ORIGINAL ARTICLE

Small Cell Lung Cancer subtypes identified by systems-level modeling of transcription factor networks

Wooten, David J.[1*] | **Maddox, Sarah F.**[2*] | **Tyson, Darren R.**[3] | **Liu, Qi**[4] | **Lim, Jing Shan**[5] | **Sage, Julien**[5] | **Quaranta, Vito**[1,3]

[1]Cancer Biology, Vanderbilt University, Nashville, TN, 37212, USA

[2]Chemical and Physical Biology, Vanderbilt University, Nashville, TN, 37212, USA

[3]Biochemistry, Vanderbilt University, Nashville, TN, 37212, USA

[4]Biomedical Informatics and Biostatistics, Vanderbilt University, Nashville, TN, 37212, USA

[5]Pediatrics and Genetics, Stanford University, Stanford, CA, 94305, USA

Correspondence
Vito Quaranta, Department of Biochemistry, Vanderbilt University, Nashville, TN, 37212, USA
Email: vito.quaranta@vanderbilt.edu

Funding information
NCI U54-CA217450 and U01-CA215845 to VQ; R01 CA201513 to JS; NSF fellowship DGE-1445197 to SFM; J.S.L. received a scholarship from the Agency for Science, Technology and Research (A*STAR) in Singapore.

Small-cell lung cancer (SCLC) is characterized by rapid progression, early metastasis, and poor standard-of-care outcomes. Recent investigations have uncovered SCLC heterogeneity, but have not resulted in alternative effective treatments. Adopting a systems approach, we identify gene expression programs and biological ontologies that define four SCLC subtypes encompassing known subtypes and a previously undescribed neuroendocrine variant (NEv2). Using gene signatures and tumor deconvolution, we find all human and mouse tumors analyzed are a composite of subtypes. This functional intratumoral heterogeneity may promote emergence of a rich microenvironment and adaptive response to treatment. Indeed, we show the four subtypes exhibit differential drug sensitivity, with NEv2 being consistently least sensitive. To illuminate sources of phenotypic heterogeneity, we infer a network of transcription factors and develop BooleaBayes, a minimally-constrained Boolean rule-fitting approach. Using probabilistic simulation, we identify master regulators and destabilizers of network attractors by *in silico* perturbations. Specific to NEv2,

* Equally contributing authors.

BooleaBayes predicts ELF3 and NR0B1 as master regulators, and TCF3 as master destabilizer. In summary, our systems approach should lead to actionable therapeutic strategies that consider SCLC intratumoral heterogeneity.

1 | INTRODUCTION

Small cell lung cancer (SCLC) makes up ~15% of lung cancer cases and accounts for over 20,000 deaths every year in the United States [1]. SCLC is characterized by early metastasis, high genomic instability, and an ability to develop resistance to all known treatments [2]. Though 60-80 % of SCLC patients initially respond to the standard of care (a chemotherapeutic regimen of etoposide and a platinum-based agent such as cisplatin), they inevitably relapse with recalcitrant disease, and less than 5% survive five years past initial diagnosis. Despite these poor outcomes, the standard of care for SCLC has not changed in decades. This is partially due to the fact that, though SCLC has obligate inactivation of *TP53* and *RB1* [2], the search for oncogenic driver mutations in SCLC has been largely futile. For example, anti-vascular endothelial growth factor targeted therapies showed no significant benefit over standard chemotherapy [3], and therapies targeting Poly (ADP-ribose) polymerase enzymes, an important enzyme for DNA damage repair, have been met with unclear results [4, 5, 6]. Clinical trials of therapies ranging from inhibition of BCL2 to Aurora Kinases (AURK) to Smoothened likewise showed little promise [6, 7, 8]. Immune checkpoint inhibitors (*e.g.*, anti-PD-L1 therapies, [9]) and antibody-drug conjugates (*e.g.*, Rova-T) are currently undergoing clinical trials [10]. However, immunotherapy can be associated with high toxicities, and therefore generally only help a small portion of patients. We propose that a new approach to SCLC treatment that incorporates a systems-level understanding of the disease may be necessary for successful therapies.

In order to understand the marked recalcitrance of SCLC, efforts to stratify patients have led to the recognition of phenotypic heterogeneity within and between SCLC tumors. SCLC subtypes exhibit unique behaviors, and while SCLC heterogeneity represents a promising avenue for development of targeted therapies, the categorization of SCLC subtypes is still in early days. 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 [11, 12, 13]. However, further classification able to explain all subpopulations of SCLC has been limited.

SCLC tumors are thought to arise from pulmonary neuroendocrine epithelial cells (PNECs) in the airway epithelium of the lung [14], though recent studies suggest additional possible cells of origin for SCLC tumors [15]. In both human and mouse tumors, most cells appear to belong to the NE subtype, corresponding to a PNEC origin, with high expression of neuroendocrine genes such as ASCL1. However, several groups have found evidence for non-NE variants within SCLC tumors. Lim et al. found a population of cells in both mouse and human samples that were positive for Hes1, a transcriptional repressor of neuroendocrine differentiation (vis-a-vis ASCL1 inhibition) and marker for Notch activity [16]. Calbo et al. found an alternative non-NE subpopulation that was CD44+ and ASCL1- and supports NE metastasis, which they termed "mesenchymal-like" [17]. A MYC-amplified, NEUROD1+ SCLC variant has also been reported both in mouse and human [18, 19, 20]. We described SCLC cell lines with hybrid expression of both NE and non-NE markers [21], and proposed they could serve as a resistant niche since drug perturbations shifted most cell lines towards hybrid phenotype(s). Huang et al. recently found that a small subset (~15%) of SCLC cell lines and tumors may be driven by POU2F3, a characteristic gene of the tuft cell lineage [15], leading to speculations of multiple cells of origin for SCLC. 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 [22].

Here, we set out to develop a systems-level approach to identify SCLC subtypes and their sources from global gene expression. Using consensus clustering and weighted gene co-expression network analysis on transcriptomics data, we identify four SCLC subtypes distinguished by several criteria, including gene expression signatures and biological ontologies. Three of these subtypes correspond to reported ones. The fourth one is a previously unreported NE variant (termed NEv2) with reduced sensitivity to drugs. We computationally derive a transcription factor (TF) network regulating this SCLC phenotypic heterogeneity [21], and characterize its four stable states by a combination of experimental and computational tools, including single-cell RNA sequencing, gene expression analysis, and Boolean modeling. 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. Thus, given the limitations of using immunohistochemistry in SCLC, we propose that defining SCLC subtypes by gene signatures should enable a comprehensive identification of SCLC tumor heterogeneity, with profound translational implications. That is, it should be possible to delineate subtype content of an SCLC tumor from transcriptomics data, which are now accessible in the clinic [**?**], coupled with computational analyses to sort out subtype proportions (e.g., CIBERSORT).

2 | RESULTS

2.1 | Consensus Clustering uncovers new SCLC variant phenotype

We previously reported that the two broad SCLC subtypes (NE and non-NE) correspond to stable states (*e.g.*, attractors) of an SCLC transcription factor (TF) regulatory network [21]. However, we and others have demonstrated the occurrence of additional variant subtypes [18, 20, 21]. To investigate whether attractors corresponding to such variant subtype(s) exist, we applied Consensus Clustering [23] to RNA-seq gene expression data from the 50 SCLC cell lines in the Cancer Cell Line Encyclopedia (CCLE) [24]. We clustered the cell lines using a k-means method with a Pearson distance metric for $k \in \{2, 20\}$ (Fig 1A). Both k=2 and k=4 gave well defined clusters with high cluster consensus, a measure of how tight and distinct the clusters are. While k=2 clusters had slightly higher cluster consensus, as mentioned above, our previous analyses and recent literature suggest that more than two subtypes are necessary to adequately describe SCLC phenotypic heterogeneity. Beyond 2, k=4 had the next best clustering consensus, and therefore we further investigated these four clusters.

We considered expression of well-studied markers of SCLC heterogeneity in the context of the 4-cluster phenotypic classification (Fig 1B). Three of the four consensus clusters can be matched to phenotypes previously identified: the canonical NE subtype, an NE variant subtype (referred to here as NEv1), and a non-NE variant subtype [16, 20, 19]. Interestingly, the non-NE subtype fell closest to non-SCLC cell lines from the CCLE dataset, rather than other SCLC cell lines (Figure 1A), suggesting a departure of non-NE from the classical NE phenotype. Each of the three previously-identified subtypes has been independently defined by expression of a few markers; However, considering only a few markers at a time has limitations. Indeed, the fourth previously unrecognized phenotypic cluster (referred to here as NEv2) could not be easily resolved using only a few markers. For example, NEv2 may be considered a tumor propagating cell (TPC) by Jachan et al; however expression of *HES1* may suggest that NEv2 cell lines should be grouped into a non-NE subtype by Lim et al [16]. Clearly, a more global approach is needed to comprehensively distinguish SCLC phenotypes.

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

To this end, we compared the activity of global gene expression programs across the four subtypes. Weighted gene co-expression network analysis (WGCNA) [25] on the RNA-seq data from the CCLE uncovered 17 modules of co-expressed genes. We computed module eigengenes, which capture trends of gene expression levels in each module. Based on comparison of eigengene values, 11 out of 17 modules statistically distinguished between the four consensus clusters (Fig 2A, Table EV1 Kruskal-Wallis, FDR-adjusted $p < 0.05$). To identify biological processes enriched in each of these 11 gene modules, we performed gene ontology (GO) enrichment analysis using the Consensus Path Database [26], which resulted in a combined total of 1,763 statistically enriched biological processes (Table EV2). To visualize these terms in an organized "phenospace" (Fig 2B), we used the GOSemSim package [27] in R to compute a pairwise dissimilarity score between all enriched GO terms (FDR-adjusted $p < 0.05$ in at least one of the 11 significant modules). Using this dissimilarity matrix as a distance metric, we projected all significant GO terms into a 2D space by t-distributed stochastic neighbor embedding (t-SNE) [28]. 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 biological process (such as neuronal development, adhesion, and signal transduction). This map shows that SCLC heterogeneity spans biological processes that can largely be grouped as related to neuronal, endocrine, or epithelial differentiation; metabolism and catabolism; cell-cell adhesion and mobility; and response to environmental stimuli, including immune and inflammatory responses.

SCLC phenotypes co-exist within this phenospace, and can be distinguished by differential module expression. In particular, the yellow, salmon, and pink modules are enriched for neuroendocrine differentiation and neurotransmitter secretion and are upregulated in the canonical NE and NEv1 phenotypes (Fig 2C). 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.

Genes within the brown, midnight blue, and green modules are upregulated in the NEv2 phenotype. 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 2C). 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.

In summary, a phenospace constructed from global gene expression analyses captures the unique characteristics of the heterogeneous SCLC phenotypes, encompassing well-known features such as cell adhesion vs. epithelial differentiation, as well as uncovering new actionable biology.

2.3 | Drug resistance is a feature of the NEv2 subtype

As mentioned above, the enrichment of drug catabolism and xenobiotic metabolism GO terms within the green module suggests that the NEv2 phenotype may have a higher ability to metabolize drugs and therefore exhibit decreased sensitivity. To test this possibility, we reanalyzed, in the context of our 4 subtype classification, drug responses of SCLC cell lines to a panel of 103 FDA-approved oncology agents and 423 investigational agents [29]. We used as a summary measure of the resultant dose-response curves the activity area (AA) (Supplemental methods). The drugs were analyzed individually, by common mechanism of action, and by target type, and the cell lines were grouped by the 4 subtypes. Across all the drugs, NEv2 was the most resistant subtype 54% of the time, while both NE and NEv1 were most resistant

each 20% of the time, and non-NE 6% (Fig 3A). These results corroborate the prediction from the GO term enrichment analysis that treatment resistance is a feature of the NEv2 subtype (Figure 2C).

We then performed ANOVA to identify drugs for which differential response across subtypes were statistically significant, which gave 123 significant drugs (Bonferroni-corrected p-value < 0.05, Table EV3). In particular, mTOR inhibitors are a class of compounds to which NEv2 was significantly more resistant (Fig 3). PI3K pathway mutations have previously been implicated as oncogenic targets for SCLC, as about a third of patients show genetic alterations in this pathway [30]. Among the four subtypes, NEv2 is also the least sensitive to AURKA, B, and C inhibitors (AURKA shown); TOPO2 inhibitors; and HSP90 inhibitors (Figure 3). These results have implications for interpreting expected or observed treatment response with respect to tumor heterogeneity in individual patients (see below in Section 2.7 and Discussion).

2.4 | Genetic mutations alone cannot account for four SCLC phenotypes

Differential drug sensitivity is often thought to be caused by oncogenic mutations. Therefore, we analyzed genomic data in the Broad Cancer Dependency Map [31] to determine whether mutations could be responsible for defining the four SCLC subtypes. We subsetted these data to the 50 SCLC cell lines with matching CCLE RNA-seq data, and using MutSigCV [32], we found 29 genes (Fig EV2) 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 (Fig EV2), either by mutational status alone or type of mutation, suggesting alternative sources of heterogeneity.

2.5 | Transcription factor network defines SCLC phenotypic heterogeneity and reveals master regulators

Along these lines, we previously identified a TF network that explained NE and non-NE SCLC subtype heterogeneity [21]. As mentioned above, that analysis suggested the existence of additional SCLC subtypes but did not specify corresponding attractors. 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 2B) 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 [21], as detailed in Fig 4A.

Using customary coarse-grained Boolean modeling approaches, such as threshold updates or inhibitory dominant rules, the updated network still did not stabilize attractors corresponding to NEv1 or NEv2 SCLC phenotypes (data not shown). Nevertheless, more complex Boolean regulatory patterns are possible, and likely to be more relevant in biological complexity. To this end, we developed BooleaBayes, a method to infer logical relationships between TF regulators (Fig 4B) by enhancing confidence in Boolean rules via a Bayes-like adjustment approach (Supplemental methods). With BooleaBayes, we were able to describe the stabilization of all four SCLC phenotypes from steady-state gene expression data (Fig 4C). An advantage of this method is that it makes predictions intrinsic to those parts of the network in which we are most confident (Supplemental Methods).

BooleaBayes rules (probabilistic logical relationships between connected nodes) were derived for each node of the SCLC TF network in Fig 4A. As an example, Figure 4B shows the rule fitting for one node, ASCL1. Rules for all other nodes are given in Supplemental Figure 4. We simulated the dynamics of the Boolean network using a general-asynchronous

update scheme [33]. This formed a state transition graph (STG), in which each state is defined by a vector of TF ON/OFF expression values.

For the initial states in the simulation, we discretized 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. 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 4C), that we refer to as pseudo-attractors.

These 10 pseudo-attractor states each correlated with, and could be assigned to, one of the 4 SCLC subtypes (stars in Fig.4C); 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.

2.6 | *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 4C, we performed random walks (algorithm described in Supplemental Methods) starting from each of the 10 pseudo-attractors. We counted how many steps were required to reach a state outside of a 4-TF neighborhood around the starting state (Figure 5). We chose a 4-TF neighborhood for a given attractor because a randomly chosen state has only a 2.5% chance of falling within this neighborhood, suggesting states in this neighborhood are significantly similar to the attractor state ($p = 0.025$). Each TF in the network was either activated (held constant at TF=1) or silenced (TF=0) in each of the stable states, and 1000 random walks were executed for each condition (Fig 5A). The percent increase or decrease of the stability relative to the unperturbed reference was calculated, resulting in a score quantifying (de)stabilization of the starting state by each TF perturbation (Fig 5B, Fig EV3). For example, either activation of GATA4 (TF = 1) or silencing FOXA1 (TF=0) are predicted to destabilize both the NE and NEv2 subtypes (Fig 5C).

TFs that, when silenced, cause destabilization greater than 20% (score $\leq$ -0.2) of a specific subtype were considered master regulators of that subtype. They include *REST* (non-NE) (in agreement with [16], *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 [34]; 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 5C).

2.7 | Neuroendocrine variants are represented in mouse and human SCLC tumors

Next, we investigated whether the four phenotypes we detected in human SCLC cell lines are also present in tumors. SCLC tumors are largely neuroendocrine, and in genetically engineered mouse models, often include a portion of non-neuroendocrine cells that interact with and support the neuroendocrine bulk [16, 17]. The NEv1 subtype has been shown to exist in a Myc-amplified mouse model and to correlate with a human phenotype of SCLC [20]. Nevertheless, there is a need for quantitative understanding of the relative prevalence of each subtype in a single tumor, in order to

calibrate treatment with respect to subtype differential drug sensitivity (Fig 3).

To fulfill this need, we used CIBERSORT [35] to generate gene signatures for each of the 4 subtypes, and deconvolved RNA-seq measurements on 81 SCLC tumors from George et al [2]. CIBERSORT predicted that a majority of tumors were comprised of all four subtype signatures, in varying proportions across tumor samples (Fig 6A). Interestingly, the 81 tumors fell into four main groups: 1) the first (top group in Fig 6A) is mostly comprised of NE; 2) the second is mostly NEv2; 3) the third is mostly non-NE with a significant NEv2 portion of each tumor; and 4) the fourth is mostly non-NE (though the high proportion of non-NE in all of these tumors may be due to infiltrating non-tumor cells; see below).

