genes (*P* value < 0.426). This pathway did not show differential expression between CCLE and TCGA (Supplementary Fig. 2). We also verified that this signal was not a disease-specific artifact by repeating the SVM after subsetting the data to the disease with the largest number of samples in TCGA (BRCA) and the smallest number of samples (DLBC). The same four cancer driver pathways were identified in both analyses (Supplementary Data 1, Tab 4). These four pathways overall have similar correlation between tumor and cell line samples when evaluated on a disease-specific basis (Fig. 2d). We also verified that this signal is not related to tumor purity by performing ESTIMATE<sup>14</sup> tumor purity measurements on all TCGA samples and repeating the SVM comparison on the solid tumor types with the highest (KIRC) and lowest (PRAD) ESTIMATE scores (Supplementary Fig. 3). The four cancer driver pathways were identified as differentially expressed in both analyses (Supplementary Data 1, Tab 5).

We next examined whether the downregulation of these pathways extended beyond gene expression into protein activity. Proteomics quantification of many CCLE and TCGA samples was

performed using Reverse Phase Protein Array (RPPA) in the MD Anderson Cell Lines Project (MCLP) and The Cancer Proteome Atlas (TCPA)<sup>15,16</sup>. Because the RPPA data include fewer than 250 proteins, we identified through literature review proteins that are normally highly expressed downstream of each cancer driver pathway, and examined their levels in the RPPA Level 4 Normalized cell line and tumor data (Fig. 2c). PIK3R1 (antibody PI3KP85) is activated subsequent to KRAS signaling, and Cyclin D1 (antibody CYCLIND1) is activated downstream of the P53 pathway<sup>17,18</sup>. STAT5 (antibody STAT5ALPHA) represents the protein counterpart of the *STAT5* gene. The antibody NFKBP65\_pS536 binds to phosphorylated p65, one of the two protein subunits of NFKB. Phosphorylation of p65 is one of several molecular mechanisms known to activate the NFKB pathway<sup>19,20</sup>.

We noted significantly lower protein expression of PIK3R1, Cyclin D1, and STAT5 in the cell-line data, consistent with our gene expression results in the corresponding pathways. This carries important implications for the applicability of preclinical drug tests against these targets in cancer cell lines. Interestingly, the phosphorylation level of p65 is higher in cell lines than tumors, opposite the gene expression of the NFKB signaling pathway. This suggests that p65 phosphorylation may be playing a different role in cell lines, and underscores the importance of examining multiple types of data to elucidate complex molecular interactions.

Because activation of the KRAS and TP53 pathways is associated with mutations in the *KRAS* and *TP53* genes<sup>21,22</sup>, we investigated whether there is a correlation between diseases with heavy mutation burden in these genes and diseases with higher correlation between tumor and cell line, with the assumption that

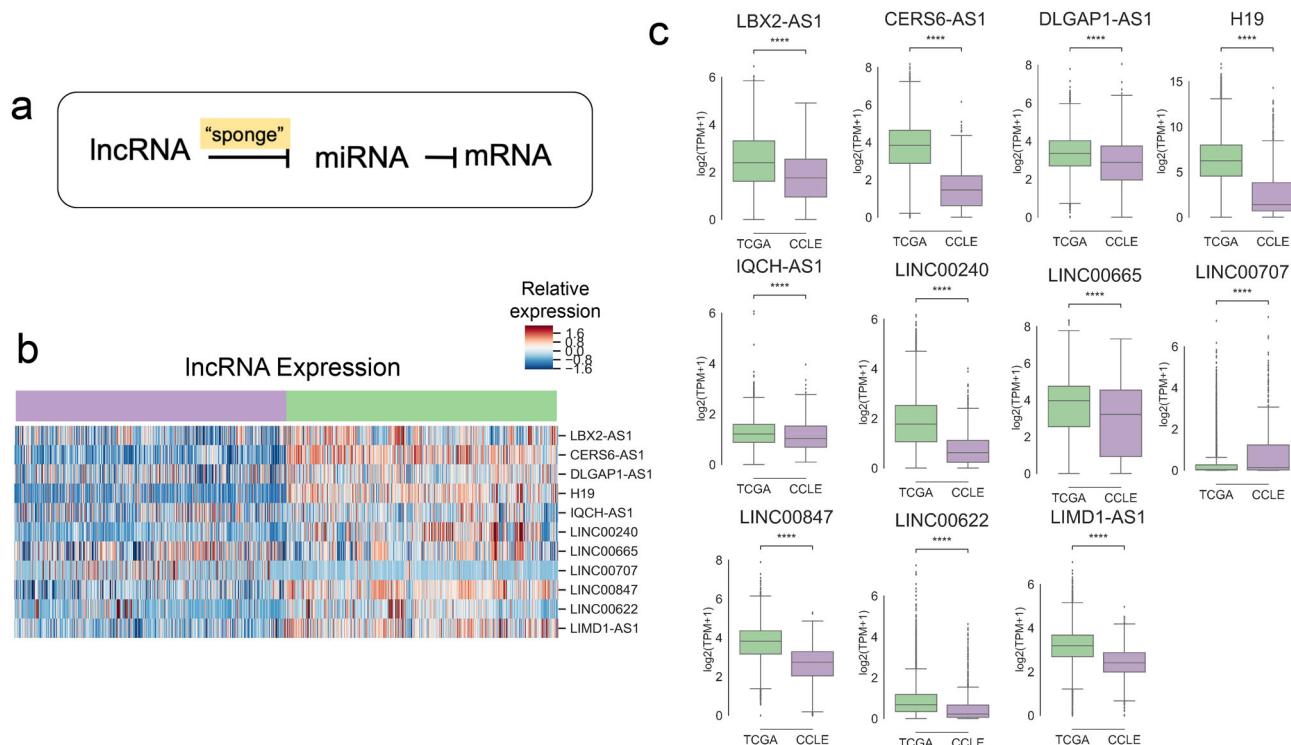


**Fig. 2 KRAS signaling, TP53 pathway, IL2/STAT5 signaling, and NFKB signaling are significantly enriched for genes with reduced expression in CCLE compared to TCGA.** **a** GSEA results for the cancer driver pathways which overlap with SVM-derived genes. *P* values are shown for the significance of gene overlap with SVM-derived protein-coding genes, and for genes linked by miRNet to SVM-derived lncRNA. **b** Heatmaps showing expression of SVM-identified genes in four cancer driver pathways in TCGA compared to CCLE. The samples shown are a random subset with equal representation from each dataset in each disease. **c** Boxplots showing overall protein quantification of representative proteins from each of the four cancer driver pathways in TCPA ( $n = 5607$ ) and MCLP ( $n = 646$ ) datasets (two-sided Mann-Whitney significance test; \**P* value < 0.05, \*\**P* value < 0.01, \*\*\*\**P* value < 0.001). **d** Boxplots show pairwise Spearman correlation scores between all CCLE and TCGA RNA-seq samples in each disease type, for all four cancer driver pathways. Plots are sorted by mean correlation.

cell lines derived from mutated tumors maintain those mutations. We found that there is no correlation between mutational burden and tumor-cell-line correlation; in fact, in the case of the TP53 pathway, there is a slight inverse correlation between the two factors (Supplementary Fig. 4). To further investigate the effect of *KRAS* and *TP53* mutation status on the differentially expressed cancer driver pathways, we subset the TCGA and CCLE datasets to samples carrying a non-silent *KRAS* or *TP53* mutation and repeated the SVM analysis (see “Methods”). In both the *KRAS*-mutant and *TP53*-mutant analysis, all four cancer driver pathways were again identified as differentially expressed in tumor compared to cell line. These results indicate that dysregulation of

these pathways occurs in cancer cell lines regardless of the mutational status of the primary tumor, and is unrelated to the activating DNA mutations.

