RESEARCH ARTICLE

Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers

David J. Wooten[1,2☯], **Sarah M. Groves**[2☯], **Darren R. Tyson**[2], **Qi Liu**[3], **Jing S. Lim**[4], **Réka Albert**[1], **Carlos F. Lopez**[2], **Julien Sage**[4], **Vito Quaranta**[2]*

1 Department of Physics, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **2** Department of Biochemistry, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, **3** Departments of Biomedical Informatics and Biostatistics, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, **4** Departments of Pediatrics and Genetics, Stanford University, Stanford, California, United States of America

☯ These authors contributed equally to this work.
* vito.quaranta@vanderbilt.edu

Citation: Wooten DJ, Groves SM, Tyson DR, Liu Q, Lim JS, Albert R, et al. (2019) Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers. PLoS Comput Biol 15(10): e1007343. https://doi.org/10.1371/journal.pcbi.1007343

Editor: Philip K Maini, Oxford, UNITED KINGDOM

Received: March 22, 2019

Accepted: August 19, 2019

Published: October 31, 2019

Data Availability Statement: All TKO mouse scRNAseq data files are available from the GEO database under the accession number GSE137749.

Funding: VQ received funding from National Institutes of Health National Cancer Institute U54-CA217450 (https://csbconsortium.org/). VQ and CL received funding from NIH NCI U01-CA215845 (https://csbconsortium.org/). JS received funding from NIH R01 CA201513 (https://grants.nih.gov/grants/funding/r01.htm). CL received funding from NSF MCB-1411482 (https://nsf.gov/mcb). RA received funding from NSF MCB-1715826

Abstract

Adopting a systems approach, we devise a general workflow to define actionable subtypes in human cancers. Applied to small cell lung cancer (SCLC), the workflow identifies four subtypes based on global gene expression patterns and ontologies. Three correspond to known subtypes (SCLC-A, SCLC-N, and SCLC-Y), while the fourth is a previously undescribed ASCL1+ neuroendocrine variant (NEv2, or SCLC-A2). Tumor deconvolution with subtype gene signatures shows that all of the subtypes are detectable in varying proportions in human and mouse tumors. To understand how multiple stable subtypes can arise within a tumor, we infer a network of transcription factors and develop BooleaBayes, a minimally-constrained Boolean rule-fitting approach. *In silico* perturbations of the network identify master regulators and destabilizers of its attractors. Specific to NEv2, BooleaBayes predicts ELF3 and NR0B1 as master regulators of the subtype, and TCF3 as a master destabilizer. Since the four subtypes exhibit differential drug sensitivity, with NEv2 consistently least sensitive, these findings may lead to actionable therapeutic strategies that consider SCLC intratumoral heterogeneity. Our systems-level approach should generalize to other cancer types.

Author summary

Small-cell lung cancer (SCLC) is an extremely aggressive disease with poor prognosis. Despite significant advances in treatments of other cancer types, therapeutic strategies for SCLC have remained unchanged for decades. We hypothesize that distinct SCLC subtypes with differential drug sensitivities may be responsible for poor treatment outcomes. To this end, we applied a computational pipeline to identify and characterize SCLC subtypes. We found four subtypes, including one (termed "NEv2") that had not previously been reported. Across a broad panel of drugs, we show that NEv2 is more resistant than other

(https://nsf.gov/mcb). SMG received funding from NSF fellowship DGE-1445197 (https://www.nsfgrfp.org/). DJW received funding from NIH T32 5T32CA009592-27 (http://grants.nih.gov/training/T_Table.htm). JSL received a scholarship from the Agency for Science, Technology and Research (A*STAR) in Singapore (https://www.a-star.edu.sg/). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: J.S. receives research funding from StemCentrx/Abbvie, Revolution Medicines, and from Pfizer, and owns stock in Forty Seven Inc.

SCLC subtypes, suggesting that this subtype may be partly responsible for poor treatment outcomes. Importantly, we validate the existence of NEv2 cells in both human and mouse tumors. Reprogramming the identity of NEv2 cells into other subtypes may sensitize these cells to existing treatments. However, deciphering global mechanisms that regulate different subtypes is generally unfeasible. To circumvent this, we developed BooleaBayes, a modeling approach that only infers local regulatory mechanisms near stable cell subtypes. Using BooleaBayes, we found master regulators and master destabilizers for each subtype. These findings predict targets that may destabilize a particular subtype, including NEv2, and lead to successful therapy, by either knocking out master regulators or turning on master destabilizers.

Introduction

A major barrier to effective cancer treatment is the occurrence of heterogeneous cell subpopulations that arise within a tumor via genetic or non-genetic mechanisms. Clonal evolution of these subpopulations via plasticity, drug-induced selection, or transdifferentiation allows tumors to evade treatment and relapse in a therapy-resistant manner. Characterizing cancer subpopulations, or subtypes, has led to breakthrough targeted treatments that significantly improve patient outcomes, as in the case of melanoma [1], breast [2], and lung cancer [3]. However, approaches to subtype identification suffer from several limitations, including: i) focus on biomarkers, which frequently possess insufficient resolving power; ii) lack of consideration for the system dynamics of the tumor as a whole; and iii) often phenomenological, rather than mechanistic, explanations for subtype sources.

To accelerate progress in cancer subtype identification, we set out to develop a general systems-level approach that considers underlying molecular mechanisms to generate multiple stable subtypes within a histological cancer type. We focused on gene regulatory networks (GRNs) comprised of key transcription factors (TFs) that could explain the rise, coexistence and possibly transdifferentiation of subtypes. To enumerate subtypes, identify key regulating TFs, and predict reprogramming strategies for these subtypes, we established the workflow shown in Fig 1. Briefly, we use consensus clustering and weighted gene co-expression network analysis on transcriptomics data to identify cancer subtypes distinguished by gene expression signatures, biological ontologies, and drug response. We validate the existence of the subtypes in both human and mouse tumors using CIBERSORT [4] and nearest neighbor analyses, and develop a GRN that can explain the existence of multiple stable subtypes within a tumor. We then introduce BooleaBayes, a Python-based algorithm to infer partially constrained regulatory interactions from steady state gene expression data. Applied to this GRN, BooleaBayes identifies and ranks master regulators and master destabilizers of each subtype. In a nutshell, starting from transcriptomics data, the workflow can predict reprogramming strategies to improve efficacy of treatment.

We applied this workflow to Small Cell Lung Cancer (SCLC), in which genetic aberrations cannot fully distinguish subtypes [5], or point toward a targeted therapy. SCLC treatment has instead remained cytotoxic chemotherapy (a regimen of etoposide and a platinum-based agent such as cisplatin) and radiation for over half a century, despite the fact that virtually all patients relapse after therapy. This has caused SCLC to be designated as a recalcitrant cancer by the Recalcitrant Cancer Research Act of 2012, with five year survival rates less than 5%.

Recently, efforts to stratify patients have led to the recognition of phenotypic heterogeneity within and between SCLC tumors, raising hopes for more efficient subtype-based treatment

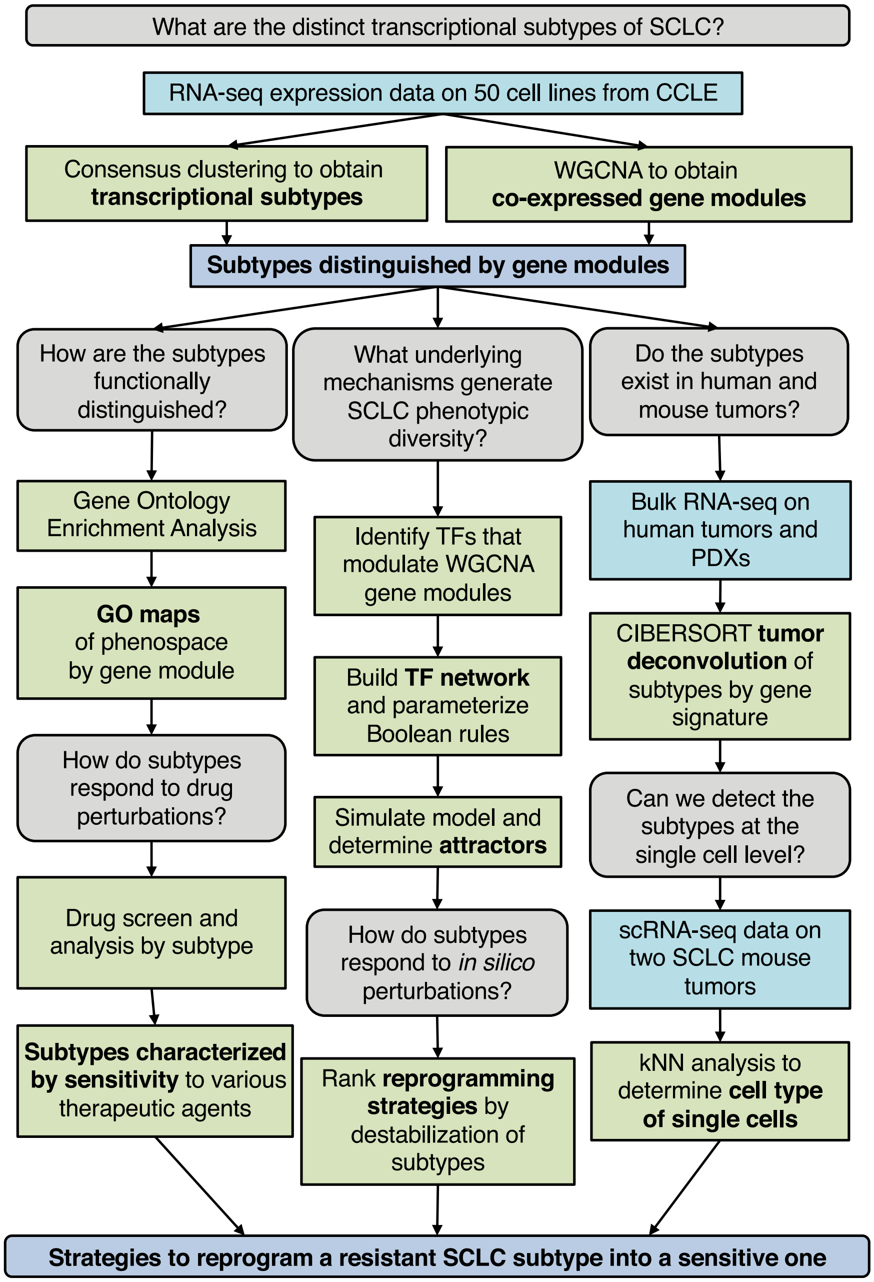

What are the distinct transcriptional subtypes of SCLC?
RNA-seq expression data on 50 cell lines from CCLE
Consensus clustering to obtain **transcriptional subtypes**
WGCNA to obtain **co-expressed gene modules**
Subtypes distinguished by gene modules
How are the subtypes functionally distinguished?
What underlying mechanisms generate SCLC phenotypic diversity?
Do the subtypes exist in human and mouse tumors?
Gene Ontology Enrichment Analysis
GO maps of phenospace by gene module
How do subtypes respond to drug perturbations?
Drug screen and analysis by subtype
Subtypes characterized by sensitivity to various therapeutic agents
Identify TFs that modulate WGCNA gene modules
Build **TF network** and parameterize Boolean rules
Simulate model and determine **attractors**
How do subtypes respond to *in silico* perturbations?
Rank **reprogramming strategies** by destabilization of subtypes
Bulk RNA-seq on human tumors and PDXs
CIBERSORT **tumor deconvolution** of subtypes by gene signature
Can we detect the subtypes at the single cell level?
scRNA-seq data on two SCLC mouse tumors
kNN analysis to determine **cell type of single cells**
Strategies to reprogram a resistant SCLC subtype into a sensitive one

Fig 1. Workflow of our analysis. We use parallel analyses to identify strategies to reprogram resistant SCLC subpopulations into sensitive ones. These strategies can then be tested *in vitro* and *in vivo*.

https://doi.org/10.1371/journal.pcbi.1007343.g001

strategies. As first described over 30 years ago, human SCLC cell lines can be categorized into two broad subtypes: a neuroendocrine (NE) stem-cell-like "classic" subtype and a distinct non-NE "variant" subtype [6–8]. In both human and mouse tumors, most cells appear to belong to the NE subtype, corresponding to a pulmonary neuroendocrine cell (PNEC) of origin [9], with high expression of neuroendocrine genes such as ASCL1. However, several groups have found evidence for non-NE variants within SCLC tumors [10–12], as well as an NE variant driven by MYC overexpression and NEUROD1 overexpression, instead of ASCL1 [13–15]. We previously described SCLC cell lines with hybrid expression of both NE and non-NE markers [16], and proposed they could serve as a resistant niche since drug perturbations shifted most cell lines towards hybrid phenotype(s). Taken together, these observations indicate the existence of a complex landscape of SCLC phenotypes that may form a tumor microenvironment robust to perturbations and treatment [10, 17]. However, previous SCLC subtype reports were limited in their ability to systematically identify subtypes and understand plasticity across them. We hypothesized that our workflow, by taking into account the dynamics of underlying GRNs, could make systems-level predictions that more accurately reflect the occurrence and transdifferentiation of coexisting subtypes within SCLC tumors.

Starting from transcriptomics data from SCLC cell lines, our pipeline identifies four transcriptional subtypes, and a GRN that describes their dynamics. Three of these correspond to known ones, the fourth is a previously unreported NE variant (termed NEv2) with reduced sensitivity to drugs. Both CIBERSORT and single-cell validation reveal that in virtually every human and mouse tumor heterogeneity encompasses NEv2, and that all other previously reported subtypes are represented across tumors. BooleaBayes identifies both master regulators and master destabilizers for each subtype, opening the way for treatment strategies that may take SCLC subtypes into account. For instance, we hypothesize that by targeting these master TFs, the NEv2 phenotype may be destabilized leading to increased treatment sensitivity of SCLC tumors.

Materials and methods

Data

Human SCLC cell line data was taken from the Broad Institute's CCLE RNA-seq expression data (version from February 14, 2018) at https://portals.broadinstitute.org/ccle/data. 81 human tumors were obtained from George et al. dataset, courtesy of R.K. Thomas [5]. The Myc-high mouse data set [15] was obtained from the NCBI GEO deposited at GSE89660. PDX/CDX mouse data [18] was obtained from the NCBI GEO deposited at GSE110853. Data from the CCLE was subsetted to only include SCLC cell lines (50). Features with consistently low read counts (< 10 in all samples) and non-protein-coding genes were removed. All expression data was then converted to TPM units and log1p normalized by dataset.