Next, we analyzed the patient/cell-derived xenograft models (PDXs/CDXs) developed by Drapkin et al [36] (Fig 6B). These tumors showed vast differences across the samples; some samples were primarily NEv1, some were mostly non-NE, and others contained varying proportions of NE and NEv2 (Fig 6B). The advantage of using this data is temporal: because some PDXs/CDXs were taken from the same patient at different times throughout treatment, we could analyze changes in tumor composition over time. For example, three samples taken from patient MGH1514, before and after treatment, indicate a change in tumor composition in favor of the NE phenotype. In contrast, patient MGH1518 shows a reduction of NEv1 and increase in NEv2 after treatment. Overall, the high variance in proportions of each subtype suggest a high degree of intertumoral, as well as intratumoral, heterogeneity.

We also investigated phenotypic patterns in mouse tumors from two different sources ([20], Supplemental Methods) to determine whether human SCLC subtype signatures are conserved across species. The first mouse model is a triple knockout (*RB1*, *TP53*, and *p130* conditionally deleted in lung cells via a Cre-Lox system, TKO) generated by the Sage laboratory [37]. The conditional knockout of all three floxed sites is initiated by intratracheal administration of an adenovirus-delivered Cre recombinase, driven by either a CMV or a CGRP promoter. Whether driven by CMV or CGRP, tumors were primarily composed of the NE and NEv2 subtypes (Fig 7Ai). A slight decrease in the NEv2 in tumors from metastatic sites was also observed. Of note is the lower percentage of non-NE cells found in each tumor in Fig 7Ai; we suspect this is due to a filtering step before sequencing (Supplemental Methods), as the non-NE subtype signature is more similar to that of other stromal cells such as tumor-associated immune cells which may increase the non-NE signal in unfiltered populations. The second mouse model was generated in the Oliver laboratory by *Myc* amplification (double knockout of RB1 and TP53, and amplification of MYC) (Fig 7Aii) [20]. Using the subtype gene signatures we provided, the *Myc*-amplified tumors showed a clear increase in the percentage of NEv1 detected compared to the triple knockout tumors in Fig 7Ai. This corroborates the correlation between our human NEv1 subtype and the previously described *Myc*-amplified mouse tumor subtype. In particular, both contain the *NeuroD1* gene, not expressed in SCLC mouse tumors without *Myc* amplification.

Lastly, we tested whether an orthogonal approach (independent of CIBERSORT) would also detect the four subtypes in human and mice tumors. We therefore analyzed two primary TKO mouse tumors by single cell RNA-seq (scRNA-seq)(Supplementary Methods). For each mouse single cell transcriptome, we computed the k=10 nearest *human cell line* neighbors (kNN with k=10, Supplemental Methods). If at least 8 of the 10 neighbors were one subtype, the mouse cell was assigned to that subtype (otherwise, not assigned), and the results were visualized with t-SNE. As shown in Fig 7B, 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 7B). Tumor A has a large proportion of the NEv2 subtype, corresponding to the tumors in 7Ai. In contrast, tumor B has a large NEv1 subpopulation, similar to the tumors in 7Aii.

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.

3 | DISCUSSION

We report that integrating bioinformatics with mathematical modeling of global gene expression data across mouse and human tumors enables efficient discrimination of SCLC into subtypes. As an initial step, independent clustering and WGCNA analyses of transcriptomics data separate human SCLC cell lines into four subtypes: NE, NEv1, NEv2, and non-NE. Transcription factors regulating each WGCNA module are then used to build a gene regulatory network (GRN) defining SCLC cell identity. Minimally constrained Boolean simulations with BooleaBayes, a new method that identifies regulatory rules for each node in the network, predicted GRN stable states that matched the empirically-observed subtypes. *In silico* network perturbations led to detection of master transcription factor (TF) regulators and destabilizers. Using CIBERSORT and scRNA-seq, we provide evidence that these subtypes can be detected in human SCLC tumor specimens, as well as in tumors from genetically engineered SCLC mouse models (GEMMs). Furthermore, GO enrichment analysis allowed us to identify defining characteristics of each of these subtypes, and analysis of drug response showed the subtypes exhibit differential drug sensitivity across a wide array of treatments.

An original advance of general value in this work is the introduction of BooleaBayes, a novel method for identifying master TFs of phenotypic heterogeneity. 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. *In silico* activation and silencing experiments suggest strategies for destabilizing each subtype. 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. By quantitatively ranking TFs based on their destabilizing capability for a given subtype, our method gives a systematic way to evaluate a large number of potential perturbations to find those that may effectively reprogram a resistant phenotype. We plan to validate these predictions with experiments in which we perturb gene expression levels and evaluate shifts in phenotype. We propose that our master TF identification method should be applicable to any biological systems in which it is desirable to probe the source of phenotypic heterogeneity at the gene expression level, and particularly for the construction of master TF networks. Furthermore, our method for predicting perturbation strategies that redirect cells towards a drug-sensitive phenotype can be applied to any cancer or transcriptionally-regulated disease.

We have positioned four subtypes (NE, NEv1, NEv2, and non-NE) within the context of the broader literature on SCLC heterogeneity, and conclude that one (NEv2) has not been described previously. However, tumor deconvolution by CIBERSORT and analysis of scRNA-seq indicate that this subtype is present in a large proportion of human and mouse SCLC tumors. Moreover, a key finding of our gene program analysis is that the NEv2 subtype is enriched in drug and xenobiotic metabolism/catabolism genes, suggesting it may be better equipped for adaptation to stress and/or perturbation, enhancing treatment resistance. Indeed, a drug screen across a broad range of compounds indicated that this subtype is broadly more resistant than the others, especially in response to *AURK* and *mTOR* inhibitors. Direct experimental verification in mouse and human tumors of the NEv2 subtype's drug resistant properties will be an important next step. For example, the PDXs from patient MGH1518 before and after drug show an increase in the NEv2 signature (Fig 6B), which 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 firm up this conclusion.

Previous markers of SCLC heterogeneity were unable to differentiate between the NE and NEv2 subtypes. Both express canonical neuroendocrine markers, such as ASCL1, CGRP, and DDC, which supports a neuroendocrine phenotype. However, unlike the canonical NE subtype, NEv2 expresses the Notch-pathway target HES1, in the absence of expression of Notch family receptors. This finding confirms the distinction between NEv2 and canonical NE phenotypes,

and suggests Notch-independent mechanisms for the activation of HES1.

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. [38] 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 [39]; however, our clusters do not correlate with location of the tumor sample from which each cell line was derived (*e.g.*, primary vs metastatic, data not shown).

Lastly, Huang et al. recently reported a new SCLC subtype defined by expression of *POU2F3* [15], a TF implicated in the differentiation of sensory tuft cells across diverse tissues O'Leary 2018. It is not clear why this subtype was not detected by our method. WGCNA placed POU2F3 into the Green-yellow module along with SOX9 and ASCL2, both of which were associated with the POU2F3 phenotype by Huang et al., and with COX1, a well-known marker of tuft cell differentiation. The consensus clustering algorithm used here may miss this signal due to the small size of the Green-yellow module (only 328 genes). However, preliminary results indicate a larger k may recover that signal and classify the four POU2F3-high cell lines into a single subtype, as a 10-cluster scheme, for example, resolves this subtype. Future studies with additional cell line and/or mouse data may be used to further investigate this question, underscoring that the delineation of four subtypes here does not preclude the existence of others.

Enrichment of an immune-enriched module (Green) in NEv2 suggests that this subtype benefits tumor growth in the presence of pro-inflammatory cytokines. Of note is the neuropeptide Calcitonin-gene-related-peptide (CGRP), encoded by CALCA, a gene significantly expressed in NEv2. CGRP has previously been implicated in modulation of inflammatory responses in the lung [40, 41]. Upon challenge with antigen, PNECs release CGRP to influence immune cells chemotactically, or to promote hyper-proliferation (potentially as a mechanism of wound healing), and we speculate that NEv2 cells may inherit this capability.

The CGRP gene promoter has been used in the construction of SCLC GEMMs, as it is thought to target PNECs specifically [16]. CGRP is a commonly used marker for identifying neuroendocrine cells, but we have shown in single cell mouse data that its expression is specific to the NEv2 subtype (Section 2.7). This suggests that CGRP, while highly and specifically expressed in PNECs, exhibits heterogeneous expression across SCLC subtypes. Therefore, CGRP may be a useful biomarker for identifying the NEv2 subtype in tumors.

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 (Figure 1B). 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.

4 | SUPPLEMENTAL 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 [2]. Myc-amplified mouse data set [20] was obtained from the NCBI GEO deposited at GSE89660. PDX/CDX mouse data [36] was obtained from the NCBI GEO deposited at GSE110853. Data 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 [42]. 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 $k \in \{2, 12\}$ (Fig EV1). 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 (Fig EV1, tracking plot, delta area plot, and average consensus scores (not shown).

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 [27], 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 [29]. 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 [29], "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 (Lubbock et al., manuscript in preparation). For fitting the dose response curve, as mentioned in the Thunor manual (https://docs.thunor.net/): "Thunor fits viability data to a three parameter log-logistic model. The three parameters are E_{max}, EC_{50}, and Hill coefficient. A constrained fit is used: E_{max} is constrained to be between 0 and 1, and the Hill coefficient (slope) is constrained to be non-negative." A measure

of activity area for each drug-cell line combination was calculated as "the area above the observed response values, up to the no response value (1 on the y-axis). Between doses, a straight line extrapolation is used (on a log10 dose x-axis). Response values above 1 are truncated at 1. If there are multiple response values for a particular concentration (replicates), the mean average is used." This takes into account both the efficacy and potency of a drug when added to each cell line. By segregating the cell lines by subtype, we were able to evaluate the relationship between drug response and subtype. Further information is available at https://docs.thunor.net/.

GENOMIC ANALYSIS

Mutational Analysis was performed by MutSigCV V1.2 from the Broad Institute [32]. First, a dataset of merged mutation calls (including coding region, germ-line filtered) from the Broad Cancer Dependency Map [31] 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, which are shown in Fig EV2.

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. Therefore we adopt a probabilistic framework, in which each component of the vector is a continuous real-value $v_i \in [0,1]$ reflecting our confidence in the output of F, based on the 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 observations 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 probabilistic vector $\vec{V}$ describing a probabilistic Boolean function F of N variables. First, we organize the input-output relationship as a binary decision 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 decision tree correspond to the probabilities that $(\overline{A} \wedge \overline{B})$, $(\overline{A} \wedge B)$, $(A \wedge \overline{B})$, and $(A \wedge B)$. Collectively, the observations define an $M \times N$ matrix $\mathbf{R} = \left[\vec{R}_1, \vec{R}_2, \ldots, \vec{R}_N\right]$ 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 target, 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 $\vec{T}'$ 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 ith observation belong to the jth leaf of the binary decision tree. For instance using the example above, the second leaf of the binary tree is $(\overline{A} \wedge B)$. 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 ith observation constrains the jth component of $\vec{V}$. Additionally, to avoid overfitting under-determined leaves, we define the uncertainty $\vec{U} = [u_1, u_2, \ldots, u_{2N}]$ 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 **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 $t'_2 = 1$.

CIBERSORT

CIBERSORT is a computational inference tool developed by Newman et al. at Stanford University [35]. 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 = .3, and 50-150 barcode genes considered when building the signature matrix.

SINGLE CELL RNA SEQUENCING OF TKO SCLC TUMORS

The p53, Rb 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 (Jahchan et al., 2016, Lim et al., 2017). 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 [43]. 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).

ACKNOWLEDGEMENTS

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.

REFERENCES

[1] Society AC. Cancer Facts and Figures 2018 2018;.

[2] 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.

[3] Tiseo M, Boni L, Ambrosio F, Camerini A, Baldini E, Cinieri S, et al. Italian, Multicenter, Phase III, Randomized Study of Cisplatin Plus Etoposide With or Without Bevacizumab as First-Line Treatment in Extensive-Disease Small-Cell Lung Cancer: The GOIRC-AIFA FARM6PMFJM Trial. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2017;35(12):1281–1287.

[4] Owonikoko TK, Dahlberg SE, Sica G, Wagner LI, Wade JL, Srkalovic G, et al. Randomized trial of cisplatin and etoposide in combination with veliparib or placebo for extensive stage small cell lung cancer: ECOG-ACRIN 2511 study. Journal of Clinical Oncology 2017;35:8505–8505.

[5] Pietanza MC, Waqar SN, Krug LM, Dowlati A, Hann CL, Chiappori A, et al. Randomized, Double-Blind, Phase II Study of Temozolomide in Combination With Either Veliparib or Placebo in Patients With Relapsed-Sensitive or Refractory Small-Cell Lung Cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2018;36(23):2386–2394.

[6] Gadgeel SM. Targeted Therapy and Immune Therapy for Small Cell Lung Cancer. Options in Oncol 2018;19(53).

[7] Gately K, Collins I, Forde L, Al-Alao B, Young V, Gerg M, et al. A Role for IGF-1Râ€"Targeted Therapies in Small-Cell Lung Cancer? 2011;.

[8] Alvarado-Luna G, Morales-Espinosa D. Treatment for small cell lung cancer, where are we now?—a review. Translational Lung Cancer Research 2016;.

[9] Horn L, Mansfield AS, Szczęsna A, Havel L, Krzakowski M, Hochmair MJ, et al. First-Line Atezolizumab plus Chemotherapy in Extensive-Stage Small-Cell Lung Cancer. New England Journal of Medicine 2018;379(23):2220–2229.

[10] Hellmann MD, Ott PA, Zugazagoitia J, Ready NE, Hann CL, De Braud FG, et al. Nivolumab (nivo) ± ipilimumab (ipi) in advanced small-cell lung cancer (SCLC): First report of a randomized expansion cohort from CheckMate 032. Journal of Clinical Oncology 2017;35:8503–8503.

[11] 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 Properties1; 1985.

[12] 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 Features1; 1985.

[13] Gazdar AF, Bunn PA, Minna JD. Small-cell lung cancer: what we know, what we need to know and the path forward. Nature Publishing Group 2017;17. https://www.nature.com/articles/nrc.2017.87.pdf.

[14] 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;19:754–764.

[15] 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;32:915–928.

[16] Lim JS, Sage J. Intratumoural heterogeneity generated by Notch signalling promotes small-cell lung cancer. Nature Letters 2017;545:360–364.

[17] Calbo J, Montfort EV, 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;19:244–256.

[18] 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. PNAS 2012;109(42):17034–17039.

[19] Borromeo MD, Savage TK, Kollipara RK, Gazdar AF, Cobb MH, Johnson Correspondence JE, et al. ASCL1 and NEUROD1 Reveal Heterogeneity in Pulmonary Neuroendocrine Tumors and Regulate Distinct Genetic Programs. Cell Reports 2016;16.

[20] Mollaoglu G, Guthrie MR, Behm S, Brugelmann 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;31(2):270–285.

[21] Udyavar A, Wooten D. Novel Hybrid Phenotype Revealed in Small Cell Lung Cancer by a Transcription Factor Network Model That Can Explain Tumor Heterogeneity 2017;77(5).

[22] Huch M, Rawlins E. Tumours build their niche. Nature News and Views 2017;.

[23] 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;52:91–118.

[24] Barretina a. The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. Nature 2012;483(7391):603–607.

[25] Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 2008;9.

[26] Kamburov A, Stelzl U, Lehrach H, Herwig R. The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Research 2013;41.

[27] 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;26(7):976–978.

[28] Van Der Maaten L, Hinton G. Visualizing Data using t-SNE. Journal of Machine Learning Research 2008;9:2579–2605.

[29] 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;108(10).

[30] Umemura S, Mimaki SM, Makinoshima H, Tada SM, 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;9:1324–1331.

[31] Consortium CCLE, of Drug Sensitivity in Cancer Consortium G. Pharmacogenomic Agreement between Two Cancer Cell Line Data Sets. Nature 2015;528(7580):84––87.

[32] 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;499:214–218.

[33] Albert I, Thakar J, Li S, Zhang R, Albert R. Boolean network simulations for life scientists. Source Code for Biology and Medicine 2008;3.

[34] 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.

[35] Newman A, Liu C, Green M, Gentles A, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature Methods 2015;12(5):453–457.

[36] 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 Research 2018;.

[37] Schaffer BE, Park KS, Yiu G, Conklin JF, Lin C, Burkhart DL, et al. Loss of p130 Accelerates Tumor Development in a Mouse Model for Human Small-Cell Lung Carcinoma. Cancer Research;.

[38] Williamson SC, Metcalf RL, Trapani F, Mohan S, Antonello J, Abbott B, et al. Vasculogenic mimicry in small cell lung cancer. Nature Communications 2016;7.

[39] 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;166(2):328–342.

[40] Russell FA, King R, Smillie SJ, Kodji X, Brain SD. Calcitonin Gene-Related Peptide: Physiology and Pathophysiology. Physiological Reviews 2014;94(4):1099–1142.

[41] Gu X, Karp PH, Brody SL, Pierce RA, Welsh MJ, Holtzman MJ, et al. Chemosensory Functions for Pulmonary Neuroendocrine Cells. American Journal of Respiratory Cell and Molecular Biology 2014;50:637–646.

[42] Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 2010;26.

[43] Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species A. nature biotechnology 2018;36(5).