**Dysregulation of a lncRNA-miRNA regulatory network in cancer cell lines is potentially associated with underexpression of key cancer pathways.** Because the four cancer driver pathways were derived in the context of lncRNA-related gene expression, we hypothesized that cell-line-specific dysregulation of lncRNA-based regulation programs may be associated with aberrant pathway-level gene expression. lncRNAs control gene expression



**Fig. 3 Long non-coding RNA associated with four cancer driver pathways are significantly underexpressed in cancer cell lines.** **a** In the “sponge” model of lncRNA gene expression regulation, lncRNA competitively inhibit miRNA which would otherwise be responsible for inhibiting mRNA. **b** Heatmap showing expression of 11 lncRNAs with miRNA-dependent associations to protein-coding genes in the four cancer driver pathways. The samples shown are a random subset with equal representation from each dataset in each disease. **c** Boxplots showing expression of the 11 lncRNA associated with four cancer driver pathways. All samples from both datasets are shown. (Two-sided Mann-Whitney significance test; \* $P$  value  $< 0.05$ , \*\* $P$  value  $< 0.01$ , \*\*\* $P$  value  $< 0.001$ ).

in a tissue-specific manner, and one of their key regulatory mechanisms is by sequestering or “sponging” miRNA through base pairing interactions<sup>23–25</sup>. miRNA directly affect gene expression by binding mRNA and targeting them for degradation<sup>26–28</sup>. In this method of expression control, lncRNA regulate miRNA, while miRNA regulate gene expression (Fig. 3a).

To investigate potential non-coding RNA dysregulation in cell lines as compared to tumors, we focused on the 54 lncRNA identified as differentially expressed through the SVM classification (Fig. 1 and Supplementary Data 1, Tab 1). We used miRNet databases to link the 54 differentially expressed lncRNA to the 4 differentially expressed cancer driver pathways via shared miRNA interactions (Supplementary Data 2). Via miRNet, we found that 77 miRNA have known interactions both with genes in the four cancer driver pathways, and with 11 of the differentially expressed lncRNA (11 lncRNA: *LBX2-AS1*, *CERS6-AS1*, *DLGAP1-AS1*, *H19*, *IQCH-AS1*, *LINC00240*, *LINC00665*, *LINC00707*, *LINC00847*, *LINC00622*, *LIMD1-AS1*). With the exception of *LINC00707*, 10 of the 11 lncRNA are significantly underexpressed in CCLE (Fig. 3b, c). We hypothesized that the reduced cell-line expression of these lncRNA may be associated with expression changes in the downstream miRNA regulatory network, which in turn may be associated with aberrant expression of the four cancer driver pathways being controlled by the miRNA network.

In order to investigate this hypothesis, we leveraged publicly available miRNA sequencing (miRNAseq) data from CCLE and TCGA. We used the ComBat method to correct for experimental batch effects (see “Methods” and Supplementary Fig. 5)<sup>29</sup>. Sixty-nine of the 77 miRNA were quantified in both miRNAseq datasets, so we used these miRNA for all downstream

analyses (Supplementary Data 4, Tab 1). We calculated the log fold change (LFC) in expression between CCLE and TCGA for these 69 miRNA. Notably, over half of the miRNA ( $n = 43$ ) are more highly expressed in cell lines than tumors. Cytoscape was used to visualize the lncRNA-miRNA-coding gene network colored by gene type or by LFC (Fig. 4a, b and Supplementary Data 4, Tab 2)<sup>30</sup>.

In keeping with the lncRNA “sponge” regulatory model, the lncRNA are underexpressed in cancer cell lines, which could be involved with the observed overexpression of a majority of the miRNA whose expression is known to be kept in check by these lncRNA. The aberrant overexpression of inhibitory miRNA could explain the observed underexpression of key genes in four important cancer driver pathways in cancer cell lines. Consistent with this idea, we noted several miRNA with higher expression in cell lines that are known to play roles in regulation of the four cancer driver pathways. *mir-497*, *mir-195*, *mir-148a*, and *mir-152* directly inhibit genes in the KRAS/RAF/MEK/ERK pathway<sup>31,32</sup>. The TP53 pathway is repressed by *mir-339*, and the TP53-associated gene *TP53INP1* is regulated by *mir-92*<sup>33,34</sup>. *mir-519d* directly represses *STAT3*, a key gene in the IL2/STAT5 signaling pathway<sup>35</sup>. The NFkB pathway is activated by *mir-301a*, which has lower expression in cell lines compared to tumors, in keeping with lower NFkB activity in cell lines<sup>36</sup>.

Because lncRNA and miRNA are known for cell-type-specific expression, we hypothesized that the observed dysregulation of lncRNA-miRNA expression networks is caused by biological selection for a subset of cancer cells which are more likely to survive the cell-line derivation process and thrive in a cell culture setting. Consistent with this hypothesis, both stem cell and epithelial cell-specific lncRNA and miRNA display reduced

expression in cancer cell lines. Specifically, CCLE samples have reduced expression of *H19*, a lncRNA strongly associated with the cancer stem cell state<sup>37</sup>, but show increased expression of *mir-1* and *mir-206*, which promote cellular differentiation by blocking anti-differentiation signaling targets<sup>38,39</sup>. In addition, CCLE samples show reduced expression of *CERS6-AS1*, *IQCH-AS1*, and *LINC00240*, lncRNA implicated in mediating tight junctions or extracellular matrix interactions, which are features of epithelial and endothelial cells<sup>40–43</sup>. At the same time, CCLE samples have comparatively high expression of *mir-9*, which directly represses E-cadherin, a well-known epithelial marker<sup>44</sup>. E-cadherin repression is known to induce the epithelial-mesenchymal-transition, a process which plays a role in cancer progression from an epithelial state to a motile and invasive metastatic state<sup>45</sup>. CCLE samples display reduced expression of E-cadherin/*CDH1*, and higher expression of mesenchymal markers, including N-cadherin/*CDH2*, *MUC1*, and claudins *CLDN1*, *CLDN2*, *CLDN3*<sup>46</sup> (Supplementary Fig. 6).

The observed reduced epithelial and stem cell expression in cancer cell lines suggests that cancer cell culture conditions select for the subset of cancer cells with a mesenchymal, invasive and metastatic phenotype. Overall, these results indicate that selection against specific cancer cell types in tumor-derived cell lines may cause global downregulation of key cell-type-specific lncRNA, potentially allowing overexpression of a variety of miRNA, many of which play important roles in regulating cancer signaling pathways. However, more work is needed to fully investigate this hypothesis.

