Network Control of Cell Identity and Plasticity in Small Cell Lung Cancer

By

Sarah Maddox Groves

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Chemical and Physical Biology

May 13, 2022

Nashville, Tennessee

Approved:

Ken Lau, Ph.D.

Emily Hodges, Ph.D.

Bradley Malin, Ph.D.

Jacob Hughey, Ph.D.

Christine Lovly, M.D., Ph.D.

Vito Quaranta, M.D.

To my grandad, Fran Balint, who taught me to love science

Acknowledgements

This work would not have been possible without the financial support of the Vanderbilt Training Program in Big Biomedical Data Science (BIDS) and an NSF Graduate Research Fellowship. I am especially grateful for my mentor, Dr. Vito Quaranta, who has been incredibly supportive of my research and career goals. Thank you for always challenging me to ask questions, to engage with other scientists, and to go outside my comfort zone. You taught me to think critically, and with your guidance I was able to devise my own research plan and grow as a scientist and a person. I am grateful to my committee members for their perceptive questions and scientific expertise. Your feedback on my research proposal and execution provided much-needed insight and guidance. I would also like to thank Dr. Bruce Damon, former Director of the Chemical and Physical Biology program, and Patty Mueller, the Chemical and Physical Biology Program Manager, for their continuous kindness and support.

To all of the members of the Quaranta and Lopez labs, I am eternally grateful. Specifically, I would like to thank Dr. David Wooten, who guided me through the initial ups and downs of grad school. Thank you for your kindness, your remarkably wise view of the world, and, of course, all of our arguments at the white board. To Dr. Christian Meyer, I am so grateful to you and your family, both personally and professionally. Thank you for being so welcoming and for being such a role model as a child of God and a scientist. To Dr. Darren Tyson, thank you for your guidance and mentorship. To Melaine Sebastian and Michelle Scott, you are missed. To all other members, past and present, of the Quaranta lab, I am thankful for your guidance, wisdom, and friendship.

I am incredibly grateful for my friends who got me through the tough times. To Dr. Corey Hayford, thank you for your friendship, mentorship, and for being in my corner the whole way.

To Caroline Stuart, thank you for reminding me that grad school is not my whole life, for late night chats, and for being a piece of Williamsburg in Nashville. To Geena Ildefonso, I am so grateful to be friends with you—Thank you for trauma-bonding, coffee-shopping, and always snacking. To Alexandria Oviatt, thank you for the long conversations, long runs, and your kind heart. To Danielle Kubicki, thank you for being like a sister to me. To Esha Dalvie, thank you for the many moments of understanding and laughter we shared. To Annika Faucon, thank you for your contagious positivity. To all other friends in Nashville and afar, thank you for believing in me and bringing me joy.

Lastly, I am forever indebted to my family. Thank you, Mom and Dad, for your support and love. Thank you for always teaching me to push myself and to stay joyful along the way. To my grandparents, thank you for your constant encouragement and kindness. To my sister, Maggie, thank you for being inspiring and understanding. To my brother, Joey, thank you for always making me laugh and for our thoughtful conversations. To my mother- and father-in-law, thank you for your compassion and for always hearing me. Finally, thank you to my husband, Matthew, for the unbelievable amount of encouragement, love, and grace you have given me. Thank you for always picking me up when I'm down. Thank you for being a listening ear, a shoulder to lean on, and a source of reassurance. I could not have done this without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1.

Introduction

1.1. Small cell lung cancer

1.1.1. History and treatment of small cell lung cancer

Small cell lung cancer (SCLC) is an extremely aggressive, fast growing form of lung cancer that predominantly develops in current or former smokers (Alexandrov et al., 2016). SCLC is characterized by high mutational burden, low immune infiltration, and early metastasis. Furthermore, SCLC is associated with more paraneoplastic syndromes, a group of disorders caused by the production of hormones or peptides by the tumor itself, that any other cancer potentially due to the neuroendocrine (NE) nature of the disease (Kanaji et al., 2014). Though SCLC makes up only 15% of lung cancer cases, it contributes to a large proportion of lung cancer deaths, making lung cancer the leading cause of cancer deaths in the United States (ACS Cancer Facts & Figures 2021).

While smoking prevalence in the United States has decreased in recent years, lung and bronchus cancers still make up over 12% of new cancer cases, second only to prostate cancer in men and breast cancer in women. Smoking is also still increasing in prevalence in developing countries, such that deaths from SCLC are increasing worldwide. Understanding how to effectively target and treat SCLC is therefore a critical step to reducing cancer mortality worldwide. Unfortunately, prospects for SCLC patients are dismal: the five-year survival rate remains around 5%, with a median survival duration of less than 2 years for patients with early-stage disease and about 1 year for patients with late-stage disease (Semenova et al., 2015).

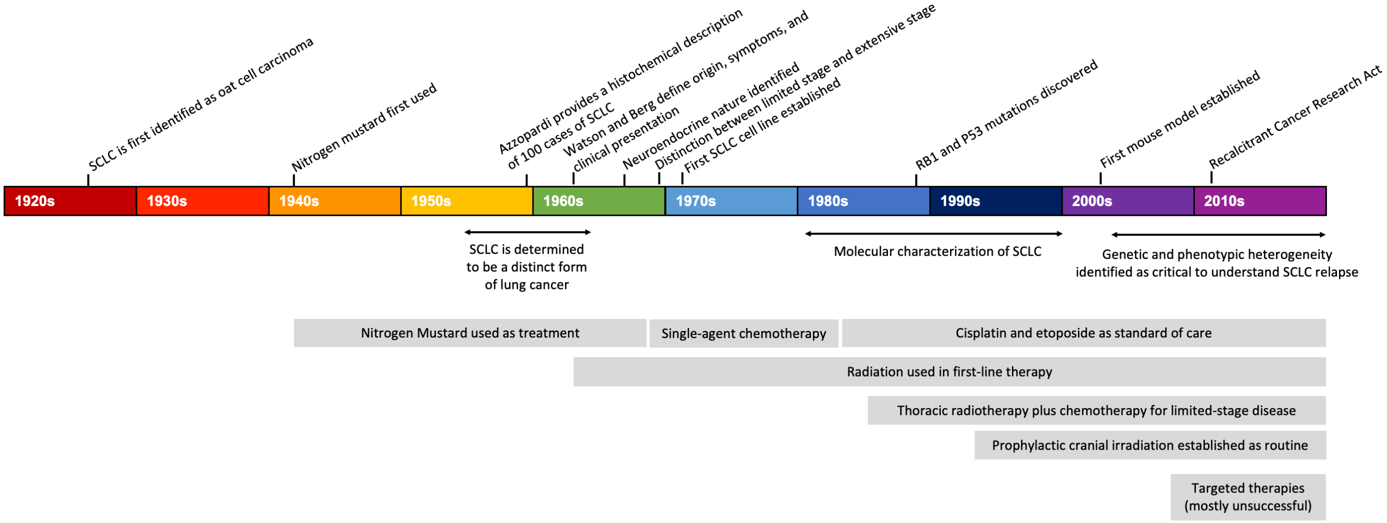

SCLC is first identified as oat cell carcinoma
Nitrogen mustard first used
Azzopardi provides a histochemical description of 100 cases of SCLC
Watson and Berg define origin, symptoms, and clinical presentation
Neuroendocrine nature identified
Distinction between limited stage and extensive stage
First SCLC cell line established
RB1 and P53 mutations discovered
First mouse model established
Recalcitrant Cancer Research Act
1920s
1930s
1940s
1950s
1960s
1970s
1980s
1990s
2000s
2010s
SCLC is determined to be a distinct form of lung cancer
Molecular characterization of SCLC
Genetic and phenotypic heterogeneity identified as critical to understand SCLC relapse
Nitrogen Mustard used as treatment
Single-agent chemotherapy
Cisplatin and etoposide as standard of care
Radiation used in first-line therapy
Thoracic radiotherapy plus chemotherapy for limited-stage disease
Prophylactic cranial irradiation established as routine
Targeted therapies (mostly unsuccessful)

Figure 1.1. Timeline of major events in SCLC history.

In 1926, SCLC was first described as “oat-celled carcinoma” due to the small, flat appearance of these cancer cells under a microscope (Gazdar et al., 2017) (**Figure 1.1**). Decades later, Watson and Berg helped to further define this type of cancer and proposed nitrogen mustard and radiation as the standard of care therapy. Because SCLC, in contrast to non-SCLC (NSCLC), is characterized by early and frequent metastases, surgery is often not an option for treatment. The resulting lack of SCLC tissue for research studies has hindered progress in treating this disease, and the standard of care has remained a cytotoxic chemotherapy similar to the nitrogen mustard treatment used almost a century ago. Today, patients receive a first-line therapeutic regimen of etoposide and a platinum-based agent such as cisplatin (EP) and radiation, despite the fact that virtually all patients relapse after therapy. Because of these statistics, SCLC was designated a recalcitrant cancer by the Recalcitrant Cancer Research Act in 2012, which reinvigorated interest, funding, and research for SCLC.

Recently, an increased understanding of the molecular characteristics of SCLC has allowed for the development of targeted therapies, although most have been unsuccessful (Gadgeel, 2018).

For example, SCLC is characterized by extremely fast-growing tumors and rapid onset of metastases, which suggests inhibition of pathways related to self-renewal (such as Hedgehog) or DNA repair (such as Poly (ADP-ribose) polymerase enzymes, PARP) may be beneficial in SCLC (Gadgeel, 2018). Unfortunately, PARP inhibitors have only shown modest improvement over the standard of care in a subset of patients, and Hedgehog pathway inhibition, such as with the Smoothened inhibitor vismodegib, did not improve progression-free or overall survival in a clinical study (Belani et al., 2016; Owonikoko et al., 2017). Other targeted therapies that have seen success, such as immune checkpoint inhibitors, are only beneficial in a minority of patients (Hellmann et al., 2017; Ott et al., 2017). Therefore, there is an urgent need for a deeper understanding of SCLC to develop better therapies for patients that both extend survival rates and increase quality of life over that seen with cytotoxic chemotherapeutic regimens.

1.1.2. Genetic and epigenetic heterogeneity in SCLC

A defining feature of SCLC is the virtually ubiquitous biallelic inactivation of TP53 and RB1 (George et al., 2015). While silencing of these two tumor suppressors seems necessary for SCLC to develop, no oncogenic drivers seen in other cancer types, such as BRAF in melanoma, are necessary for SCLC development. Due to carcinogenic cigarette smoke, SCLC tends to have an extremely high mutational burden, with an average of 7.4 protein-changing mutations per million base pairs (Peifer et al., 2012). While the role of many of these mutations is still incompletely understood (George et al., 2015), some signaling pathways show recurrent mutations in a subset of patients.

For example, all three members of the MYC family genes are often mutated, with expression of MYC, MYCL, and MYCN in a mutually exclusive manner. MYCL tends to be amplified or highly expressed in the majority of neuroendocrine-high SCLC, while MYC is amplified in about 20% of SCLC tumors, which tend to have a more non-neuroendocrine phenotype. Notch pathway regulates NE differentiation and progression, and thus acts as a tumor suppressor against ASCL1+ tumors. Genes in this pathway tend to be mutated in SCLC, with about 25% of SCLC harboring loss-of-function mutations in Notch receptors. Unfortunately, mutations have been incapable of defining clinically relevant subtypes of SCLC. Therefore, our current understanding of genetic heterogeneity in SCLC is not enough to identify clinically actionable subgroups of patients.

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 strategies (**Figure 1.2**). 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 (Carney et al., 1985; Gazdar et al., 1985, 2017). In both human and mouse tumors, most cells appear to belong to the NE subtype, corresponding to a pulmonary neuroendocrine cell (PNEC) of origin (Sutherland et al., 2011), with high expression of neuroendocrine genes such as ASCL1. However, several groups have found evidence for non-NE variants within SCLC tumors (Calbo et al., 2011; Huang et al., 2018; Lim et al., 2017), as well as an NE variant driven by MYC overexpression and NEUROD1 overexpression, instead of ASCL1 (Borromeo et al., 2016; Mollaoglu et al., 2017; Sos et al., 2012). Non-NE subtypes have further been described by driving transcription factors, such as YAP1 and POU2F3. Our lab previously described SCLC cell lines with hybrid expression of both NE and non-NE markers (Udyavar et al., 2017) 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 (Lim et al., 2017; Tammela et al., 2017). The relationship between these subtypes is less clear; some research has shown evidence of phenotypic transitions between subtypes, such as from the NE subtype to the non-NE subtype (Lim et al., 2017). Understanding the plasticity of these subtypes in different conditions, such as after chemotherapy, is one of the goals of this dissertation. To do so, we use the theoretical systems biology notion of a phenotypic landscape, undergirded by gene regulatory network dynamics, combined with bioinformatics approaches to analyze patterns of behavior of single cells that reside within the landscape.

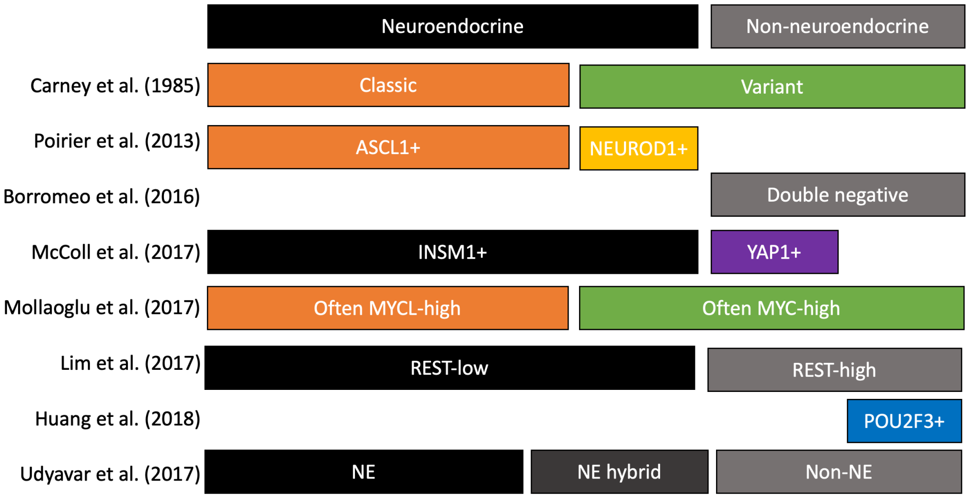


Figure 1.2: History of subtype identification in SCLC. *Over the last half century, SCLC heterogeneity has been clustered in various ways. The first classification of phenotype heterogeneity considered classic and variant subgroups. Since then, gene signatures and driving transcription factors have been used to identify subtypes.*

1.2. Understanding cell identity in cancer

1.2.1. Layers of heterogeneity in cancer

Heterogeneity within tumors has been shown to be critical for acquired resistance to therapy in several cancer types. As described for SCLC, several layers of heterogeneity exist: genetic heterogeneity dependent on selection of mutants, and non-genetic heterogeneity dependent on epigenetic regulation of phenotype and the inherent stochasticity in biology (Hayford et al., 2021).

Variability in genotype is necessary for clonal selection through Darwinian evolution, where clones with a higher fitness outcompete others. Genomic heterogeneity is often driven by genomic instability, a hallmark of cancer, which generates random mutations including rearrangement of entire chromosomes (Hanahan and Weinberg, 2000). While normal cells have an extraordinary ability to detect and repair defects in DNA and therefore keep the rate of spontaneous mutation to a minimum, cancer cells often increase the rates of mutation and therefore decrease genomic stability. This may be achieved in a cancer cell by increased sensitivity to mutagenic agents, such as cigarette smoke in the case of SCLC, and by defects in the repair machinery that maintains genomic fidelity. In SCLC, virtually universal inactivation of TP53, the “guardian of the genome,” plays a central role in compromising the repair mechanisms that help to reduce mutagenesis (Lane, 1992).

However, it is becoming increasing clear that genetics alone cannot explain the variability seen between and within tumors. Instead, non-genetic heterogeneity must also be considered. Two main types of non-genetic heterogeneity exist: deterministic heterogeneity, which is often used synonymously with epigenetic heterogeneity, and stochastic heterogeneity, often called stochasticity, noise, or randomness. Deterministic heterogeneity describes the existence of multiple stable phenotypic states given a particular genome. Epigenetic heterogeneity can be attributed to several sources, including variability in chromatin accessibility, DNA methylation, and DNA-binding proteins that regulate transcription levels of genes (**Figure 1.3)**. In this work, phenotypic state refers to the transcriptomic state of a cell or tumor, with an implicit understanding that other levels of epigenomic regulation, such as chromatin modifications, help

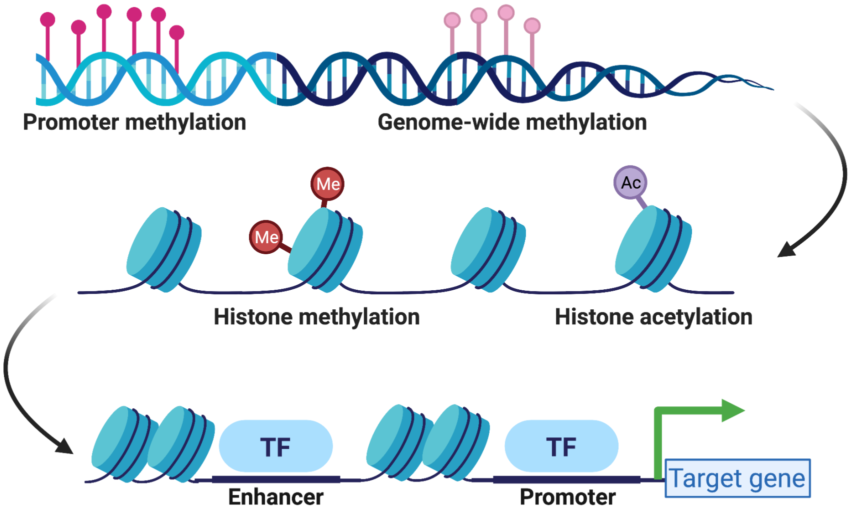


Figure 1.3 Epigenetic modifications control gene expression. *Various post-translational modifications are mechanisms of gene regulation. Promoter and genome-wide methylation of DNA decreases gene expression; in normal development, this controls tissue-specific gene expression, and in cancer, methylation often occurs near tumor suppressor genes to silence them (Wajed et al., 2001). Histone acetylation tends to increase gene expression. Transcription factors (TFs) bind to promoters and enhancers of genes to promote or inhibit transcription. Each of these modifications can therefore lead to an increase (or decrease) in transcription (green arrow) of a target gene.*

to determine a cell's transcriptomic profile. We visualize and quantify this heterogeneity in the following sections using epigenetic landscapes.

There is also a certain amount of variation in gene expression that can be expected between isogenic cells in the same phenotypic state, which is attributed to stochasticity. Stochasticity arises from intrinsic sources, such as the probabilistic nature of biochemical reactions within a cell, or extrinsic sources, such as local fluctuations in chemical concentrations in the microenvironment. While transient, this variability can probabilistically drive transitions between phenotypes (Feinberg and Irizarry, 2010; Gupta et al., 2011; Hayford et al., 2021; Liao et al., 2012).

Together, these layers of heterogeneity—genetic, epigenetic, and stochastic— define the variability in cell identity, or phenotype. A critical need in cancer research, particularly for SCLC, is understanding these levels of heterogeneity quantitatively, as each phenotype will respond differently to treatment, and heterogeneity can lead to acquired resistance. To some extent, much of this phenotypic variability was hidden before the advent of single-cell sequencing, which has greatly amplified the ability of researchers to investigate heterogeneity within a population. Previously, bulk sequencing methods that average transcriptomic profiles of thousands of cells obfuscated the heterogeneity within cell populations of a single cell "type." Now, with single cell technology, it is clear that heterogeneity within cell "types" plays a huge role in cancer progression and relapse (Marjanovic et al., 2020; Neftel et al., 2019; Stewart et al., 2020; Wahl and Spike, 2017).

1.2.2. Cell type identification of single cells

With the explosion of information regarding phenotypic variability, single cell data has redefined what is meant by cellular identity. It is often unclear whether cell phenotype should be described by a continuous phenotype quantified by expression of gene signatures along major axes of variance, or by discrete subtypes that can easily be teased apart (Wagner et al., 2016).

In SCLC, a small number of biomarkers are often used to describe cell identity in different systems—for example, ASCL1 and NEUROD1 for NE subtypes. However, it is becoming increasingly clear that a few biomarkers cannot fully capture the heterogeneity seen in SCLC cell populations, and there is a need to use new tools for understanding cell identity more comprehensively. Single cell sequencing allows for systematic identification of gene expression signatures and biological signals, which is quickly becoming the standard for identifying cell phenotypes. Furthermore, single cell data provides a source of phenotypic classification for populations where biomarkers are not yet known, and can lead to identification of novel cell types, such as recognition of phenotypic subtypes in various cancers that could not be teased apart with previous methods (Abdelaal et al., 2019; Baron et al., 2016; Karaayvaz et al., 2018; Lieberman et al., 2018; Marjanovic et al., 2020; Pellin et al., 2019; Plasschaert et al., 2018; Travaglini et al., 2020; Weinreb et al., 2020; Wooten et al., 2019).

Discrete clustering is a powerful way to classify data into actionable, functional identities in cancer and other diseases (Balanis et al., 2019; Bebber et al., 2021; Borromeo et al., 2016; Bramsen et al., 2017; Gay et al., 2021; Karaayvaz et al., 2018; Poirier et al., 2015; Rudin et al., 2019; Schafer et al., 2020; Simbolo et al., 2019; Simpson et al., 2020; Wang et al., 2019b; Yeo and Guan, 2017). However, the variance in expression seen in single cell data is often more continuous than binary, and many systems with biomarker- or cluster-defined cell types can contain hidden diversity (Trapnell, 2015). In some cases, hybrid cell types between distinctly defined clusters may exist (Antebi et al., 2013; Patel et al., 2014; Udyavar et al., 2017). Furthermore, cell identity in several cancer types has been shown to be continuous, such as in glioblastoma (Neftel et al., 2019; Wang et al., 2019a) and lung adenocarcinoma (LaFave et al., 2020). Continuous states can be characterized based on prior knowledge, such as gene signatures that are upregulated in a cell type of interest. For example, matrix factorization methods find biological patterns across samples or single cells, which can then be characterized by function (Stein-O'Brien et al., 2018). Alternatively, the expression level of pre-defined signatures representing functions of interest can place single cells on a spectrum, such as a spectrum between neuroendocrine (NE) and non-NE cells in SCLC (Zhang et al., 2018).

One of the benefits of a continuous paradigm for cell identity is that cell dynamics are generally thought be continuous, i.e., state transitions occur smoothly through space from one cell type to another (Brackston et al., 2018; Eizenberg-Magar et al., 2017; Mulas et al., 2021; Su et al., 2017; Zhou et al., 2021). Continuums therefore allow us to understand the intermediate, transitioning states and better characterize transition paths for dynamic processes. The lack of a continuous description of phenotype in SCLC over the last few decades may have hindered our understanding of how SCLC cells might move through these transition paths. Therefore, a more comprehensive and continuous view of phenotypic heterogeneity in SCLC should lead to better understanding of how SCLC tumors change in response to treatment (Udyavar et al., 2017).

To better understand how cell identity is regulated, this dissertation uses a combination of top-down and bottom-up modeling approaches. Bottom-up approaches, which build models from the underlying theory of biochemical interactions, can predict strategies for controlling cell identity. In this dissertation, gene regulatory network inference is used to characterize cell identity control by transcription factors. However, these models can be limited to less complex dynamics of cell identity. Therefore, we supplement the network inference modeling with top-down approaches, which build statistical models of cell identity directly from high dimensional data. Together, these approaches are used synergistically to characterize and understand the regulatory control of SCLC cell identity via the metaphor of an epigenetic landscape.

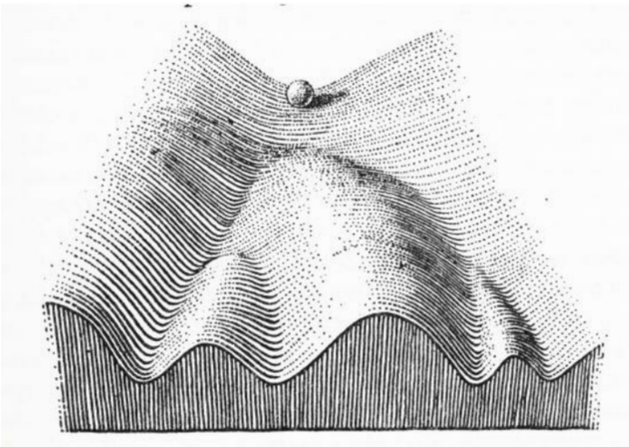


Figure 1.4: Epigenetic landscape. (Waddington, 1957)

1.2.3. Waddington's epigenetic landscape in cancer

In 1957, C.H. Waddington proposed the concept of an epigenetic landscape for understanding the regulation of phenotype in the context of biological differentiation (Waddington, 1957). During development of an organism with a specific genome, cells can change into very distinct-looking cells. Waddington proposed that differentiation could be thought of as an epigenetic landscape, "a rough and ready picture of the developing embryo" (Waddington, 1957). In this analogy, cells roll downhill through canalized channels or "chreods" representing differentiation pathways. Cells at the top of the landscape are pluripotent stem cells, and as they travel down the landscape, they gradually become more committed to a particular cell fate.

In normal development, cells are generally isogenic. In cancer, however, where the mutation rate is higher and multiple subclones may exist within a single tumor, genetic heterogeneity can be represented by a "fitness landscape" (**Figure 1.5, bottom**). In this landscape, mutants with higher fitness will be selected for via Darwinian evolution. For each location in the fitness landscape (each genome), an entire Waddington landscape of phenotypes exists (**Figure 1.5, top**). Similar to Waddington's original conception, cells in the epigenetic landscape in **Figure**

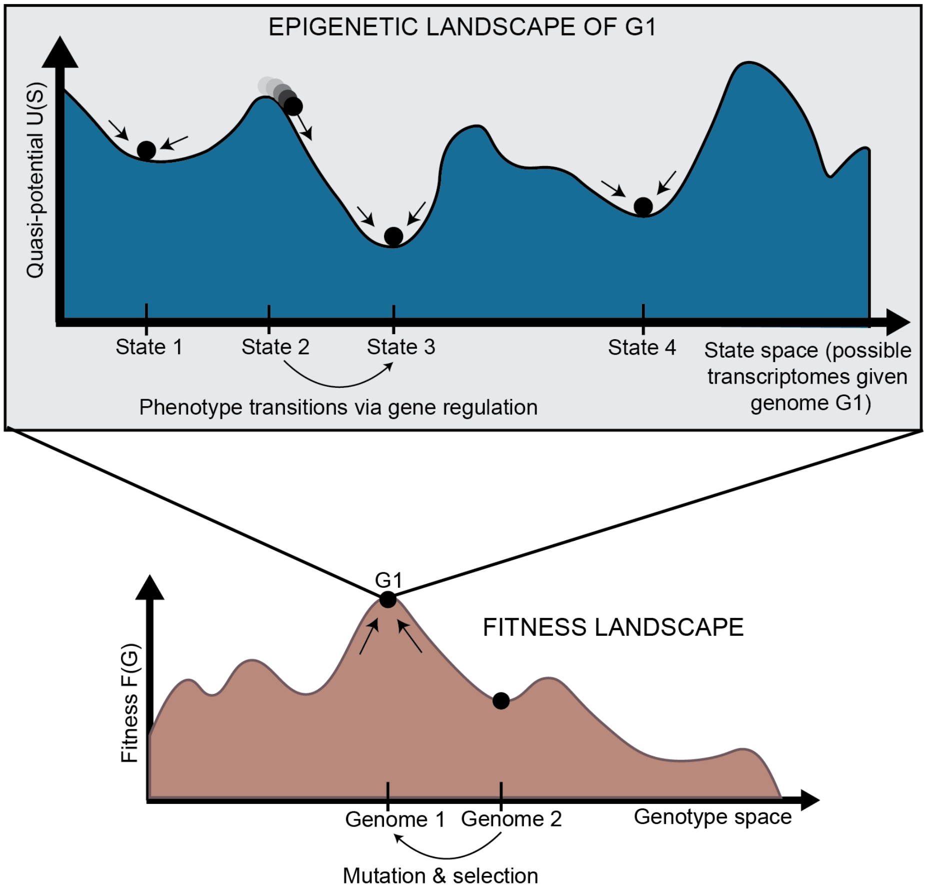


Figure 1.5: Relationship between the fitness landscape and epigenetic landscapes. *Each epigenetic landscape is associated with a single genome (G1). Selection of high-fitness mutants can be represented by cells "climbing" up a fitness landscape, where each point along the horizontal axis is a different genome. For a specific genome, we can imagine an entire epigenetic landscape that characterizes the phenotypes associated with that genome (since there is not a one-to-one, but one-to-many, relationship between genotype and phenotype). Phenotypic transitions through epigenetic mechanisms allow for movement through the epigenetic landscape.*

1.5 "fall downhill" towards the states with the lowest "potential." These phenotypic transitions depend on the instability of each cell state, and a cell's ability to transition can be defined by its plasticity.

1.2.4. System attractors, instability, and plasticity

The notion of plasticity goes hand in hand with the dynamical systems theoretical concept of instability. In dynamical systems, stability of a state requires more than stationarity; a stable state is one that is resilient to perturbations such that, after external influences such as changing microenvironmental conditions, the system returns to its original state. This idea is represented in the potential landscape, in which cells roll downhill toward local minima, as shown in **Figure 1.6 (top)**. While there may be steady states throughout the landscape, such as the top of a flat hill or the bottom of a valley, a small push to a cell on top of a hill will cause it to roll down to a local minimum, far from its original starting state. On the contrary, a cell in a local minimum is resilient to small perturbations: it is in a stable "attractor" state of the landscape. The high-dimensional region around the attractor where a cell will roll towards the attractor is called the basin of attraction (**Figure 1.6, bottom**). Cell states with larger basins of attraction can withstand larger perturbations to their cell state, thereby demonstrating resilience of the system.

In dynamical systems theory, plasticity is a weaker kind of stability, in which a perturbed system neither returns to its original state nor escapes from it, but instead tracks the environmental change (Huang, 2013). However, in biology, plasticity and instability are often thought of as interchangeable: a more plastic cell state responds to an external perturbation by changing its state to a larger degree. In this view, cells with higher potential on the landscape are often more plastic. I therefore utilize this biological interpretation of plasticity when discussing plasticity in future chapters.

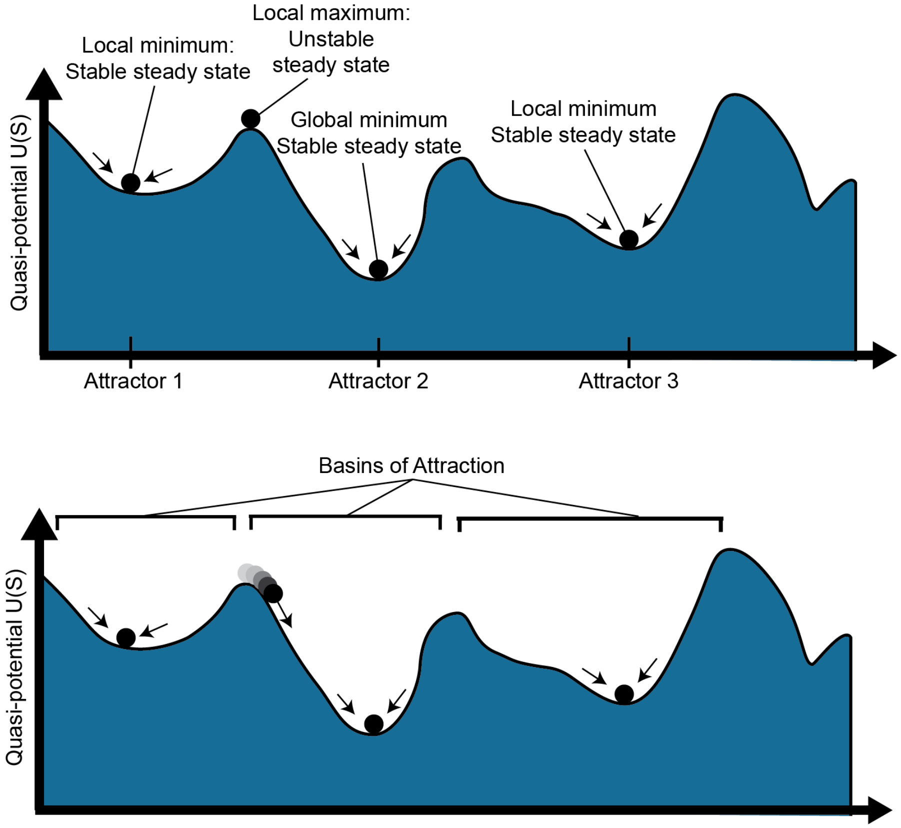


Figure 1.6: Phenotype stability and attractors. *The epigenetic landscape shown above has multiple stable and unstable steady states. While a cell at a local maximum could technically be stable, small stochastic perturbations to the cell will quickly push it one direction or another towards a local minimum. Attractor 2 has the lowest potential as the global minimum. The region around each attractor where cells will move towards the attractor is known as the basin of that steady state.*

1.2.5. Quantifying quasi-potential of the landscape

While Waddington intended this picture as purely a metaphor, it has now been quantified in various ways, borrowing ideas from physics and dynamical systems theory to describe the underlying regulation of these processes (Wang et al., 2008, 2011; Zhou et al., 2012). The height of the landscape describes instability of each phenotype as a “quasi-potential,” a correlate of gravitational potential in a physical landscape (**Figure 1.5, top**). Quantification of this quasi-potential is informative for processes in which plasticity and instability plays a central role, including SCLC and other cancer systems. By modeling phenotypic potential of heterogeneous populations within an SCLC tumor, we can better determine ways to control the permissivity of phenotype and prevent reprogramming of cell identity from a sensitive phenotype to a resistant one.

Borrowing from physics, movement of cells in the landscape (i.e., changes in $\mathbf{x_i}(x_1,x_2,\ldots,x_N)$ over time) may be due to some "force" $\mathbf{F}(\mathbf{x})$, similar to the effect of gravity on movement through a physical landscape. A potential, $\mathbf{U}(\mathbf{x})$, can be defined such that the change in phenotype is equal to the gradient of this potential:

$$\frac{d\vec{\boldsymbol{x}}}{dt} = \boldsymbol{F}(\vec{\boldsymbol{x}}) = -\nabla\mathbf{U}(\vec{\boldsymbol{x}})$$

Cells will therefore "roll down" the gradient of $\mathbf{U}(\vec{\boldsymbol{x}})$ towards states with lower potential. It is worth noting that most high-dimensional, non-equilibrium biological systems are not simple gradient systems, and therefore the vector field $\boldsymbol{F}(\vec{\boldsymbol{x}})$ is sometimes decomposed into two components: the gradient of some quasi-potential, $\widetilde{\boldsymbol{U}}(\vec{\boldsymbol{x}})$, and a remainder term. Still, the gradient term has been successfully used to understand pathways of transition through epigenetic landscapes, describing everything from differentiation to cell fate reprogramming (Li and Wang, 2014; Luo et al., 2017; Wang, 2015; Wang et al., 2006, 2010a; Wu and Wang, 2013a, 2013b; Yan et al., 2019; Zhou and Huang, 2010; Zhou et al., 2012, 2016a).

Several systems biology approaches have been developed to determine the driving force $\mathbf{F}(\mathbf{x})$ that shapes the epigenetic landscape and defines phenotypic heterogeneity and paths of transition. Classical dynamical systems modeling of underlying gene regulatory networks is a bottom-up approach that can explain how phenotypic transitions are dependent on regulation of gene expression by transcription factors (TFs). Alternatively, phenomenological top-down approaches based on analysis of large datasets can approximate the potential landscape. For example, single-cell sequencing sample cell density in the landscape, and trajectory inference methods uncover transition paths between attractors. These two orthogonal approaches are detailed in the following two sections (**1.3** and **1.4**).

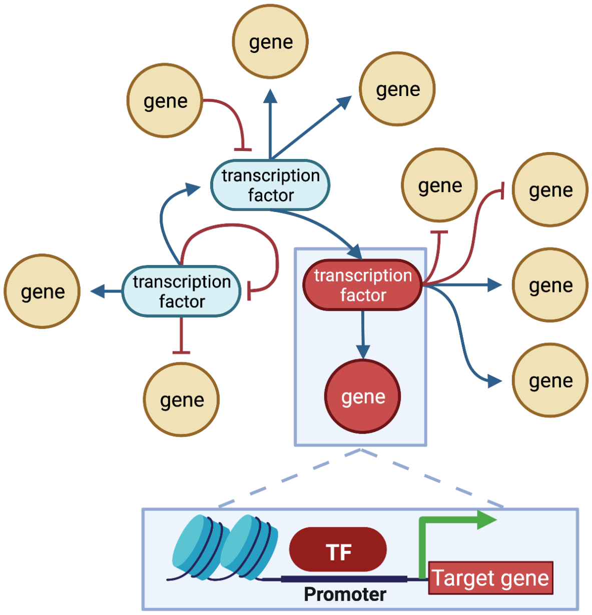


Figure 1.7: A gene regulatory network (GRN) constructed from interactions between DNA-binding TFs and target genes. *Each connection in the GRN represents a physical interaction: the "parent node" is a transcription factor (protein) that binds to the promoter or enhancer region associated with a target gene. When the target gene is also a transcription factor, the connection is part of the GRN; otherwise, if the gene does not make a protein that feeds back into the network, it is often pruned, since the transcription and translation of that gene will not affect the network dynamics.*

1.3. Modeling epigenetic landscapes via gene regulatory networks

1.3.1. Gene regulatory network structure and dynamics

To understand the driving force $\mathbf{F}(\mathbf{x})$ that defines the landscape quasi-potential, it is first important to understand gene regulatory networks (GRNs). A GRN is established by the fact that certain genes encode transcription factors (TF), which are capable of binding to DNA and regulating transcription of genes into RNA (**Figure 1.3**). Because TFs can control the transcription of other TFs (and sometimes themselves), a network of TFs and the genes they regulate can be constructed (**Figure 1.7**). The structure of the GRN for a particular cell is hardcoded in the genome of a cell, as shown in **Figure 1.8 (left)**, since each interaction in the network is a molecular interaction between a DNA-binding protein and the cis-regulatory loci (such as promoter and enhancer regions) for a particular gene (Huang, 2013). On the other hand, the dynamics of the network are described by the collective changes in gene expression over time. The dynamics of a GRN allow for various stable states dependent on the expression of genes in the network (**Figure 1.8, right**). Therefore, the state of the GRN, given by the expression levels of the genes within it, maps to a single location on the epigenetic landscape—the phenotypic state. We note that using the transcriptomic profile of single cells as a proxy for phenotypic state is only one lens through which we view the epigenetic state of a cell.

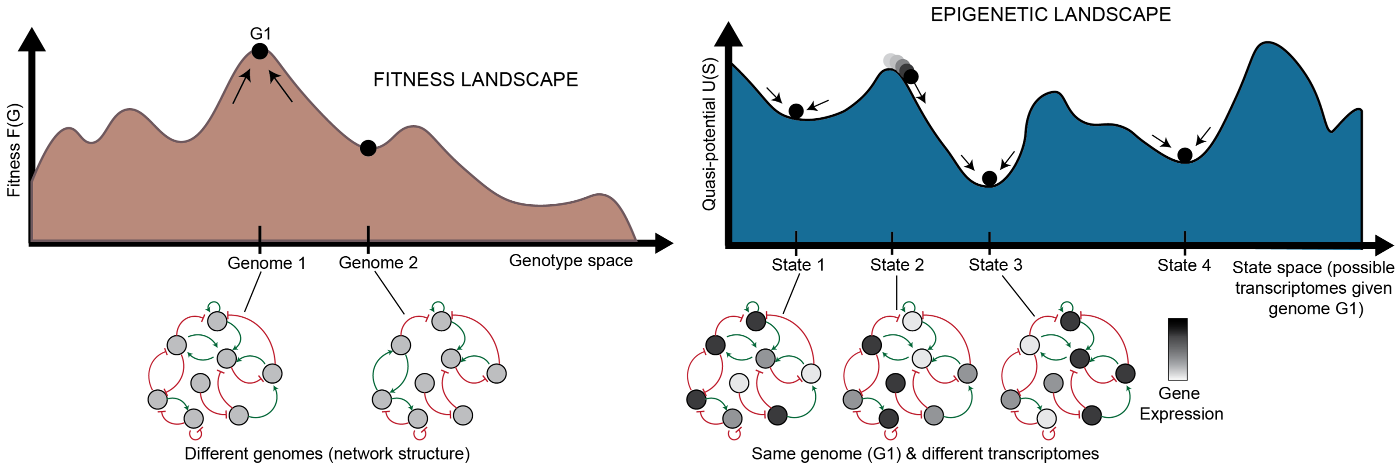


Figure 1.8: Relationship between landscapes and GRNs. *Each state in the fitness landscape (a single genome) is associated with a different GRN structure; mutations can affect the physical interactions between TFs and their target genes, causing the addition or removal of nodes or connections. Each state in the epigenetic landscape, alternatively, has the same genome, and thus the same structure of a GRN. The states in the landscape here represent different states of the same network, where the same nodes in the network are expressed at different levels. The stability of each pattern of expression partially determines the shape of the landscape.*

Quantifying the dynamics of TF binding can be calculated by adapting Hill kinetics to describe the rate at which a target gene is transcribed when regulated by TFs (Hill, 1913). The Hill equation is a sigmoidal function that describes activation (or repression) of a gene as dependent on the concentration of a regulator until it reaches saturation. This is a relatively realistic description of many gene control functions and can be derived directly from the binding of the TF to the promoter site. The dynamics, or the change over time, of each TF in the network can therefore be represented as a function of all "upstream" parent nodes in the network that influence its transcription. The system of such differential equations, where each TF in the network has a corresponding equation for its rate of change, defines the complete dynamics of the system. Based on this system of equations, GRN dynamics are equivalent to the driving force that pushes cells down the gradient of the potential in the landscape. Often, for high-dimensional systems, this system of equations becomes intractable. Instead, we can approximate network interactions, for example, by using Boolean functions or probabilistic dependencies.

1.3.2. Boolean and Bayesian network models of GRNs

In 1969, Stuart Kauffman introduced the idea of Boolean network models for biological systems (Kauffman, 1969a, 1969b). Kauffman posited that, "while finely-graded intermediate levels of gene activity could occur," genes tended to be very active or very inactive (Kauffman, 1971), consistent with switch-like Hill kinetics with a high Hill coefficient. Therefore, it is often useful to idealize the control of gene expression as a binary switch. A Boolean network model suggests that these binary genes interact with one another under Boolean functions, such as AND, OR, or NOT. Boolean logic determines the activity level of each gene given the binary states of its regulating TFs by approximating the Hill equation, turning the smooth, monotonic function into a step function with activation (or repression) threshold of S (Glass and Kauffman, 1973; Kauffman, 1971; Thieffry and Thomas, 1998). An example of a Boolean network is shown in **Figure 1.9**, where each TF in the network is some Boolean function of activation levels of other TFs.

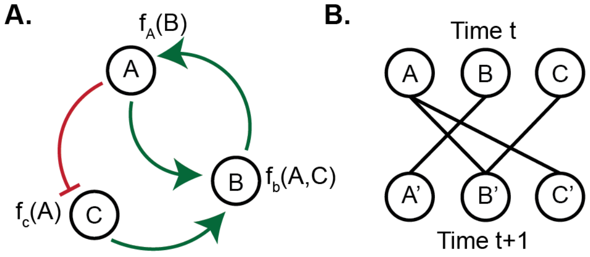


Figure 1.9: Boolean network and wiring diagram. A. *The network has connections between nodes that are Boolean functions of parent nodes. For example, the function for node B might be f_b(A, C) = A AND C, meaning the expression of B is "ON" if A and C are both "ON," and otherwise the expression of B is "OFF."* ***B.*** *The "unraveled" network over two timepoints. The states of all of the nodes (A, B, C) at time t will determine the states (A', B', C') at time t+1.*

Since Kauffman's original idea, several studies have shown the utility of conceptualizing gene regulation as a set of binary genes coupled together through Boolean functions (Albert et al., 2008; Correia et al., 2018; Joo et al., 2018; Masoudi-Nejad et al., 2015; Pomerance et al., 2009; Saadatpour and Albert, 2013; Steinway et al., 2015; Wooten et al., 2019; Yachie-Kinoshita et al., 2018; Zhou et al., 2016b). While a Boolean approximation for transcriptional regulation is realistic for many biological systems, some genes are regulated by multiple TFs in a manner that does not use Boolean logic. For example, Kalir and Alon (2004) showed that gene regulation in an *E. coli* network of flagella biosynthesis follows a summation function (SUM), rather than Boolean logic gates (AND, OR, and NOT). Several studies have shown other functions, including complex functions with many inputs, are also possible (Beer and Tavazoie, 2004; Istrail and Davidson, 2005; Yuh et al., 1998).

In order to understand systems with non-Boolean gene regulatory functions, probabilistic methods of network inference, known as probabilistic graphical models (PGMs), may be used. These models have multiple advantages over Boolean approaches. For example, they can infer non-linear relationships between TFs, such that the rule of interaction is not required *a priori* to have a particular form such as a Boolean function. One such PGM, known as Bayesian network inference, considers a GRN to be a network (or graph) where each directed edge represents the probabilistic dependence among genes. The goal in GRN inference is always two-fold: (1) to determine the structure of the network (how the TFs interact) and (2) to quantify the causal relationships between connected nodes so that the future state of the system can be predicted based on its current state. In Bayesian networks, these causal relationships are completely encoded in the joint probability distribution of the network, which details the probability of each state for each TF conditioned on the state of its regulators:

$$P(X_1, X_2, \ldots, X_N) = \prod_{i=1}^{N} P(X_i | X_{pa(i)})$$

where Pa(i) is the set of parents of the node X_i for all X_i in the network. PGMs are more phenomenological than Hill kinetics or Boolean network modeling, but they can mine information from transcriptomic data—for example, RNA-seq profiles for the TFs in the network—without assumptions about the relationships between TFs.

Regardless of the limitations or assumptions of network inference algorithms, these methods require biological data to fully characterize a system. Transcriptomics data are often used, sometimes in combination with other types of epigenomic or proteomic information (Duren et al., 2017; Langfelder and Horvath, 2008; Liu et al., 2016; Margolin et al., 2004; Ramirez et al., 2017; Wooten et al., 2019). Today, single-cell RNA-sequencing (scRNA-seq) is commonly used to obtain a more granular picture of transcriptional regulation and stable phenotypes in a system than bulk sequencing data can provide. Several computational algorithms have been developed to infer single-cell dynamics based on Boolean, Bayesian, or other types of networks (Aibar et al., 2017; Chen and Mar, 2018; Pratapa et al., 2020; Sanchez-Castillo et al., 2017; Sande et al., 2020).

Top-down, phenomenological approaches for modeling the epigenetic landscape can also utilize scRNA-seq data directly to find empirical patterns of expression. Because intratumoral heterogeneity and plasticity are relevant to acquire resistance in SCLC, it is important to determine how cells change in phenotype in various contexts. These top-down approaches work towards the long-term goal of personalizing SCLC treatment by providing a framework for understanding plasticity in an individual patient's tumor.

1.4. Modeling epigenetic landscapes via single-cell dynamics

Single-cell sequencing methods have paved the way for data-driven approaches to quantifying plasticity. While classical dynamical systems modeling—i.e., modeling a GRN that determines a quasi-potential landscape—has the advantage of being predictive, it can be difficult or impossible to model the complete dynamics of a high-dimensional system. Alternatively, it is possible to use a data-driven, bottom-up approach by modeling single-cell dynamics as a Markovian process, which can identify transition paths heuristically from scRNA-seq data.

Borrowing once again from physics, a drift-diffusion equation can model the change in cell density for a given region of gene expression space (or, analogically, the phenotypic landscape):

$$\frac{\partial c}{\partial t} = -\nabla(cv) + Rc$$

where c is cell density of a given region of gene expression space, R describes the rate of accumulation and loss due to cell proliferation, death, and movement through the region, and v is the net average velocity (Weinreb et al., 2018). With additional assumptions, we can model the velocity as related to the deterministic average velocity field (due to the epigenetic landscape, for example) and a stochastic component related to diffusion. This velocity field may be calculated heuristically from pseudo-temporal information using trajectory inference methods and can predict cell-state transitions in the epigenetic landscape (Qiu et al., 2022). Furthermore, drift-diffusion modeling of cell dynamics along a high-dimensional manifold in gene expression space can be used to infer dynamics through a Markov chain, with defined transition probabilities between cell states (Weinreb et al., 2018).

1.4.1. Trajectory inference and RNA velocity

Trajectory inference algorithms also aim to understand changes in cell density by ordering cells along a trajectory based on transcriptomic similarity, empirically determining transition paths in the system (Guo et al., 2016; Haghverdi et al., 2016; Herring et al., 2018; Qiu et al., 2017; Saelens et al., 2018; Trapnell et al., 2014; Welch et al., 2016; Wolf et al., 2019). These trajectory inference algorithms tend to search for an underlying manifold of the data to delineate graph-based trajectories. By interrogating the structure of the single cell data in gene expression space, multifurcations, trees, and other graph structures can be identified. While these methods are unbiased and often unsupervised, they tend to require identification of a "root cell" to determine the directionality of transitions, as multiple trajectories could be explained by the same graph structure.

Some methods utilize time-series data to determine directionality by optimal transport-based algorithms (Kimmel et al., 2019; Marjanovic et al., 2020; Schiebinger et al., 2019). Because scRNA-seq is a destructive method, the same single cells cannot be sequenced over time. Optimal transport-based methods overcome this by inferring "temporal couplings" across timepoints to determine the most likely phenotypic "descendants" of each cell at later timepoints. Ultimately, lineage tracing provides a benchmark for interrogating trajectories, as cell lineages across timepoints are identified via "barcodes," thereby linking cell state in early timepoints to cell fate in later timepoints (Griffiths et al., 2018; Wagner and Klein, 2020; Wang et al., 2021).

In 2018, a novel approach to trajectory inference was developed based on RNA splicing dynamics (La Manno et al., 2018). By fitting an ordinary differential equation (ODE) model of RNA transcription, splicing, and degradation, La Manno et al. discovered that it was possible to infer short-term dynamics on a cell-by-cell basis (**Figure 1.13**). RNA velocity infers a steady-state ratio of unspliced to spliced counts of RNA on a gene-by-gene basis to fit the ODE model parameters, such as the degradation rate of the mRNA. As shown in **Figure 1.12**, an increase in transcription of a particular gene is followed by a slow increase in unspliced RNA, followed by a delayed increase in spliced RNA. Therefore, by then comparing the unspliced and spliced counts of a gene in each cell in this model, it is possible to determine the future state of each cell.

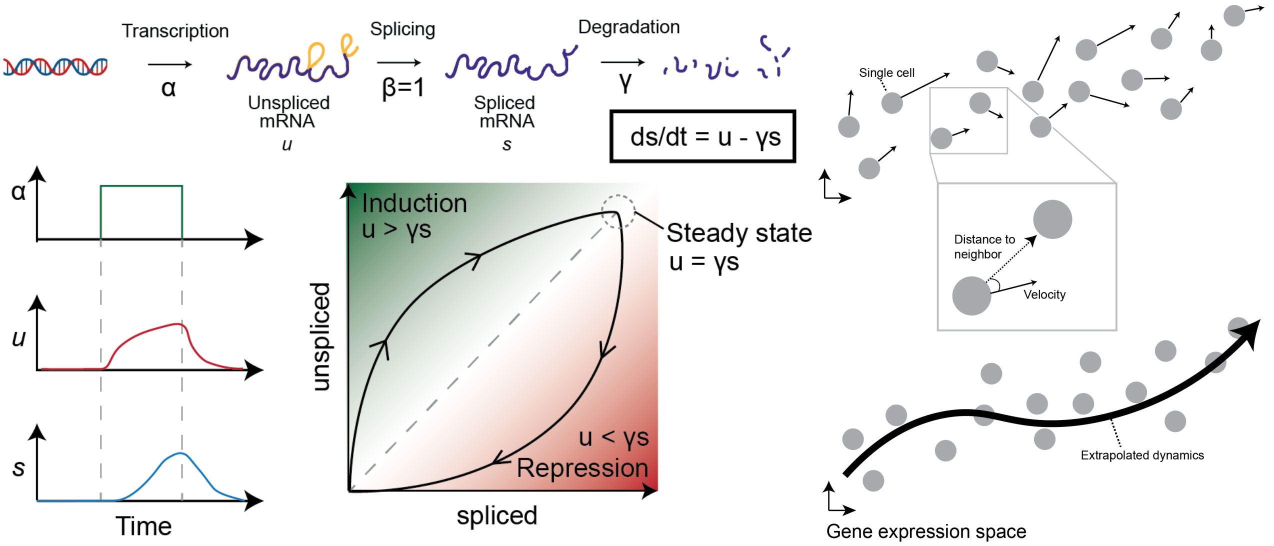


Figure 1.10: RNA velocity model. *By modeling transcription, splicing, and degradation of RNA as ODEs, we can determine the steady state proportion of unspliced and spliced RNA and infer dynamics of single cells. An increase in transcription leads to an increase in unspliced and then spliced RNA, with lag time. This difference helps to determine whether a snapshot proportion of unspliced and spliced counts of a particular gene is increasing (induction) or decreasing (repression). Velocity vectors in gene expression space are calculated for each individual cell. By comparing each velocity vector to the distance to neighboring sampled cells, we can predict the probability of the cell transitioning to other states (defined by sampled cells). This allows us to generate a Markov chain model and infer dynamics through the single cell data.*

The timeframe for dynamic predictions is on the order of a few hours, similar to the average splicing rate. However, RNA velocity can be extrapolated to longer timeframes by considering the relationship between a cell's velocity vector—i.e., the directionality and magnitude of its inferred change in gene expression—and the location of neighboring cells (**Figure 1.12**). These extrapolated dynamics can be used to make predictions about the future state of cells near the beginning of the trajectory. Because this method does not rely on multiple sampled timepoints or prior knowledge about the "root" cell of a trajectory, it is optimal for understanding the dynamics of systems for which a temporal series of samples is not possible, such as tumor dynamics from single biopsies. Together, these analysis methods can uncover an empirical epigenetic landscape by defining stable phenotypes and transition paths in scRNA-seq data sampled from various cancer systems, including human biopsies, to complement or replace quantification of GRN dynamics.

1.5. Outline of dissertation

In this dissertation, I detail the results of my investigation into the relationships between phenotypic heterogeneity, plasticity, and SCLC. Chapter 2 reports my work on the gene regulatory network that defines SCLC phenotypes. I use clustering methods to delineate phenotypic subtypes of SCLC and identify overexpressed gene programs in cell lines, mouse models, and human tumors. The development of a gene regulatory network that defines these subtypes allows for characterization of driving transcription factors that stabilize (or destabilize) each subtype, suggesting methods of perturbation that may reprogram a cell from one phenotype to another. Chapter 3 follows up this work by combining bioinformatics and evolutionary dynamics approaches to investigate why multiple stable subtypes of SCLC may arise within a single tumor. I quantify phenotypic transitions within these heterogeneous tumor populations by developing a novel metric of plasticity and identify subpopulations that may drive resistance to chemotherapy. Chapter 4 then applies the methods developed in these two chapters to elucidate the role of network dynamics in a variant mouse model of SCLC and the role of plasticity in evading targeted therapy. In total, this work uses mathematical modeling and bioinformatic methods to further our understanding of the relationship between cell identity, system dynamics, and treatment resistance in SCLC.

Chapter 2.

Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers[1]

2.1. 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, breast, and lung cancer (Hauschild et al., 2012; Robert et al., 2017; Travis et al., 2011). 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 **Figure 2.1**.

[1] Adapted from Wooten, D. J.*, Groves, S.M.* et al. Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers. Plos Comput Biol 15, e1007343 (2019). *authors contributed equally to this work.

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 (Newman et al., 2015) 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 the efficacy of treatment.

We applied this workflow to Small Cell Lung Cancer (SCLC), in which genetic aberrations cannot fully distinguish subtypes or point toward a targeted therapy (George et al., 2015). 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 strategies. These observations indicate the existence of a complex landscape of SCLC phenotypes that may form a tumor microenvironment robust to perturbations and treatment (Lim et al., 2017; Tammela et al., 2017). 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 considering the dynamics of underlying GRNs, could make systems-level predictions that more accurately reflect the occurrence and transdifferentiation of coexisting subtypes within SCLC tumors.