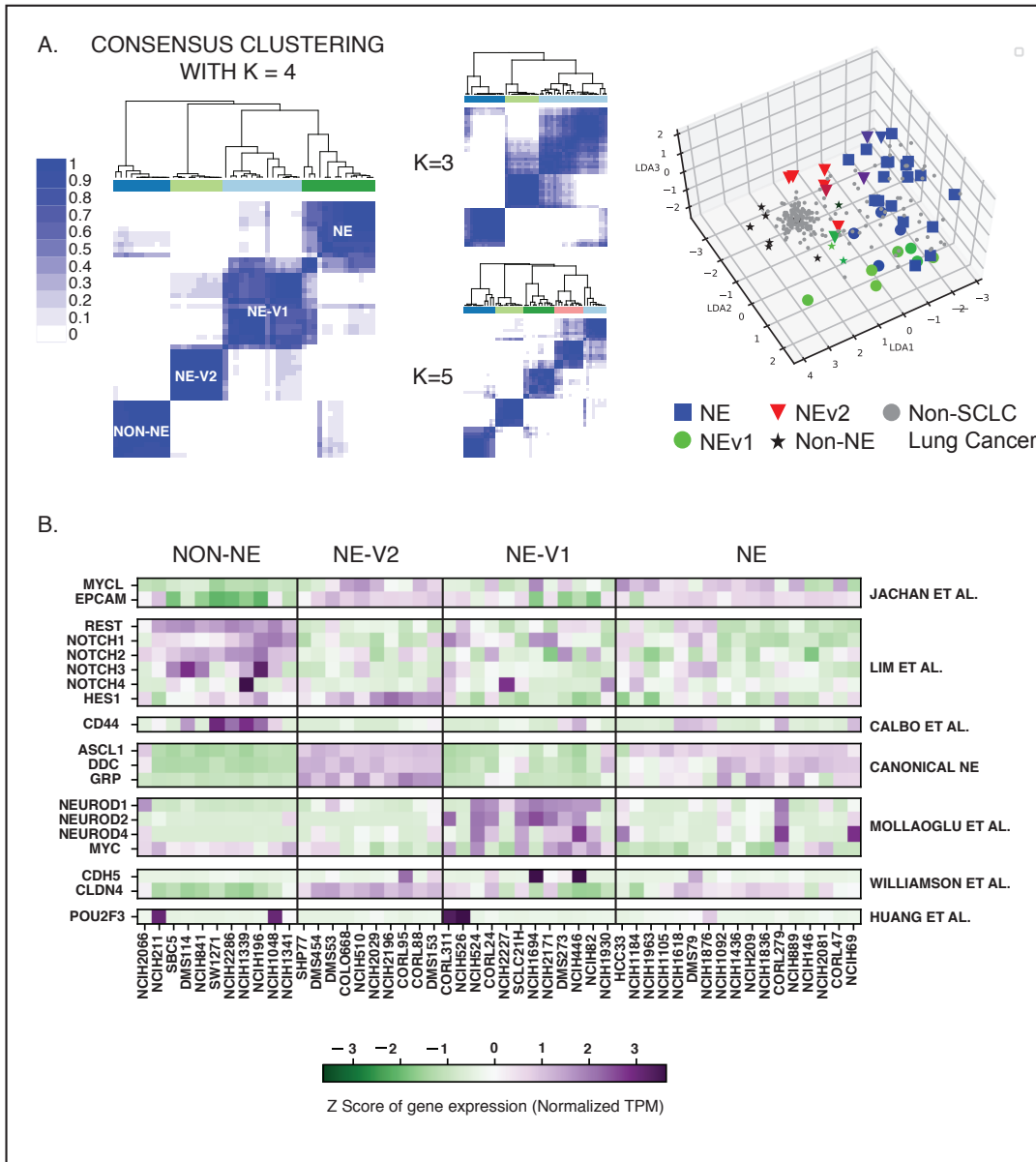


FIGURE 1

CLUSTERING AND WGCNA

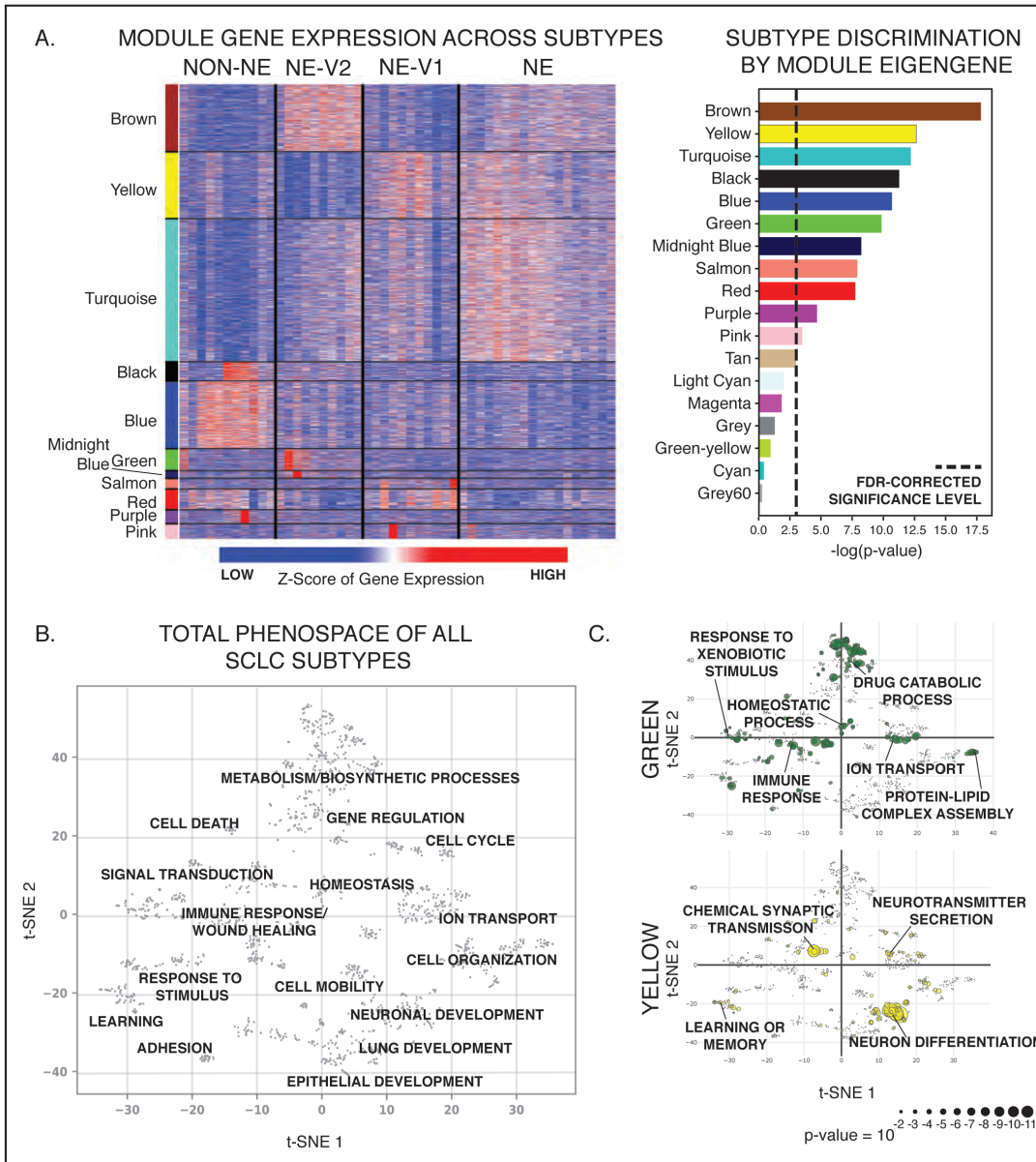


FIGURE 2

GENE ONTOLOGY MAPS

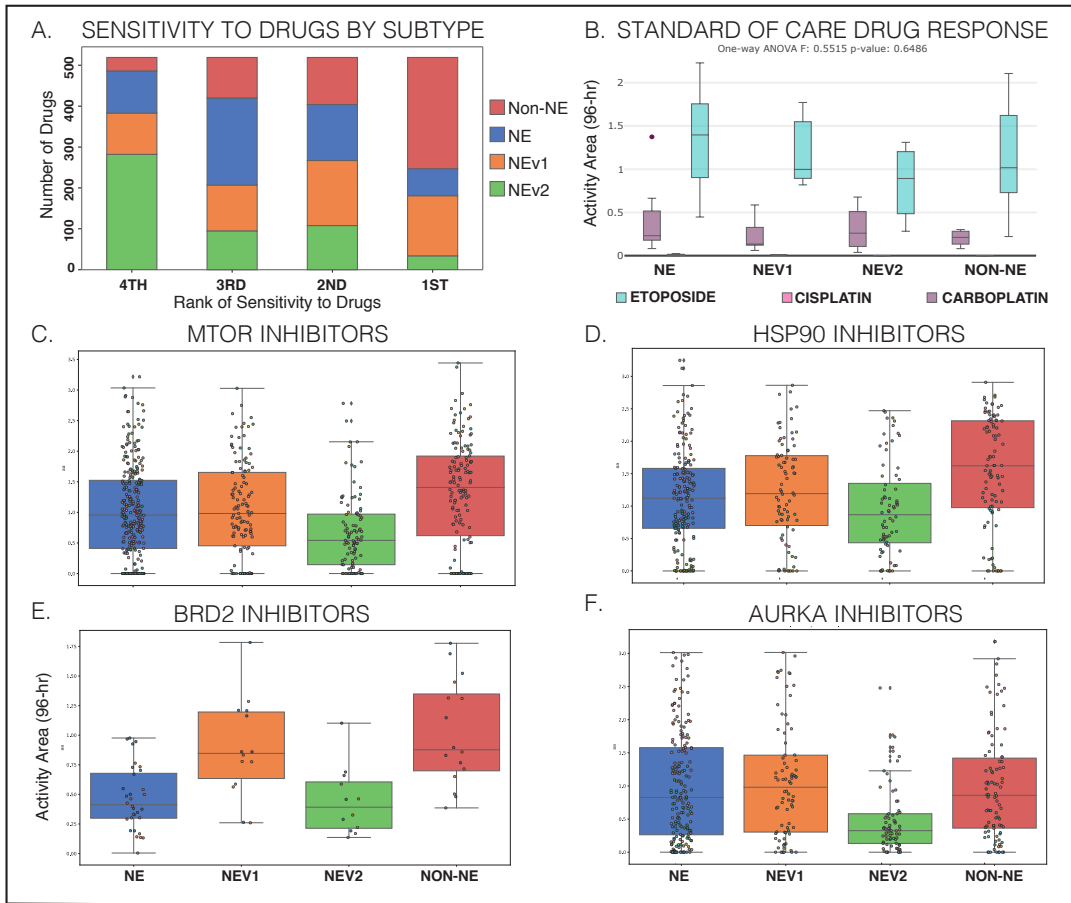


FIGURE 3

Drug Response

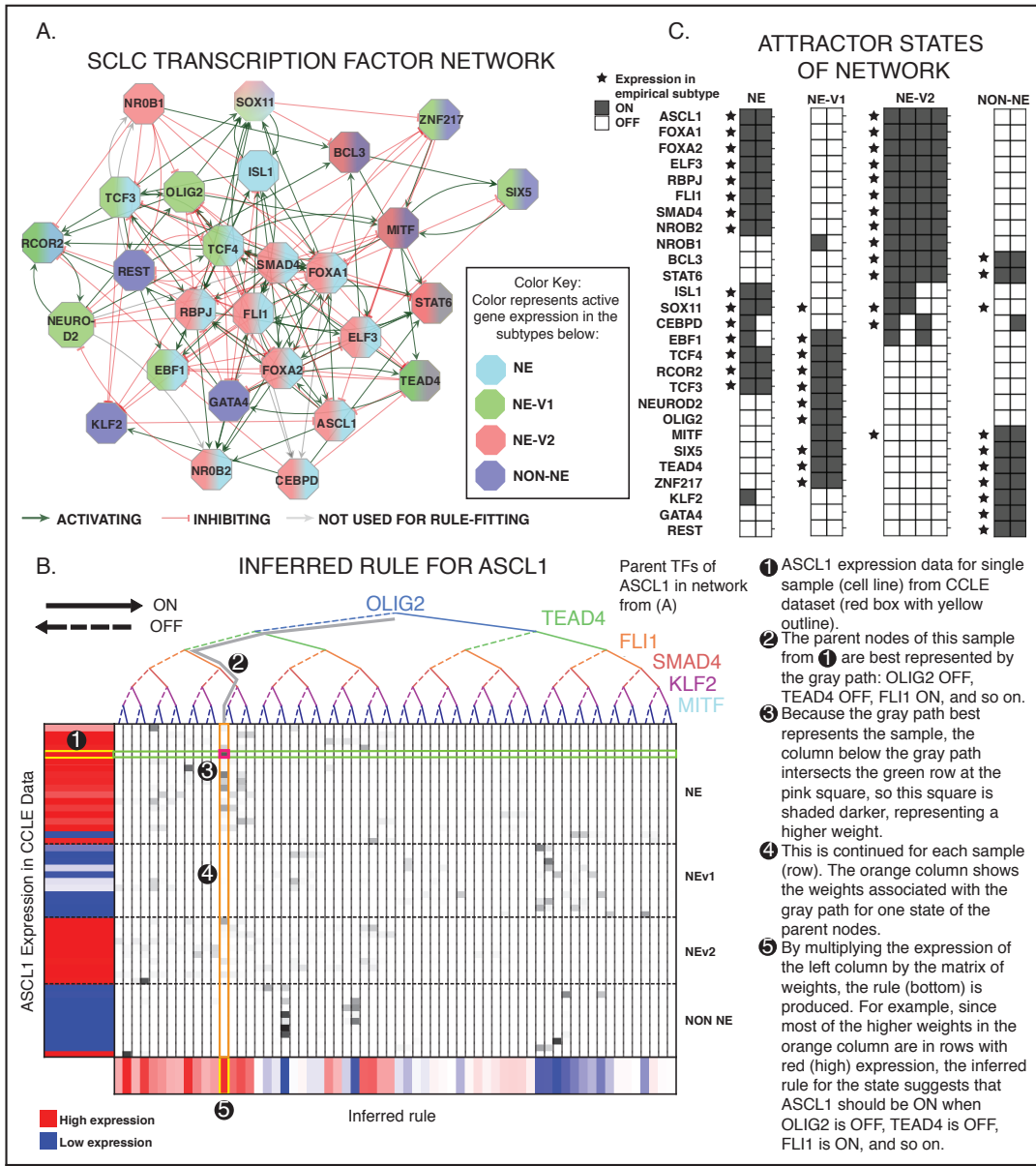


FIGURE 4

NETWORK

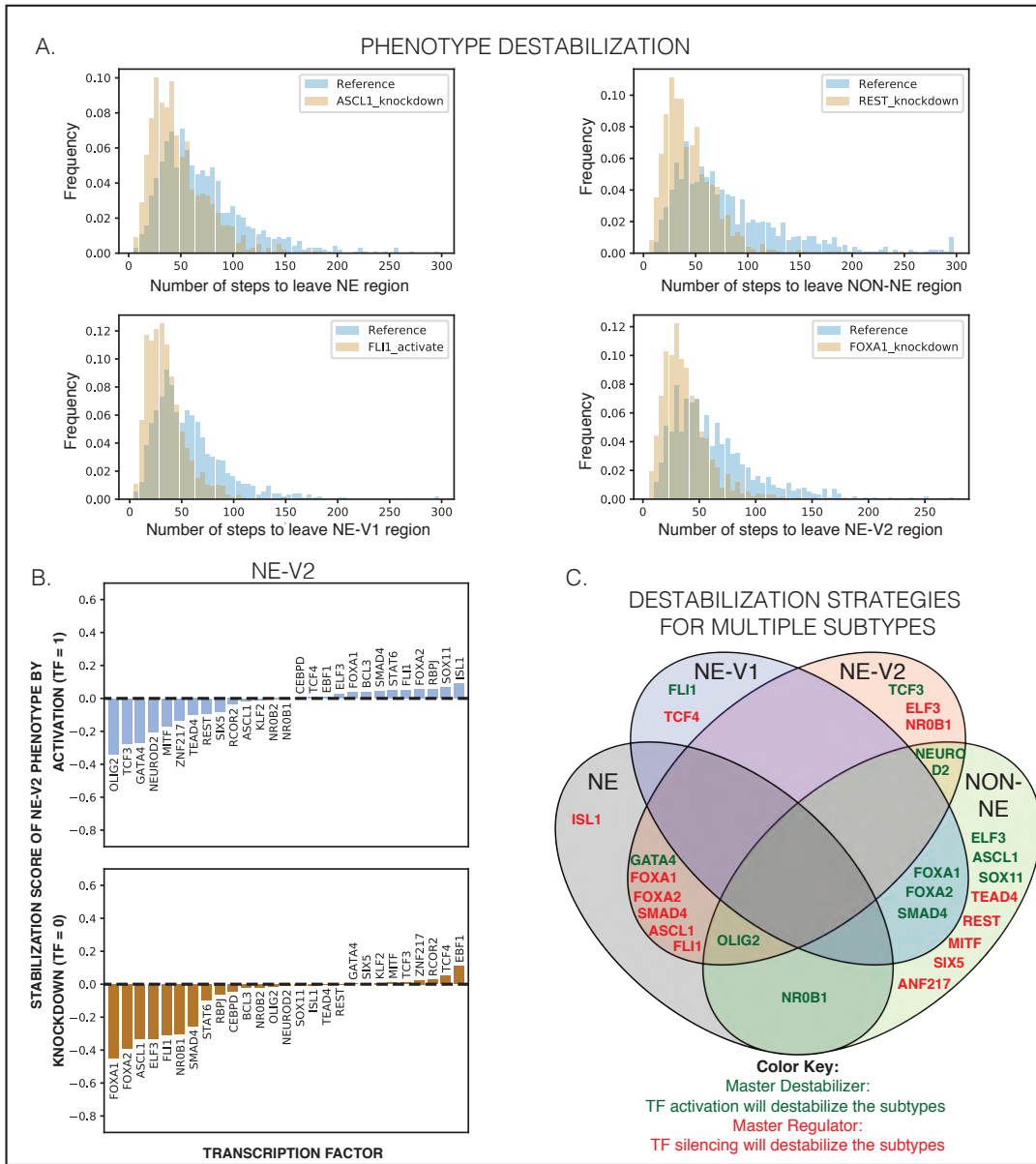


FIGURE 5

REPROGRAMMING

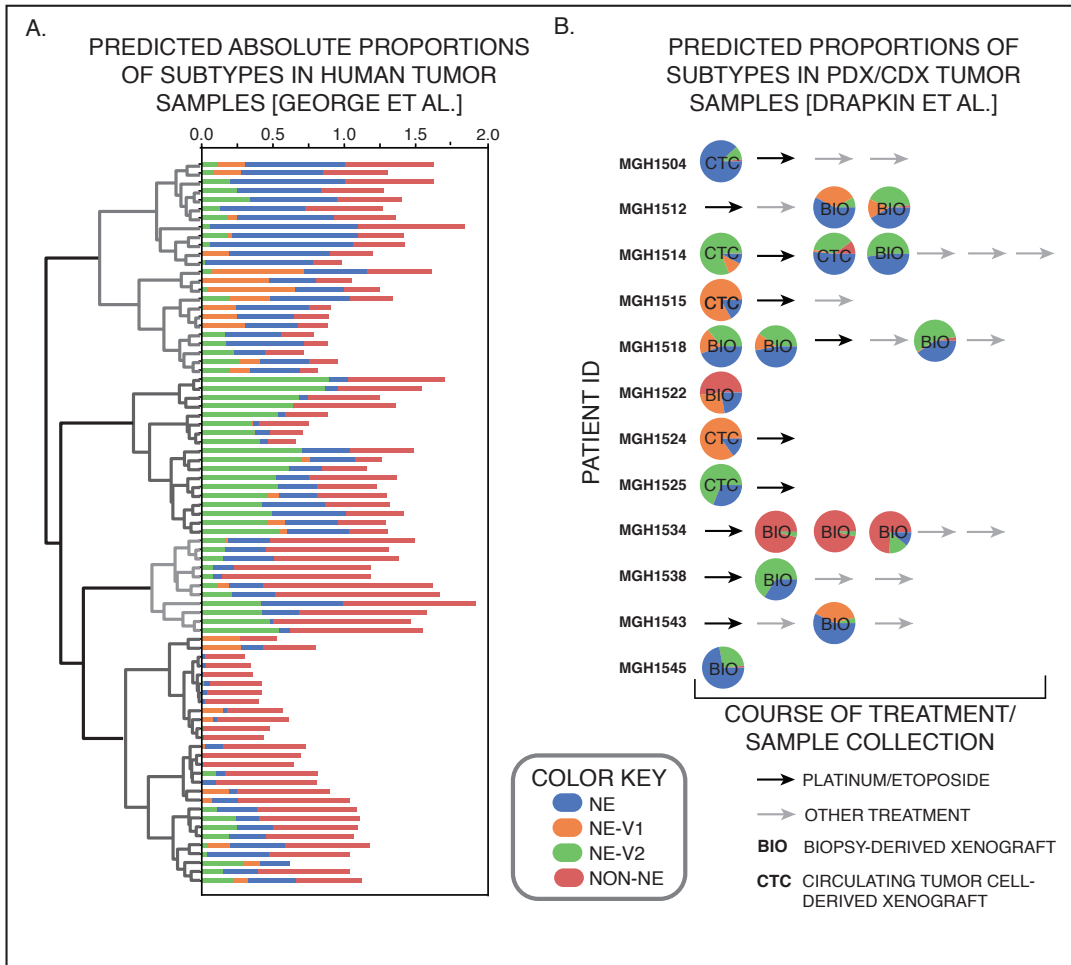