Clustering and WGCNA

We applied Consensus Clustering to RNA-seq gene expression data from the 50 SCLC cell lines in the Cancer Cell Line Encyclopedia (CCLE) using the ConsensusClusterPlus R package [19]. Gene expression (TPM) was median-centered prior to clustering, and we clustered the cell lines using a k-means method with a Pearson distance metric for $k \in \{2, 12\}$. Other parameters were set as follows: reps = 1000, pItem = 0.8, pFeature = 0.8, seed = 1. Best k value was chosen heuristically based on the cumulative distributive function plot, tracking plot, delta area plot, and consensus scores.

To identify gene programs driving the distinction between the four SCLC phenotypic clusters, we performed weighted gene co-expression network analysis (WGCNA) on the same RNA-seq data. The softPower threshold was chosen as 12 to generate a signed adjacency matrix from gene expression. A topological overlap matrix (TOM) was created using this adjacency matrix as input. Hierarchical clustering on 1-TOM using method = 'average,' and the function cutTreeDynamic was used to find modules with parameters: deepSplit = 2, pamRespectsDendro = TRUE, minClusterSize = 100. These settings were chosen based on an analysis of module stability and robustness. We then computed an ANOVA comparing the four subtypes for each module. 11 out of 18 modules were able to statistically distinguish between the four clusters with an FDR-adjusted p-value < 0.05.

Gene ontology enrichment analysis

We ran a gene ontology (GO) enrichment analysis on each module that was significantly able to distinguish the phenotypes (11 total). The terms that were significantly enriched in at least one module were culminated into a general list of terms enriched in SCLC, which had 1763 terms. To visualize these terms, we computed a distance matrix between pairs of GO terms using GoSemSim [20], and used this matrix to project the terms into a low dimensional space using t-SNE. t-SNE is a popular method that computes a low-dimensional embedding of data points and seeks to preserve the high-dimensional distance between points in the low-dimensional space.

Drug sensitivity analysis

Our drug sensitivity analysis used the freely available drug screen data from Polley, et al [21]. This screen included 103 Food and Drug Administration-approved oncology agents and 423 investigational agents on 63 human SCLC cell lines and 3 NSCLC lines. We subsetted the data to the 50 CCLE cell lines used for our previous analyses that had defined phenotypes according to Consensus Clustering (above). As described in [21], "the compounds were screened in triplicate at nine concentrations with a 96-hour exposure time using an ATP Lite endpoint." Curve fitting, statistical analysis, and plotting was done by Thunor Web, a web application for managing, visualizing and analyzing high throughput screen (HTS) data developed by our lab at Vanderbilt University [22]. To fit a dose response curve for each drug and cell line pair, we fit percent viability data from the screen to a three parameter log-logistic model. The three parameters are E_{max}, EC_{50}, and the Hill coefficient, where each coefficient is constrained to reasonable ranges (E_{max} is constrained to be between 0 and 1, and the Hill coefficient (slope) is constrained to be non-negative.) Activity area (AA) was calculated as described in [23]. Briefly, AA is the area (on a log-transformed x-axis) between $y = 1$ (no response) and linear extrapolations connecting the average measured response at each concentration. A larger activity area indicates greater drug sensitivity, characterized either by greater potency or greater efficacy, or both. By segregating the cell lines by subtype, we were able to evaluate the relationship between drug response and subtype.

CIBERSORT

CIBERSORT is a computational inference tool developed by Newman et al. at Stanford University [4]. We utilized the interactive user interface of CIBERSORT Jar Version 1.06 at https://cibersort.stanford.edu/runcibersort.php. Gene signatures were automatically determined by the software from a provided sample file with a matching phenotype class file. For this sample file and class file, the RNA-seq data from 50 human SCLC cell lines were inputted with their consensus clustering class labels. For each run, 500 permutations were performed. Relative and absolute modes were run together, with quantile normalization disabled for RNA-seq data, kappa = 999, q-value cut-off = 0.3, and 50-150 barcode genes considered when building the signature matrix.

Single cell RNA sequencing of TKO SCLC tumors

The Tp53, Rb1 and p130 triple-knockout (TKO) SCLC mouse model with the Rosa26membrane-Tomato/membrane-GFP (Rosa26mT/mG) reporter allele has been described (Denny and Yang et al., 2016). Tumors were induced in 8-weeks old TKO; Rosa26mT/mG mice by intratracheal administration of 4x107 PFU of Adeno-CMV-Cre (Baylor College of Medicine, Houston, TX). 7 months after tumor induction, single tumors (one tumor each from two mice) were dissected from the lungs and digested to obtain single cells for FACS as previously described [10, 24]. DAPI-negative live cells were sorted using a 100 μm nozzle on a BD FACSAria II, spundown and resuspended in PBS with 10% bovine growth serum (Fisher Scientific) at a concentration of 1000 cells/μl. Single cell capture and library generation was performed using the Chromium Single Cell Controller (10x Genomics) and sequencing was performed using the NextSeq High-output kit (Illumina).

Single cell analysis

Cells with $\leq$ 500 detected genes per cell or with $\leq$ 10% of transcripts corresponding to mitochondria-encoded genes were removed. Low abundance genes that were detected in less than 10 cells were excluded. Each cell was normalized to a total of 10,000 UMI counts, and log2-transformed after the addition of 1. Top 1000 highly variable genes were selected and clusters of cells were identified by the shared nearest neighbor modularity optimization based on the top 10 PCs using the highly variable genes and visualized by t-SNE in R package Seurat [25]. The k-nearest neighbors (kNN) with k = 10 of human cell lines was detected for each mouse cell to predict subtypes of the individual cell based on signature genes of each subtype. If at least 80% nearest human cell line neighbors for a mouse cell belong to one subtype, the mouse cell was assigned to that subtype. Otherwise, the subtype was undetermined (not assigned).

Genomic analysis

Mutational Analysis was performed by MutSigCV V1.2 from the Broad Institute [26]. First, a dataset of merged mutation calls (including coding region, germ-line filtered) from the Broad Cancer Dependency Map [27] was subsetted to only include SCLC cell lines. Background mutation rates were estimated for each gene-category combination based on the observed silent mutations for the gene and non-coding mutations in the surrounding regions. Using a model based on these background mutation rates, significance levels of mutation were determined by comparing the observed mutations in a gene to the expected counts based on the model. MutSigCV was run on the GenePattern server using this mutation table, the territory file for the reference human exome provided for the coverage table file, the default covariate table file (gene.covariates.txt), and the sample dictionary (mutation_type_dictionary_file.txt). Only genes with an FDR-corrected q-value $<$ 0.25 were considered significant.

Gene regulatory network construction

Transcription factors from significantly differentiating gene modules were used as input to network structure construction. A list of connections between these TFs was curated from the literature and added as edges between the TF nodes. The ChEA database of ChIP-seq-derived interactions [28] was queried to add additional connections between TFs that may not have been found in the literature. Our edge list thus comprises the literature-based connections that are verified from ChEA, and additional connections from the ChEA database directly. The network was built using NetworkX software [29].

BooleaBayes inference of logical relationships in the TF network

A Boolean function of N input variables is a function F: $\{0, 1\}^N \mapsto \{0, 1\}$. The domain of F is a finite set with 2^N elements, and therefore F is completely specified by a 2^N dimensional vector in the space $\{0,1\}^{2^N}$ in which each component of the vector corresponds to the output of F for one possible input. In general, knowledge of the steady states of F is unlikely to be sufficient to fully constrain all 2^N components of the vector describing F. BooleaBayes is a practical approach that constrains F in the neighborhood of stable fixed points based on steady-state gene expression data. In practice, we let each component of the vector be a continuous real-value $v_i \in [0, 1]$ reflecting our confidence in the output of F, based on available constraints. Components of F that are near 0.5 will indicate uncertainty about whether the output should be 0 or 1, given the available constraining data.

Given M observations (in our case, each observation is a measurement of gene expression of the N regulator TFs and the target TF in $M = 50$ cell lines), we want to compute this vector ($\vec{V}$) describing a probabilistic Boolean function F of N variables. First, we organize the input-output relationship as a binary tree with N layers leading to the 2^N leaves, each of which corresponds to a component of vector $\vec{V}$. For instance, given two regulators A and B ($N = 2$), the leaves of the binary tree correspond to the probabilities that $(\bar{A} \wedge \bar{B})$, $(\bar{A} \wedge B)$, $(A \wedge \bar{B})$, and $(A \wedge B)$. Collectively, the observations define an $M \times N$ matrix $\mathbf{R} = [\vec{R}_1, \vec{R}_2, \ldots, \vec{R}_N]$ quantifying the input regulator variables (columns) for each observation (rows), as well as an M dimensional vector $\vec{T} = [t_1, t_2, \ldots, t_M]$ quantifying the output variable. A Gaussian mixed model is then used to transform the columns of $\mathbf{R}$ (regulator variables) and the vector $\vec{T}$ into probabilities $\mathbf{R}'$ and $\vec{T}\prime$ of the variables being OFF or ON in each observation (row).

Let $P_j = (\vec{R}'_i)$ be a function that quantifies the probability that the input variables of the i^{th} observation belong to the j^{th} leaf of the binary tree. For instance using the example above, the second leaf of the binary tree is $(\bar{A} \wedge B)$. Therefore, $P_{j=2}(A, B) = (1 - A) \cdot B$. Note that by this definition, $\sum_{j=1}^{2^N} P_j(\vec{R}'_i) = 1$. Using this, we define an $M \times 2^N$ weight matrix $\mathbf{W} = w_{i,j}$ as

$$w_{i,j} = P_j(\vec{R}'_i)$$

that describes how much the i^{th} observation constrains the j^{th} component of $\vec{V}$. Additionally, to avoid overfitting under-determined leaves, we define the uncertainty $\vec{U} = [u_1, u_2, \ldots, u_{2^N}]$ of each leaf

$$u_j = 1 - \max_{i \in \{1,\ldots,M\}} (w_{i,j})$$

From these, we then define the vector $\vec{V}$ describing function F as

$$v_j = \frac{\sum_{i=1}^{M} t'_i \cdot w_{i,j} + 0.5 \cdot u_j}{\sum_{i=1}^{M} w_{i,j} + u_j}$$

Thus, each component of $\vec{V}$ is the average of the output target variable $\vec{T}$ weighted by $\mathbf{W}$, with an additional uncertainty term $\vec{U}$ to avoid overfitting. For leaves j of the binary tree that are poorly constrained by any of the observables, $v_j \approx 0.5$, indicating maximal uncertainty in the output of F at those leaves. Uncertainty of a leaf j also arises when observations i with large weight $w_{i,j}$ have inconsistent values for t'_i, such as if $t'_1 = 0$ and $t'_2 = 1$.

BooleaBayes network simulations

As input to the BooleaBayes simulations, we know the network structure defining regulatory relationships as described above, and regulatory rules (from BooleaBayes algorithm for rule fitting, see above). We first pick a random initial state by choosing a vector $V = [v_1, v_2, \ldots, v_g]$, where g is the number of genes in the network. We initialize each v_i in this vector to be 0

(OFF) or 1 (ON). For *in silico* perturbation experiments, this initial state is be chosen as one of the pseudo-attractors corresponding to a specific subtype. We then randomly pick one transcription factor x (where each gene has probability $\frac{1}{g}$ of getting picked) to update.

Using the rule for x given by the rule fitting method above, find the column that corresponds to the current state (V) of the parent genes ($pa(x)$) of x (in other words, find the column corresponding to ($V[pa(x)]$). This column is defined by the state of the parent nodes of x, and it has some probability associated with it for how likely it is to turn on x when in the state $V[pa(x)]$. In Results section *Transcription factor network defines SCLC phenotypic heterogeneity and reveals master regulators*, this probability is visualized as a color (blue to red) at the bottom of the figures. We then flip a weighted coin with this probability, and turn x ON or turn x OFF based on the outcome. This will result in moving to a state 1 step away (if we do indeed flip the expression of x from 0 to 1 or 1 to 0), or in staying in the same state (if we "flip" from 0 to 0 or 1 to 1). The state has now moved to a new state in the state transition graph. If all transition probabilities to neighboring states are less than 0.5, this state is considered a pseudo-attractor. For the *in silico* perturbation experiments, the number of steps in the shortest path from the current state to the starting state is recorded instead.

See Algorithm 1 for pseudo-code describing the pseudo-attractor finding algorithm, and Algorithm 2 for pseudo-code describing the random-walk stability scores.

Ethics statement

Mice were maintained according to practices prescribed by the NIH (Bethesda, MD) at Stanford's Research Animal Facility, accredited by the Association for the Assessment and Accreditation of Laboratory Animal Care (AAALAC). All animal studies were conducted following approval from the Stanford Animal Care and Use Committee (protocol 13565).

Results

Consensus clustering uncovers new SCLC variant phenotype

Recently, the occurrence of variant SCLC subtypes has been reported [13, 15, 16]. Given the translational value of defining subtypes, a more global approach to comprehensively define SCLC subytpes would be desirable. To this end, we devised the workflow described in Fig 1. First, we applied Consensus Clustering [30] to RNA-seq gene expression data from the 50 SCLC cell lines in the CCLE [31]. Here, the underlying assumption of bulk RNA-seq data is that single-cells from each cell line belong to one cellular state. While this is consistent with our previous findings that SCLC cell lines resolve into discrete clusters by flow cytometry [16], future cell-line analysis at single-cell resolution may refine our results, and it will be interesting to see to what extent subtype heterogeneity may be reflected within one cell line. We clustered the cell lines using a k-means method with a Pearson distance metric for k ∈ {2, 20} (Fig 2A, S1 Table). Consensus Clustering is a method in which multiple k-means clustering partitions have been obtained for each k. Consensus Clustering is then used to determine the consensus (or best) clustering across these multiple runs of the k-means algorithm, in order to determine the number and stability of clusters in the data. Using criterion such as the tracking plot and delta area plot (S1 Fig), both k = 2 and k = 4 gave well defined clusters. Since recent literature suggests that more than two subtypes are necessary to adequately describe SCLC phenotypic heterogeneity, we selected k = 4 for further analyses.