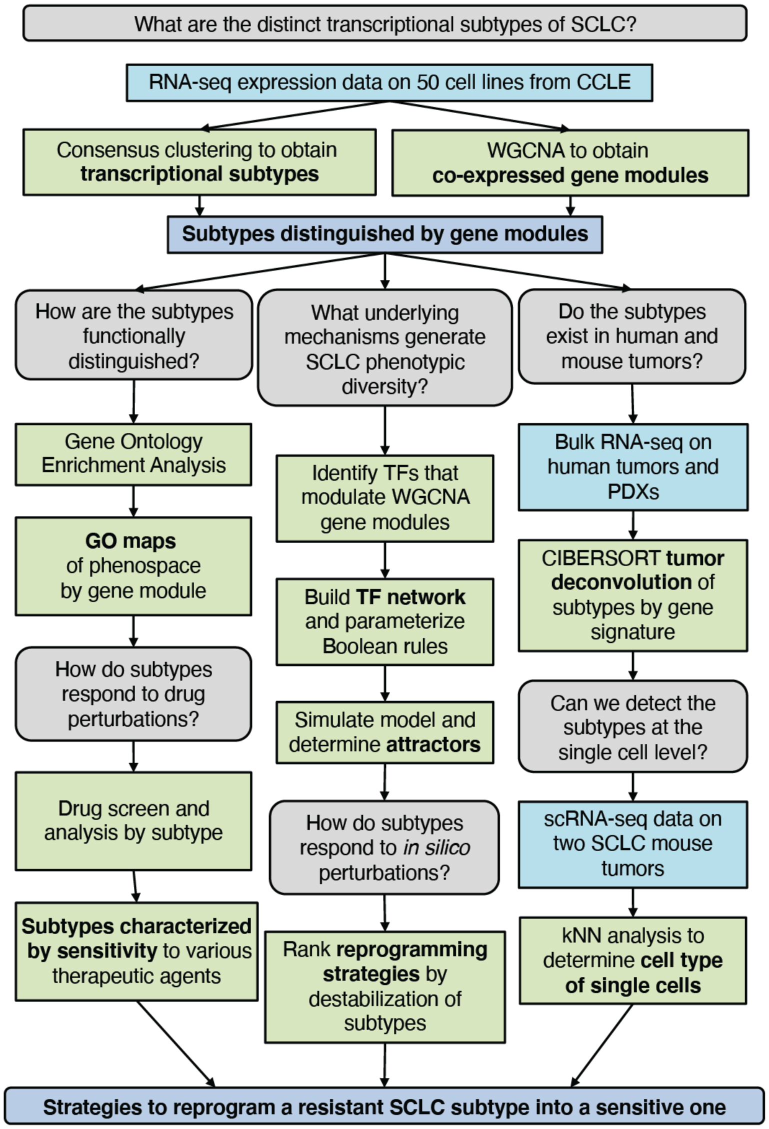

What are the distinct transcriptional subtypes of SCLC?
RNA-seq expression data on 50 cell lines from CCLE
Consensus clustering to obtain transcriptional subtypes
WGCNA to obtain co-expressed gene modules
Subtypes distinguished by gene modules
How are the subtypes functionally distinguished?
What underlying 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

Figure 2.1: 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.*

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 previously identified subtypes (ASCL1+ NE, a NEUROD1+ NE variant, and a YAP1+ non-NE variant). 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.

2.2. Results

2.2.1. Consensus clustering uncovers new SCLC variant phenotype

Recently, the occurrence of variant SCLC subtypes has been reported (Mollaoglu et al., 2017; Sos et al., 2012; Udyavar et al., 2017). Given the translational value of defining subtypes, a more global approach to comprehensively define SCLC subtypes would be desirable. To this end, we devised the workflow described in **Figure 2.1**. First, we applied Consensus Clustering (Monti et al., 2003) to RNA-seq gene expression data from the 50 SCLC cell lines in the CCLE (Barretina et al., 2012). 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 (Udyavar et al., 2017), 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 $k \in \{2, 20\}$ (**Figure 2.2A**). 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 criteria such as the tracking plot and delta area plot (**Figure 2.2B**), both $k = 2$

and k = 4 gave well-defined clusters. Since recent literature suggests that more than two subtypes are necessary to describe SCLC phenotypic heterogeneity, we selected k = 4 for further analyses.

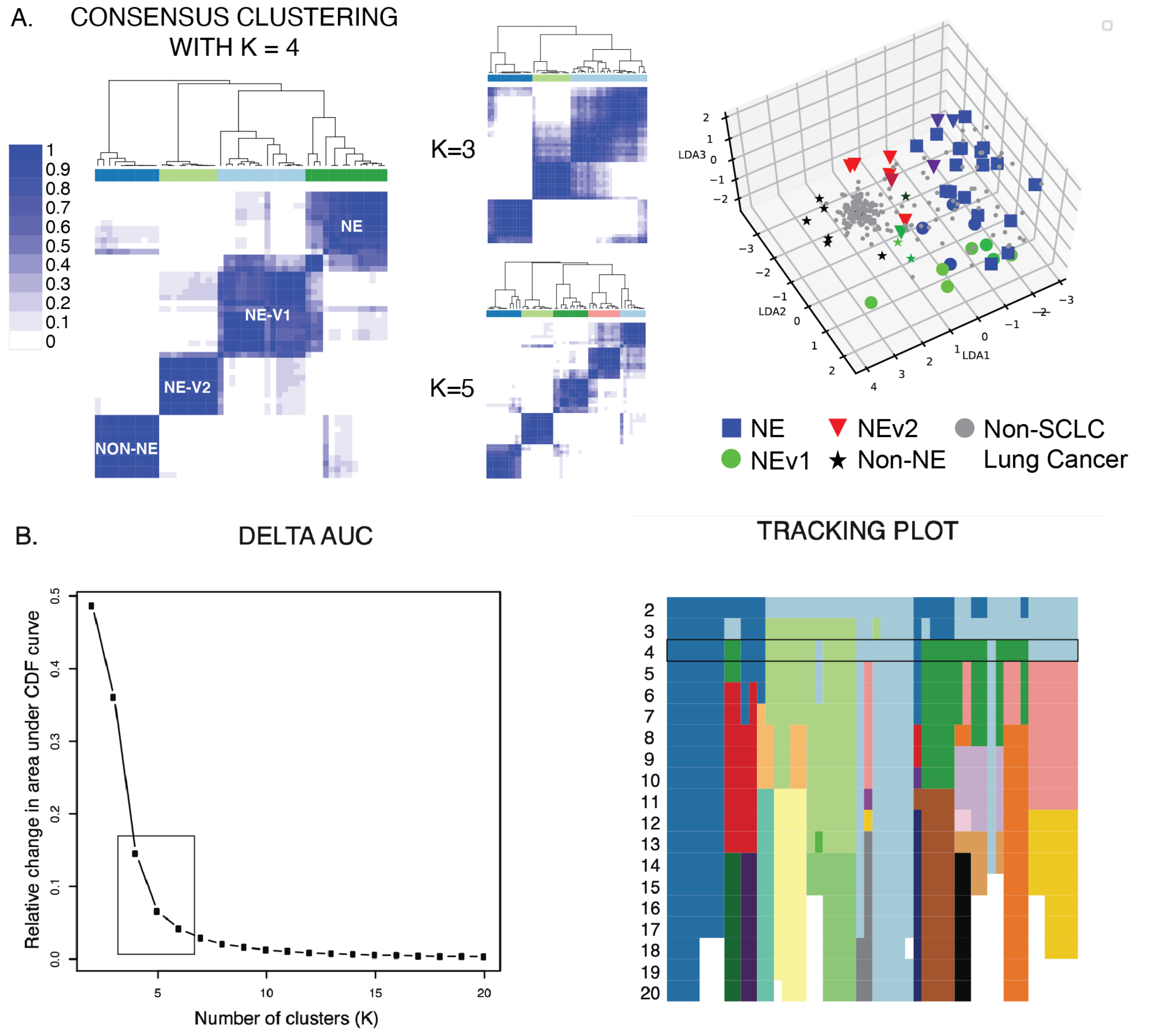


Figure 2.2: Consensus clustering and WGCNA of 50 SCLC cell lines reveal four subtypes differentiated by gene modules. A. *Consensus clustering with k = 4 gives most consistent clusters. K = 3 and K = 5 add complexity without a corresponding increase in accuracy. LDA plot shows separation of 4 clusters, with non-SCLC cell lines falling near non-NE cell lines.* ***B.*** *The delta area plot shows the relative change in the area under the CDF curve. 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). 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.*

To align the 4-cluster classification (**Figure 2.3**) with existing literature, we considered well-studied biomarkers of SCLC heterogeneity across the clusters. Three of the four consensus clusters could be readily matched to subtypes previously identified with 2 to 5 biomarkers: the canonical NE subtype (SCLC-A) (Borromeo et al., 2016; Rudin et al., 2019), an NE variant subtype (referred to here as NEv1, corresponding to SCLC-N in Mollaoglu et al. (2017), and a non-NE variant subtype (SCLC-Y) (Lim et al., 2017; Poirier et al., 2015). The fourth cluster (referred to here as NEv2) could not be easily resolved using a 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 (2016), yet the expression of a single biomarker, HES1, would suggest this subtype falls outside of the NE subtype according to Lim et al (2017). Discrepancies like this drove us to consider broader patterns of gene expression, rather than a limited number of biomarkers, to characterize each subtype.

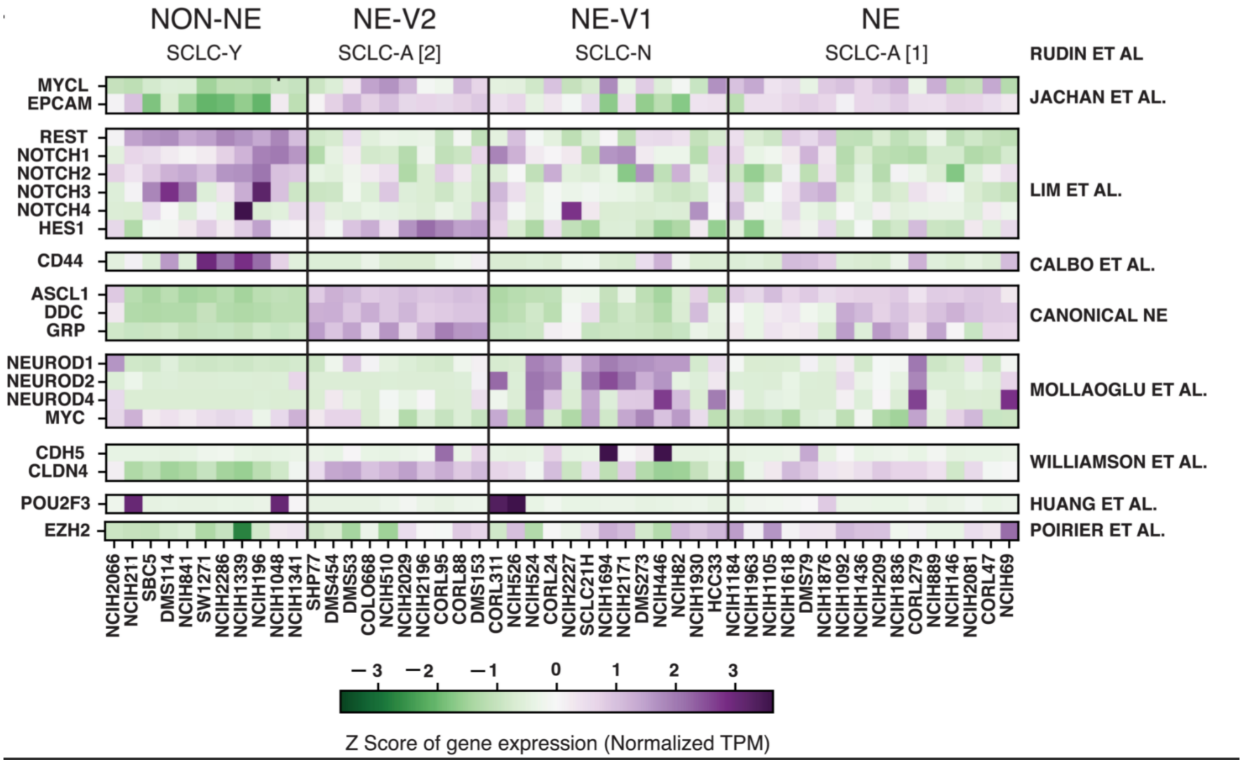


Figure 2.3: SCLC Biomarkers. *Current biomarkers in the field of SCLC are able to distinguish between three of the subtypes; The fourth subtype, NEv2, is not separable from NE using markers from SCLC literature.*

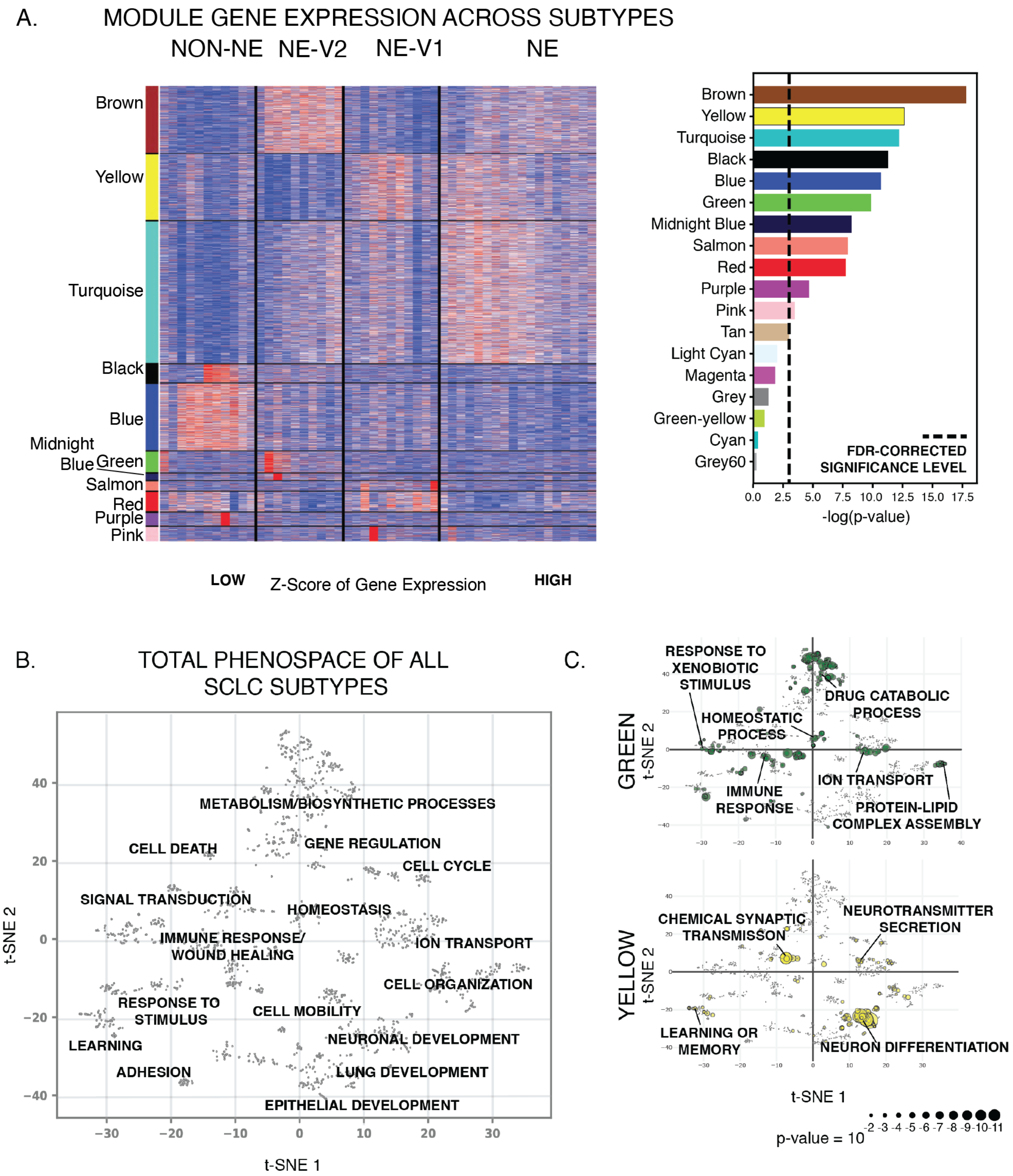


Figure 2.4: SCLC subtypes can be distinguished by gene expression patterns. *Transcriptional patterns that distinguish the four subtypes are captured in WGCNA analysis. Gene modules by color show patterns of expression that are consistent across the subtypes. Only modules that significantly distinguish 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.*

2.2.2. 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) (Langfelder and Horvath, 2008) 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 (**Figure 2.4A**, 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 (Kamburov et al., 2013), which resulted in a combined total of 1,763 statistically enriched biological processes (**Figure 2.4B**).

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 (Subramanian et al., 2005) (**Figure 2.4C and 2.5**). 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” (**Figure 2.4C**). Evidently, such functions may be maintained in NE and NE-v1 subtypes, as reflected by the frequent occurrence of paraneoplastic syndromes in SCLC patients (Paraschiv et al., 2015). 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 (**Figure 2.5**).

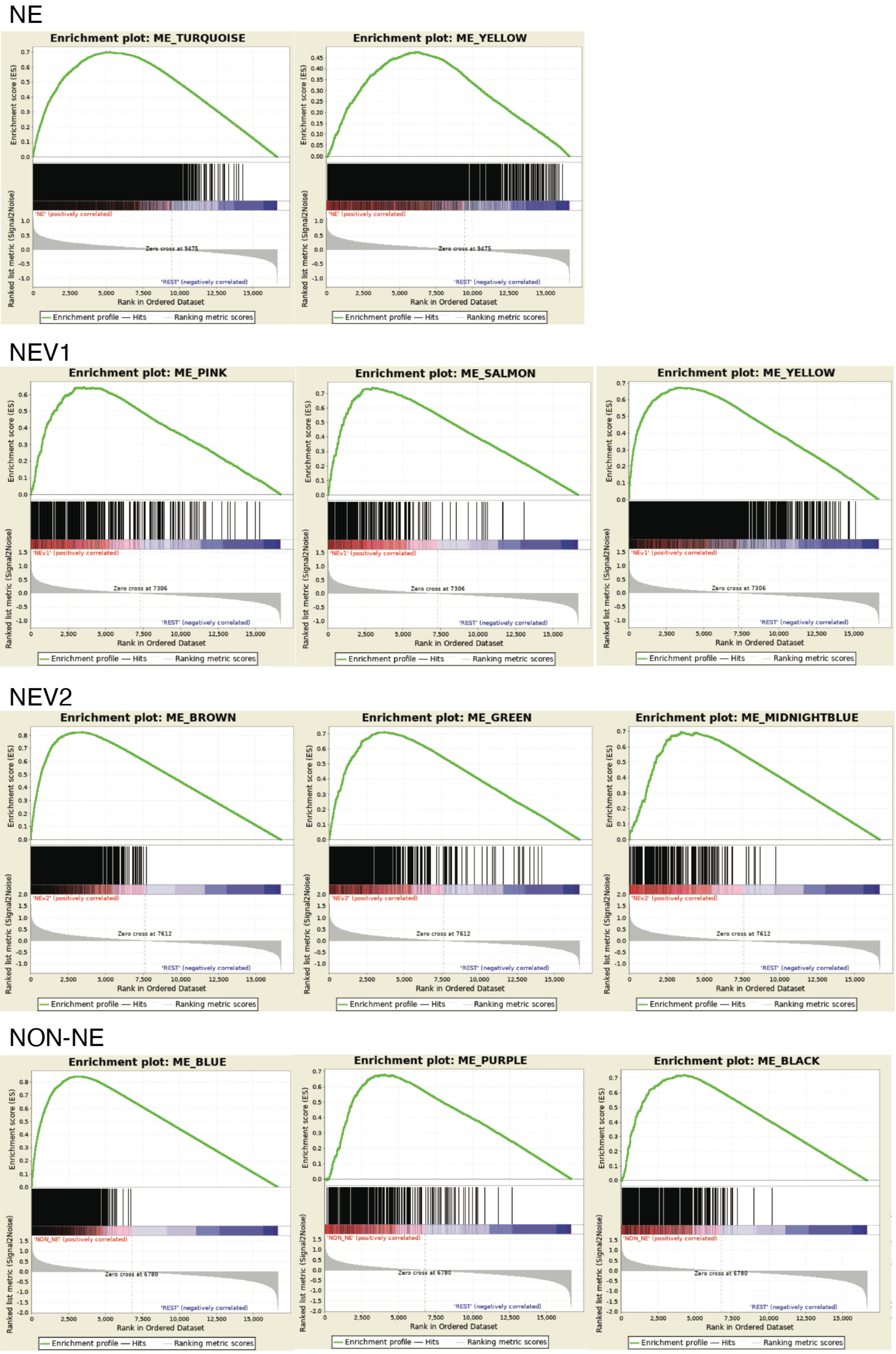


Figure 2.5: Enrichment of WGCNA gene modules by subtype using GSEA.

Genes within the brown, midnight blue, and green modules are upregulated in the NEv2 phenotype (**Figure 2.4A and 2.5**). 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 (**Figure 2.4C**). Enrichment of these GO terms suggests 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 (**Figure 2.4B**), we used the GOSemSim package (Yu et al., 2010) 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) (Van Der Maaten and Hinton, 2008). 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. 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.

2.2.3. 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 (Polley et al., 2016). We used the Activity Area (AA) metric as a measure of the resultant dose-response curves. The drugs

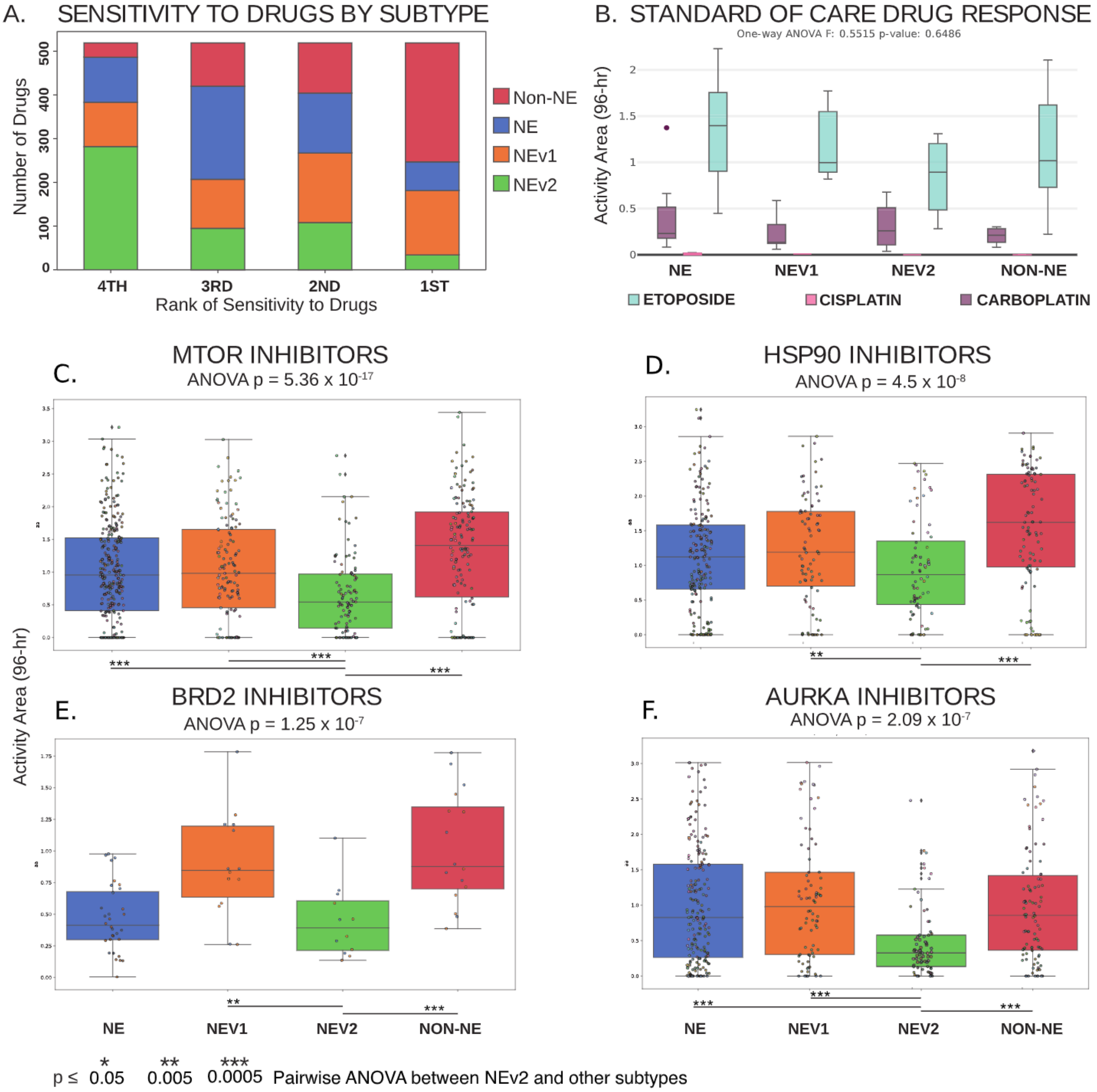


Figure 2.6: Differential response of SCLC subtypes to a wide variety of oncology drugs and investigational agents. A. *Ranked sensitivity of subtypes across 526 compounds. NEv2 is least sensitive for over half of the drugs tested.* ***B****. No significant differences can be seen in response to etoposide 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.*

were analyzed individually and clustered by common mechanism of action and target type, and the cell lines were grouped by the four subtypes.

As shown in **Figure 2.6A,** cross 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%). 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 (**Figure 2.4C**), even though the subtypes do not show differential response to the standard of care (etoposide and platinum-based agents, **Figure 2.6B**). In particular, mTOR inhibitors are a class of compounds to which NEv2 was significantly more resistant (**Figure 2.6C**). PI3K pathway mutations have previously been implicated as oncogenic targets for SCLC, as about a third of patients show genetic alterations in this pathway (Umemura et al., 2014). Among the four subtypes, NEv2 is also the least sensitive to AURKA, B, and C inhibitors (AURKA shown); TOPO2 inhibitors; and HSP90 inhibitors (**Figure 2.6D-F**). These results have implications for interpreting expected or observed treatment response with respect to tumor heterogeneity in individual patients.

2.2.4. 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 (Newman et al., 2015) 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. (2015) to specify the 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 (Yeo and Guan, 2017), CIBERSORT predicted that a majority of tumors were comprised of all four subtype signatures, in varying proportions across tumor samples (**Figure 2.7A**). We then analyzed the patient/cell-derived xenograft models (PDXs/CDXs) developed by Drapkin et al. (2018), and the tumors also showed vast differences across samples (**Figure 2.7B**). 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 an increase in NEv2 after treatment. Similar observations of phenotypic changes over treatment time courses, made in breast cancer patients (Yeo and Guan, 2017) have recently been explained in the context of a mathematical model of epithelial to mesenchymal transition (EMT) (Bocci et al., 2019). It is possible that the tumor composition changes we observe may also be explained by molecular level and/or cell population-level models (Harris et al., 2019). Overall, the high variance in proportions of each subtype suggests a high degree of intertumoral, as well as intratumoral, dynamic heterogeneity and plasticity.

We also investigated phenotypic patterns in mouse tumors from two different sources to determine whether human SCLC subtype signatures are conserved across species (see Methods) (Mollaoglu et al., 2017; Schaffer et al., 2010). 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 (**Figure 2.7Ci**). Of note is the lower percentage of non-NE cells found in each tumor in **Figure 2.7Ci**; we suspect this is due to a filtering step before sequencing (see Methods), as the non-NE subtype signature is more similar to tumor-associated immune cells in an unfiltered tumor population. The second mouse model shown in **Figure 2.7Cii** was generated with Myc overexpression (double knockout of Rb1 and Tp53, and

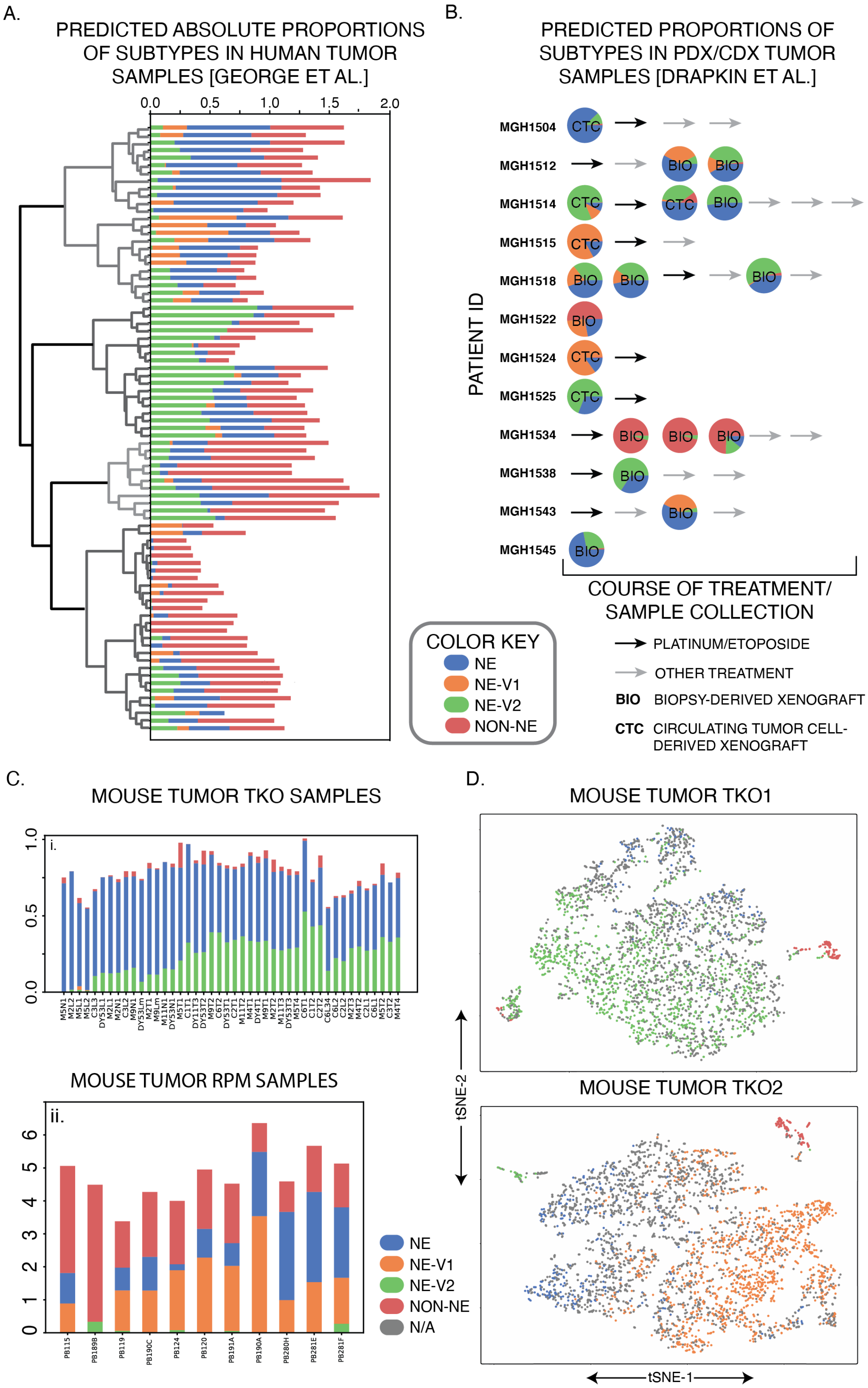

A.
PREDICTED ABSOLUTE PROPORTIONS OF SUBTYPES IN HUMAN TUMOR SAMPLES [GEORGE ET AL.]
0.0
0.5
1.0
1.5
2.0
B.
PREDICTED PROPORTIONS OF SUBTYPES IN PDX/CDX TUMOR SAMPLES [DRAPKIN ET AL.]
PATIENT ID
MGH1504
CTC
MGH1512
BIO
BIO
MGH1514
CTC
CTC
BIO
MGH1515
CTC
MGH1518
BIO
BIO
BIO
MGH1522
BIO
MGH1524
CTC
MGH1525
CTC
MGH1534
BIO
BIO
BIO
MGH1538
BIO
MGH1543
BIO
MGH1545
BIO
COURSE OF TREATMENT/ SAMPLE COLLECTION
COLOR KEY
NE
NE-V1
NE-V2
NON-NE
PLATINUM/ETOPOSIDE
OTHER TREATMENT
BIO BIOPSY-DERIVED XENOGRAFT
CTC CIRCULATING TUMOR CELL-DERIVED XENOGRAFT
C.
MOUSE TUMOR TKO SAMPLES
i.
1.0
0.5
0.0
M5N1
M2L2
M5L1
M5L2
C3L3
DY53L1
M2L1
M2N1
C3L2
M9N1
DY53Lm
M2T1
M9Lm
M11N1
DY53N1
M5T1
C1T1
DY11T3
DY53T2
M9T2
C6T2
DY53T1
C2T1
M11T2
M4T1
DY4T1
M9T1
M2T2
M11T3
DY53T3
M5T4
C6T1
C1T2
C2T2
C6L34
C6L2
C2L2
M2T3
M4T2
C2L1
C6L1
M5T2
C3T2
M4T4
MOUSE TUMOR RPM SAMPLES
ii.
6
5
4
3
2
1
0
PB115
PB189B
PB119
PB190C
PB124
PB120
PB191A
PB190A
PB280H
PB281E
PB281F
NE
NE-V1
NE-V2
NON-NE
N/A
D.
MOUSE TUMOR TKO1
MOUSE TUMOR TKO2
tSNE-2
tSNE-1

Figure 2.7: Computational evidence for existence of subtypes in human and mouse 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. (2018). Ci. TKO (Rb1, Tp53, P130 floxed) mouse tumors showing a high proportion of NE and NEv2 subtypes.* ***C.*** *Ci. As described in Mollaoglu et al. (2017), these mouse tumors were generated by crossing Rb1 fl/fl Trp53 fl/fl (RP) animals to knockin Lox-Stop-Lox (LSL)-MycT58AIRES-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 proportions 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.* ***D.*** *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.*

overexpression of Myc) as reported previously (Mollaoglu et al., 2017). 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 knockout tumors in **Figure 2.7Ci**, 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 **Figure 2.7D**, 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 (**Figure 2.7D**). Tumor 1 has a large proportion of the NEv2 subtype, corresponding to the tumors in **Figure 2.7Ci.** In contrast, tumor 2 has a large NEv1 subpopulation, similar to the tumors in **Figure 2.7Cii**. 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.

2.2.5. 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 (Stransky et al., 2015) We subsetted these data to the 50 SCLC cell lines with matching CCLE RNA-seq data, and using MutSigCV (Lawrence et al., 2013), we found 29 genes (**Figure 2.8**) 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 (**Figure 2.8**), suggesting alternative sources of heterogeneity.

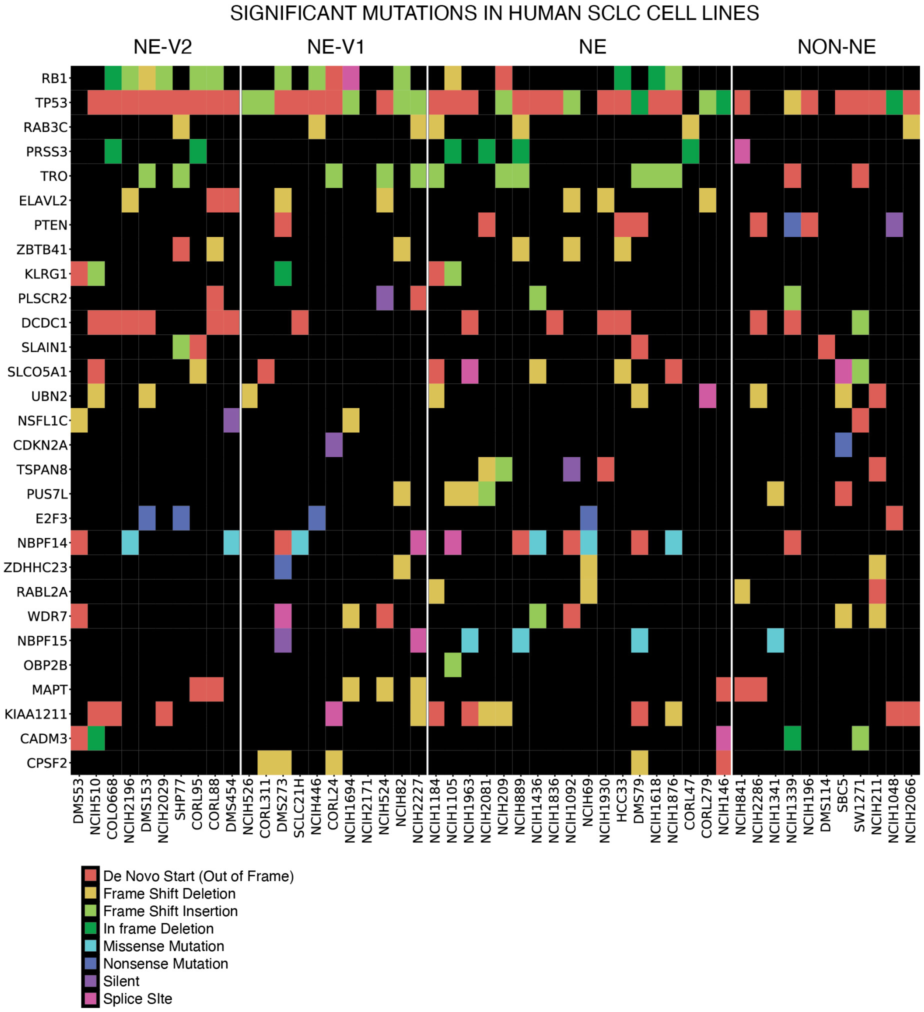


Figure 2.8: Significant mutations across subtypes. *Significantly mutated genes across 50 SCLC cell lines, as determined by MutSigCV, ordered by significance. As expected, significant mutations were found in both the Rb1 and Tp53 genes. Inspection by eye shows that no significant mutations can distinguish 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 ≤ 0.5 shown.*

2.2.6. 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 (Udyavar et al., 2017). That analysis suggested the existence of additional SCLC subtypes but did not specify corresponding attractors (Udyavar et al., 2017). 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 (**Figure 2.4B**) 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 (Udyavar et al., 2017), as detailed in **Figure 2.9A**.

Following the procedure we previously used (Udyavar et al., 2017)., 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 At+1 = Bt or Ct, 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 (Albert et al., 2008), or majority rules (Albert et al., 2008; Font-Clos et al., 2018). 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. 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. Using the network in **Figure 2.9A**, neither of these approximations stabilized attractors corresponding to either the NEv1 or NEv2 phenotypes, suggesting that the regulatory rules governing stability of these phenotypes are more complex.

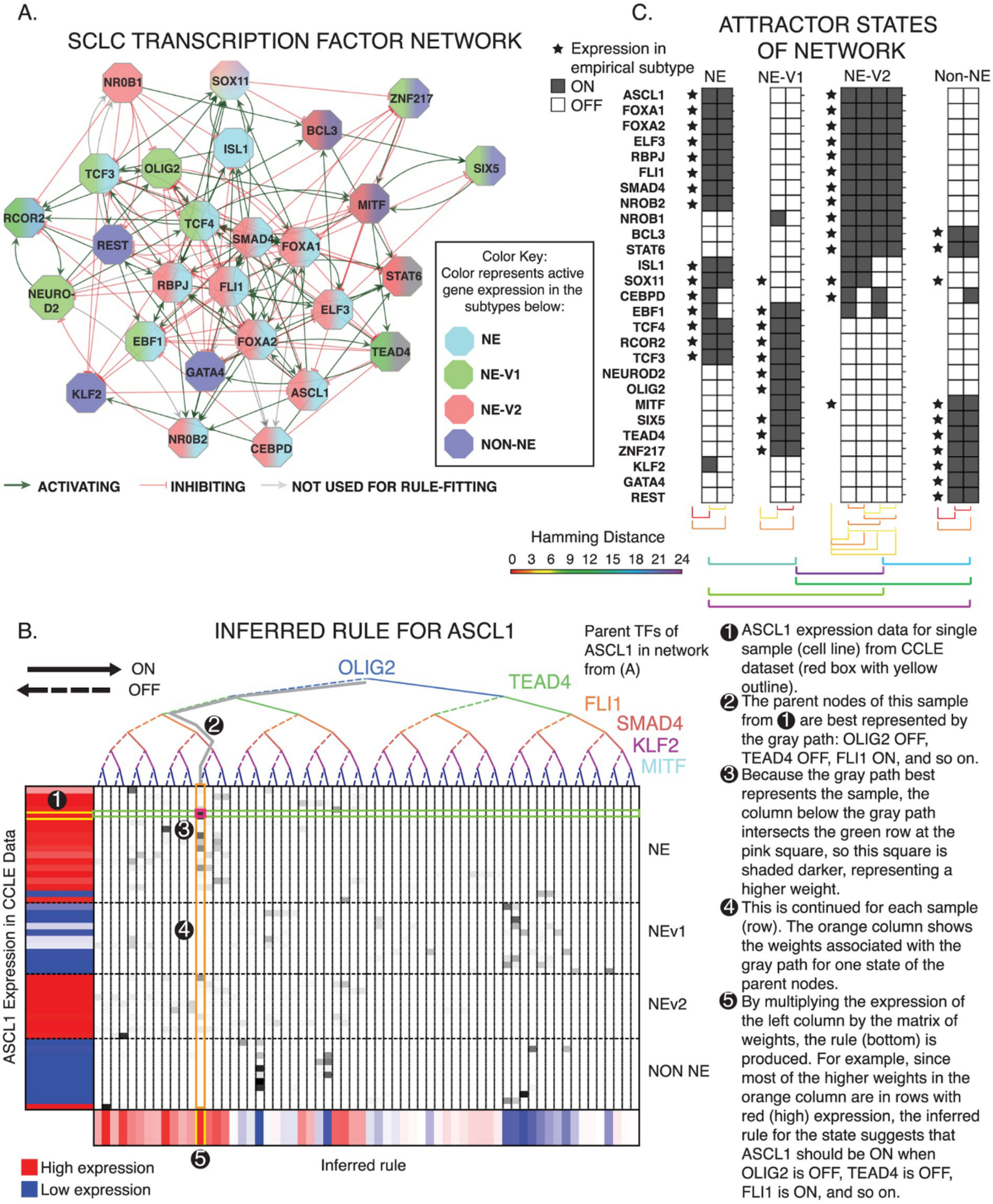
To address this complexity, we developed BooleaBayes, a method to infer logical relationships in gene regulatory networks (**Figure 2.9B**) using gene expression data, by enhancing confidence in Boolean rules via a Bayes-like adjustment approach (see Methods). BooleaBayes leverages sparsity (the in-degree 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 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 **Figure 2.9A**. As an example, **Figure 2.9B** shows the rule fitting for one node, ASCL1. Cross-validation suggested BooleaBayes did not overfit the data. We simulated the dynamics of the Boolean network using a general-asynchronous update scheme (Albert et al., 2008). 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 (**Figure 2.9C**), which 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.

These 10 pseudo-attractor states each correlated with, and could be assigned to, one of the 4 SCLC subtypes (stars in **Figure 2.9C**); this indicates 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.

Figure 2.9: TF network simulations reproduce subtypes as attractors. (Next page) A. *Regulatory network of differentially expressed TFs from each of the 11 co-expressed gene modules in Figure 2.4. 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.*



	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		□	□		□	□		□	□	□	□	★	■	■

❶ ASCL1 expression data for single sample (cell line) from CCLE dataset (red box with yellow outline).

❷ The parent nodes of this sample from ❶ are best represented by the gray path: OLIG2 OFF, TEAD4 OFF, FLI1 ON, and so on.

❸ 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.

❹ 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.

❺ 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.

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

To quantify the baseline stability of the steady states in **Figure 2.9C**, 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 (**Figure 2.10**, Algorithm 2). We chose a 4-TF neighborhood to account for the models' greatest intra-subtype attractor variability (**Figure 2.9C**, 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 (**Figure 2.9C**). 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 (**Figure 2.10A**). We defined (de)stabilization as the percent decrease or increase of the average number of steps under perturbation relative to the unperturbed reference (**Figure 2.10B**). For example, either activation of GATA4 or silencing FOXA1 are predicted to destabilize both the NE and NEv2 subtypes (**Figure 2.10B-C**).

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 (Lim et al., 2017)), TEAD4 (non-NE), ISL1 (NE), and TCF4 (NEv1). TEAD4 is a downstream mediator of YAP1 action, which has been previously identified as a possible phenotypic modulator in a subset of SCLC cell lines (Horie et al., 2016); 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.

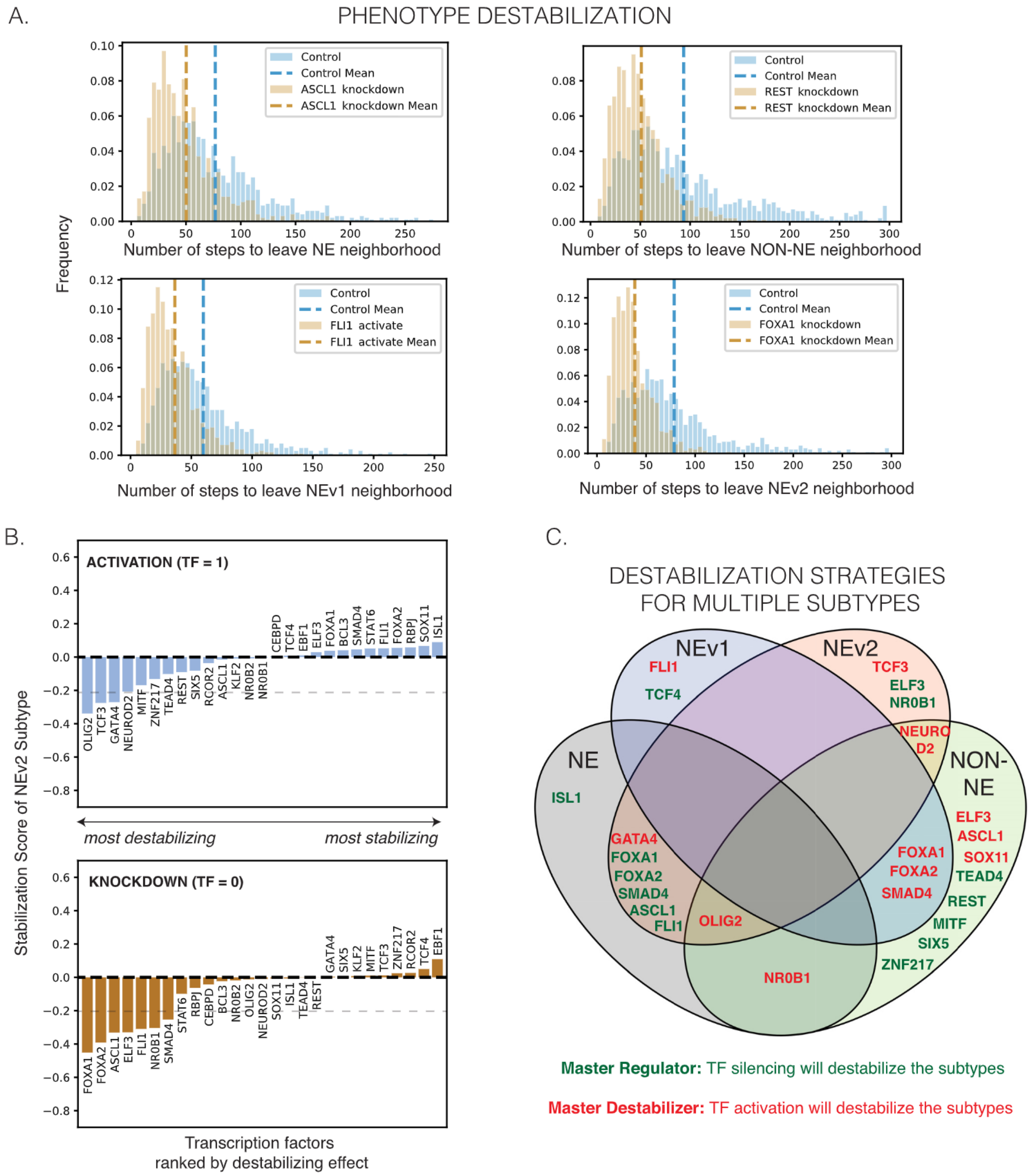


Figure 2.10: Destabilization of subtypes by perturbation to network. A. *Random walks starting from the attractors in Figure 2.9C will eventually leave the start state due to uncertainty in the Boolean rules. Control histogram shows how many random steps are required to reach a state with a Hamming distance ≤ 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 indicate destabilizing perturbations, while positive indicates increasing stability. Results are shown for 1000 iterations starting from NEv2. 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.*

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 (**Figure 2.10C**). 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.

2.3. 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 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 (Rudin et al., 2019). Significantly, we find that one (NEv2) has not been described previously and 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 (Rudin et al., 2019).

Tumor deconvolution by CIBERSORT and scRNA-seq data indicate that a large proportion of human and mouse tumors comprise more than one subtype (**Figure 2.7**). While MutSigCV mutational analysis did not find any significant differences in mutated genes between subtypes

(**Figure 2.8**) 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 (Bocci et al., 2019). 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 (**Figure 2.6 and 2.7**) 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 (Jolly et al., 2018; Kröger et al., 2019; Lu et al., 2013). 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 (Hinohara and Polyak, 2019). 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 (Jolly et al., 2018; Sharma et al., 2010). 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 (**Figure 2.7**) may be responsible for acquired drug resistance in this patient. However, this study was underpowered 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 the 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 (Lang et al., 2014), 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 (Hong et al., 2015; Jia et al., 2015).

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. (2016) 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. (2016) have previously reported that Nfib amplification promotes metastasis; however, our clusters do not correlate with the 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) (Poirier et al., 2015), and the correspondence of these subtypes with the ones described here is intriguing and remains to be defined. Finally, Huang et al. (2018) recently reported an SCLC subtype defined by the expression of POU2F3. 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 (Udyavar et al., 2017), 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 the expression of one or few markers, which has been customary in the field but has limited ability to discriminate between phenotypes (**Figure 2.3**). 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.

2.4. Materials and Methods

2.4.1. 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 (George et al., 2015). The Myc-high mouse data set (Mollaoglu et al., 2017) was obtained from the NCBI GEO deposited at GSE89660. PDX/CDX mouse data (Drapkin et al., 2018) 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.

2.4.2. 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 (Wilkerson and Hayes, 2010). 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.

2.4.3. 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 (Yu et al., 2010) 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.

2.4.4. Drug sensitivity analysis

Our drug sensitivity analysis used the freely available drug screen data from Polley et al. (2016). 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 Polley et al. (2016), "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., 2021). 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 Emax, EC50, and the Hill coefficient, where each coefficient is constrained to reasonable ranges (Emax 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 Harris et al. (2016). 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.

2.4.5. 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.

2.4.6. 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).

2.4.7. Single-cell analysis

Cells with ≤ 500 detected genes per cell or with ≤ 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 log2transformed 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 the 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).

2.4.8. 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, germline 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.

2.4.9. 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 (Hagberg et al., 2008).

2.4.10. BooleaBayes inference of logical relationships in the TF network

A Boolean function of N input variables is a function F: $\{0,1\}^N \rightarrow \{0,1\}$. 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 number $v_i \in [0,1]$

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 ×N matrix $\boldsymbol{R} = [\vec{R}_1, \vec{R}_2, \dots, \vec{R}_N]$ quantifying the input regulator variables (columns) for each observation (rows), as well as an M dimensional vector $\vec{T} = [t_1, t_2, \dots, t_M]$ quantifying the output variable. A Gaussian mixed model is then used to transform the columns of **R** (regulator variables) and the vector $\vec{T}$ into probabilities **R'** and $\overrightarrow{T'}$ 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 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 **W** = $w_{i,j}$ as:

$$w_{i,j} = P_j(\vec{R'}_i) \tag{2.1}$$

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 $\vec{U} = [u_1, u_2, \dots, u_{2^N}]$ of each leaf:

$$u_j = 1 - \max_{i \in \{1, \ldots M\}} (w_{i,j}) \tag{2.2}$$

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

$$v_j = \frac{\sum_{i=1}^{M} t'_i \cdot w_{i,j} + 0.5 \cdot u_j}{\sum_{i=1}^{M} w_{i,j} + u_j} \tag{2.3}$$

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'_j, such as if $t'_1 = 0$ and $t'_2 = 1$.

2.4.11. 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, . . ., 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.