In light of recent research identifying a panel of 110 CCLE cell lines with the highest correlation to their primary tumor samples, the TCGA-110-CL<sup>7</sup>, we examined whether these cell lines show more representative expression of the 4 cancer driver pathways. We repeated the SVM after subsetting the CCLE dataset to the TCGA-110-CL and the TCGA dataset to the tumor types in the TCGA-110-CL (Supplementary Data 1, Tab 6). Interestingly, the same four cancer driver pathways were again identified as differentially expressed (Supplementary Fig. 7). However, several metabolic, cellular response and developmental pathways that were identified in the original analysis were not identified here, including Hedgehog signaling, apical junction, and fatty acid metabolism (Supplementary Data 1, Tab 3). Overall, these results indicate that while the TCGA-110-CL cell-line panel is indeed more representative of its primary tumors by overall gene expression, our pathway-level examination reveals that caution must still be used when interpreting results involving targeting these four cancer driver pathways. This result is consistent with our hypothesis that the dysregulation of cancer driver signaling is driven by a loss of cellular heterogeneity overall in cancer cell lines.

**Single-cell RNA-seq analysis of hepatocellular carcinoma cell lines and tumor samples highlights that the differences in the expression of key cancer pathways are tumor-specific.** In order to further investigate the hypothesis that our results are driven by the selection of a specific malignant cell type in cancer cell line derivation, we leveraged previously published single-cell RNA sequencing (scRNAseq) data from hepatocellular carcinoma (HCC) cell lines and patient samples. Within the scRNAseq patient tumor data, we differentiated the malignant cells from the normal cell infiltrate (e.g., immune and stromal cells; Supplementary Data 3, Tab 1), in order to assess the impact of tumor purity on the observed cancer driver pathway dysregulation.

We used publicly available scRNAseq data from HCC cell lines HuH1 and HuH7<sup>47</sup> and from seven samples biopsied from two different HCC patients<sup>48</sup>. We assigned cell types based on the

expression of published gene markers (Fig. 5a and Supplementary Data 3, Tab 1).

To see whether we could recapitulate the findings of our bulk RNA sequencing data analysis, we performed differential expression analysis between the cell line and tumor-cell populations. GSEA of the top 100 genes identified the KRAS, TP53, TNFA via NFKB, and IL6/STAT3 signaling pathways as enriched in genes differentially expressed in HCC tumor cells ( $P$  value < 0.05), very similar to the bulk RNA sequencing analysis (Fig. 5b).

We then removed the non-malignant cells from the HCC tumor data and repeated the differential expression analysis and GSEA (Fig. 5c and Supplementary Data 3, Tab 2). We observed that the P53 Pathway and TNFA signaling via NFKB remained significantly enriched with  $P$  value < 0.05, whereas the KRAS signaling and IL6/STAT3 signaling pathways were only significantly enriched with a less stringent  $P$  value < 0.1. These results indicate that the observed overexpression of the TNFA signaling via NFKB and P53 Pathway in bulk RNA sequencing data from tumor samples is likely not related to tumor purity, whereas the observed KRAS signaling and IL2/STAT5 signaling overexpression may be in part attributable to lower tumor purity. We were unable to investigate the observed lncRNA and miRNA expression dysregulation due to the low sequencing depth of single-cell RNA sequencing data and low overall expression of non-coding RNA.

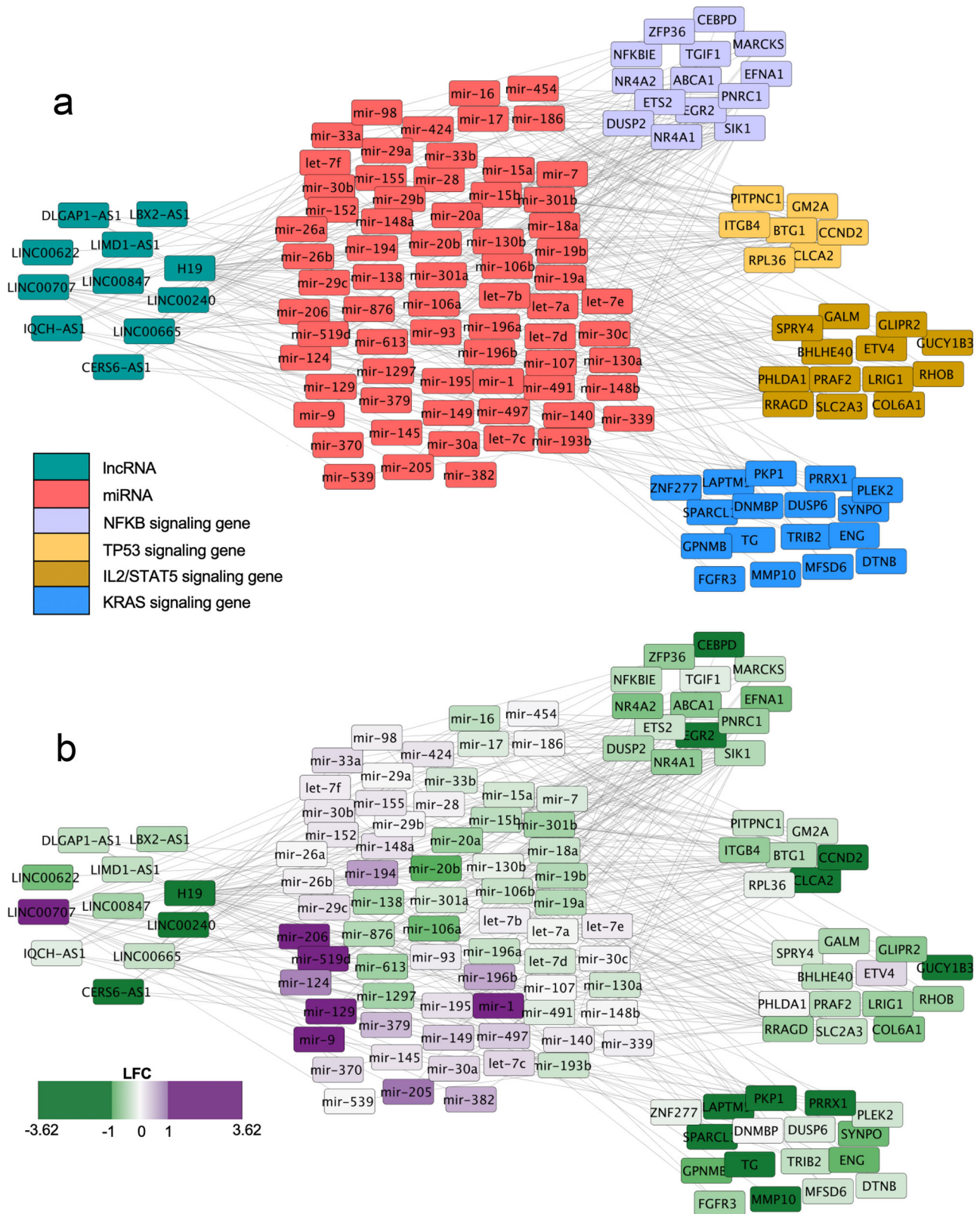
Finally, we leveraged the HCC scRNAseq data to investigate our hypothesis that cancer cell culture conditions select for the subset of cancer cells with a mesenchymal, invasive, and metastatic phenotype. We evaluated three well-known molecular subtypes of HCC tumors: S1 subtype is invasive and characterized by poor survival; S2 subtype tumors are larger, with poor survival; and S3 subtype is lower grade, with overall better survival<sup>49</sup>. These three subtypes are all represented in the seven scRNAseq HCC patient samples<sup>48</sup>.