To align the 4-cluster classification (Fig 2B) with existing literature, we considered well-studied biomarkers of SCLC heterogeneity across the clusters. Three of the four consensus

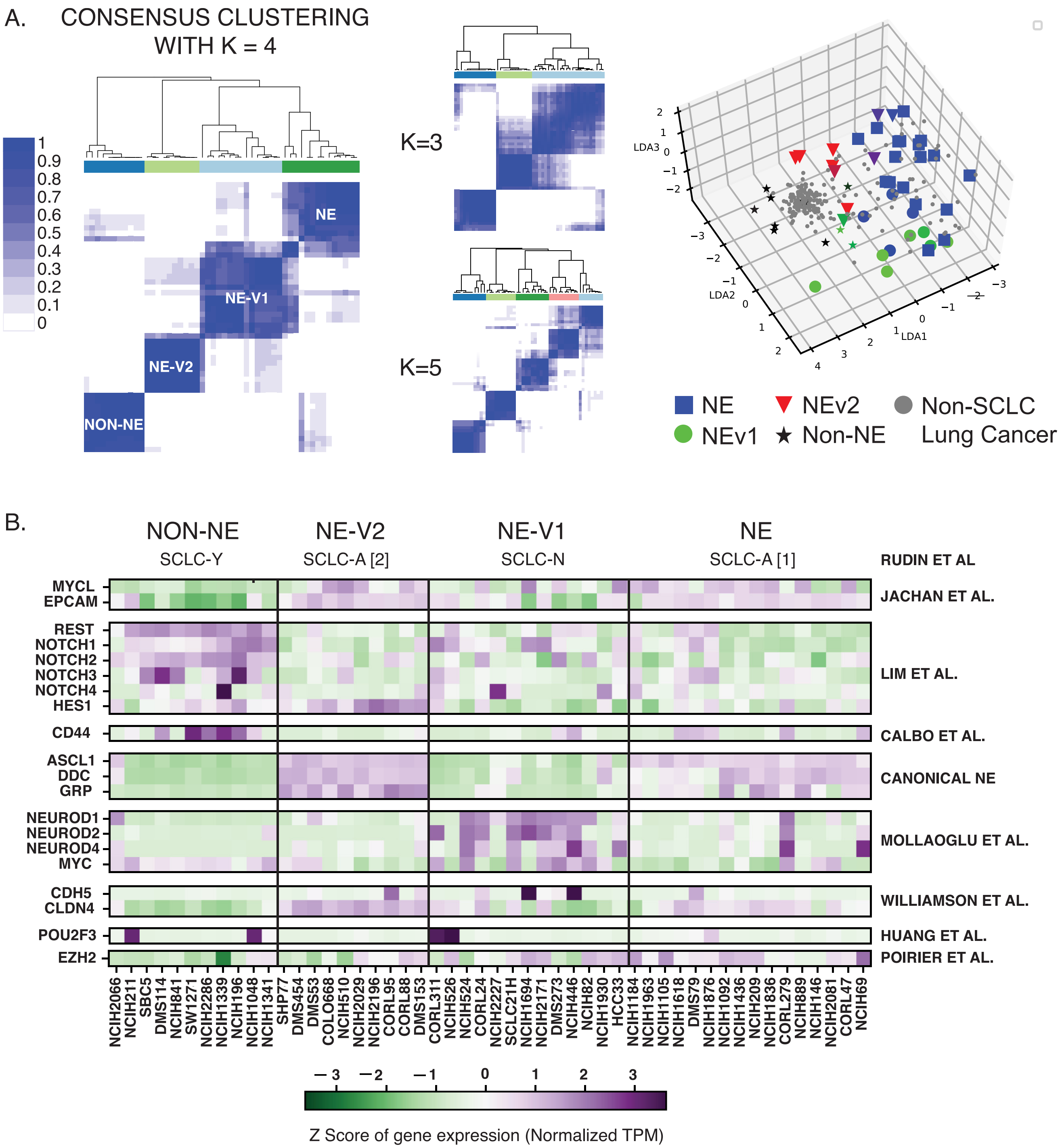


Fig 2. Consensus clustering and WGCNA of 50 SCLC cell lines reveals four subtypes differentiated by gene modules. A. Consensus clustering with k = 4 gives most consistent clusters. K = 3 and K = 5 add complexity without a corresponding increase in accuracy. LDA plot shows separation of 4 clusters, with non-SCLC cell lines falling near non-NE cell lines. **B**. Current biomarkers in the field of SCLC are able to distinguish between three of the subtypes; The fourth subtype, NEv2, is not separable from NE using markers from SCLC literature.

https://doi.org/10.1371/journal.pcbi.1007343.g002

clusters could be readily matched to subtypes previously identified with 2 to 5 biomarkers: the canonical NE subtype (SCLC-A [14, 32]), an NE variant subtype (referred to here as NEv1, corresponding to SCLC-N in [15, 32]), and a non-NE variant subtype (SCLC-Y) [10, 32, 33]. The fourth cluster (referred to here as NEv2) could not be easily resolved using few markers. For example, NEv2 may be considered a tumor propagating cell (TPC, which encompasses the NE, or SCLC-A, subtype) by biomarkers in Jachan et al [24], yet expression of a single biomarker, *HES1*, would suggest this subtype falls outside of the NE subtype according to Lim et al [10]. Discrepancies like this drove us to consider broader patterns of gene expression, rather than a limited number of biomarkers, to characterize each subtype.

SCLC phenotypes are differentially enriched in diverse biological processes, including drug catabolism and immuno-modulation

To capture global gene expression patterns, we applied Weighted Gene Co-expression Network Analysis (WGCNA) [34] to RNA-seq data from CCLE for multiple SCLC cell lines (See Methods). This analysis revealed 17 groups, or modules, of co-expressed genes. Module eigengenes could be used to describe trends of gene expression levels. 11 of these 17 groups of co-expressed genes could statistically distinguish between the four consensus clusters (Fig 3A, S2 Table, Kruskal-Wallis, FDR-adjusted $p < 0.05$). To specify the biological processes enriched in each of these 11 gene modules, we performed gene ontology (GO) enrichment analysis using the Consensus Path Database [35], which resulted in a combined total of 1,763 statistically enriched biological processes (Fig 3B, S3 Table). In particular, the turquoise, yellow, salmon, and pink modules are enriched for neuroendocrine differentiation and neurotransmitter secretion and are upregulated in the canonical NE and NEv1 phenotypes, as quantified by Gene Set Enrichment Analysis [36] (Fig 3C, S2 and S3 Figs, S4 Table). PNECs, the presumed cell of origin for SCLC, group into neuroendocrine bodies (NEBs) that are innervated by sensory nerve fibers and secrete neuropeptides that affect responses in the autonomic and/or central nervous system. This is consistent with the NE- and NEv1-enriched GO terms “learning or memory” and “chemical synaptic transmission” (Fig 3C). Evidently, such functions may be maintained in NE and NE-v1 subtypes, as reflected by the frequent occurrence of paraneoplastic syndromes in SCLC patients [37]. In contrast, the blue, black, and purple modules, enriched for cell adhesion and migration processes, are upregulated in the non-NE variant phenotype, in agreement with the observed adherent culture characteristics of these cell lines (S4 Fig, S1 File).

Genes within the brown, midnight blue, and green modules are upregulated in the NEv2 phenotype (Fig 3A, S3 Fig). The brown module is enriched for canonical phenotypic features of SCLC, particularly cellular secretion and epithelial differentiation, and accordingly is also upregulated in the canonical NE subtype. The midnight blue module, enriched in nervous system processes and lipid metabolism, is highly expressed in the NEv2 cell lines. The green module is enriched for immune/inflammatory response, wound healing, homeostasis, drug/xenobiotic metabolism, and cellular response to environmental signals (Fig 3C). Enrichment of these GO terms suggest that NEv2 cells may more easily adapt to external perturbations such as therapeutic agents, and potentially show higher drug resistance.

To visualize these enriched GO terms in an organized way (Fig 3B), we used the GOSemSim package [20] in R to compute a pairwise dissimilarity score, or distance, between all enriched GO terms (FDR-adjusted $p < 0.05$ in at least one of the 11 significant modules). We then projected all significant GO terms into a 2D space by t-distributed stochastic neighbor embedding (t-SNE) [38]. In this t-SNE projected phenospace, GO terms that describe semantically similar biological processes are placed close to one another and grouped into a general

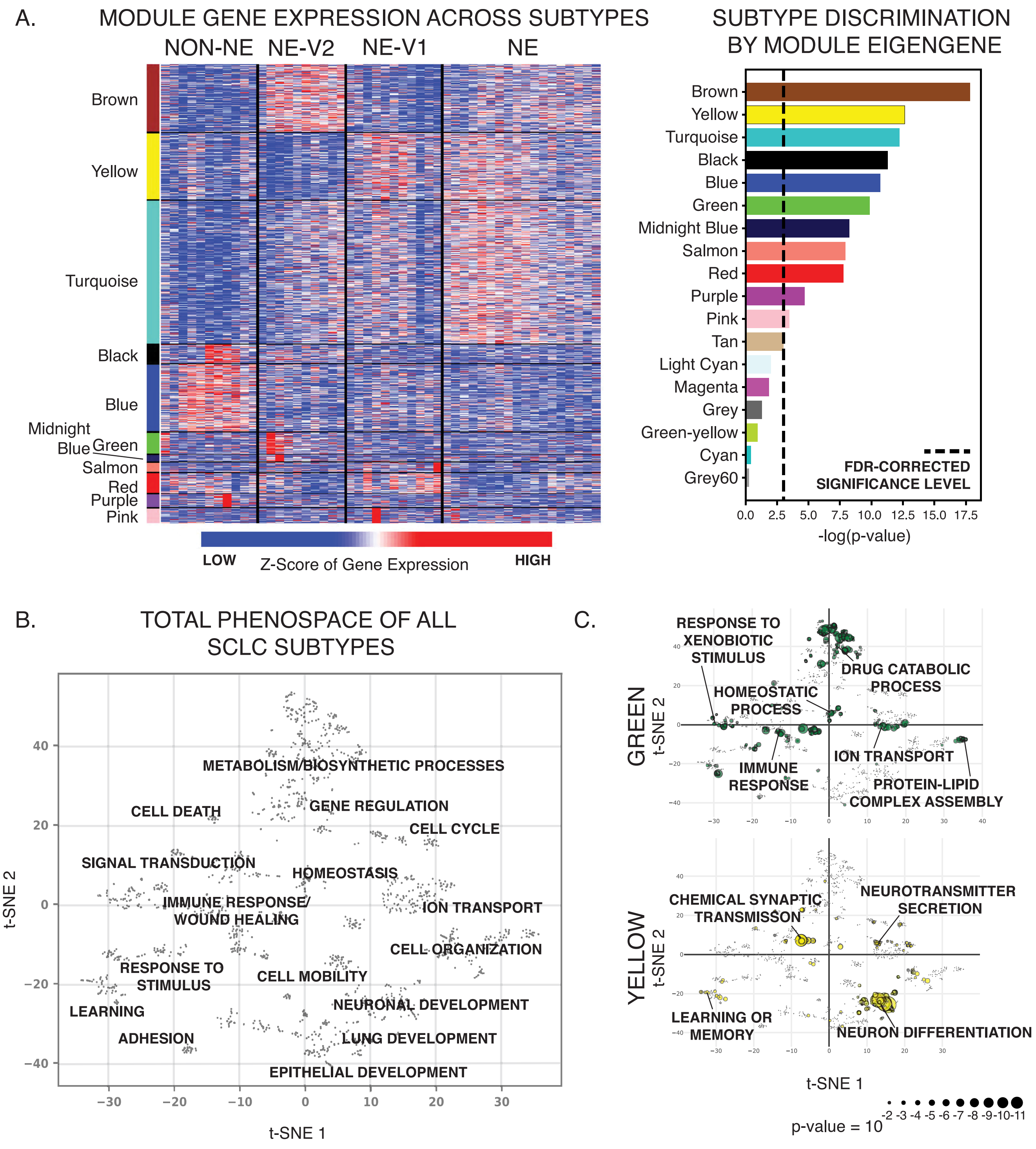


Fig 3. SCLC subtypes can be distinguished by gene expression patterns. A. Transcriptional patterns that distinguish the four subtypes are captured in WGCNA analysis. Gene modules by color show patterns of expression that are consistent across the subtypes. Only modules that significantly distinguish between the subtypes are shown (ANOVA, FDR-corrected p-value < 0.05). **B**. SCLC heterogeneity biological process phenospace. A dissimilarity score between pairs of SCLC-enriched GO terms was calculated using GoSemSim, and used to create a t-SNE projection grouping similar biological processes together. Several distinct clusters of related processes can be seen. **C**. Module-specific phenospace. A breakdown of where some of the 11 statistically significant WGCNA modules fall in the GO space from A. Of particular interest, the green module, which is highly upregulated in the NEv2 phenotype, is enriched in metabolic ontologies, including drug catabolism and metabolism and xenobotic metabolism. The yellow module is enriched in canonical neuronal features.

https://doi.org/10.1371/journal.pcbi.1007343.g003

biological process. This map allows exploration of biological processes enriched in individual gene modules or subtypes, and it shows that SCLC heterogeneity spans biological processes that can largely be grouped as 1) related to neuronal, endocrine, or epithelial differentiation; 2) metabolism and catabolism; 3) cell-cell adhesion and mobility; and 4) response to environmental stimuli, including immune and inflammatory responses. In summary, the phenospace constructed from global gene expression patterns captures the unique characteristics of each SCLC subtype.