Algorithm 1 Limited Pseudo-Attractor Search

```
procedure SEARCH ENTIRE STG IN NEIGHBORHOOD OF GIVEN STATE TO FIND PSEUDO-ATTRACTORS
Inputs:
            $state\_init \in \{0,1\}^N$   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%.)
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
Inputs:
              $state\_init \in \{0,1\}^N$   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)
Output:
Number of steps taken before the random walk is a distance greater than $R$ from $state\_init$

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

Chapter 3.

Archetype tasks link intratumoral heterogeneity to plasticity in recalcitrant small cell lung cancer[2]

3.1. Introduction

3.1.1. The role of phenotypic plasticity in SCLC

Accumulating molecular and functional evidence has led to the identification of distinct SCLC transcriptional subtypes across several model systems, including cell lines, human tumors, and genetically engineered mouse models (GEMMs) (Borromeo et al., 2016; Gazdar et al., 1985; Huang et al., 2018; Mollaoglu et al., 2017). Phenotypic heterogeneity, both genetic and non-genetic, is intensively studied across cancer types because of its perceived impact on progression, acquired resistance, and relapse (Altschuler and Wu, 2010; Gupta et al., 2011; Howard et al., 2018; Jia et al., 2017; Pisco and Huang, 2015; Sáez-Ayala et al., 2013; Su et al., 2019). As described in Chapter 2, phenotypic heterogeneity is becoming increasingly more important to understand SCLC cell identity, and several subtypes of SCLC have been described. Dynamics of these subtypes within tumors are especially relevant for SCLC, since cooperativity and transitions among SCLC subtypes have been postulated to underlie its recalcitrant features, i.e., early metastatic spread, and inevitable relapse after therapy response (Ireland et al., 2020; Lim et al., 2017; Rudin et al., 2019).

In Chapter 2, we defined four subtypes of SCLC: a classic NE subtype, two NE variants (NEv1 and NEv2), and a non-NE variant. Soon after the publication of this work, Rudin et al. (2019) published a review summarizing the impact of phenotypic heterogeneity in SCLC. This review suggested consensus nomenclature for the subtypes, which can easily be aligned with the

[2] Adapted from Groves, S. M. *et al.* Cancer Hallmarks Define a Continuum of Plastic Cell States between Small Cell Lung Cancer Archetypes. *bioRxiv* (2021) doi:10.1101/2021.01.22.427865.

work described in Chapter 2. Whereas the consensus nomenclature underscored four SCLC subtypes, our recognition of two ASCL1+, NE subtypes (NE and NEv2) suggests that there are five main subtypes of SCLC: SCLC-A (NE), SCLC-N (NEv1), SCLC-A2 (NEv2), SCLC-Y (non-NE), and SCLC-P (which may have been missed in Chapter 2 due to its rarity).

While this delineation of SCLC into discrete subtypes can be useful for understanding the subpopulation composition of a given tumor, this stark subtyping, either at bulk or single-cell level, is difficult because often multiple or none of the eponymous TFs are expressed in a population of SCLC cells. For instance, our work in Chapter 2 using CIBERSORT decomposition (Newman et al., 2015) showed all tested SCLC tumors are composed of multiple NE and non-NE subtypes, and several studies have reported changes of subtype prevalence during tumor progression or in response to treatment (Ireland et al., 2020; Stewart et al., 2020; Wooten et al., 2019). Bulk RNA-seq and immunohistochemistry (IHC) analyses confirm some samples are positive for more than one TF, such as tumors that are positive for both ASCL1 and NEUROD1 (Simpson et al., 2020; Zhang et al., 2018). In bulk data, it is unclear if this is due to a mix of discrete NE and non-NE cells, or if intermediate phenotypes exist. These layers of heterogeneity suggest that single-cell data may be necessary to fully parse subtype prevalence in SCLC cell lines and tumors.

3.1.2. Discrete versus continuous methods of subtype identification

Clustering methods, which identify prototypical gene expression profiles of cluster centers, have often been used to characterize subtypes. These clusters are easily interpretable but are often too rigidly defined in the case of mixed or intermediate samples. One method that is more flexible than clustering and yet remains easily interpretable is Archetypal Analysis (AA) (Mørup and Hansen, 2012; Shoval et al., 2012). AA characterizes heterogeneous gene expression by finding archetypes, or “pure subtypes,” in gene expression space that best explain the heterogeneity seen across samples in a dataset. Using AA on SCLC cells from cell lines and tumors gives us the flexibility to identify and characterize the stability of intermediate states, which cannot be described in a discrete-clustering framework, that may arise as cells transition between subtype extremes.

The low-dimensional geometry of data uncovered by AA may be attributed to evolutionary tradeoffs between multiple functional tasks (**Figure 3.1**) (Hausser et al., 2019; Shoval et al., 2012). When cancer cells with limited resources (e.g. metabolic constraints) must optimize fitness in the face of multiple competing tasks, such as proliferation and migration, they fill a polygonal shape between archetypes in gene expression space (Gallaher et al., 2019; Hatzikirou et al., 2010). We analyzed the low-dimensional polytope of single-cell data from SCLC cell lines within this context and found the main trade-offs of SCLC cells include proliferation, migration, chemosensation, secretion, and lung epithelium regeneration, which mirror the tasks performed by pulmonary neuroendocrine cells (PNECs) under different environmental conditions (Garg et al., 2019; Gu et al., 2014; Lommel, 2001). Where each cell falls with respect to the archetypes determines how specifically it optimizes a single task (specialists near an archetype), or how it has generalized to complete several tasks (near the center of the polytope or along an edge or face between two or more tasks). If the proportion of tasks needed to optimize fitness changes rapidly, such as during tumor evolution and metastasis, a population of generalists may have an advantage. We show here that SCLC cell lines and tumors comprise continuums of cell states with both specialists and generalists.

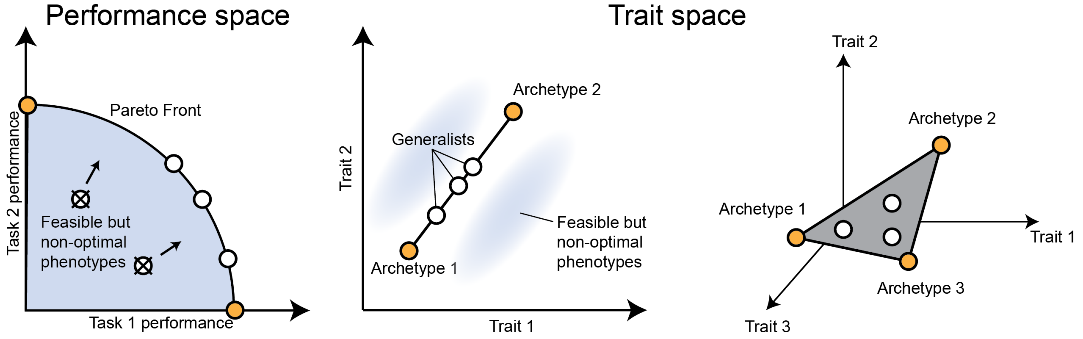


Figure 3.1: The Pareto Front is the set of all optimal phenotypes that cannot simultaneously improve at multiple tasks. *In performance space, where each axis is the performance level for a specific task. Feasible but non-optimal phenotypes are less optimal than phenotypes on the Pareto Front for at least task, so increasing fitness will push these cells towards the Pareto Front. Once on the Front, a cell cannot become more optimal at one task without becoming less optimal for another. In trait space, trade-off between two tasks force cells onto a line between archetypes, which are each optimal at a single task. Generalists in between the archetypes optimize multiple tasks at once. For three tasks, the archetypes form a triangle.*

3.1.3. Plasticity of cells within a phenotypic continuum

A continuum of transcriptomic states between archetypal extremes suggests that SCLC cells may easily diversify and shift between archetypes to optimize fitness by fulfilling different evolutionary tasks. To analyze changes in phenotype, we model single-cell dynamics as a Markovian process along an underlying state manifold (Teschendorff and Feinberg, 2021), from which we can calculate metrics of plasticity. We quantify the average change in expression over the phenotypic transition from source states to terminal states, which we term Cell Transport Potential (CTrP). Using this metric, we found that SCLC cells from human cell lines diversify across the archetype space; in the cell lines studied, we delineate subpopulations of high and low plasticity within each sample. We saw that, under some circumstances, such as MYC activation, NE subtypes can acquire plasticity. Importantly, our subtyping and plasticity framework allows for the characterization of transitions between intermediate phenotypes, which cannot be adequately captured by a 4-TF framework. We also quantify multipotency of MYC-driven transitioning cells and show that cells can transition towards two lineages: SCLC-Y, or Archetype X, which does not match any of our previously defined subtypes of SCLC.

3.1.4. Plasticity of pulmonary neuroendocrine cells

In this chapter, we hypothesize that SCLC phenotypes may gain plastic capabilities from the cell of origin from which they are derived. Over the last half century, our understanding of the cell of origin of SCLC has been greatly refined. Because SCLC cells share many features with pulmonary neuroendocrine cells (PNECs), such as small, dense core granules characteristic of neuroendocrine cells, and expression of NE genes like ASCL1, Bensch and colleagues in the 1960s postulated that SCLC arose from PNECs (Bensch et al., 1968).

Using a Cre-Lox system, in which an adenovirus expressing Cre-recombinase driven by a cell type-specific promoter is administered to a mouse, researchers can direct SCLC-promoting inactivations in P53 and RB1 into specific lung cell types of interest. In 2011, researchers in the Berns' lab used a double knockout model with adenoviruses that targeted three distinct respiratory epithelial populations: CGRP targeted NE cells, SPC targets AT2 cells, and CC10 targeted club cells (Sutherland et al., 2011). While SPC promoters induced SCLC with low efficiency, and CC10 produced no SCLC, CGRP was most effective in generating SCLC, with 27 of 30 models developing the cancer. The same year, the Kim and Sage laboratories found similar results with a tamoxifen-inducible Cre^{ER} system, where all non-PNEC cells of origin did not develop into SCLC (Park et al., 2011).

The following year, Song et al. (2012) used lineage tracing in CGRP-Cre^{ER} mice crossed to a double knockout mouse model (P53 and RB1) that labeled PNEC-lineages with eGFP. Almost all mice that developed hyperplastic lesions were lineage-labeled, providing more definitive evidence that PNECs are the predominant cell of origin for SCLC. However, this does not preclude the existence of a non-NE cell of origin for SCLC. More recently, several groups have shown that SCLC may develop from non-PNECs, which influences the eventual phenotypic landscape of the tumor (Ferone et al., 2020; Park et al., 2018; Yang et al., 2018). Furthermore, evidence regarding a newer subtype of SCLC driven by POU2F3 expression suggests these tumors may derive from tuft-like cells in the lung (Huang et al., 2018). Overall, this evidence seems to suggest that lung cells may have broad plasticity for SCLC transformation, but the most common cell of origin remains the PNEC. Therefore, while we explore other cells of origin in Chapter 4, in this chapter we focus on PNECs, which are still considered the main cell of origin and the most efficient source of SCLC in mouse models.

PNECs themselves have been shown to have broad plasticity in terms of functional state. PNECs are capable of proliferating and transdifferentiating, particularly after injury to the lung epithelium. Following exposure to naphthalene, which ablates club cells, PNECs repopulate the club and ciliated cells in the lung, as experimentally shown by Song et al. (2012). The number of PNECs, which were labeled by eGFP in a CGRP$^{CreER/+}$;ROSA26$^{mTmG/+}$ mouse model, significantly increased after injury. Furthermore, some of the club and ciliated cells were labeled with eGFP, suggesting that they could be derived directly from PNECs. Recently, Ouadah et al. (2019) showed that a subset of PNECs have stem cell capabilities, and injury induces PNEC stem cells to self-renew and disperse to other areas in the lung. Through Notch signaling, these PNECs are then able to deprogram into a transit-amplifying state. Finally, activation of Notch signaling late in repair can induce reprogramming to mature cell fates such as club cells. This deprogramming ability inherent to PNECs in response to injury may manifest in SCLC as phenotypic plasticity.

3.1.5. Overview of chapter

Overall, our work advances the field’s understanding of SCLC heterogeneity and plasticity by revealing the prevalence of cells that fall in between extreme subtypes, thus demonstrating a need for a more flexible method of phenotype characterization. We enumerate the extreme archetypal phenotypes as optimizing PNEC-related functions that require tradeoffs. We provide a theoretical basis for the existence of generalist cells, which are supported by recent evidence of dual positive cells (such as ASCL1+/NEUROD1+ cells), by characterizing these tradeoffs that SCLC tumors must make to thrive. We quantify plasticity as CTrP and show that SCLC NE cell types under MYC activation may be multipotent. These findings suggest that SCLC tumors work as a complex ecosystem of plastic NE and non-NE cells that can phenotypically transition under different environmental constraints to optimize tumor fitness and overcome therapy.

3.2. Results

3.2.1. Archetype analysis defines a 5-vertex polytope for SCLC

As described in previous Chapters, SCLC subtypes have recently been classified into neuroendocrine (NE) and non-NE subtypes by expression of eponymous transcription factors: ASCL1+ (NE), NEUROD1+ (NE), POU2F3+ (non-NE), and triple-negative non-NE subtypes, often but not always YAP1+ (Baine et al., 2020; Lim et al., 2017; Rudin et al., 2019; Simpson et al., 2020). To further examine relationships between these discrete subtypes, we analyzed a dataset of bulk RNA-seq on 120 human SCLC cell lines from two sources: the Cancer Cell Line Encyclopedia (CCLE), and cBioPortal (Barretina et al., 2012; Cerami et al., 2012; Gao et al., 2013). This combined dataset includes cell lines with overexpression of each of the four subtype-driving TFs, suggesting it adequately covers the relevant phenotypic space for SCLC. Furthermore, we defined the SCLC phenotypic space on cell line data under the assumption they are less heterogeneous, and therefore may better capture SCLC cell-specific phenotypes, than tumor samples that may contain other cell types like immune cells. In Chapter 2, we analyzed gene expression profiles (RNA-seq) of human SCLC cell lines using Weighted Gene Co-expression Network Analysis (WGCNA) and showed that each subtype expressed gene programs (modules) enriched in distinct cellular functions, such as immune response or neuronal differentiation.

In this Chapter, we update this characterization to include additional human cell line samples and the SCLC-P subtype (**Figure 3.2**). We found groups of genes (gene modules) with coordinated expression across the subtypes, each enriched in a distinct set of cellular functions. Furthermore, a subset of the gene modules corresponded to each SCLC subtype. This diversity of functions across subtypes may arise when cells, under selective pressure to optimize survival tasks, cannot optimize all tasks at once and must tradeoff between them (Hausser et al., 2019; Shoval et al., 2012). Therefore, we asked whether the presence of distinct subtype gene programs in SCLC cell lines might similarly suggest trade-offs between functional tasks. To this end, we applied

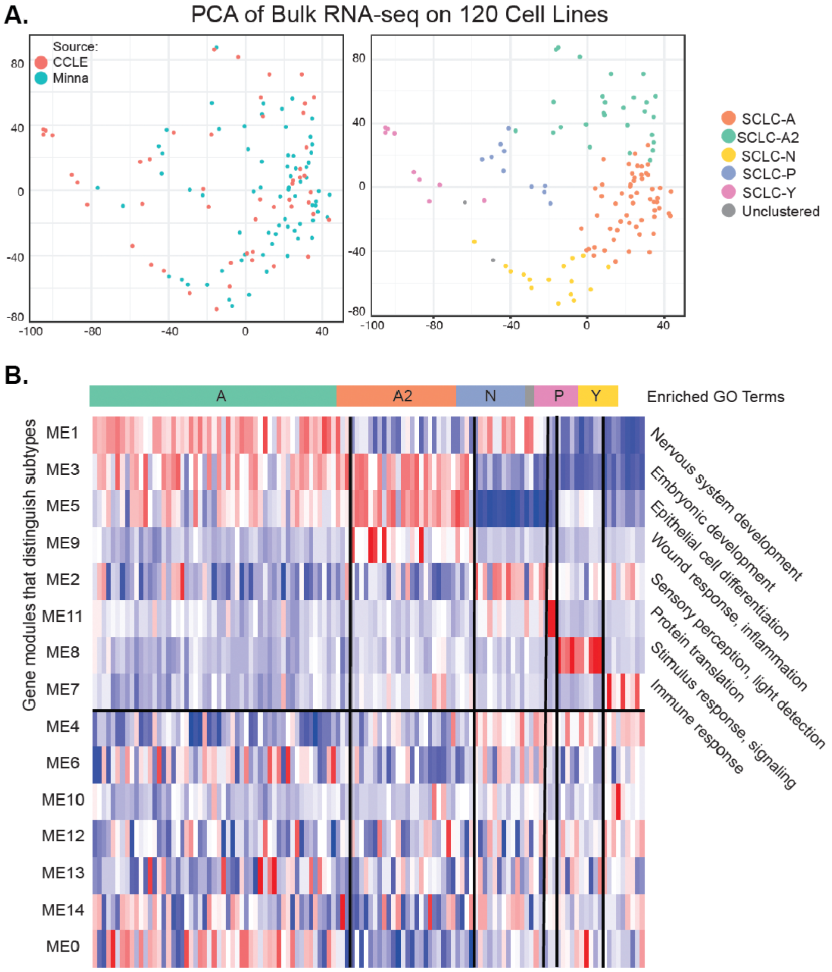


Figure 3.2: Clustering and WGCNA on Bulk Cell Lines (Updated from Chapter 2). A. *Cell line source in PCA, and clustering shown by color on PCA.* ***B.*** *WGCNA on cell lines shows genes can be grouped into 15 coexpressed gene modules. Several of the modules (above black line) distinguish subtype clusters and are labeled with enriched gene ontology terms describing each gene program (see methods).*

Archetype Analysis (AA), which allows for a flexible characterization of gene expression space constrained by functional phenotypic features (Mørup and Hansen, 2012).

Briefly, AA approximates the cell phenotype space as a low dimensional shape, or polytope, that envelops gene expression data. The vertices of this multi-dimensional shape represent archetypes, constrained to be linear mixtures of some set of data points, that are each optimal in a specific functional task. To determine the optimal number and location of the archetype vertices in gene-expression space, we applied the Matlab package *ParTI* (Hart et al., 2015). We used the Principal Convex Hull Analysis (PCHA) algorithm (Mørup 2012), which finds k points on the convex hull, or bounding envelope, of the data that enclose as much of the data as possible (See Methods) (Korem et al., 2015). Using this method, we determined whether the cell line data was enclosed within a low dimensional polytope and compared the fit to randomized datasets to calculate statistical significance.

First, to determine how well the data is fit by polytopes of varying dimensionality, we computed the variance in the data that is explained (Explained Variance, EV) by polytopes with different possible numbers of k vertices (k = 2-15). We found that EV saturates around 5 or 6 archetypes, such that the variance explained by additional archetypes was minimal (**Figure 3.3A**). This was confirmed by identifying the elbow, k*, in the EV versus k curve, which suggested k*=4, 5, or 6 (See Methods). Therefore, we fit the data to polytopes of each order (4, 5, or 6 vertices), and computed the t-ratio, a measure comparing the volume enclosed by the data to that of a polytope. As described in Korem et al. (2015), a larger t-ratio suggests that the data is more similar to the polytope (**Figure 3.3A**). The t-ratio of the data can be compared to that of randomly shuffled datasets to quantify the significance of the fit as a p-value.

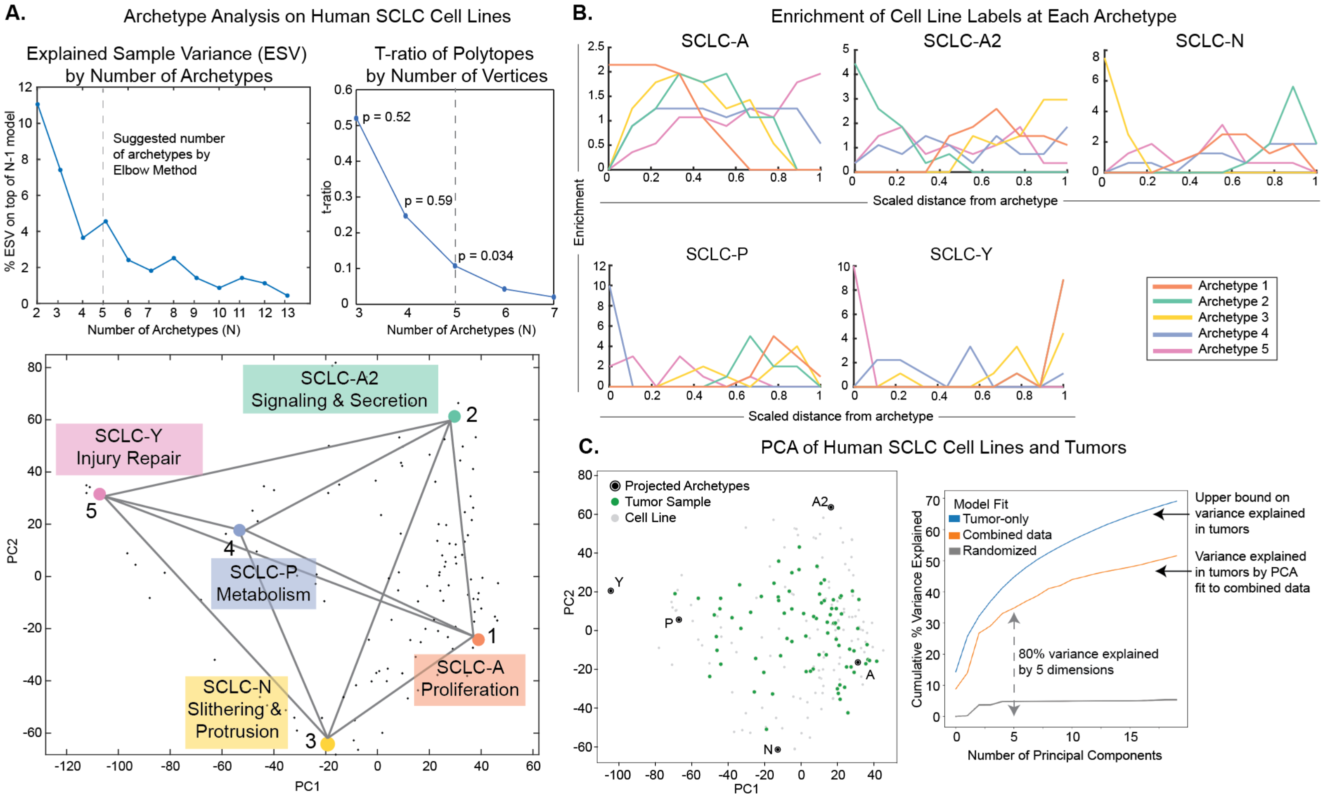


***Figure 3.3: Archetype analysis on bulk RNA-seq data shows human cell lines and tumors fall in a polytope with five archetypes. A.** Archetype analysis of bulk RNA-seq from 120 human cell lines shows 5 archetypes fit the cell line data well (p = 0.034). Explained sample variance increases for 5 archetypes compared to 4, and 5 archetypes is the lowest number with a significant p-value by a t-ratio test. **B.** Subtype label enrichment. Data were binned by distance from archetype (x-axis), and enrichment of each subtype label (y-axis) was computed. Enriched subtypes are highest at x=0, in the bin closest to one of the archetypes, and lowest near all other archetypes. Each archetype shows enrichment in one of the five SCLC subtypes from literature. **C**. PCA of full human RNA-seq dataset (tumors and cell lines). Projection of 5 archetypes by this PCA shows that tumors are mainly contained within the same archetype space as cell lines. Variance explained by this combined-data PCA, a tumor-data PCA, and a randomized model shows that the top 5 components of the combined-data PCA explains a significant percentage, around 80%, of the variance explained by the tumor-only PCA.*

To avoid overfitting, the lowest number of archetypes that reached significance was chosen. Therefore, a polytope with five archetypes best fits the data (**Figure 3.3A,** p-value = 0.034, t-ratio test). Mathematically, each of the consensus SCLC subtypes was enriched at an archetype (**Figure 3.3B**, p < 10-6 for each subtype) such that there is a one-to-one correspondence between archetypes and consensus subtypes and the nomenclature is interchangeable. Fitting the data to a polytope with fewer vertices, such as a tetrahedron (four-vertex polytope) did not achieve a statistically significant t-ratio (p-value = 0.059). Furthermore, the only difference between the 4- and 5-vertex polytopes was the SCLC-P archetype, which is a distinct, and not an intermediate phenotype (Huang et al., 2018). When we compared the archetypes of the 5- and 6-vertex polytopes (see Methods), we found that the 6-vertex polytope did not identify any distinct archetypes, as two of the new archetypes matched one in the 5-vertex polytope, and each other vertex matched one-to-one between polytopes. We used bootstrapping tests where we resampled the data with replacement 1000 times to evaluate the robustness of the archetypes found. We found the five archetypes were robust to data sampling and not dependent on any extreme points in the dataset.

To determine if cell-line archetypes could adequately describe the variance in human tumors, we batch-corrected 81 human SCLC tumor samples (George et al., 2015) to the cell line data. If tumors are heterogeneous mixtures of different cell types, we would expect each of their bulk (averaged) expression profiles to fall closer to the center of the polytope. When we project the archetypes by a PCA fit to the combined dataset, we find that most tumors are contained by the same phenotypic space as cell lines (**Figure 3.3C**). Furthermore, the variance explained by this PCA is a significant proportion of the variance explained in a tumor-only PCA, with the top five components explaining 80% of the tumor variance (**Figure 3.3C**). The polytope best fit to the combined dataset of cell lines and tumors had 5-archetypes ($p = 0.09$), and each archetype matched at least one of the cell line archetypes.

In summary, AA explained SCLC heterogeneity in bulk transcriptomics data as a low-dimensional phenotypic space between five archetype vertices corresponding to five major SCLC phenotypes (SCLC-A, -A2, -N, -P, and -Y). The archetype space enables the placement of any bulk transcriptome profile, including those that may not adhere to any of the canonical subtypes, along a continuum anywhere within the polytope, rather than either remain unclassifiable or forced into a class not fully reflective of their transcriptomic profile. Specifically, samples ill-defined due to lack of expression of any of the eponymous TFs can be classified in this polytope based on distance from the archetypes. In addition, functional tasks optimized by each subtype can be inferred based on the expressed gene programs of cell lines nearest the archetypes.

3.2.2. The SCLC phenotypic polytope is bounded by functional tasks reminiscent of PNECs

Pulmonary neuroendocrine cells (PNECs), the physiological counterpart of SCLC in the normal lung, are plastic cells that can trade-off between functions in response to microenvironmental conditions, including lung epithelium repair in response to injury, chemosensation, and secretion of neuro- and immuno-modulatory peptides (**Figure 3.4A**) (Garg et al., 2019; Song et al., 2012). Therefore, we hypothesized that SCLC cells may be innately programmed to fulfill similar tasks, albeit in a dysregulated manner, geared toward optimized tumor fitness and increased survival.

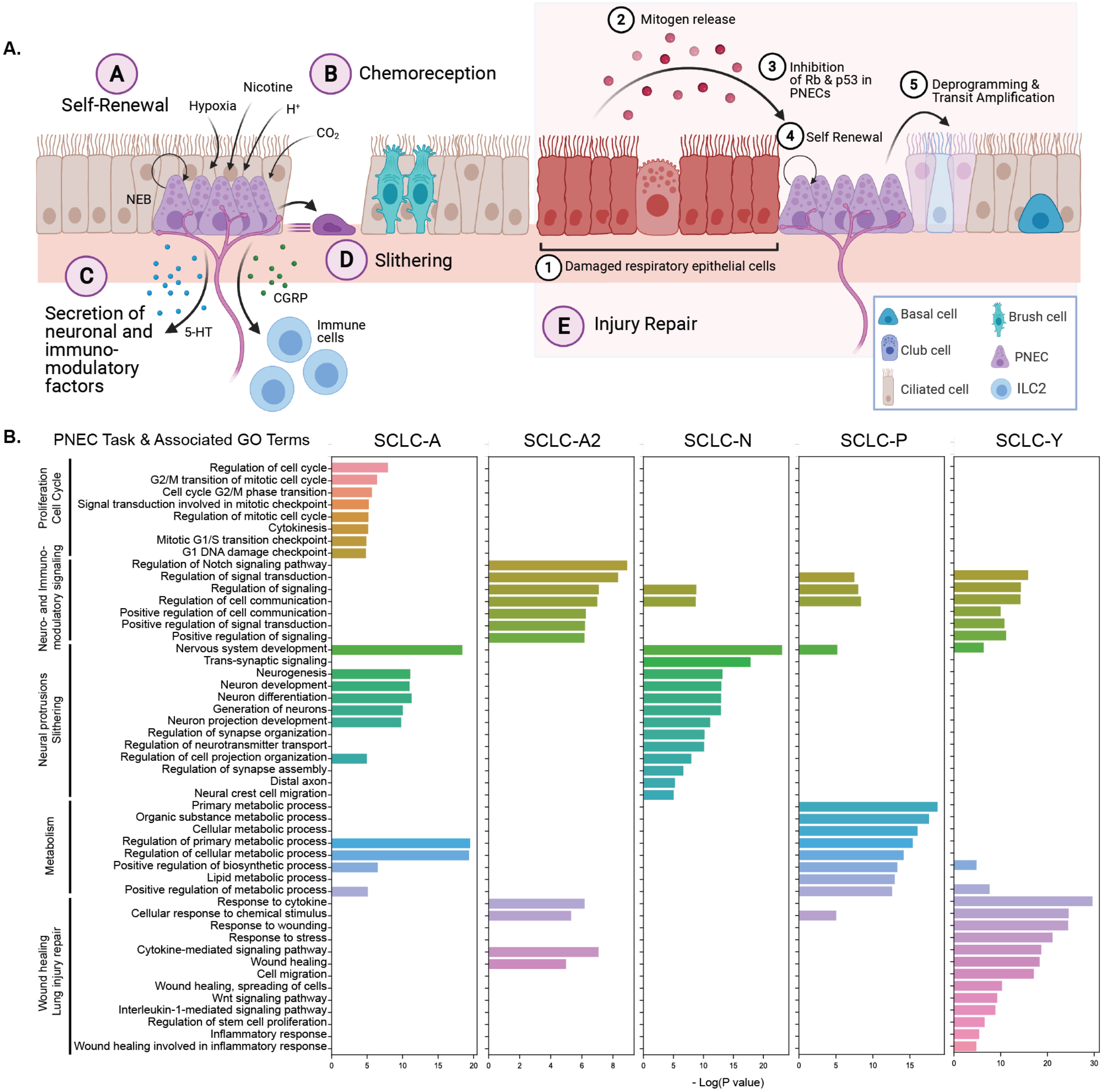


Figure 3.4: SCLC archetypes are enriched for PNEC-related gene programs.* *A. *Pulmonary neuroendocrine cell (PNEC) related tasks. PNECs can trade-off between these tasks to regenerate injured lung epithelium, respond to chemical signals in the microenvironment, affect the nervous and immune systems, and migrate to new regions of the lung airways. A. A subset of PNECs have been shown to act like stem cells that can proliferate under lung injury (Ouadah et al., 2019). B. PNECs and brush cells both respond to chemicals and cytokines in the lung (Lommel et al., 2001). C. PNECs are innervated and can send neuronal signals by releasing neurotransmitters and peptides such as serotonin (5-HT) (Lommel et al., 2001). They also have been shown to interact with the immune system by releasing proteins such as CGRP, which can activate IL2 cells (Branchfield et al., 2016). D. A subset of PNECs can "slither," or migrate, by transiently downregulating epithelial genes to move toward and form neuroendocrine bodies (NEBs), or clusters of PNECs (Kuo and Krasnow, 2015). E. After injury to the lung epithelium (ablation of club cells), PNEC stem cells can deprogram into a transit amplifying cell type that can then differentiate into other lung types to regenerate the epithelium (Ouadah et al., 2019).* ***B.*** *Each archetype is enriched in gene ontology terms related to PNEC tasks.*

To define functional tasks optimized by each archetype, we evaluated enrichment of genes at each SCLC archetype location (Bonferroni-Hochberg-corrected q < 0.1). We then used ConsensusPathDB on the most enriched genes to find enriched gene ontologies and used the molecular signatures database (MSigDB) to evaluate the enrichment of cancer hallmarks (Kamburov et al., 2013; Liberzon et al., 2011; Zhang et al., 2020). As shown in **Figure 3.4B** and **Table 3.1**, each archetype optimized a task previously associated with PNECs and performed functions to promote tumor survival.

Table 3.1: Archetype tasks are related to PNEC functions and increase tumor fitness through optimizing cancer hallmarks.

*Cancer hallmark is inferred from GO term enrichment rather than the enrichment of Cancer Hallmark Gene Sets.

Archetype	Associated PNEC task	Optimized function for increased tumor fitness
SCLC-A	Proliferation	Increased cell proliferation*
SCLC-A2	Neuro- and immuno-modulatory signaling	Evading immune destruction & tumor-promoting inflammation
SCLC-N	Slithering and axon-like protrusions	Activating invasion and metastasis*
SCLC-P	Chemosensation and metabolism	Reprogramming energy metabolism
SCLC-Y	Transdifferentiation to non-NE state in response to injury	Inducing angiogenesis & resisting cell death

Archetype 1, corresponding to the SCLC-A subtype (**Figure 3.3B**), is enriched in cell cycle GO terms. This enrichment may reflect the self-renewal potential of PNECs, which proliferate after lung injury and/or chronic hypoxia (McGovern et al., 2010; Noguchi et al., 2020). Previous studies on ASCL1 positive, HES1 negative cells similar to the SCLC-A archetype have shown them to be more proliferative than other SCLC cell types (Lim et al., 2017). Therefore, these archetype tasks are consistent with the highly proliferative nature of the SCLC-A subtype, evidenced by its often-larger proportion in primary tumors (Alam et al., 2020; Carney et al., 1985).

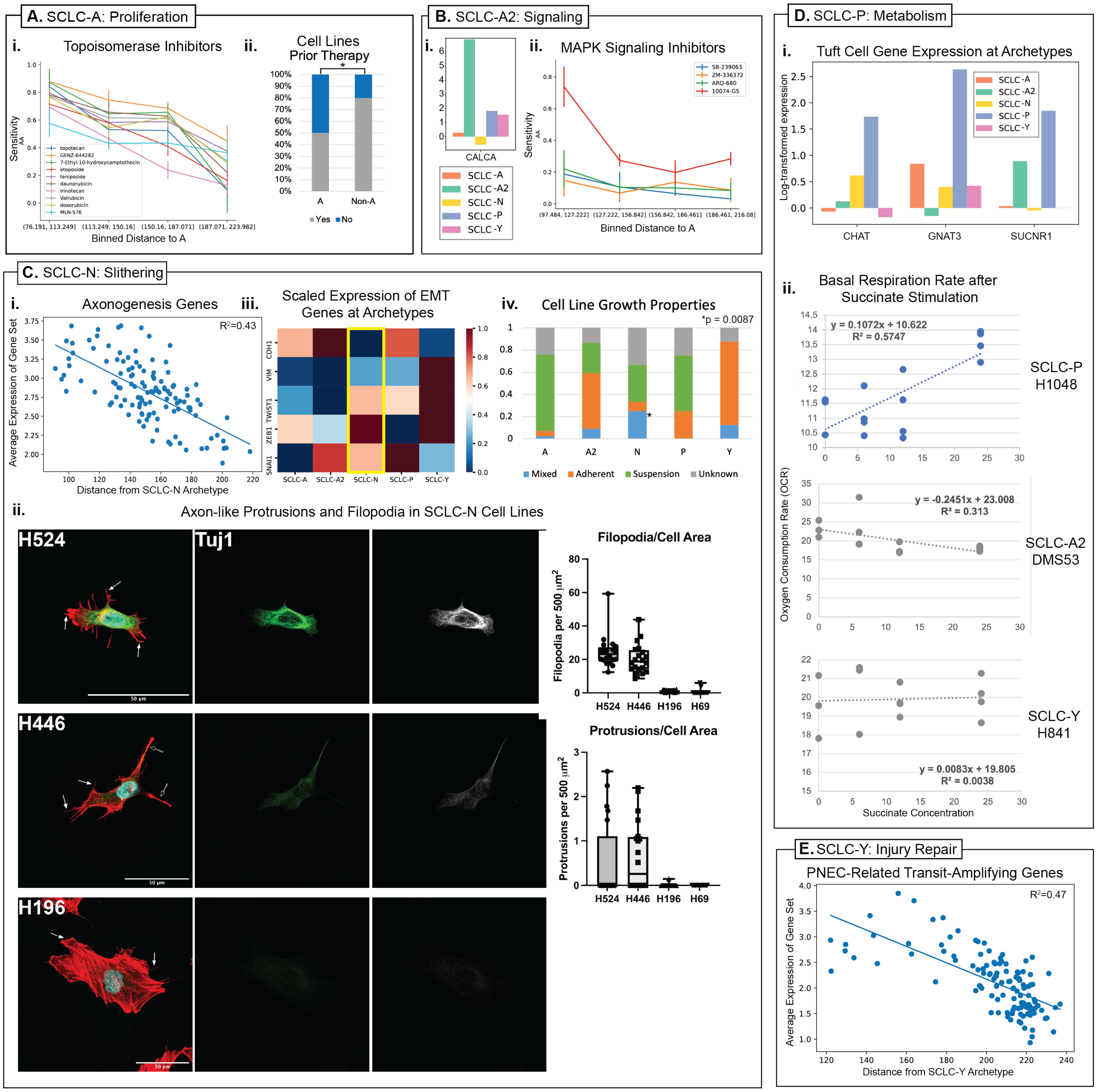


Figure 3.5: SCLC cell line archetypes optimize PNEC-related tasks. A. *SCLC-A is enriched for proliferation.* ***i.*** *Normalized activity area (AA, a measure of sensitivity) to topoisomerase inhibitors. Cell lines in the bin closest to the SCLC-A archetype are more sensitive (p <0.05).* ***ii.*** *Cell lines closest to A are less likely to have had prior therapy (p = 0.019).* ***B.*** *SCLC-A2 is enriched for signaling.* ***i****. CALCA expression is highest at SCLC-A2 archetype.* ***ii****. Cell lines closest to SCLC-A2 are most sensitive to MAPK signaling inhibitors (p < 0.05).* ***C****. SCLC-N is enriched for slithering-related tasks.* ***i****. Average expression of an axonogenesis gene set from Yang et al. as a function of distance from the SCLC-N archetype, showing a correlation between expression and closeness to the SCLC-N archetype.* ***ii****. Axon-like protrusions and filopodia are more prevalent in SCLC-N cell lines. Open arrows = protrusions, closed arrows = filopodia.* ***iii****. EMT genes are shown in a heat map across archetypes. SCLC-N cells express some mesenchymal markers at intermediate levels and downregulate CDH1.* ***iv****. SCLC-N cell lines are more likely to be mixed (3/12) than non-N cell lines (3/80) with p = 0.0087.* ***D****. SCLC-P is enriched for tuft cell-like features and metabolism tasks.* ***i****. Genes upregulated in the SCLC-P archetype that are expressed in tuft cells. CHAT, GNAT3, and SUCNR1 are part of the pathway by which succinate stimulation affects the metabolism of intestinal tuft cells and the stimulation of type 2 immunity (Banerjee et al., 2020).* ***ii****. Basal respiration rate (OCR) after overnight (12 hour) stimulation by succinate. H1048, which is closest to the SCLC-P archetype, increases OCR after stimulation, while SCLC-A2 and SCLC-Y cell lines do not.* ***E****. SCLC-Y is enriched in injury repair tasks. Average expression of genes related to the transit-amplifying subpopulation of PNEC stem cells from Ouadah et al. (2019) under lung injury is correlated with closeness to the SCLC-Y archetype.*

Furthermore, classic tumors containing mostly proliferative SCLC-A cells are initially sensitive to DNA damaging agents that selectively kill fast-growing cells (Sen et al., 2018), such that ASCL1 is downregulated in post-chemotherapy tumors and chemoresistant cell lines (Wagner et al., 2018). Analysis of drug sensitivity to topoisomerase inhibitors, DNA alkylators, and cell cycle inhibitors shows that cell lines closest to SCLC-A are more sensitive to these drug classes (**Figure 3.5Ai**). This is reflected in the cell line data analyzed here: cell lines near the SCLC-A archetype are more likely to be untreated than cell lines near other archetypes ($p = 0.019$), and similarly treated cell lines are less likely to be near SCLC-A ($p = 0.03$, one-tailed binomial tests on treatment status of cell lines, see Methods, **Figure 3.5Aii**). Together, this evidence suggests that the SCLC-A archetype optimizes the cancer hallmark of *increased cell proliferation* (**Table 3.1**).

Archetype 2 (SCLC-A2), also driven by NE gene programs, is enriched for stimulus-response, cytokine-mediated signaling, and signal transduction, suggesting these cells specialize in the PNEC task of neuronal and immune-modulatory signaling and secretion. Together, optimization of these tasks may allow SCLC-A2 cells to interact with the tumor microenvironment quickly and effectively by sensing and responding to signals from other cells. This is consistent with previous work from our co-authors that showed the SCLC-A2 subtype is enriched in GO terms related to neuronal secretion and response to environmental signals (Wooten et al., 2019). The SCLC-A2 archetype is enriched for *CALCA* expression (**Figure 3.5Bi)**, overexpression of which has been shown to modulate the immune system (Branchfield et al., 2016). Furthermore, SCLC-A2 cell lines are preferentially sensitive to MAPK signaling inhibitors (**Figure 3.5Bii**). Together, this evidence corroborates enrichment of the cancer hallmarks *tumor-promoting inflammation* and *evading immune destruction* (**Table 3.1**).

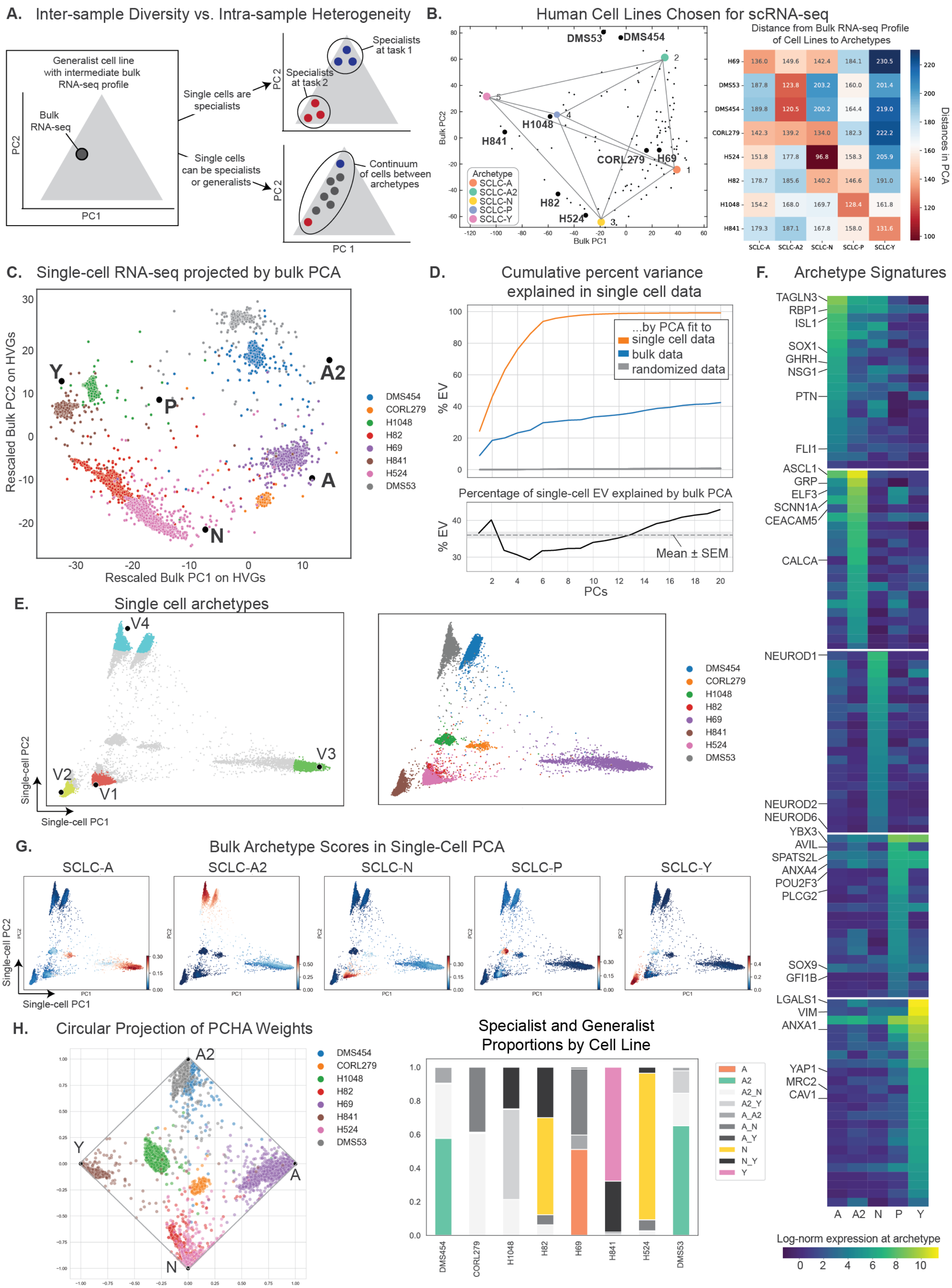
Archetype 3 (SCLC-N) is enriched in neurogenesis terms, including synapse and distal axon terms. These functions may enhance tumor survival by specifying a protruding, axon-like morphology. Yang et al. (2019) previously determined that some SCLC cells are capable of forming axon-like protrusions, and disruption of protrusion formation impairs cell migration. Therefore, we compared the expression of axon guidance-related genes from this study across cell lines and found that distance to SCLC-N was inversely correlated to expression (**Figure 3.5Ci**). To substantiate this, we imaged cell lines close to SCLC-N (H524 and H446) and found that they had substantially more protrusions than cell lines far from the SCLC-N archetype (**Figure 3.5Cii**). Furthermore, these protrusions were positive for Tuj1, which is a marker for neuronal protrusions, suggesting the protrusions we see here are truly axon-like (Yang et al., 2019). Such a morphology may be related to the slithering observed in PNECs, whereby cells transiently downregulate adhesion genes and use axon-like protrusions to migrate (Kuo and Krasnow, 2015; Osborne et al., 2013). We therefore considered the expression of adhesion, migration, and epithelial-to-mesenchymal transition (EMT) genes in the SCLC-N phenotype. We found that the mesenchymal genes *ZEB1*, *SNAI1*, and *TWIST1* are upregulated in SCLC-N, but not VIM, which may suggest a hybrid E/M or M phenotype (**Figure 3.5Ciii**). This may be reflected in the growth of SCLC cell lines, where cell lines close to SCLC-N are significantly more likely to have a mixed morphology than non-N cell lines ($p = 0.0087$, **Figure 3.5Civ**). Thus, by performing the PNEC task of slithering, Archetype 3 may optimize the hallmark *activating invasion and metastasis* to promote survival (**Table 3.1**).

Archetype 4 (SCLC-P) is enriched in metabolic GO terms. While the cell of origin of SCLC-P cells remains unclear, the phenotype has been described as tuft-like and shows remarkable similarity to brush cells in the lung (Huang et al., 2018), which recent evidence suggests may act as precursors for PNECs (Goldfarbmuren et al., 2020). Chemosensory tuft cells respond to the metabolite succinate through the receptor SUCNR1, promoting type 2 inflammation through ILC2 activation (Nadjsombati et al., 2018). We found that SCLC-P upregulates the receptor SUCNR1 and gustducin (GNAT3) (**Figure 3.5Di**). Therefore, we tested the ability of SCLC-P cells to respond to succinate metabolically by measuring their basal oxygen consumption rate (OCR) after an overnight stimulation. In response to succinate, SCLC-P cells (but not SCLC-A2 or -Y) adapted their metabolism by increasing basal respiration rate (**Figure 3.5Dii**). Therefore, SCLC cells close to the SCLC-P archetype may respond to metabolites like succinate, similar to the function of chemosensory tuft cells. These functions validate our findings that the SCLC-P archetype shows gene set enrichment of the cancer hallmark *reprogramming energy metabolism* (**Table 3.1**).

Archetype 5 (SCLC-Y) was enriched in GO terms such as stress response, wound healing, and cell migration. This archetype showed gene set enrichment of the most cancer hallmark gene sets, corroborating previous findings that it may be key to understanding resistance (Cai et al., 2021; Lim et al., 2017; Wagner et al., 2018). The cancer hallmarks of *inducing angiogenesis* and *resisting cell death* showed the greatest enrichment in SCLC-Y cell lines compared to others (**Table 3.1**). Notably, normal PNECs transdifferentiate to a transit-amplifying (TA) state to repair the lung epithelium after injury (Ouadah et al., 2019). We compared the transcriptomes of SCLC cell lines to different cell types arising from this process. While sequencing data is limited for understanding the transit-amplifying (TA) state of de-differentiated PNECs, we evaluated the expression of genes related to the small number of transdifferentiated PNECs in Ouadah et al. (2019). There is a clear correspondence between the SCLC-Y subtype and the TA signature (**Figure 3.5E**). We conclude that this archetype is a dysregulated version of the transit-amplifying cell type whose task is lung repair after injury. Upregulation of genes in the NOTCH and WNT pathways provides further evidence that SCLC-Y corresponds to PNECs that regenerate the lung epithelium in response to Notch signaling after injury (Lim et al., 2017; Shi et al., 2015; Wagner et al., 2018). In summary, these data indicate that SCLC subtypes in cell lines and tumors are reminiscent of the functional tasks of normal PNECs. Furthermore, the functions can be tied to the enrichment of cancer hallmark tasks, illustrating how SCLC cells may utilize PNEC functions for survival (**Table 3.1**).

3.2.3. Intra-sample heterogeneity is aligned with inter-sample diversity

By considering bulk RNA-seq data, we characterized the diversity of SCLC cell line and tumor samples and identified five archetypal gene programs enriched at the extremes of this phenotypic space. However, it is unclear if single cells within each sample can be both generalists and specialists. While specialist cell lines are most likely made up of specialist single cells, a generalist cell line could comprise multiple specialist subpopulations or generalist single cells (**Figure 3.6A**). Therefore, we considered the relationship between this inter-sample diversity and intra-sample heterogeneity. To do this, we analyzed single-cell expression data from a panel of 8 cell lines, selected to maximally span the archetype space (see Methods, **Figure 3.6B**). The axes of maximal variance with this dataset can be defined with Principal Components Analysis (PCA) fit to the single-cell expression data. We compared the variance explained by this model to the variance explained by projecting the single-cell data onto the space defined by the bulk data-derived archetypes (**Figure 3.6C**). If intra-sample heterogeneity perfectly aligns with inter-sample diversity, we would expect the single-cell variance explained by inter-sample diversity to equal the single-cell variance explained by the single-cell PCA. Therefore, percent variance explained by the single-cell PCA is an upper bound on the variance explained by inter-sample diversity.



	SCLC-A	SCLC-A2	SCLC-N	SCLC-P	SCLC-Y
H69	136.0	149.6	142.4	184.1	230.5
DMS53	187.8	123.8	203.2	160.0	201.4
DMS454	189.8	120.5	200.2	164.4	219.0
CORL279	142.3	139.2	134.0	182.3	222.2
H524	151.8	177.8	96.8	158.3	205.9
H82	178.7	185.6	140.2	146.6	191.0
H1048	154.2	168.0	169.7	128.4	161.8
H841	179.3	187.1	167.8	158.0	131.6

Figure 3.6: SCLC archetype gene signatures reveal generalists and specialists in cell lines at the single-cell level. A. *Inter-sample diversity is supported by intra-sample heterogeneity. Generalist cell lines may comprise several specialist subpopulations or both specialists and generalists in a continuum of single cells.* ***B.*** *To investigate intra-sample heterogeneity, human cell lines for scRNA-seq were chosen to span the phenotypic space of SCLC. Two cell lines from each neuroendocrine subtype (A, A2, and N) were chosen, and one from each non-neuroendocrine subtype (P and Y) was chosen. Left: chosen cell lines in bulk PCA space. Right: Distance of each bulk cell line gene expression profile to each archetype in PCA.* ***C.*** *Single-cell RNA-seq on sampled cell lines projected by PCA fit to bulk RNA-seq on cell lines in A. Each sample occupies a distinct region, and many samples fall in between archetypes.* ***D.*** *Top: Variance explained in single-cell data by PCA fit to bulk cell line data. Orange: Upper bound of EV for each number of components is given by PCA fit to single-cell data. Blue: EV by the bulk PCA is a significant proportion of this, as compared to a randomized model (gray). Bottom: Inter-sample diversity explains a significant percentage of the intra-sample variance, around 36%. This fraction stays relatively constant for varying numbers of PCs. Black line: intra-sample variance explained by inter-sample diversity as a percentage of upper bound. Grey dotted line: Mean +/- SEM (grey box).* ***E.*** *Left: single-cell archetypes from PCHA on imputed cell line scRNA-seq data in single-cell PCA. 5% of cells closest to each archetype are colored; generalists are shown in gray. Right: Cell lines labeled in single-cell PCA.* ***F.*** *Gene signature used for single-cell subtyping. Expression of genes at archetype location is shown, with genes of interest highlighted.* ***G.*** *Using least-squares approximation, we score single cells by 5 bulk archetype signatures in (F).* ***H.*** *Using a permutation test (see Methods), we compare average archetype scores of each single-cell specialist subpopulation to background distributions (orange) from non-specialists to label archetypes. Circular a posteriori (CAP) plot of single-cell archetype weights for each cell (see Methods), with archetypes labeled by enriched bulk signature.*

Projection of the single cells onto the archetype-defined space suggests that inter-sample diversity in human SCLC cell lines explains 36% of the intra-sample variance (**Figure 3.6D**). The alignment between bulk and single-cell variation is not likely to be due to random chance: PCA models fit to shuffled bulk data only explained about 0.26 +/- 0.008% of the single-cell variance (50 shuffles, see Methods). The remaining unexplained single-cell variation may be due to the inherent stochasticity of RNA expression in single cells (Hayford et al., 2021). Overall, intra-sample variation in SCLC cell lines is well explained by variation between human SCLC samples.

3.2.4. Single cells in SCLC cell lines can be task specialists or generalists

Our analyses so far suggest that single cancer cells fit into the phenotypic space defined by population-level measurements. We next sought to grade single cells along a continuum of specialists and generalists in the bulk-derived archetype space. To do so, we compared a polytope fit to single-cell data with the bulk data-derived archetypes. We first applied PCHA to the single-cell data directly to determine if the geometry of the data was bounded by a polytope. We found the sampled cell lines fall in a shape with four vertices with a t-ratio test p-value of 0.001 (**Figure 3.6E**, see note about SCLC-P in Methods). This suggests that cancer cells trade-off between multiple functions at the individual, and not just the population, level.

To align these single-cell archetypes with our previously defined bulk archetype space, we asked whether each single-cell archetype was enriched for a bulk archetypal gene signature. We generated gene expression signatures characteristic of each bulk archetype location by finding genes enriched in the bulk expression profiles of cell lines closest to each archetype (Mann-Whitney Test, $q < 0.1$, see Methods). We then perform feature selection by considering the condition number of the gene signature matrix, which measures the sensitivity of the matrix to changes, or errors, in input (i.e., the bulk RNA-seq profiles). A well-conditioned matrix with a low condition number is better able to discriminate between archetypes and therefore can be used to project other data into this lower-dimensional space more accurately. By minimizing the condition number, we found a small signature matrix of 105 genes that can sufficiently define archetype space (**Figure 3.6F**).

The resulting signature contains several NE and non-NE genes that have previously been associated with SCLC subtypes. For example, Transgelin 3 (TAGLN3), growth-hormone-releasing hormone (GHRH), and gastrin-releasing peptide (GRP) are all neuropeptides previously associated with neuroendocrine tumors including SCLC (Bepler et al., 1988; Bostwick and Bensch, 1985; Gola et al., 2006; Ratié et al., 2014; Wang and Conlon, 1993; Zhang et al., 2018), while ASCL1, ISL1, ELF3, and FLI1 are NE transcription factors that drive distinct transcriptional programs in SCLC-A and SCLC-A2 subtypes (Agaimy et al., 2013; Borromeo et al., 2016; Li et al., 2017; Wooten et al., 2019). Several NEUROD family genes are enriched at the SCLC-N archetype, as expected (Borromeo et al., 2016; Osborne et al., 2013; Wooten et al., 2019). The top genes for the SCLC-P archetype have previously been associated with this SCLC subtype and tuft cells (Huang et al., 2018). The top two genes enriched in the SCLC-Y archetype, LGALS1 and

VIM, are associated with a mesenchymal phenotype and have previously been implicated with SCLC chemoresistance (Krohn et al., 2014; Tripathi et al., 2017).

We therefore use this signature matrix to score single cells by least-squares approximation and tested enrichment of these scores near each single-cell archetype. The bulk archetype with the greatest significant enrichment (family-wise error rate $q < 0.1$) labeled each single-cell archetype. Each single-cell archetype was enriched in one of four SCLC signatures: A, A2, N, or Y (**Figure 3.6G**). We visualized the location of the single cells in relationship to these archetypes in two-dimensional space by a Circular A Posteriori (CAP) projection.

Each cell line occupies a distinct region in archetype space, as expected from the bulk transcriptomes (**Figure 3.6B**). While each cell line was predominantly a single subtype, some included single cells that could be classified as generalists, as they fell in between multiple archetypes (**Figure 3.6H**). Critically, these cells are not predicted to be doublets, a technical artifact of scRNA-seq, suggesting they have a truly intermediate cell type. For example, CORL279 forms a continuum of A/N and A2/N generalists, consistent with its dual positivity for ASCL1 and NEUROD1 at the bulk expression level. **(Figure 3.7**). In contrast, H841 is composed entirely of SCLC-Y specialists and non-NE generalists (between Y and another archetype), consistent with its sole expression of YAP1. Our classification was consistent with the bulk expression of the canonical TFs (ASCL1, NEUROD1, POU2F3, and YAP1) in each cell line (**Figure 3.7**). Some intermediate cell types were more common, such as A-N and N-Y generalists, while others were not found or were extremely rare, such as A-Y. Interestingly, H82 spanned states between the A, N, and Y archetypes, which has been shown to be a possible transition path in mouse models (Ireland et al., 2020) and is consistent with its bulk expression of ASCL1, NEUROD1, and YAP1 (**Figure 3.7**).

In conclusion, SCLC cell lines may each comprise archetypal specialists and generalists at the single-cell level. The relative proportion of specialists and generalists varies in each cell line, and generalist cell types may represent intermediate phenotypes or cells transitioning between two archetypes.

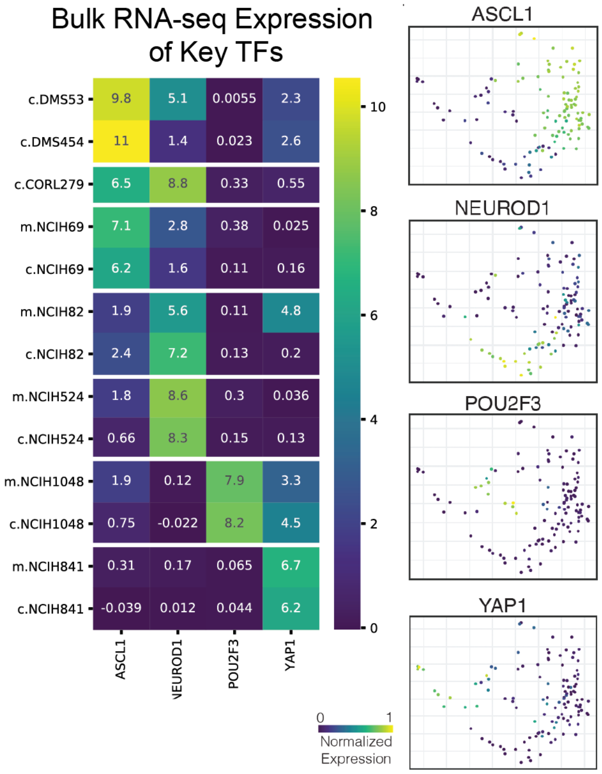


Figure 3.7: Bulk expression of key TFs in human cell lines.

3.2.5. A phenotypic continuum of specialist and generalist cells is detected in SCLC tumors

To determine whether generalists exist in tumors as well, we used the same method of labeling single cell archetypes by bulk archetype gene signatures to analyze scRNA-seq data from SCLC human tumors and genetically engineered mouse models (GEMMs) (**Figure 3.8**).

We sequenced single cells from human tumors from the lungs of two patients who had been treated with and relapsed from the standard of care therapy (etoposide and a platinum-based agent, EP; patient 1 also received prophylactic cranial irradiation; see Methods). Archetype analysis showed that the two tumors fit within a triangle polytope (p = 0.008). Tumor 1 spanned two of the archetypes, one of which was enriched for ASCL1 expression (p = 4.19e-6) and the NE subtypes SCLC-A and SCLC-A2 (**Figure 3.8A-C**). Interestingly, the second archetype did not show significant enrichment in any bulk archetype signatures. Tumor 2 spanned the region between the same A/A2 archetype and an archetype most enriched in the SCLC-Y signature and YAP1 (p=2.1e-49). This is also reflected in the projection of the tumors using the bulk archetype space; Tumor 2 is closer to the SCLC-Y archetype, while most of the variance in Tumor 1 spans the NE archetypes. In both samples, we found subpopulations of generalist cells spanning the archetypes to different degrees, again supporting the existence of intermediate cell states.