We wanted to see whether these three subtypes were also fully represented in the HCC cell-line data. We evaluated the expression of gene signatures associated with each subtype in the HCC cell line and malignant tumor cells, and observed higher overall expression of the S1 signature in cell lines, with higher expression of S2 and S3 signatures in the tumor samples (Fig. 5d). The enrichment of each signature in the top 100 differentially expressed genes in each sample is shown in Fig. 5e (Supplementary Data 3, Tab 2). The HCC cell-line population is only enriched for the most invasive, metastatic subtype S1, while the HCC tumor cells display variable enrichment for each subtype. These results are consistent with our hypothesis that cancer cell culture conditions select for the most invasive cell subtypes.

A limitation of the HCC analysis is that cancer and cell-line samples were not matched from the same patients. Therefore, we performed the same analysis using two sets of matched brain tumor and cell-line samples (melanoma brain metastases; Supplementary Figs. 8 and 9, respectively). We again observe enrichment of the same four cancer driver pathways, with the exception of the KRAS signaling pathway in the first analysis.

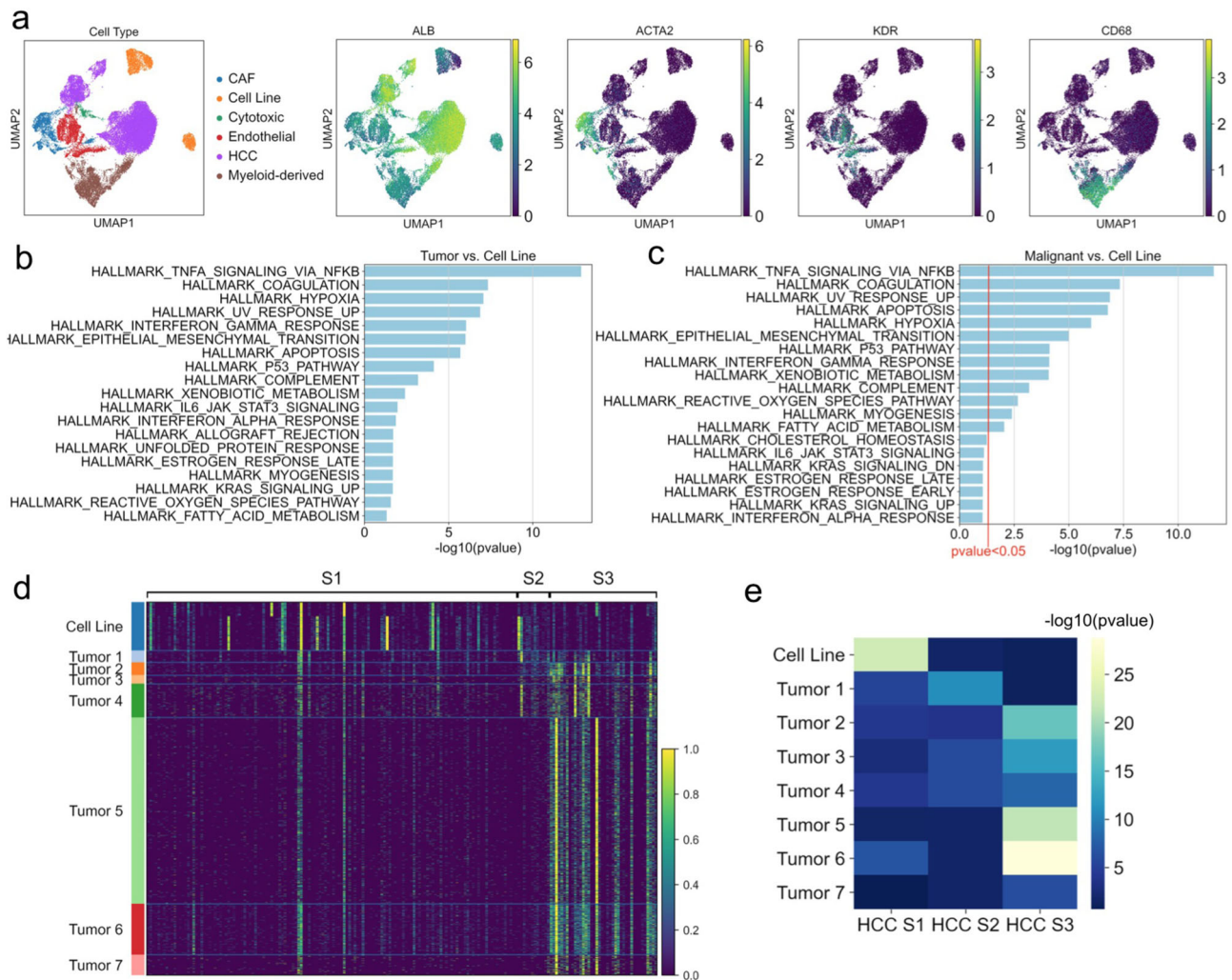
## Discussion

The ability to model and manipulate cancer cells has empowered therapeutic discovery since the derivation of the first cancer cell line<sup>50</sup>. Two-dimensional cancer cell cultures have enabled researchers to discover how cancers arise, characterize cancer cell types and growth patterns, and identify effective pharmaceuticals through drug screens<sup>1</sup>. Today, personalized tumor-derived 2D and 3D cultures are increasingly in use for the identification of precision therapies for individual patients<sup>51–53</sup>. The rise of



**Fig. 4 Dysregulation of lncRNA-miRNA regulatory network causes downregulation of key cancer driver pathways in tumor-derived cell lines. a** Types of genes are identified by color and positioning in the Cytoscape graph. Gene interactions from miRNet databases are denoted by gray lines. lncRNA are on the left, miRNA in the center, and differentially expressed protein-coding genes from each of the four cancer driver pathways are on the right side of the graph. **b** Positive LFC (purple) denotes higher expression in CCLE. Negative LFC (green) denotes higher expression in TCGA.





**Fig. 5** Single-cell RNA sequencing analysis of hepatocellular carcinoma tumor samples and cell lines. **a** Single-cell RNA sequencing data clustering of hepatocellular carcinoma (HCC) cell line and tumor cells. **b** Gene set enrichment analysis of genes differentially overexpressed in all HCC tumor cells compared to HCC cell-line cells ( $P$  value < 0.05). **c** Gene set enrichment analysis of genes differentially overexpressed in only malignant HCC tumor cells compared to HCC cell-line cells ( $P$  value < 0.1, red line denotes  $P$  value < 0.05 cutoff). **d** Heatmap showing relative expression of genes from HCC molecular subtype gene signatures, in HCC cell line and tumor single-cell samples. **e** Heatmap showing  $-\log_{10} P$  values for enrichment of differentially expressed genes in each HCC sample overlapping with the gene signatures for each HCC molecular subtype.

precision medicine and small-molecule inhibitor development brings an urgent need for molecular characterization of the cell culture models used widely for therapeutic development. Large-scale genomic efforts such as The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE) have enabled comprehensive comparison of cancer cell cultures and primary tumors.