Drug resistance is a feature of the NEv2 subtype

As mentioned previously, the enriched GO terms for drug catabolism and xenobiotic metabolism in the green module suggest that the NEv2 phenotype may have a higher ability to metabolize drugs and therefore exhibit decreased sensitivity. To test this possibility, we reanalyzed drug responses of SCLC cell lines to a panel of 103 FDA-approved oncology agents and 423 investigational agents in the context of our four subtype classification [21]. We used the Activity Area (AA) metric as a measure of the resultant dose-response curves. The drugs were analyzed individually and clustered by common mechanism of action and target type, and the cell lines were grouped by the four subtypes (S5 Table). Across all evaluated drugs, the NEv2 subtype exhibited the most resistance (54% of drugs showed NEv2 as most resistant). In contrast, both NE and NEv1 exhibited less resistance (20%), with non-NE exhibiting the least resistance (6%) (Fig 4A). Taken together, these results confirm that based on the prediction from the gene-regulation based classification, the subtypes exhibit different levels of resistance and that high resistance is a feature of the NEv2 subtype (Fig 3C), even though the subtypes do not show differential response to the standard of care (etoposide and platinum based agents, Fig 4B). In particular, mTOR inhibitors are a class of compounds to which NEv2 was significantly more resistant (Fig 4C). PI3K pathway mutations have previously been implicated as oncogenic targets for SCLC, as about a third of patients show genetic alterations in this pathway [39]. Among the four subtypes, NEv2 is also the least sensitive to AURKA, B, and C inhibitors (AURKA shown); TOPO2 inhibitors; and HSP90 inhibitors (Fig 4C–4F). These results have implications for interpreting expected or observed treatment response with respect to tumor heterogeneity in individual patients.

Neuroendocrine variants are represented in mouse and human SCLC tumors

Next, we investigated whether the four subtypes we detected in human SCLC cell lines are also present in tumors. We used CIBERSORT [4] to generate gene signatures for each of the 4 subtypes. These gene signatures could then deconvolve RNA-seq measurements on 81 SCLC tumors from George et al [5] to specify relative prevalence of each subtype within a single tumor. Consistent with studies of intra-tumoral heterogeneity in other types of cancer, such as breast cancer [40], CIBERSORT predicted that a majority of tumors were comprised of all four subtype signatures, in varying proportions across tumor samples (Fig 5A). We then analyzed the patient/cell-derived xenograft models (PDXs/CDXs) developed by Drapkin et al [18], and the tumors also showed vast differences across samples (Fig 5B). Some of these samples were taken across multiple time points from the same patient, thus enabling us to test both tumor composition and dynamic changes in tumor subpopulations. Three samples taken from patient MGH1514, before and after treatment, indicated a change in tumor composition in favor of the NE phenotype. In contrast, patient MGH1518 showed a reduction of NEv1 and increase in NEv2 after treatment. Similar observations of phenotypic changes over treatment time courses, made in breast cancer patients [40] have recently been explained in the context

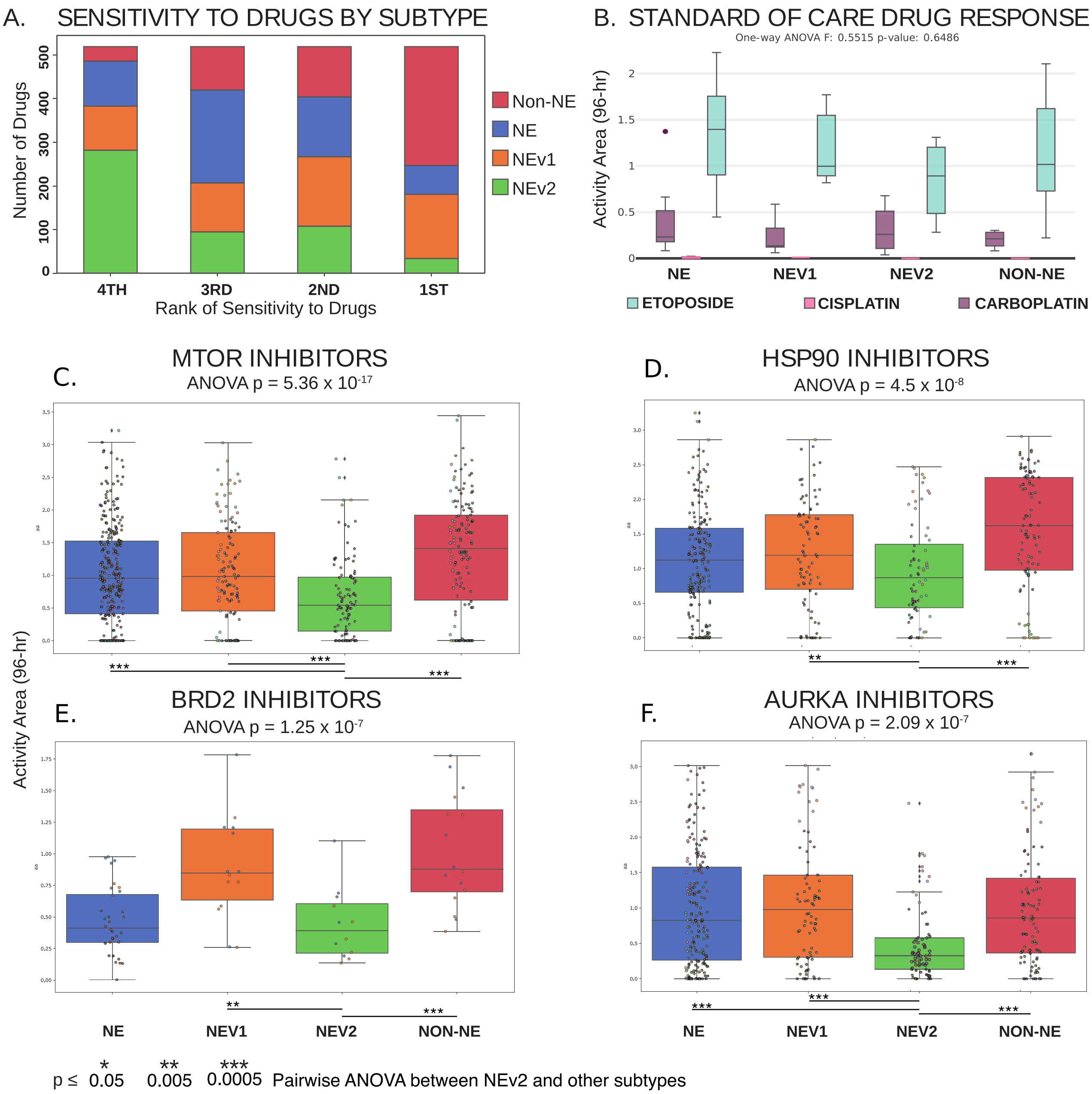


Fig 4. Differential response of SCLC subtypes to a wide variety of oncology drugs and investigational agents. A. Ranked sensitivity of subtypes across 526 compounds. NEv2 is least sensitive for over half of the drugs tested. **B**. No significant differences can be seen in response to etoposide and platinum-based agents cisplatin and carboplatin, the standard of care for SCLC. **C-F**. Significantly differential response by ANOVA, $p < 0.05$, shown in drugs that target **C**. mTOR, **D**. HSP90, **E**. BRD2, and **F**. AURKA. NEv2 is significantly more resistant to all of these drugs.

https://doi.org/10.1371/journal.pcbi.1007343.g004

of a mathematical model of epithelial to mesenchymal transition (EMT) [41]. It is possible that the tumor composition changes we observe may also be explained by molecular level and/or cell population level models [42]. Overall, the high variance in proportions of each subtype suggest a high degree of intertumoral, as well as intratumoral, dynamic heterogeneity and plasticity.

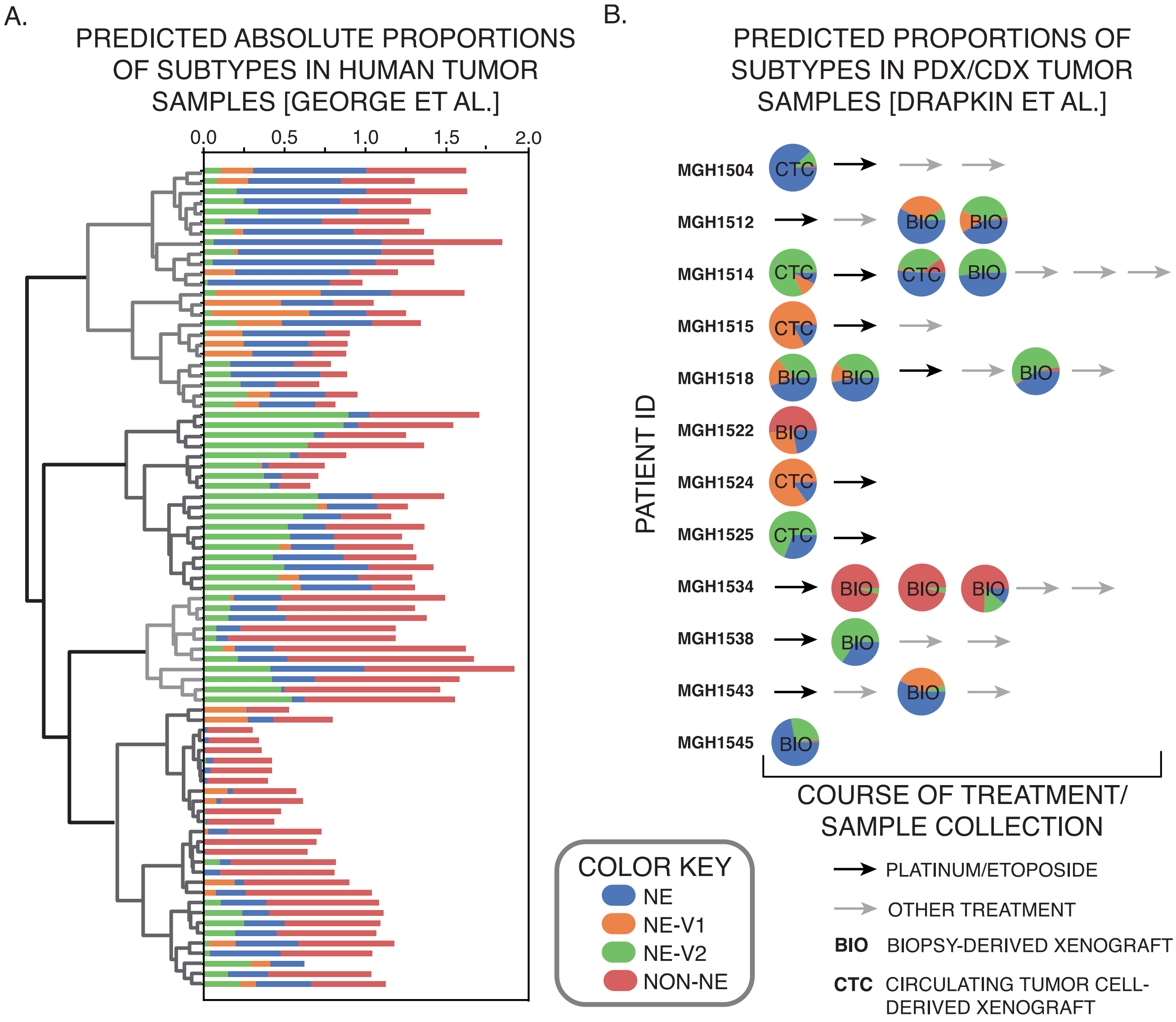


Fig 5. Computational evidence for existence of subtypes in human tumors. A. Absolute proportion of each subtype in 81 human tumors as determined by CIBERSORT. The 81 tumors can then be sorted by hierarchical clustering, which finds four main groups of subtype patterns across tumors. **B.** Similar analysis in mouse PDX/CDX tumors from Drapkin et al. [18].

https://doi.org/10.1371/journal.pcbi.1007343.g005

We also investigated phenotypic patterns in mouse tumors from two different sources to determine whether human SCLC subtype signatures are conserved across species [15] (Methods). The first mouse model is a triple knockout (*Rb1*, *Tp53*, and *P130*, conditionally deleted in lung cells via a Cre-Lox system, TKO), and these tumors were primarily composed of the NE and NEv2 subtypes (Fig 6Ai). Of note is the lower percentage of non-NE cells found in each tumor in Fig 6Ai; we suspect this is due to a filtering step before sequencing (Methods), as the non-NE subtype signature is more similar to tumor-associated immune cells in an unfiltered tumor population. The second mouse model was generated with *Myc* overexpression (double knockout of *Rb1* and *Tp53*, and overexpression of *Myc*) (Fig 6Aii) as reported previously [15]. Using the subtype gene-signatures developed in the previous sections, the *Myc*-high tumors showed a clear increase in the percentage of NEv1 detected compared to the triple