We next analyzed three tumors isolated from an Rb1$^{fl/fl}$/Tp53$^{fl/fl}$/Rbl2$^{fl/fl}$ mouse model (labeled TKO1, 2, and 3). TKO1 and TKO2 were primary tumors from independent replicates, and TKO3 was a metastatic tumor from the same mouse as TKO2. Archetype analysis showed that the three tumors fit within a four-vertex polytope (p = 0.001). In both the primary and metastatic tumors, archetype signatures revealed a large proportion of SCLC-A2 (TKO1) or SCLC-A (TKO2 and TKO3) specialists (**Figure 3.8D-F)**. TKO2 and TKO3 also comprised specialists with a high signature score for SCLC-P. In each mouse tumor analyzed, regardless of relative specialist composition, a large proportion of cells were generalists. Thus, intermediate cell states are a staple of GEMM tumors, further supporting the notion of a cell-state continuum.

Taken together, single-cell gene expression data indicated that SCLC cell lines, human tumors, and GEMMs each comprised archetypal specialists and generalists. This characterization of single cells into a continuous phenotypic spectrum between archetypes reveals critical facets of cellular identity that cannot be captured in a discrete framework and may provide insights into the adaptive, dynamic nature postulated for SCLC cells.

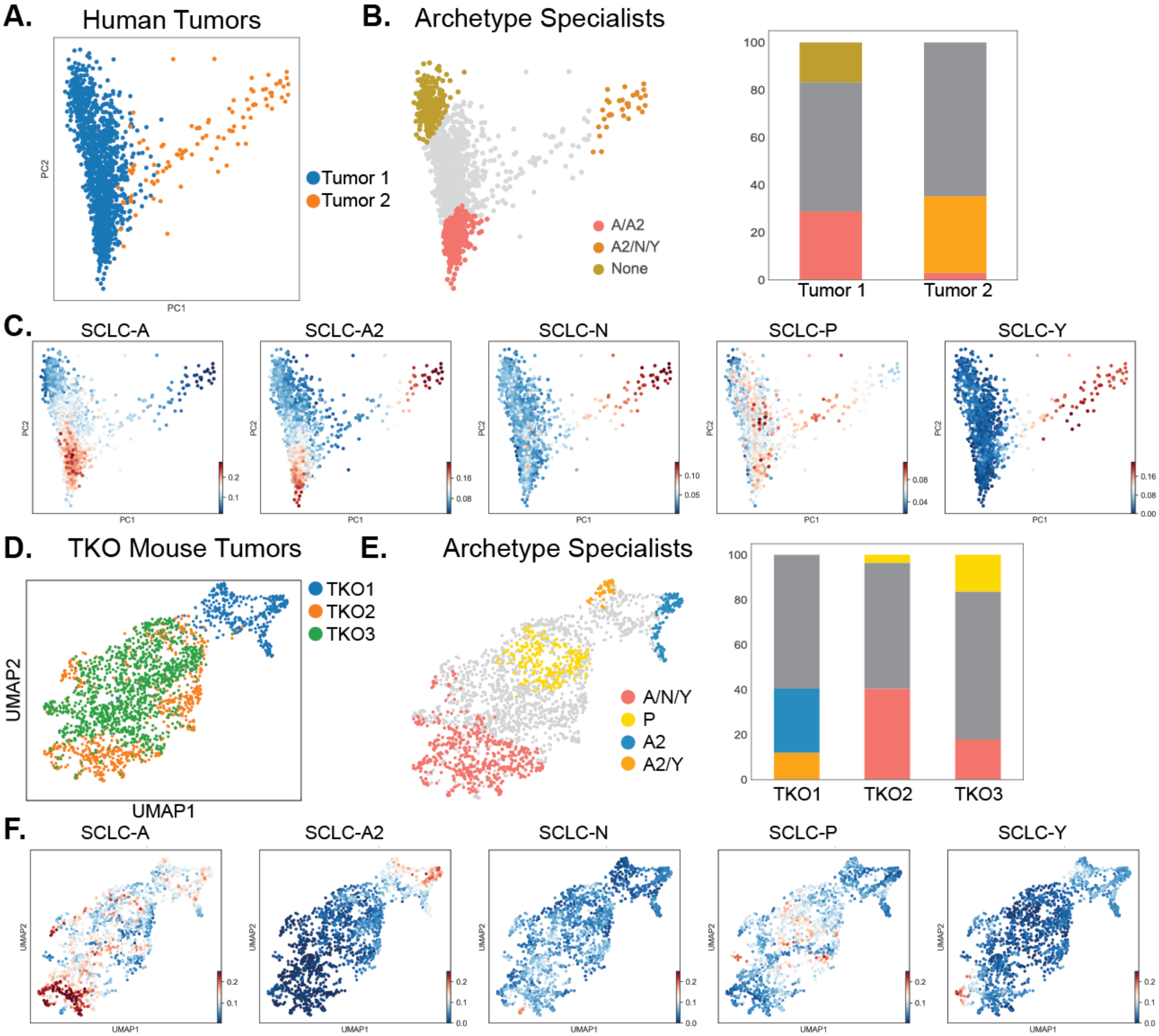


Figure 3.8: Archetype analysis of human tumors and triple knockout (TKO) mouse models. A. *PCA of imputed scRNA-seq from two human tumors.* ***B.*** *Three archetypes best fit the data. Specialists with scores > 0.9 are shown on the PCA projection. Bar plots show proportions of specialists and generalists in each tumor.* ***C.*** *Bulk archetype scores used to label specialists in B.* ***D.*** *Three TKO mouse tumors in a UMAP projection. TKO2 and TKO3 are from the same mouse, contributing to their overlap in the UMAP.* ***E.*** *Four archetypes fit the three TKO tumors. Archetype specialists are shown by color; generalists are shown in grey. Bar plots show proportions of specialists and generalists in each tumor.* ***F.*** *Bulk archetype scores used to label specialists in E.*

3.2.6. Task trade-offs drive transitions in SCLC tumors

The intra-tumoral heterogeneity seen in the mouse and human tumors we analyzed may have arisen due to the phenotypic plasticity of single cancer cells. Phenotypic plasticity, in the context of SCLC archetype space, is tantamount to dynamics of task trade-offs, i.e., state transitions between specialists and/or generalists. We previously showed that a highly plastic non-NE subpopulation emerges in human tumors (Gay et al., 2021). These tumors are largely ASCL1+ before treatment, which raises the possibility of a trade-off between the injury repair task optimized by SCLC-Y specialists, and the proliferation task of SCLC-A, which is susceptible to chemotherapy.

To test this possibility in independent datasets, we focused on task trade-offs along the SCLC-A and SCLC-Y axis, using cell plasticity as a proxy. Previous studies from our co-authors and others showed that SCLC cells transition between A and Y subtypes under certain perturbations, such as Notch pathway activation (Lim et al., 2017) and c-Myc hyperactivation (Ireland et al, 2020; Patel et al., 2021). In these studies, classical neuroendocrine (NE) cells, such as SCLC-A, -A2, and -N, acquire non-NE properties such as variant morphology and expression of non-NE markers (such as YAP1). These observations show c-Myc may be able to control SCLC lineage plasticity and suggest NE subtypes could exhibit increased plasticity under c-Myc activation.

To investigate whether task trade-offs could explain these dynamics, we analyzed a time-course of a genetically engineered mouse tumor with hyperactivation of c-Myc (Rb1$^{fl/fl}$;TP53$^{fl/fl}$; Lox-Stop-Lox [LSL]-Myc^{T58A}, RPM tumors, six time points, **Figure 3.9A**) (Ireland et al., 2020). To align previous subtyping of these time points based on key transcription factors, we tested the enrichment of our bulk archetypal signatures in the single-cell time series dataset (**Figure 3.9B**). Using PCHA, we found that a six-vertex polytope best fit the data (p = 0.001), and 5 of the 6 archetypes were enriched for SCLC signatures (**Figure 3.9C**). This suggests that multi-objective optimization of tasks may have a causal role in the time course.

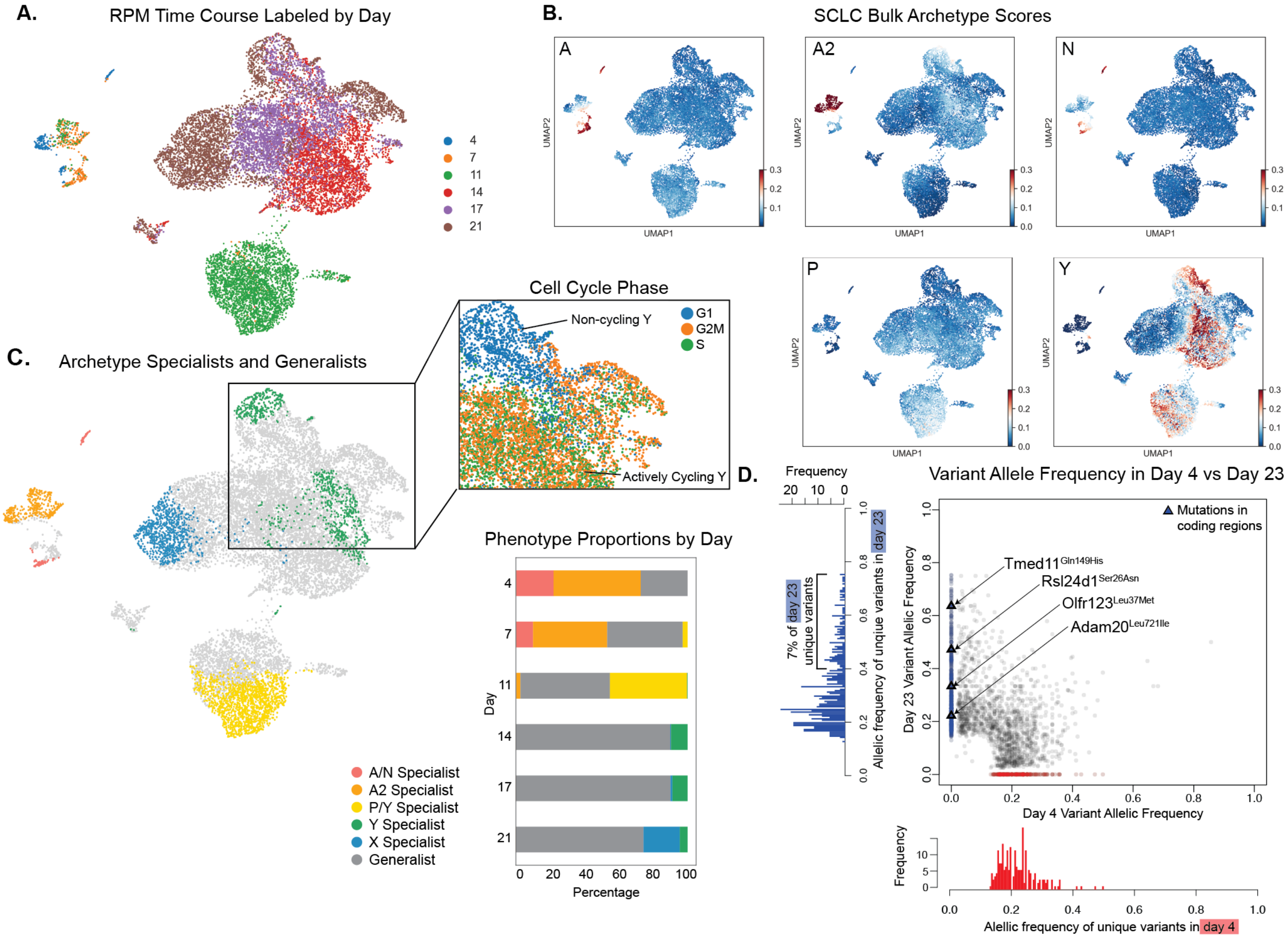


Figure 3.9: MYC-driven tumor progression transitions from NE to non-NE archetypes. A. *UMAP of RPM time course with timepoints labeled. Days 4 and 7 fall in the same region of the UMAP; Day 11 is mostly distinct; and Days 14-21 fall in the same large cluster.* ***B.*** *Bulk archetype signature scores for single cells in time course. Days 4 and 7 are enriched in SCLC-A, -A2, and -N archetype signatures; Day 11 is slightly enriched for SCLC-P and -Y signatures; and a subpopulation of Days 14 to 21 is enriched in the SCLC-Y signature.* ***C.*** *Left: Specialists for 6 archetypes are shown on UMAP, with generalists in grey. 5 of 6 archetypes are enriched in SCLC signatures; the sixth archetype (Blue) is labeled as X. Top right: Two archetypes are enriched for the SCLC-Y signature. One of these archetypes is actively cycling, with cells in the G2M and S phases of the cell cycle. The other is non-cycling. Bottom right: Stacked bar plots show overall subtype composition change.* ***D.*** *Variant allele frequency for beginning (Day 4) and end (Day 23) of an independent RPM time course. Only four variants unique to Day 23 are in coding regions (triangles), and less than 7% of variants are high frequency, suggesting minimal clonal evolution. This supports the notion that phenotype transitions, rather than clonal selection, drive movement from NE to non-NE archetypes.*

As expected, there was a shift from NE subtype cells to non-NE (**Figure 3.9C**). Specifically, at the earliest time points (day 4 and day 7) tumors were largely composed of SCLC-A/N and SCLC-A2 specialist cells (>50%), forming a continuum of specialists and generalists near the NE archetypes. By day 11 the population of cells was near an SCLC-P/Y archetype **(Figure**

3.9C). Two archetypes in the dataset were enriched in the SCLC-Y signature (green in **Figure 3.9C**). While the transcriptomic profiles were similar, one key difference between these archetypes was whether they were actively cycling; one was dominated by G2M and S genes, while the other contained cells mostly in the G1 phase. From day 14 to 21, cells move towards these SCLC-Y archetypes, consistent with the increase in YAP1 expression found in Ireland et al. (2020). Interestingly, by day 21, cells fall near a new archetype that is not enriched in any of the SCLC signatures (X specialists, blue in **Figure 3.9C**). Gene set enrichment analysis (GSEA) showed that archetype X is enriched for the following hallmark gene sets: MYC targets, oxidative phosphorylation, reactive oxygen species (ROS) pathway, and glycolysis. Archetype X is significantly depleted in hallmark gene sets related cell cycle terms (mitotic spindle and G2M checkpoint) and hypoxia. Further research will be necessary to characterize this non-NE archetype. The changing proportions of archetypal subpopulations over the time course suggests that cells maybe trading off between the NE and non-NE archetypal tasks.

We sought to validate that cell state transitions, rather than clonal selection, were responsible for the shift in phenotype from NE to non-NE. To this end, we performed whole-genome sequencing on independent samples from day 4 and day 23 (**Figure 3.9D**). We filtered variants by read depth and compared the frequency of variants across the two time points. If clonal selection of a pre-existing non-NE subpopulation was driving the dynamics of the time course, we would expect to see a substantial number of subclonal variants in the day 4 sample increase in allelic frequency in the day 23 sample. Instead, we found that only 7% of the total somatic variants were unique to, and had high allelic frequencies, on day 23 (greater than 0.4). Furthermore, only four of the variants unique to day 23 are in coding regions of genes (shown as triangles in **Figure 3.9D**). None of the four genes are associated with SCLC phenotype identity and show low to no expression dynamics in the scRNA-seq data, suggesting these variants do not drive phenotypic evolution. Together, this shows there is minimal genetic evolution between days 4 and 23, and the transformation of cell state over this time course is due to phenotypic transitions rather than clonal selection. Together, these results show that RPM tumor cells can transition between NE and non-NE states as a result of archetype task optimization.

3.2.7. Plasticity analysis identifies regulators of task trade-offs

We next sought a method that could deconvolve two aspects of plasticity, reflecting two distinct qualities of the underlying phenotypic landscape: containment potential and drift potential (Weinreb et al., 2018). Containment potential should be reflected in the multipotency of cells. Therefore, we examined whether cell progressed along multiple lineages using CellRank (Lange et al., 2022). To approximate drift potential, we calculate an expected distance of transition for every single cell, here termed Cell Transport Potential (CTrP), to reflect movement across phenotypic space (see Methods).

First, to determine the transition paths of cells along the time course, we applied RNA velocity analysis using scVelo (**Figure 3.10A**) (Bergen et al., 2020; Manno et al., 2018). We fit each gene using a dynamical model and investigated the genes with top fit likelihoods (see Methods). A gene set enrichment analysis (GSEA) shows that genes ranked by their fit likelihood were enriched for MYC target genes ($q = 0.000$), corroborating that MYC is critical for driving the transitions across timepoints. We next used EnrichR (Chen et al., 2013) to investigate transcription factors that regulate the top fit genes (fit likelihood > 0.3) which validated MYC as an important regulator of the velocity dynamics (**Figure 3.10B**). E2F family proteins, REST, and SMAD4 were also identified as regulators (**Figure 3.10C**), which have previously been implicated in the progression of SCLC (Lim et al., 2017; Wang et al., 2017; Wooten et al., 2019).

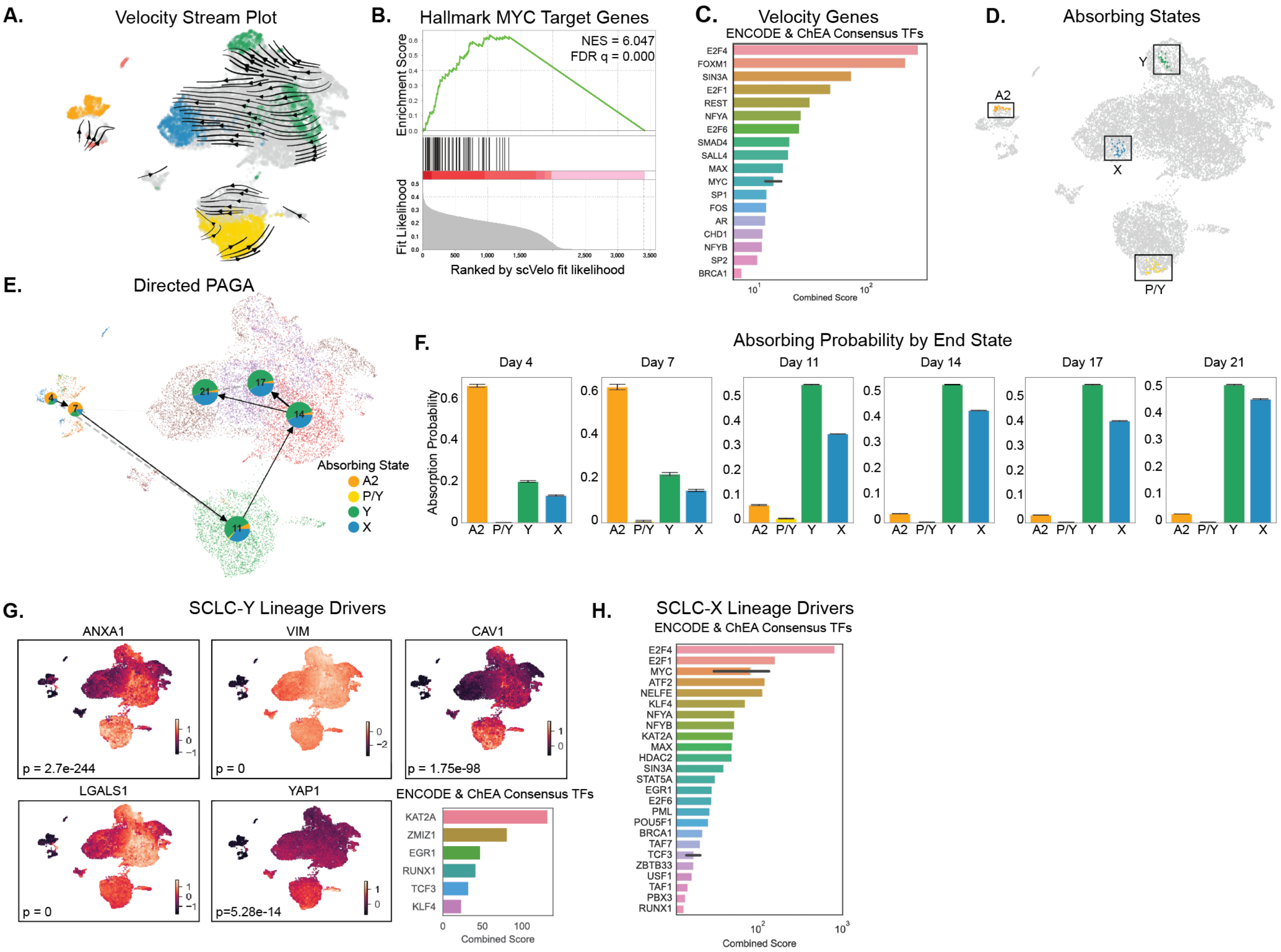


Figure 3.10: RNA velocity analysis identifies lineage drivers and high plasticity cells in RPM time course. A. *RNA velocity shows transition across the time course in UMAP projection.* ***B.*** *Hallmark gene set of MYC targets is enriched in gene set with high fit likelihoods for dynamical RNA velocity model.* ***C.*** *ENCODE and ChEA consensus TFs from EnrichR analysis of top fit likelihood genes (likelihood > 0.3). Consensus score from EnrichR shown. For genes from both sources (i.e. ENCODE and ChEA both have the TF), a black bar shows 95% confidence interval on mean consensus score. E2F family genes and MYC are key drivers of the transition.* ***D.*** *Using CellRank, we fit a Markov transition matrix to these dynamics using a weighted kernel of the RNA velocity (weight = 0.8) and diffusion pseudotime (DPT) calculated in Ireland et al. (2020) (weight = 0.2). Using the CellRank implementation of a GPCCA estimator, we find end states for the Markov chain model and display the top 30 most likely cells for each absorbing (end) state.* ***E.*** *PAGA plot shows significant transitions between time points. Pie plots overlaid on PAGA show aggregate lineage probabilities by timepoint.* ***F.*** *Aggregate lineage probabilities by timepoint shown as bar plot, with absorption probability on y-axis.* ***G.*** *Lineage drivers of the SCLC-Y lineage. Genes correlated to absorption probabilities for the SCLC-Y lineage are considered drivers of that lineage. UMAP with expression of select lineage drivers from the SCLC-Y archetype signature are shown. EnrichR analysis shows TF regulators, ranked by consensus score, of the top 40 significant lineage drivers sorted by correlation with lineage. TCF3 is in the SCLC network described in Wooten et al. (2019); RUNX1 was predicted to regulate an intermediate osteogenic state in an RPM mouse model with inactivated ASCL1 (Olsen et al., 2021).* ***H.*** *TF regulators of lineage drivers for the X absorbing state. As in (K), EnrichR was used to rank regulators by consensus score. E2F family genes, MYC, and RUNX1 are regulators of the X lineage.* ***I.*** *Cell transport potential shows most plastic subtypes across the time course. Cells closer to the NE archetypes SCLC-A and -A2 have higher plasticity in earlier time points. CTrP decreases over time, consistent with cells that transition from NE phenotypes to non-NE phenotypes with lower plasticity.*

Using CellRank (Lange et al., 2022), we fit a Markov chain model by combining two sources of dynamic information: diffusion pseudotime calculated in Ireland et al. (2020) and RNA velocity. We find four regions of end states (absorbing states, **Figure 3.10D**), two in earlier timepoints (days 7 and 11) and two in later timepoints (days 17 and 21). Interestingly, all of the absorbing states are in specialist regions rather than generalists (SCLC-A2, P/Y, Y, and X specialists). A coarse-grained PAGA graph shows significant transitions between timepoints as expected, with varying proportions of cells in each timepoint transitioning towards each end state (**Figure 3.10E**). While about two thirds of the cells in days 4 and 7 transition towards the A2 end state in day 7, the remaining third transitions towards the Y and X end states. The remaining timepoints (11-21) are split between the SCLC-Y and X lineages (**Figure 3.10F**).

We then correlated probabilities of absorption at either end state with gene expression to find potential lineage drivers for SCLC-Y and X and applied EnrichR to investigate transcription factors (TFs) regulating these genes (**Figure 3.10G-H**). VIM and LGALS1 were both top SCLC-Y lineage drivers, consistent with their presence in our SCLC-Y archetype signature (**Figure 3.10G**). In fact, 19 of 24 genes from the SCLC-Y signature (**Figure 3.6F**) were identified as significant lineage drivers ($q < 0.05$), confirming their role in driving this phenotype. The top SCLC-Y lineage drivers were regulated by TCF3 and RUNX1, which we previously showed may be important in SCLC progression (**Figure 3.10G**) (Olsen et al., 2021; Wooten et al., 2019).

SCLC-X lineage drivers are regulated by MYC, RUNX1, and E2F family genes, suggesting MYC activation is key to reaching this archetype (**Figure 3.10H**). Furthermore, ChEA identified as X lineage regulators several TFs that are important for maintenance of pluripotent stem cells, such as OCT4, NANOG, and SOX2 (**Figure 3.11A**). To determine if the TF regulators

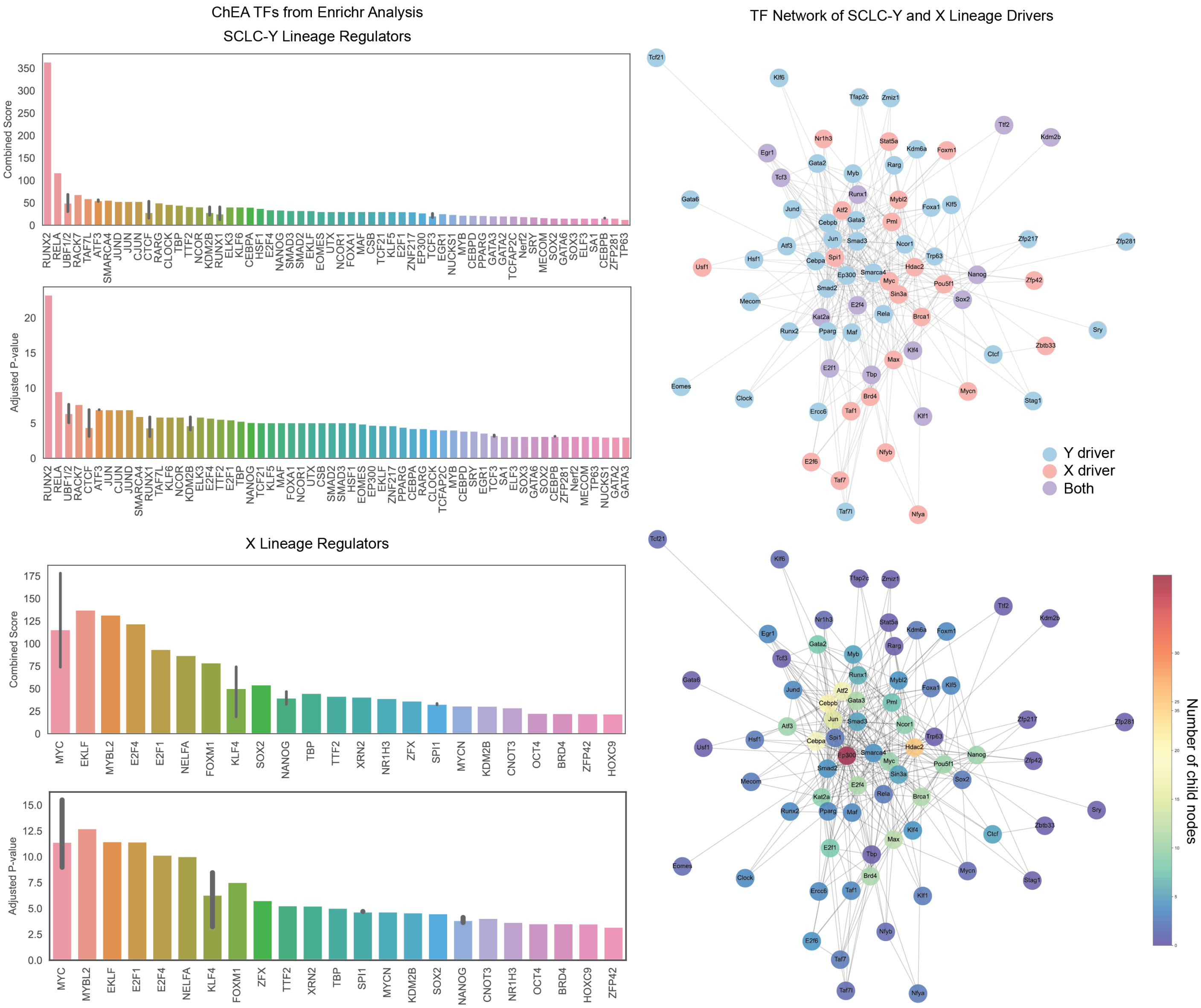


Figure 3.11: SCLC-Y and SCLC-X share TF drivers in a gene regulatory network A. *TFs that are significant regulators of SCLC-Y and X lineages from Figure 3.10. Shown are TFs from ChEA only.* ***B.*** *TF network of lineage drivers from Figure 3.10 and 3.11A.* ***Top****: TFs are colored by lineage; some TFs are shared between lineages and shown in purple. Only TFs that connect to this single main network are shown.* ***Bottom****: TFs are colored by number of child nodes they regulate in the network. P300 regulates the most child nodes at 38, making it the most central node to the lineage drivers in this time course. Other central TFs include MYC, CEBP family genes, JUN, and RUNX1/2.*

of the SCLC-Y and X lineages interact, we used STRING to construct a regulatory network (**Figure 3.11B**) (Snel, 2000; Szklarczyk et al., 2020). Twelve of the 86 drivers regulated both lineages, including SOX2, RUNX1, and KLF and E2F family genes. An analysis of centrality demonstrated that p300, which is often mutated in SCLC and may be associated with poor prognosis (Gao et al., 2014; George et al., 2015; Hou et al., 2018; Jia et al., 2018), regulates the most child nodes (38) in the network. Other central TFs include MYC, JUN, which is important for the SCLC-to-NSCLC transition (Risse-Hackl et al., 1998; Shimizu et al., 2008), and CEBP family genes, which have been shown to play a vital role in inflammatory diseases, including cancer (Chi et al., 2021).

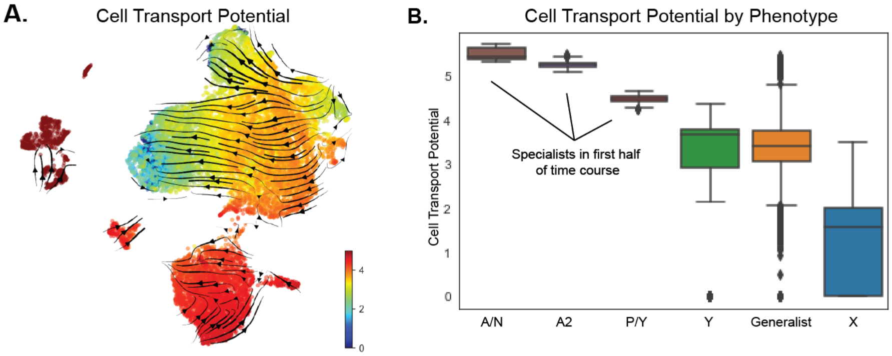


Figure 3.12: Cell Transport Potential shows NE archetypes have highest plasticity. A. *CTrP shown on UMAP with RNA velocity overlaid.* ***B.*** *CTrP decreases from early-timepoint archetypes, including A/N, A2, and P/Y to the X archetype. While SCLC-Y is an absorbing state in the system, many of the Y specialists still have high plasticity.*

Finally, we applied our CTrP pipeline. As expected for a time course of phenotype-transitioning cells, transport potential decreased steadily over the time course (**Figure 3.12A**). Despite the presence of early timepoint end states (A2 and P/Y), all specialist cells in early timepoints had higher CTrP than later timepoints (**Figure 3.12B)**. Together, our plasticity analysis suggests that MYC may be capable of increasing the plasticity of early time-point cells, or NE specialists, allowing them to transition to the non-NE SCLC-Y archetype and a new archetype regulated by pluripotency TFs.

3.2.8. Network analysis validates the role of MYC in driving SCLC plasticity

To gain mechanistic insights into the effect of MYC on plasticity, we introduced MYC into an SCLC-specific transcription factor (TF) network (**Figure 3.13A**). As described in Wooten et al. (Wooten et al., 2019), computer simulations of this TF network dynamics reveal attractors (i.e., network equilibrium states) that correspond well to the experimentally defined SCLC subtypes. The stability of these attractors (i.e. subtypes) can be quantified with the BooleaBayes algorithm. To mirror the experimental conditions of Ireland et al., we imposed constitutive activation to the MYC node in simulations of dynamics of the SCLC TF network. This modification decreased the number of steps needed to leave the NE attractors; in other words, MYC activation destabilized the SCLC-A and SCLC-A2 attractors but did not significantly destabilize the SCLC-N or the non-NE SCLC-Y attractors (**Figure 3.13B**). The *in-silico* perturbations suggest that activation of MYC and the subsequent epigenetic regulations may be able to shift an NE-phenotype cell to a non-NE one by destabilizing the NE attractor (cell state). Further experimental validation is needed to determine whether MYC activation is sufficient for this phenotype shift to occur.

We next investigated this effect of MYC by measuring plasticity at the single-cell level in a human tumor dataset comprising two PDXs from the same SCLC patient, generated before and after relapse following chemotherapy. At the single-cell level, MYC expression was higher in the PDX after relapse than the PDX before treatment, consistent with genomic amplification of MYC in the tumor following treatment. CTrP analysis confirmed that plasticity was correlated with MYC expression in each of the tumors.

The above independent lines of evidence indicate (1) MYC hyperactivation can drive phenotype transitions from NE to non-NE states, as confirmed by whole genome sequencing showing little clonal evolution; (2) MYC may be capable of increasing the plasticity of NE subtypes, as demonstrated by *in silico* simulations and RNA velocity analysis; and (3) a correlation in MYC expression and plasticity after treatment may point to MYC's role in SCLC tumors acquiring resistance. Together, this suggests that upregulation or activation of Myc can increase NE cell plasticity to promote cell state transitions toward a non-NE state, which may help cancer cells overcome treatment.

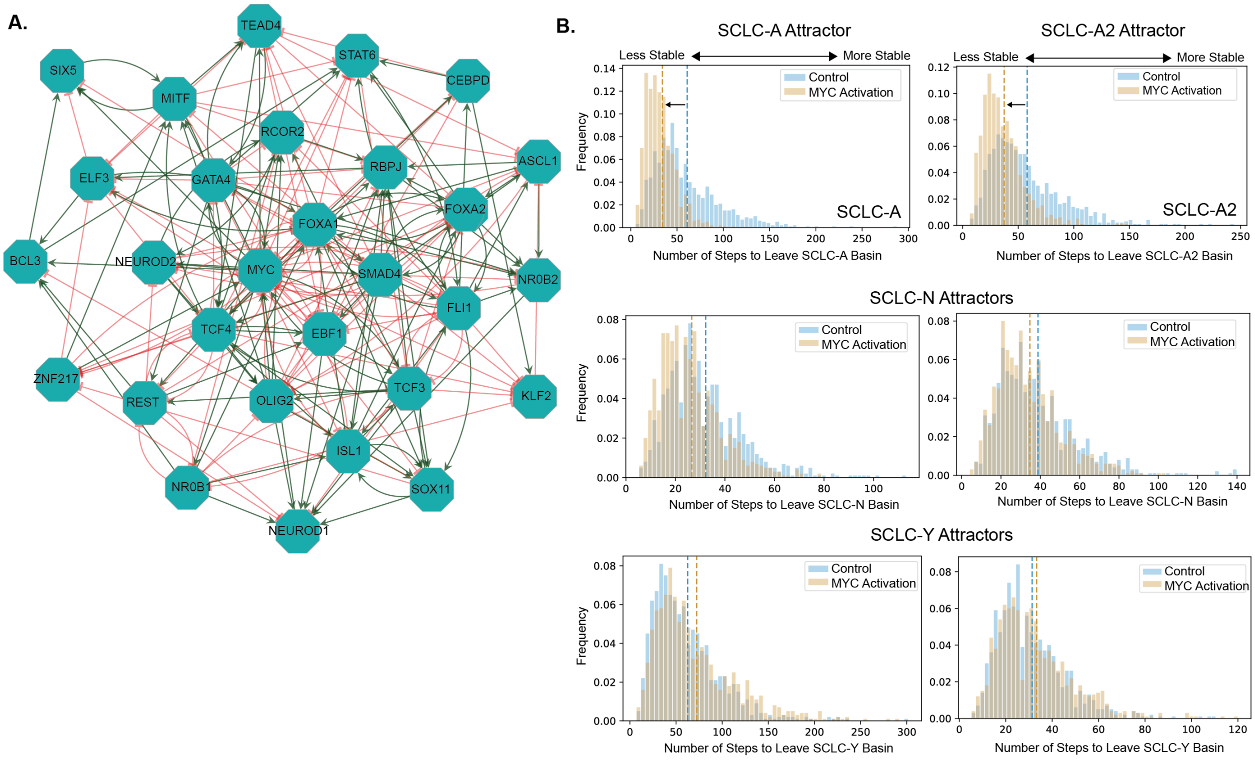


Figure 3.13: MYC activation destabilizes NE states. A. *Transcription factor network adapted from Wooten et al. to incorporate MYC activity.* ***B.*** *In silico destabilization of NE specialists by MYC activation. Using BooleaBayes simulations (Wooten et al., 2019), we performed random walks with activated MYC and found that SCLC-A and SCLC-A2 states are destabilized, i.e. MYC activation is capable of increasing plasticity of these subtypes in RPM tumors. SCLC-N and SCLC-Y attractors were not significantly destabilized.*

3.3. Discussion

SCLC is a heterogeneous cancer comprising neuroendocrine (NE) and non-neuroendocrine (non-NE) subtypes, classified by eponymous transcription factors (Rudin et al., 2019). Our goal in this study was to understand dynamics amongst these subtypes since plasticity is likely to play a crucial role in supporting the aggressive features of SCLC (Ireland et al., 2020; Lim et al., 2017; Stewart et al., 2020). In analyzing SCLC datasets from diverse sources, we realized that applying the current discrete subtype classification is insufficient to capture subtype dynamics because many SCLC cells in cell lines and tumors fall between distinct subtypes. Therefore, our understanding of SCLC plasticity was limited by the lack of (i) continuous definitions of cell state and (ii) quantitative metrics for single-cell plasticity.

We propose an alternative, continuous view of SCLC heterogeneity based on SCLC archetypes defined by functional tasks. While there was a high concordance between archetypes and canonical subtypes, the archetype-bounded phenotypic space paradigm presented several advantages that better represent SCLC heterogeneity. First, the transcriptional profile of every single cell can be evaluated based on distance from archetypes and graded as a specialist or generalist (e.g., a cell between archetypes N and Y has a generalist phenotype with a high degree of N and Y character). Second, the plasticity of phenotypes can be quantified by tracing transition paths between archetypes and identifying regions of high SCLC cell plasticity. Third, cell state transitions are rooted in multi-objective evolutionary theory such that movement across the continuum fulfills the goal of trading off between tasks, providing a functional interpretation of SCLC phenotypes. Lastly, we can identify epigenetic strategies for targeting plastic SCLC cells, which we propose is a high priority for effective SCLC treatment.

3.3.1. Cooperation of SCLC Archetypal Tasks

Using gene set enrichment analysis, we identified tasks optimized by each specialist cell type that mirror tasks fulfilled by normal PNECs. We then projected single-cell data into an archetype-defined polytope and found intratumoral heterogeneity aligns with intra-sample diversity, with single cells capable of optimizing varying tasks within a single tumor. This palette of biological tasks within a cell line or tumor agrees with recent reports indicating that lung tumors are capable of building their own microenvironment, where SCLC cell types (NE and non-NE) were found to interact in a way that is mutually beneficial to the growth of the tumor (Calbo et al., 2011; Huch and Rawlins, 2017; Kwon et al., 2015; Lim et al., 2017). Similarly, we expect SCLC cells optimizing archetypal functions to cooperate *in vivo* by performing PNEC-related tasks that contribute to the growth of a tumor in the face of changing external conditions, such as treatment. It remains to be seen whether the normal functions of PNECs represent an actionable constraint for SCLC cells.

Our analysis suggests that multi-objective optimization under Pareto theory shapes SCLC phenotypic space, supported by the enriched gene programs and experimentally tested tasks of each archetype. However, a polytope could result from other phenomena. For example, each archetype could correspond to a weighted average of five transcriptional profiles. While we show preliminary experimental evidence that each archetype optimizes a specific task, further work is needed to validate task trade-offs characteristic of Pareto optimality. Phenotypic perturbation experiments may help determine the cost trade-off between archetypes and uncover the relationship between archetypal task optimization and tumor fitness. For example, Archetype 1 (SCLC-A) cells optimize proliferation (function) and, therefore, are highly chemosensitive (cost). In contrast, a transition to Archetype 5 (SCLC-Y) under chemotherapy may decrease the rate of growth of a tumor (cost) but are better able to respond to cell injury and may therefore better survive treatment (function).

3.3.2. SCLC and PNEC Plasticity

We find that this heterogeneous ecosystem of phenotypes arises in SCLC tumors due to cell state transitions. By quantifying Cell Transport Potential, we uncovered subpopulations of high plasticity, capable of transitioning to multiple other phenotypes. We speculate that the plasticity of SCLC cells may derive from dysregulation of the innate plasticity in normal PNECs.

After injury to the lung epithelium, “specialist” stem-like PNECs can transdifferentiate to perform repair tasks and regenerate “specialist” club cells, whose main task is the secretion of protective proteins, most likely through non-genetic mechanisms (Oudah et al., 2019). As shown in the tumors analyzed here, SCLC cells can likewise transition between NE and non-NE phenotypes.

It is tempting to speculate that such levels of adaptability may be responsible for the highly aggressive features of SCLC tumors. For instance, an altered balance in favor of the wound-healing SCLC-Y specialists may be expected in tumors immediately after treatment, supported by our limited treated tumor data here and could be further tested experimentally in GEMM or PDX tumors. These dynamics could explain the initial response to chemotherapy seen in patients, which is inevitably followed by relapse as cells transition to generalist and non-NE specialist cells better equipped to overcome chemotherapy.

3.3.3. Controlling plasticity in SCLC

Previously, a subset of SCLC cells has been shown to be capable of long-term propagation of tumors (Tumor propagating cells, TPCs) (Jahchan et al., 2016), and it is unclear how these cells relate to the archetypes described here or our definition of plastic potential. While SCLC-A cells express markers for TPCs (positive for *EPCAM, MYCL*, and *CD24*, and negative for *CD44*), it remains to be seen whether SCLC-A cells correspond to TPCs functionally or if TPCs can span archetype space. Similarly, a PLCG2-expressing stem-like subpopulation was recently reported in a survey of human SCLC tumors (Chan et al., 2021). This stem-like cell may be consistent with a diverse, stem-like functional state since it is present across SCLC-A, -N, and -P tumors. Interestingly, PLCG2, enriched in SCLC-P, was present in our archetype signature. Further work is needed to understand the relationship between this archetype and stemness.

Plasticity is dependent on the underlying genetics that determine the shape of the phenotypic landscape, the particular cellular state in which a cell resides, and any external conditions that may transiently distort the landscape. For this reason, epigenetic methods may directly target plasticity, such as gene regulatory network perturbations. For example, two PDX models from a single patient show elevated levels of MYC after relapse that is correlated with plasticity. There are two possibilities for why MYC expression may correlate with plasticity: (1) MYC may be driving an increase in plasticity, or (2) another driving genetic or epigenetic mechanism may increase both MYC and plasticity. Our analyses of a hyperactivated MYC GEMM show that MYC may promote NE cell plasticity, consistent with studies that MYC overexpression in human ASCL1+ cells can promote N and Y subtypes (Ireland et al., 2020; Patel et al., 2021). Therefore, MYC itself seems capable of driving the increase in plasticity seen after treatment.

Furthermore, previous research suggests that MYC may play a role in genome-wide transcriptional upregulation, allowing cells to change expressed gene programs and thus phenotype (Lin et al., 2012). In other words, MYC may allow cells to "move further" in gene expression space, exhibited by increased CTrP. However, future studies, such as using an inducible MYC model in GEMMs or PDXs, will be necessary to determine the complete mechanism underlying the relationship between MYC and phenotype plasticity.

3.3.4. Task trade-offs and acquired resistance

The current standard of care for SCLC is predicated upon targeting highly proliferative cells. However, this treatment inevitably results in resistant relapse. Highly plastic cells detected in SCLC cell lines and tumors suggests that plasticity may drive resistance in SCLC, consistent with a recent study showing increased intratumoral heterogeneity upon chemotherapy relapse (Stewart et al., 2020). The phenotypic continuum also shows that plasticity enables SCLC cells to trade-off PNEC-related tasks, which translates to a high level of adaptability to diverse microenvironments. Thus, plasticity may also be responsible for SCLC aggressive traits, such as local invasion and early metastatic spread. Therefore, we propose epigenetic strategies to target plasticity directly that can be derived from analyses of TF network dynamics, such as MYC inhibition. Given the primary role of TFs in driving SCLC phenotype (Wooten et al., 2019), SCLC should be a prime candidate for plasticity-targeted therapy.

3.4. Methods

3.4.1. Bulk SCLC Cell Line RNA-seq Data Preprocessing

Bulk RNA sequencing expression data on SCLC cell lines were taken from two sources: 50 cell lines were taken from the Cancer Cell Line Encyclopedia (as in Chapter 2) and 70 cell lines (not including H69 variants) were taken from cBioPortal (Cerami et al., 2012; Gao et al., 2013) deposited by Dr. John Minna (2017). Access to data from cBioPortal was provided by participation in the NCI SCLC Consortium. 29 cell lines overlapped between the datasets, so a "c" (CCLE) or "m" (Minna) was used to denote the source of each cell line. Each dataset was filtered and normalized independently and then batch corrected together. For each source, genes and cell lines with all NAs were removed, as well as mitochondrial genes. The counts data was then normalized by library size and transcript abundance to TPM values. The two datasets were combined using overlapping genes and log-transformed, and genes with low expression across all samples were removed (cutoff of log(TPM) >= 1). The two datasets were then batch corrected using the *sva* R package, which includes a ComBat-based integration method (Johnson et al., 2007; Leek and Storey, 2007). SVA, or surrogate variable analysis, uses a null model and a full model to derive hidden variables, such as batch, that may contribute to gene expression variance across samples. The four SCLC TF factors that define broad subtypes— ASCL1 (A), NEUROD1 (N), YAP1 (Y), and POU2F3 (P)— were used to align the two datasets to each other. The resulting dataset contained 120 samples and 15,950 genes.

For labeling cell lines by subtype cluster in **Figure 3.1A**, the clustering method in Chapter 2 was adapted for the expanded dataset. Briefly, hierarchical clustering with the Spearman distance metric was calculated. Cell lines previously characterized as SCLC-A in Chapter 2 comprised two branches of the dendrogram separated by SCLC-N cell lines, most likely due to the dual positivity of ASCL1 and NEUROD1 in some SCLC-A cell lines. Cell line H82 (from both data sources) was considered "unclustered," as it was considered an SCLC-N cell in Chapter 2 but was clustered with SCLC-Y cell lines here. PCA was run on the bulk RNA-seq dataset and the elbow method on explained variance per component was used to choose 12 principal components for downstream analysis. The top 12 principal components were able to explain ~50% of the variance in the dataset, suggesting a low-dimensional representation of the data was possible.

To identify gene programs associated with the five SCLC phenotypes found with clustering, we performed weighted gene co-expression network analysis (WGCNA) on the same RNA-seq data (Langfelder and Horvath, 2008) as in Chapter 2.

3.4.2. Archetypal Analysis using PCHA

Using an AA method known as Principle Convex Hull Analysis (PCHA) we found a low-dimensional Principal Convex Hull (PCH) for the cell line dataset (Mørup and Hansen, 2012). The convex hull is the minimal convex set of data points that can envelop the whole dataset. The PCH is a subset of the convex hull comprising a set of vertices, or archetypes, that form a polytope able to capture the shape of the data. The vertices are constrained to be a weighted average of the data points, and the data points are then approximated as a weighted average of the vertices. The algorithm solves the optimization problem of minimizing the norm of the approximated data subtracted from the original data. The algorithm constraints can be relaxed such that the vertices can be found within a certain volume around the convex hull; i.e. relaxing the constraint that the vertices must fall *on* the convex hull. By comparing the convex hull to the PCH, we can determine how well a low-dimensional shape fits the dataset with a statistical test, and thus ascertain the optimal number of vertices, or archetypes, that best define the shape.

Archetypal Analysis was done using the Matlab package *ParTI* (Hart et al., 2015). To determine the best k, we found the explained variance (EV) for each number of archetypes *k,* which is computed by the PCHA algorithm (in the function *ParTI_lite*) as previously described (Cutler and Breiman, 1994; Korem et al., 2015) Then, we chose a number of archetypes, *k**, for which the EV doesn't increase by much when adding additional archetypes. In practice, this is done by finding the "elbow" of the EV versus *k* curve, which is the most distant point from the line that passes through the first (k=2) and the last ($k = k_{max}= 15$) points in the graph. Because this could be dependent on our choice of k_{max}, we varied k_{max} between 8 and 15 and found *k** in each case. *K** = 5 for 6 of 8 EV versus k plots; *k** = 4 when $k_{max} = 12$, and *k** = 6 when $k_{max} = 11$. Therefore, we proceeded with our analysis using k* = 4, 5, or 6.

The function *ParTI* was then used for the full analysis for each k* with parameters dim = 12 (dimensions) and algNum = 5 (PCHA) to find the location of the archetypes in gene expression space. To measure the similarity between the data and a polytope is its t-ratio. This was calculated by comparing the ratio of the polytope volume to that of the convex hull. For PCHA, a bigger t-ratio suggests the polytope is more similar to the data (and thus a better fit). Empirical p-values were calculated by comparing the t-ratio of the data to that of 1000 sets of shuffled data, as described in Korem et al. (2015). The p-value was defined as the fraction of sets for which the t-ratio is equal to or larger than that of the data. The p-value for 5 archetypes, $p = 0.034$, suggests the polytope fits the data well. As a side note, we also tested k* = 3, even though it was not suggested by ParTI software. We found this polytope has an insignificant p-value of 0.58, suggesting it does not fit the data well.

We adapted a method from Hausser et al. (2019) to compare the archetypes for 4, 5, and 6-vertex polytopes. Briefly, we compare the significantly enriched (FDR <10% and log2 fold change > 0.1) gene sets at each archetype (described below in "Gene and ontology enrichment at archetype locations") using a hypergeometric test. This allows us to test if the number of overlapping enriched sets for the two archetypes is significantly higher than expected, given the null hypothesis of random sampling from the union of gene sets found at any archetype in a given polytope.

Errors were then calculated on each archetype location by sampling the data with replacement and calculating the archetypes on the bootstrapped data sets (1000 times). Error on the archetypes gives an idea of the variance in archetype position expected, and a smaller variance suggests the archetype is robust to outlier samples. In the 5-vertex polytope, the errors on each archetype are relatively small, suggesting none are dependent on outliers in the data.

3.4.3. Enrichment of cluster labels, genes, and gene ontologies at archetypes

Enrichment of subtype labels was determined using the *ParTI* function *DiscreteEnrichment*. Cell lines were binned into 10 bins according to distance from each archetype. For each subtype label (from hierarchical clustering), the percentage of labels in the bin closest to the archetype was compared to the percentage in the rest of the data using a hypergeometric test. Enrichment was considered significant if the bin closest to the archetype was maximal for that label and the FDR-corrected p-value for the hypergeometric test was significant (Benjamini-Hochberg, $q < 0.1$). After binning the data into ten bins by distance to archetype, we found that each archetype was enriched in cell lines from one of the five SCLC subtypes.

To find genes enriched by each archetype, we tested the enrichment of each feature on the bin closest to archetypes versus the rest of the data. The *ParTI* function *ContinuousEnrichment* was used to analyze gene expression of all 15,950 genes and Cancer Hallmark Gene Sets. The expression of each feature was compared between the closest bin to each archetype and the rest of the data using a Mann-Whitney test (FDR-corrected p-value, q<0.1). To determine PNEC functions enriched at each archetype, we used ConsensusPathDB (Kamburov et al., 2013) on the top 300 most enriched genes for each archetype, as well as for PNECs and other airway cell types from Montoro et al (Montoro et al., 2018). (using *Scanpy's* function rank_genes_groups). As described in Chapter 2, we used t-SNE to cluster the GO terms, using distances from GOSemSim (Yu et al., 2010). We then chose clusters with GO terms related to PNEC tasks and evaluated enrichment of these terms at each archetype.

3.4.4. Archetype analysis on bulk RNA-seq from 81 human tumors

We chose to define archetypes on cell lines because cell lines are generally thought to be less heterogeneous than tumor samples, and therefore may better represent extreme, pure phenotypes rather than mixed (averaged) phenotypes of tumors. To test whether it is true that cell lines better represent extreme phenotypes in our particular SCLC samples, we combined our dataset of 120 cell lines with an independent dataset of 81 human SCLC tumors (George et al., 2015) and analyzed the relationships between the cell lines and tumor samples. First, we batch corrected this data using SVA as described above. Cell line and tumor RNA-seq datasets were preprocessed as described above. The ComBat method with a scale adjustment best aligned the distributions of log-transformed expression.

We next fit a principal components analysis (PCA) model to this combined dataset. We find that the tumor samples tend to be contained within the same archetype space as defined by cell lines. Next, we compared a PCA model fit to tumor samples only to this PCA on the combined dataset, to determine how variance in the cell lines differs from variance in the tumors. We find that the top PCs of the tumor-only model match the top PCs in the combined-dataset model.

Once we determined that the variance in SCLC cell lines and tumors is aligned, we applied archetype analysis directly to the tumor data. We use the ParTI MATLAB package to run PCHA and find archetypes associated with (a) the combined dataset and (b) the tumor data only, and used a method described in Hausser et al. (2019) to match the archetypes to our 5 cell-line archetypes.

To find the best number of archetypes $k*$ that explains this combined dataset, we found the "elbow" of the Explained Variance (EV) versus k curve, which is the most distant point from the line that passes through the first (k=2) and the last ($k = k_{max}= 15$) points in the graph. Because this could be dependent on our choice of k_{max}, we varied k_{max} between 8 and 15 and found $k*$ in each case. The Elbow method on the EV vs k plot suggests k* = 4 (when $k_{max} = 8$), 5 (k_{max} = 9, 10, 11, or 15) or 6 (k_{max} = 12, 13, or 14) archetypes best fit the tumor data. We ran ParTI for k* = 4, 5, or 6 and found that 5 archetypes gave the most significant p-value (p = 0.59 for k*=4; p = 0.09 for k*=5; p =0.33 for k* = 6).

We then compared tumor-only archetypes to the archetypes defined by cell lines alone. Using the elbow method and varying k_{max} between 8 and 15, PCHA suggested k* = 3 (k_{max} = 8, 10), k* = 4 (k_{max} = 9, 12), k* = 5 (k_{max} = 11, 13), or k* = 6 (k_{max} = 14, 15). Interestingly, no best fit polytope with 3 to 7 vertices was significant according to its t-ratio. We wondered if this was due to our hypothesis that tumors are more mixed than cell lines as mixing cell types in different proportions should produce polytopes in linear but not logarithmic space. We, therefore, looked for polytopes in linear gene expression space by exponentiating gene expression and subtracting 1 (inverse operation of log(x+1)) before mean-centering the data and performing a PCA. In linear space, the elbow method suggested k* = 5 for all k_{max} between 8 and 15. To ensure k* = 5 is the best fit polytope, we tested k* between 3 and 7. We found k*= 5, 6, or 7 were all significant with p-values of 0.021, 0, and 0.002 respectively. Therefore, while no polytope was significant in the log-transformed dataset, at least three polytopes significantly fit the tumor samples in linear space; generally, the lowest number of vertices that reach significance is chosen, which would mean that a 5-vertex polytope best fit the tumor samples in linear space.

While these results do not preclude the possibility that a polytope in linear space best fits the tumor data due to technical variation, as described in the ParTI Manual Caveats, the two analyses together (combined data and tumor samples only) suggest the tumor samples may be linear mixtures of cell types, defined by the original cell-line-based archetype analysis and the combined-data analysis.

We then found the enriched gene sets (GO Biological Processes v7.2) for the combined dataset, as described above for cell lines. We compared the combined dataset archetypes to the original cell line archetypes using a hypergeometric test, as described in “Finding the best fit polytope and evaluating significance with a t-ratio test,” above. We found that each cell-line archetype matches at least one combined-dataset archetype, and each combined-dataset archetype matches at least one cell-line archetype.

3.4.5. Pareto Task Inference

We used the same data from Polley, et al (2016) as in Chapter 2 and considered the activity area (AA). We scored each drug using an adapted version of the method described in Hausser et al. (2019). Briefly, for each archetype *x*, we binned the cell lines into 4 bins by distance to *x*. We then calculated a score as a product over the difference between bins:

$$S = \prod_{i}^{3} AA_i - AA_{i+1} + 2 * SE_i$$

Where i is the bin, AA_i is the median activity area for bin i, and SE is the standard error of the median of bin i. This method gives us a way to rank drugs for each archetype x where cell lines closest to x are most sensitive to the drug, and there is an inverse relationship between sensitivity and distance to x. If the difference between consecutive bins increases by more than twice the SE of the first bin, such that $AA_{i+1} - AA_i > 2 * SE_i$, the product is set to 0. Using the standard error of the mean for each bin in this way allows for small increases in consecutive bins, but if the increase between bins is too large, the score will be set to 0. Positive scores can then be ranked to find drugs for which archetype x is most sensitive. Using the drug-archetype combinations with positive scores, we then ran a one-tailed Mann Whitney test comparing AA of the closest bin to AA of the remaining bins to determine which drug sensitivity was significantly higher for the bin closest to an archetype. An FDR (Bonferroni-Hochberg) correction to these tests showed that no corrected q-values were lower than 0.18. This may be due to the large number of comparisons being made with relatively low numbers of samples (37 total cell lines). Keeping this in mind, we report drugs for each archetype with a p-value < 0.05, which suggest trends in drug sensitivity that should be further confirmed by additional experiments.

3.4.6. Binomial test on treatment of SCLC-A and growth properties of SCLC-N cell lines

To determine if there was a relationship between chemotherapy treatment status of cell lines and the SCLC-A archetype, we ran two binomial tests. First, we mined the ATCC and CCLE databases, and past literature, to determine the treatment status of as many cell lines in our dataset as possible. We found this information for 43 cell lines near the SCLC-A archetype, and 49 non-SCLC-A cell lines. Of the 43 SCLC-A cell lines, 6 had prior therapy and 6 did not. Of the 49 non-A cell lines, 16 had prior therapy and 4 did not. We therefore tested the hypothesis that SCLC-A cells are more likely to be untreated with the logic that, if chemotherapy selectively kills SCLC-A cells, we would be more likely to see SCLC-A cell lines from tumors prior to treatment. We compared the probability of being untreated given an A cell line (6 out of 12) to the expected distribution of untreated cell lines from non-A cell lines (4 out of 20, or p = 0.2). A one-tailed binomial test with an alternative hypothesis that the probability of SCLC-A cells being untreated is *greater* than non-A cells showed that we can reject the null with a p-value = 0.019.

Similarly, we asked whether an untreated cell line is more likely to have a phenotype near the SCLC-A archetype. We compared the probability of being SCLC-A given an untreated cell line (6 out of 10) to the expected distribution of SCLC-A from treated cell lines (6 out of 22). Again, we reject the null of a one-tailed binomial test with a p-value of 0.03, suggesting that untreated cell lines are more likely to be near SCLC-A.

We used a similar analysis to test whether SCLC-N cell lines are more likely to be mixed in culture than non-N cell lines.

3.4.7. Filopodia staining