Here, we provide a comparative multi-omic analysis of cancer cell lines and primary tumors by leveraging several types of genomic data from large public compendia. Our gene expression analysis reveals reduced expression of key cancer driver pathways, including KRAS signaling, TP53 pathway, IL2/STAT5 signaling, and NF $\kappa$ B signaling in cancer cell lines. These results are recapitulated in a comparative analysis of protein levels for key proteins from each pathway. Our analysis indicates the need for caution when interpreting in vitro preclinical testing results of inhibitors targeting these pathways.

In fact, the consequences of preclinical testing in cell lines against these pathways have already been felt. For example, during preclinical testing several MDM2/TP53 small-molecule inhibitors displayed potent cancer cell-line inhibitory activity, but

could only achieve partial tumor regression in xenograft models<sup>54</sup>. Subsequent optimization of these compounds led to success in clinical trials, but initial results in cell lines did not predict in vivo results. Targeting both wild-type and mutant KRAS in cancer has been notably unsuccessful; in particular, several high-throughput screens of KRAS-mutant cancer cell lines identified compounds which subsequently only partially reduced tumor volumes in xenograft models<sup>55,56</sup>. In contrast, targeting the cholesterol biosynthesis pathway, which is not differentially expressed in cancer cell lines by our analysis, has been promising both in vitro and in vivo through the usage of the statin family of drugs<sup>57,58</sup>. In addition, targeted inhibition of cyclin-dependent kinases (CDKs) by agents such as PD-0332991/palbociclib has been promising in phase I and II clinical trials, and this drug was originally identified in cancer cell lines<sup>59,60</sup>. CDKs are active during cell cycle entry from G0 and during the G2M checkpoint pathway, which were not differentially expressed in cancer cell lines by our analysis. We note that apart from the four cancer driver pathways identified as dysregulated in cell lines, preclinical testing in cell lines can in many cases reliably predict in vivo responses.

To identify potential causes of dysregulation of cancer driver signaling in cell lines, we analyzed the expression of lncRNA and miRNA implicated in regulation of these pathways. In cell lines, we found reduced expression of a set of lncRNA predicted to regulate a downstream network of regulatory miRNA, which are in turn overexpressed. Several of these miRNA are directly involved in specific inhibition of these cancer driver pathways, linking their overexpression to the observed reduced expression of cancer driver pathways.

We speculate that our results may indicate a partial loss of cancer stem cells in cancer cell culture due to the presence of serum in culture media. All CCLE cell lines were cultured in RPMI or DMEM media with 10% fetal bovine serum<sup>61</sup>. It is well known that the presence of serum in culture media encourages cellular differentiation, and cancer cell lines grown in serum-free conditions contain larger populations of cancer stem cells<sup>62,63</sup>. This hypothesis is supported by the markedly lower expression of stem cell-specific lncRNA in cell culture, and higher expression of pro-differentiation miRNA. Because cancer stem cell populations are known for their chemoresistance, and even small populations are thought to be capable of tumor recurrence<sup>64–66</sup>, it is essential that preclinical models accurately model the response of cancer stem cells to potential therapeutics.

A recent study by Yu et al. also included a pan-cancer analysis of TCGA and CCLE data, finding that tumor type is not a large factor in overall tumor-cell-line correlation, but that some cell lines are poor models of their primary tumor type and tumor purity plays a large part in tumor-cell-line correlations<sup>7</sup>. Our study provides additional information by identifying the set of genes which accounts for the highest variation between human tumors and cell culture models, and by identifying specific, clinically important cancer driver pathways which are poorly recapitulated in cancer cell culture models. Additionally, our finding of global cancer driver pathway dysregulation in cell lines may help to explain biologically the previous study's finding of low correlation between some tumors and cell lines. We investigated the TCGA-110-CL cancer cell-line panel identified in the previous study, which has a higher overall cell-line-tumor gene expression correlation. The same four cancer driver pathways were identified by an SVM comparing these cell lines to disease-matched primary tumors, and in particular, the KRAS pathway had the lowest overall cell-line-tumor correlation. This indicates that even this subset of cancer cell lines does not fully recapitulate the cancer driver signaling of primary tumors, and studies on these pathways must be interpreted with caution.

In light of this finding that tumor purity impacts the correlation between tumor and cell line<sup>7</sup>, we sought to assess the impact of tumor purity on our cancer driver pathway dysregulation findings by leveraging single-cell RNA sequencing data from HCC patients and cell lines. The TP53 and NFKB pathways remained significantly enriched in HCC tumor cells compared to cell lines when normal infiltrating cells were removed from the tumor-cell population, while the KRAS and IL6/STAT3 pathways were only significantly enriched with a less stringent statistical cutoff. These results indicate that the KRAS and IL2/STAT5 findings from the TCGA-CCLE gene expression comparison may be partially due to lower tumor purity. Moreover, this analysis indicates that normal cell infiltrate contributes a molecular signature to bulk gene expression data and highlights the utility of single-cell data. Finally, we found that the HCC cell lines most closely resemble the most invasive and metastatic HCC subtype, unlike the HCC tumor cells which resemble all three HCC subtypes, consistent with our hypothesis that cancer cell culture selects for the most invasive malignant cells. We saw similar results in an analysis of matched melanoma brain metastasis cell lines and tumors. A caveat of this analysis is that it focuses on only two types of solid tumor; a

broader investigation of single-cell RNA sequencing data in this context from multiple tumor types and cell lines will be informative when those data become available.

Taken together, our results underscore the need for caution when interpreting preclinical cancer testing results in multiple model types, and point to specific signaling networks which can serve as litmus tests for the accuracy of past and future cancer laboratory models. We suggest several potential solutions to improve the efficacy of tumor-derived cell lines. Cancer cell culture in serum-free conditions may improve the maintenance of tumor stem cell populations and reverse the dysregulation of important regulatory gene networks. Specific efforts to model the immune microenvironment in cancer-derived organoids may improve cancer driver pathway expression related to the tumor microenvironment. A potential solution may be genetic manipulation of tumor-derived models with an emphasis on preserving or rescuing the intrinsic cancer driver pathway expression which is most at risk for dysregulation. Overall, this study provides much-needed genomics-based considerations for future pre-clinical cancer model development and result interpretation.