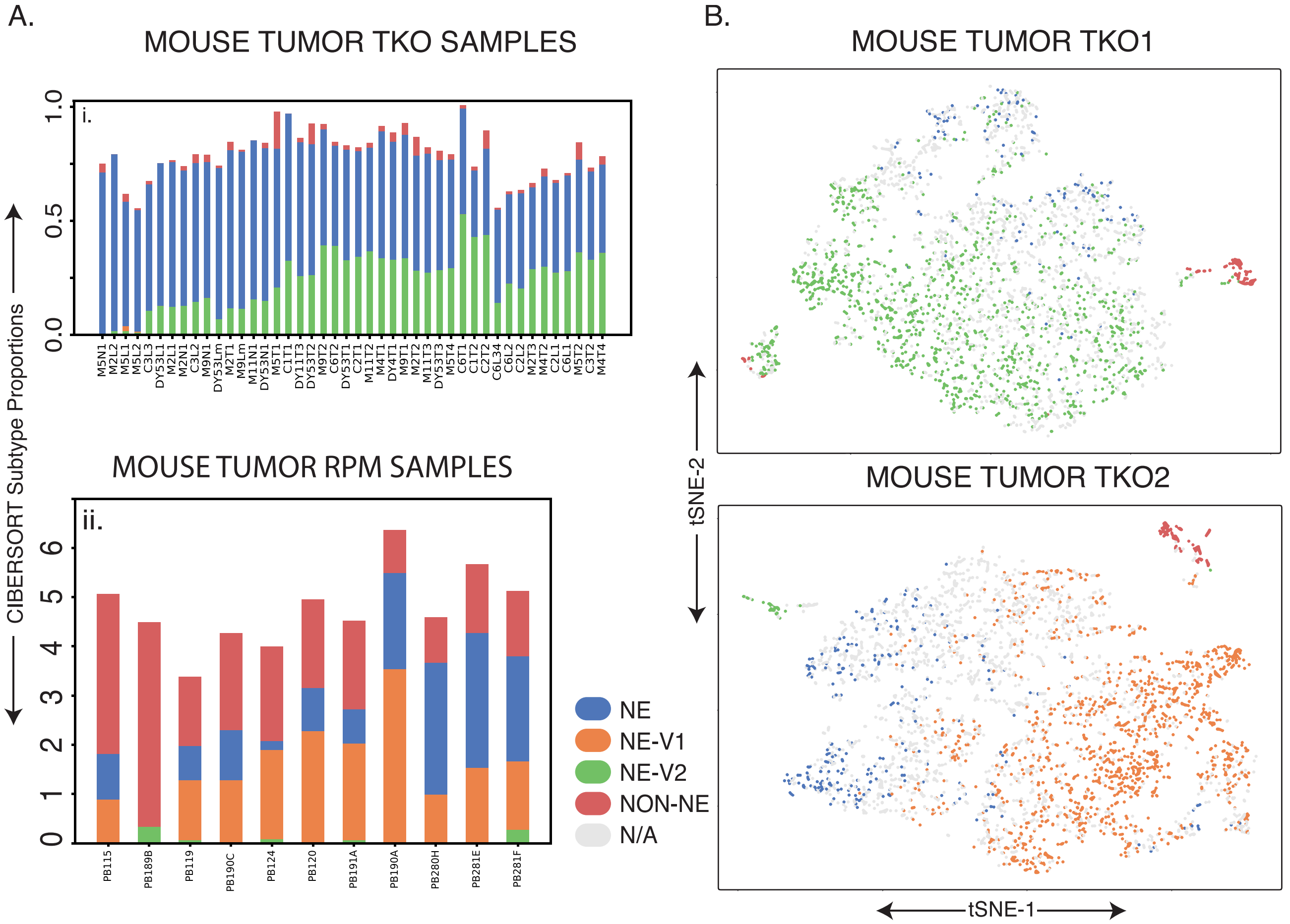


Fig 6. Similar analysis in mouse tumors. A. Ai. TKO (*Rb1, Tp53, P130* floxed) mouse tumors showing a high proportion of NE and NEv2 subtypes. Aii. As described in [15], these mouse tumors were generated by crossing Rb1 fl/fl Trp53 fl/fl (RP) animals to knockin Lox-Stop-Lox (LSL)-MycT58A-IRES-Luciferase mice. These Rb1 fl/fl Trp53 fl/fl Myc LSL/LSL (RPM) mice have overexpressed Myc and have been shown to be driven towards a variant phenotype, which is corroborated in this CIBERSORT analysis. It is clear that RPM mice contain greater portions of NEv1 compared to the tumors in Ai., which seems to correspond to the Aurora-Kinase-inhibitor-sensitive, *Myc*-high phenotype published by Mollaoglu et al. **B**. t-SNE plots of single cell RNA-seq from two TKO mouse tumors. The k-nearest neighbors (kNN) with k = 10 was computed for each mouse cell to predict subtypes of individual cell using signature genes of each subtype. If at least 8 of the 10 nearest human cell line neighbors for a mouse cell were of one subtype, the cell was assigned that subtype. Large amounts of intratumoral and intertumoral heterogeneity are evident.

https://doi.org/10.1371/journal.pcbi.1007343.g006

knockout tumors in Fig 6Ai, corroborating the correlation between NEv1 and a previously described *Myc* -high mouse tumor subtype.

Lastly, we analyzed two primary TKO mouse tumors by single cell RNA-seq (scRNA-seq). For each mouse single cell transcriptome, we computed the k = 10 nearest *human cell line* neighbors (kNN with k = 10), and assigned each mouse cell to a subtype based on its neighbors (Methods). As shown in Fig 6B, a large portion of the cells from each tumor correspond to one of the four human subtypes. A small non-NE population can be seen in both tumors, and about a third of the assigned cells correspond to the NE subtype (Fig 6B). Tumor 1 has a large proportion of the NEv2 subtype, corresponding to the tumors in Fig 6Ai. In contrast, tumor 2 has a large NEv1 subpopulation, similar to the tumors in Fig 6Aii. Taken together, these results indicate that subtypes in SCLC tumors are conserved across species, and can be categorized either by CIBERSORT analysis of bulk transcriptomics data, or by kNN analysis of scRNA-seq data.

Genetic mutations alone cannot account for four SCLC phenotypes

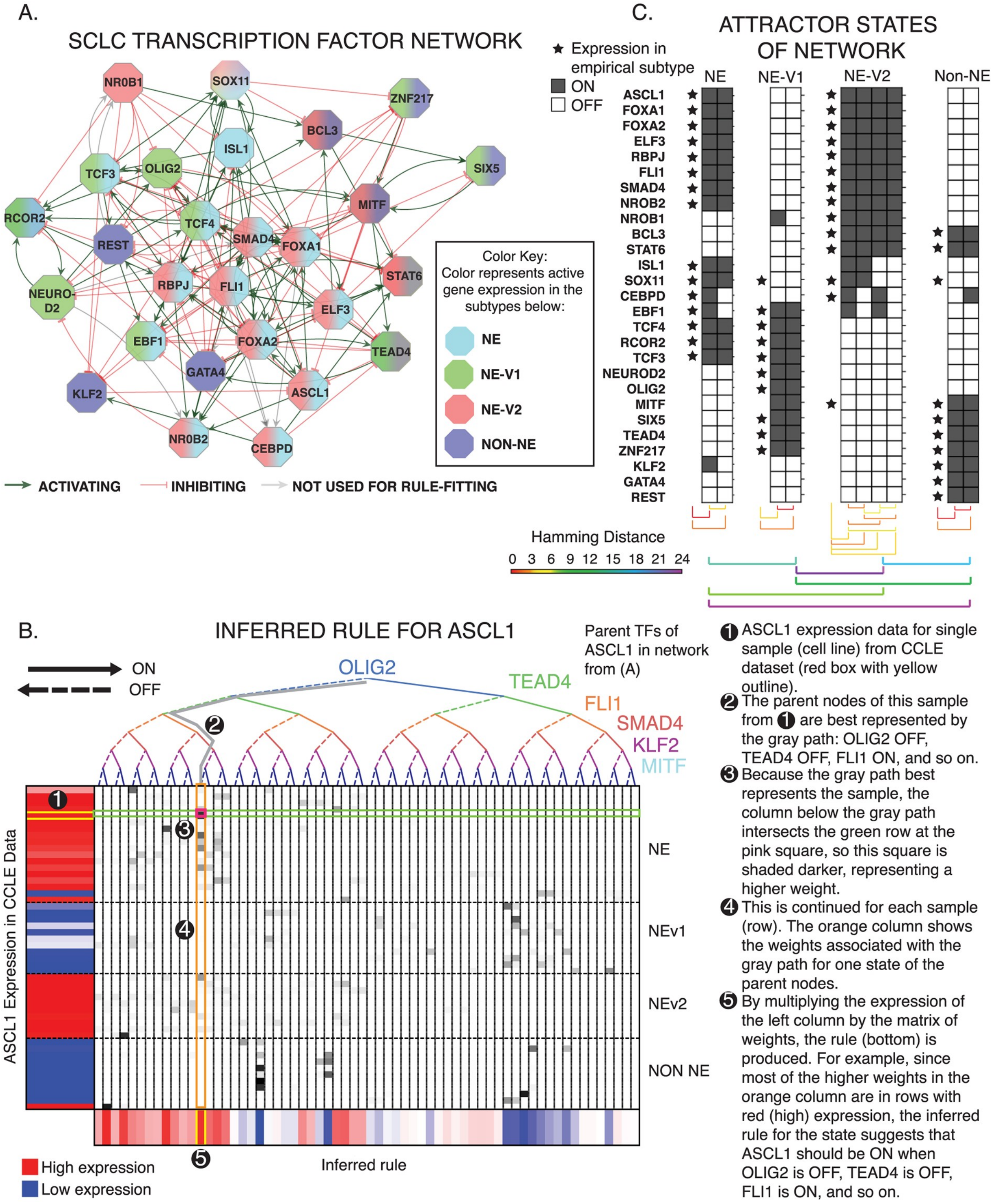
The evidence above for intratumoral and intertumoral heterogeneity led us to investigate how the subtypes arise and coexist in both human and mouse SCLC tumors. To determine whether mutations could be responsible for defining the four SCLC subtypes, we analyzed genomic data in the Broad Cancer Dependency Map [27]. We subsetted these data to the 50 SCLC cell lines with matching CCLE RNA-seq data, and using MutSigCV [26], we found 29 genes (S5 Fig) mutated more often than expected by chance (using a significance cutoff of q-value ≤ 0.5 to be as inclusive as possible). However, none of these genes were able to separate the four subtypes by mutational status alone (S5 Fig), suggesting alternative sources of heterogeneity.

Transcription factor network defines SCLC phenotypic heterogeneity and reveals master regulators

To investigate these alternative sources of heterogeneity, we hypothesized that different SCLC subtypes emerge from the dynamics of an underlying TF network. We previously identified a TF network that explained NE and non-NE SCLC subtype heterogeneity [16]. That analysis suggested the existence of additional SCLC subtypes but did not specify corresponding attractors [16]. Here, we performed an expanded TF network analysis to find stable attractors for all four SCLC subtypes. As an initial step, we identified putative master TF regulators within each of the 11 WGCNA modules (Fig 3B) based on differential expression. Regulatory interactions between these TFs were extracted from public databases, including ChEA, TRANSFAC, JASPAR, and ENCODE, based on evidence of TF-DNA binding sites in the promoter region of a target TF, as well as several sources from the literature. This updated network largely overlaps with, but contains several refinements compared to our previous report [16], as detailed in Fig 7A.

Following the procedure we previously used [16], we simulated the network as a dynamic Boolean model. In a Boolean model, the state of the network at a given time, t, is defined by the value of all TFs, each of which can be either ON or OFF. Each TF can be updated to determine its value at time $t + 1$ based on a Boolean rule, or logical statement, that represents how that TF is regulated by its regulators. For example, if $A_{t+1} = B_t$ or C_t, and if $A(t)$ = OFF, $B(t)$ = ON, and $C(t)$ = OFF, then updating A will give $A(t + 1)$ = ON or OFF = ON, so A turns ON. Boolean models are powerful tools to investigate the regulation of attractors corresponding to stable subtypes or oscillators of biological systems. Because precise update rules are often not known, one of two approximations are commonly applied: inhibitory dominant [43], or majority rules [43, 44]. Inhibitory dominant rules assert that the target node turns ON only when at least one activator is ON and all inhibitors are OFF, otherwise the target turns OFF (S6C Fig). Majority rules, conversely, assert that the target node turns ON as long as it has more activators ON than inhibitors, otherwise the target turns OFF (S6C Fig). Using the network in Fig 7A, neither of these approximations stabilized attractors corresponding to either the NEv1 or NEv2 phenotypes (S6 Fig), suggesting that the regulatory rules governing stability of these phenotypes are more complex.

To address this complexity, we developed BooleaBayes, a method to infer logical relationships in gene regulatory networks (Fig 7B) using gene expression data, by enhancing confidence in Boolean rules via a Bayes-like adjustment approach (see Methods). BooleaBayes leverages sparsity (the in-degree of any node is much less than the total number of nodes) in the underlying regulatory network structure, allowing it to make partially constrained predictions about regulatory dynamics, even in regions of state space that are not represented in the data. An advantage of this method is that its predictions are intrinsic to the parts of the



		NE			NE-V1			NE-V2					Non-NE	
ASCL1	★	■	■		□	□	★	■	■	■	■		□	□
FOXA1	★	■	■		□	□	★	■	■	■	■		□	□
FOXA2	★	■	■		□	□	★	■	■	■	■		□	□
ELF3	★	■	■		□	□	★	■	■	■	■		□	□
RBPJ	★	■	■		□	□	★	■	■	■	■		□	□
FLI1	★	■	■		□	□	★	■	■	■	■		□	□
SMAD4	★	■	■		□	□	★	■	■	■	■		□	□
NROB2	★	■	■		□	□	★	■	■	■	■		□	□
NROB1		□	□		■	□	★	■	■	■	■		□	□
BCL3		□	□		□	□	★	■	■	■	■	★	■	■
STAT6		□	□		□	□	★	■	■	■	■	★	■	■
ISL1	★	■	■		□	□		■	■	□	□		□	□
SOX11	★	■	■	★	□	□	★	■	■	□	□	★	□	□
CEBPD	★	■	□		□	□	★	■	□	■	□		□	■
EBF1	★	■	□	★	■	■		■	□	■	□		□	□
TCF4	★	■	■	★	■	■		□	□	□	□		□	□
RCOR2	★	■	■	★	■	■		□	□	□	□		□	□
TCF3	★	■	■	★	■	■		□	□	□	□		□	□
NEUROD2		□	□	★	■	■		□	□	□	□		□	□
OLIG2		□	□	★	■	■		□	□	□	□		□	□
MITF		□	□		■	■	★	□	□	□	□	★	■	■
SIX5		□	□	★	■	■		□	□	□	□	★	■	■
TEAD4		□	□	★	■	■		□	□	□	□	★	■	■
ZNF217		□	□	★	■	■		□	□	□	□	★	■	■
KLF2		■	□		□	□		□	□	□	□	★	■	■
GATA4		□	□		□	□		□	□	□	□	★	■	■
REST		□	□		□	□		□	□	□	□	★	■	■

1. ASCL1 expression data for single sample (cell line) from CCLE dataset (red box with yellow outline).
2. The parent nodes of this sample from 1 are best represented by the gray path: OLIG2 OFF, TEAD4 OFF, FLI1 ON, and so on.
3. Because the gray path best represents the sample, the column below the gray path intersects the green row at the pink square, so this square is shaded darker, representing a higher weight.
4. This is continued for each sample (row). The orange column shows the weights associated with the gray path for one state of the parent nodes.
5. By multiplying the expression of the left column by the matrix of weights, the rule (bottom) is produced. For example, since most of the higher weights in the orange column are in rows with red (high) expression, the inferred rule for the state suggests that ASCL1 should be ON when OLIG2 is OFF, TEAD4 is OFF, FLI1 is ON, and so on.