Glass coverslips were sterilized and then coated with 5 μg/mL Laminin (mouse, Corning Cat# 354232) diluted in PBS overnight at 4 degrees in a 12 well plate. PBS/Laminin solution was aspirated from the coverslips the next day and cells were seeded into the wells at a concentration of 5x10^4 cells per well for H524, H446, and H69, and 1x10^4 cells per well for H196 (due to their larger size). 24 hours post-seeding, cells were fixed with 4% paraformaldehyde, permeabilized with 0.2% saponin, and blocked for 1 hour with 5% BSA + 0.1% saponin, with PBS washes in between. Cells were incubated with anti-Tubulin beta 3 (TUBB3 Clone TUJ1, Cat# 801213, Biolegend) diluted 1:500 in Blocking solution for 1 hour at room temperature. Cells were washed with PBS and then incubated for 1 hour in the dark at room temperature with secondary antibodies diluted 1:1000 in Blocking solution as follows: Hoechst 33342, Rhodamine phalloidin (Cat# R415, Invitrogen), and donkey anti-mouse Alexa Fluor 488 (Cat# A-21202, Invitrogen). Finally, cells were washed with PBS and mounted onto glass slides for imaging. Images were acquired using a Nikon-A1R-HD25 confocal microscope (ran by NIS-Elements) equipped with an Apo TIRF 60x/1.49 NA oil immersion lens. Twenty images of each cell type were acquired per experiment, and each cell analyzed was isolated and not directly touching another cell, to ensure accurate filopodia and protrusion counts per cell. Analysis was done using Fiji software by manually tracing the cell border and using Analyze tab/Measure to quantify cell area and fluorescence intensity. "Filopodia" were counted manually and are defined as the slender protrusions from the cell body that are phalloidin-positive and TUBB3-negative. "Protrusions" were counted manually and are defined as long slender protrusions of the cell body that are TUBB3-positive. Data was graphed in GraphPad Prism as the number of filopodia or protrusions per 500 μm^2 cell area.

3.4.8. Seahorse XF Cell Mito Stress Test

Response to succinate was tested by incubating 25,000-50,000 living cells per well onto Seahorse cell culture plates. Sodium succinate dibasic hexahydrate was diluted into water at 0.25 M (6.75 g/mL), and HCl and KOH were added until the pH was between 7.2 and 7.4. Succinate was then added to cell culture plates at three different concentrations (6mM, 12mM, and 24mM, plus control) and left to incubate overnight for 12 hours. Oxygen consumption rate testing was performed as described in the Seahorse XF Cell Mito Stress test kit User Guide. Cells were then imaged using Hoechst and propidium iodide to count live cells for normalization.

3.4.9. Gene signature matrix generation

After testing each gene with a Mann-Whitney test as described above, genes that are not maximized in the bin closest to an archetype or with a p-value higher than 0.05 are considered insignificant and are removed from the analysis. The remaining genes are assigned to the archetype for which the mean difference (log-ratio) of log-transformed gene expression in the closest bin to the archetype compared to the rest is highest. The matrix (with size [G, n], where G = total number of genes and n = number of archetypes) is populated with the archetype gene expression profiles (i.e. the average location in gene expression space after bootstrapping the archetype analysis). To reduce the size of the gene matrix and choose the most salient genes for each archetype, an algorithm is used to optimize the condition number, or stability, of the matrix. The condition number of the matrix, κ(A), is the value of the asymptotic worst-case relative change in output for a relative change in input:

$$A(x + \delta x) = b + \delta b$$

$$\frac{||\delta x||}{||x||} \leq \kappa(A) \frac{||\delta b||}{||b||}$$

Where A is the signature matrix, b is the input expression matrix, x is the output signature score, and δ is the error. The condition number κ thus gives an upper bound on the output error given a perturbation to the input. Minimizing this value ensures that genes that do not well-distinguish between the archetypes are not included in the matrix. With genes sorted by mean difference for each archetype, the top g genes are chosen for each archetype, with g ranging from 20 to 200. For each g, the condition number of the matrix is calculated using the Python function *cond* from *numpy.linalg* (Oliphant, 2006) using the 2-norm (largest singular value, p = 2). The gene signature matrix size with the lowest condition number, which includes g* genes, is chosen. For our archetypes, g* = 21, so the resulting size of the gene signature matrix is [g* x n, n] = [105, 5]. This method can be extended to other sorted lists of genes, such as genes sorted by adjusted p-value in an ANOVA test between archetypes. For SCLC, the gene signature included the four consensus TFs: ASCL1, NEUROD1, POU2F3, and YAP1.

3.4.10. Single-cell RNA sequencing

Eight SCLC human cell lines from the bulk data above were chosen for single-cell RNA-sequencing. SCLC human cells lines were obtained from ATCC. We chose two cell lines from each NE subtype (A: H69 and CORL279, A2: DMS53 and DMS454; N: H82 and H524) and one cell line from each non-NE subtype (P: H1048; Y: H841). This approximates the distribution of subtypes seen in bulk tumor data, where most tumors are largely NE. We also aimed to pick cell lines that ranged in their distance from their "assigned" archetype, to better understand intermediate samples as compared to ones close to an archetype location.

Cell lines were grown in the preferred media by ATCC in incubators at 37 degrees Celsius and 5% CO2. In preparation for single-cell RNA-sequencing, cells were dissociated with TrypLE, washed with PBS three times, and then the cells were counted, and concentration was adjusted to 100 cells/μL. Droplet-based single-cell encapsulation and barcoding were performed using the inDrop platform (1CellBio), with an in vitro transcription library preparation protocol (Klein et al., 2015). After library preparation, the cells were sequenced using the NextSeq 500 (Illumina). DropEst pipeline was used to process scRNA-seq data and to generate count matrices of each gene in each cell (Petukhov et al., 2018). Specifically, cell barcodes and UMIs were extracted by dropTag, reads were aligned to the human reference transcriptome hg38 using STAR (Dobin et al., 2013) and cell barcode errors were corrected and gene-by-cell count matrices and three other count matrices for exons, introns, and exon/intron spanning reads were measured by dropEst. Spliced and unspliced reads were annotated and RNA expression dynamics of single cells were estimated by velocyto (Manno et al., 2018). SCLC human cell lines have been validated by matching transcript abundance to the bulk RNA-seq data from CCLE.

Patients with SCLC were prospectively identified and consented using an Institutional Review Board (IRB, #030763) approved protocol for collection of tissue plus clinical information and treatment history. All samples were de-identified and protected health information was reviewed according to the Health Insurance Portability and Accountability Act (HIPAA) guidelines. The two human SCLC tumors were collected in collaboration with Vanderbilt University Medical Center. Tumor #1 was a relapsed tumor collected via bronchoscopy with transbronchial needle aspiration of a left hilar mass. The patient had completed carboplatin and etoposide and then prophylactic cranial irradiation. The tissue was washed in an RBC lysis buffer, passed through a 70 μm filter, and washed in PBS. Cells were dissociated with cold DNAse and proteases and titrated every 5-10 minutes to increase dissociation. Library preparation for scRNA-seq was performed according to previous protocols (Banerjee et al., 2020), and cells were sequenced using BGI MGI-seq. Human tumor #2 was a stage 1B SCLC tumor with a mixed large cell NE component treated with etoposide and cisplatin and was surgically removed via right upper lobectomy. The tumor was immediately placed in cold RPMI on ice for dissociation. Library preparation for scRNA-seq was performed as described previously (Banerjee et al., 2020). Cells were prepared for sequencing using TruDrop (Southard-Smith et al., 2020) and sequenced on Nova-seq. As with cell lines, the DropEst pipeline was used to process scRNA-seq data and to generate count matrices of each gene in each cell (Petukhov et al., 2018).

The Rb1/p53/Myc (RPM) mice are available at JAX#029971; RRID: IMSR_JAX:029971 and all experiments with RPM cells were previously performed as in Ireland, et al. TKO mouse lines used were the triple-knockout (TKO) SCLC mouse model bearing deletions of floxed (fl) alleles of p53, Rb, and p130 as previously described (PMID: 20406986). For in vivo SCLC tumor studies with this model, 8 to 12 weeks old mice were used for cancer initiation, and tumors were collected 6-7 months later.

3.4.11. Preprocessing of single-cell RNA-seq data

Single-cell RNA-seq counts matrices were primarily analyzed using the Python packages *Scanpy* and *scVelo* (Bergen et al., 2020; Wolf et al., 2018). First, the scRNA-seq read counts, including both spliced and unspliced counts, for each of the samples in each dataset were generated using the command line interface (*run_dropest)* from *velocyto* on the BAM files generated from DropEst, as described above. This tool generates loom files that can be used with Scanpy and scVelo for preprocessing and velocity calculations. We then used Scanpy to read in the loom files as an AnnData object (anndata.readthedocs.io).

For cell lines, we used a combination of Scanpy and Dropkick v1.2.6 (https://github.com/KenLauLab/dropkick) to label and filter out low-quality cells from each sample. We then concatenate the datasets with a batch key for each cell line. We hierarchically perform the filtering, following the *recipe_dropkick* function from the Dropkick package, We initially filter out cells with < 100 genes using *scanpy.pp.filter_cells* with min_genes = 100. This reduces the total number of cells (in all 8 samples) from 86,492 to 86,349. We remove genes found in < 3 total spliced counts across all samples with *scanpy.pp.filter_genes* with min_counts < 3. This reduces the number of genes from 63,677 to 22,475. These filtering steps ensure we remove any cells or genes with low or no reads, to prepare for further filtering steps below.

We normalize the data using *scanpy.pp.normalize_total*, log-transform with *numpy.log1p* Finally, the log-transformed, normalized counts are then scaled using *scapy.pp.scale*. We then compute a 50-component PCA embedding of the data and use *scvelo.pp.neighbors* and *scvelo.tl.umap* to generate a UMAP dimensionality reduction. After removing so many cells, we re-filter the genes with a low threshold (min_cells = 3) to remove any genes that were only expressed in the low-quality cells. This reduces the number of genes from 22,475 to 20,446.

We remove doublets by following the best practices for Scrublet (Wolock et al., 2019) at github.com/swolock/scrublet and apply the tool on each sample independently. We run Scrublet directly on the raw data with default parameters. Interestingly, in all samples except CORL279, the detected doublet rate was exactly or near 0%, with less than 10 total doublets detected across all 7 samples. In CORL279, however, Scrublet detected 20.8% of the cells as doublets (3148 cells). We analyzed this cell line further by plotting the histograms of the doublet scores for the observed data and the simulated data, which suggested that log-transformation may be more accurate. We thus use the doublet prediction from the log-transformed data, which detected doublets at a rate of 28.6% (4328 cells). Imputation of single-cell data with a tool such as MAGIC has been shown to improve archetype detection (Dijk et al., 2018). We, therefore, use MAGIC to build a model with the default parameters (knn =5, decay = 1, t= 3).

Similar to human cell lines, for human tumors we use Dropkick v1.2.6 to label and filter out low-quality cells from each sample and concatenate the two tumors. We initially filter out cells with < 100 genes using *scanpy.pp.filter_cells* with min_genes = 100, and we remove genes found in < 3 total spliced counts across all samples. This reduces the number of genes from 43,306 to 15,344. We normalize, log transform, and rescale the data. We filter out the low-quality cells that have low Dropkick scores. We use Scrublet to determine the number of possible doublets in the data. One tumor had 6 predicted doublets (out of 7741 original cells before filtering); the other had 3 (out of 580 original cells), which were removed.

To remove immune cells, we filtered Leiden clusters by expression of PTPRC. To remove fibroblasts, we filtered cells using COL1A1 expression, and we used CLDN5 expression to remove endothelial cells. We also used EPCAM to identify epithelial cells. We found several small clusters of immune cells and a single small population of likely fibroblasts. A single cluster had a few cells with low expression of CLDN5, and higher average expression of EPCAM, so we chose not to remove this cluster.

We use MAGIC to impute the dataset for archetype analysis. After cell cycle scoring, one of the single cell archetypes (described below) was enriched for G2M cells. Therefore, we chose to regress out the difference between the G2M and S scores (*cell_cycle_diff*).

TKO tumors were filtered using the same steps as above. TKO1 was predicted to have 11 doublets, which were removed with the filtering above. TKO2 was predicted to have 0 doublets. TKO3 was predicted to have 1,052 doublets, or 27% of the samples, using the default parameters for *scrub_doublets*. We, therefore, investigated this sample further and found that the observed and simulated doublet score histograms were not bimodal, which would be expected if there were true doublets in the data. This may be due to the homogeneity of the sample because Scrublet can only detect neotypic doublets, "which are generated by cells with distinct gene expression (e.g., different cell types) and are expected to introduce more artifacts in downstream analyses" (from scrublet_basics tutorial). Log-transforming the data, as suggested by the demuxlet_example tutorial, gave a doublet percentage of 0.3% (12 cells). Therefore, we removed only these 12 doublets from the dataset.

For the RPM time series, we utilized the data from Ireland et al (2021). While calculating velocity requires realignment to the genome (as described above) and therefore the counts matrices are slightly different from those deposited by Ireland et al. (2020), we chose preprocessing steps and parameters to be as consistent as possible with the original publication of the data, following the above preprocessing steps. Because stringent filtering was already done in Ireland et al. (2020)

to remove low-quality cells and non-cancer populations such as immune and stromal cells, we filtered to the same cells. Ireland et al. (2020) regressed out cell cycle effects to remove variation due to the cell cycle phase, so we repeat this preprocessing step here. We remove variability due to location in the cell cycle by regressing out the cell_cycle_diff attribute. Finally, we use MAGIC to impute the data.

We used the same filtering method as for the samples above for the PDX tumors. Scrublet detected doublets at rates of 0% (1 cell) for MGH1518-1B3 and 0.1% (3 cells) for MGH1518-3A2. Scoring cell cycle genes demonstrated that variance in the data was not highly dependent on the cell cycle phase, so we did not regress out this effect.

3.4.12. Projection of single-cell data on principal components of bulk RNA-seq

We sought to assess whether variation across SCLC samples is aligned to variation within SCLC samples. To do this, we adapt the method described in Hausser et al. (2019) for comparing intra-tumor heterogeneity and inter-tumor diversity in human cell lines. We reduce the dimensionality of the data by focusing on the genes that are profiled in the bulk cell line data and are highly variable in the single-cell data. This reduces the dataset down to 3033 genes and 13,945 cells. We then project the single-cell data into the space defined by the bulk data-fitted PCA, focusing on the top 7 principal components due to an elbow in the explained variance curve for this number of components.

If intra-sample heterogeneity were perfectly aligned with inter-sample diversity, we would expect the single-cell variance explained by the principal components computed on the bulk data to equal the single-cell variance explained by the principal components computed on the single-cell data. In other words, the percentage of variance explained by the single-cell PCA is an upper bound on the variance explained by inter-sample diversity. We, therefore, computed each of these fractions for the single-cell data and compared them. We compare the fraction of variance explained by the bulk PCA to the upper bound of variance explained by the single-cell PCA for the imputed scRNA-seq data on cell lines. When we investigate this ratio for the top 7 components, we find that inter-sample diversity explains 32% of the variance explained by the first 7 single-cell PCs, while a PCA fit to shuffled data explains about 0.28 +/- 0.009% of this variance (50 shuffles). If we consider varying numbers of PCs (1 to 20), the percentage of intra-sample heterogeneity explained by inter-sample diversity stays stable at 36 +/- 0.94%.

3.4.13. Bulk gene signature scoring of single cells using archetype signature matrix

To score the gene signature matrix in single cancer cells, we first subset the single-cell data to the genes in the archetype signatures. Due to dropouts, the intersection of genes from the signatures and the single-cell data may be less than the full signature (105 genes), and we refer to this intersection as "shared genes." We scale the gene signature and the single-cell gene expression data by the L2 norm for each archetype and cell, respectively, to remove differences caused by different platforms (bulk vs. single-cell sequencing). Each archetypal gene expression vector and each cell's gene expression vector were therefore scaled to have a length of 1 so that the archetype space has basis vectors of length 1. We then transformed each cell signature into archetype space using the least-squares approximate solution to $Ax = b$, where A is the signature matrix (shared genes g by subtypes s) and b is the single-cell matrix (shared genes g by cells c). This is solved using *numpy.linalg.lstsq* to generate a "pattern matrix" x (subtypes s by cells c).

3.4.14. Single cell PCHA

We used the R package ParetoTI (Kleshchevnikov, 2019) to run Archetype Analysis on the scRNA-seq from 8 human cell lines, which we found to be more computationally efficient than the original MATLAB package ParTI for the high-dimensional single-cell datasets. To reduce the dimensionality of the data, we fit a PCA and find the number of PCs where the additional explained variance (variance explained on top of n-1 model) is less than 0.1%. We use 11 components, which explain over 85% of the variance in the imputed data. We fit k* = 2 to 8 vertices to the 11-component PCA of the imputed data, with delta = 0. Looking at the variance explained vs. the number of archetypes k, we find that k* = 4 or k* = 6 based on the elbow method. We also considered the mean variance in position of vertices upon bootstrapping (200 iterations with data downsampled to 75%). We find that k* = 4 and k* =6 give variances close to 0, while k* = 5 gives the highest mean variance, suggesting that k* = 5 is not fitting the data geometry well. Therefore, we move forward with k* = 4 and k* = 6 for t-ratio tests. We randomized the data 1000 times as described previously to generate a background distribution of t-ratios. We find four archetypes give significant polytopes using the t-ratio test as described in Methods (t-ratio = 0.3743, p = 0.001). Six archetypes were not significant with a t-ratio of 0.0012 and p = 0.940. Based on these results, it seems k*=4 best fits the cell line data. H82, CORL279, and H1048 are all more central in the polytope, suggesting they may comprise single-cell generalists. These results validate that the human cell lines can be fit by a polytope, suggesting Pareto optimality applies to single cancer cells in these cell lines.

Interestingly, we did not find an SCLC-P archetype. Even using k* = 6 did not result in an archetype by the cell line expressing POU2F3 (H1048), but instead gives two SCLC-A archetypes and two SCLC-A2 archetypes. This suggests that the lack of an SCLC-P archetype is not due to the number of archetypes chosen. A few explanations exist for why there is not a POU2F3 enriched archetype: First, bulk RNA-seq profiles of SCLC-P cell lines may give a spurious archetype in our analysis that is actually a mixture of the other four archetypal transcriptomic profiles in varying proportions. Secondly, due to our small sample size, we may miss an SCLC-P archetype in this single-cell data. Third, SCLC-P cell lines and tumors may represent a valid subtype that does not fit within the Pareto theoretical polytope found here. In other words, the SCLC-P subtype may be a distinct subtype not confined by the Pareto front of the other four SCLC subtypes. If SCLC-P tumors are derived from an alternative cell of origin, we would not expect the functional tasks to trade off with the others found here. This does not explain how some tumors may contain markers for multiple subtypes, including SCLC-P, such as NEUROD1 (SCLC-N, Ireland et al., 2020).

We followed the same preprocessing steps as described for human cell lines above to run PCHA on human tumors. After imputation, we fit a PCA to the data and found that 8 PCs explain over 85% of the variance. The knee of the explained variance vs. PC plot is 7, suggesting a low dimensional representation of the data is possible. We also consider the number of PCs where the additional explained variance is less than 0.1%; this gives 11 components. We fit k* = 2 to 8 vertices to the 11-component PCA of the imputed data with delta = 0. We find that three or more archetypes explain over 80% of the variance in the reduced dimensional data (2 archetypes explain less than 70% of the variance). When considering the mean variance in position of vertices, we find that K* = 2-4 archetypes show little variance (less than .05), suggesting the archetype locations are robust to bootstrapping. K* = 5 gives the highest mean variance in position, at 0.8. Furthermore, the t-ratio of the polytope dips significantly for k*= 5 (where a higher t-ratio close to 1 is a better fit). Therefore, k* = 3 or 4 seems most likely to fit the data. We move forward with these k* for t-ratio tests. As expected, we find k*= 3 is the smallest number of archetypes that significantly fits the data, with a t-ratio of 0.596 (closer to 1 is better) and p = 0.008. K* = 4 had a much lower t-ratio of 0.396 (p = 0.001) and k* = 5 was insignificant, with a p-value of 0.753. Based on these results, k* = 3 best fits the data.

We wanted to ensure that mouse tumors could also be well described by Pareto theory and applied archetype analysis to TKO mouse tumors. In this dataset of three tumors, nine principal components explain over 85% of the variance in the imputed data. Furthermore, 16 PCs are needed for the variance explained by additional components to be less than 0.1%. We, therefore, reduced the dataset dimensionality to 16 dimensions for archetype analysis. We fit k* = 2 to 8 vertices to this PCA with delta = 0 and found that three archetypes can explain over 65% of the variance in the dataset. Using a bootstrapping method, we found that the mean variance in position of the vertices significantly increases after four archetypes; 4 archetypes have a mean variance of less than 0.025, while 5 gives over 0.25, and 6 gives over 0.175. Lastly, the t-ratio drops dramatically after four archetypes, from over 0.2 to under 0.05. These results suggest that the number of archetypes that best fit the data is k* = 3 or 4. We then performed a t-ratio test with the same parameters as above to determine which was the better fit. Three archetypes were insignificant, with a t-ratio of 0.33 and p = 0.999. Contrarily, four archetypes gave a significant polytope with p = 0.001 and t-ratio = 0.21. Therefore, four archetypes best fit the data.

We next analyzed the time course of RPM tumor cells. We found that the top 9 PCs explain over 90% of the variance in the data, and the top 22 PCs are needed for the variance explained by an additional PC to be less than 0.1%. We fit K* = 2-8 vertices with delta = 0 to 22 PCs and found a knee in the EV vs. number of archetypes plot at k* = 3. Using a bootstrapping method, we found that the mean variance in position of the vertices increases for k* = 4 and k* =5, and decreases for k* = 6, suggesting k* = 3 or 6 are the most robust number of archetypes. We therefore ran a t-ratio test on k* = 3-6 to determine the best number of archetypes. We found that k* = 6 was the only polytope with a significant t-ratio (p-value = 0.001) and had a larger t-ratio than fewer archetypes (k*=5). Therefore, k* = 6 archetypes best fit the data.

Lastly, we test whether the two PDX tumors (MGH1518-1B3 and MGH1518-3A2) fit into a polytope predicted by Pareto theory. We find that eight components represent the imputed data, explaining 72% of the variance in the imputed data. We fit k* = 2 to 8 vertices to the 8-component PCA of the imputed data, with delta = 0 and find that at least three archetypes explain over 85% of the data, while two explain less than 55%. We also considered the mean variance in position of vertices upon bootstrapping (200 iterations with data down-sampled to 75%). We find that k* = 2, 3, and 5 give variances close to 0 (less than 0.1), while k* = 4 gives the highest mean variance at over 0.6. Therefore, we move forward with k* = 3-6 for t-ratio tests. We find five archetypes is the smallest k* that give significant polytopes using the t-ratio test (t-ratio = 0.11, p = 0.001). Three and four archetypes were insignificant (p = 0.35 and p=.39, respectively). Similarly, a six-archetype polytope was not significant (p = 0.105). Based on these results, it seems k*=5 best fits the PDX tumor data.

3.4.15. Alignment of bulk and single-cell archetypes

PCHA constrains the archetype vertices to be a weighted average of the data points and approximates the data points by a weighted average of the archetypes. Therefore, each cell has archetypal weights given by a matrix S, such that the weights for each cell sum to 1. We can use these weights to directly "score" the single cells and label them by archetype. Each cell is given weights summing to 1, and we consider the cells with a score above 0.95 for a single archetype to be a specialist. In order to align the single-cell archetypes with our predefined archetype space, we consider the scores for each cell described in the Method section "Bulk gene signature scoring of single cells using archetype signature matrix" above. For each bulk signature *x* and for each single-cell archetype *a*, we ran the following significance test:

1. Find the mean bulk score *x* for *a* specialists, *m*.

2. Choose a random sample of size n_a, where n_a is the number of a specialists, with replacement from the remaining cells (i.e. cells that are not a specialists, including generalists and other specialist cells). Find the mean bulk score for this sample. N.B. Because some timepoints have very few cells, we sample evenly from each timepoint to ensure adequate representation across the timepoints.
3. Repeat this random selection 1000 times.
4. Generate a p-value, which is equal to the percentage of means from this random distribution above *m*.
5. Using *statsmodels.states.multitest,* correct p-values for multiple tests. We used the Bonferroni-Holm method to control the family-wise error rate. Consider $q < 0.1$ significant.

3.4.16. Visualization by Circular A Posteriori (CAP) Projection Plots

To display archetype scores or probabilities $\mathbf{p_i}$ of archetype labels for each cell, we used a method based on circular a posteriori (CAP) projection adapted from Jaitin et al. (2014) and Velten et al. (2017). For the five-dimensional vectors **p** of archetype scores (shape of **p** is the number of cells **C** X number of bulk archetypes **N**), we first arrange the archetypes on the edge of a circle such that each archetype **k** is assigned an angle $\mathbf{a_k}$. The class probabilities $\mathbf{p_{ik}}$ for cell **i** are transformed to Cartesian coordinates by

$$x_i = \sum_k p_{ik} \cos a_k$$

and

$$y_i = \sum_k p_{ik} \sin a_k$$

Because the archetypes could be arranged in several different orders around the circle, we wish to find the best arrangement such that the most similar archetypes are placed next to each other. In practice, this is done by calculating the proximity between archetypes, given for archetypes **l** and **k** by

$$D_{lk} = \sum_{i} p_{il} \times p_{ik}$$

We calculate the proximity for each arrangement of archetypes as the sum of the proximity for each pair of neighboring archetypes; for example, the arrangement of archetypes A → B → C → D → E gives the proximity

$$D_{ABCDE} = D_{AB} + D_{BC} + D_{CD} + D_{DE} + D_{EA}$$

We test all possible arrangements (**N!** for **N** archetypes) and choose the arrangement with the highest proximity.

3.4.17. RNA Velocity using scVelo and CellRank

We interrogated the dynamics of SCLC cells and tumors by analyzing RNA velocity with the Python packages *scVelo* and *CellRank* (Bergen et al., 2020; Lange et al., 2022). RNA velocity uses a splicing model to predict directionality and magnitude of gene expression change in the near future for each cell sampled. Using the data without MAGIC imputation (because there are no standardized ways to incorporate imputation and RNA velocity in the field), we used scVelo packages to fit a neighborhood graph (adjacency matrix) and first-order moments with scvelo.pp.neighbors and *scvelo.pp.moments*, respectively. We then used scVelo's dynamical modeling pipeline as described in https://scvelo.readthedocs.io/DynamicalModeling/, with the velocity calculation grouped by timepoint. We then computed velocity graphs, confidences (coherence of velocities), and velocity lengths, which indicate how coherent and significant the velocity vectors are across gene expression space. This gives an idea of how much movement is in the dataset.

The dynamical model also reports fitting parameters and fit likelihoods for each gene. We used the fit likelihood to rank-order the velocity genes, and visualizing investigated the top genes for each dataset to determine if the fit could be used to make predictions about transitions. As described in the tutorial, the plots of unspliced versus spliced counts for each gene should have a characteristic “almond” shape. To determine possible regulators of the highly fit genes that are driving transitions, we used EnrichR on the genes with a fit likelihood > 0.3 and report the significant transcription factors from the “ENCODE and ChEA Consensus TF” list (Chen et al., 2013; Kuleshov et al., 2016).

For the RPM dataset, the data samples span 17 days, and therefore the dynamics of the data cover a longer timescale than that of splicing dynamics, which typically occurs on the timescale of a few hours (Manno et al., 2018). To overcome this challenge, we use CellRank. CellRank is capable of incorporating velocity information fit to each timepoint and alternative measures of temporal dynamics such as pseudotime. We therefore use a previously calculated diffusion pseudotime (Ireland et al., 2020) which adds information about the longer timescale dynamics across days. We adapted the CellRank tutorial on “Kernels and estimators" to combine these two sources of dynamical information into a combined kernel. We used the same weights as in the tutorial—0.8 for the velocity kernel and 0.2 for the DPT kernel—though our results were robust to these parameters from [.1 –.9] for each combination of velocity and DPT. The combined kernel is used to compute a cell-cell transition matrix as a representation of the Markov chain underlying the dynamics. We used the Generalized Perron Cluster Cluster Analysis (GPCCA) estimator, which computes aggregate dynamics based on the Markov chain transition matrix by projecting the Markov chain onto a small set of macrostates. We computed a Schur decomposition with 20 components and default parameters. Finally, we computed the macrostates and terminal states on the phenotype clusters (specialists and generalists).

We use the “gmres” solver to compute absorption probabilities using the solvers from petsc4py. We computed driver genes for the X and SCLC-Y lineages and used EnrichR to investigate the regulators of the top 40 drivers for each lineage. Because the significance of the driver genes is quantified by a q-value, we used this to determine whether genes in the SCLC-Y bulk archetype signature were drivers of the SCLC-Y lineage, including only gene with a positive correlation to the lineage.

3.4.18. Cell Transport Potential Calculation and Analysis

In quantifying plasticity, we wanted to capture the local likelihood of phenotypic transition (that is, change in gene expression profile) for each transcriptional state sampled. Furthermore, we would like to consider size and variance of the phenotypic change. Colloquially, “plastic” cells, such as stem cells, generally are considered plastic because they have at least these two characteristics: they are poised to change their gene expression profile by a large amount (differentiation potency), and they are able to change into multiple different end states (multipotency). Weinreb et al. showed that single cell transitions could be quantified via the velocity field of a phenotypic landscape, which is the gradient of a potential function. This potential can be decomposed into two terms: a “transport” term and a “constraint” term. The deterministic transport term counteracts sources and sinks in the landscape to keep the cell density in dynamic equilibrium, assuming the population is at steady state. As a proxy for this potential term, we calculate the Cell Transport Potential (CTrP). CTrP is the expected value for the movement of each individual cell. More formally, it is the expected distance of travel for a cell, weighted by the time spent in each other cell state before absorption (reaching an end state). The method is detailed below.

CTrP is a measure of the average distance a cell may travel according to its RNA velocity. For each independent sample (untreated or treated), we ran the following pipeline:

1. **Using RNA velocity calculated as described above, and for each category ('treatment'), compute a Markov Chain Model transition matrix**. This is calculated using an adapted version of ScVelo's transition_matrix function, in which transition probabilities between each two cells, i and j, is calculated from the velocity graph pairwise. Each entry is a probability describing the likelihood of moving from state i to state j, and each row is the probability distribution of transitions from state i. RNA velocity is compared to distances between other cells to get a pairwise cosine correlation matrix (velocity graph). A scale parameter (default 10) is used to scale a Gaussian kernel applied to the velocity graph, restricted to transitions in the PCA embedding. This transition matrix, P, has dimensions nxn, where n = number of cells. It is then normalized to ensure each row adds to 1 (because each row is the probability of cell i transitioning to any other cell j, which should total 1). Diffusion for P is scaled to 0 (i.e., ignored). Alternatively (for RPM time course), we used CellRank to compute a transition matrix as described above in "RNA Velocity Calculation and Analysis."

2. **Calculate absorbing states (end states) using eigenvectors**. Eigenvalues are calculated for the transition matrix. Any eigenvalue $l = 1$ (here, with a tolerance of 0.01), is associated with an end state distribution (eigenvector **v**); i.e., $P(v) = l v$ implies that a distribution of states v will not change under further transformation (transitions) from P. If the Markov Chain is an absorbing Markov Chain, it will contain both transient states (t = number of transient states, where $P(i,i) < 1$), and absorbing states (r = number of absorbing states, where $P(i,i) = 1$). For every absorbing state in

the matrix, there will be an associated eigenvalue/vector pair, with l =1, because any initial configuration of states will continue to evolve until every cell has reached an absorbing state. Therefore, the multiplicity of $l = 1$ is equal to the number of end states (absorbing states, or irreducible cycles). The associated eigenvectors **v** thus correspond to the absorbing states in the Markov Chain, within the tolerance of 0.01.

3. **Calculate the fundamental matrix.** In an absorbing Markov Chain, it is possible for every cell to reach an absorbing state in a finite number of steps. Let us rewrite P, the transition matrix, that has *t* transient states and *r* absorbing states, as:

$$P = \begin{bmatrix} Q \; R \\ 0 \; I_r \end{bmatrix}$$

where Q is a t x t matrix, R is a non-zero t x r matrix, 0 is an r x t zero matrix, and I_r is an r x r identity matrix. Thus, Q describes the probability of transitioning between transient states, and R describes the probability of transitioning from a transient state to an absorbing state. The fundamental matrix N of P describes the expected number of visits to a transient state j from a transient state i before being absorbed. Because the Markov Chain is absorbing, this number is the sum for all k of Q^k for k in {0,1,2,…}:

$$N = \sum_{k=0}^{inf} Q^k = (I_t - Q)^{-1}$$

Because Q^k eventually goes to the zero matrix (all cells are absorbed), this sum converges for all absorbing chains. Furthermore, each row of the fundamental matrix describes the expected amount of time (i.e. number of steps in the Markov random walk) spent in state j starting from state i, and thus the row can be thought of as a distribution of weights associated with each state j for each starting state i. N is calculated as written above: the inverse of Q subtracted from the identity matrix. In practice, Numpy's function numpy.linalg.inv(I-Q) is used to calculate N.

4. **Calculate a distance matrix.** A distance matrix *D (n x n)* is then calculated using scipy's function *scipy.spatial.distance.cdist* (Virtanen et al., 2020). Here, we calculate the Euclidean distance on the PCA embedding of each sample. Distance may also be calculated directly on the high dimensional data; alternatively, it may be calculated on nonlinear dimensionality reduction techniques, such as UMAP and tSNE, but these distances tend to break down for samples that are highly discontinuous (discrete clusters) and should only be applied to continuous data that falls on a single manifold.

5. **Calculate Cell Transport Potential**. Finally, CTrP is calculated as the inner product of each row in fundamental matrix N, and each row in distance matrix D. This gives an expected distance (sum of distances to j from i, weighted by time or number of steps spent in j before absorption).

The advantage of this metric over similar techniques, such as pseudotime and other trajectory inference metrics, is that CTrP is an expected distance in linear (PCA) space, which can be compared across samples (assuming they have been embedded in the same PCA).

3.4.19. Whole Genome Sequencing (WGS)

As described previously in Ireland et al. (2020), "30X WGS data was collected from Day 4 and Day 23 samples, as well as from a blood sample from RPM mice as the normal control. Genomic DNA was extracted from flash frozen cell pellets of Day 4 and 23 cells along with whole blood from the same RPM mouse using Qiagen's DNeasy Blood and Tissue kit (Qiagen cat#69504). Libraries were prepared using the Nextera DNA Flex Library Prep Kit (Illumina cat#20018705). Libraries were sequenced on a NovaSeq 6000 instrument targeting 300 million read-pairs on a 2 x 150 bp run (30x coverage of whole genome). Sequencing reads were aligned to mouse genome mm10 by BWA 0.7.17-r1188 (Li and Durbin, 2009). Rb1and Trp53 deletions were examined in the Integrated Genome Viewer (IGV) software v2.5.0. SNVs were jointly called by Freebayes 1.2.0." Somatic SNVs were filtered by the following criteria: DP >15 and AO =0 in the normal sample. Somatic SNVs were further filter by AO>15 and AO<110 in day 4 and day 23 samples. Variants were annotated by SnpEff 4.3 (Cingolani et al., 2012).

Chapter 4.

Elucidating the role of phenotypic plasticity and gene regulatory network dynamics in therapy resistance of variant SCLC

4.1. Introduction

In the previous two chapters, we developed a comprehensive framework for understanding phenotypic heterogeneity and plasticity in Small Cell Lung Cancer. This work showed that SCLC tumors in humans and mouse models are often mixes of subtypes defined by gene regulatory network dynamics. Perturbations to these dynamics are predicted to destabilize attractors associated with each subtype, and therefore may be useful in designing treatment in the face of acquired resistance, which is ubiquitous in SCLC patients. Furthermore, a continuous, archetype-based framework allowed us to understand how cells may transition from one subtype to another to optimize trade-offs between various survival tasks, such as proliferation or injury repair.

In this chapter, we apply the methods developed in the previous two chapters to elucidate the role of network dynamics in a variant mouse model of SCLC and the role of plasticity in evading targeted therapy. First, we use a novel approach to subtype SCLC patients and find a variant SCLC subtype enriched in inflammatory signatures and immune infiltration (SCLC-I). This class of SCLC is particularly susceptible to immune checkpoint blockade therapy, suggesting relevance of our subtyping approach that may be beneficial to determine treatment options for patients. While this non-NE subtype of patients seems distinct from the SCLC-Y, single cells within this class are also enriched in SCLC-Y signatures. Furthermore, we see an enri8chment of SCLC-I cells after cisplatin resistance that seem capable of regenerating the rest of the tumor, suggesting relapse after therapy may be due to phenotypic plasticity. We therefore applied the plasticity pipeline developed in Chapter 3 and found that these cells are indeed more plastic after therapy. Therefore, targeting this plasticity may be a good strategy for treatment of such tumors.

Finally, we investigate a distinct variant subtype of SCLC that arises in MYC-driven tumors after ASCL1 loss. Using the RPM model discussed in Chapters 2 and 3, we knocked out ASCL1 to form an RPMA model with distinct and striking morphological and transcriptomic features from RPM tumors. Specifically, RPMA tumors become ossified and turn on gene programs related to bone ossification, suggesting that ASCL1 represses a latent osteogenic program. Compared to the archetype space that defines classic SCLC tumors, this can be thought of as an additional archetype that arises specifically under the genetic modification of ASCL1 loss. While these tumors seem to progress through similar stages as RPM tumors (i.e., they start in a more NE state similar to RPM tumors, as shown by single cell data in Olsen et al., 2021), they end up "escaping" the classic archetype space to reach a mesenchymal stem cell like state that seems capable of chondrogenesis and osteogenesis. We use gene regulatory network analysis, similar to Chapter 2, to analyze the mechanism by which ASCL1 loss may result in activation of an osteogenic program. We then use *in silico* simulations of the network to determine master regulators and destabilizers of the RPM and RPMA states. Overall, these sections demonstrate that phenotypic transitions are critical to understanding both relapse to therapy and the acquisition of a variant, bone-developmental gene program that may begin to explain the propensity of SCLC tumors to metastasize preferentially to bone.

4.2. Quantifying plasticity of an inflammatory subtype in resistant SCLC tumors[3]

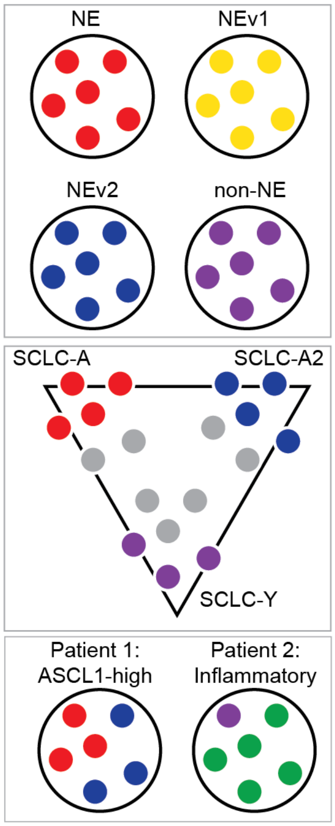
4.2.1. Motivation

As mentioned previously, SCLC heterogeneity has been characterized by expression of a few driving transcription factors. In Chapters 2 and 3, we saw that transcription factors (ASCL1, NEUROD1, POU2F3, and YAP1) are not sufficient to fully capture the heterogeneity within SCLC cell lines and tumors. Instead, gene signatures or programs (modules) and transcription factor (TF) networks may be used to identify key phenotypes that define cell identity in SCLC. In those chapters, we utilized cell line expression data to define the phenotypic space of SCLC, as these are generally considered to be less heterogeneous that tumor populations. This approach was used to understand the phenotypes of *single cells*; in other words, an SCLC cell could acquire any of these phenotypes and potentially be forced to transition between them, which was modeled using transcription factor network simulations.

Rather than subtyping based on cell lines, which may include ASCL1+, NEUROD1+, POU2F3+, and YAP1+ populations, we can also subtype human tumors directly. There is some evidence that the YAP1+ phenotype is not prevalent in tumors (Baine et al., 2020). Furthermore, it is currently unclear if subtype classifications predict responses to chemo-, targeted-, and immune-based therapies. To address both of these gaps in knowledge, we took an alternative approach to those in previous chapters to see if the subtypes of human tumors match the individual subtypes we have found previously. This is helpful for defining classes of *patients*, rather than that of *individual cells*, that may correspond to or predict response to therapy. For example, we found

[3] This section is adapted from "Patterns of transcription factor programs and immune pathway activation define four major subtypes of SCLC with distinct therapeutic vulnerabilities" published in Cancer Cell and has been reproduced with the permission of the publisher and my co-authors. Gay, C. M., Stewart, C.A., Park, E.M., Diao, L., **Groves, S.M.** *et al.* Patterns of transcription factor programs and immune pathway activation define four major subtypes of SCLC with distinct therapeutic vulnerabilities. *Cancer Cell* 39, 1-15 (2021) doi:10.1016/j.ccell.2020.12.014.

an inflamed subtype of SCLC tumors which we term SCLC-I, and gene expression profiles of tumors in this class have greater immune cell infiltration, as determined by CIBERSORTx (Newman et al., 2019). Our approaches are summarized in **Figure 4.1**.



Chapter 2 Approach:
Discrete clustering on cell lines
at bulk and single cell level
to identify stable attractor
states in phenotypic space

Chapter 3 Approach:
Refine subtype identification
using archetype analysis
on cell lines and tumors
at bulk and single cell level

Chapter 4 Approach:
Classify SCLC patients
using Non-negative Matrix
Factorization (NMF) to identify
clinically-relevant classes