## Methods

**RNA sequencing data.** Gene expression transcripts per million (TPM) matrices from TCGA ( $n$  samples = 10,535) and CCLE ( $n$  samples = 933) were downloaded from the UCSC Xena browser. These data were processed uniformly through the TOIL UCSC RNA sequencing data processing pipeline (TCGA on v2.0.8; CCLE on v3.3.4) to remove technical batch effects<sup>67</sup>. Both datasets were normalized by  $\log_2(\text{TPM} + 1)$  and duplicate genes were averaged. Genes not expressed in 80% of samples were removed, and 20% of the lowest varying remaining genes were removed, leaving 46865 remaining genes. Both datasets were subset to the 19 overlapping cancer types for subsequent analysis (BRCA, LUSC, LIHC, DLBC, THCA, PRAD, OV, STAD, BLCA, KIRC, UCEC, COAD, SARC, CESC, SKCM, PAAD, HNSC, ESCA, GBM). All heatmaps use a random subset of samples from each dataset with equal numbers from each disease. The random subset method is used because the CCLE data holds approximately 1/10 the number of samples as the TCGA data, making it impossible to visually detect gene expression differences in heatmaps containing all samples from both datasets.

**Mutation data.** TCGA somatic mutation data were downloaded from UCSC Xena ([xena.ucsc.edu](http://xena.ucsc.edu)). CCLE somatic mutation data were downloaded from the Broad Institute CCLE Database ([portals.broadinstitute.org/ccle/data](http://portals.broadinstitute.org/ccle/data)).

**micro-RNA sequencing data.** TCGA micro-RNA (miRNA) Illumina sequencing read counts data were downloaded from the Genomic Data Commons Data Portal<sup>68</sup>. CCLE Nanostring probe miRNA quantification data were downloaded from the Broad Institute CCLE database: <https://portals.broadinstitute.org/ccle/data><sup>69</sup>. For dataset comparability, miRNA naming formats were harmonized, and duplicates were averaged. Because different miRNA sequencing methods were used in each dataset, ComBat was used to batch-correct the data<sup>29</sup>. Pre- and post-batch effect correction data were then  $\log_2(\text{count} + 1)$  normalized for downstream visualization and analysis. Supplementary Fig. 5 shows pre- and post-batch effect correction expression distributions of several housekeeping genes to validate successful correction<sup>70,71</sup>.

**RPPA data.** Level 4 Reverse Phase Protein Array (RPPA) data for the TCGA and CCLE samples were downloaded from the The Cancer Proteome Atlas (TCPA) portal (<https://tcpaportal.org/tcpa/download.html> and <http://tcpaportal.org/mclp/#/download>). Both datasets were subset to the 16 overlapping cancer types for subsequent analysis (BLCA, BRCA, COAD, DLBC, HNSC, KIRC, LGG, LIHC, LUAD, OV, PAAD, PRAD, SARC, SKCM, STAD).

**Single-cell RNA sequencing data.** Hepatocellular carcinoma (HCC) circulating tumor cell and cell line 10x Genomics single-cell RNA sequencing datasets were downloaded from the Gene Expression Omnibus (accession GSE103867). The cell-line cells were selected based on the expression of *ALDH1A1*, based on the original study findings (see Fig. 4, Zheng et al.<sup>23</sup>). HCC tumor 10x Genomics single-cell RNA sequencing datasets from seven samples from two patients were downloaded from the Gene Expression Omnibus (accession GSE112271). All single-cell gene expression data analysis was performed using scanpy (v1.6.7) in Python (v3.6.8). The datasets were combined, and cells expressing >5000 genes and/or >30 mitochondrial genes were excluded, leaving 52,630 cells. Leiden clustering was performed using the top 40 Principal Components, cell types assigned based on cell-type markers (Supplementary Data 3, Tab 1), and the top 100 upregulated genes in the primary tumors were derived by a Wilcoxon rank-sum test. Primary patient-

derived melanoma brain metastasis tissue (CI0000035650 and CI0000035270) were obtained from the UAMS Tissue Biorepository and Procurement Core within a few hours of surgical removal where tumors were cultured and bio-banked. Consent was obtained from all patients prior to surgery (IRB# 228443). A board-certified neuropathologist confirmed all tumor histopathological diagnoses. Tumors were harvested intraoperatively with myriad device which mechanically dissociates the tumor (NICO corporation, Indianapolis, IN, USA). Tumor specimens were then digested with collagenase solution. Cells were then prepared for single-cell sequencing using the 10x Genomics platform (10x Genomics, Pleasanton, CA, USA) as previously described<sup>72</sup>. The remaining cells, following digestion, were cultured in Roswell Park Memorial Institute culture medium (RPMI) containing 10% fetal bovine serum (FBS) and 1% (v/v) antibiotic/antimitotic. Early passage cells (<P3) were grown to confluency and then harvested for single-cell RNA sequencing analysis as described above.

**Statistical gene selection via support vector machine.** In Python (v3.6.8), the sklearn module (v0.21.2) was used with linear kernel to train a support vector machine (SVM) on 50 random 80/20 splits of the merged TCGA-CCLE gene expression dataset, in which 50 different training sets (and corresponding test sets) were resampled from the same data. The top 10% of genes from each training run based on the magnitude of their feature weights coefficients were merged in a non-duplicate manner, to account for variation in prediction results based on the training and corresponding test set, resulting in 1858 genes. Gene set enrichment from this analysis revealed 26/100 enriched pathways were immune-related, so a non-redundant immune gene list was created by merging all genes from the 26 enriched immune pathways (Supplementary Data 1, Tab 1). A second SVM analysis was conducted on a set of 50 random 80/20 splits of the merged TCGA-CCLE gene expression dataset with the immune-related gene list removed, and the top 10% of genes from each run based on feature importance coefficients were merged, resulting in 1854 genes which were used in all downstream analysis. We verified that the cancer driver pathway signal identified in the second SVM analysis is not related to immune signaling by calculating pairwise correlation between SVM-identified cancer driver genes and the immune genes which were removed from the analysis (mean = 0.1, Supplementary Fig. 1).

**Statistics and reproducibility.** Expression comparisons of individual genes between datasets were performed using a two-sided Mann-Whitney significance test with significance defined as  $P$  value < 0.05. Significantly enriched gene sets were identified using gene set enrichment analysis<sup>41</sup> with FDR  $q$ -value < 0.05. For bulk RNA-seq, heatmaps are generated using seaborn clustermap method in Python with  $z$ -scores calculated per gene. For single-cell RNA-seq, heatmaps are generated using scanpy heatmap method in Python with values scaled between 0 and 1. Figures were produced using pandas v0.25.3, matplotlib v3.0.3, and seaborn v0.9.0 in Python v3.6.8. Numerical data used to generate figures are available here: [https://drive.google.com/drive/u/1/folders/16pqRgDlx50Wj\\_H9rIwYjB11a0IE6TAXH](https://drive.google.com/drive/u/1/folders/16pqRgDlx50Wj_H9rIwYjB11a0IE6TAXH) All data were used for each figure, except when otherwise indicated.