Fig 7. TF network simulations reproduce subtypes as attractors. **A**. Regulatory network of differentially expressed TFs from each of the 11 co-expressed gene modules in Fig 2B. Colors indicate which phenotype each TF is upregulated in. Red edges indicate inhibition (on average), and green activation (on average). **B**. Probabilistic Boolean rule fits for ASCL1. The target gene is a function of all the genes along the binary tree at the top, while expression of the target is shown on the left. Each row represents one cell line, each column represents one possible input state, and the bottom shows the inferred function F for every possible input state. Color ranges from 0 = blue (highly confident the TF is off), to 0.5 = white, to 1 = red (highly confident the TF is on). Rows are organized by subtype (top to bottom: NE, NEv1, NEv2, non-NE). **C**. Attractors found with asynchronous updates of Boolean network. 10 attractors were found, and each correlates highly with one of the four defined

subtypes (represented by stars). Hamming distance between intra-subtype attractors and inter-subtype attractors are shown. The average distance between intra-subtype attractors was around 2.5, while the average distance between subtype attractors was around 16, signifying that the variation between subtypes is much greater that that within a single subtype. Specifics of the probabilistic simulation are described in Results.

https://doi.org/10.1371/journal.pcbi.1007343.g007

network in which we are most confident, based only on relationships between each TF and its parent nodes. See Methods for more details about the BooleaBayes algorithm.

BooleaBayes rules, like the Boolean example above, describe when a target node will be ON or OFF, given that state of all its regulators. Unlike the Boolean example, BooleaBayes rules are probabilistic, accounting for the (un)certainty with which we can state a target node will turn ON or OFF. For instance, values of 0 means it is certain the target node will turn OFF, 1 means it is certain the target node will turn ON, 0.5 means it is equally likely the target node will turn ON or OFF. BooleaBayes rules were derived for each node of the SCLC TF network in Fig 7A. As an example, Fig 7B shows the rule fitting for one node, ASCL1. Rules for all other nodes are given in S7 Fig. Cross-validation suggested BooleaBayes did not overfit the data (S6 Fig). We simulated the dynamics of the Boolean network using a general-asynchronous update scheme [43]. This formed a state transition graph (STG), in which each state is defined by a vector of TF ON/OFF expression values.

Initial states for simulation were chosen near where we expected the four subtypes would be, by discretizing the average TF expression for each of the four SCLC subtypes. We exhaustively searched the neighborhood of each of these starting states out to a distance of 6 TF changes in the STG (Algorithm 1). Within these neighborhoods, we found 10 states for which all 27 TFs had at least a 50% chance of remaining unchanged. Transitions into these states are therefore more likely, and escapes less likely. Thus, these 10 states represent semi-stable states of the network dynamics (Fig 7C), that we refer to as pseudo-attractors. We also searched within neighborhoods of over 200 random initial states (allowing us to search over 200,000 total additional states), and found no additional pseudo-attractors (S6 Table).

These 10 pseudo-attractor states each correlated with, and could be assigned to, one of the 4 SCLC subtypes (stars in Fig 7C); this indicates that the updated network structure and BooleaBayes rules are sufficient to capture stability of the four SCLC phenotypes. Having identified network dynamics that closely match experimental observations, we are now in a position to perform *in silico* (de)stabilizing perturbations and predict the resulting trajectory through the STG for each subtype. We do so in the next section.

In silico SCLC network perturbations identify master regulators and master destabilizers of SCLC phenotypes

To quantify the baseline stability of the steady states in Fig 7C, we performed random walks (algorithm described in Methods) starting from each of the 10 pseudo-attractors. We counted how many steps were required to reach a state more than 4 TFs away (Hamming distance greater than 4) from the starting state (Fig 8, Algorithm 2). We chose a 4-TF neighborhood to account for the models' greatest intra-subtype attractor variability (Fig 7C, Hamming Distance), and therefore movement within the 4-TF neighborhood of a starting state is still considered reflective of that subtype. For each simulation, one TF in the network was either activated (held constant at TF = 1) or silenced (TF = 0) in each of the stable states (Fig 7C). 1000 random walks were executed for each condition. The number of steps in each random walk required to leave the 4-TF neighborhood was recorded in a histogram (Fig 8A). We defined (de)stabilization as the percent decrease or increase of the average number of steps

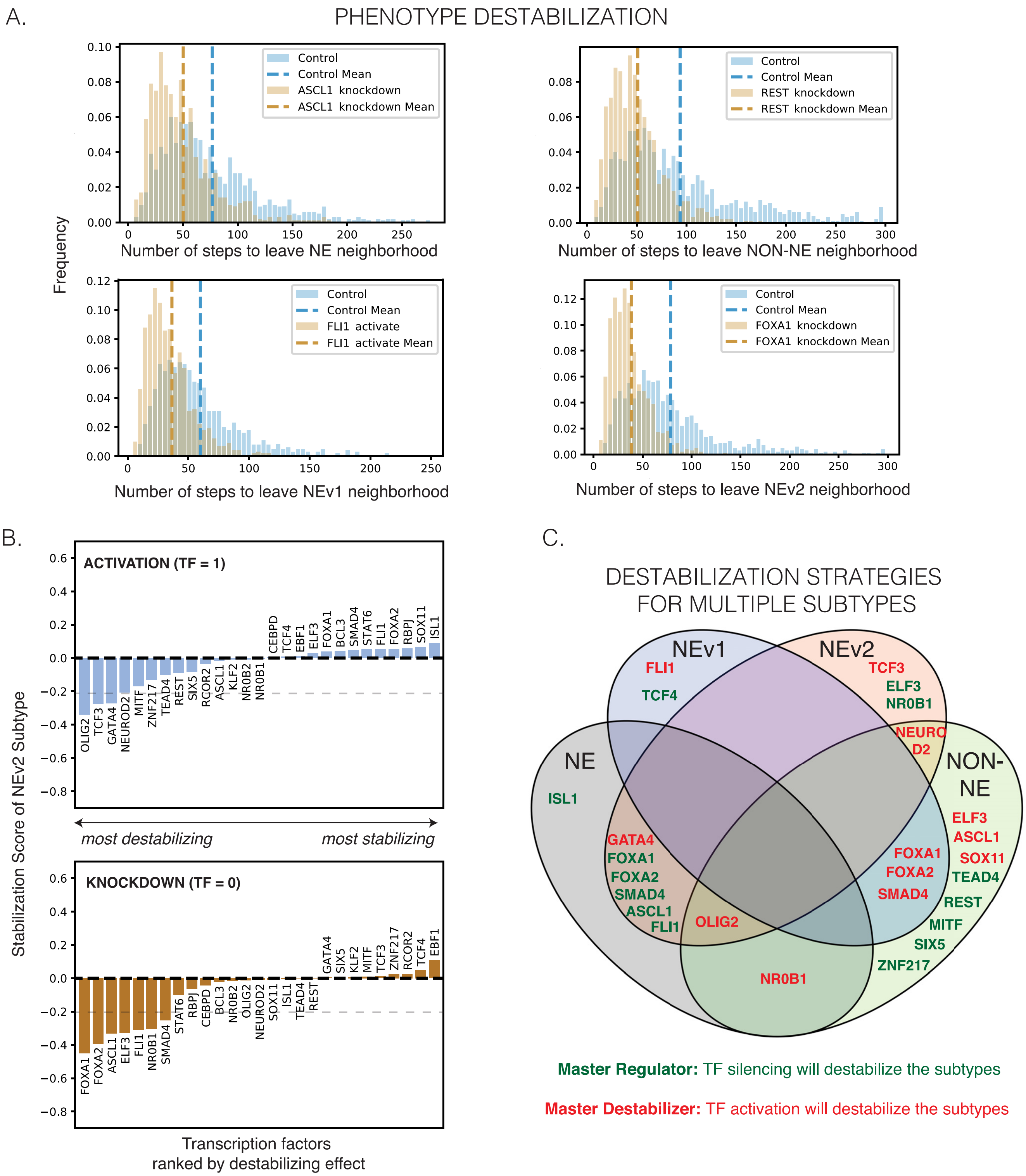


Fig 8. Destabilization of subtypes by perturbation to network. A. Random walks starting from the attractors in Fig 7C will eventually leave the start state due to uncertainty in the Boolean rules. Control histogram shows how many random steps are required to reach a state with a Hamming distance ≥ 4 under the network's natural dynamics. The knockdowns and activations shown here hold expression of the perturbed gene OFF or ON in an attempt to destabilize the start state, such that the random walk leaves the neighborhood sooner. A shift to the left in the perturbed distribution signifies that the perturbation "pushed" the simulated cell out of the 4-TF neighborhood more quickly, and the perturbation thus "destabilized" the subtype represented by the start state. This indeed occurs for several perturbations, shown for NE, NEv1, NEv2, and non-NE starting states. Dotted line shows mean for each histogram, which is

used to calculate the change in average number of steps under perturbation. **B**. Ranking of phenotype stabilization of NEv2 by TF activation and knockdown. The percent change of stability measures the percent change in the average number of steps needed to leave the neighborhood of the stable states. Negative stabilization scores indicates destabilizing perturbations, while positive indicates increasing stability. Results are shown for 1000 iterations starting from NEv2. Similar plots for the other subtypes can be found in S8 Fig. Dotted line at $y = -0.2$ signifies the cutoff for "destabilizing" perturbations shown in C. **C**. A Venn diagram demonstrating overlap of destabilization strategies. A single activation (green text) or knockdown (red text) can sometimes destabilize multiple phenotypes.

https://doi.org/10.1371/journal.pcbi.1007343.g008

under perturbation relative to the unperturbed reference (Fig 8B, S8 Fig). For example, either activation of GATA4 or silencing FOXA1 are predicted to destabilize both the NE and NEv2 subtypes (Fig 8C).

TFs that, when silenced, cause destabilization greater than 20% (score ≤ -0.2) of a specific subtype were considered master regulators of that subtype. They include *REST* (non-NE) (in agreement with [10]), *TEAD4* (non-NE), *ISL1* (NE), and *TCF4* (NEv1). TEAD4 is downstream mediator of YAP1 action, which has been previously identified as a possible phenotypic modulator in a subset of SCLC cell lines [45]; our analyses suggest that expression of TEAD4 may be able to stabilize this phenotype. Simulations of the network also identified the novel NEv2 master regulators, *ELF3* and *NR0B1*.

Our network simulations further identified TFs that can be considered master "destabilizers", i.e., *activation* of these TFs destabilizes a specific phenotype by at least 20%. For instance, activation of *ELF3* is predicted to destabilize non-NE, while activation of *NR0B1* would destabilize both non-NE and NE subtypes. Simulations identified a single master destabilizer for NEv2, the TF *TCF3* (Fig 8C). Taken together, our pipeline, which includes subtype identification, drug response analysis, and network simulations, suggests possible therapeutic perturbations that could shift the phenotypic landscape of SCLC into a more sensitive state for treatment.

Discussion

We report a systems approach to understanding SCLC heterogeneity that integrates transcriptional, mutational, and drug-response data. Our findings culminate in discrimination and mechanistic insight into the four SCLC subtypes shown in Table 1: NE, non-NE, NEv1, and NEv2. Within the context of the broader literature on SCLC heterogeneity, we showed that

Table 1. Comprehensive framework for SCLC heterogeneity. Using our workflow, we have characterized 4 SCLC subtypes by gene expression, drug response, and master regulators and destabilizers.

Subtype	GO Terms	Gene Modules	Drug Response	Regulators	Destabilizers	Nomenclature adapted from [32]
NE	Neuron differentiation, Synaptic transmission, Cell-cycle regulation, Epithelial cell differentiation, Sensory perception	Turquoise, Brown, Yellow		ISL1, ASCL1, FLI1, SMAD4, FOXA1, FOXA2	GATA4, OLIG2, NR0B1	SCLC-A1
NEv1	Neuron differentiation, neurotransmitter secretion, Synapse organization, Cardiac muscle cell contraction, Mechanoreceptor differentiation, Sensory perception	Yellow, Salmon, Pink, Red	Sensitive to AKIs	TCF4	FLI1, FOXA1, FOXA2, SMAD4	SCLC-N
NEv2	Immune response, Drug catabolism, Ion transport, Homeostasis, Epithelial cell differentiation, Glycosylation	Brown, Green, Midnight blue	Least sensitive	ELF3, NR0B1, FOXA1, FOXA2, SMAD4, ASCL1, FLI1	TCF3, GATA4, NEUROD2, OLIG2	SCLC-A2
NON-NE	Immune Response, Stress Response, Defense response, Macrophage activation, Myeloid cell differentiation and activation, neutrophil-mediated immunity, vesicle-mediated transport, Angiogenesis, Viral processes	Black, Blue, Red	Most sensitive	TEAD4, REST, MITF, SIX5, ZNF217	ELF3, ASCL1, SOX11, FOXA1, FOXA2, SMAD4, NR0B1, OLIG2	SCLC-Y

https://doi.org/10.1371/journal.pcbi.1007343.t001

NE, non-NE, and NEv1 correspond to several subtypes that have been previously reported based on a few markers–more specifically, SCLC-A, SCLC-Y, and SCLC-N, respectively [32]. Significantly, we find that one (NEv2) has not been described previously, and which is nearly indistinguishable from NE based on currently used markers of SCLC heterogeneity. Because this subtype has high expression of ASCL1, it would be SCLC-A2 in the nomenclature used in a recent review [32].

Tumor deconvolution by CIBERSORT and scRNA-seq data indicate that a large proportion of human and mouse tumors comprise more than one subtype (Fig 6). While MutSigCV mutational analysis did not find any significant differences in mutated genes between subtypes (S5 Fig), we cannot rule them out, and future studies may uncover genomic mechanisms interfacing with the epigenetic heterogeneity reported here. Existing examples of epigenetic intratumoral heterogeneity are often framed in the context of transitions between epithelial and mesenchymal differentiation states [41]. Mechanisms underlying SCLC differentiation heterogeneity remain to be defined, and they may include functional states of PNECs, distinct cells of origin, or response to microenvironmental factors. It remains to be seen whether changes in tumor composition after treatment (Fig 5B) are due to phenotypic transitions, selection, or both.