FIGURE 6

CIBERSORT

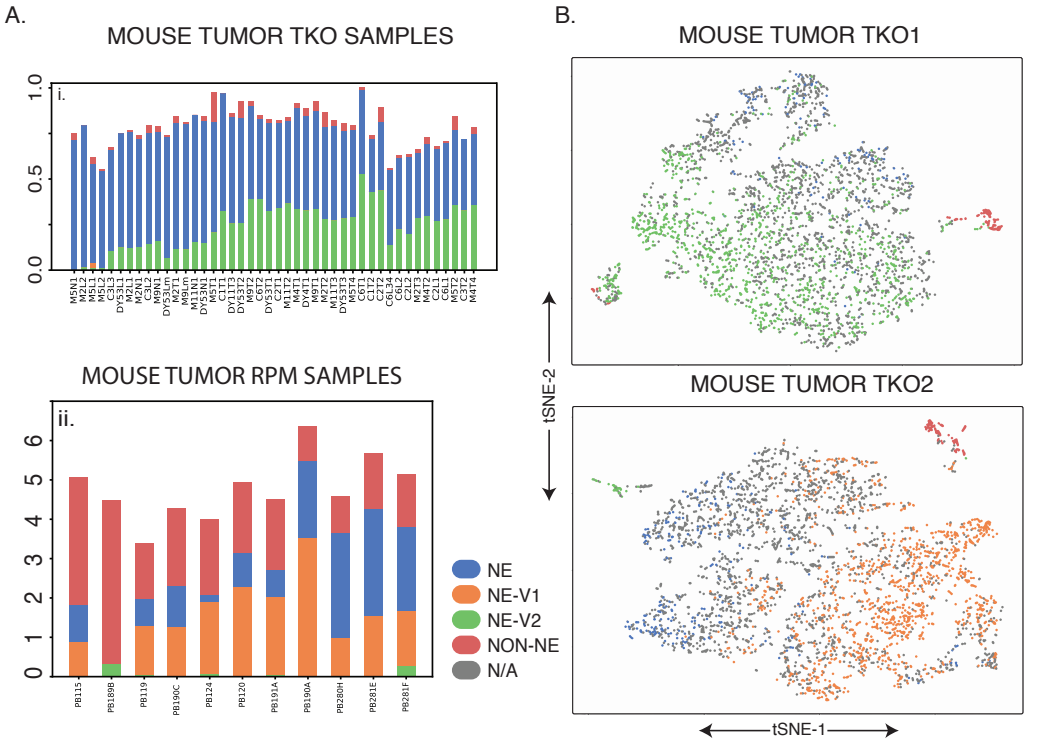
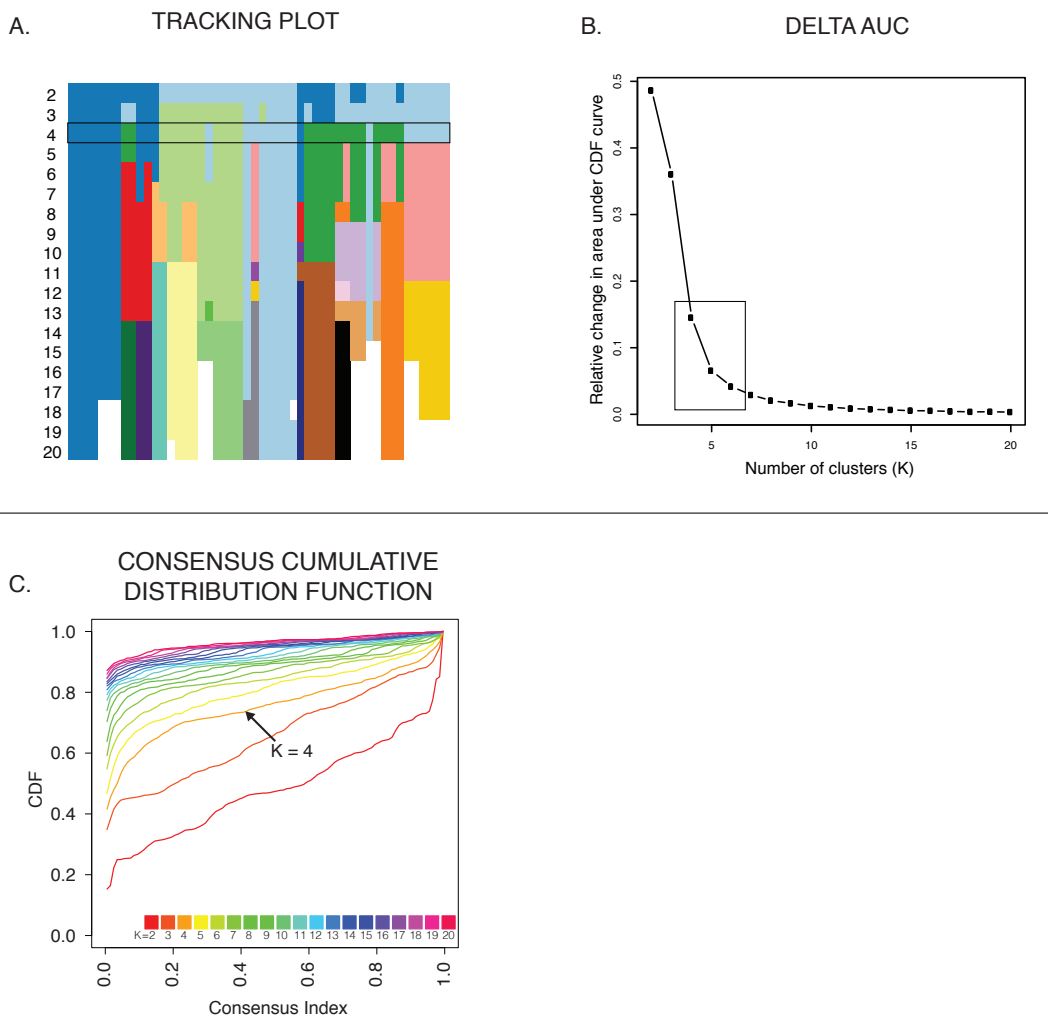
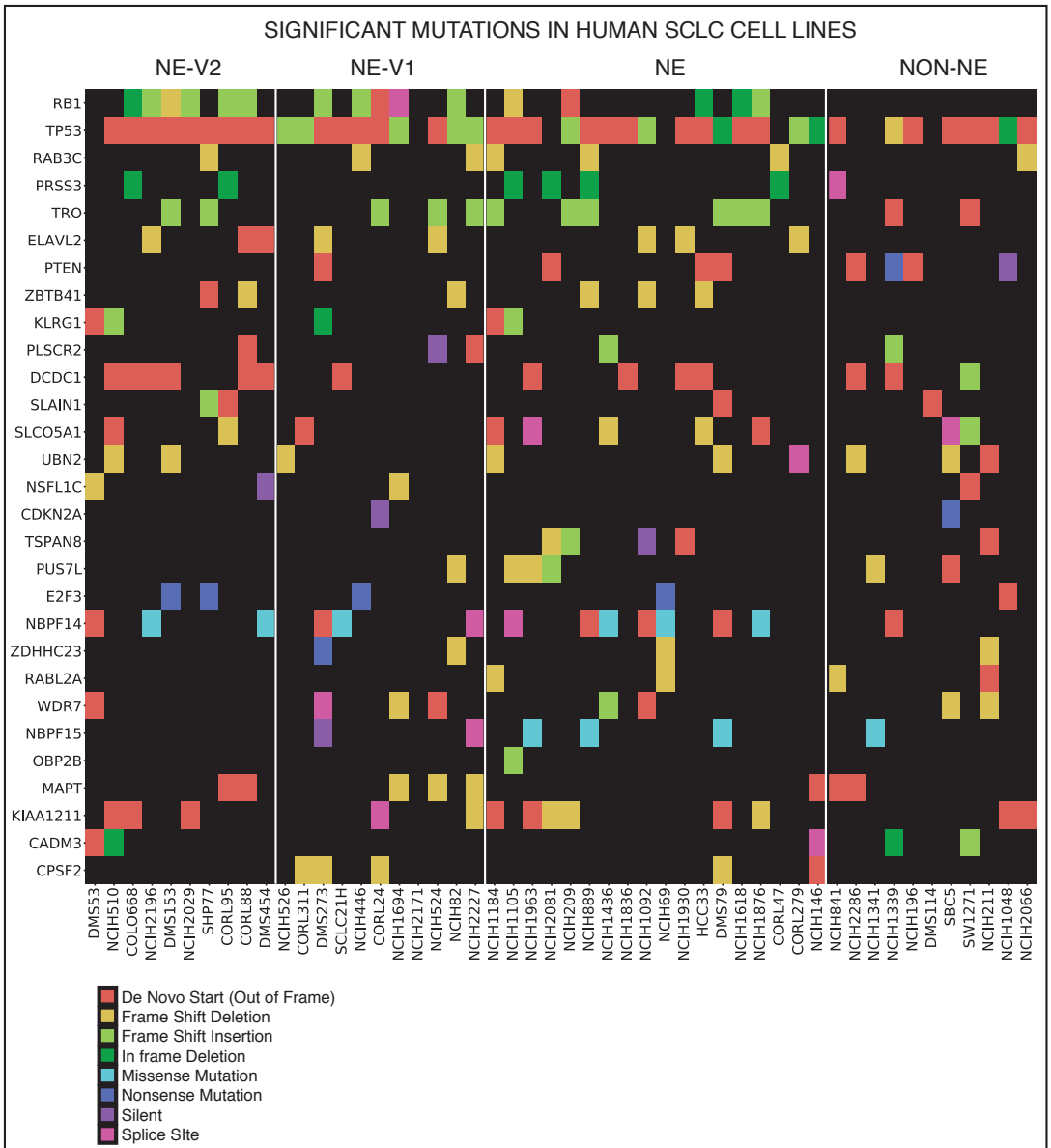

A.
MOUSE TUMOR TKO SAMPLES
i.
1.0
0.5
0.0
MOUSE TUMOR RPM SAMPLES
ii.
0
1
2
3
4
5
6
PB115
PB189B
PB119
PB190C
PB124
PB120
PB191A
PB190A
PB280H
PB281E
PB281F
NE
NE-V1
NE-V2
NON-NE
N/A
B.
MOUSE TUMOR TKO1
MOUSE TUMOR TKO2
tSNE-2
tSNE-1