Figure 4.1: Evolution of approaches to characterizing SCLC heterogeneity. *In Chapter 2, we use a discrete clustering approach to find attractor states of a phenotypic landscape. In Chapter 3, we supplement this with archetype analysis, which is capable of characterizing intermediate, generalist cells and provides an evolutionary theoretical underpinning for the existing SCLC phenotypes. In this Chapter, we use patient data to find clinically relevant classes of SCLC patient, including an inflamed subtype, SCLC-I. It is unclear if tumors with an inflamed subtype comprise cells of the SCLC-Y subtype, a novel subtype, or a mix of both.*

Lastly, it is now evident that intratumoral heterogeneity (ITH) and plasticity may impact the natural history of a tumor. We and others have shown that multiple transcriptional subtypes may exist within a single tumor and may switch between phenotypes during tumor progression or under treatment pressure (Ireland et al., 2020; Simpson et al., 2020; Stewart et al., 2020). Here, we investigate whether transcriptional subtyping of SCLC intertumoral heterogeneity can identify molecular and immune subtypes with discrete therapeutic vulnerabilities. Furthermore, we consider whether subtype-specific ITH in response to treatment may underlie acquired therapeutic resistance.

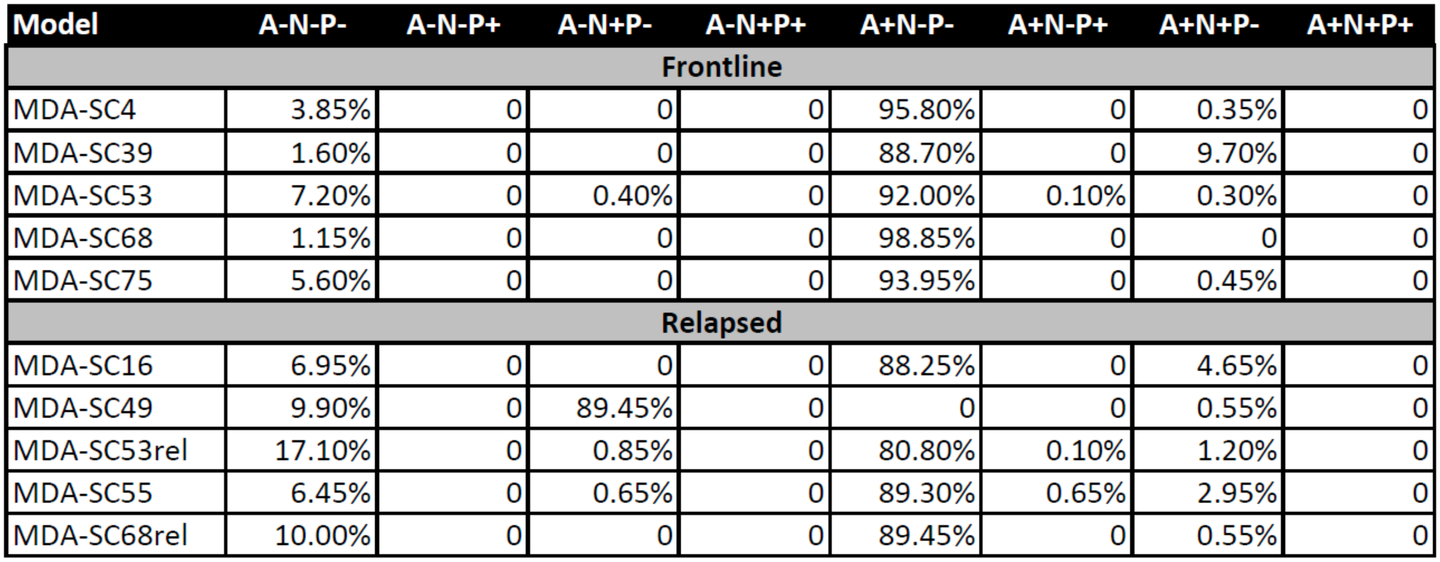
4.2.2. SCLC-I is a novel inflamed subtype of SCLC tumors

Non-negative matrix factorization (NMF) (Skoulidis et al., 2015) was applied to previously published RNAseq data from 81 surgically resected, mostly limited-stage SCLC (LS-SCLC) tumors (George et al., 2015). This method showed that three or four-cluster options best fit the tumor data. Both three- and four-cluster options included an ASCL1-high and a NEUROD1-high group; the four-cluster scheme separated a POU2F-high cluster from a cluster with lower expression of all three TFs. Interestingly, this fourth subtype had several uniquely expressed, immune-related genes such as immune checkpoints and human leukocyte antigens (HLAs). The subtype was therefore designated as inflamed, or SCLC-I. As expected, there was a clear distinction between NE subtypes (SCLC-A and SCLC-N) and non-NE subtypes (SCLC-P and SCLC-I) when considering canonical NE genes, such as *Chromogranin A* (*CHGA*) and *Synaptophysin* (*SYP*). Furthermore, the expression of *YAP1* and its transcriptional targets were higher in both SCLC-P and SCLC-I compared to the other two subtypes. This is consistent with recent immunohistochemistry (IHC) characterization of SCLC tumors, which did not find a subtype of tumors exclusively defined by *YAP1* expression (Baine et al., 2020).

SCLC-I tumors had significantly higher expression of both CD8A and CD8B, suggesting greater cytotoxic T cell infiltration. CIBERSORTx deconvolution (Newman et al., 2019) confirmed that SCLC-I tumors have the highest total immune infiltration. Several immune cell populations were markedly increased in SCLC-I, including T-cells, NK cells, and macrophages. Finally, SCLC-I tumors had consistently higher expression of an interferon-γ-related T cell gene expression profile (GEP) (Ayers et al., 2017), which predicts response to ICB in solid tumors independent of tumor mutational burden (TMB) (Ott et al., 2019), and of countless immune checkpoint molecules, including CD274 (encoding Programmed Death Ligand 1, PD-L1), as well as PDCD1, (encoding PD-1). This may be significant due to alternative proposed mechanisms of ICB resistance in SCLC, which include low expression of HLAs, interferon signatures, and immune checkpoints (Hamilton and Rath, 2019).

SCLC-I tumors made up approximately 17% of the tumor samples in the George et al. (2015) dataset. This proportion was corroborated in independent datasets, such as RNA-seq profiles from treatment-naïve patients enrolled in the Phase 3 Impower133 trial (n=276) and a published RNA microarray dataset from 23 SCLC tumor samples (Sato et al., 2013). Subtypes in these two datasets were determined using the same NMF-derived gene signature (n = 1300) applied to the George et al. (2015) dataset, and four subtypes were observed in each validation cohort. While the Impower133 trial was not statistically powered for subtype-specific subgroup analyses, overall survival (OS) hazard ratios (HRs) for standard of care (etoposide and a platinum-based agent, EP) plus atezolizumab support a modest trend toward improved overall survival in

Table 4.1: Single-cell expression of ASCL1/NEUROD1/POU2F3 in patient-derived SCLC xenografts.



Model	A-N-P-	A-N-P+	A-N+P-	A-N+P+	A+N-P-	A+N-P+	A+N+P-	A+N+P+
Frontline								
MDA-SC4	3.85%	0	0	0	95.80%	0	0.35%	0
MDA-SC39	1.60%	0	0	0	88.70%	0	9.70%	0
MDA-SC53	7.20%	0	0.40%	0	92.00%	0.10%	0.30%	0
MDA-SC68	1.15%	0	0	0	98.85%	0	0	0
MDA-SC75	5.60%	0	0	0	93.95%	0	0.45%	0
Relapsed								
MDA-SC16	6.95%	0	0	0	88.25%	0	4.65%	0
MDA-SC49	9.90%	0	89.45%	0	0	0	0.55%	0
MDA-SC53rel	17.10%	0	0.85%	0	80.80%	0.10%	1.20%	0
MDA-SC55	6.45%	0	0.65%	0	89.30%	0.65%	2.95%	0
MDA-SC68rel	10.00%	0	0	0	89.45%	0	0.55%	0

SCLC-I compared to the other three subtypes that is not observed in the EP + placebo arm. Finally, RNA-seq from 62 SCLC cell lines (Stewart et al., 2021) contained cell line models of all four subtypes, confirming that subtype can be defined in the absence of tumor microenvironment.

4.2.3. Emergence of SCLC-I accompanies platinum resistance

The prior analyses focused on subtype heterogeneity among SCLC tumors. To investigate intratumoral heterogeneity, a series of CDX models from SCLC patients (Stewart et al., 2020) was analyzed. Based on single-cell expression of ASCL1, NEUROD1, and POU2F3, both SCLC-A and SCLC-N dominate the models within our xenograft library.

Single cell RNA-seq also permits exploration of co-expression of subtype-defining transcription factors. Each cell can be classified into one of seven categories on the basis of the binary presence or absence of ASCL1, NEUROD1, and POU2F3 expression (and co-expression) (**Table 4.1**). While most cells express only one of the transcription factors, the expression is not entirely mutually exclusive. As discussed in Chapter 3, this co-expression of key transcription factors is to be expected, as SCLC cells may have gene expression profiles intermediate between archetypal, transcription-factor defined extremes.

In our scRNA-seq data, there was an increase in triple-negative, SCLC-I cells in platinum-relapsed models (**Table 4.1**). This suggested that intratumoral shifts toward increasing SCLC-I may underlie platinum resistance. We selected two platinum sensitive, ASCL1-predominant CDX models developed from treatment-naïve patients (MDA-SC53 and MDA-SC68) to further analyze this shift in phenotype. As described in Stewart et al. (2020), these models were treated with cisplatin to maximal response and then throughout relapse (cis-relapsed) and collected for scRNA-seq along with a matched vehicle treated tumor of same size (treatment-naïve). In both models, there was a reduction in ASCL1-positive proportion of cells and an emergence of a distinct “island” cluster that contained a majority of the ASCL1-negative cells that emerge post-relapse. These ASCL1-negative cells do not gain expression of NEUROD1, POU2F3, or even YAP1, but

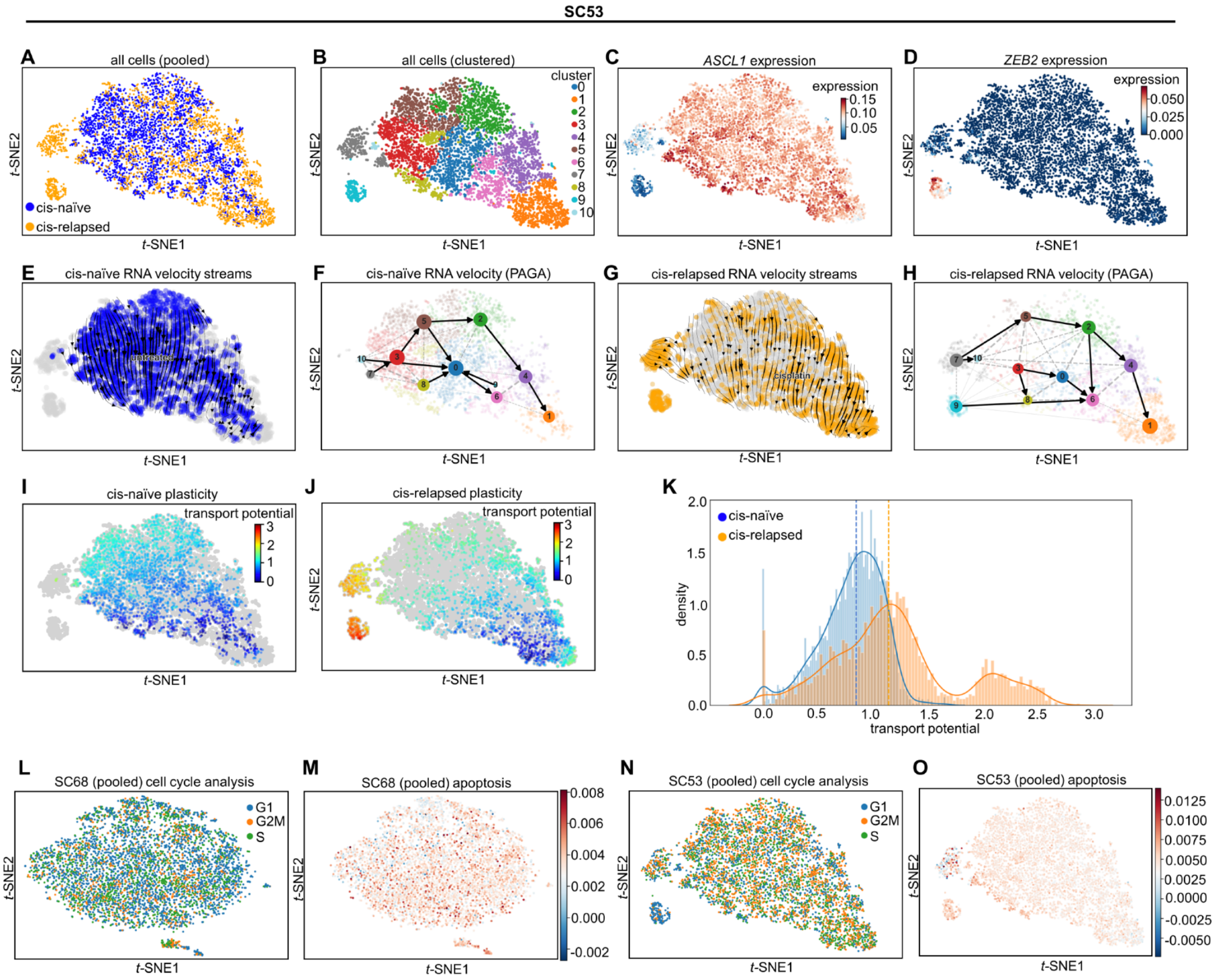


Figure 4.2: Plasticity of emerging SCLC-I populations in cisplatin-resistant SC53 model. A and B. *tSNE projection of all cells from MDA-SC53 CDXs with treatment history (A) or Leiden clustering assignment (B)denoted.* ***C.*** *Expression of ASCL1.* ***D.*** *Expression of ZEB2.* ***E-H.*** *RNA velocity vector streams and PAGA maps for cells from cisplatin-naïve (E-F) and cisplatin-relapsed (G-H) CDX tumors.* ***I and J.*** *Cell plasticity, as measured by cell transport potential, for cells from cisplatin-naïve (I) and cisplatin-relapsed (J) tumors highlighting areas of greatest plasticity appear within island cluster within relapsed tumor.* ***K.*** *Comparison of transport potential between cisplatin-naïve and –relapsed cells demonstrating higher overall plasticity in cisplatin-relapsed cells.* ***L-O.*** *Expression of cell cycle-specific and apoptosis-specific gene lists in pooled cells from MDA-SC68 (L-M) and MDA-SC53 (N-O). Sample sizes: n=4000 cells total (pooled) (A-O).*

instead are largely triple negative (SCLC-I).

4.2.4. SCLC-I populations support tumor-wide resistance via transcriptional plasticity

It is not clear whether the small populations of SCLC-I cells that emerge following cisplatin-relapse are sufficient to drive the observed platinum resistance. There may be two explanations for how the observed subtype switching could drive global resistance. Firstly, the limitations of the binary assessment of single-cell subtype used here may underestimate the level of subtype evolution. As discussed in previous chapters, subtyping by a few transcription factors can lead to insights regarding broad changes in subtype but may be insufficient to fully capture the evolution of phenotypic heterogeneity in SCLC. In both models (SC53 and SC68), we found that ASCL1 decreased in expression and the EMT score increased across the entire tumor following

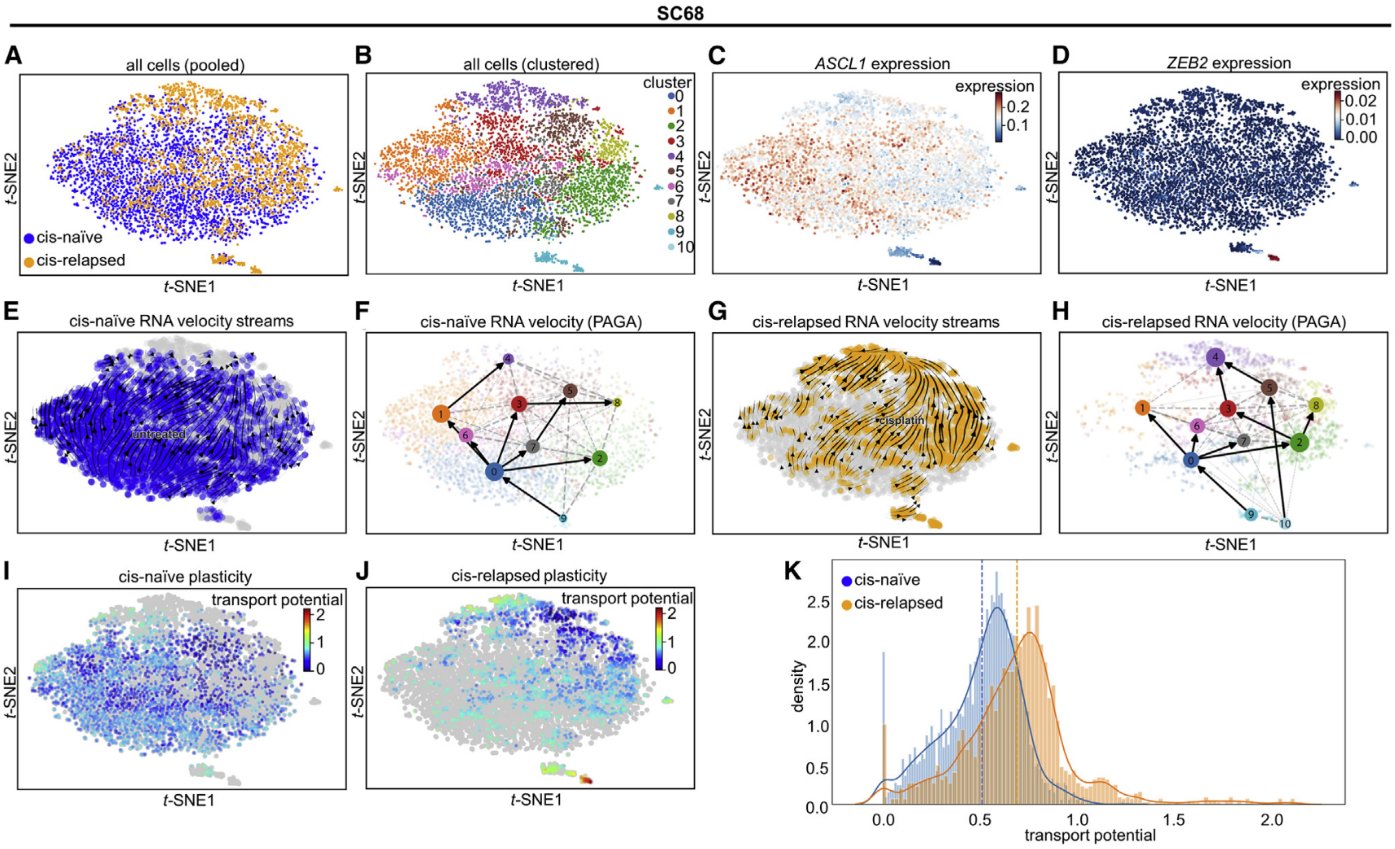


Figure 4.3: Plasticity of emerging SCLC-I populations in cisplatin-resistant SC68 model. A–D. *t-SNE projection of all cells from MDA-SC68 CDXs with treatment history (A) or Leiden clustering assignment (B) denoted.* ***C and D.*** *Expression of ASCL1 (C) and ZEB2 (D) in these cells.* ***E–H***. *Note the upper right, composed largely of cisplatin-relapsed cells, demonstrates lower ASCL1 expression, while the island clusters are essentially ASCL1null. RNA velocity vector streams and PAGA maps for cells from cisplatin-naïve (E-F) and cisplatin-relapsed (G-H) CDX tumors.* ***I and J***. *Cell plasticity, as measured by cell transport potential, for cells from cisplatin-naïve (I) and cisplatin-relapsed (J) tumors highlighting areas of greatest plasticity in island cluster within relapsed tumor.* ***K.*** *Comparison of transport potential between cisplatin-naïve and -relapsed cells demonstrating higher overall plasticity in cisplatin-relapsed cells. Sample sizes: n = 2000 cells per arm.*

platinum resistance, suggesting that these changes are not solely due to the small SCLC-I subpopulation. Together, these data suggest that even outside of the SCLC-I cluster containing now fully triple-negative cells, there is ongoing evolution toward lower ASCL1 expression and increasing features of SCLC-I (e.g., EMT) that may account for decreasing platinum sensitivity.

Secondly, the SCLC-I cells that emerge following platinum resistance may serve as a highly resistant and plastic population with the potential to replenish the tumor even as the remaining platinum-sensitive cells undergo cell death. To address this question, we explored quantitative measures of single-cell plasticity using RNA velocity. First, we applied t-SNE and a clustering algorithm to pooled cells from both cisplatin-naïve and cisplatin-relapsed tumors (**Figures 4.5A-B and 4.6A-B**). In each case, we again identified non-NE islands, similar to **Figure 4.4**, here almost exclusively composed of relapsed cells and showing an absence of ASCL1 and presence of SCLC-I features (e.g., ZEB2, a transcriptional mediator of EMT) (**Figures 4.5C and D; 4.6C and D**). Outside of the island clusters in both PDX models, sensitive and resistant cells often cluster together, but a significant portion of cells from the resistant model shifts to occupy an adjacent region of phenotypic space (such as clusters 1 and 4 in SC53, **Figure 4.5A and B**, and clusters 4, 5, and 8 in SC68 **Figure 4.6A and B**). RNA velocity shows that both cisplatin-naïve and cisplatin-relapsed samples exhibit positive velocity trending toward these clusters (cluster 1 in SC53 and cluster 4 in SC68), making them sinks in the phenotypic landscape (**Figures 4.5E-H and 4.6 E-H**). To fully visualize and investigate the movement in each model, we used Partition-Based Graphical Abstraction (PAGA) (Wolf et al., 2019). PAGA is a computational tool that reconciles clustering with continuous cell transitions inferred from RNA velocity. In each model, subpopulations enriched in cisplatin-resistant cells and SCLC-I features act as sources for the rest of the population (clusters 7, 9 and 10 in **Figures 4.5F** and clusters 9 and 10 in **Figure 4.6F**). As shown in **Figure 4.5F**, the SCLC-I cluster in the resistant sample is able to transition to cluster 6 in the main population, and in **Figure 4.6F**, a similar resistant SCLC-I cluster transitions to cluster 5. This suggests that, under dynamic equilibrium, SCLC-I cells act as a source that transitions toward the resistant sink clusters enriched with relapsed cells.

These observations suggest that the SCLC-I clusters in each model are highly plastic and may give rise to a more proliferative SCLC population. To test this possibility, we used our metric Cell Transport Potential (CTrP). Because the CTrP metric is a distance, it can be compared across samples, allowing us to compare the plasticity of SCLC-I cells to the rest of the resistant and the sensitive cells. The island clusters composed of relapsed, SCLC-I cells have markedly higher CTrP and, thus, plasticity (**Figures 4.5I-J and 4.6I-J**). Furthermore, resistant tumors show a near-universal increase in plasticity compared with sensitive tumors, and therefore SCLC-I cells are the most plastic phenotype regardless of treatment (**Figure 4.5K and 4.6K**). This suggests plasticity may be a defining characteristic of the resistant SCLC-I cells.

We also considered whether higher proliferation or lower death rates could explain SCLC-I as a source of progenitor, treatment-refractory cells. SCLC-I cells are not significantly upregulated in cell cycle (S or G2M) genes, nor do they significantly downregulate cell death markers (**Figure 4.5L-O**). In fact, assigning a phase to each cell suggests most SCLC-I cells are quiescent, with increased G1 genes, and may upregulate death markers slightly. Thus, SCLC-I cells appear to be neither more proliferative nor less death-susceptible than the rest of the population.

Together, these single-cell analyses suggest that cisplatin resistance coincides with the emergence of a cluster of cells that typify the SCLC-I subtype, apparently derived from cells originally SCLC-A that have undergone subtype switching associated with fluctuations in Notch pathway activation. These SCLC-I cells, in turn, are plastic such that the small population may be able to drive a resistant tumor phenotype.

4.2.5. Discussion

By analyzing tumor data directly, we were able to identify classes of patients that seem to have differential response to treatment. In particular, SCLC-I is an inflamed subtype of SCLC tumors that is preferentially sensitive to immune checkpoint blockade therapy. While SCLC-A, SCLC-N, and SCLC-P subtypes align with those we previously found in human cell lines in Chapters 2 and 3, we did not previously detect an inflamed subtype. This may be due to the fact that our previous methods, using cell lines that are generally considered to be less heterogeneous than tumors, may be better at detecting single-cell subtypes, whereas the methods described in this chapter find classes that are clinically relevant for subtyping patients. While a strict YAP1+ subtype was not found in tumors in this chapter, SCLC-P and SCLC-I tumors both showed an increase in YAP1 expression over SCLC-A and SCLC-N tumors. This may be because SCLC-P and SCLC-I tumor subtypes comprise SCLC-P *and* SCLC-Y cells in varying proportions, as well as other intermediate phenotype cells (generalists), immune infiltrate (as demonstrated by CIBERSORTx), and potentially true SCLC-I-subtype single-cells. Because both SCLC-P and SCLC-I tumors are enriched in SCLC-Y single-cells, YAP1 does not emerge as enriched in a single class of human SCLC tumors. Further analysis is needed to delineate the SCLC-Y single cell subtype, SCLC-I tumor subtype, and presence of immune populations.

Single-cell analyses presented here corroborate the intratumoral heterogeneity and plasticity predicted in previous chapters. Specifically, in the context of platinum resistance, a small, plastic, non-NE subpopulation may be capable of regenerating the rest of the tumor to overcome therapy. In favor of subtype switching as the underlying mechanism, one of the two models analyzed (MDA-SC68) had virtually no SCLC-I cells before treatment relapse, as the few that are triple negative lack other features of SCLC-I. On the other hand, those triple-negative cells that emerge following relapse show consistent, robust features of SCLC-I, and we observe evidence of transcriptional shifts away from SCLC-A and toward SCLC-I even among cells that remain ASCL1 positive. This work reinforces the clinical implications of phenotypic plasticity, and the need for treatment strategies that target plasticity directly to reduce acquired resistance in SCLC.

4.3. Network dynamics of variant SCLC mouse models with ASCL1 loss[4]

4.3.1. Motivation

Classic SCLC tends to be driven by the neuroendocrine transcription factor *ASCL1*. In contrast, about 25% of SCLC tumors have variant features, often driven by *NEUROD1* or non-neuroendocrine transcription factors such as *YAP1* or *POU2F3* (see Chapters 2 and 3). Furthermore, *MYC* family genes are often mutated and/or overexpressed, suggesting they play a critical role in SCLC progression. While ASCL1-driven SCLC is often characterized by high expression of the MYC family gene *L-MYC*, variant, *ASCL1*-low SCLC is associated with high *C-MYC* expression. As discussed in Chapter 3, the *ASCL1*-high subtype of SCLC can transition to variant subtypes in a *MYC*-driven genetically engineered mouse model (GEMM, RPM model

[4] This section is adapted from "ASCL1 represses a SOX9+ neural crest stem-like state in small cell lung cancer" published in Genes & Development and has been reproduced with the permission of the publisher and my co-authors. Olsen, R. R., Ireland, A.S., Kastner, D.W., **Groves, S.M**. et al. ASCL1 represses a SOX9+ neural crest stem-like state in small cell lung cancer. Genes & Dev 35:1-23 (2021) doi:10.1101/gad.348295.121.

discussed in Chapters 2 and 3). However, it is unclear if ASCL1 is a necessary precursor of SCLC-N or other variant subtypes.

ASCL1 and NEUROD1 are lineage-specifying transcription factors necessary for neural differentiation (Borromeo et al., 2016; Rudin et al., 2019). ASCL1, but not NEUROD1, is necessary for the development of pulmonary neuroendocrine cells (PNECs)(Ito et al., 2003), which, as discussed in Chapter 3, are often the cell of origin for SCLC. ASCL1, but not NEUROD1, is also required for the development of classical SCLC, as conditional deletion of ASCL1 was sufficient to abolish tumor formation in the classical, triple knockout mouse model of SCLC (RPR2) (Borromeo et al., 2016). However, in MYC-driven GEMMs, a similar deletion has not been tested, which would determine whether ASCL1 is also required for tumor formation in the RPM mouse model (Rudin et al., 2019). Here, we used GEMMs to determine the function of ASCL1 on cell fate and plasticity in MYC-driven SCLC derived from multiple cells of origin.

4.3.2. ASCL1 represses a SOX9+ neural crest stem-like state in small cell lung cancer

Ireland et al. (2020) showed that a MYC-driven mouse model of SCLC (RPM) can transition from neuroendocrine SCLC-A cells to non-neuroendocrine SCLC-Y cells via SCLC-N. In this model, mice are intratracheally infected with adenoviruses carrying cell type-specific promoters driving *Cre recombinase* expression, such as a general CMV promoter (Ad-CMV-Cre). Using various promoters to target different cell types (general CMV promoter, PNEC-specific CGRP, club cell-specific CCSP, and AT2-specific SPC), we initiated tumorigenesis in RPM mice. Interestingly, in situ tumors from each cohort were dominated by ASCL1-high cells, regardless of cell of origin. In contrast, invasive tumors were dominated by an ASCL1-low phenotype, with higher levels of NEUROD1 and YAP1.

RPM mice were next crossed to *Ascl1*-floxed animals to generate *Rb1*$^{fl/fl}$*;Trp53*$^{fl/fl}$*;Myc-T58A*$^{LSL/LSL}$*;Ascl1*$^{fl/fl}$ (RPMA) mice. All promoter-driven models developed tumors, though with significantly delayed latencies compared with RPM mice. This suggests that ASCL1 is not necessary for tumor formation in the MYC-driven mouse model. Interestingly, RPMA mice developed tumors with a tissue density consistent with bone, and bone analysis by microCT imaging confirmed that the tumors within the lung were bone-like. Histopathological analysis of H&E-stained tissues further validated that the lungs contained high-grade osteosarcoma with well-developed osteoid. These data suggest that ASCL1 represses a latent osteogenic fate in the context of MYC-driven SCLC.

RNA sequencing of RPR2, RPM, and RPMA tumors confirmed that all three tumor types are distinct, as they are well-separated in a principal components analysis. ASCL1 target genes were indeed reduced in RPMA tumors, and NEUROD1 was one of the most significantly down-regulated genes in RPMA versus RPM tumors. GSEA showed a significant positive enrichment for bone development genes in RPMA tumors, supporting the observed osteoid formation in these tumors. Furthermore, ossification-related processes were significantly upregulated in RPMA compared with RPM tumors, and neuronal development-related processes were significantly downregulated, as revealed by gene ontology (GO) enrichment analysis. Together, these data highlight a critical role of ASCL1 in promoting NE cell fate and repressing an underlying osteosarcoma-like fate in RPM mice.

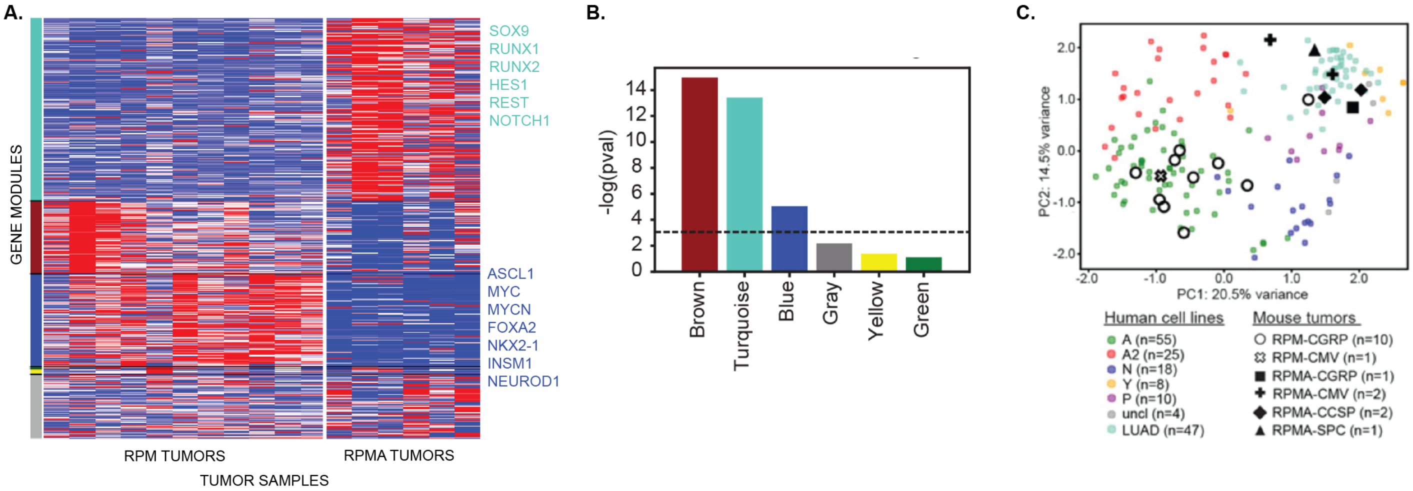


Figure 4.4: WGCNA reveals co-expressed gene modules that distinguish RPM and RPMA tumors. A. *WGCNA analysis on RNA-seq from RPM and RPMA tumors shows distinct gene programs that regulate each tumor type. Key transcription factors in the turquoise and blue gene modules shown.* ***B.*** *ANOVA statistical analysis of co-regulated gene modules identified by WGCNA in RPM (n=11) vs RPMA (n=6) tumors corresponding with the color code in Fig. 4A. Three modules had significant differential gene expression. The turquoise module is highly expressed in RPMA tumors, and brown and blue modules are highly expressed in RPM tumors. Data is shown as negative log10-transformed p-value. Dotted line indicates p = 0.05.* ***C.*** *Principal component (PC) analysis comparing bulk RNA-seq expression in human SCLC cell lines (SCLC lines from CCLE and cBioPortal and lung adenocarcinoma (LUAD) cell lines from CCLE) with mouse RPM or RPMA tumors initiated with the indicated viruses. Human SCLC cell lines were classified into subtypes as described in Chapter 3. Mouse tumors were harvested at the following time points post infection: RPM-CMV 55 d, RPM-CGRP 47–61 d, RPMA-CMV 85–86 d, RPMA-CGRP 111 d, RPMA-CCSP 120–204 d, and RPMA-SPC 204 d.*

4.3.3. Network analyses predict transcriptional regulators that drive osteosarcoma cell fate upon ASCL1 loss

To identify the key transcription factors responsible for this dramatic change in cell fate upon ASCL1 loss, we turned to weighted gene co-expression network analysis (WGCNA). We used an adapted version of the method described in Chapter 2 to generate a co-expression network of all genes, which allowed the identification of distinct gene modules across all RPM and RPMA samples (**Figure 4.4A and B**). Strikingly, approximately one-third of the transcriptome was altered upon ASCL1 loss (**Figure 4.4A**). Using PCA, we compared mouse tumors to the human SCLC cell lines from Chapter 3 and 47 lung adenocarcinoma cell lines from CCLE. We found that RPMA tumors clustered with the non-NE POU2F3 and YAP1 SCLC subtypes and lung adenocarcinoma (**Figure 4.4C**), suggesting a similarity between RPMA tumors and human non-NE tumors.

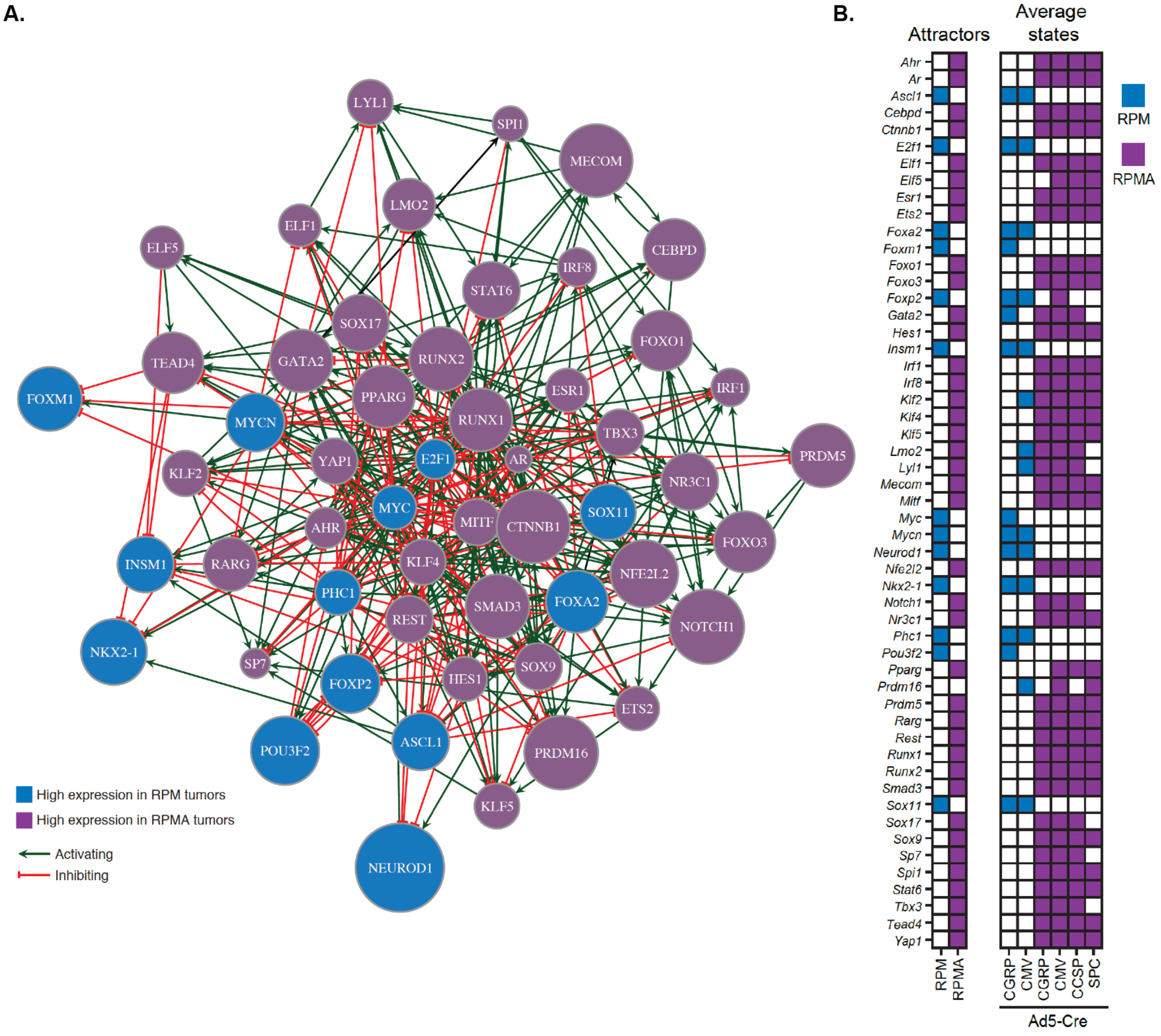


Figure 4.5: Network analyses predict transcriptional regulators that drive osteosarcoma cell fate upon ASCL1 loss. A. *Predicted transcription factor interaction network from RPM vs RPMA tumors based on the genes most central to each identified DEG module and data from ChIP-seq databases.* ***B****. Binarized average states and state-attractors in RPM (blue) and RPMA (purple) tumors initiated with the indicated Cre viruses. For each gene, colored squares are ON and white squares are OFF.*

To generate a gene regulatory network, we focused on transcription factors predicted by WGCNA to be central to differentially expressed gene (DEG) modules, as well as known regulators of lung cancer cell fate. Using the BooleaBayes pipeline developed in Chapter 2, we determined rules of interaction between transcription factors (**Figure 4.5A**). For example, ASCL1 is regulated by eight parent nodes (AR, E2F1, HES1, KLF4, MITF, NR3C1, PHC1, and RUNX1) where each ON/OFF combination of these parent nodes determines ASCL1 expression. Likewise, ASCL1 regulates expression of a number of downstream transcription factors. These regulations define how a cell may change its identity or reach a stable phenotype (an "attractor state").

Dynamic simulations identified two attractor states, each corresponding to either RPM or RPMA tumors (**Figure 4.5B**). Using the Hamming distance between states (the difference between two binary data strings), the RPM attractor was only one state away from the average RPM-CGRP state, the adenovirus promoter of which is specific to PNECs. The RPM-CMV model, with a general promoter, had an average state that was 7 states away from the RPM attractor. The RPMA attractor was closest to the average RPMA-CMV and RPMA-CGRP states (Hamming distance = 2). This may reflect the propensity of SCLC to develop from PNECs versus other lung cell types, as the resulting tumors are closer to an attractor and thus more stable.

4.3.4. TF network determines master regulators and destabilizers of RPMA tumors

Using the method described in Chapter 2, we ran a random walk simulation to predict regulators driving these steady states. Satisfyingly, *in silico* silencing of the ASCL1 node destabilized the RPM attractor, consistent with experimental results (**Figure 4.6A**). Conversely, activation of ASCL1 or NEUROD1 in the RPMA attractor destabilized that steady state, reminiscent of human SCLC, in which ASCL1 is destabilizes SCLC-Y as shown in Chapter 2.

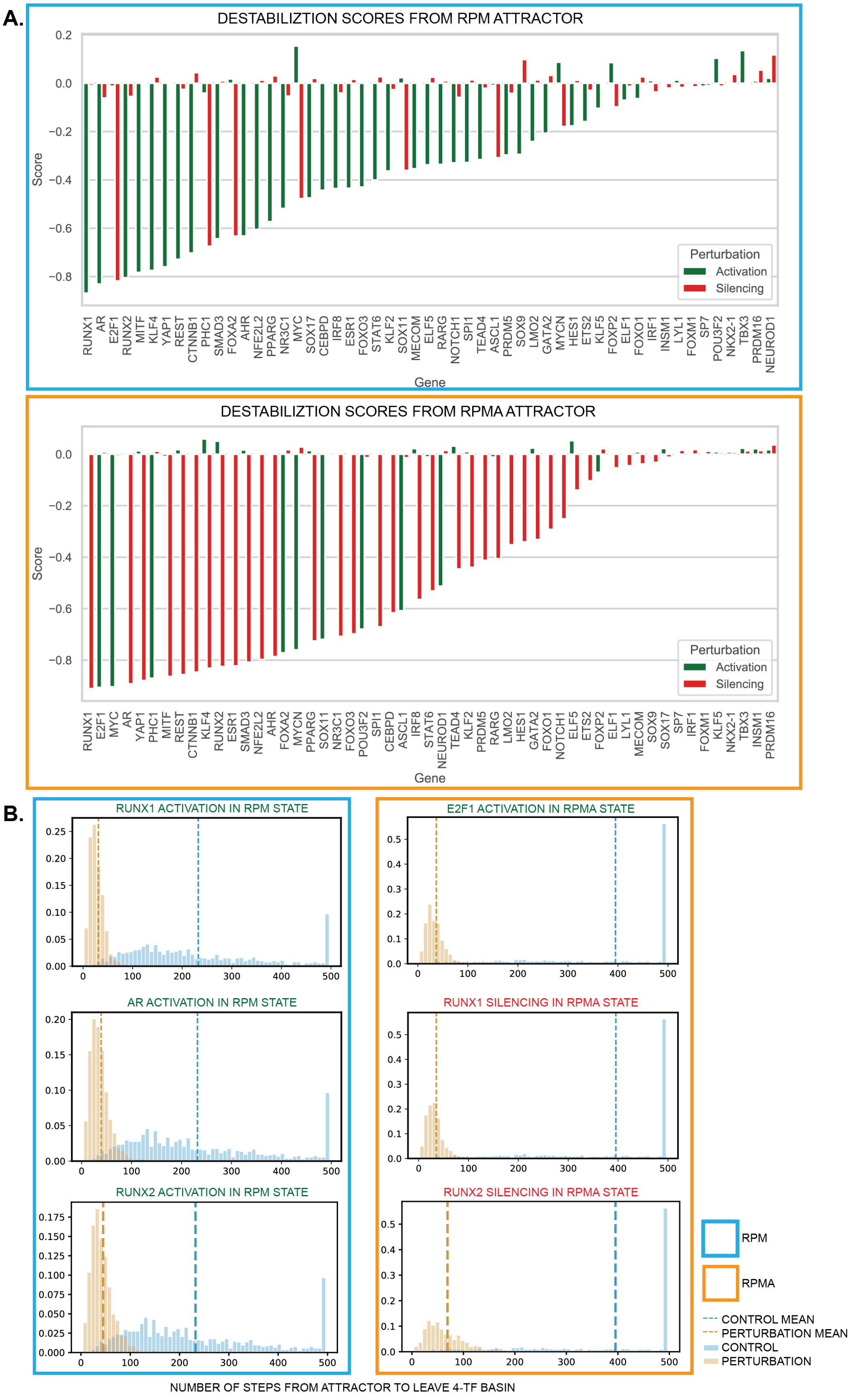

A.
DESTABILIZTION SCORES FROM RPM ATTRACTOR
0.2
0.0
−0.2
−0.4
−0.6
−0.8
Score
Perturbation
Activation
Silencing
RUNX1
AR
E2F1
RUNX2
MITF
KLF4
YAP1
REST
CTNNB1
PHC1
SMAD3
FOXA2
AHR
NFE2L2
PPARG
NR3C1
MYC
SOX17
CEBPD
IRF8
ESR1
FOXO3
STAT6
KLF2
SOX11
MECOM
ELF5
RARG
NOTCH1
SPI1
TEAD4
ASCL1
PRDM5
SOX9
LMO2
GATA2
MYCN
HES1
ETS2
KLF5
FOXP2
ELF1
FOXO1
IRF1
INSM1
LYL1
FOXM1
SP7
POU3F2
NKX2-1
TBX3
PRDM16
NEUROD1
Gene
DESTABILIZTION SCORES FROM RPMA ATTRACTOR
0.0
−0.2
−0.4
−0.6
−0.8
Score
Perturbation
Activation
Silencing
RUNX1
E2F1
MYC
AR
YAP1
PHC1
MITF
REST
CTNNB1
KLF4
RUNX2
ESR1
SMAD3
NFE2L2
AHR
FOXA2
MYCN
PPARG
SOX11
NR3C1
FOXO3
POU3F2
SPI1
CEBPD
ASCL1
IRF8
STAT6
NEUROD1
TEAD4
KLF2
PRDM5
RARG
LMO2
HES1
GATA2
FOXO1
NOTCH1
ELF5
ETS2
FOXP2
ELF1
LYL1
MECOM
SOX9
SOX17
SP7
IRF1
FOXM1
KLF5
NKX2-1
TBX3
INSM1
PRDM16
Gene
B.
RUNX1 ACTIVATION IN RPM STATE
0.25
0.20
0.15
0.10
0.05
0.00
0
100
200
300
400
500
AR ACTIVATION IN RPM STATE
0.20
0.15
0.10
0.05
0.00
0
100
200
300
400
500
RUNX2 ACTIVATION IN RPM STATE
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
100
200
300
400
500
E2F1 ACTIVATION IN RPMA STATE
0.5
0.4
0.3
0.2
0.1
0.0
0
100
200
300
400
500
RUNX1 SILENCING IN RPMA STATE
0.5
0.4
0.3
0.2
0.1
0.0
0
100
200
300
400
500
RUNX2 SILENCING IN RPMA STATE
0.5
0.4
0.3
0.2
0.1
0.0
0
100
200
300
400
500
RPM
RPMA
CONTROL MEAN
PERTURBATION MEAN
CONTROL
PERTURBATION
NUMBER OF STEPS FROM ATTRACTOR TO LEAVE 4-TF BASIN

Figure 4.6: Destabilization of attractors. A. *Destabilization score of each perturbation (activation or silencing) of each TF.* ***B.*** *Select perturbations with large destabilizations show a clear decrease in steps to leave an attractor basin.*

Based on these results, we included other known NE fate specifiers like INSM1 and transcription factors important in endodermal and lung adenocarcinoma fate, such as NKX2-1 and FOXA2, in our network. While all of these genes were significantly reduced in invasive RPMA tumors (Fig. 4D), only FOXA2 was predicted to significantly affect the dynamics of the transcription factor network. In fact, activation of RPMA genes was more destabilizing, on average, than silencing of RPM genes like INSM1 (**Figure 4.6A**). RUNX1 activation was most destabilizing to the RPM attractor and silencing was most destabilizing to the RPMA attractor, suggesting it may play a central role in regulation of the RPMA fate (**Figure 4.6B**). Activation of the androgen receptor gene, AR, significantly destabilized RPM (**Figure 4.6B**). Interestingly, RUNX1 is a target of AR and both function in chondrogenic lineage commitment of mesenchymal progenitor cells (Hui et al., 2021; Smith et al., 2005). Furthermore, E2F1 is predicted to be a master regulator of RPM and destabilizer of RPMA, such that silencing this gene destabilized RPM whereas activating this gene pushes cells away from the RPMA attractor (**Figure 4.4B**). Constitutive overexpression of E2F1 has previously been shown to delay bone formation by inhibiting chondrocyte differentiation (Scheijen et al., 2003), which is consistent with our network predictions and strengthens the connection between the RPMA state and chondrocyte and bone differentiation.

4.3.5. Network simulations probe mesenchymal stem cell-like reprogramming of RPMA tumors

Several other transcription factors in the network were predicted to be master regulators or destabilizers (**Figure 4.6A**). To deepen our understanding of these predicted regulators, we considered their role in normal development. The lung epithelium is believed to derive largely from endoderm, whereas bone and cartilage fates are derived from mesoderm or ectoderm (Serizawa et al., 2019; Zepp and Morrisey, 2019). Mesenchymal stem cells (MSCs) from mesodermal tissue and neural crest stem (NCS) cells from ectodermal tissue are capable of differentiating into neuron, bone, or cartilage fates (Achilleos and Trainor, 2012; Jiang et al., 2002). In the RPMA tumors analyzed here, GSEA revealed that both MSC and NCS cell signatures were significantly enriched, suggesting that ASCL1 is associated with a more endodermal tumor cell fate and may repress an MSC/NC-stem-like fate. We therefore considered the developmental pathways that are known to drive bone fate from MSC and NCS progenitor cells, including Indian hedgehog (Ihh), Hippo/Yap1, Transforming growth factor-β (Tgf-β), Bone morphogenic protein (Bmp), Wnt, Notch, and others (Long, 2012; Pan et al., 2018; Vanyai et al., 2020). Consistent with our GSEA analyses, several components of these pathways were upregulated in RPMA tumors, and network analyses predicted that several destabilizers of the RPM phenotype fell within these pathways (YAP1, CTNNB1, HES1, and REST). Together, these data suggest that ASCL1 represses the emergence of an MSC/NC-stem-like state as well as multiple developmental pathway regulators in MYC-driven SCLC.

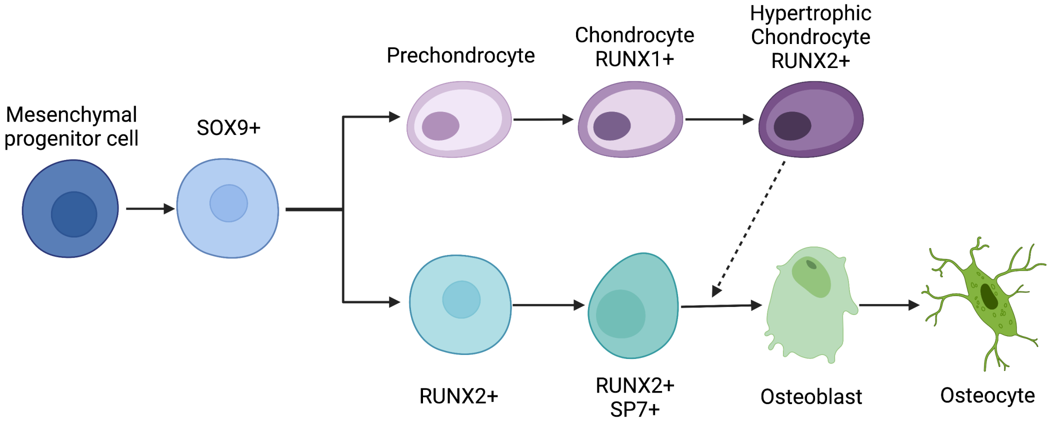


Figure 4.7: Role of SOX9 in chondrocyte and bone differentiation.

We examined the role of SOX9 in regulating the RPMA state, as pathway analysis indicated that several enriched genes in RPMA compared to RPM are targets of SOX9. In normal development, SOX9 marks osteoblast progenitors and precedes RUNX2 expression and bone differentiation (Long, 2012) (**Figure 4.7**). Furthermore, RUNX1, a master regulator of the RPMA state, can directly regulate SOX9 and RUNX2 during bone development. In our data, RUNX1 and SOX9 expression were enriched in noncalcified, “soft” RPMA tumors compared with more differentiated osteosarcomas, whereas RUNX2 levels were high in both types of RPMA tumors. SP7 expression, which normally follows RUNX2 expression in differentiation of bone-producing osteoblasts, was predominantly expression in the more differentiated osteosarcomas. While ASCL1 knockdown induced SOX9 expression in human SCLC cell lines, ASCL1 ChIP-seq data did not identify significant ASCL1 binding sites near SOX9, which suggests that SOX9 repression by ASCL1 is likely indirect.

To further investigate the transcriptional regulation of RPM to RPMA reprogramming, we ran 1000 network simulations starting from the RPM attractor under unperturbed dynamics and with ASCL1 silenced during the entire walk (**Figure 4.8**). A PCA transformation of all of the binary walks shows that a single dimension (PC1) is capable of separating the RPM and RPMA attractors, which is corroborated by the loadings of each TF (**Figure 4.8A and B**). In PC2, however, there is less of a clear split between RPM and RPMA TFs. Instead, RUNX1 and SPI1 dominate PC2, while FOXM1, CTNNB1, SOX9, and AR are negatively correlated with PC2. In **Figure 4.8C**, the random walks that successfully reach the RPMA attractor (within 10,000 steps) are shown in a contour plot, with example random walks overlaid on the left plot. From these plots, it is clear that most paths move somewhat monotonically through PC1, but first decrease and the increase in PC2, finally evening out around the same location in PC2 as the RPM attractor once they reach the RPMA attractor. This may indicate that most paths first activate SOX9, AR, and CTNNB1 (first dense region of states the paths reach), followed by an increase in SPI1, RUNX1, and other RPMA-related TFs (second dense region). This ordering is consistent with activation in MSC differentiation, where SOX9 precedes RUNX1 and RUNX2 expression (**Figure 4.7**).

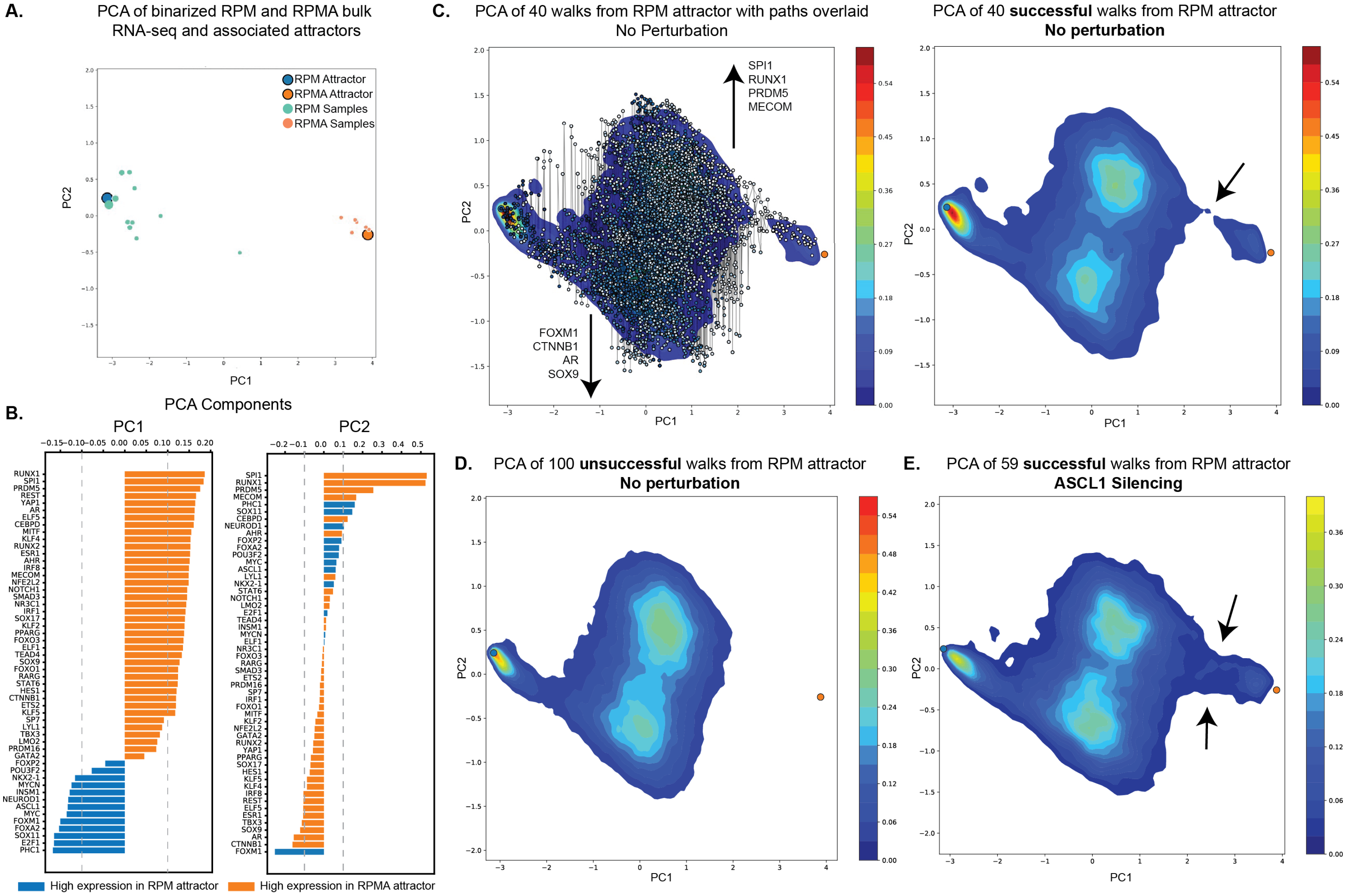


Figure 4.8: Network simulations of RPM to RPMA reprogramming by ASCL1 silencing. ***A.*** *We ran 1000 random walk simulation from the RPM attractor, stopping if it reached the RPMA attractor or reached 10,000 steps. We then fit a PCA to all states in the random walks. Shown are the attractors and tumor samples in this PCA space.* ***B.*** *By analyzing the principal components, we get a sense as to how each TF varies in each dimension. TFs are colored by their expression, enriched in the RPM (blue) or RPMA (orange) attractor as determined in* ***Figure 4.5B****.* ***C****. Paths of 40 unperturbed, successful random walks (that reach the RPMA attractor) are overlaid on a contour plot (left). The paths, which generally move from left to right in the first PC, tend to first decrease in PC2 and then increase before reaching the RPMA attractor (which is similar to the RPM attractor in PC2). The highest density of states along the paths (after leaving the RPM attractor) appears as two "hills" in the PCA. The first dense region, which decreases in PC2, suggests cells first enter a state with expression of FOXM1, CTNNB1, AR, and SOX9. After reaching this state, cells move to a second dense region which increases in PC2, suggesting a state increasing expression of SPI1, RUNX1, PRDM5, and MECOM. The successful walks then cross a "thin bridge" to the RPMA attractor, suggesting a bottleneck of possible states before full reprogramming to the RPMA state.* ***D.*** *Compared to C, paths that do not successfully reach the RPMA attractor (before maximum of 10,000 steps) are still capable of reaching the two dense regions in the PCA space. This suggests that the thin bridge in C is truly a bottleneck for accessing the RPMA attractor.* ***E.*** *With ASCL1 silenced in silico (ASCL1 is kept OFF during the simulations), more walks reach the RPMA attractor. This may be due to the fact that the bottleneck is widened by ASCL1 loss, making it easier to reach the attractor successfully.*

Even more notable is the fact that, after reaching each dense region in the PCA, the paths that successfully reach the RPMA attractor seem to cross a “thin bridge” that acts as a bottleneck to full reprogramming (**Figure 4.8C, right**). Compared to **Figure 4.8D,** it becomes clear that this bottleneck is what prevents the unsuccessful simulations from reaching the RPMA attractor, as the unsuccessful walks still reach both intermediate dense regions. This suggests that it may be possible to reach the RPMA state under the normal dynamics of RPM development, but the transition is unlikely due to this bottleneck in the reprogramming trajectory.

To mimic the RPMA tumor model, we silenced ASCL1 *in silico* (i.e., ASCL1 remains OFF during the entire walk) and reran the simulations (**Figure 4.8E**). Interestingly, the paths follow much the same trajectory as the unperturbed simulations, again reaching two dense regions intermediately during reprogramming. However, ASCL1 silencing removes the bottleneck, allowing more of the simulations to successfully reach the RPMA attractor. This suggests that ASCL1 is indeed constraining evolution of RPM tumors from fully reprogramming to the RPMA state, as suggested by ASCL1 repression of the MSC-like state. While it is still unclear exactly what the mechanism of this bottleneck is, further experimental analysis of the RPM to RPMA reprogramming pathway, such as through lineage tracing or temporally controlled activation of particular genes, may provide insight.

4.3.6. Discussion

In this work, we investigate the question of whether ASCL1 is necessary for SCLC development. While it appears to be required for classic SCLC, as conditional deletion of ASCL1 is sufficient to abolish tumor formation in a classical mouse model (RPR2) (Borromeo et al., 2016), ASCL1 is not necessary for SCLC development in a MYC hyperactivated mouse model (RPM). The data here suggest that loss of ASCL1 could potentially convert SCLC to an alternative cell fate, but fate-tracking approaches will be needed to definitively address this possibility. Interestingly, RPMA tumors with ASCL1 silenced began to ossify in the lung, suggesting that ASCL1 is capable of repressing a latent osteogenic program. This surprising result is supported by increases in mesenchymal stem cell-like and neural crest stem cell-like gene programs, as well as activation of Notch pathway genes and bone differentiation driving transcription factors, such as RUNX2.

Using a transcription factor network fit to RPM and RPMA tumor data, we made predictions of master destabilizers and regulators that control the possible NE and bone fates of these tumors. Interestingly, activation of many of the major drivers of bone fate significantly destabilize the RPM attractor, whereas silencing of RPM drivers has a smaller effect. This signifies that turning on bone and chondrocyte developmental genes, such as SOX9, RUNX1, RUNX2, is critical to reprogramming RPM tumors to an RPMA state under ASCL1 loss.

The importance of SOX9 and possibility of a mesenchymal stem cell-like phenotype in this work points toward the relevance of developmental biology in understand SCLC tumor dynamics (**Figure 4.9**). Until recently, PNECs (and, as a correlate, SCLC) were thought to arise from the neural crest, which is ectodermal in origin (similar to other neuronal cells) (Pearse and Polak, 1971). It is now clear that PNECs share the development origin of other lung cell types, arising from local multipotent stem cells (Gazdar et al., 2017; Nikolić et al., 2018; Rosai, 2011). Answering the question of how cells with an ectodermal origin could potentially transition to a mesenchymal stem cell-like state, which is mesodermal in origin, will be critical to our understanding of RPMA dynamics. Do RPMA cells reprogram from a lung progenitor-like state to fully pluripotent stem cells before transitioning to an MSC state? Calcification and ossification of the lung is rare but possible (Chan et al., 2012), and SCLC tumors preferentially metastasize to the bone, suggesting the relevance of this question (Ko et al., 2021; Nakazawa et al., 2012). Furthermore, recent research shows that some mesenchymal stem cells may reside in the lung and possess lung-specific properties (Enes et al., 2016). These lung-resident mesenchymal stem cells may play a role in tissue remodeling under conditions of oxidative stress and lung cancer development (Chow et al., 2013; Sentek and Klein, 2021; Sveiven and Nordgren, 2020). Future studies should investigate the relationship between the RPMA osteogenic state and these MSC populations.

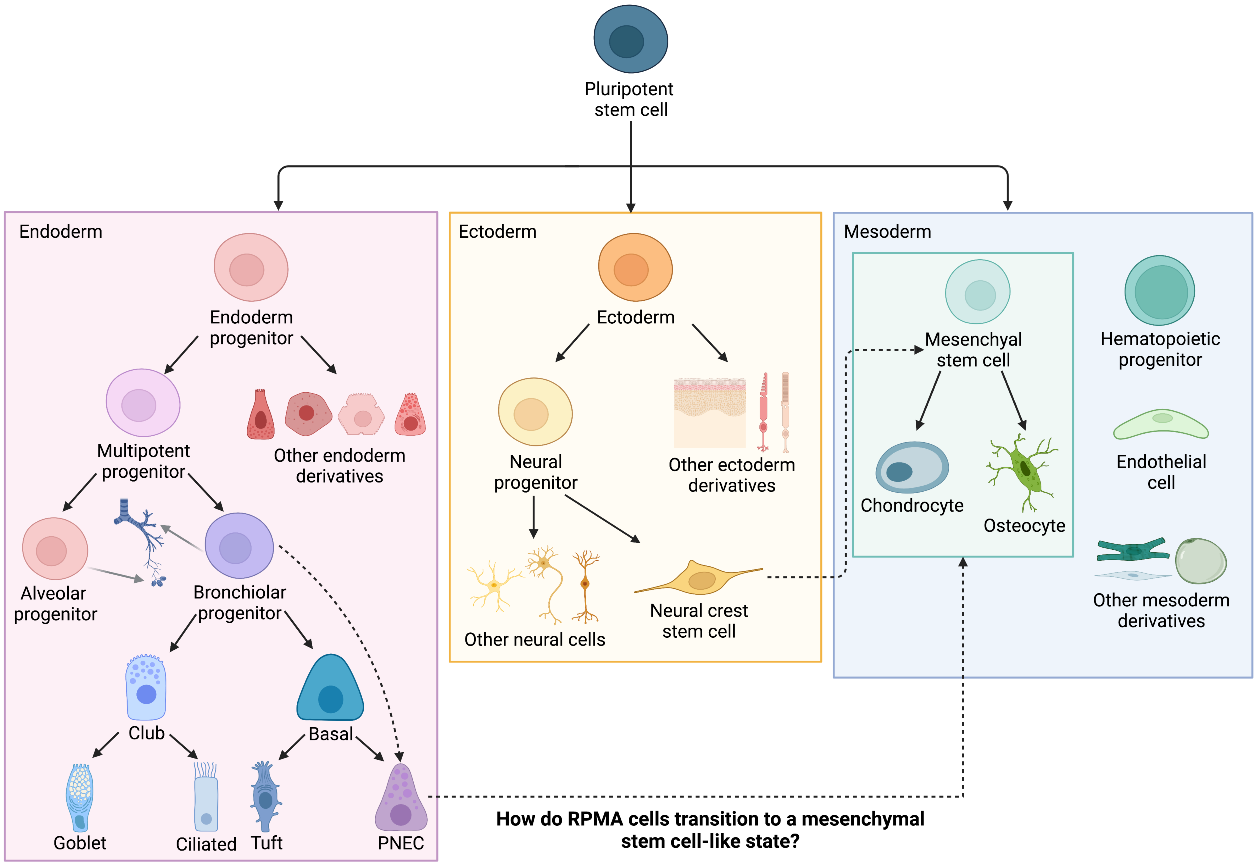


Figure 4.9: RPMA progression to an osteogenic state can be understood in terms of normal developmental biology.

4.4. Conclusions

In this chapter, we explore variant SCLC tumors through the lens of systems biology. While most SCLC are ASCL1+, about a third can be considered variant, with different morphology, expression of non-NE markers, and inflammatory properties. We first identified a new SCLC subtype of tumors, SCLC-I, that has increased immune infiltrate and higher levels of immune checkpoints and human leukocyte antigens. This class of tumors from SCLC patients has not been fit into our current paradigm of SCLC heterogeneity, as explored in Chapters 2 and 3. There are a few possible explanations for this. Since we explored the same dataset of 81 human tumors in Chapter 3 and showed that variance between these tumors is well explained by the archetypes in that chapter, we may find that the SCLC-I subtype is describing a subset of tumors close to the non-NE subtypes that do not have significantly increased YAP1 expression over the SCLC-P tumors. While we define one of the archetypes in Chapter 3 by YAP1 expression due to its enrichment over other archetypes, not all samples near the SCLC-Y archetype show high expression of YAP1. In this Chapter, the SCLC-I subtype shows positive expression of SCLC-Y that is not significantly enriched over the SCLC-P tumors, suggesting that the tumors were samples not sufficiently close to, but still could be described by their distance from, the SCLC-Y archetype. Alternatively, these SCLC-I tumors may be generalists in the archetype space, and therefore not enriched for any of the four defining TFs in the field, including YAP1.

Another hypothesis, which was introduced in Chapter 3, is that the SCLC tumors are more heterogeneous mixtures of phenotypes across archetype space, and it is the relative proportions of these subpopulations that define the patterns seen in this Chapter. This is consistent with our evidence in Chapter 3 that tumors form a polytope in linear space but not logarithmic space, which is what would be seen if tumors were heterogeneous mixtures of distinct phenotypes, rather than samples fit within a Pareto optimal polytope. In this case, the SCLC-I class of tumors would not optimize a distinct task in trading off with other phenotypes, but instead these tumors would be a mix of those other phenotypes, plus a higher number of immune cells, in such proportions that the samples are distinct in bulk RNA-seq and IHC. Non-NE cancer cells in these tumors may be pulled away from YAP1+ and POU2F3+ phenotypes due to their interactions with more immune infiltrate.

Another possible explanation for SCLC-I that is less consistent with our work in Chapter 3 is that it is indeed a completely distinct, novel phenotype that individual cells can acquire. Further investigation into these hypotheses should be able to determine whether SCLC-I tumors comprise cells within the archetype space defined in Chapter 3 with varying amounts of immune infiltrate, or individual cells of a novel phenotype.

In this chapter, we explore the dynamics of tumors that increase in SCLC-I after treatment. We use our plasticity pipeline described in Chapter 3 to calculate Cell Transport Potential and show that the SCLC-I-enriched subpopulation after treatment may be capable of regenerating the rest of the tumor. If this is true, SCLC-I and non-NE cell types may be capable of transitioning to NE subtypes, particularly after treatment. Lineage tracing studies should be able to determine whether non-NE cells can transition in this direction, as opposed to NE to non-NE transitions which are more common (Lim et al., 2017). Critically, our plasticity analysis shows that the majority of cells after treatment increase in plasticity, which may explain the aggressiveness of relapsed SCLC. This increase in plasticity could be tested in mouse models of SCLC before and after treatment in which distinct lineages can be labeled with a fluorescent reporter.

While many tumors, including the ones discussed in this chapter, decrease in ASCL1 expression after relapse from chemotherapy, it has been unclear whether progression of SCLC is initially dependent on ASCL1 activation. In this chapter, we show that mouse models without ASCL1 (RPMA) can form SCLC tumors from a variety of lung cell types. Furthermore, ASCL1 inactivation leads to a variant SCLC phenotype that is enriched in osteogenic and mesenchymal stem cell-like gene programs.

Because inactivating mutations of ASCL1 are not common in SCLC (George et al., 2015), this novel phenotype may “escape beyond” the archetypes defined in Chapter 3. If a mutation is capable of “rewiring” the cell’s gene regulatory network in such a way that the mutant is more optimal at more than one task, the Pareto front that constrains cells may shift or become obsolete (Li et al., 2019). In the case of ASCL1 inactivation, about a third of the transcriptome is altered, suggesting that this single mutation is capable of significantly changing the accessible phenotypes for MYC-driven SCLC and therefore may change the Pareto front and the associated fitness function, which is dependent on trade-offs in performance of SCLC tasks. Interestingly, the network analysis on the RPM to RPMA transition suggests that RPM cells can reach the RPMA state without specific inactivation of ASCL1; the natural dynamics of the system may allow for cells to turn off ASCL1 expression without a genetic modification. This suggests that the RPMA state falls farther along the same phenotypic trajectory as normal MYC-driven SCLC, which is corroborated by pseudotime analysis of scRNA-seq from RPM and RPMA tumors in Olsen et al. (2021). Furthermore, if the RPMA attractor does match a new archetype in SCLC phenotypic space, it may be accessible for RPM tumor cells under epigenetic modification. Lastly, there may be a connection between the RPMA steady state and the archetype X absorbing state found in Chapter 3, which was not enriched in any of the five canonical SCLC signatures. Further investigation into both of these phenotypes is needed to uncover the specific relationship, if any, that exists between them.

Together, this chapter investigates two phenotypes that have not been previously described by our SCLC network or archetype analysis. A comprehensive understanding of SCLC phenotypes and transitions between them will require incorporation of these distinct states into a broader framework of heterogeneity.

Chapter 5.

Conclusion

5.1. Discussion