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

## Data availability

The TCGA and CCLE processed gene expression data are publicly available at UCSC Xena ([xena.ucsc.edu](http://xena.ucsc.edu)). The processed micro-RNA sequencing data are publicly available at the Genomic Data Commons and the Broad Institute CCLE database ([portals.broadinstitute.org/ccle/data](http://portals.broadinstitute.org/ccle/data)). The protein quantification data are publicly available at the The Cancer Proteome Atlas (TCPA) portal ([tcpaportal.org/tcpa/download.html](http://tcpaportal.org/tcpa/download.html) and [tcpaportal.org/mclp/#/download](http://tcpaportal.org/mclp/#/download)). The mutation data are publicly available at UCSC Xena and the Broad Institute CCLE database. The hepatocellular carcinoma single-cell RNA sequencing data are publicly available at the following accessions: GSE103867 and GSE112271. The brain tumor single-cell RNA sequencing data are publicly available at the following Gene Expression Omnibus accession: GSE213519.

Received: 17 June 2020; Accepted: 6 October 2022;

Published online: 13 December 2022

## References

- Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6**, 813–823 (2006).
- Ross, D. T. & Perou, C. M. A comparison of gene expression signatures from breast tumors and breast tissue derived cell lines. *Dis. Markers* **17**, 99–109 (2001).

- Domcke, S., Sinha, R., Levine, D. A., Sander, C. & Schultz, N. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat. Commun.* **4**, 2126 (2013).
- Li, H. et al. Genomic analysis of head and neck squamous cell carcinoma cell lines and human tumors: a rational approach to preclinical model selection. *Mol. Cancer Res.* **12**, 571–582 (2014).
- Mouradov, D. et al. Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer. *Cancer Res.* **74**, 3238–3247 (2014).
- Goodspeed, A., Heiser, L. M., Gray, J. W. & Costello, J. C. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. *Mol. Cancer Res.* **14**, 3–13 (2016).
- Yu, K. et al. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat. Commun.* **10**, 3574 (2019).
- Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- Furey, T. S. et al. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **16**, 906–914 (2000).
- George, V. S. & Raj, C. Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile. *Int. J. Comput. Sci. Eng. Surv.* **2**, 16–27 (2011).
- Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
- Fan, Y. & Xia, J. miRNet—functional analysis and visual exploration of miRNA–target interactions in a network context. in *Computational Cell Biology: Methods and Protocols* (eds. von Stechow, L. & Santos Delgado, A.) 215–233 (Springer New York, 2018).
- Park, J., Lee, J. & Choi, C. Evaluation of drug-targetable genes by defining modes of abnormality in gene expression. *Sci. Rep.* **5**, 13576 (2015).
- Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
- Li, J. et al. Characterization of human cancer cell lines by reverse-phase protein arrays. *Cancer Cell* **31**, 225–239 (2017).
- Li, J. et al. TCPA: a resource for cancer functional proteomics data. *Nat. Methods* **10**, 1046–1047 (2013).
- Eser, S., Schnieke, A., Schneider, G. & Saur, D. Oncogenic KRAS signalling in pancreatic cancer. *Br. J. Cancer* **111**, 817–822 (2014).
- Eser, S. et al. Selective requirement of PI3K/PDK1 signaling for Kras oncogene-driven pancreatic cell plasticity and cancer. *Cancer Cell* **23**, 406–420 (2013).
- Sasaki, C. Y., Barberi, T. J., Ghosh, P. & Longo, D. L. Phosphorylation of RelA/p65 on serine 536 defines an I $\kappa$ B $\alpha$ -independent NF- $\kappa$ B pathway. *J. Biol. Chem.* **280**, 34538–34547 (2005).
- Wang, J. et al. Activation of NF- $\kappa$ B by TMPRSS2/ERG fusion isoforms through toll-like receptor-4. *Cancer Res.* **71**, 1325–1333 (2011).
- Gemignani, M. L. et al. Role of KRAS and BRAF gene mutations in mucinous ovarian carcinoma. *Gynecol. Oncol.* **90**, 378–381 (2003).
- Zhu, J. et al. Gain-of-function p53 mutants co-opt chromatin pathways to drive cancer growth. *Nature* **525**, 206–211 (2015).
- Liz, J. & Esteller, M. lncRNAs and microRNAs with a role in cancer development. *Biochim. Biophys. Acta* **1859**, 169–176 (2016).
- Militello, G. et al. Screening and validation of lncRNAs and circRNAs as miRNA sponges. *Brief. Bioinform.* **18**, 780–788 (2017).
- Paraskevopoulou, M. D. & Hatzigeorgiou, A. G. Analyzing MiRNA–lncRNA interactions. in *Long Non-Coding RNAs: Methods and Protocols* (eds. Feng, Y. & Zhang, L.) 271–286 (Springer New York, 2016).
- Gregory, R. I. & Shiekhattar, R. MicroRNA biogenesis and cancer. *Cancer Res.* **65**, 3509–3512 (2005).
- Huang, Y. et al. Biological functions of microRNAs: a review. *J. Physiol. Biochem.* **67**, 129–139 (2011).
- Cai, Y., Yu, X., Hu, S. & Yu, J. A brief review on the mechanisms of miRNA regulation. *Genomics Proteom. Bioinforma.* **7**, 147–154 (2009).
- Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
- Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
- Li, D. et al. Analysis of MiR-195 and MiR-497 expression, regulation and role in breast cancer. *Clin. Cancer Res.* **17**, 1722–1730 (2011).
- Xu, Q. et al. A regulatory circuit of miR-148a/152 and DNMT1 in modulating cell transformation and tumor angiogenesis through IGF-IR and IRS1. *J. Mol. Cell Biol.* **5**, 3–13 (2013).
- Jansson, M. D., Damas, N. D., Lees, M., Jacobsen, A. & Lund, A. H. miR-339-5p regulates the p53 tumor-suppressor pathway by targeting MDM2. *Oncogene* **34**, 1908–1918 (2015).