A drug screen across a broad range of compounds indicated that the NEv2 subtype is more resistant than the others, especially in response to *AURK* and *mTOR* inhibitors. This is reminiscent of a new hybrid EMT phenotype recently identified as more aggressive and drug resistant than other phenotypes [46–48]. More broadly, recent reviews have suggested that both genetic mutations and epigenetic regulators such as histone demethylases may affect intratumoral heterogeneity and modulate therapeutic response [49]. Additionally, non-genetic processes such as phenotypic plasticity and stochastic cell-to-cell variability may enable tumor cells to evade therapy and give rise to drug-tolerant persisters [47, 50]. Our findings of differential drug response across subtypes corroborate the significance of these reports. *In vivo* verification of NEv2's drug resistant properties in mouse and human tumors will be an important next step. Along these lines, it is tempting to speculate that the increase of the NEv2 signature in patient MGH1518 after drug treatment (Fig 5B) may be responsible for acquired drug-resistance in this patient. However, this study was under-powered for our analyses, and more experimental data will be necessary to strengthen this conclusion.

A significant advance of our work is the introduction of BooleaBayes, which we developed to infer mechanistic insights into regulation of the heterogeneous SCLC subtypes. By considering the distinct subtype clusters as attractors of a gene regulatory network, BooleaBayes infers partially-constrained mechanistic models. A key benefit of this method is that it does not overfit data: predictions are based only on parts of the network for which available data can constrain the dynamics, while states that lack constraining data diffuse randomly. With this method we were able to recapitulate known master regulators of SCLC heterogeneity, as well as identify novel ones such as *ISL1* (NE) and *TEAD4* (non-NE). Additionally, we predict *ELF3* and *NR0B1* to be master regulators of the NEv2 phenotype. Furthermore, we introduce the label of "master destabilizers" to describe TFs whose activation will destabilize a phenotype. Our method gives a systematic way to rank perturbations that may destabilize a resistant phenotype. We emphasize that BooleaBayes provides an adaptive roadmap to systematically walk the circle from prediction to experimental validation and back. Thus, a prediction from BooleaBayes about stabilizers can be experimentally tested, and the outcome will inform a new datapoint to further constrain the BooleaBayes model to refine predictions. For instance, if cells become stuck in a previously unknown partially reprogrammed attractor [51], expression data from these cells may be added to constrain BooleaBayes in a region where no data previously existed. In ongoing work, we are validating these predictions experimentally. We propose that with BooleaBayes, our approach for identifying master TFs could be applicable to other systems, including other cancer types or transcriptionally-regulated diseases. This approach parallels other modeling techniques to identify phenotypic stability factors, such as recent bifurcation analysis on an EMT network [52, 53].

While many of the previously reported subtypes of SCLC fit into our framework, a few are noticeably absent, and will require further study. The vasculogenic subtype of SCLC described by Williamson et al. [54] did not emerge from our analysis. We speculate that this may be due to the rarity and/or instability of this CTC-derived phenotype among the available SCLC cell lines. Denny and Yang et al. have previously reported that *Nfib* amplification promotes metastasis [55]; however, our clusters do not correlate with location of the tumor sample from which each cell line was derived (*e.g.*, primary vs metastatic, S1 Table). Poirier et al., using a similar clustering approach to ours, identified highly methylated SCLC subtypes (M1 and M2) [33], and the correspondence of these subtypes with the ones described here is intriguing and remains to be defined. Finally, Huang et al. recently reported an SCLC subtype defined by expression of *POU2F3* [12]. In our data, POU2F3 was highly expressed in only four cell lines and was placed into a small (328 genes, green-yellow) module, and therefore represented only a small signal in our data. Overall, future studies with additional cell line and/or mouse data may be used to further investigate these different subtypes, underscoring that the delineation of four subtypes here does not preclude the existence of others.

To identify subtype clusters and BooleaBayes rules, we rely on the underlying assumption of bulk RNA-seq data that single-cells from each cell line belong to one cellular state. While this is consistent with our previous findings that SCLC cell lines resolve into discrete clusters by flow cytometry [16], future cell-line analysis at single-cell resolution may refine our results, and it will be interesting to see to what extent subtype heterogeneity may be reflected within one cell line.

An advantage of our analyses is that each subtype is defined by distinct co-expressed gene programs, rather than by expression of one or few markers, which has been customary in the field but has limited ability to discriminate between phenotypes (Fig 2B). In addition, these modules participate in unique biological processes (e.g., as identified by GO), such that the systems-level approach presented here may provide a comprehensive framework to understand the regulation and functional consequences of SCLC heterogeneity in a tumor. This understanding can be actionable since SCLC subtypes show differential drug sensitivity; for example, our analyses in this paper support the hypothesis that NEv2 may be a drug-resistant phenotype of SCLC. We propose that identification of drugs targeting the NEv2 subtype, or perturbagens that reprogram it toward less recalcitrant states, may lead to improved treatment outcomes for SCLC patients.

Algorithm 1 Limited Pseudo-Attractor Search

```
procedure SEARCH ENTIRE STG IN NEIGHBORHOOD OF GIVEN STATE TO FIND PSEUDO-ATTRACTORS
```

state_init $\in \{0, 1\}^N$	Initial Boolean state (*N* dimensional vector of 0's and 1's, where *N* is the number of TFs)
f: $\{0, 1\}^N \mapsto [0, 1]^N$	Probabilistic update rules mapping the current *state* $\in \{0, 1\}^N$ to a probability (value between 0 and 1) for each TF to flip (ON to OFF, or OFF to ON)
$d : \{0,1\}^N \times \{0,1\}^N \mapsto \mathbb{R}$	Function to calculate distance between two states (we use the Hamming distance)
R	Maximum radius to search from *state_init*
P_T	Threshold probability used to define pseudo-attractors (we used $P_T = 0.5$ so that pseudo-attractors are defined to have out-transitions with probability less than 50%.)

```
Inputs:
Output:
PseudoAttractors—A set of strongly connected components of the state transition graph for which transitions in have probability greater than $P_T$
  pending $\leftarrow$ {state_init} (A set containing the initial state)
  STG $\leftarrow$ empty directed graph
  oob $\leftarrow$ dummy vertex (this will be the "out-of-bounds" vertex—all states in the STG with distance greater than R from the initial state will point to this vertex, preventing them from being detected as attractors)
  Add oob to STG
  Add state_init to STG
  while pending is not empty do
    state $\leftarrow$ POP any state from pending
    if $d(state, state\_init) > R$ then
      Add edge from state $\rightarrow$ oob with weight = 1
    else
      for i in 1..N do
        neighbor $\leftarrow$ state
        $neighbor_i \leftarrow$ NOT $state_i$ (Flip TF i to get the neighbor)
        if neighbor is not in STG then
          Add neighbor to STG
          Add neighbor to pending
        Add edge from state $\rightarrow$ neighbor with $weight = f(state)_i$ (add the transition, with probability given by f)
  $STG_{pruned} \leftarrow$ STG
  Remove all edges with $weight < P_T$ (prune edges with probability less than given threshold)
  PseudoAttractors $\leftarrow$ empty set
  for SCC in strongly connected components of $STG_{pruned}$ do
    ### First make sure this SCC does not contain the dummy vertex, which by definition has no out-transitions
    if obb is not in SCC then
      if There are no edges in $STG_{pruned}$, from any node within SCC to any node not within SCC then
        Add SCC to PseudoAttractors
  Return: PseudoAttractors
```

Algorithm 2 Probabilistic Boolean Random Walk

```
procedure RANDOM WALK TO DETERMINE STABILITY OF INITIAL CONDITION
```

$state\_init \in \{0, 1\}^N$	Initial Boolean state (*N* dimensional vector of 0's and 1's, where *N* is the number of TFs)
$f: \{0, 1\}^N \mapsto [0, 1]^N$	Probabilistic update rules mapping the current $state \in \{0, 1\}^N$ to a probability (value between 0 and 1) for each TF to flip (ON to OFF, or OFF to ON)
$d : \{0,1\}^N \times \{0,1\}^N \mapsto \mathbb{R}$	Function to calculate distance between two states (we use the Hamming distance)
R	Maximum allowed distance from *state_init*
fixed_TFs	Set of TFs that are held constant (*i.e.*, perturbed to be ON or OFF)

```
Inputs:
Output:
Number of steps taken before the random walk is a distance greater than R from state_init
```

```
  state ← state_init
  steps ← 0
  while $d(state, state\_init) \le R$ do
    steps ← steps + 1
    i ← a random integer between 1 and N, excluding fixed_TFs (randomly chose one, non-fixed, TF to update)
    probability_update ← $f(state)_i$ (probability of flipping TF i)
    r ← a uniform random number between 0 and 1
    if $r <$ probability_update then
      $state_i$ ← NOT $state_i$ (flip TF i from ON to OFF, or OFF to ON)
  Return: steps
```

Supporting information

S1 Fig. Tracking plot, delta area plot, and CDF for consensus clustering and ProgenyClust scores with different values of k. A. Tracking plot shows slight inconsistency for cell lines with k = 3. One of these is assigned to the "light green" cluster in the k = 3 clustering scheme, whereas when k = 4, it returns to the "light blue" cluster. The others are in the "dark blue" cluster when k = 2 and "light blue" cluster when k = 3. Thus k = 3 is not a good fit to the data **B**. The delta area plot shows the relative change in the area under the CDF curve (Fig 2A). The largest changes in area occur between k = 2 and k = 4, at which point the relative increase in area becomes noticeably smaller (from an increase of 0.5 and 0.4 to 0.15). This suggests that k = 4,5, or 6 are the best clustering that maximizes detail (more, smaller clusters present a more detailed picture than a few large clusters) and minimizes noise (by minimizing average pairwise consensus values and maximizing extreme pairwise consensus values). Average cluster consensus scores (CCS) across clusters show that k = 4 may be the best choice because it has the highest average (k = 4 average CCS: 0.848, k = 5 average CCS: 0.814, k = 6 average CCS: 0.762). **C**. Consensus Cumulative Distribution Function. This CDF show that k = 4 has more black cells and white cells than gray, suggesting the consensus clusters are more robust. **D**. ProgenyClust suggests that the optimal number of clusters is 3 or 4, as determined by the maximum of the Gap and Score criteria values. As k = 3 was already ruled out above, we chose k = 4 to describe the heterogeneity between cell lines.
(EPS)

S2 Fig. Gene set enrichment analysis. Enrichment plots for significantly enriched gene co-expression modules, as determined by Gene Set Enrichment Analysis, for the NE and NEv1 subtypes.
(EPS)

S3 Fig. Same as S2 Fig, for the NEv2 and non-NE subtypes.
(EPS)

S4 Fig. GO maps for each gene module. GO phenospace maps, as in Fig 3C, for all modules.
(PDF)

S5 Fig. Significant mutations across subtypes. Significantly mutated genes across 50 SCLC cell lines, as determined by MutSigCV, ordered by significance. As expected, significant mutations were found in both the *Rb1* and *Tp53* genes. Inspection by eye shows that no significant mutations can distinguish completely between two or more phenotypes. This suggests an alternate source of heterogeneity, such as transcriptional regulation. Significance cut-off: q (p-value corrected for multiple comparisons) ≤ 0.25. $q \le 0.5$ shown.
(EPS)

S6 Fig. Cross-validation of network inference and comparison to other Boolean rule schemes. A. To test for overfitting, we partitioned the samples into training sets (80% of samples) used to fit BooleaBayes rules, and testing sets (20%), over 70 iterations. We calculated the squared error between BooleaBayes predictions and the true values in both the training and testing sets. The squared error was similar for the training and testing data, suggesting the results are not due to overfitting. **B**. Correlation between prediction and measured expression is a function of predicted expression from the BooleaBayes rule. For values $x < 0.5$ along the x-axis, correlation is calculated on a subset of the data for which prediction $< x$ or prediction $> 1 - x$ (*e.g.*, at $x = 0.1$, correlation is calculated on subset with prediction 0 to 0.1 or 0.9 to 1. When $x = 0.5$, all data are included. When $x = 0.9$, data subset is identical to when $x = 0.1$). In cases where the BooleaBayes predictions are confident whether a gene was ON (near 1) or OFF (near 0), the correlation between the prediction and the actual expression approaches 1. When BooleaBayes is not confident (near 0.5), the correlation decreases. Both (A) and (B) are consistent with our interpretation that BooleaBayes is fitting rules only where the data are well constrained. BooleaBayes on both the training and testing data sets predicts expression with high confidence, especially when the actual expression is near 0 or 1. **C**. Inhibitory dominant and majority rule update schemes both converge onto the same, single attractor (left-most state). This attractor has a Hamming distance of 4-6 away from both the NE and NEv2 attractors found by BooleaBayes (right 4 states), suggesting that it may represent an "average" NE attractor. We speculate it is an artifact of the coarse-graining imposed by those rules, as they both assume simplistic TF interactions.
(EPS)

S7 Fig. Rules for all transcription factors in network. Rules for all TFs, as in Fig 7B.
(PDF)

S8 Fig. Stabilization of SCLC phenotypes by TF knockdown and activation. The percent change of stability measures the percent change in the average number of steps needed to leave the neighborhood of the stable states. Negative indicates destabilizing, while positive indicates increasing stability. Results are shown for 1000 iterations starting from **A**. NE, **B**. NEv1, and **C**. non-NE. NEv2 is shown in Fig 8.
(EPS)

S1 File. Interactive GO maps, as in Fig 3C.
(ZIP)

S1 Table. Cell line characteristics and consensus clusters.
(CSV)

S2 Table. WGCNA module eigengenes.
(CSV)

S3 Table. Enriched gene ontology terms in subtype modules.
(CSV)

S4 Table. Enriched modules in subtypes: Gene set enrichment analysis statistics.
(CSV)

S5 Table. Drug response of subtypes grouped by drug target.
(CSV)

S6 Table. Network simulations starting at random states in the state transition graph. We ran multiple simulations searching various regions of the state transition graph, and did not

find any additional attractors. This suggests that BooleaBayes was capable of identifying all of the biologically relevant attractors, but does not preclude the possibility of additional unseen attractors using our network structure and BooleaBayes rules.
(CSV)

Acknowledgments

We thank members of the Quaranta laboratory for useful suggestions; members of Carlos Lopez laboratory (Vanderbilt University) for critical feedback and support with computation; and Wade Iams, Christine Lovly, and Jonathan Lehman (Vanderbilt University Medical Center) for their clinical expertise. We also thank the Stanford Functional Genomics Facility for technical support with the 10x Genomics experiments. J.S. is the Harriet and Mary Zelencik Scientist in Children's Cancer and Blood Diseases.

Author Contributions

Conceptualization: David J. Wooten, Sarah M. Groves, Darren R. Tyson, Carlos F. Lopez, Vito Quaranta.