While several cancer types have seen marked improvement in treatment options in recent years, SCLC patients are still relegated to the same fate as half a century ago. Fortunately, the Recalcitrant Cancer Research Act signed in 2012 imbued SCLC research with new funding and enthusiasm that led to rapid developments in our understanding of the disease. In the last decade alone, SCLC research has seen a substantial increase in resources, including new genetically engineered mouse models (double- and triple-knockout models, and RPM/RPMA variant models), xenograft models (PDXs and CDXs), and genomic and epigenomic data directly from human tumors (Cui et al., 2014; Drapkin et al., 2018; George et al., 2015; Hodgkinson et al., 2014; McFadden et al., 2014; Mollaoglu et al., 2017; Olsen et al., 2021; Peifer et al., 2012; Rudin et al., 2012; Schaffer et al., 2010; Simpson et al., 2020; Song et al., 2012; Stewart et al., 2020; Sutherland et al., 2011; Tlemsani et al., 2020; Williamson et al., 2016; Zhang et al., 2018). The rapid progress of new translational discoveries has been called “The Second Golden Age” of SCLC (Gazdar et al., 2017).

This increase in resources has been supported by a simultaneous explosion in technological advancements in biology, due in part to the decreasing cost of genomic sequencing (Wetterstrand) and the development of single cell sequencing (Hong et al., 2020). As described in this dissertation, single cell data can be incredibly useful to understanding the epigenetic heterogeneity present within SCLC cell lines and tumors and is beginning to illustrate the importance of phenotypic plasticity in SCLC relapse (Gay et al., 2021; Groves et al., 2021; Ireland et al., 2020; Lim et al., 2017; Wooten et al., 2019). These recent improvements in our understanding of the biology of SCLC impart optimism for therapies that can better treat patients, particularly after relapse from the standard of care regimen. This work contributes to a larger goal of understanding how phenotypic heterogeneity and plasticity may promote acquired resistance in SCLC and proposes strategies to overcome plasticity-driven relapse.

In this work, we examine SCLC phenotypic heterogeneity by delineating subtypes (Chapter 2), placing those subtypes in the context of multi-objective evolutionary theory (Chapter 3), and exploring new phenotypes that arise in variant SCLC models (Chapter 4). With this comprehensive view of SCLC subtypes as occupying basins in an epigenetic landscape, we can begin to develop new strategies for overcoming treatment evasion caused by phenotypic plasticity.

5.1.1. Systems biological approaches uncover strategies to control SCLC response

A systems-level understanding of biology ultimately requires the ability to control the system (Kitano, 2002). In this dissertation, I both make and test predictions for ways to control SCLC phenotypic heterogeneity. For example, in chapter 2, we make several predictions for master stabilizers and destabilizers of each phenotype; experimental perturbations of these TFs are critical to validate their importance in SCLC cell identity regulation. In chapter 3, we show preliminary evidence that SCLC phenotypes may optimize functions related to normal PNECs, causing differential sensitivity to drugs targeting those functions. Future work is needed to determine how well cells can change phenotype in response to these drugs towards more-resistant states. For example, if we find that an inactivation of a master regulator forces a phenotypic transition from phenotype A to phenotype B, that same perturbation should affect the sensitivity of the cells to treatments targeting phenotype A. Together, these ideas suggest a framework of manipulating phenotypic identity to “corral” cells into a sensitive state for a given drug.

In chapter 4, we use a transcription factor network to predict master regulators that may be able to drive a transition from an RPM state to an RPMA state. We then investigate the role of a key driver of the mesenchymal stem cell fate, SOX9, and find that ASCL1 represses expression of SOX9 in SCLC. Further work is needed to test other predicted drivers, such as RUNX1 and RUNX2. We also identify a new class of SCLC tumors that are inflamed and use a plasticity analysis to show that cells high in the SCLC-I signature seem to be more plastic. This may be tested in GEMMs or PDX models by lineage tracing to determine if SCLC-I cells after treatment can truly regenerate the NE subpopulation. Finally, a true understanding of this system would require the ability to push tumors in and out of the SCLC-I state, possibly in response to treatment.

5.1.2. Unlocking phenotypic plasticity may be key to relapse in SCLC

Cell state plasticity was first recognized in organogenesis where a stem cell, without changing its genetics, is capable of terminally differentiating into various cell fates. Waddington's landscape, a metaphor now evoked in systems ranging from the cell cycle to cancer, was originally advanced as an abstract picture of the developmental process, where cells roll down canalizations in a landscape as they become more and more differentiated, less plastic, and therefore less able to acquire new phenotypes (Waddington, 1957). The landscape metaphor has been shown to be a useful model for non-genetic heterogeneity in several cancer systems (Brock et al., 2009; Huang, 2013; Huang and Kauffman, 2013; Huang et al., 2009; Menendez, 2015; Pisco and Huang, 2015; Wooten and Quaranta, 2017; Zhou et al., 2016a). In particular, quantification of landscape potential can help uncover key insights into the regulation of systems with multiple stable subtypes (i.e. attractors), such as in SCLC. Our results in Chapter 2 show that attractors of the inferred network—i.e. the most stable states as defined by regulatory network dynamics—match the empirical phenotypes seen across cell lines and tumors, both in our work and as defined by other research groups (Baine et al., 2020; Borromeo et al., 2016; McColl et al., 2015; Rudin et al., 2019). This concordance demonstrates that the phenotypic landscape is a good model for non-genetic heterogeneity in SCLC that can be utilized to make predictions regarding control of cell identity in this cancer..

Phenotypic heterogeneity and plasticity are now considered hallmarks of cancer (Hanahan, 2022). Heterogeneity in cancer may lead to resistance through selection of a pre-existing resistant subpopulation, as shown in **Figure 5.1.** In contrast, phenotype plasticity allows cells to traverse the epigenetic landscape to evade treatment by changing phenotype. Plasticity may be reflected in the epigenetic landscape as lower barriers between cells and less stable attractors between them, particularly after treatment (**Figure 5.1**).

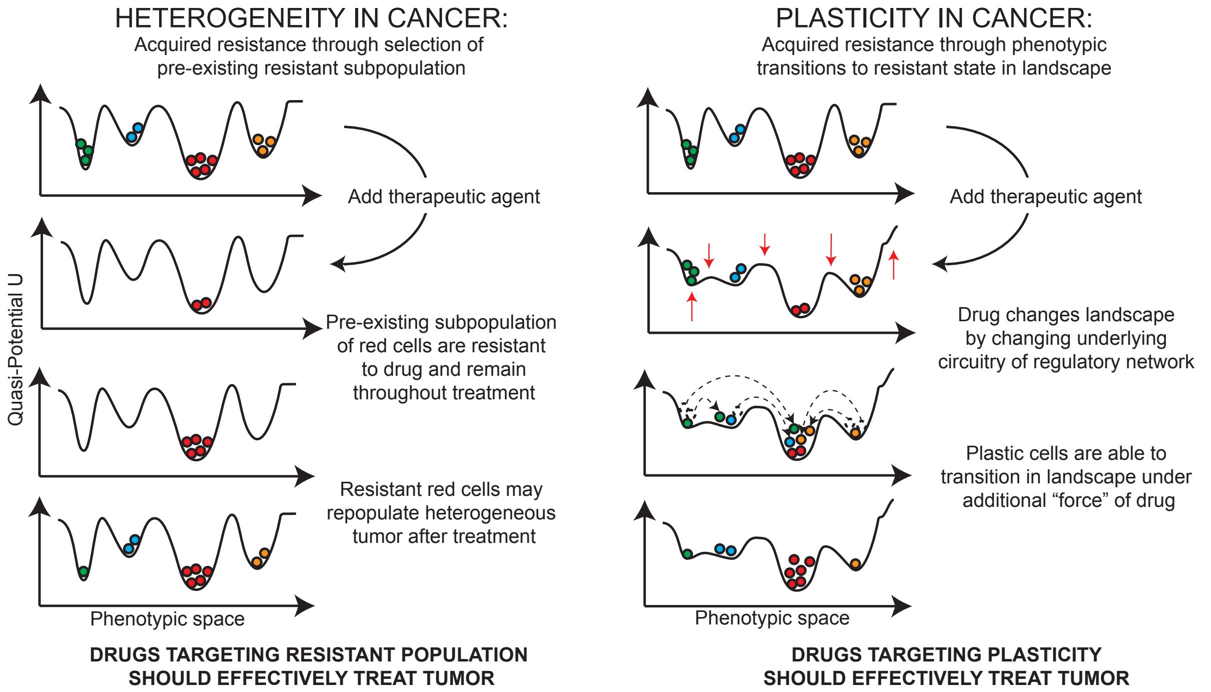

HETEROGENEITY IN CANCER:
Acquired resistance through selection of pre-existing resistant subpopulation
Add therapeutic agent
Pre-existing subpopulation of red cells are resistant to drug and remain throughout treatment
Resistant red cells may repopulate heterogeneous tumor after treatment
Quasi-Potential U
Phenotypic space
DRUGS TARGETING RESISTANT POPULATION SHOULD EFFECTIVELY TREAT TUMOR
PLASTICITY IN CANCER:
Acquired resistance through phenotypic transitions to resistant state in landscape
Add therapeutic agent
Drug changes landscape by changing underlying circuitry of regulatory network
Plastic cells are able to transition in landscape under additional "force" of drug
Phenotypic space
DRUGS TARGETING PLASTICITY SHOULD EFFECTIVELY TREAT TUMOR

Figure 5.1: Heterogeneity vs plasticity in cancer relapse and treatment strategies.

As demonstrated in this dissertation, both phenotypic heterogeneity and plasticity are critical to SCLC development, survival, and relapse. Furthermore, characterization of phenotype plasticity via a quantifiable landscape leads to predictions for possible regulation strategies that could lead to better treatment. Chapter 2 shows how intra-tumoral heterogeneity is undergirded by GRN dynamics and predicts strategies for controlling plasticity.

In chapter 4, plasticity plays a clear role in relapse of SCLC-I tumors. Furthermore, we identified a relationship between MYC-driven SCLC with ASCL1 loss and a mesenchymal stem cell-like state. Together, these results suggest that the phenotypic transitions of SCLC in various contexts may be explained as "unlocking" plastic de-differentiation capabilities, similar to normal multipotent progenitors.

5.1.3. Normal biology can inform our understanding of SCLC dynamics

The epigenetic landscape framework provides a quantifiable metaphor for understanding the relationship between cancer and normal cell development. The idea of cancer states as attractors can be integrated into a normal developmental landscape, where the genomic instability of cancer cells allows them to reach new attractors previously inaccessible in the landscape. This may explain how SCLC cells, in some contexts, seem to be able to replicate the de-differentiation of normal cells; for example, the SCLC-A to SCLC-Y transition is reminiscent of the lung injury repair ability of normal PNECs that trans-differentiate to other cell types to replenish the lung epithelium (Ouadah et al., 2019). Similarly, the transition to an MSC-like state in RPMA tumors may reflect the de-differentiation capability of some normal cell types. It is unclear how transition paths from new cancer attractors within this landscape are capable of reaching these inherent progenitor attractors.

While there seems to be a relationship between the landscape of PNECs, which can optimize various functions through distinct gene regulatory programs, and SCLC subtypes, it is not yet clear if PNEC functions represent an actionable constraint, or if they are simply a starting point that SCLC cells can easily expand beyond. Mutations that allow for optimization of multiple objectives

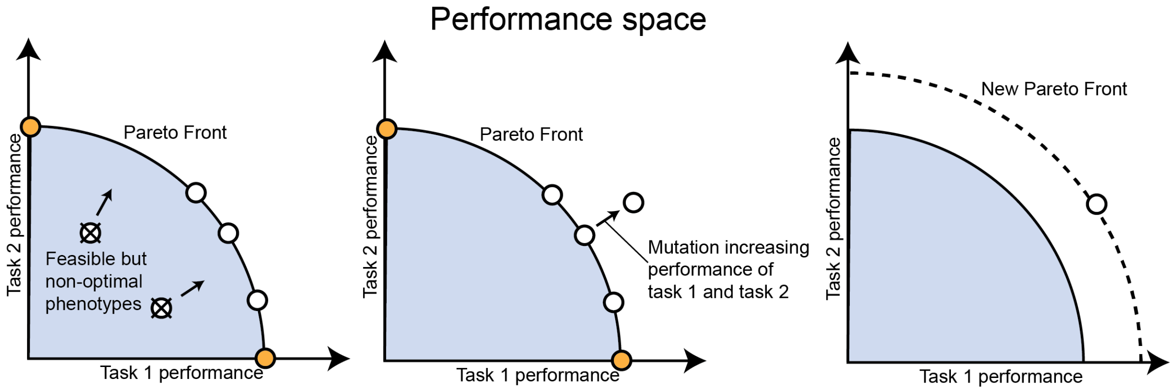


Figure 5.2: Mutations are capable of expanding the Pareto Front. *A mutation may allow a cell to move beyond the Pareto Front if it can become more optimal at both tasks. Selection of this more fit mutant will, over time, expand the Pareto Front (right).*

at once may allow cells to move "beyond" archetypes, towards phenotypes that have higher fitness. This is akin to changing the shape of the Pareto front, which may be achieved by the high mutational burden of SCLC tumors (**Figure 5.2**).

The high mutational burden in SCLC would seem to promote completely unconstrained progression of phenotypic heterogeneity, so it is interesting that the same five archetypes seem to recur in various SCLC models, in various conditions like human cell lines, human tumors, classic and variant GEMMs, and PDX models. Some of these models may pinpoint the mechanism of expansion to novel archetypes— for example, RPM tumors seem to move beyond the five archetypes defining SCLC phenotypic space toward an unknown Archetype X. Still, the recurrent pattern of phenotypes across models suggests that the defined functions are key to SCLC development and survival. It is currently unclear if any novel phenotypes, such as Archetype X or SCLC-I, play a large role in acquired resistance, or if characterization of the five main SCLC archetypes (A, A2, N, P, and Y) is sufficient to develop better treatment strategies targeting plasticity of these cells within a phenotypic landscape. Further experimental work is necessary to define the quantitative relationship between task performance and fitness in SCLC and may help to explain why "escaping" the polytope defined by SCLC archetypes seems to be the exception more than the rule.

5.1.4. Phenotypic landscapes and archetype analysis are complementary approaches to understanding cell identity

Several different approaches to understanding phenotypic heterogeneity are used in this work. In chapter 3, we introduced the idea of Pareto task inference. Because cells make a limited amount of biological material, like proteins, they can only optimize a limited number of functional programs. Therefore, cells face tradeoffs between tasks, and multi-objective evolution gives a theoretical underpinning for the existence of multiple stable phenotypes in cancer. While an epigenetic landscape and the underlying GRN dynamics describe the mechanism by which cells change their phenotype and explain the stability of various cell states, it does not on its own provide an evolutionary understanding for *why* cells may occupy different states. Archetype analysis, on the other hand, provides this rationale, explaining that multiple stable subtypes exist, as defined by GRN dynamics, to optimize the fitness of the tumor. Therefore, the archetype analysis is complementary to the epigenetic landscape paradigm: while Pareto theory explains why multiple stable subtypes exist, epigenetic landscapes explain how they exist.

Furthermore, the theory behind trading off between distinct gene programs is consistent with many qualities of gene regulatory networks. As described, a network of interacting TFs dictates the complex relationship between different gene programs regulated by TFs that a cell can turn on or off to change its phenotype. Often, these networks are highly *modular* to accomplish the trade-off, i.e., subgroups of the network are more highly connected within the subgroup than between, and subgroups may be mutually inhibitory (Galvão et al., 2010; Kashtan and Alon, 2005; Newman, 2006; Wagner et al., 2007). Essentially, many networks contain expanded "bowtie motifs" (the network structure described in the Introduction), where each node represents a subgroup of coregulatory TFs specific to some cell fate rather than a single TF (Wang et al., 2011). This allows for cells to easily turn on and off full gene programs to change phenotype identity efficiently in response to environmental perturbations and paracrine signaling from other cells and therefore may be more adaptable in new environments (Clune et al., 2013; Friedlander et al., 2013; Kashtan and Alon, 2005). While it is not yet clear how the TF network structure defining SCLC cell identity may be related to trade-offs between archetypal tasks, further computational analysis should be able to clarify this important relationship. In other systems such as E. coli, for example, evaluation of the evolutionary constraints of GRNs using Pareto optimality showed that populations could expand along the Pareto front by fine-tuning relationships between TFs in the network and could expand beyond the Pareto front by changing the network structure (Kogenaru et al., 2020). While SCLC tumors could be considered much more complex than E. coli, a similar approach may be possible to investigate the evolutionary constraints of the SCLC GRN and define the relationships between response to environmental signals, network dynamics, task performance of states in the epigenetic landscape, and evolutionary fitness (**Figure 5.3**). A quantitative metric

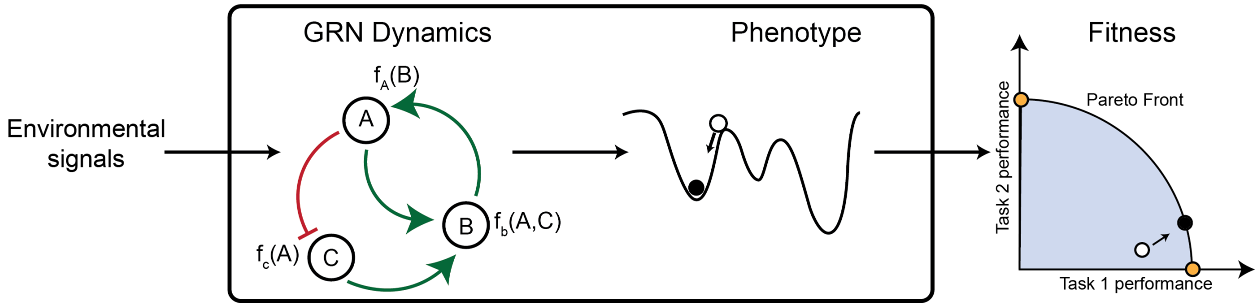


Figure 5.3: Relationship between gene regulatory network dynamics, epigenetic landscape potential, and multi-objective evolutionary fitness. *The GRN defining cell identity, which may be determined using network inference algorithms as in chapters 2 and 4, responds to environmental signals and intrinsic stochasticity to determine the phenotypic location of a cell in the epigenetic landscape. While the GRN dynamics and epigenetic landscape must be inferred indirectly using computational modeling, environmental signals and fitness of the cell state can be observed directly using experimental perturbations. Movement through the epigenetic landscape corresponds to changes in task performance, which determines evolutionary fitness. Over time, cells will move towards the more fit states along the Pareto front. It remains to be seen how locations in the SCLC epigenetic landscape (and states of the GRN) map to locations in performance and fitness space. This is dependent on the measure of fitness, and cell survival in response to the standard of care therapies may be a relevant quantitative metric for fitness in SCLC.*

of fitness, such as cell proliferation in response to drug (i.e. drug-induced proliferation rate, DIP rate) is needed to fully understand the relationship between GRN modeling and archetype analysis.

5.1.5. Integration of top-down and bottom-up modeling

GRN modeling (in chapters 2 and 4), archetype analysis, and the empirical plasticity quantification (in chapters 3 and 4) represent distinct modeling paradigms along the continuum of top-down to bottom-up modeling (**Figure 5.4**). In top-down modeling, analytical approaches are used find empirical biological patterns in high-dimensional data, such as sequencing data (Bruggeman and Westerhoff, 2007; Oulas et al., 2017). The machine learning methods we use in chapters 2 (hierarchical clustering) and 4 (NNMF) to cluster SCLC cell lines and tumors are an unsupervised, top-down approaches to characterize the heterogeneity in SCLC. The goal of these methods is to integrate and analyze experimental data to characterize phenomenological dynamics or patterns.

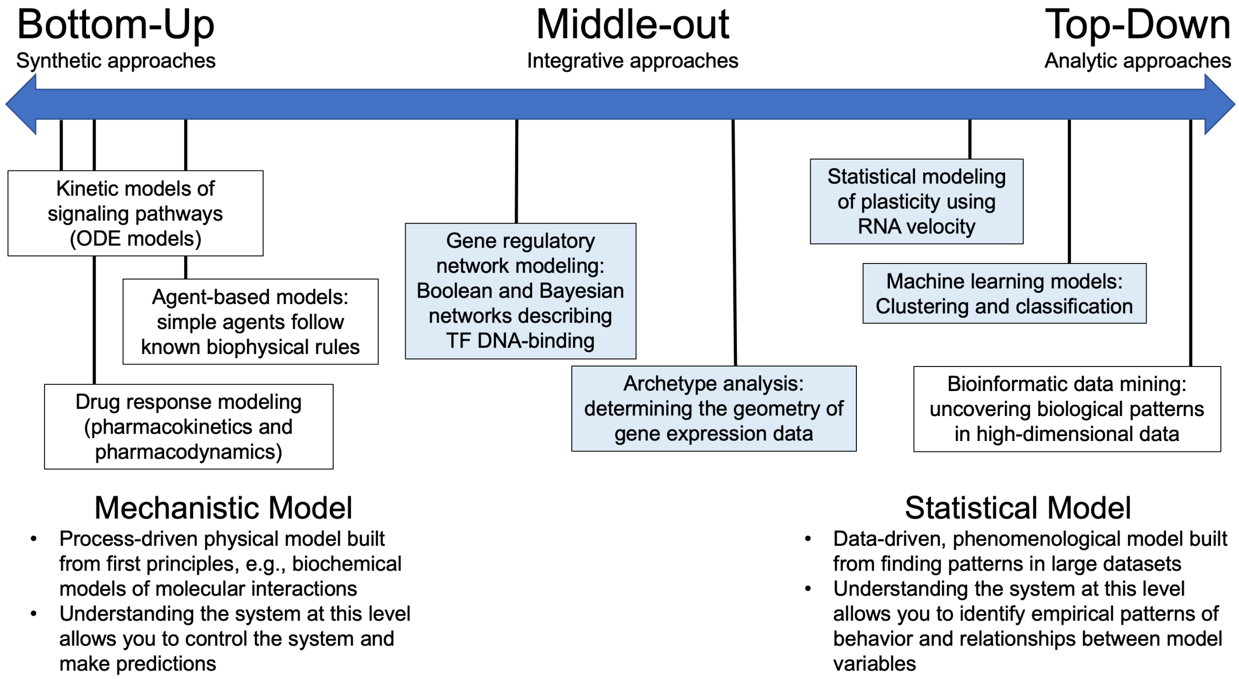

Bottom-Up
Synthetic approaches
Middle-out
Integrative approaches
Top-Down
Analytic approaches
Kinetic models of signaling pathways (ODE models)
Agent-based models: simple agents follow known biophysical rules
Drug response modeling (pharmacokinetics and pharmacodynamics)
Gene regulatory network modeling: Boolean and Bayesian networks describing TF DNA-binding
Archetype analysis: determining the geometry of gene expression data
Statistical modeling of plasticity using RNA velocity
Machine learning models: Clustering and classification
Bioinformatic data mining: uncovering biological patterns in high-dimensional data
Mechanistic Model
Process-driven physical model built from first principles, e.g., biochemical models of molecular interactions
Understanding the system at this level allows you to control the system and make predictions
Statistical Model
Data-driven, phenomenological model built from finding patterns in large datasets
Understanding the system at this level allows you to identify empirical patterns of behavior and relationships between model variables

Figure 5.4: Types of modeling in systems biology along the axis of bottom-up and top-down approaches.

The statistical modeling in chapter 3, where single cell dynamics are approximated as a Markov chain model, is another top-down approach used in this work. While RNA velocity itself may be considered a bottom-up approach due to the underlying ODE model of splicing dynamics, imputation of single cell dynamics uses an RNA velocity-based Markov model that is dependent on patterns in the entire dataset (such as the geometry of the data in transcriptomic space, used to infer Markov state transition probabilities) (Bergen et al., 2020; La Manno et al., 2018; Teschendorff and Feinberg, 2021).

Bottom-up modeling formulates interactions between parts as mechanistic equations that can then be used to predict the emergent behavior of a system. The approach to GRN networking presented here uses a pseudo-Boolean approximation of TF DNA binding. Interactions in this network imply a biochemical interaction between a TF and the promoter region of a target gene, which could be modeled by Hill kinetics using differential equations. While not quite mechanistic, our current implementation of network modeling can predict future behavior, similar to mechanistic models. Therefore, future work could expand the network model to include dynamics of transcription, translation, and DNA binding, to derive true mechanistic explanations of cell identity dynamics.

Archetype analysis could be considered in between these two modeling extremes. While the method of identifying archetypes by determining the geometry of gene expression data is purely a top-down approach, multi-objective optimality is an underlying evolutionary mechanism for task trade-offs that could be modeled with a bottom-up approach. For example, archetype fitness, as it relates to task performance, could be quantified by drug kinetic models based on first principles of the biochemical effect of drug.

Together, the various models in this work can inform our understanding of how a phenotype is regulated. While network models may be used to make predictions about changing phenotype, analyzing patterns in single cell data can show how cells behave phenomenologically. Therefore, these complementary modeling approaches can uncover the mechanism and empirical patterns of phenotype transitions in SCLC, particularly in response to treatment.

5.2. Future Directions

5.2.1. Network structure inference using epigenetic information

In Chapter 2, we develop a novel method for gene regulatory network (GRN) inference that utilizes ChIP-seq data to identify connections between transcription factors and the genes they regulate. However, ChIP-seq-based network construction requires mining databases of TFs assayed in various cell contexts that may not be relevant to the cell type or process of interest. Therefore, one possible improvement on network structure inference in the BooleaBayes algorithm is to use a system-specific chromatin accessibility assay known as Assay of Transposase-Accessible Chromatin with Sequencing (ATAC-seq) (Buenrostro et al., 2013). This method could be applied to SCLC cell line and tumor samples to curate information about the chromatin landscape of SCLC specifically. DNA footprinting, a method that detects regions of chromatin that are protected due to direct occupation by DNA-binding proteins, can be applied to ATAC-seq data to infer TF-TF relationships. Several packages, such as Wellington (Piper et al., 2013), are available to determine the locations of such chromatin footprints. This method would allow for the identification of transcriptional regulator binding sites specific to phenotypic identity in SCLC, which could then be used to prune connections that are spurious or unrelated to SCLC phenotypic identity from our preliminary network structure.

5.2.2. Single cell network inference

One future direction for this work is to apply our network inference pipeline to single cell data. As described in a review by Pratapa et al. (2020), many network inference methods built for bulk transcriptomics are not suitable for single cell data. In contrast to bulk data, single cell data is often plagued by dropouts, resulting in spurious zero counts of network genes that may influence dynamics inference. Single cell data can also be dominated by noise, whereas the averaging effect of bulk transcriptomics data helps to greatly reduce the variance attributed to stochastic processes such as mRNA transcriptional bursting dynamics (Golding et al., 2005; Hsu and Moses, 2022). Pratapa et al. (2020), suggests using the BEELINE framework for benchmarking an algorithm, which could be useful in determining the stability of predictions from single-cell BooleaBayes. Further experimental work is needed to validate our predictions of phenotype regulators. For example, the CRISPR-CAS9 system could be used to turn off (or on) regulators of each attractor in SCLC cell lines with different phenotypes, and sequencing before and after the perturbation at multiple timepoints could corroborate predictions for transitions to other states.

In order to improve the rule inference in our single cell datasets, we took precautionary steps and tested each method using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. As explained in Chapter 4, this gives an evaluation between 0 and 1 for how well the inference tool accurately labels transcription factors as ON or OFF according to its parent nodes. If the rule were completely random (i.e., no significant inference was made), the rule would have an AUC of 0.5, suggesting that the false positive rate and true positive rate was equal on average. By comparing this metric for various methods, we chose the best preprocessing steps for accurate prediction of transcription factor activity. For example, we used MAGIC imputation (Dijk et al., 2018), which can recover signal from noisy or sparse single-cell data using data diffusion. In a test between this imputation method and others, we found that MAGIC gave

AUC values closest to 1 for a majority of the inferred rules. Similarly, other preprocessing steps to help recover signal from the single cell data may improve the robustness of network rule inference. One alternative method for inferring transcription factor activity in single cell data would be to identify cells with expression of that transcription factor's target gene set (its regulon) using AUCell (Sande et al., 2020). As suggested by Pratapa et al. (2020), single cell network inference tools tend to perform better when utilizing larger gene sets. While a single transcription factor may be dropped out (and therefore appear to be turned OFF in network inference), presence of that transcription factor, even in small quantities, should activate expression of its regulon which, due to the size of the regulated gene set, would be less likely to be dropped out completely. These suggested updates to the algorithm may be able to refine rule inference for single cell datasets to make predictions of master regulators and destabilizers more accurate. Together, improving single-cell network inference for SCLC datasets would provide a context-specific understanding of regulatory mechanisms underlying phenotypic heterogeneity and plasticity.

5.2.3. Integrating stability of single cell-derived networks and plasticity quantification

There is a clear relationship between GRN dynamics, which determine cell phenotype, and epigenetic landscape potential, as described in the Introduction and chapter 2. However, classical modeling of epigenetic landscapes relating Hill kinetics to the gradient of the potential landscape is only realistic for relatively small networks. Our implementation of BooleaBayes does not provide the same quantitative measure of potential, but instead calculates the stability of pseudo-attractors based on the time to leave basins of various sizes. Theoretically, this should be inversely proportional to potential: the lower the potential, the higher our stability metric.

In chapter 3, we take an alternative, top-down approach to quantifying single cell dynamics based on probabilistic modelling as a Markovian process, which we use to approximate the drift potential and multipotency of cells (Teschendorff and Feinberg, 2021). One relationship between the two modeling approaches is the identification of relatively stable states; in network inference, these are pseudo-attractors, and in single-cell Markov chain dynamics, these are known as absorbing states. Further work is needed to understand how this empirical measurement of plasticity in single cells relates to the stability metric derived from network inference. We expect that there should be some mapping between the two such that the states with the most stability based on network inference should be absorbing states in the Markov chain.

To do so, it will be critical to define what is meant by cell state in each modeling approach. For example, the cell states in the probabilistic modeling of single-cell dynamics are sampled single cells, i.e., each single cell is its own state. When discussing classical dynamical systems from GRNs, a state is generally considered to be a broader region in the landscape, i.e., the entire basin of attraction around an attractor may be considered a single state. Mapping definitions of cell state between the modeling approaches will be necessary to determine the relationship between our metrics of RNA velocity-inferred plasticity and network-inferred instability.

5.2.4. Targeting epigenetic heterogeneity and plasticity to overcome acquired resistance to chemotherapy in SCLC

While the standard of care therapy for SCLC has not changed in half a century, new targeted therapies are currently being developed and undergoing clinical trials (Coles et al., 2020; Gardner et al., 2017; Horn et al., 2018; Iams et al., 2020; Jia et al., 2018; Rudin et al., 2021; Saunders et al., 2015; Sen et al., 2017, 2018; Taniguchi et al., 2020). Unfortunately, the success of these therapies is often limited by mechanisms of acquired resistance. Non-genetic plasticity has emerged as a major cause of acquired resistance in several cancer types (Chan et al., 2021; Hanahan, 2022; Marjanovic et al., 2013, 2020; Mu et al., 2017; Pisco and Huang, 2015; Qin et al., 2020; Quintanal-Villalonga et al., 2020; Su et al., 2017; Zou et al., 2017). Targeting plasticity directly has been suggested as a possible treatment option for several of these cancers, including melanoma, breast cancer, and prostate cancer (Ahmed and Haass, 2018; Arozarena and Wellbrock, 2019; Boumahdi and Sauvage, 2020; Chapman et al., 2019; Kemper et al., 2014; Risom et al., 2018; Sáez-Ayala et al., 2013; Yabo et al., 2021). As discussed in this dissertation, phenotypic plasticity also seems to play a central role in the progression and acquired resistance of SCLC. Therefore, targeting plasticity may be a realistic future option for combating SCLC relapse.

A few different methods for targeting plasticity can be envisioned. First, cell plasticity could be used advantageously to reprogram cells towards more drug-sensitive states (Yuan et al., 2019). For example, we propose several master regulators and destabilizers in chapter 2, which could be used to direct phenotype switching to attractors that better respond to treatment, which has been shown to be an effective strategy in melanoma (Sáez-Ayala et al., 2013). Further experimental validation is necessary to confirm the role of these predicted TFs in determining SCLC cell identity.

Second, preventing phenotype switching may be more desirable (Boumahdi and Sauvage, 2020). Phenotypic plasticity is intrinsic to the epigenetic landscape: GRN dynamics that shape the landscape form transition paths and unused attractors, and cells may enter transition paths between stable attractors due to extrinsic perturbations or intrinsic stochasticity (Huang, 2013). The barrier to exit attractors may be lower in cancer than normal cells, with “de-canalized,” shallow valleys and attractor basins, enabling cancer cells to stochastically sample the landscape and find new attractors that evade treatment (Jia et al., 2017). Targeting the mechanisms that allow for this stochastic search of drug-tolerant states in the landscape may lower plasticity and acquired resistance to therapy.

For example, chromatin remodeling may be a key mechanism by which cells reprogram to other fates, and therefore promoting repressive chromatin organization may be able to keep cells from transitioning. Several epigenetic modifiers that can control transcription of various gene expression programs may be possible therapeutic targets in SCLC (Poirier et al., 2020). It also may be possible to directly target molecular pathways that reactivate developmental programs in cancer, such as the Wnt and Notch signaling pathways discussed in chapter 3 (associated with the SCLC-Y archetypal task of transdifferentiation in response to injury) (Qin et al., 2020).

In chapter 3, we explore the role of MYC in driving plasticity of a variant SCLC mouse model, which developed tumors much faster than classic SCLC mouse models. We hypothesize that targeting MYC may be capable of reducing the plasticity of NE cells to prevent the switch to a non-NE phenotype. As the non-NE phenotype has been shown to be more mesenchymal and possibly correlated with poor prognosis (McColl et al., 2015; Song et al., 2020), preventing this transition could have huge effects clinically.

In chapter 4, loss of ASCL1 reprograms NE cells into a mesenchymal stem cell-like state. This single molecular perturbation allows cells to transition from an NE attractor in the landscape to a state replicating a multipotent progenitor cell type with higher “potential.” Our work shows that, even in the absence of direct ASCL1 loss, MYC-driven SCLC cells may be capable of reaching this attractor. Therefore, therapeutic strategies to prevent this transition could be essential to “re-canalization” of the landscape and stabilization of attractors. In this chapter, we also find that a small non-NE subpopulation that arises in some tumors after treatment relapse may be capable of regenerating drug-resistant NE cells. While future work is needed to understand the plasticity of this population, one hypothesis explaining relapse in these tumors is that the small non-NE population has a slow-cycling drug-tolerant persister phenotype that can reprogram to a re-proliferative, drug-resistant state, much like the drug-tolerant persisters in several other cancer types (Jolly et al., 2018; Liau et al., 2017; Paudel et al., 2018; Rehman et al., 2021; Risom et al., 2018; Sharma et al., 2010).

Together, the modeling approaches used in this work increase our understanding of cell identity and plasticity, particularly in cancer. Development of strategies that target plasticity and systematically reprogram cell identity may finally be able to overcome the recalcitrance of SCLC.

REFERENCES

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J.T., and Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biology *20*.

Achilleos, A., and Trainor, P.A. (2012). Neural crest stem cells: discovery, properties and potential for therapy. Cell Res *22*, 288–304.

Agaimy, A., Erlenbach-Wünsch, K., Konukiewitz, B., Schmitt, A.M., Rieker, R.J., Vieth, M., Kiesewetter, F., Hartmann, A., Zamboni, G., Perren, A., et al. (2013). ISL1 expression is not restricted to pancreatic well-differentiated neuroendocrine neoplasms, but is also commonly found in well and poorly differentiated neuroendocrine neoplasms of extrapancreatic origin. Modern Pathol *26*, 995–1003.

Ahmed, F., and Haass, N.K. (2018). Microenvironment-Driven Dynamic Heterogeneity and Phenotypic Plasticity as a Mechanism of Melanoma Therapy Resistance. Frontiers Oncol *8*, 173.

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.-C., Geurts, P., Aerts, J., et al. (2017). SCENIC: single-cell regulatory network inference and clustering. Nat Methods *14*, 1083.

Albert, I., Thakar, J., Li, S., Zhang, R., and Albert, R. (2008). Boolean network simulations for life scientists. Source Code Biology Medicine *3*, 16.

Alexandrov, L.B., Ju, Y.S., Haase, K., Loo, P.V., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., et al. (2016). Mutational signatures associated with tobacco smoking in human cancer. Science *354*, 618–622.

Alon, U. (2007). Network motifs: theory and experimental approaches. Nat Rev Genet *8*, 450–461.

American Cancer Society. Cancer Facts & Figures 2021. Atlanta: American Cancer Society; 2021.

Antebi, Y.E., Reich-Zeliger, S., Hart, Y., Mayo, A., Eizenberg, I., Rimer, J., Putheti, P., Pe’er, D., and Friedman, N. (2013). Mapping Differentiation under Mixed Culture Conditions Reveals a Tunable Continuum of T Cell Fates. Plos Biol *11*, e1001616.

Arozarena, I., and Wellbrock, C. (2019). Phenotype plasticity as enabler of melanoma progression and therapy resistance. Nat Rev Cancer 377–391.

Ayers, M., Lunceford, J., Nebozhyn, M., Murphy, E., Loboda, A., Kaufman, D.R., Albright, A., Cheng, J.D., Kang, S.P., Shankaran, V., et al. (2017). IFN-γ–related mRNA profile predicts clinical response to PD-1 blockade. J Clin Invest *127*, 2930–2940.

Baine, M.K., Hsieh, M.-S., Lai, W.V., Egger, J.V., Jungbluth, A.A., Daneshbod, Y., Beras, A., Spencer, R., Lopardo, J., Bodd, F., et al. (2020). SCLC Subtypes Defined by ASCL1, NEUROD1, POU2F3, and YAP1: A Comprehensive Immunohistochemical and Histopathologic Characterization. J Thorac Oncol *15*, 1823–1835.

Balanis, N.G., Sheu, K.M., Esedebe, F.N., Patel, S.J., Smith, B.A., Park, J.W., Alhani, S., Gomperts, B.N., Huang, J., Witte, O.N., et al. (2019). Pan-cancer Convergence to a Small-Cell Neuroendocrine Phenotype that Shares Susceptibilities with Hematological Malignancies. Cancer Cell *36*, 17—34.

Baron, M., Veres, A., Wolock, S.L., Faust, A.L., Gaujoux, R., Vetere, A., Ryu, J.H., Wagner, B.K., Shen-Orr, S.S., Klein, A.M., et al. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst *3*, 346-360.e4.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Bebber, C.M., Thomas, E.S., Stroh, J., Chen, Z., Androulidaki, A., Schmitt, A., Höhne, M.N., Stüker, L., Alves, C. de P., Khonsari, A., et al. (2021). Ferroptosis response segregates small cell lung cancer (SCLC) neuroendocrine subtypes. Nat Commun *12*, 2048.

Beer, M.A., and Tavazoie, S. (2004). Predicting Gene Expression from Sequence. Cell *117*, 185–198.

Belani, C.P., Dahlberg, S.E., Rudin, C.M., Fleisher, M., Chen, H.X., Takebe, N., Velasco, M.R., Tester, W.J., Sturtz, K., Hann, C.L., et al. (2016). Vismodegib or cixutumumab in combination with standard chemotherapy for patients with extensive-stage small cell lung cancer: A trial of the ECOG-ACRIN Cancer Research Group (E1508). Cancer *122*, 2371–2378.

Bensch, K.G., Corrin, B., Pariente, R., and Spencer, H. (1968). Oat-cell carcinoma of the lung. Its origin and relationship to bronchial carcinoid. Cancer *22*, 1163–1172.

Bepler, G., Rotsch, M., Jaques, G., Haeder, M., Heymanns, J., Hartogh, G., Kiefer, P., and Havemann, K. (1988). Peptides and growth factors in small cell lung cancer: production, binding sites, and growth effects. J Cancer Res Clin *114*, 235–244.

Bergen, V., Lange, M., Peidli, S., Wolf, F.A., and Theis, F.J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol 1408–1414.

Berns, K., and Berns, A. (2017). Awakening of ‘‘Schlafen11’’ to Tackle Chemotherapy Resistance in SCLC. Cancer Cell *31*, 169—171.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J Statistical Mech Theory Exp *2008*, P10008.

Bocci, F., Jolly, M.K., and Onuchic, J.N. (2019). A Biophysical Model Uncovers the Size Distribution of Migrating Cell Clusters across Cancer Types. Cancer Res *79*, 5527–5535.

Borromeo, M.D., Savage, T.K., Kollipara, R.K., Gazdar, A.F., Cobb, M.H., Correspondence, J.E.J., He, M., Augustyn, A., er, Osborne, J.K., et al. (2016). ASCL1 and NEUROD1 Reveal Heterogeneity in Pulmonary Neuroendocrine Tumors and Regulate Distinct Genetic Programs. Cell Reports *16*, 1259-1272.

Bostwick, D.G., and Bensch, K.G. (1985). Gastrin releasing peptide in human neuroendocrine tumours. J Pathology *147*, 237–244.

Boumahdi, S., and Sauvage, F.J. de (2020). The great escape: tumour cell plasticity in resistance to targeted therapy. Nat Rev Drug Discov *19*, 39–56.

Brackston, R.D., Lakatos, E., and Stumpf, M.P.H. (2018). Transition state characteristics during cell differentiation. Plos Comput Biol *14*, e1006405.

Brägelmann, J., Böhm, S., Guthrie, M.R., Mollaoglu, G., Oliver, T.G., and Sos, M.L. (2017). Family matters: how MYC family oncogenes impact small cell lung cancer. Cell Cycle *16*, 1489-1498.

Bramsen, J.B., Rasmussen, M.H., Ongen, H., Mattesen, T.B., Ørntoft, M.-B.W., Árnadóttir, S.S., Sandoval, J., Laguna, T., Vang, S., Øster, B., et al. (2017). Molecular-Subtype-Specific Biomarkers Improve Prediction of Prognosis in Colorectal Cancer. Cell Reports *19*, 1268–1280.

Brock, A., Chang, H., and Huang, S. (2009). Non-genetic heterogeneity — a mutation-independent driving force for the somatic evolution of tumours. Nat Rev Genet *10*, nrg2556.

Bruggeman, F.J., and Westerhoff, H.V. (2007). The nature of systems biology. Trends Microbiol *15*, 45–50.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods *10*, 1213–1218.

Cai, L., Liu, H., Huang, F., Fujimoto, J., Girard, L., Chen, J., Li, Y., Zhang, Y.-A., Deb, D., Stastny, V., et al. (2021). Cell-autonomous immune gene expression is repressed in pulmonary neuroendocrine cells and small cell lung cancer. Commun Biology *4*, 314.

Calbo, J., Montfort, E.V., Proost, N., Drunen, E.V., Beverloo, H.B., Meuwissen, R., and Berns, A. (2011). A Functional Role for Tumor Cell Heterogeneity in a Mouse Model of Small Cell Lung Cancer. Cancer Cell *19*, 244—256.

Carney, D.N., Gazdar, A.F., Bepler, G., Guccion, J.G., Marangos, P.J., Moody, T.W., Zweig, M.H., and Minna, J.D. (1985). Establishment and identification of small cell lung cancer cell lines having classic and variant features. Cancer Res *45*, 2913–2923.

Chan, E.D., Morales, D.V., Welsh, C.H., McDermott, M.T., and Schwarz, M.I. (2012). Calcium Deposition with or without Bone Formation in the Lung. Am J Resp Crit Care *165*, 1654–1669.

Chan, J.M., Quintanal-Villalonga, Á., Gao, V.R., Xie, Y., Allaj, V., Chaudhary, O., Masilionis, I., Egger, J., Chow, A., Walle, T., et al. (2021). Signatures of plasticity, metastasis, and immunosuppression in an atlas of human small cell lung cancer. Cancer Cell.

Chapman, M.P., Risom, T., Aswani, A.J., Langer, E.M., Sears, R.C., and Tomlin, C.J. (2019). Modeling differentiation-state transitions linked to therapeutic escape in triple-negative breast cancer. Plos Comput Biol *15*, e1006840.

Chen, S., and Mar, J.C. (2018). Evaluating methods of inferring gene regulatory networks highlights their lack of performance for single cell gene expression data. Bmc Bioinformatics *19*, 232.

Chen, E.Y., Tan, C.M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G.V., Clark, N.R., and Ma'ayan, A. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. Bmc Bioinformatics *14*, 128.

Chi, J.-Y., Hsiao, Y.-W., Liu, H.-L., Fan, X.-J., Wan, X.-B., Liu, T.-L., Hung, S.-J., Chen, Y.-T., Liang, H.-Y., and Wang, J.-M. (2021). Fibroblast CEBPD/SDF4 axis in response to chemotherapy-induced angiogenesis through CXCR4. Cell Death Discov *7*, 94.

Chow, K., Fessel, J.P., Ihida-Stansbury, K., Schmidt, E.P., Gaskill, C., Alvarez, D., Graham, B., Harrison, D.G., Wagner, D.H., Nozik-Grayck, E., et al. (2013). Dysfunctional Resident Lung Mesenchymal Stem Cells Contribute to Pulmonary Microvascular Remodeling. Pulm Circ *3*, 31–49.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly *6*, 80–92.

Clune, J., Mouret, J.-B., and Lipson, H. (2013). The evolutionary origins of modularity. Proc Royal Soc B Biological Sci *280*, 20122863.

Coles, G.L., Cristea, S., Webber, J.T., Levin, R.S., Moss, S.M., He, A., Sangodkar, J., Hwang, Y.C., Arand, J., Drainas, A.P., et al. (2020). Unbiased Proteomic Profiling Uncovers a Targetable GNAS/PKA/PP2A Axis in Small Cell Lung Cancer Stem Cells. Cancer Cell.

Correia, R.B., Gates, A.J., Wang, X., and Rocha, L.M. (2018). CANA: A Python Package for Quantifying Control and Canalization in Boolean Networks. Front Physiol *9*, 1046.

Cui, M., Augert, A., Rongione, M., Conkrite, K., Parazzoli, S., Nikitin, A.Yu., Ingolia, N., and MacPherson, D. (2014). PTEN Is a Potent Suppressor of Small Cell Lung Cancer.

Denny, S.K., Yang, D., Chuang, C.-H., Brady, J.J., Lim, J.S., Grüner, B.M., Chiou, S.-H., Schep, A.N., Baral, J., Hamard, C., et al. (2016). Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. Cell *166*, 328--342.

Dijk, D. van, Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A.J., Burdziak, C., Moon, K.R., Chaffer, C.L., Pattabiraman, D., et al. (2018). Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. Cell *174*, 716-729.e27.

Drapkin, B.J., George, J., Christensen, C.L., Mino-Kenudson, M., Dries, R., Sundaresan, T., Phat, S., Myers, D.T., Zhong, J., Igo, P., et al. (2018). Genomic and functional fidelity of small cell lung cancer patient-derived xenografts. Cancer Discov *8*, CD-17-0935.

Duren, Z., Chen, X., Jiang, R., Wang, Y., and Wong, W.H. (2017). Modeling gene regulation from paired expression and chromatin accessibility data. Proc National Acad Sci *114*, E4914--E4923.

Eizenberg-Magar, I., Rimer, J., Zaretsky, I., Lara-Astiaso, D., Reich-Zeliger, S., and Friedman, N. (2017). Diverse continuum of CD4+ T-cell states is determined by hierarchical additive integration of cytokine signals. Proc National Acad Sci *114*, E6447–E6456.

Enes, S.R., Sjöland, A.A., Skog, I., Hansson, L., Larsson, H., Blanc, K.L., Eriksson, L., Bjermer, L., Scheding, S., and Westergren-Thorsson, G. (2016). MSC from fetal and adult lungs possess lung-specific properties compared to bone marrow-derived MSC. Sci Rep-Uk *6*, 29160.

Feinberg, A.P., and Irizarry, R.A. (2010). Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. Proc National Acad Sci *107*, 1757–1764.

Ferone, G., Song, J.-Y., Krijgsman, O., Vliet, J. van der, Cozijnsen, M., Semenova, E.A., Adams, D.J., Peeper, D., and Berns, A. (2020). FGFR1 Oncogenic Activation Reveals an Alternative Cell of Origin of SCLC in Rb1/p53 Mice. Cell Reports *30,* 3837—3850.

Flavahan, W.A., Gaskell, E., and Bernstein, B.E. (2017). Epigenetic plasticity and the hallmarks of cancer. Science *357*, eaal2380.

Font-Clos, F., Zapperi, S., and Porta, C.A.M.L. (2018). Topography of epithelial–mesenchymal plasticity. Proc National Acad Sci *115*, 5902—5907.

Friedlander, T., Mayo, A.E., Tlusty, T., and Alon, U. (2013). Mutation Rules and the Evolution of Sparseness and Modularity in Biological Systems. Plos One *8*, e70444.

Gadgeel, S.M. (2018). Targeted Therapy and Immune Therapy for Small Cell Lung Cancer. Curr Treat Option On *19*, 53.

Galvão, V., Miranda, J.G.V., Andrade, R.F.S., Andrade, J.S., Gallos, L.K., and Makse, H.A. (2010). Modularity map of the network of human cell differentiation. Proc National Acad Sci *107*, 5750–5755.

Gao, Y., Geng, J., Hong, X., Qi, J., Teng, Y., Yang, Y., Qu, D., and Chen, G. (2014). Expression of p300 and CBP is associated with poor prognosis in small cell lung cancer. Int J Clin Exp Patho *7*, 760–767.

Gardner, E.E., Lok, B.H., Schneeberger, V.E., Desmeules, P., Miles, L.A., Arnold, P.K., Ni, A., Khodos, I., Stanchina, E. de, Nguyen, T., et al. (2017). Chemosensitive Relapse in Small Cell Lung Cancer Proceeds through an EZH2-SLFN11 Axis. Cancer Cell *31*, 286-299.

Gately, K., Collins, I., Forde, L., Al-Alao, B., Young, V., Gerg, M., Feuerhake, F., and O'Byrne, K. (2011). A Role for IGF-1R–Targeted Therapies in Small-Cell Lung Cancer? Clin Lung Cancer *12*, 38–42.

Gay, C.M., Stewart, C.A., Park, E.M., Diao, L., Groves, S.M., Heeke, S., Nabet, B.Y., Fujimoto, J., Solis, L.M., Lu, W., et al. (2021). Patterns of transcription factor programs and immune pathway activation define four major subtypes of SCLC with distinct therapeutic vulnerabilities. Cancer Cell *39,* 1-15.

Gazdar, A.F., Carney, D.N., Nau, M.M., and Minna, J.D. (1985). Characterization of variant subclasses of cell lines derived from small cell lung cancer having distinctive biochemical, morphological, and growth properties. Cancer Res *45*, 2924—2930.

Gazdar, A.F., Bunn, P.A., and Minna, J.D. (2017). Small-cell lung cancer: what we know, what we need to know and the path forward. Nat Rev Cancer *17*, 725.