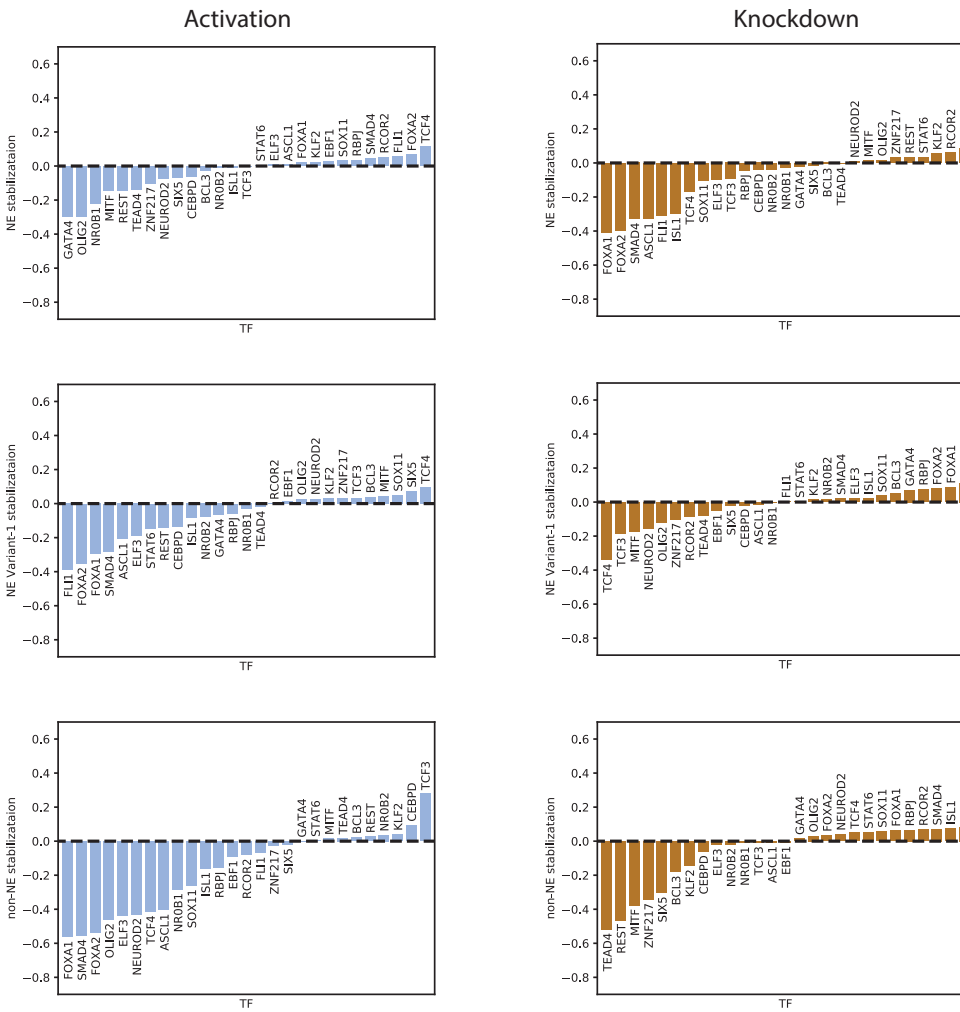
FIGURE 7



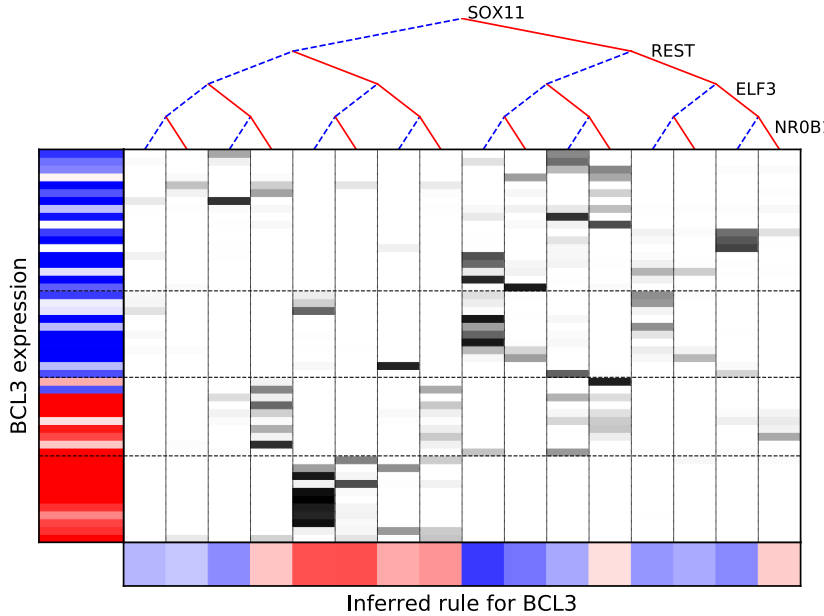
SUPPLEMENTARY FIGURE 1

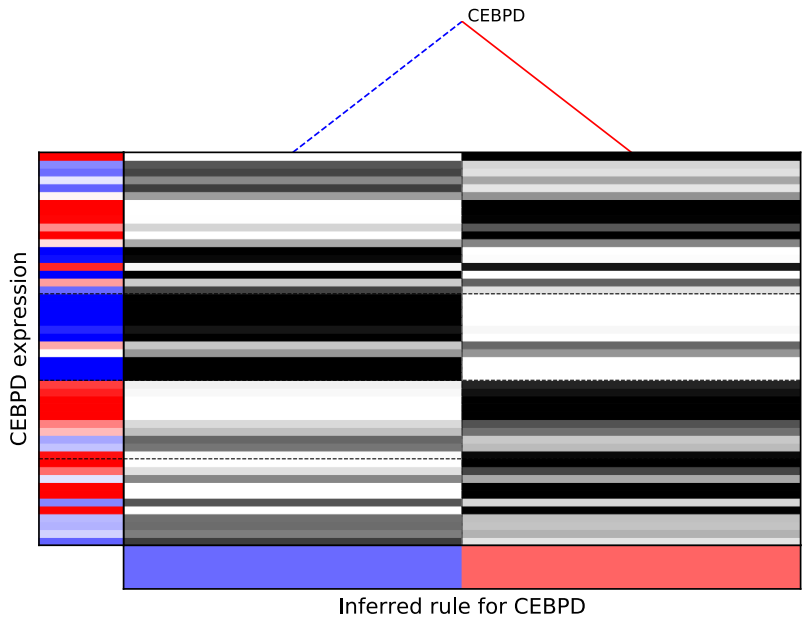


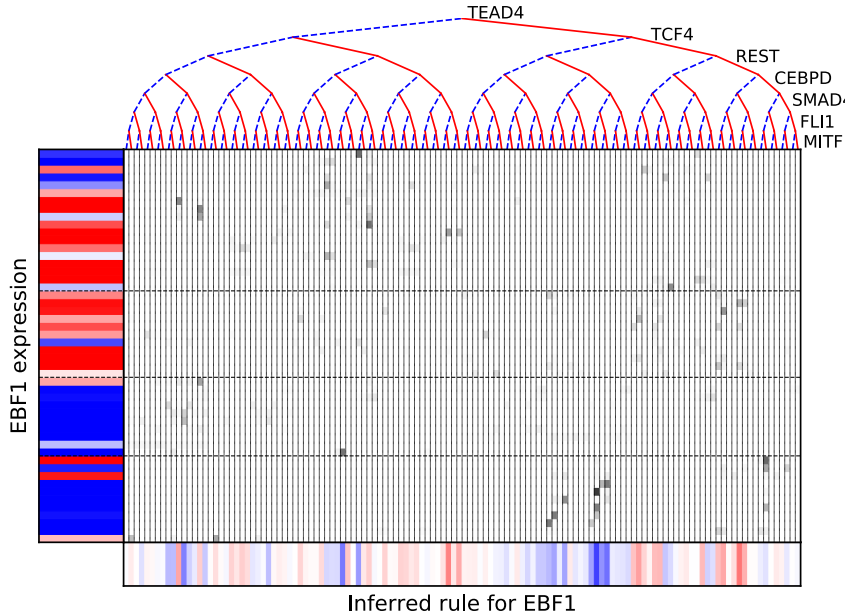
SUPPLEMENTARY FIGURE 2

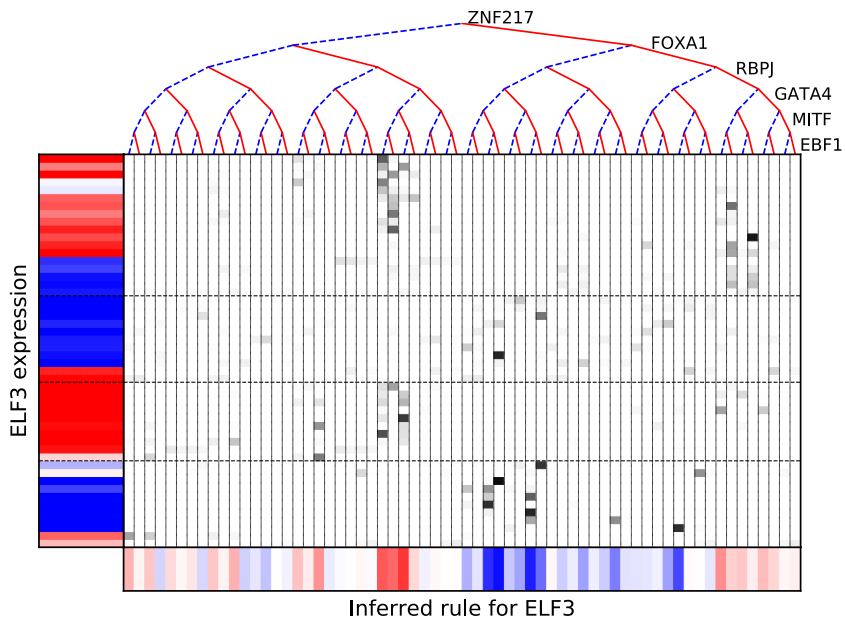


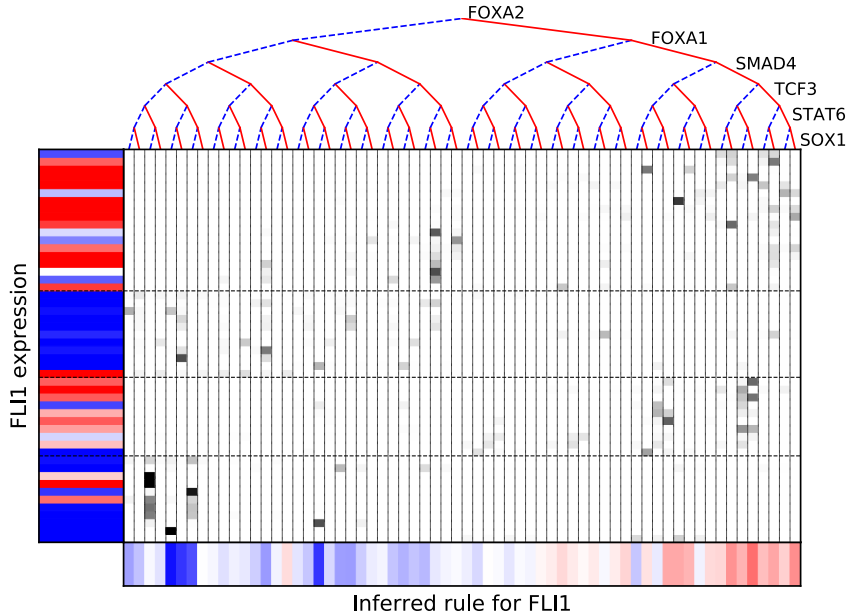
SUPPLEMENTARY FIGURE 3

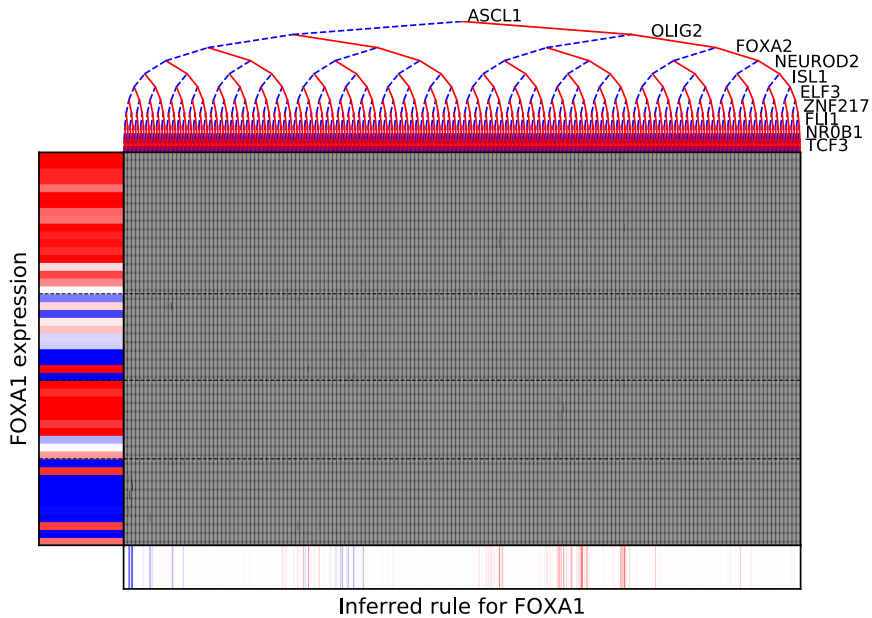

SOX11
REST
ELF3
NR0B1
BCL3 expression
Inferred rule for BCL3

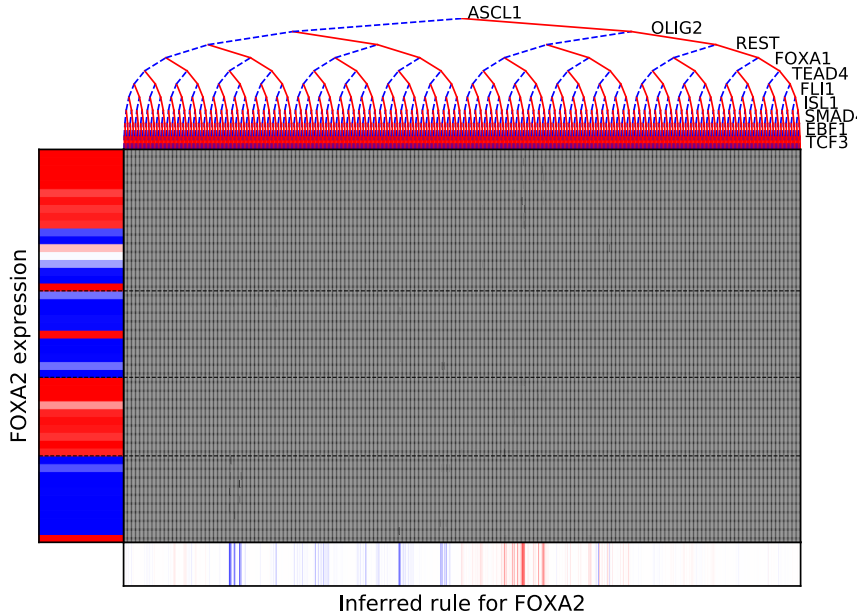

CEBPD
CEBPD expression
Inferred rule for CEBPD

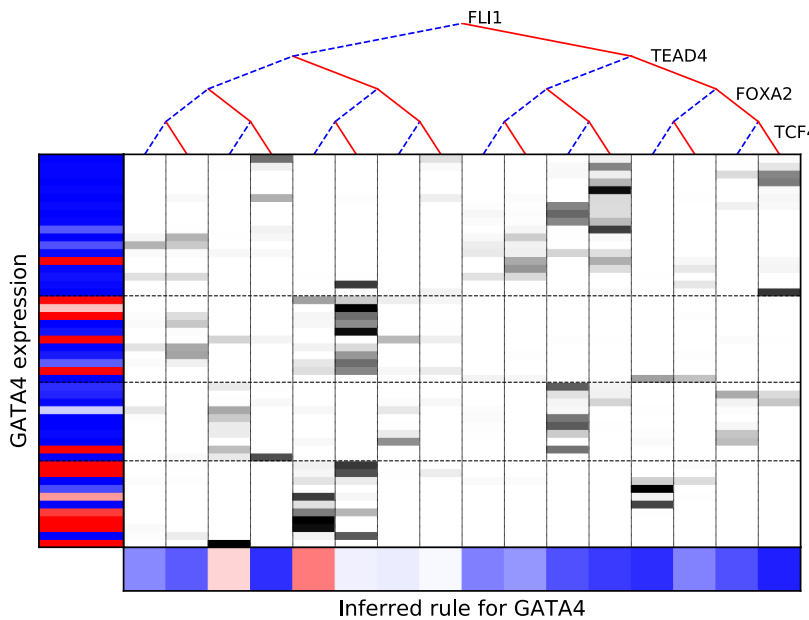

TEAD4
TCF4
REST
CEBPD
SMAD4
FLI1
MITF
EBF1 expression
Inferred rule for EBF1

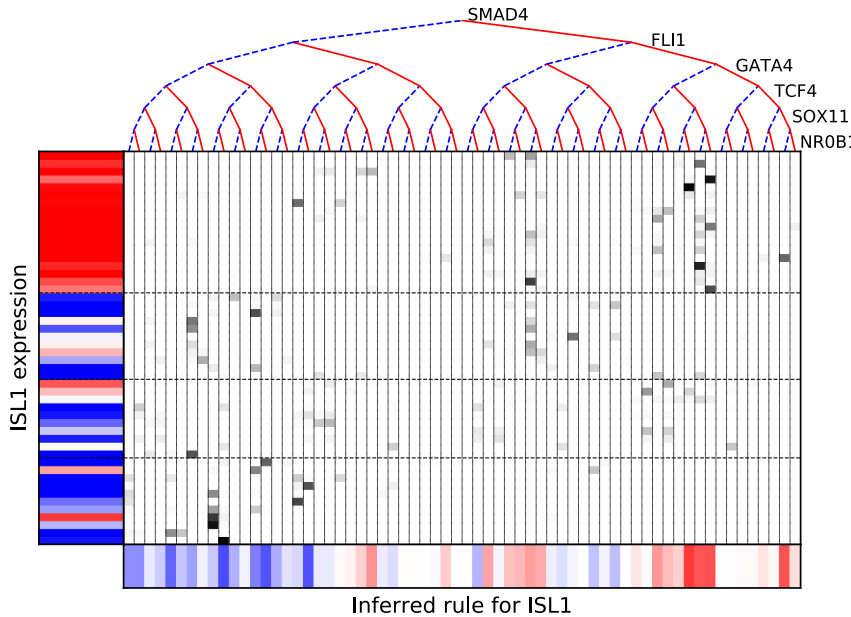

ZNF217
FOXA1
RBPJ
GATA4
MITF
EBF1
ELF3 expression
Inferred rule for ELF3

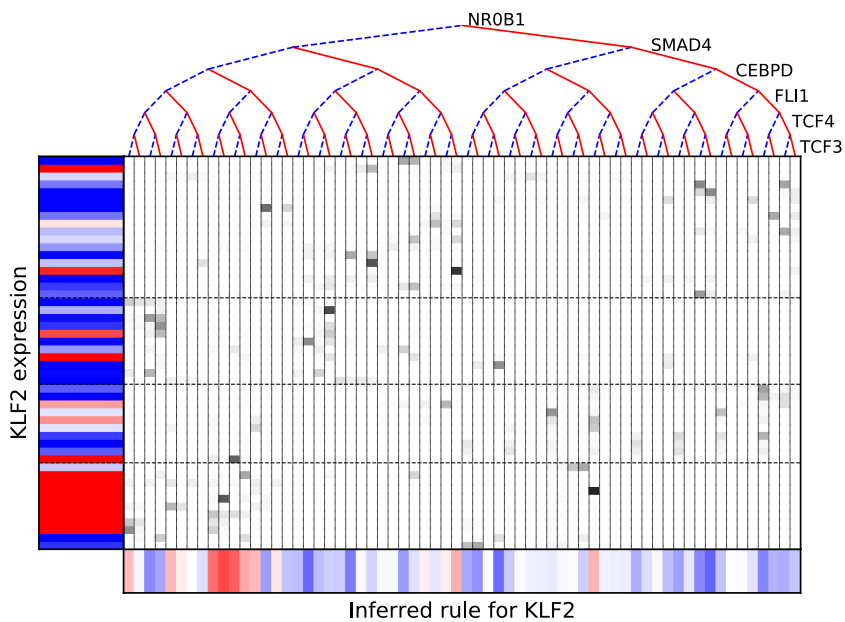

FOXA2
FOXA1
SMAD4
TCF3
STAT6
SOX11
FLI1 expression
Inferred rule for FLI1

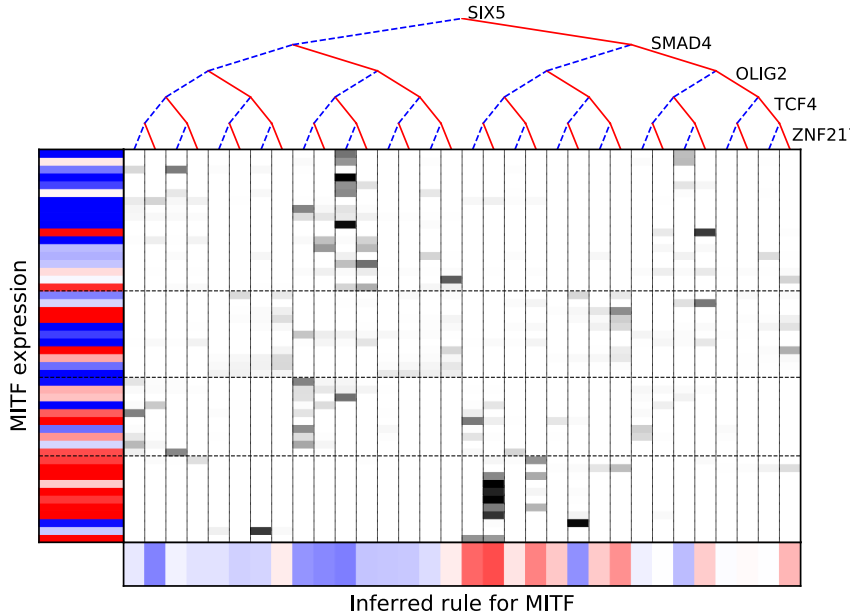

ASCL1
OLIG2
FOXA2
NEUROD2
ISL1
ELF3
ZNF217
FLI1
NR0B1
TCF3
FOXA1 expression
Inferred rule for FOXA1