34. Yeung, M. L. et al. Roles for microRNAs, miR-93 and miR-130b, and tumor protein 53-induced nuclear protein 1 tumor suppressor in cell growth dysregulation by human T-cell lymphotropic virus 1. *Cancer Res.* **68**, 8976–8985 (2008).
35. Deng, X., Zhao, Y. & Wang, B. miR-519d-mediated downregulation of STAT3 suppresses breast cancer progression. *Oncol. Rep.* **34**, 2188–2194 (2015).
36. Lu, Z. et al. miR-301a as an NF- $\kappa$ B activator in pancreatic cancer cells. *EMBO J.* **30**, 57–67 (2011).
37. Peng, F. et al. H19/let-7/LIN28 reciprocal negative regulatory circuit promotes breast cancer stem cell maintenance. *Cell Death Dis.* **8**, e2569 (2017).
38. Sweetman, D. et al. Specific requirements of MRFs for the expression of muscle specific microRNAs, miR-1, miR-206 and miR-133. *Dev. Biol.* **321**, 491–499 (2008).
39. Goljanek-Whysall, K. et al. Regulation of multiple target genes by miR-1 and miR-206 is pivotal for C2C12 myoblast differentiation. *J. Cell Sci.* **125**, 3590–3600 (2012).
40. Wang, L. et al. Identifying the crosstalk of dysfunctional pathways mediated by lncRNAs in breast cancer subtypes. *Mol. Biosyst.* **12**, 711–720 (2016).
41. Yang, L. et al. Genome-wide identification of long non-coding RNA and mRNA profiling using RNA sequencing in subjects with sensitive skin. *Oncotarget* **8**, 114894–114910 (2017).
42. Yang, S. et al. Construction of differential mRNA-lncRNA crosstalk networks based on ceRNA hypothesis uncover key roles of lncRNAs implicated in esophageal squamous cell carcinoma. *Oncotarget* **7**, 85728–85740 (2016).
43. Salvador, E., Burek, M. & Förster, C. Y. Tight junctions and the tumor microenvironment. *Curr. Pathobiol. Rep.* **4**, 135–145 (2016).
44. Ma, L. et al. miR-9, a MYC/MYCN-activated microRNA, regulates E-cadherin and cancer metastasis. *Nat. Cell Biol.* **12**, 247–256 (2010).
45. Yang, J. & Weinberg, R. A. Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis. *Dev. Cell* **14**, 818–829 (2008).
46. Redmer, T. et al. E-cadherin is crucial for embryonic stem cell pluripotency and can replace OCT4 during somatic cell reprogramming. *EMBO Rep.* **12**, 720–726 (2011).
47. Zheng, H. et al. Single-cell analysis reveals cancer stem cell heterogeneity in hepatocellular carcinoma. *Hepatology* **68**, 127–140 (2018).
48. Losic, B. et al. Intratumoral heterogeneity and clonal evolution in liver cancer. *Nat. Commun.* **11**, 291 (2020).
49. Hoshida, Y. et al. Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res.* **69**, 7385–7392 (2009).
50. Gillet, J.-P., Varma, S. & Gottesman, M. M. The clinical relevance of cancer cell lines. *J. Natl Cancer Inst.* **105**, 452–458 (2013).
51. Phan, N. et al. A simple high-throughput approach identifies actionable drug sensitivities in patient-derived tumor organoids. *Commun. Biol.* **2**, 78 (2019).
52. Pauli, C. et al. Personalized in vitro and in vivo cancer models to guide precision medicine. *Cancer Discov.* **7**, 462–477 (2017).
53. Kodack, D. P. et al. Primary patient-derived cancer cells and their potential for personalized cancer patient care. *Cell Rep.* **21**, 3298–3309 (2017).
54. Zhao, Y., Aguilar, A., Bernard, D. & Wang, S. Small-molecule inhibitors of the MDM2-p53 protein-protein interaction (MDM2 Inhibitors) in clinical trials for cancer treatment. *J. Med. Chem.* **58**, 1038–1052 (2015).
55. Kempf, E., Rousseau, B., Besse, B. & Paz-Ares, L. KRAS oncogene in lung cancer: focus on molecularly driven clinical trials. *Eur. Respir. Rev.* **25**, 71–76 (2016).
56. Wang, Y., Kaiser, C. E., Frett, B. & Li, H.-Y. Targeting mutant KRAS for anticancer therapeutics: a review of novel small molecule modulators. *J. Med. Chem.* **56**, 5219–5230 (2013).
57. Sleijfer, S., van der Gaast, A., Planting, A. S. T., Stoter, G. & Verweij, J. The potential of statins as part of anti-cancer treatment. *Eur. J. Cancer* **41**, 516–522 (2005).
58. Cho, S.-J. et al. Simvastatin induces apoptosis in human colon cancer cells and in tumor xenografts, and attenuates colitis-associated colon cancer in mice. *Int. J. Cancer* **123**, 951–957 (2008).
59. Fry, D. W. et al. Specific inhibition of cyclin-dependent kinase 4/6 by PD 0332991 and associated antitumor activity in human tumor xenografts. *Mol. Cancer Ther.* **3**, 1427–1438 (2004).
60. Cicenas, J. & Valius, M. The CDK inhibitors in cancer research and therapy. *J. Cancer Res. Clin. Oncol.* **137**, 1409–1418 (2011).
61. Barretina, J. et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
62. Chin, A. C. P., Padmanabhan, J., Oh, S. K. W. & Choo, A. B. H. Defined and serum-free media support undifferentiated human embryonic stem cell growth. *Stem Cells Dev.* **19**, 753–761 (2010).
63. Lee, J. et al. Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell* **9**, 391–403 (2006).
64. Clarke, M. F. et al. Cancer stem cells—perspectives on current status and future directions: AACR Workshop on cancer stem cells. *Cancer Res.* **66**, 9339–9344 (2006).
65. Clevers, H. The cancer stem cell: premises, promises and challenges. *Nat. Med.* **17**, 313–319 (2011).
66. Sakaguchi, M. et al. miR-137 regulates the tumorigenicity of colon cancer stem cells through the inhibition of DCLK1. *Mol. Cancer Res.* **14**, 354–362 (2016).
67. Vivian, J. et al. Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **35**, 314–316 (2017).
68. Chu, A. et al. Large-scale profiling of microRNAs for The Cancer Genome Atlas. *Nucleic Acids Res.* **44**, e3 (2016).
69. Ghandi, M. et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* <https://doi.org/10.1038/s41586-019-1186-3> (2019).
70. Masè, M. et al. Selection of reference genes is critical for miRNA expression analysis in human cardiac tissue. A focus on atrial fibrillation. *Sci. Rep.* **7**, 41127 (2017).
71. Shen, Y. et al. Identification of miR-23a as a novel microRNA normalizer for relative quantification in human uterine cervical tissues. *Exp. Mol. Med.* **43**, 358–366 (2011).
72. Couturier, C. P. et al. Single-cell RNA-seq reveals that glioblastoma recapitulates a normal neurodevelopmental hierarchy. *Nat. Commun.* **11**, 3406 (2020).
73. Chen, X. et al. Targeting oxidative stress in embryonal rhabdomyosarcoma. *Cancer Cell* **24**, 710–724 (2013).

## Acknowledgements

This study was funded by American Association for Cancer Research NextGen Grant for Transformative Cancer Research Award (O.M.V.), St Baldrick's Foundation Consortium Award and Emily Beazley Kures for Kids Fund Hero Award (D.H.), Alex's Lemonade Stand Foundation for Childhood Cancer Research, Unravel Pediatric Cancer, Team G Childhood Cancer Foundation, the California Initiative to Advance Precision Medicine and Live for Others Foundation. D.H. is a Howard Hughes Medical Institute Investigator. O.M.V. holds a Colligan Presidential Chair in Pediatric Genomics. This study makes use of data generated by the St. Jude Children's Research Hospital—Washington University Pediatric Cancer Genome Project; the St. Jude Rhabdomyosarcoma data were part of SJC-DS-1001, which was obtained with permission from St. Jude Cloud (<https://www.stjude.cloud>)—a publicly accessible pediatric genomic data resource requiring approval for controlled data access: Chen et al.<sup>73</sup>.

## Author contributions

L.M.S. contributed to conception, design, interpretation, and writing. R. Chandra, D.V., A.G.L., N.Z., and H.C.B. contributed data analysis. E.T.K., J.P., A.C., K.L., R. Currie, and A.R. contributed to data acquisition. L.G., D.H., S.R.S., and O.M.V. contributed interpretation and manuscript revision.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s42003-022-04075-4>.

**Correspondence** and requests for materials should be addressed to Lauren M. Sanders or Olena M. Vaske.

**Peer review information** *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Erik Knudsen and Luke R. Grinham. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022