Data curation: David J. Wooten, Sarah M. Groves, Jing S. Lim, Julien Sage.

Formal analysis: David J. Wooten, Sarah M. Groves, Darren R. Tyson, Qi Liu.

Funding acquisition: David J. Wooten, Sarah M. Groves, Carlos F. Lopez, Julien Sage, Vito Quaranta.

Investigation: David J. Wooten, Sarah M. Groves.

Methodology: David J. Wooten, Sarah M. Groves, Réka Albert.

Project administration: Vito Quaranta.

Resources: Carlos F. Lopez, Julien Sage.

Software: David J. Wooten, Sarah M. Groves, Qi Liu.

Supervision: Darren R. Tyson, Vito Quaranta.

Validation: David J. Wooten, Sarah M. Groves.

Visualization: David J. Wooten, Sarah M. Groves, Qi Liu.

Writing – original draft: David J. Wooten, Sarah M. Groves, Vito Quaranta.

Writing – review & editing: David J. Wooten, Sarah M. Groves, Darren R. Tyson, Qi Liu, Jing S. Lim, Réka Albert, Carlos F. Lopez, Julien Sage, Vito Quaranta.

References

1. Hauschild A, Grob JJ, Demidov LV, Jouary T, Gutzmer R, Millward M, et al. Dabrafenib in BRAF-mutated metastatic melanoma: a multicentre, open-label, phase 3 randomised controlled trial. The Lancet. 2012; 380(9839):358–365. https://doi.org/10.1016/S0140-6736(12)60868-X
2. Robert M, Frenel JS, Gourmelon C, Patsouris A, Augereau P, Campone M. Olaparib for the treatment of breast cancer. Expert Opinion on Investigational Drugs. 2017; 26(6):751–759. https://doi.org/10.1080/13543784.2017.1318847 PMID: 28395540
3. Travis WD, Brambilla E, Noguchi M, Nicholson AG, Geisinger KR, Yatabe Y, et al. International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society International Multidisciplinary Classification of Lung Adenocarcinoma. Journal of Thoracic Oncology. 2011; 6(2):244–285. https://doi.org/10.1097/JTO.0b013e318206a221 PMID: 21252716

4. Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature methods. 2015; 12(5):453–7. https://doi.org/10.1038/nmeth.3337 PMID: 25822800

5. George J, Lim JS, Jang SJ, Cun Y, Ozretić L, Kong G, et al. Comprehensive genomic profiles of small cell lung cancer. Nature. 2015; 524(7563):47–53. https://doi.org/10.1038/nature14664 PMID: 26168399

6. Gazdar AF, Carney DN, Nau MM, Minna JD. Characterization of variant subclasses of cell lines derived from small cell lung cancer having distinctive biochemical, morphological, and growth properties. Cancer research. 1985; 45(6):2924–30. PMID: 2985258

7. Carney DN, Gazdar AF, Bepler G, Guccion JG, Marangos PJ, Moody TW, et al. Establishment and identification of small cell lung cancer cell lines having classic and variant features. Cancer Research. 1985; 45(6):2913–2923. PMID: 2985257

8. Gazdar AF, Bunn PA, Minna JD. Small-cell lung cancer: what we know, what we need to know and the path forward. Nature Reviews Cancer. 2017; 17(12):725–737. https://doi.org/10.1038/nrc.2017.87 PMID: 29077690

9. Sutherland KD, Proost N, Brouns I, Adriaensen D, Song JY, Berns A. Cell of origin of small cell lung cancer: inactivation of Trp53 and Rb1 in distinct cell types of adult mouse lung. Cancer Cell. 2011; 19 (6):754–764. https://doi.org/10.1016/j.ccr.2011.04.019 PMID: 21665149

10. Lim JS, Ibaseta A, Fischer MM, Cancilla B, O'Young G, Cristea S, et al. Intratumoral heterogeneity generated by Notch signaling promotes small cell lung cancer. Nature. 2017; 545(7654):360–364. https://doi.org/10.1038/nature22323 PMID: 28489825

11. Calbo J, van Montfort E, Proost N, van Drunen E, Beverloo HB, Meuwissen R, et al. A functional role for tumor cell heterogeneity in a mouse model of small cell lung cancer. Cancer Cell. 2011; 19(2):244–256. https://doi.org/10.1016/j.ccr.2010.12.021 PMID: 21316603

12. Huang YH, Klingbeil O, He XY, Wu XS, Arun G, Lu B, et al. POU2F3 is a master regulator of a tuft cell-like variant of small cell lung cancer. Genes and Development. 2018. https://doi.org/10.1101/gad.314815.118

13. Sos ML, Dietlein F, Peifer M, Schöttle J, Balke-Want H, Müller C, et al. A framework for identification of actionable cancer genome dependencies in small cell lung cancer. Proceedings of the National Academy of Sciences of the United States of America. 2012; 109(42):17034–9. https://doi.org/10.1073/pnas.1207310109 PMID: 23035247

14. Borromeo MD, Savage TK, Kollipara RK, He M, Augustyn A, Osborne JK, et al. ASCL1 and NEUROD1 Reveal Heterogeneity in Pulmonary Neuroendocrine Tumors and Regulate Distinct Genetic Programs. Cell Reports. 2016. https://doi.org/10.1016/j.celrep.2016.06.081 PMID: 27452466

15. Mollaoglu G, Guthrie MR, Böhm S, Brägelmann J, Can I, Ballieu PM, et al. MYC Drives Progression of Small Cell Lung Cancer to a Variant Neuroendocrine Subtype with Vulnerability to Aurora Kinase Inhibition. Cancer Cell. 2017. https://doi.org/10.1016/j.ccell.2016.12.005 PMID: 28089889

16. Udyavar AR, Wooten DJ, Hoeksema M, Bansal M, Califano A, Estrada L, et al. Novel hybrid phenotype revealed in small cell lung cancer by a transcription factor network model that can explain tumor heterogeneity. Cancer Research. 2017; 77(5). https://doi.org/10.1158/0008-5472.CAN-16-1467 PMID: 27932399

17. Tammela T, Sanchez-Rivera FJ, Cetinbas NM, Wu K, Joshi NS, Helenius K, et al. A Wnt-producing niche drives proliferative potential and progression in lung adenocarcinoma. Nature. 2017; 545 (7654):355–359. https://doi.org/10.1038/nature22334 PMID: 28489818

18. Drapkin BJ, George J, Christensen CL, Mino-Kenudson M, Dries R, Sundaresan T, et al. Genomic and Functional Fidelity of Small Cell Lung Cancer Patient-Derived Xenografts. Cancer discovery. 2018; 8 (5):600–615. https://doi.org/10.1158/2159-8290.CD-17-0935 PMID: 29483136

19. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics (Oxford, England). 2010; 26(12):1572–1573. https://doi.org/10.1093/bioinformatics/btq170

20. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: An R package for measuring semantic similarity among GO terms and gene products. Bioinformatics. 2010. https://doi.org/10.1093/bioinformatics/btq064

21. Polley E, Kunkel M, Evans D, Silvers T, Delosh R, Laudeman J, et al. Small Cell Lung Cancer Screen of Oncology Drugs, Investigational Agents, and Gene and microRNA Expression. Journal of the National Cancer Institute. 2016. https://doi.org/10.1093/jnci/djw122 PMID: 27247353

22. Lubbock ALR, Harris LA, Quaranta V, Tyson DR, Lopez CF. Visualization and analysis of high-throughput in vitro dose-response datasets with Thunor. bioRxiv. 2019; p. 530329.

23. Harris LA, Frick PL, Garbett SP, Hardeman KN, Paudel BB, Lopez CF, et al. An unbiased metric of antiproliferative drug effect in vitro. Nature Methods. 2016; 13(6):497–500. https://doi.org/10.1038/nmeth.3852 PMID: 27135974

24. Jahchan NS, Lim JS, Bola B, Morris K, Seitz G, Tran KQ, et al. Identification and Targeting of Long-Term Tumor-Propagating Cells in Small Cell Lung Cancer. Cell Reports. 2016; 16(3):644–656. https://doi.org/10.1016/j.celrep.2016.06.021 PMID: 27373157

25. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018; 36(5):411–420. https://doi.org/10.1038/nbt.4096 PMID: 29608179

26. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013; 499(7457):214–218. https://doi.org/10.1038/nature12213 PMID: 23770567

27. Stransky N, Ghandi M, Kryukov GV, Garraway LA, Lehár J, Liu M, et al. Pharmacogenomic agreement between two cancer cell line data sets. Nature. 2015; 528(7580):84–7. https://doi.org/10.1038/nature15736

28. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. Bioinformatics (Oxford, England). 2010; 26(19):2438–2444. https://doi.org/10.1093/bioinformatics/btq466

29. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: Proceedings of the 7th Python in Science Conference (SciPy); 2008.

30. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learning. 2003. https://doi.org/10.1023/A:1023949509487

31. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483 (7391):603–607. https://doi.org/10.1038/nature11003 PMID: 22460905

32. Rudin CM, Poirier JT, Byers LA, Dive C, Dowlati A, George J, et al. Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data; 2019.

33. Poirier JT, Gardner EE, Connis N, Moreira AL, De Stanchina E, Hann CL, et al. DNA methylation in small cell lung cancer defines distinct disease subtypes and correlates with high expression of EZH2. Oncogene. 2015. https://doi.org/10.1038/onc.2015.38 PMID: 25746006

34. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC bioinformatics. 2008; 9:559. https://doi.org/10.1186/1471-2105-9-559 PMID: 19114008

35. Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 Update. Nucleic Acids Research. 2013. https://doi.org/10.1093/nar/gks1055 PMID: 23143270

36. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005; 102(43):15545–15550. https://doi.org/10.1073/pnas.0506580102

37. Paraschiv B, Diaconu CC, Toma CL, Bogdan MA. Paraneoplastic syndromes: The way to an early diagnosis of lung cancer; 2015.

38. Van Der Maaten LJP, Hinton GE. Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research. 2008.

39. Umemura S, Mimaki S, Makinoshima H, Tada S, Ishii G, Ohmatsu H, et al. Therapeutic priority of the PI3K/AKT/mTOR pathway in small cell lung cancers as revealed by a comprehensive genomic analysis. Journal of Thoracic Oncology. 2014. https://doi.org/10.1097/JTO.0000000000000250 PMID: 25122428

40. Yeo SK, Guan JL. Breast Cancer: Multiple Subtypes within a Tumor?; 2017.

41. Bocci F, Jolly MK, Onuchic JN. A biophysical model of Epithelial-Mesenchymal Transition uncovers the frequency and size distribution of Circulating Tumor Cell clusters across cancer types. bioRxiv. 2019.

42. Harris LA, Beik S, Ozawa PMM, Jimenez L, Weaver AM. Modeling heterogeneous tumor growth dynamics and cell-cell interactions at single-cell and cell-population resolution. Current Opinion in Systems Biology;In Review.

43. Albert I, Thakar J, Li S, Zhang R, Albert R. Boolean network simulations for life scientists. Source Code for Biology and Medicine. 2008; 3:16. https://doi.org/10.1186/1751-0473-3-16 PMID: 19014577

44. Font-Clos F, Zapperi S, La Porta CAM. Topography of epithelial–mesenchymal plasticity. Proceedings of the National Academy of Sciences. 2018; 115(23):5902–5907. https://doi.org/10.1073/pnas.1722609115

45. Horie M, Saito A, Ohshima M, Suzuki HI, Nagase T. YAP and TAZ modulate cell phenotype in a subset of small cell lung cancer. Cancer Science. 2016; 107(12):1755–1766. https://doi.org/10.1111/cas.13078 PMID: 27627196

46. Lu M, Jolly MK, Levine H, Onuchic JN, Ben-Jacob E. MicroRNA-based regulation of epithelial-hybrid-mesenchymal fate determination. Proceedings of the National Academy of Sciences. 2013; 110 (45):18144–18149. https://doi.org/10.1073/pnas.1318192110

47. Jolly MK, Kulkarni P, Weninger K, Orban J, Levine H. Phenotypic Plasticity, Bet-Hedging, and Androgen Independence in Prostate Cancer: Role of Non-Genetic Heterogeneity. Frontiers in Oncology. 2018; 8. https://doi.org/10.3389/fonc.2018.00050

48. Kröger C, Afeyan A, Mraz J, Eaton EN, Reinhardt F, Khodor YL, et al. Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. Proceedings of the National Academy of Sciences. 2019; 116(15):7353–7362. https://doi.org/10.1073/pnas.1812876116

49. Hinohara K, Polyak K. Intratumoral Heterogeneity: More Than Just Mutations. Trends in Cell Biology. 2019; 29(7):569–579. https://doi.org/10.1016/j.tcb.2019.03.003 PMID: 30987806

50. Sharma SV, Lee DY, Li B, Quinlan MP, Takahashi F, Maheswaran S, et al. A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. Cell. 2010; 141(1):69–80. https://doi.org/10.1016/j.cell.2010.02.027 PMID: 20371346

51. Lang AH, Li H, Collins JJ, Mehta P. Epigenetic Landscapes Explain Partially Reprogrammed Cells and Identify Key Reprogramming Genes. PLoS Computational Biology. 2014. https://doi.org/10.1371/journal.pcbi.1003734 PMID: 25122086

52. Hong T, Watanabe K, Ta CH, Villarreal-Ponce A, Nie Q, Dai X. An Ovol2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-step Transition between Epithelial and Mesenchymal States. PLOS Computational Biology. 2015; 11(11):e1004569. https://doi.org/10.1371/journal.pcbi.1004569 PMID: 26554584

53. Jia D, Jolly MK, Boareto M, Parsana P, Mooney SM, Pienta KJ, et al. OVOL guides the epithelial-hybrid-mesenchymal transition. Oncotarget. 2015; 6(17). https://doi.org/10.18632/oncotarget.3623

54. Williamson SC, Metcalf RL, Trapani F, Mohan S, Antonello J, Abbott B, et al. Vasculogenic mimicry in small cell lung cancer. Nature Communications. 2016. https://doi.org/10.1038/ncomms13322

55. Denny SK, Yang D, Chuang CH, Brady JJ, Lim JS, Grüner BM, et al. Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. Cell. 2016. https://doi.org/10.1016/j.cell.2016.05.052 PMID: 27374332