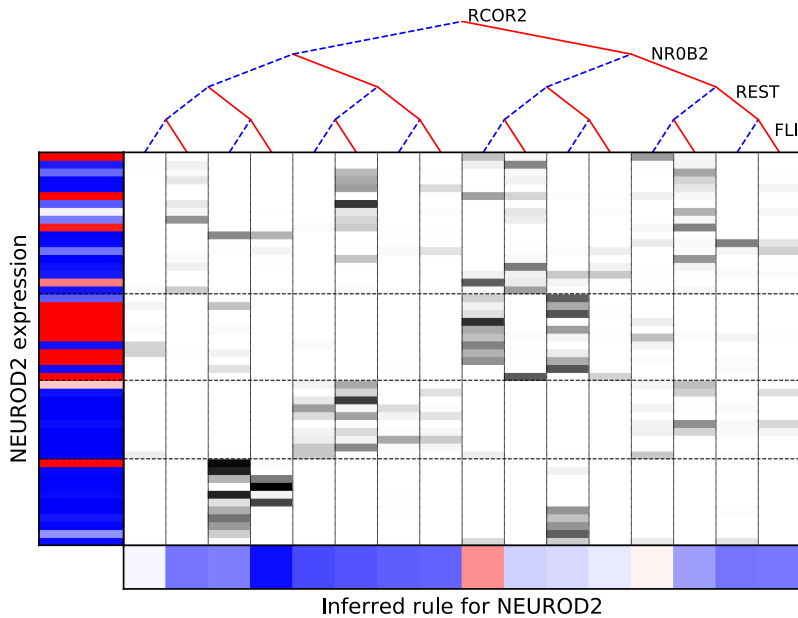

ASCL1
OLIG2
REST
FOXA1
TEAD4
FLI1
ISL1
SMAD4
EBF1
TCF3
FOXA2 expression
Inferred rule for FOXA2

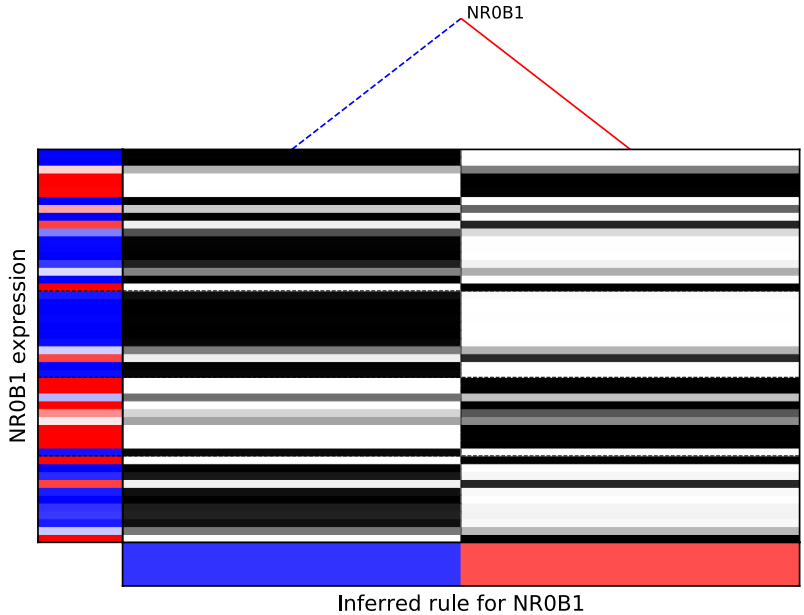

FLI1
TEAD4
FOXA2
TCF4
GATA4 expression
Inferred rule for GATA4

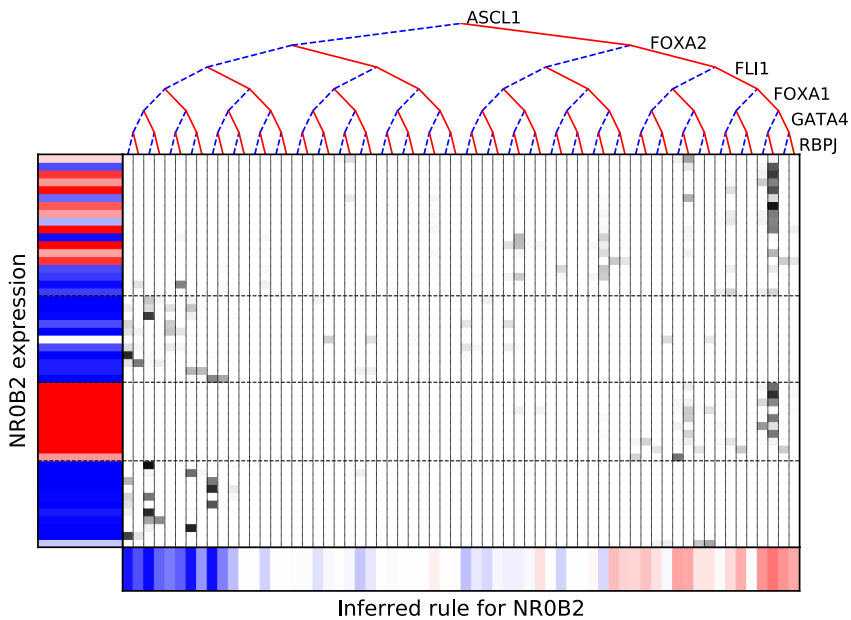

SMAD4
FLI1
GATA4
TCF4
SOX11
NR0B1
ISL1 expression
Inferred rule for ISL1

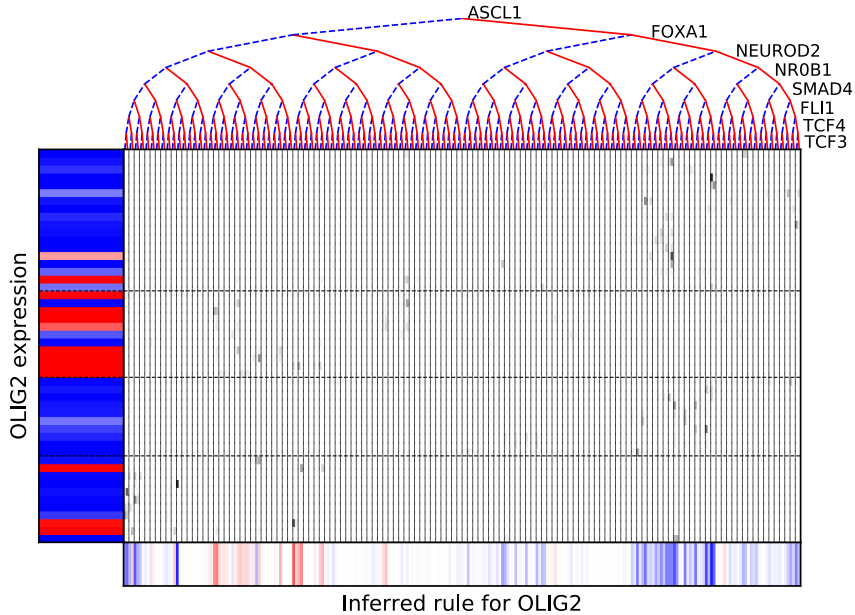

NR0B1
SMAD4
CEBPD
FLI1
TCF4
TCF3
KLF2 expression
Inferred rule for KLF2

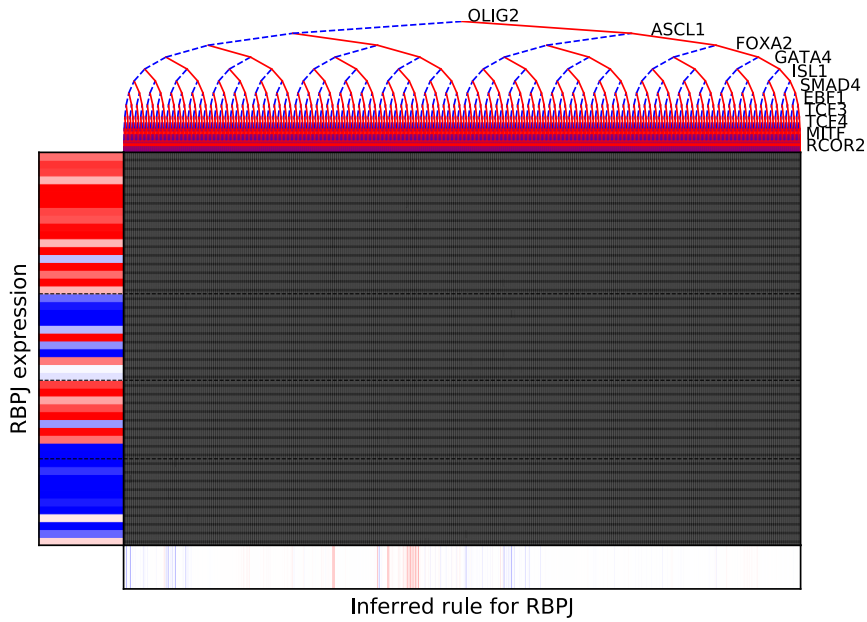

SIX5
SMAD4
OLIG2
TCF4
ZNF217
MITF expression
Inferred rule for MITF

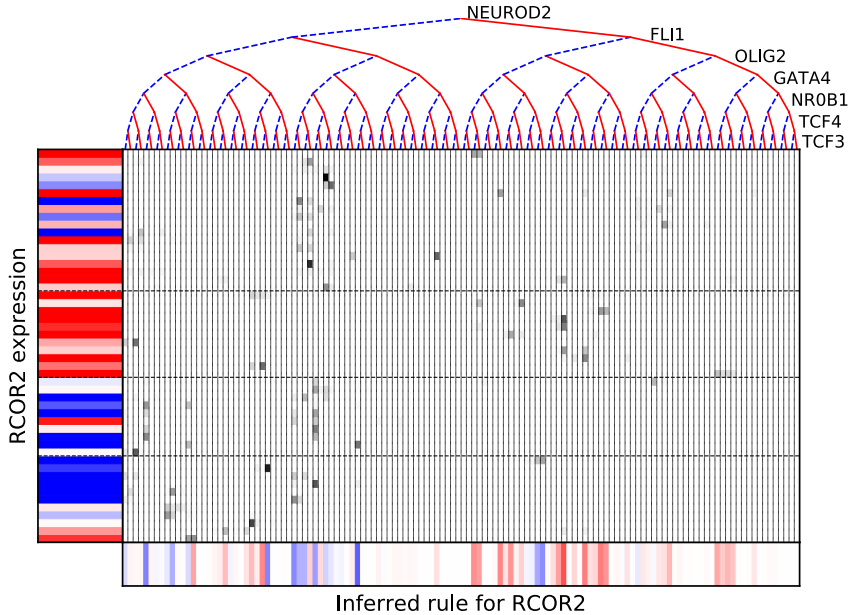

RCOR2
NR0B2
REST
FLI1
NEUROD2 expression
Inferred rule for NEUROD2

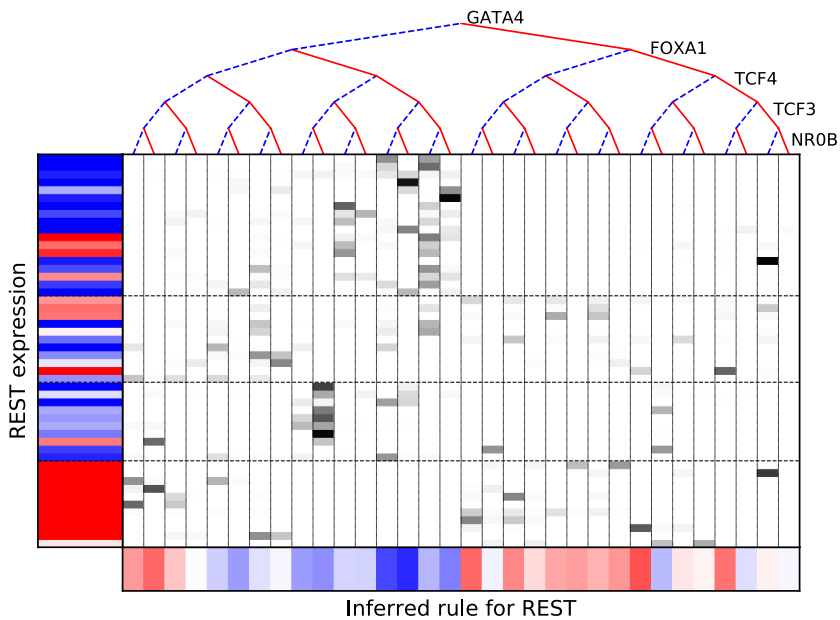

NR0B1
NR0B1 expression
Inferred rule for NR0B1

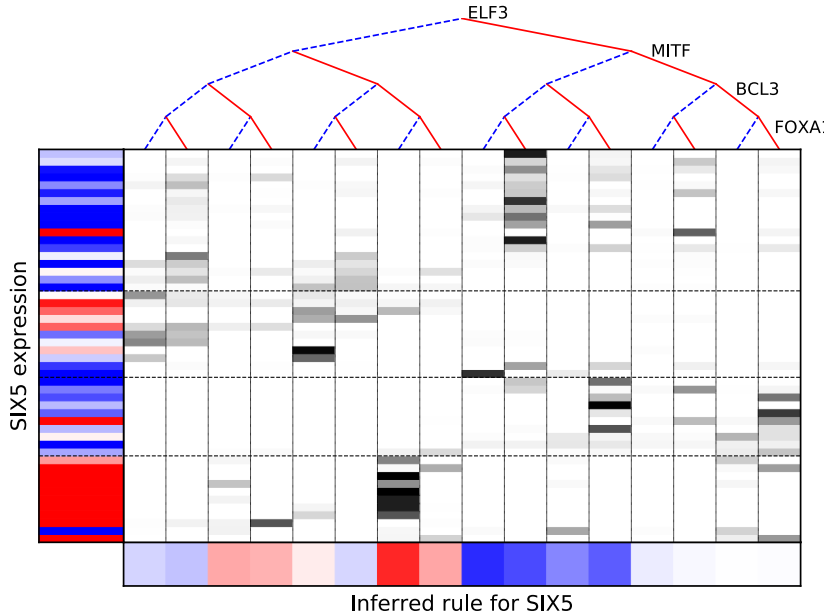

ASCL1
FOXA2
FLI1
FOXA1
GATA4
RBPJ
NR0B2 expression
Inferred rule for NR0B2

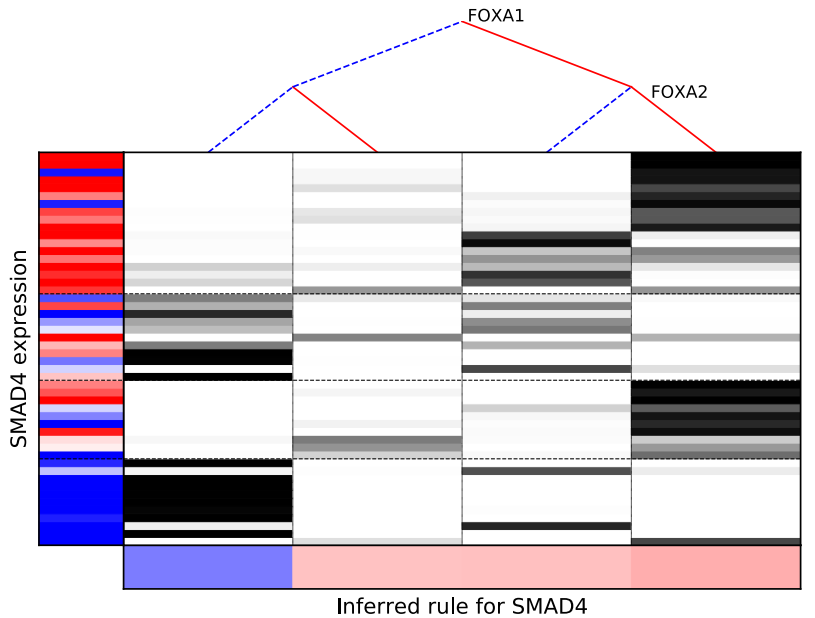

ASCL1
FOXA1
NEUROD2
NR0B1
SMAD4
FLI1
TCF4
TCF3
OLIG2 expression
Inferred rule for OLIG2

