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 ($E_{i,j} > 3.5$) 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.