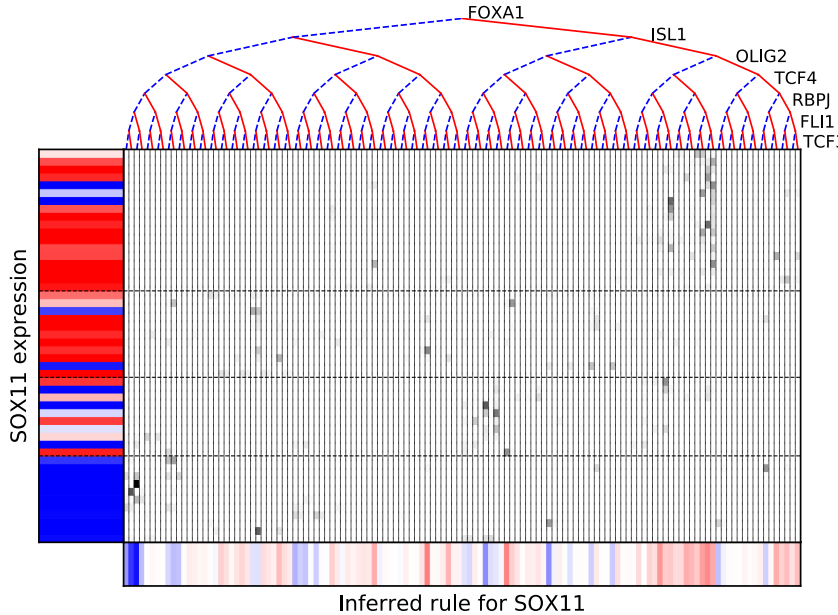

OLIG2
ASCL1
FOXA2
GATA4
ISL1
SMAD4
EBF1
TCF3
TCF4
MITF
RCOR2
RBPJ expression
Inferred rule for RBPJ

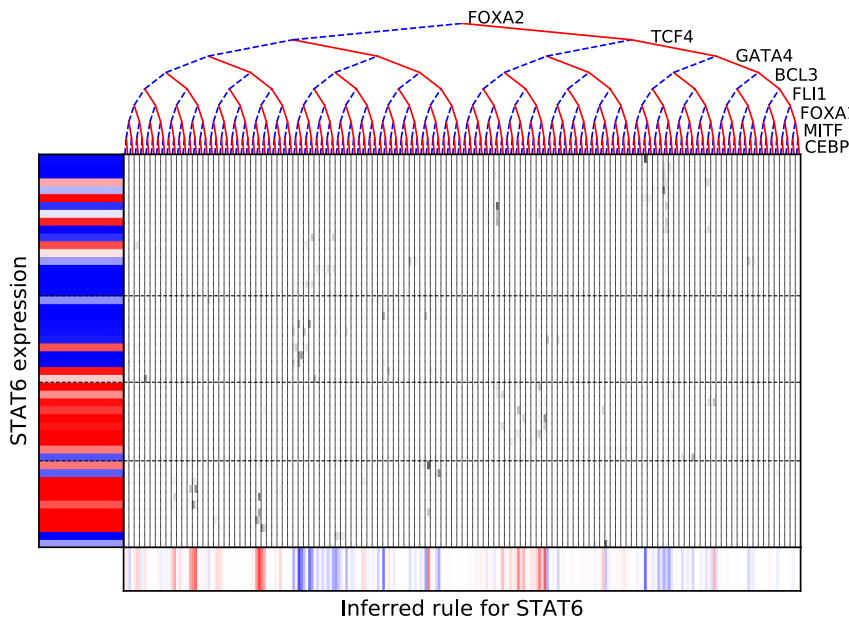

NEUROD2
FLI1
OLIG2
GATA4
NR0B1
TCF4
TCF3
RCOR2 expression
Inferred rule for RCOR2

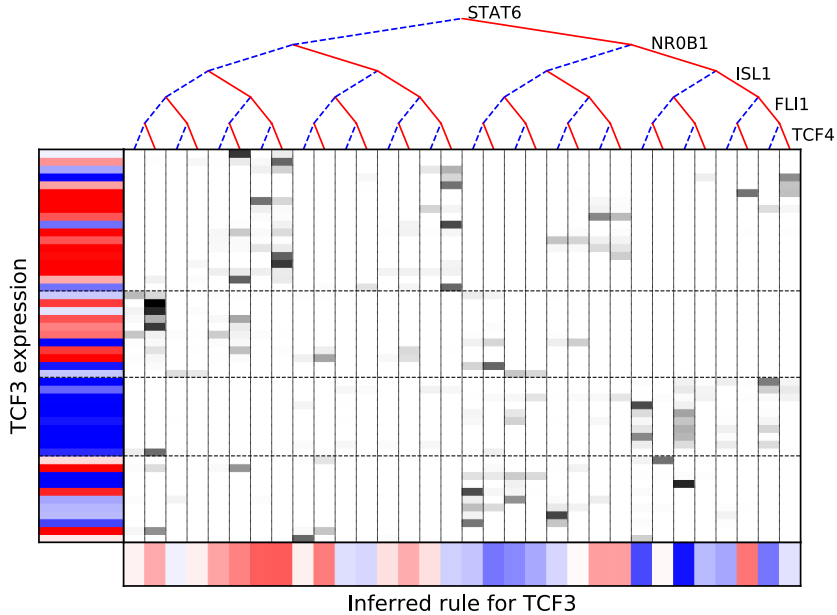

GATA4
FOXA1
TCF4
TCF3
NR0B1
REST expression
Inferred rule for REST

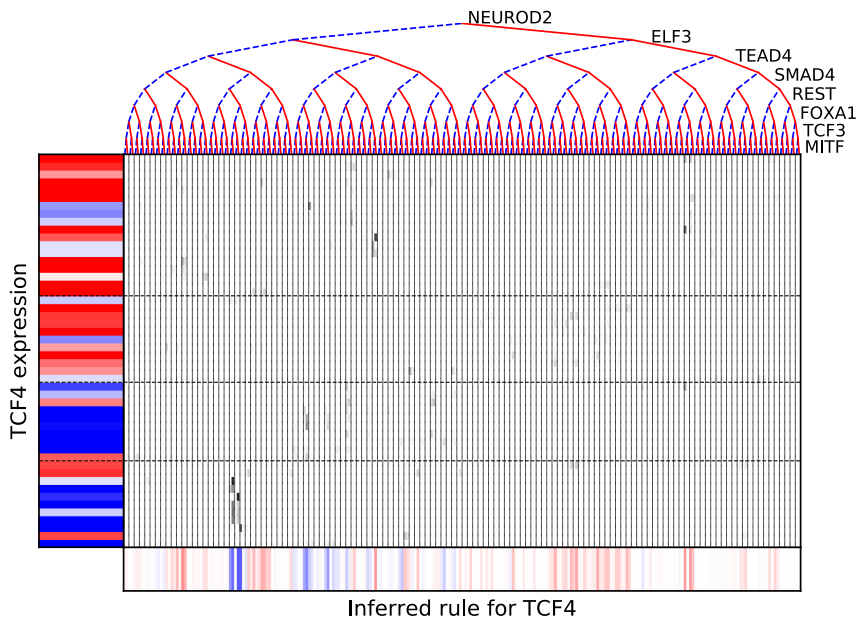

ELF3
MITF
BCL3
FOXA1
SIX5 expression
Inferred rule for SIX5

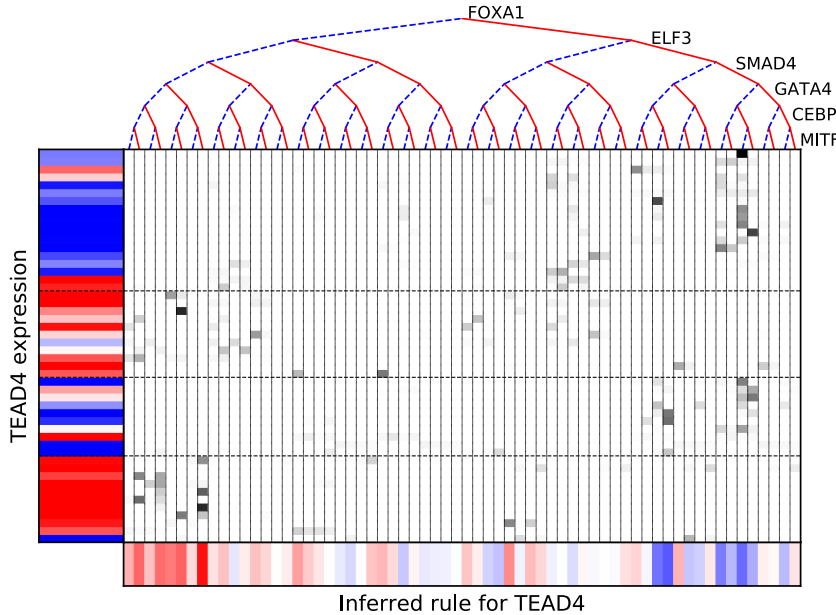

FOXA1
FOXA2
SMAD4 expression
Inferred rule for SMAD4

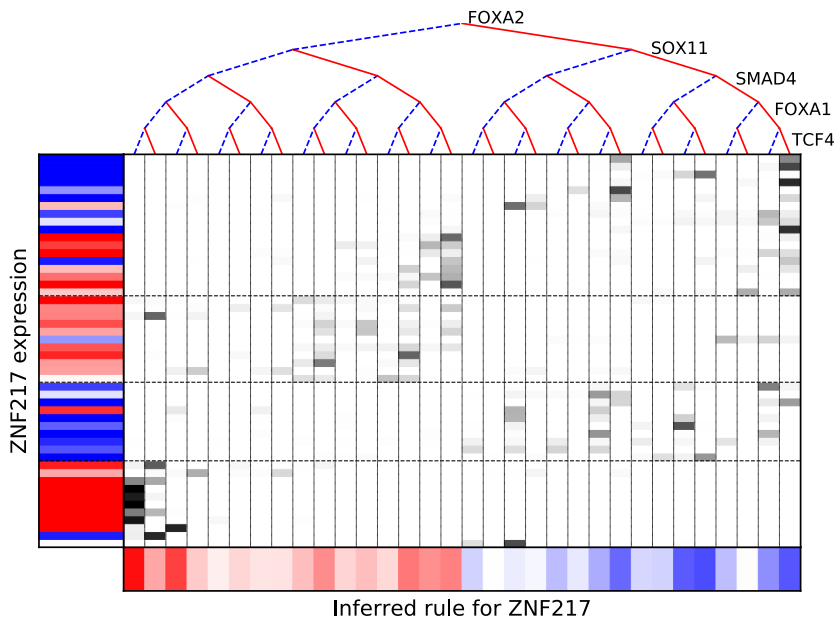

FOXA1
ISL1
OLIG2
TCF4
RBPJ
FLI1
TCF3
SOX11 expression
Inferred rule for SOX11

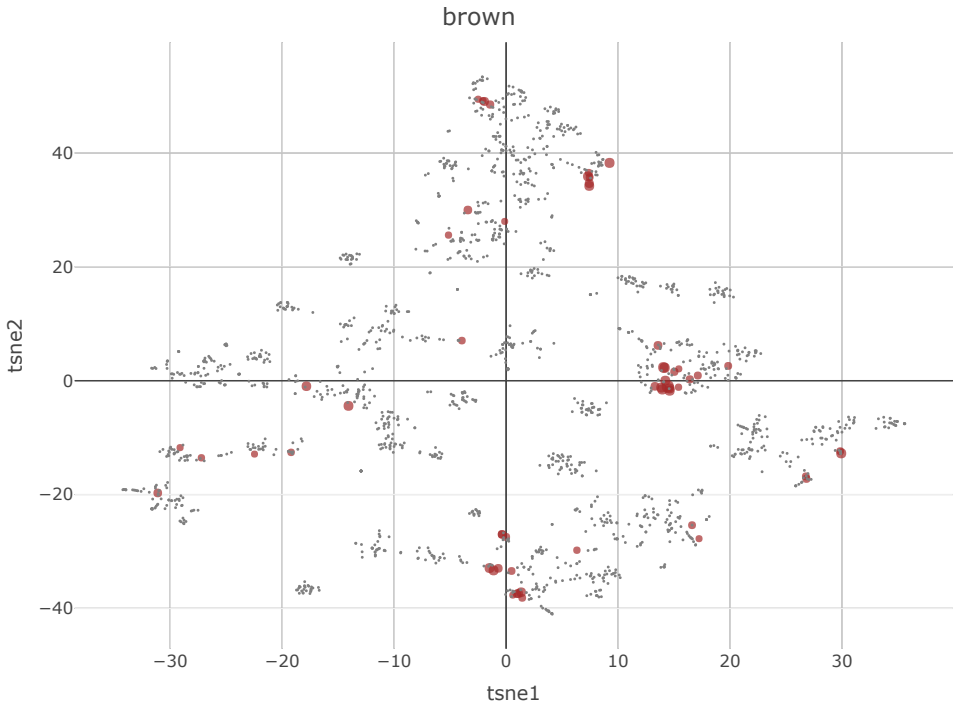

FOXA2
TCF4
GATA4
BCL3
FLI1
FOXA1
MITF
CEBPD
STAT6 expression
Inferred rule for STAT6

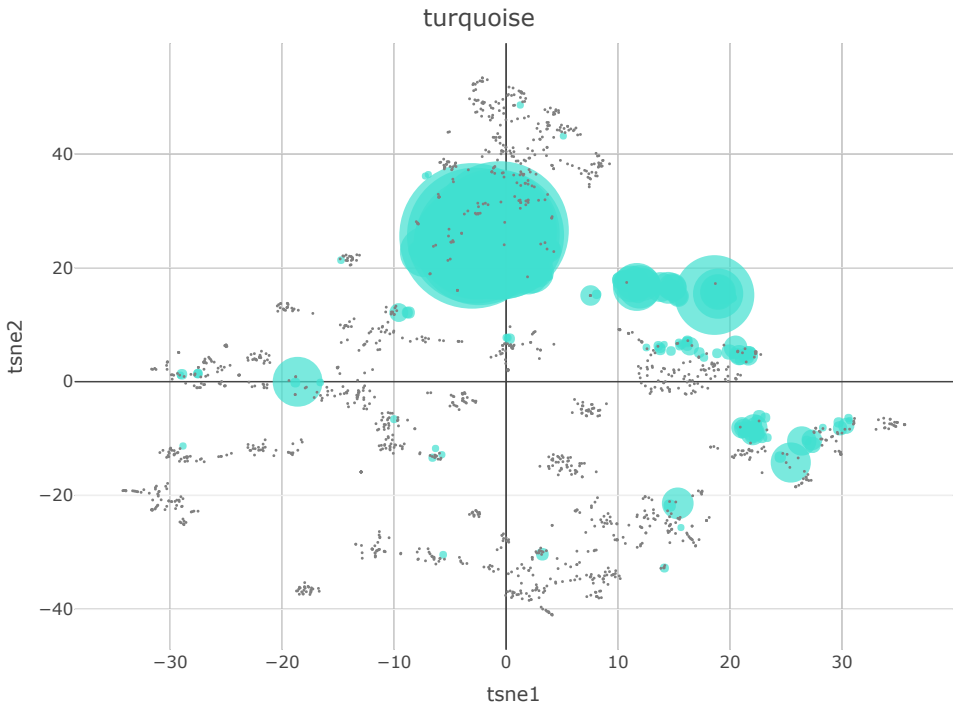

STAT6
NR0B1
ISL1
FLI1
TCF4
TCF3 expression
Inferred rule for TCF3

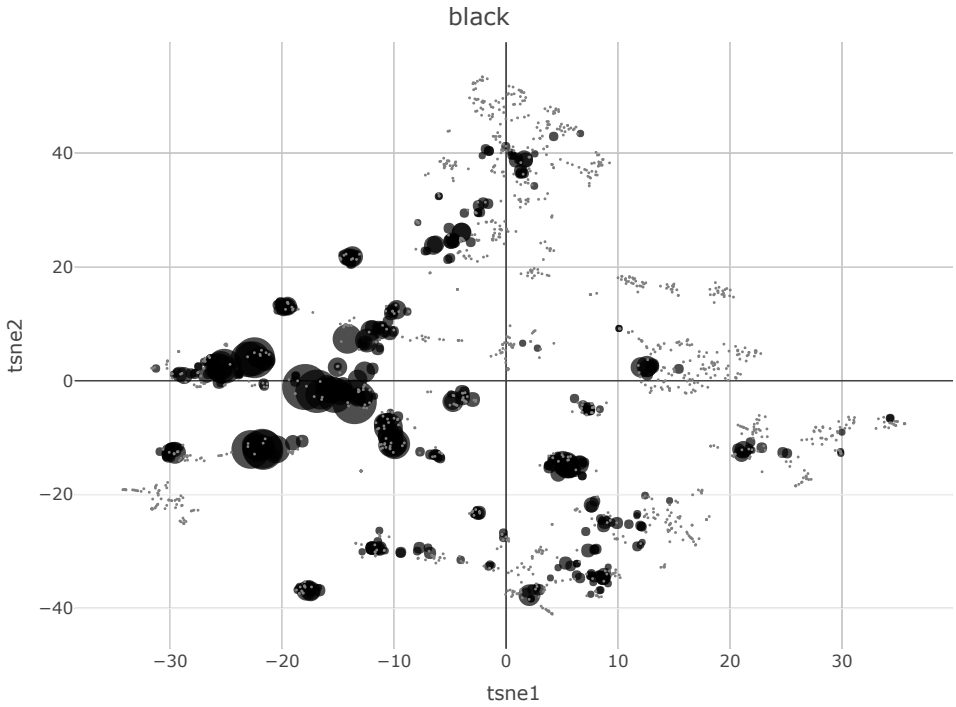

NEUROD2
ELF3
TEAD4
SMAD4
REST
FOXA1
TCF3
MITF
TCF4 expression
Inferred rule for TCF4

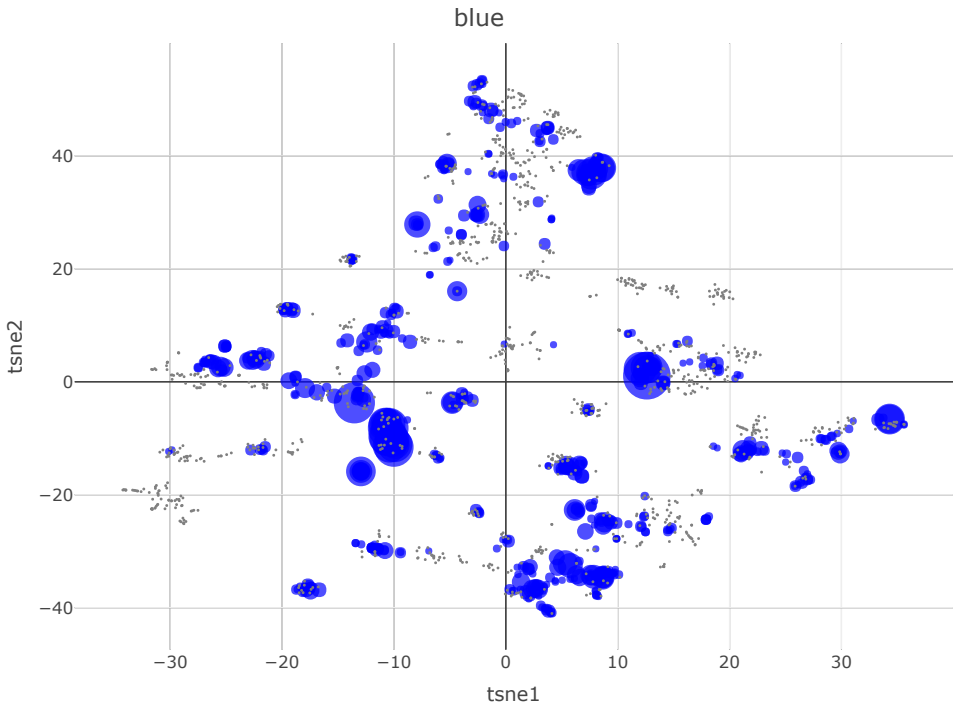

FOXA1
ELF3
SMAD4
GATA4
CEBPD
MITF
TEAD4 expression
Inferred rule for TEAD4

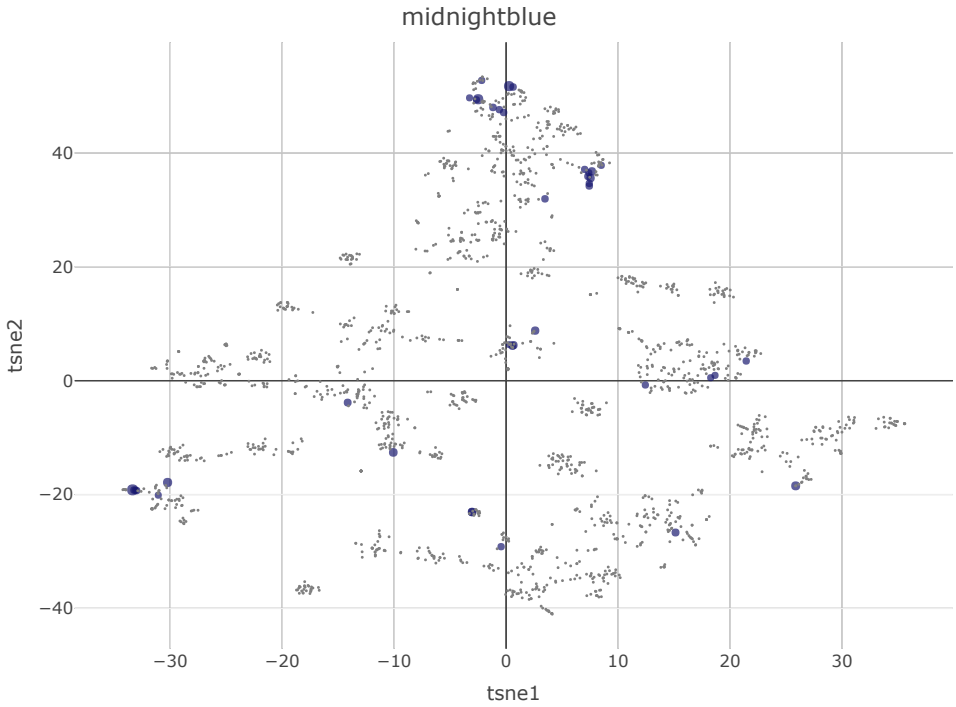

FOXA2
SOX11
SMAD4
FOXA1
TCF4
ZNF217 expression
Inferred rule for ZNF217

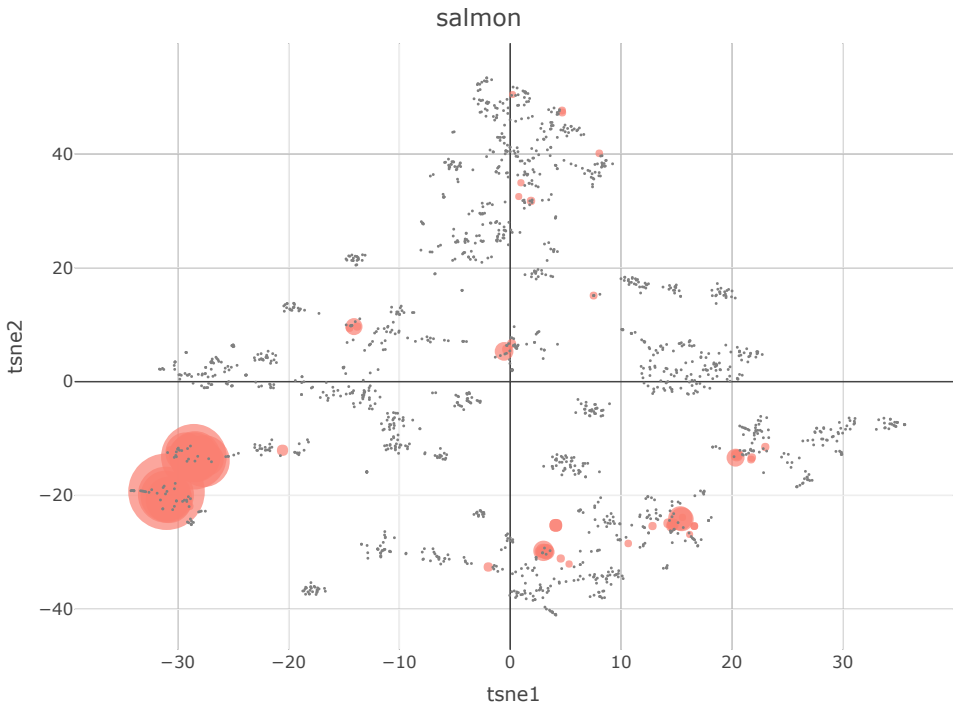

brown
40
20
0
−20
−40
tsne2
−30
−20
−10
0
10
20
30
tsne1

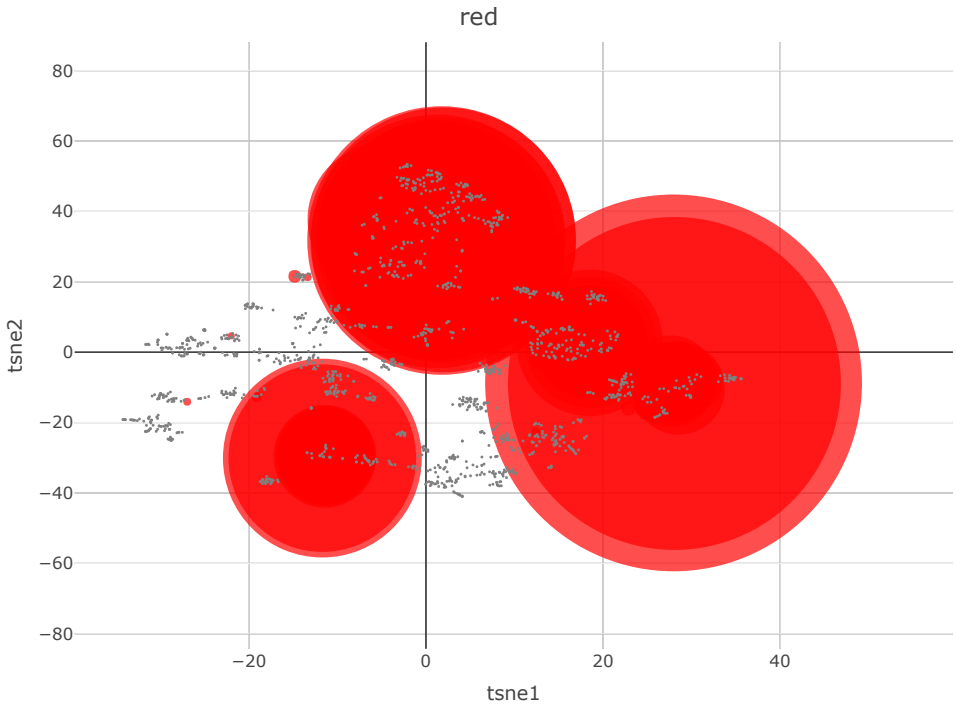

turquoise
40
20
0
−20
−40
tsne2
−30
−20
−10
0
10
20
30
tsne1

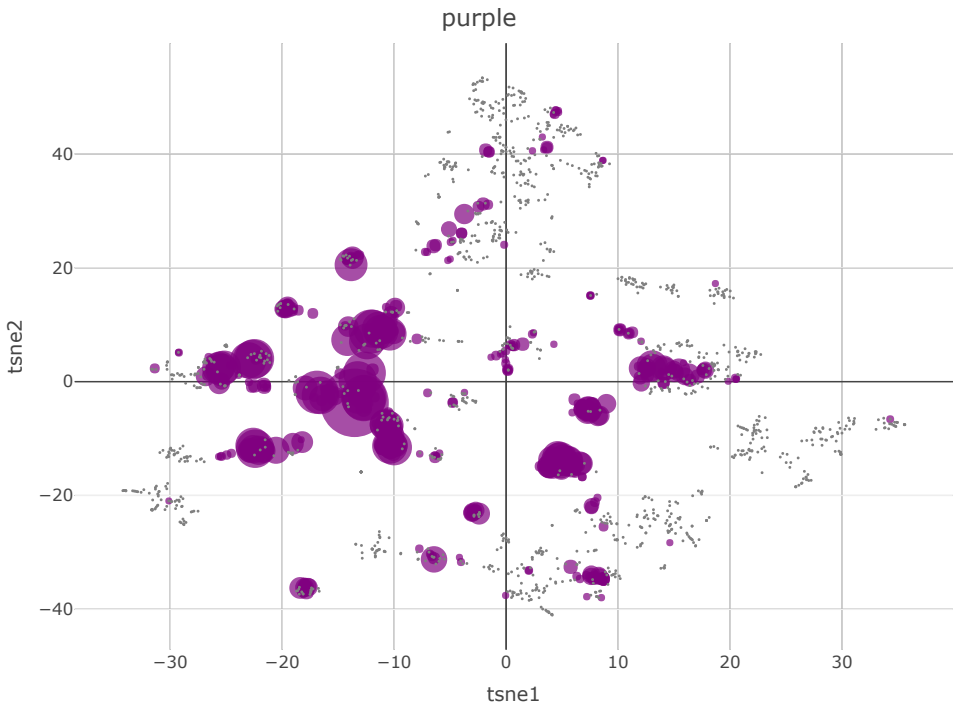

black
40
20
0
−20
−40
tsne2
−30
−20
−10
0
10
20
30
tsne1


blue
40
20
0
−20
−40
tsne2
−30
−20
−10
0
10
20
30
tsne1


midnightblue
40
20
0
−20
−40
tsne2
−30
−20
−10
0
10
20
30
tsne1


salmon
40
20
0
−20
−40
tsne2
−30
−20
−10
0
10
20
30
tsne1


red
80
60
40
20
0
−20
−40
−60
−80
tsne2
−20
0
20
40
tsne1


purple
40
20
0
−20
−40
tsne2
−30
−20
−10
0
10
20
30
tsne1

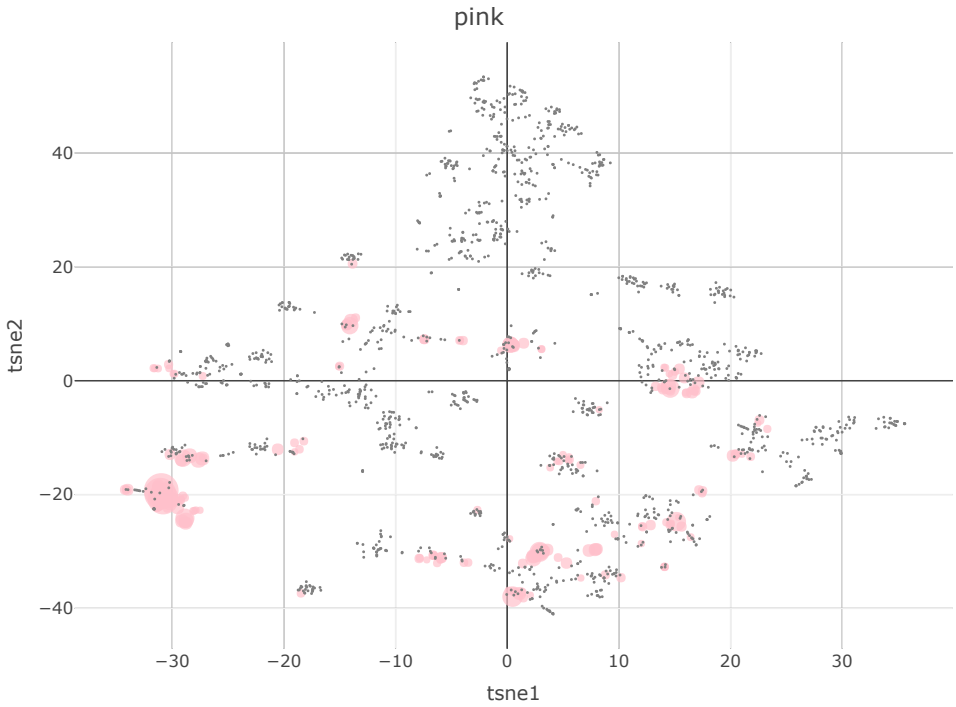

pink
40
20
0
−20
−40
tsne2
−30
−20
−10
0
10
20
30
tsne1

FIGURE 1 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.

FIGURE 2 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. Each blue dot is a GO term, with selected terms highlighted. Several distinct clusters of related processes can be seen.

C. Module-specific phenospace. A breakout 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.

FIGURE 3 Differential response of SCLC subtypes to a wide variety of oncology drugs and investigational agents.

A.

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.

FIGURE 4 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 subtypes (represented by stars). Specifics of the probabilistic simulation are described in Results.

FIGURE 5 Destabilization of subtypes by perturbation to network.

A.Random walks starting from the attractors in Fig 4C will eventually leave the start state due to uncertainty in the Boolean rules. Reference histogram shows how many random steps are required for a distance greater than 4 TF changes away from the start state under the network's natural dynamics. The knockdowns and activations shown here hold expression of the perturbed gene OFF or ON. The perturbation destabilizes the start state, such that the random walk leaves the neighborhood sooner, shown for NE, NEv1, NEv2, and non-NE starting states. **B.** Stabilization of SCLC phenotype 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 indicates destabilizing, while positive indicates increasing stability. Results are shown for 1000 iterations starting from NEv2. Similar plots for the other subtypes can be found in Fig EV3.

C. A Venn diagram demonstrating overlap of destabilization strategies. A single activation (green text) or knockdown (red text) can sometimes destabilize multiple phenotypes. TF perturbations that have a stabilization score of less than -0.2 were considered destabilizing.

FIGURE 6 Computational evidence for existence of subtypes in both human and mouse tumors.

A. Absolute percentage 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. [36]. As shown, the tumors vary greatly in composition, and hierarchical clustering of the patterns result in four clusters (clustering not shown).

FIGURE 7

A. Similar analysis in mouse tumors. Ai. TKO (*RB1*, *TP53*, *P130* floxed) mouse tumors showing a high proportion of NE and NEv2 subtypes. Aii. As described in [20], these mouse tumors were generated by crossing Rb1fl/fl Trp53fl/fl (RP) animals to knockin Lox-Stop-Lox (LSL)-MycT58A-IRES-Luciferase mice. These Rb1fl/fl Trp53fl/fl MycLSL/LSL (RPM) mice have amplified 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 Ci., 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 triple knockout (RB1, TP53, P130) 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.

FIGURE EV1 Tracking plot and delta area plot for consensus clustering 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.

B. The delta area plot shows the relative change in the area under the CDF curve (Fig 1A). 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.

FIGURE EV2 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 *P53* 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 < 0.5 shown.

FIGURE EV3 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.