George, J., Lim, J.S., Jang, S.J., Cun, Y., Ozretić, L., Kong, G., Leenders, F., Lu, X., Fernández-Cuesta, L., Bosco, G., et al. (2015). Comprehensive genomic profiles of small cell lung cancer. Nature *524*, 47.

Glass, L., and Kauffman, S.A. (1973). The logical analysis of continuous, non-linear biochemical control networks. J Theor Biol *39*, 103–129.

Gola, M., Doga, M., Bonadonna, S., Mazziotti, G., Vescovi, P.P., and Giustina, A. (2006). Neuroendocrine tumors secreting growth hormone-releasing hormone: Pathophysiological and clinical aspects. Pituitary *9*, 221–229.

Golding, I., Paulsson, J., Zawilski, S.M., and Cox, E.C. (2005). Real-Time Kinetics of Gene Activity in Individual Bacteria. Cell *123*, 1025–1036.

Griffiths, J.A., Scialdone, A., and Marioni, J.C. (2018). Using single-cell genomics to understand developmental processes and cell fate decisions. Mol Syst Biol *14*, e8046.

Groves, S.M., Ireland, A., Liu, Q., Simmons, A.J., Lau, K., Iams, W.T., Tyson, D., Lovly, C.M., Oliver, T.G., and Quaranta, V. (2021). Cancer Hallmarks Define a Continuum of Plastic Cell States between Small Cell Lung Cancer Archetypes. BioRxiv.

Guo, M., Bao, E.L., Wagner, M., Whitsett, J.A., and Xu, Y. (2016). SLICE: determining cell differentiation and lineage based on single cell entropy. Nucleic Acids Res *45*, gkw1278.

Gupta, P.B., Fillmore, C.M., Jiang, G., Shapira, S.D., Tao, K., Kuperwasser, C., and Lander, E.S. (2011). Stochastic State Transitions Give Rise to Phenotypic Equilibrium in Populations of Cancer Cells. Cell *147*, 1197.

Hagberg, A.A., Schult, D.A., and Swart, P.J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. Proceedings of the 7th Python in Science Conference (SciPy 2008).

Haghverdi, L., Büttner, M., Wolf, F.A., Buettner, F., and Theis, F.J. (2016). Diffusion pseudotime robustly reconstructs lineage branching. Nat Methods *13*, nmeth.3971.

Hanahan, D. (2022). Hallmarks of Cancer: New Dimensions. Cancer Discov *12*, 31–46.

Hanahan, D., and Weinberg, R.A. (2000). The Hallmarks of Cancer. Cell *100*, 57–70.

Harris, L.A., Beik, S., Ozawa, P.M.M., Jimenez, L., and Weaver, A.M. (2019). Modeling heterogeneous tumor growth dynamics and cell-cell interactions at single-cell and cell-population resolution. Curr Opin Syst Biology *17*, 24–34.

Hauschild, A., Grob, J.-J., Demidov, L.V., Jouary, T., Gutzmer, R., Millward, M., Rutkowski, P., Blank, C.U., Miller, W.H., Kaempgen, E., et al. (2012). Dabrafenib in BRAF-mutated metastatic melanoma: a multicentre, open-label, phase 3 randomised controlled trial. Lancet *380*, 358–365.

Hausser, J., Szekely, P., Bar, N., Zimmer, A., Sheftel, H., Caldas, C., and Alon, U. (2019). Tumor diversity and the trade-off between universal cancer tasks. Nature Communications *10*, 5423.

Hayford, C.E., Tyson, D.R., Robbins, C.J., Frick, P.L., Quaranta, V., and Harris, L.A. (2021). An in vitro model of tumor heterogeneity resolves genetic, epigenetic, and stochastic sources of cell state variability. Plos Biol *19*, e3000797.

Hellmann, M.D., Ott, P.A., Zugazagoitia, J., Ready, N.E., Hann, C.L., Braud, F.G.D., Antonia, S.J., Ascierto, P.A., Moreno, V., Atmaca, A., et al. (2017). Nivolumab (nivo) ± ipilimumab (ipi) in advanced small-cell lung cancer (SCLC): First report of a randomized expansion cohort from CheckMate 032. J Clin Oncol *35*, 8503–8503.

Herring, C.A., Banerjee, A., McKinley, E.T., Simmons, A.J., Ping, J., Roland, J.T., Franklin, J.L., Liu, Q., Gerdes, M.J., Coffey, R.J., et al. (2018). Unsupervised Trajectory Analysis of Single-Cell RNA-Seq and Imaging Data Reveals Alternative Tuft Cell Origins in the Gut. Cell Syst *6*.

Hill, A.V. (1913). The Combinations of Haemoglobin with Oxygen and with Carbon Monoxide. I. Biochem J *7*, 471–480.

Hinohara, K., and Polyak, K. (2019). Intratumoral Heterogeneity: More Than Just Mutations. Trends Cell Biol.

Hodgkinson, C.L., Morrow, C.J., Li, Y., Metcalf, R.L., Rothwell, D.G., Trapani, F., Polanski, R., Burt, D.J., Simpson, K.L., Morris, K., et al. (2014). Tumorigenicity and genetic profiling of circulating tumor cells in small-cell lung cancer. Nat Med *20*, 897.

Hong, M., Tao, S., Zhang, L., Diao, L.-T., Huang, X., Huang, S., Xie, S.-J., Xiao, Z.-D., and Zhang, H. (2020). RNA sequencing: new technologies and applications in cancer research. J Hematol Oncol *13*, 166.

Hong, T., Watanabe, K., Ta, C.H., Villarreal-Ponce, A., Nie, Q., and Dai, X. (2015). An Ovol2-Zeb1 Mutual Inhibitory Circuit Governs Bidirectional and Multi-step Transition between Epithelial and Mesenchymal States. Plos Comput Biol *11*, e1004569.

Horie, M., Saito, A., Ohshima, M., Suzuki, H.I., and Nagase, T. (2016). YAP and TAZ modulate cell phenotype in a subset of small cell lung cancer. Cancer Sci *107*, 1755—1766.

Horn, L., Mansfield, A.S., Szczęsna, A., Havel, L., Krzakowski, M., Hochmair, M.J., Huemer, F., Losonczy, G., Johnson, M.L., Nishio, M., et al. (2018). First-Line Atezolizumab plus Chemotherapy in Extensive-Stage Small-Cell Lung Cancer. New Engl J Med *379*, 2220—2229.

Hou, X., Gong, R., Zhan, J., Zhou, T., Ma, Y., Zhao, Y., Zhang, Y., Chen, G., Zhang, Z., Ma, S., et al. (2018). p300 promotes proliferation, migration, and invasion via inducing epithelial-mesenchymal transition in non-small cell lung cancer cells. Bmc Cancer *18*, 641.

Hsu, I.S., and Moses, A.M. (2022). Stochastic models for single-cell data: Current challenges and the way forward. Febs J *289*, 647–658.

Huang, S. (2013). Genetic and non-genetic instability in tumor progression: link between the fitness landscape and the epigenetic landscape of cancer cells. Cancer Metast Rev *32*, 423—448.

Huang, S., and Kauffman, S. (2013). How to escape the cancer attractor: Rationale and limitations of multi-target drugs. Semin Cancer Biol *23*, 270–278.

Huang, S., Guo, Y.-P., May, G., and Enver, T. (2007). Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. Developmental Biology *305*, 695–713.

Huang, S., Ernberg, I., and Kauffman, S. (2009). Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective. Semin Cell Dev Biol *20*, 869—876.

Huang, Y.-H., Klingbeil, O., He, X.-Y., Wu, X.S., Arun, G., Lu, B., Somerville, T.D.D., Milazzo, J.P., Wilkinson, J.E., Demerdash, O.E., et al. (2018). POU2F3 is a master regulator of a tuft cell-like variant of small cell lung cancer. Gene Dev *32*, 915–928.

Huch, M., and Rawlins, E.L. (2017). Cancer: Tumours build their niche. Nature *545*, 292.

Hui, L., Shoumei, X., Zhoujing, Z., Kuang, G., Duohong, Z., Jiacai, H., and Yong, Z. (2021). Effects of Androgen Receptor Overexpression on Chondrogenic Ability of Rabbit Articular Chondrocytes. Tissue Eng Regen Med *18*, 641–650.

Iams, W.T., Porter, J., and Horn, L. (2020). Immunotherapeutic approaches for small-cell lung cancer. Nat Rev Clin Oncol *17*, 300–312.

Ireland, A.S., Micinski, A.M., Kastner, D.W., Guo, B., Wait, S.J., Spainhower, K.B., Conley, C.C., Chen, O.S., Guthrie, M.R., Soltero, D., et al. (2020). MYC Drives Temporal Evolution of Small Cell Lung Cancer Subtypes by Reprogramming Neuroendocrine Fate. Cancer Cell *38*.

Istrail, S., and Davidson, E.H. (2005). Logic functions of the genomic cis-regulatory code. P Natl Acad Sci Usa *102*, 4954–4959.

Ito, T., Udaka, N., Okudela, K., Yazawa, T., and Kitamura, H. (2003). Mechanisms of neuroendocrine differentiation in pulmonary neuroendocrine cells and small cell carcinoma. Endocr Pathol *14*, 133–139.

Jahchan, N.S., Lim, J.S., Bola, B., and Peifer, M. (2016). Identification and Targeting of Long-Term Tumor- Propagating Cells in Small Cell Lung Cancer. Cell Reports *16*, 644—656.

Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., et al. (2014). Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types. Science *343*, 776–779.

Jia, D., Jolly, M.K., Boareto, M., Parsana, P., Mooney, S.M., Pienta, K.J., Levine, H., and Ben-Jacob, E. (2015). OVOL guides the epithelial-hybrid-mesenchymal transition. Oncotarget *6*, 15436–15448.

Jia, D., Jolly, M.K., Kulkarni, P., and Levine, H. (2017). Phenotypic Plasticity and Cell Fate Decisions in Cancer: Insights from Dynamical Systems Theory. Cancers *9*, 70.

Jia, D., Augert, A., Kim, D.-W., Eastwood, E., Wu, N., Ibrahim, A.H., Kim, K.-B., Dunn, C.T., Pillai, S.P.S., Gazdar, A.F., et al. (2018). Crebbp loss drives small cell lung cancer and increases sensitivity to HDAC inhibition. Cancer Discov *8*.

Jiang, Y., Jahagirdar, B.N., Reinhardt, R.L., Schwartz, R.E., Keene, C.D., Ortiz-Gonzalez, X.R., Reyes, M., Lenvik, T., Lund, T., Blackstad, M., et al. (2002). Pluripotency of mesenchymal stem cells derived from adult marrow. Nature *418*, 41–49.

Jolly, M.K., Kulkarni, P., Weninger, K., Orban, J., and Levine, H. (2018). Phenotypic Plasticity, Bet-Hedging, and Androgen Independence in Prostate Cancer: Role of Non-Genetic Heterogeneity. Frontiers Oncol *8*, 50.

Joo, J.I., Zhou, J.X., Huang, S., and Cho, K.-H. (2018). Determining Relative Dynamic Stability of Cell States Using Boolean Network Model. Sci Rep-Uk *8*, 12077.

Kalir, S., and Alon, U. (2004). Using a Quantitative Blueprint to Reprogram the Dynamics of the Flagella Gene Network. Cell *117*, 713–720.

Kamburov, A., Stelzl, U., Lehrach, H., and Herwig, R. (2013). The ConsensusPathDB interaction database: 2013 update. Nucleic Acids Res *41*, D793–D800.

Kanaji, N., Watanabe, N., Kita, N., Bandoh, S., Tadokoro, A., Ishii, T., Dobashi, H., and Matsunaga, T. (2014). Paraneoplastic syndromes associated with lung cancer Nobuhiro Kanaji, Naoki Watanabe, Nobuyuki Kita, Shuji Bandoh, Akira Tadokoro, Tomoya Ishii, Hiroaki Dobashi, Takuya Matsunaga. World J Clin Oncol *5*, 197.

Karaayvaz, M., Cristea, S., Gillespie, S.M., Patel, A.P., Mylvaganam, R., Luo, C.C., Specht, M.C., Bernstein, B.E., Michor, F., and Ellisen, L.W. (2018). Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. Nat Commun *9*, 3588.

Kashtan, N., and Alon, U. (2005). Spontaneous evolution of modularity and network motifs. P Natl Acad Sci Usa *102*, 13773–13778.

Kauffman, S. (1969a). Homeostasis and Differentiation in Random Genetic Control Networks. Nature *224*, 177–178.

Kauffman, S. (1971). Chapter 5 Gene Regulation Networks: A Theory For Their Global Structure and Behaviors. Curr Top Dev Biol *6*, 145–182.

Kauffman, S.A. (1969b). Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol *22*, 437–467.

Kemper, K., Goeje, P.L. de, Peeper, D.S., and Amerongen, R. van (2014). Phenotype Switching: Tumor Cell Plasticity as a Resistance Mechanism and Target for Therapy. Cancer Res *74*, 5937–5941.

Kim, K.-Y., and Wang, J. (2007). Potential Energy Landscape and Robustness of a Gene Regulatory Network: Toggle Switch. Plos Comput Biol *3*, e60.

Kimmel, J.C., Penland, L., Rubinstein, N.D., Hendrickson, D.G., Kelley, D.R., and Rosenthal, A.Z. (2019). Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging. Genome Res *29*, 2088–2103.

Kitano, H. (2002). Systems Biology: A Brief Overview. Science *295*, 1662–1664.

Kleshchevnikov, V. (2019). vitkl/ParetoTI: Beta release 2.

Ko, J., Winslow, M.M., and Sage, J. (2021). Mechanisms of small cell lung cancer metastasis. Embo Mol Med *13*, e13122.

Kogenaru, M., Nghe, P., Poelwijk, F.J., and Tans, S.J. (2020). Predicting Evolutionary Constraints by Identifying Conflicting Demands in Regulatory Networks. Cell Syst. *10*, 1-9.

Korem, Y., Szekely, P., Hart, Y., Sheftel, H., Hausser, J., Mayo, A., Rothenberg, M.E., Kalisky, T., and Alon, U. (2015). Geometry of the Gene Expression Space of Individual Cells. Plos Comput Biol *11*, e1004224.

Kröger, C., Afeyan, A., Mraz, J., Eaton, E.N., Reinhardt, F., Khodor, Y.L., Thiru, P., Bierie, B., Ye, X., Burge, C.B., et al. (2019). Acquisition of a hybrid E/M state is essential for tumorigenicity of basal breast cancer cells. Proc National Acad Sci *116*, 201812876.

Krohn, A., Ahrens, T., Yalcin, A., Plönes, T., Wehrle, J., Taromi, S., Wollner, S., Follo, M., Brabletz, T., Mani, S.A., et al. (2014). Tumor Cell Heterogeneity in Small Cell Lung Cancer (SCLC): Phenotypical and Functional Differences Associated with Epithelial-Mesenchymal Transition (EMT) and DNA Methylation Changes. Plos One *9*, e100249.

Krushkal, J., Silvers, T., Reinhold, W.C., Sonkin, D., Vural, S., Connelly, J., Varma, S., Meltzer, P.S., Kunkel, M., Rapisarda, A., et al. (2020). Epigenome-wide DNA methylation analysis of small cell lung cancer cell lines suggests potential chemotherapy targets. Clin Epigenetics *12*, 93.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res *44*, W90–W97.

Kwon, M., Proost, N., Song, J.-Y., Sutherland, K.D., Zevenhoven, J., and Berns, A. (2015). Paracrine signaling between tumor subclones of mouse SCLC: a critical role of ETS transcription factor Pea3 in facilitating metastasis. Gene Dev *29*, 1587–1592.

LaFave, L.M., Kartha, V.K., Ma, S., Meli, K., Priore, I.D., Lareau, C., Naranjo, S., Westcott, P.M.K., Duarte, F.M., Sankar, V., et al. (2020). Epigenomic State Transitions Characterize Tumor Progression in Mouse Lung Adenocarcinoma. Cancer Cell *38,* 1-17.

Lane, D.P. (1992). p53, guardian of the genome. Nature *358*, 15–16.

Lang, A.H., Li, H., Collins, J.J., and Mehta, P. (2014). Epigenetic Landscapes Explain Partially Reprogrammed Cells and Identify Key Reprogramming Genes. Plos Comput Biol *10*, e1003734.

Lange, M., Bergen, V., Klein, M., Setty, M., Reuter, B., Bakhti, M., Lickert, H., Ansari, M., Schniering, J., Schiller, H.B., et al. (2022). CellRank for directed single-cell fate mapping. Nat Methods 1–12.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. Bmc Bioinformatics *9*, 559.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214.

Lenhart, R., Kirov, S., Desilva, H., Cao, J., Lei, M., Johnston, K., Peterson, R., Schweizer, L., Purandare, A., Ross-Macdonald, P., et al. (2015). Sensitivity of Small Cell Lung Cancer to BET Inhibition Is Mediated by Regulation of ASCL1 Gene Expression. Mol Cancer Ther *14*, 2167–2174.

Li, C., and Wang, J. (2014). Landscape and flux reveal a new global view and physical quantification of mammalian cell cycle. Proc National Acad Sci *111,* 14130-14135.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760.

Li, L., Song, W., Yan, X., Li, A., Zhang, X., Li, W., Wen, X., Zhou, L., Yu, D., Hu, J.-F., et al. (2017). Friend leukemia virus integration 1 promotes tumorigenesis of small cell lung cancer cells by activating the miR-17-92 pathway. Oncotarget *8*, 41975–41987.

Li, Y., Petrov, D.A., and Sherlock, G. (2019). Single nucleotide mapping of trait space reveals Pareto fronts that constrain adaptation. Nat Ecol Evol *3*, 1539–1551.

Liao, D., Estévez-Salmerón, L., and Tlsty, T.D. (2012). Generalized principles of stochasticity can be used to control dynamic heterogeneity. Phys Biol *9*, 065006.

Liau, B.B., Sievers, C., Donohue, L.K., Gillespie, S.M., Flavahan, W.A., Miller, T.E., Venteicher, A.S., Hebert, C.H., Carey, C.D., Rodig, S.J., et al. (2017). Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. Cell Stem Cell *20*, 233-246.e7.

Lieberman, Y., Rokach, L., and Shay, T. (2018). CaSTLe – Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. Plos One *13*, e0205499.

Lim, J.S., Ibaseta, A., Fischer, M.M., Cancilla, B., O'Young, G., Cristea, S., Luca, V.C., Yang, D., Jahchan, N.S., Hamard, C., et al. (2017). Intratumoural heterogeneity generated by Notch signalling promotes small-cell lung cancer. Nature *545*, 360.

Liu, F., Zhang, S.-W., Guo, W.-F., Wei, Z.-G., and Chen, L. (2016). Inference of Gene Regulatory Network Based on Local Bayesian Networks. Plos Comput Biol *12*, e1005024.

Long, F. (2012). Building strong bones: molecular regulation of the osteoblast lineage. Nature Reviews Molecular Cell Biology *13*, 27–38.

Lu, H., and Jiang, Z. (2017). Advances in antiangiogenic treatment of small-cell lung cancer. Oncotargets Ther *10*, 353–359.

Lu, M., Jolly, M.K., Levine, H., Onuchic, J.N., and Ben-Jacob, E. (2013). MicroRNA-based regulation of epithelial–hybrid–mesenchymal fate determination. Proc National Acad Sci *110*, 18144–18149.

Lubbock, A.L.R., Harris, L.A., Quaranta, V., Tyson, D.R., and Lopez, C.F. (2021). Thunor: visualization and analysis of high-throughput dose–response datasets. Nucleic Acids Res *49*, W633–W640.

Luo, X., Xu, L., Han, B., and Wang, J. (2017). Funneled potential and flux landscapes dictate the stabilities of both the states and the flow: Fission yeast cell cycle. Plos Comput Biol *13*, e1005710.

Maaten, L.V.D., and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research *9*, 2579--2605.

Macaulay, I.C., Svensson, V., Labalette, C., Ferreira, L., Hamey, F., Voet, T., Teichmann, S.A., and Cvejic, A. (2016). Single-Cell RNA-Sequencing Reveals a Continuous Spectrum of Differentiation in Hematopoietic Cells. Cell Reports *14*, 966–977.

Manno, G.L., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., et al. (2018). RNA velocity of single cells. Nature *560*.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R.D., and Califano, A. (2004). ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. Bmc Bioinformatics *7*, 1471—2105.

Marjanovic, N.D., Weinberg, R.A., and Chaffer, C.L. (2013). Cell Plasticity and Heterogeneity in Cancer. Clin Chem *59*, 168–179.

Marjanovic, N.D., Hofree, M., Chan, J.E., Canner, D., Wu, K., Trakala, M., Hartmann, G.G., Smith, O.C., Kim, J.Y., Evans, K.V., et al. (2020). Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. Cancer Cell *38*.

Masoudi-Nejad, A., Bidkhori, G., Ashtiani, S.H., Najafi, A., Bozorgmehr, J.H., and Wang, E. (2015). Cancer systems biology and modeling: Microscopic scale and multiscale approaches. Semin Cancer Biol *30*, 60—69.

McColl, K., Wildey, G., Sakre, N., Lipka, M.B., Behtaj, M., Kresak, A., Chen, Y., Yang, M., Velcheti, V., Fu, P., et al. (2015). Reciprocal expression of INSM1 and YAP1 defines subgroups in small cell lung cancer. Oncotarget *5*, 73745–73756.

McFadden, D.G., Papagiannakopoulos, T., Taylor-Weiner, A., Stewart, C., Carter, S.L., Cibulskis, K., Bhutkar, A., McKenna, A., Dooley, A., Vernon, A., et al. (2014). Genetic and Clonal Dissection of Murine Small Cell Lung Carcinoma Progression by Genome Sequencing. Cell *156*, 1298–1311.

Meadows, D. (2008). Thinking in Systems: A Primer.

Menendez, J.A. (2015). Metabolic control of cancer cell stemness: Lessons from iPS cells. Cell Cycle *14*, 3801–3811.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. Science *298*, 824–827.

Mollaoglu, G., Guthrie, M.R., Böhm, S., Brägelmann, J., Can, I., Ballieu, P.M., Marx, A., George, J., Heinen, C., Chalishazar, M.D., et al. (2017). MYC Drives Progression of Small Cell Lung Cancer to a Variant Neuroendocrine Subtype with Vulnerability to Aurora Kinase Inhibition. Cancer Cell *31*, 270—285.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. Mach Learn *52*, 91–118.

Mu, P., Zhang, Z., Benelli, M., Karthaus, W.R., Hoover, E., Chen, C.-C., Wongvipat, J., Ku, S.-Y., Gao, D., Cao, Z., et al. (2017). SOX2 promotes lineage plasticity and antiandrogen resistance in TP53- and RB1-deficient prostate cancer. Science *355*, 84–88.

Mulas, C., Chaigne, A., Smith, A., and Chalut, K.J. (2021). Cell state transitions: definitions and challenges. Development *148*.

Nadjsombati, M.S., McGinty, J.W., Lyons-Cohen, M.R., Jaffe, J.B., DiPeso, L., Schneider, C., Miller, C.N., Pollack, J.L., Gowda, G.A.N., Fontana, M.F., et al. (2018). Detection of Succinate by Intestinal Tuft Cells Triggers a Type 2 Innate Immune Circuit. Immunity *49*, 33-41.e7.

Nakazawa, K., Kurishima, K., Tamura, T., Kagohashi, K., Ishikawa, H., Satoh, H., and Hizawa, N. (2012). Specific organ metastases and survival in small cell lung cancer. Oncol Lett *4*, 617–620.

Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., et al. (2019). An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. Cell *178*, 835-849.e21.

Newman, M.E.J. (2006). Modularity and community structure in networks. Proc National Acad Sci *103*, 8577–8582.

Newman, A.M., Liu, C.L., Green, M.R., Gentles, A.J., Feng, W., Xu, Y., Hoang, C.D., Diehn, M., and Alizadeh, A.A. (2015). Robust enumeration of cell subsets from tissue expression profiles. Nat Methods *12*, nmeth.3337.

Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., et al. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol *37*, 773–782.

Nikolić, M.Z., Sun, D., and Rawlins, E.L. (2018). Human lung development: recent progress and new challenges. Dev Camb Engl *145*, dev163485.

Olsen, R.R., Ireland, A.S., Kastner, D.W., Groves, S.M., Spainhower, K.B., Pozo, K., Kelenis, D.P., Whitney, C.P., Guthrie, M.R., Wait, S.J., et al. (2021). ASCL1 represses a SOX9+ neural crest stem-like state in small cell lung cancer. Gene Dev *35,* 1—23.

Osborne, J.K., Larsen, J.E., Gonzales, J.X., Shames, D.S., Sato, M., Wistuba, I.I., Girard, L., Minna, J.D., and Cobb, M.H. (2013). NeuroD1 regulation of migration accompanies the differential sensitivity of neuroendocrine carcinomas to TrkB inhibition. Oncogenesis *2*, e63–e63.

Ott, P.A., Elez, E., Hiret, S., Kim, D.-W., Morosky, A., Saraf, S., Piperdi, B., and Mehnert, J.M. (2017). Pembrolizumab in Patients With Extensive-Stage Small-Cell Lung Cancer: Results From the Phase Ib KEYNOTE-028 Study. J Clin Oncol *35*, JCO.2017.72.506.

Ott, P.A., Bang, Y.-J., Piha-Paul, S.A., Razak, A.R.A., Bennouna, J., Soria, J.-C., Rugo, H.S., Cohen, R.B., O'Neil, B.H., Mehnert, J.M., et al. (2019). T-Cell–Inflamed Gene-Expression Profile, Programmed Death Ligand 1 Expression, and Tumor Mutational Burden Predict Efficacy in Patients Treated With Pembrolizumab Across 20 Cancers: KEYNOTE-028. J Clin Oncol *37*, 318–327.

Ouadah, Y., Rojas, E.R., Riordan, D.P., Capostagno, S., Kuo, C.S., and Krasnow, M.A. (2019). Rare Pulmonary Neuroendocrine Cells Are Stem Cells Regulated by Rb, p53, and Notch. Cell *179*, 403-416.e23.

Oulas, A., Minadakis, G., Zachariou, M., Sokratous, K., Bourdakou, M.M., and Spyrou, G.M. (2017). Systems Bioinformatics: increasing precision of computational diagnostics and therapeutics through network-based approaches. Brief Bioinform *20*, 806–824.

Owonikoko, T.K., Dahlberg, S.E., Sica, G., Wagner, L.I., Wade, J.L., Srkalovic, G., Lash, B.W., Leach, J.W., Leal, T.A.B., Aggarwal, C., et al. (2017). Randomized trial of cisplatin and etoposide in combination with veliparib or placebo for extensive stage small cell lung cancer: ECOG-ACRIN 2511 study. J Clin Oncol *35*, 8505–8505.

Pan, J.-X., Xiong, L., Zhao, K., Zeng, P., Wang, B., Tang, F.-L., Sun, D., Guo, H., Yang, X., Cui, S., et al. (2018). YAP promotes osteogenesis and suppresses adipogenic differentiation by regulating β-catenin signaling. Bone Res *6*, 18.

Paraschiv, B., Diaconu, C.C., Toma, C.L., and Bogdan, M.A. (2015). Paraneoplastic syndromes: the way to an early diagnosis of lung cancer. Pneumologia Buchar Romania *64*, 14–19.

Park, J.W., Lee, J.K., Sheu, K.M., Wang, L., Balanis, N.G., Nguyen, K., Smith, B.A., Cheng, C., Tsai, B.L., Cheng, D., et al. (2018). Reprogramming normal human epithelial tissues to a common, lethal neuroendocrine cancer lineage. Science *362*, 91–95.

Park, K.-S., Liang, M.-C., Raiser, D.M., Zamponi, R., Roach, R.R., Curtis, S.J., Walton, Z., Schaffer, B.E., Roake, C.M., Zmoos, A.-F., et al. (2011). Characterization of the cell of origin for small cell lung cancer. Cell Cycle *10*, 2806–2815.

Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., et al. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science *344*, 1396–1401.

Paudel, B.B., Harris, L.A., Hardeman, K.N., Abugable, A.A., Hayford, C.E., Tyson, D.R., and Quaranta, V. (2018). A Nonquiescent "Idling" Population State in Drug-Treated, BRAF-Mutated Melanoma. Biophys J *114*, 1499–1511.

Pearse, A.G.E., and Polak, J.M. (1971). Neural crest origin of the endocrine polypeptide (APUD) cells of the gastrointestinal tract and pancreas. Gut *12*, 783.

Peifer, M., Fernández-Cuesta, L., Sos, M.L., George, J., Seidel, D., Kasper, L.H., Plenker, D., Leenders, F., Sun, R., Zander, T., et al. (2012). Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat Genet *44*, 1104.

Pellin, D., Loperfido, M., Baricordi, C., Wolock, S.L., Montepeloso, A., Weinberg, O.K., Biffi, A., Klein, A.M., and Biasco, L. (2019). A comprehensive single cell transcriptional landscape of human hematopoietic progenitors. Nat Commun *10*, 2395.

Piper, J., Elze, M.C., Cauchy, P., Cockerill, P.N., Bonifer, C., and Ott, S. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Res *41*, e201–e201.

Pisco, A.O., and Huang, S. (2015). Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: 'What does not kill me strengthens me.' Brit J Cancer *112*, 1725—1732.

Plasschaert, L.W., Žilionis, R., Choo-Wing, R., Savova, V., Knehr, J., Roma, G., Klein, A.M., and Jaffe, A.B. (2018). A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. Nature *560*, 377–381.

Poirier, J., Gardner, E., Connis, N., Moreira, A., Stanchina, E.D., Hann, C., and Rudin, C. (2015). DNA methylation in small cell lung cancer defines distinct disease subtypes and correlates with high expression of EZH2. Oncogene *34*, 5869—5878.

Poirier, J.T., George, J., Owonikoko, T.K., Berns, A., Brambilla, E., Byers, L.A., Carbone, D., Chen, H.J., Christensen, C.L., Dive, C., et al. (2020). New approaches to small cell lung cancer therapy: from the laboratory to the clinic. J Thorac Oncol *15*, 520—540.

Polley, E., Kunkel, M., Evans, D., Silvers, T., Delosh, R., Laudeman, J., Ogle, C., Reinhart, R., Selby, M., Connelly, J., et al. (2016). Small Cell Lung Cancer Screen of Oncology Drugs, Investigational Agents, and Gene and microRNA Expression. Jnci J National Cancer Inst *108*, djw122.

Pomerance, A., Ott, E., Girvan, M., and Losert, W. (2009). The effect of network topology on the stability of discrete state models of genetic control. P Natl Acad Sci Usa *106*, 8209–8214.

Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., and Murali, T.M. (2020). Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nat Methods *17*, 147–154.

Qin, S., Jiang, J., Lu, Y., Nice, E.C., Huang, C., Zhang, J., and He, W. (2020). Emerging role of tumor cell plasticity in modifying therapeutic response. Signal Transduct Target Ther *5*, 228.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. Nat Methods *14*, 979–982.

Qiu, X., Zhang, Y., Martin-Rufino, J.D., Weng, C., Hosseinzadeh, S., Yang, D., Pogson, A.N., Hein, M.Y., Min, K.H. (Joseph), Wang, L., et al. (2022). Mapping transcriptomic vector fields of single cells. Cell *185*, 690-711.e45.

Quintanal-Villalonga, Á., Chan, J.M., Yu, H.A., Pe'er, D., Sawyers, C.L., Sen, T., and Rudin, C.M. (2020). Lineage plasticity in cancer: a shared pathway of therapeutic resistance. Nat Rev Clin Oncol 1–12.

Ramirez, R.N., El-Ali, N.C., Mager, M.A., Wyman, D., Conesa, A., and Mortazavi, A. (2017). Dynamic Gene Regulatory Networks of Human Myeloid Differentiation. Cell Syst *4*, 416—429.e3.

Ratié, L., Ware, M., Jagline, H., David, V., and Dupé, V. (2014). Dynamic expression of Notch-dependent neurogenic markers in the chick embryonic nervous system. Front Neuroanat *8*, 158.

Rehman, S.K., Haynes, J., Collignon, E., Brown, K.R., Wang, Y., Nixon, A.M.L., Bruce, J.P., Wintersinger, J.A., Mer, A.S., Lo, E.B.L., et al. (2021). Colorectal Cancer Cells Enter a Diapause-like DTP State to Survive Chemotherapy. Cell *184*, 226-242.e21.

Risom, T., Langer, E.M., Chapman, M.P., Rantala, J., Fields, A.J., Boniface, C., Alvarez, M.J., Kendsersky, N.D., Pelz, C.R., Johnson-Camacho, K., et al. (2018). Differentiation-state plasticity is a targetable resistance mechanism in basal-like breast cancer. Nat Commun *9*, 3815.

Risse-Hackl, G., Adamkiewicz, J., Wimmel, A., and Schuermann, M. (1998). Transition from SCLC to NSCLC phenotype is accompanied by an increased TRE-binding activity and recruitment of specific AP-1 proteins. Oncogene *16*, 3057–3068.

Robert, M., Frenel, J.-S., Gourmelon, C., Patsouris, A., Augereau, P., and Campone, M. (2017). Olaparib for the treatment of breast cancer. Expert Opin Inv Drug *26*, 751–759.

Rosai, J. (2011). The origin of neuroendocrine tumors and the neural crest saga. Modern Pathol *24*, S53–S57.

Rudin, C.M., Durinck, S., Stawiski, E.W., Poirier, J.T., Modrusan, Z., Shames, D.S., Bergbower, E.A., Guan, Y., Shin, J., Guillory, J., et al. (2012). Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet *44*, 1111.

Rudin, C.M., Poirier, J.T., Byers, L.A., Dive, C., Dowlati, A., George, J., Heymach, J.V., Johnson, J.E., Lehman, J.M., MacPherson, D., et al. (2019). Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. Nat Rev Cancer *19*, 289–297.

Rudin, C.M., Brambilla, E., Faivre-Finn, C., and Sage, J. (2021). Small-cell lung cancer. Nat Rev Dis Primers *7*, 3.

Saadatpour, A., and Albert, R. (2013). Boolean modeling of biological regulatory networks: A methodology tutorial. Methods *62*, 3–12.

Saelens, W., Cannoodt, R., Todorov, H., and Saeys, Y. (2018). A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. Biorxiv 276907.

Sáez-Ayala, M., Montenegro, M.F., Sánchez-del-Campo, L., Fernández-Pérez, M.P., Chazarra, S., Freter, R., Middleton, M., Piñero-Madrona, A., Cabezas-Herrera, J., Goding, C.R., et al. (2013). Directed Phenotype Switching as an Effective Antimelanoma Strategy. Cancer Cell *24*, 105–119.

Sanchez-Castillo, M., Blanco, D., Tienda-Luna, I.M., Carrion, M.C., and Huang, Y. (2017). A Bayesian framework for the inference of gene regulatory networks from time and pseudo-time series data. Bioinformatics *34,* 964—970.

Sande, B.V. de, Flerin, C., Davie, K., Waegeneer, M.D., Hulselmans, G., Aibar, S., Seurinck, R., Saelens, W., Cannoodt, R., Rouchon, Q., et al. (2020). A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nat Protoc *15*, 2247–2276.

Sato, T., Kaneda, A., Tsuji, S., Isagawa, T., Yamamoto, S., Fujita, T., Yamanaka, R., Tanaka, Y., Nukiwa, T., Marquez, V.E., et al. (2013). PRC2 overexpression and PRC2-target gene repression relating to poorer prognosis in small cell lung cancer. Sci Rep-Uk *3*, 1911.

Saunders, L.R., Bankovich, A.J., Anderson, W.C., Aujay, M.A., Bheddah, S., Black, K., Desai, R., Escarpe, P.A., Hampl, J., Laysang, A., et al. (2015). A DLL3-targeted antibody-drug conjugate eradicates high-grade pulmonary neuroendocrine tumor-initiating cells in vivo. Sci Transl Med *7*, 302ra136.

Schafer, J.M., Lehmann, B.D., Gonzalez-Ericsson, P.I., Marshall, C.B., Beeler, J.S., Redman, L.N., Jin, H., Sanchez, V., Stubbs, M.C., Scherle, P., et al. (2020). Targeting MYCN-expressing triple-negative breast cancer with BET and MEK inhibitors. Sci Transl Med *12*, eaaw8275.

Schaffer, B.E., Park, K.-S., Yiu, G., Conklin, J.F., Lin, C., Burkhart, D.L., Karnezis, A.N., Sweet-Cordero, E.A., and Sage, J. (2010). Loss of p130 Accelerates Tumor Development in a Mouse Model for Human Small-Cell Lung Carcinoma. Cancer Res *70*, 3877–3883.

Scheijen, B., Bronk, M., Meer, T. van der, and Bernards, R. (2003). Constitutive E2F1 Overexpression Delays Endochondral Bone Formation by Inhibiting Chondrocyte Differentiation. Mol Cell Biol *23*, 3656–3668.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. Cell *176*, 928-943.e22.

Semenova, E.A., Nagel, R., and Berns, A. (2015). Origins, genetic landscape, and emerging therapies of small cell lung cancer. Gene Dev *29*, 1447—1462.

Sen, T., Tong, P., Diao, L., Li, L., Fan, Y., Hoff, J., Heymach, J.V., Wang, J., and Byers, L.A. (2017). Targeting AXL and mTOR Pathway Overcomes Primary and Acquired Resistance to WEE1 Inhibition in Small-Cell Lung Cancer. Clin Cancer Res *23*, 6239–6253.

Sen, T., Gay, C.M., and Byers, L.A. (2018). Targeting DNA damage repair in small cell lung cancer and the biomarker landscape. Transl Lung Cancer Res *7*, 50–68.

Sentek, H., and Klein, D. (2021). Lung-Resident Mesenchymal Stem Cell Fates within Lung Cancer. Cancers *13*, 4637.

Serizawa, T., Isotani, A., Matsumura, T., Nakanishi, K., Nonaka, S., Shibata, S., Ikawa, M., and Okano, H. (2019). Developmental analyses of mouse embryos and adults using a non-overlapping tracing system for all three germ layers. Development *146*, dev174938.

Sharma, S.V., Lee, D.Y., Li, B., Quinlan, M.P., Takahashi, F., Maheswaran, S., McDermott, U., Azizian, N., Zou, L., Fischbach, M.A., et al. (2010). A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations. Cell *141*, 69–80.

Shi, Y., Shu, B., Yang, R., Xu, Y., Xing, B., Liu, J., Chen, L., Qi, S., Liu, X., Wang, P., et al. (2015). Wnt and Notch signaling pathway involved in wound healing by targeting c-Myc and Hes1 separately. Stem Cell Res Ther *6*, 120.

Shimizu, Y., Kinoshita, I., Kikuchi, J., Yamazaki, K., Nishimura, M., Birrer, M.J., and Dosaka-Akita, H. (2008). Growth inhibition of non-small cell lung cancer cells by AP-1 blockade using a cJun dominant-negative mutant. Brit J Cancer *98*, 915–922.

Simbolo, M., Barbi, S., Fassan, M., Mafficini, A., Ali, G., Vicentini, C., Sperandio, N., Corbo, V., Rusev, B., Mastracci, L., et al. (2019). Gene expression profiling of lung atypical carcinoids and large cell neuroendocrine carcinomas identifies three transcriptomic subtypes with specific genomic alterations. J Thor Onc *14*, 1651—1661.

Simpson, K.L., Stoney, R., Frese, K.K., Simms, N., Rowe, W., Pearce, S.P., Humphrey, S., Booth, L., Morgan, D., Dynowski, M., et al. (2020). A biobank of small cell lung cancer CDX models elucidates inter- and intratumoral phenotypic heterogeneity. Nat Cancer *1*, 437–451.

Skoulidis, F., Byers, L.A., Diao, L., Papadimitrakopoulou, V.A., Tong, P., Izzo, J., Behrens, C., Kadara, H., Parra, E.R., Canales, J.R., et al. (2015). Co-occurring Genomic Alterations Define Major Subsets of KRAS-Mutant Lung Adenocarcinoma with Distinct Biology, Immune Profiles, and Therapeutic Vulnerabilities. Cancer Discov *5*, 860–877.

Smith, N., Dong, Y., Lian, J.B., Pratap, J., Kingsley, P.D., Wijnen, A.J. van, Stein, J.L., Schwarz, E.M., O'Keefe, R.J., Stein, G.S., et al. (2005). Overlapping expression of Runx1(Cbfa2) and Runx2(Cbfa1) transcription factors supports cooperative induction of skeletal development. J Cell Physiol *203*, 133–143.

Snel, B. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res *28*, 3442–3444.

Song, H., Yao, E., Lin, C., Gacayan, R., Chen, M.-H., and Chuang, P.-T. (2012). Functional characterization of pulmonary neuroendocrine cells in lung development, injury, and tumorigenesis. Proc National Acad Sci *109*, 17531–17536.

Song, Y., Sun, Y., Lei, Y., Yang, K., and Tang, R. (2020). YAP1 promotes multidrug resistance of small cell lung cancer by CD74-related signaling pathways. Cancer Med *9*, 259–268.

Sos, M.L., Dietlein, F., Peifer, M., Schöttle, J., Balke-Want, H., Müller, C., Koker, M., Richters, A., Heynck, S., Malchers, F., et al. (2012). A framework for identification of actionable cancer genome dependencies in small cell lung cancer. Proc National Acad Sci *109*, 17034–17039.

Stein-O'Brien, G.L., Arora, R., Culhane, A.C., Favorov, A.V., Garmire, L.X., Greene, C.S., Goff, L.A., Li, Y., Ngom, A., Ochs, M.F., et al. (2018). Enter the Matrix: Factorization Uncovers Knowledge from Omics. Trends Genet *34*, 790–805.

Steinway, S.N., Zañudo, J.G.T., Michel, P.J., Feith, D.J., Loughran, T.P., and Albert, R. (2015). Combinatorial interventions inhibit TGFβ-driven epithelial-to-mesenchymal transition and support hybrid cellular phenotypes. Npj Syst Biology Appl *1*, 15014.

Stewart, C.A., Tong, P., Cardnell, R.J., Sen, T., Li, L., Gay, C.M., Masrorpour, F., Fan, Y., Bara, R.O., Feng, Y., et al. (2014). Dynamic variations in epithelial-to-mesenchymal transition (EMT), ATM, and SLFN11 govern response to PARP inhibitors and cisplatin in small cell lung cancer. Oncotarget *5*, 28575—28587.

Stewart, C.A., Gay, C.M., Xi, Y., Sivajothi, S., Sivakamasundari, V., Fujimoto, J., Bolisetty, M., Hartsfield, P.M., Balasubramaniyan, V., Chalishazar, M.D., et al. (2020). Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. Nature Cancer *1*, 423-436.

Stewart, C.A., Gay, C.M., Ramkumar, K., Cargill, K.R., Cardnell, R.J., Nilsson, M.B., Heeke, S., Park, E.M., Kundu, S.T., Diao, L., et al. (2021). Lung cancer models reveal SARS-CoV-2-induced EMT contributes to COVID-19 pathophysiology. Biorxiv 2020.05.28.122291.

Stransky, N., Ghandi, M., Kryukov, G.V., Garraway, L.A., Lehár, J., Liu, M., Sonkin, D., Kauffmann, A., Venkatesan, K., Edelman, E.J., et al. (2015). Pharmacogenomic agreement between two cancer cell line data sets. Nature *528*, 84–87.

Su, Y., Wei, W., Robert, L., Xue, M., Tsoi, J., Garcia-Diaz, A., Moreno, B.H., Kim, J., Ng, R.H., Lee, J.W., et al. (2017). Single-cell analysis resolves the cell state transition and signaling dynamics associated with melanoma drug-induced resistance. Proc National Acad Sci *114*, 13679–13684.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. P Natl Acad Sci Usa *102*, 15545–15550.

Sutherland, D, K., Proost, N., Brouns, I., Adriaensen, D., Song, J.-Y., and Berns, A. (2011). Cell of Origin of Small Cell Lung Cancer: Inactivation of Trp53 and Rb1 in Distinct Cell Types of Adult Mouse Lung. Cancer Cell *19*, 754—764.

Sveiven, S.N., and Nordgren, T.M. (2020). Lung-resident mesenchymal stromal cells are tissue-specific regulators of lung homeostasis. Am J Physiol-Lung C *319*, L197–L210.

Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., et al. (2020). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res *49*, D605–D612.

Tammela, T., Sanchez-Rivera, F.J., Cetinbas, N.M., Wu, K., Joshi, N.S., Helenius, K., Park, Y., Azimi, R., Kerper, N.R., Wesselhoeft, R.A., et al. (2017). A Wnt-producing niche drives proliferative potential and progression in lung adenocarcinoma. Nature *545*, 355.

Taniguchi, H., Sen, T., and Rudin, C.M. (2020). Targeted Therapies and Biomarkers in Small Cell Lung Cancer. Frontiers Oncol *10*, 741.

Taromi, S., Kayser, G., Catusse, J., Elverfeldt, D. von, Reichardt, W., Braun, F., Weber, W.A., Zeiser, R., and Burger, M. (2014). CXCR4 antagonists suppress small cell lung cancer progression. Oncotarget *5*, 85185–85195.

Teschendorff, A.E., and Feinberg, A.P. (2021). Statistical mechanics meets single-cell biology. Nat Rev Genet *22,* 459–476.

Thieffry, D., and Thomas, R. (1998). Qualitative analysis of gene networks. Pac Symposium Biocomput Pac Symposium Biocomput 77–88.

Tlemsani, C., Pongor, L., Elloumi, F., Girard, L., Huffman, K.E., Roper, N., Varma, S., Luna, A., Rajapakse, V.N., Sebastian, R., et al. (2020). SCLC-CellMiner: A Resource for Small Cell Lung Cancer Cell Line Genomics and Pharmacology Based on Genomic Signatures. Cell Reports *33*, 108296.

Traag, V.A., Waltman, L., and Eck, N.J. van (2019). From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep *9*, 5233.

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Res *25*, 1491–1498.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., and Rinn, J.L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol *32*, 381–386.

Travaglini, K.J., Nabhan, A.N., Penland, L., Sinha, R., Gillich, A., Sit, R.V., Chang, S., Conley, S.D., Mori, Y., Seita, J., et al. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. Nature *587*, 619–625.

Travis, W.D., Brambilla, E., Noguchi, M., Nicholson, A.G., Geisinger, K.R., Yatabe, Y., Beer, D.G., Powell, C.A., Riely, G.J., Schil, P.E.V., et al. (2011). International Association for the Study of Lung Cancer/American Thoracic Society/European Respiratory Society International Multidisciplinary Classification of Lung Adenocarcinoma. J Thorac Oncol *6*, 244–285.

Tripathi, S.C., Fahrmann, J.F., Celiktas, M., Aguilar, M., Marini, K.D., Jolly, M.K., Katayama, H., Wang, H., Murage, E.N., Dennison, J.B., et al. (2017). MCAM Mediates Chemoresistance in

Small-Cell Lung Cancer via the PI3K/AKT/SOX2 Signaling Pathway. Cancer Res *77*, 4414–4425.

Udyavar, A.R., Wooten, D.J., Hoeksema, M., Bansal, M., Califano, A., Estrada, L., Schnell, S., Irish, J.M., Massion, P.P., and Quaranta, V. (2017). Novel Hybrid Phenotype Revealed in Small Cell Lung Cancer by a Transcription Factor Network Model That Can Explain Tumor Heterogeneity. Cancer Res *77*, 1063–1074.

Umemura, S., Mimaki, S.M., Makinoshima, H., Tada, S.M., Ishii, G., Ohmatsu, H., Niho, S., Yoh, K., Matsumoto, S., Takahashi, A., et al. (2014). Therapeutic Priority of the PI3K/AKT/mTOR Pathway in Small Cell Lung Cancers as Revealed by a Comprehensive Genomic Analysis. J Thorac Oncol *9*, 1324—1331.

Vanyai, H.K., Prin, F., Guillermin, O., Marzook, B., Boeing, S., Howson, A., Saunders, R.E., Snoeks, T., Howell, M., Mohun, T.J., et al. (2020). Control of skeletal morphogenesis by the Hippo-YAP/TAZ pathway. Development *147*, dev187187.

Velten, L., Haas, S.F., Raffel, S., Blaszkiewicz, S., Islam, S., Hennig, B.P., Hirche, C., Lutz, C., Buss, E.C., Nowak, D., et al. (2017). Human haematopoietic stem cell lineage commitment is a continuous process. Nat Cell Biol *19*, 271–281.

Waddington, C.H. (1957). The Strategy Of The Genes.

Wagner, D.E., and Klein, A.M. (2020). Lineage tracing meets single-cell omics: opportunities and challenges. Nat Rev Genet *21*, 410–427.

Wagner, A., Regev, A., and Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol *34*, nbt.3711.

Wagner, A.H., Devarakonda, S., Skidmore, Z.L., Krysiak, K., Ramu, A., Trani, L., Kunisaki, J., Masood, A., Waqar, S.N., Spies, N.C., et al. (2018). Recurrent WNT pathway alterations are frequent in relapsed small cell lung cancer. Nat Commun *9*, 3787.

Wagner, G.P., Pavlicev, M., and Cheverud, J.M. (2007). The road to modularity. Nat Rev Genet *8*, 921–931.

Wahl, G.M., and Spike, B.T. (2017). Cell state plasticity, stem cells, EMT, and the generation of intra-tumoral heterogeneity. Npj Breast Cancer *3*, 14.

Wajed, S.A., Laird, P.W., and DeMeester, T.R. (2001). DNA Methylation: An Alternative Pathway to Cancer. Ann Surg *234*, 10–20.

Wang, J. (2015). Landscape and flux theory of non-equilibrium dynamical systems with application to biology. Adv Phys *64*, 1–137.

Wang, Y., and Conlon, J.M. (1993). Neuroendocrine peptides (NPY, GRP, VIP, somatostatin) from the brain and stomach of the alligator. Peptides *14*, 573–579.

Wang, J., Huang, B., Xia, X., and Sun, Z. (2006). Funneled Landscape Leads to Robustness of Cell Networks: Yeast Cell Cycle. Plos Comput Biol *2*, e147.

Wang, J., Xu, L., and Wang, E. (2008). Potential landscape and flux framework of nonequilibrium networks: Robustness, dissipation, and coherence of biochemical oscillations. Proc National Acad Sci *105*, 12271–12276.

Wang, J., Xu, L., Wang, E., and Huang, S. (2010a). The Potential Landscape of Genetic Circuits Imposes the Arrow of Time in Stem Cell Differentiation. Biophys J *99*, 29--39.

Wang, J., Li, C., and Wang, E. (2010b). Potential and flux landscapes quantify the stability and robustness of budding yeast cell cycle network. Proc National Acad Sci *107*, 8195–8200.

Wang, J., Zhang, K., Xu, L., and Wang, E. (2011). Quantifying the Waddington landscape and biological paths for development and differentiation. Proc National Acad Sci *108*, 8257–8262.

Wang, L., Babikir, H., Müller, S., Yagnik, G., Shamardani, K., Catalan, F., Kohanbash, G., Alvarado, B., Lullo, E.D., Kriegstein, A., et al. (2019a). The Phenotypes of Proliferating Glioblastoma Cells Reside on a Single Axis of Variation. Cancer Discov *9*.

Wang, L., Zhang, Q., Qin, Q., Trasanidis, N., Vinyard, M., Chen, H., and Pinello, L. (2021). Current progress and potential opportunities to infer single-cell developmental trajectory and cell fate. Curr Opin Syst Biology *26*, 1–11.

Wang, T., Chen, X., Qiao, W., Kong, L., Sun, D., and Li, Z. (2017). Transcription factor E2F1 promotes EMT by regulating ZEB2 in small cell lung cancer. Bmc Cancer *17*, 719.

Wang, X.-D., Hu, R., Ding, Q., Savage, T.K., Huffman, K.E., Williams, N., Cobb, M.H., Minna, J.D., Johnson, J.E., and Yu, Y. (2019b). Subtype-specific secretomic characterization of pulmonary neuroendocrine tumor cells. Nat Commun *10*, 3201.

Weinreb, C., Wolock, S., Tusi, B.K., Socolovsky, M., and Klein, A.M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. Proc National Acad Sci *115*, 201714723.

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., and Klein, A.M. (2020). Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science *367,* eaaw3381.

Welch, J.D., Hartemink, A., er J, and Prins, J.F. (2016). SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. Genome Biol *17*, 106.

Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed February 13, 2021.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573.

Williamson, S.C., Metcalf, R.L., Trapani, F., Mohan, S., Antonello, J., Abbott, B., Leong, H.S., Chester, C.P.E., Simms, N., Polanski, R., et al. (2016). Vasculogenic mimicry in small cell lung cancer. Nat Commun *7*, 13322.

Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., and Theis, F.J. (2019). PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biol *20*, 59.

Wooten, D.J., and Quaranta, V. (2017). Mathematical models of cell phenotype regulation and reprogramming: Make cancer cells sensitive again! Biochim Biophys Acta *1867*.

Wooten, D.J., Groves, S.M., Tyson, D.R., Liu, Q., Lim, J.S., Albert, R., Lopez, C.F., Sage, J., and Quaranta, V. (2019). Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers. Plos Comput Biol *15*, e1007343.

Wu, W., and Wang, J. (2013a). Landscape Framework and Global Stability for Stochastic Reaction Diffusion and General Spatially Extended Systems with Intrinsic Fluctuations. J Phys Chem B *117*, 12908–12934.

Wu, W., and Wang, J. (2013b). Potential and flux field landscape theory. I. Global stability and dynamics of spatially dependent non-equilibrium systems. J Chem Phys *139*, 121920.

Yabo, Y.A., Niclou, S.P., and Golebiewska, A. (2021). Cancer cell heterogeneity and plasticity: A paradigm shift in glioblastoma. Neuro-Oncology noab269.

Yachie-Kinoshita, A., Onishi, K., Ostblom, J., Langley, M.A., Posfai, E., Rossant, J., and Zandstra, P.W. (2018). Modeling signaling-dependent pluripotency with Boolean logic to predict cell fate transitions. Mol Syst Biol *14*, e7952.

Yan, H., Li, B., and Wang, J. (2019). Non-equilibrium landscape and flux reveal how the central amygdala circuit gates passive and active defensive responses. J Roy Soc Interface *16*, 20180756.

Yang, D., Denny, S.K., Greenside, P.G., Chaikovsky, A.C., Brady, J.J., Ouadah, Y., Granja, J.M., Jahchan, N.S., Lim, J.S., Kwok, S., et al. (2018). Intertumoral Heterogeneity in SCLC Is Influenced by the Cell Type of Origin. Cancer Discov *8,* 1316—1331.

Yang, D., Qu, F., Cai, H., Chuang, C.-H., Lim, J.S., Jahchan, N., Grüner, B.M., Kuo, C.S., Kong, C., Oudin, M.J., et al. (2019). Axon-like protrusions promote small cell lung cancer migration and metastasis. Elife *8*, e50616.

Yeo, S.K., and Guan, J.-L. (2017). Breast Cancer: Multiple Subtypes within a Tumor? Trends Cancer *3*, 753–760.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics *26*, 976–978.

Yuan, S., Norgard, R.J., and Stanger, B.Z. (2019). Cellular Plasticity in Cancer. Cancer Discov *9*, 837–851.

Yuh, C.-H., Bolouri, H., and Davidson, E.H. (1998). Genomic Cis-Regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene. Science *279*, 1896–1902.

Zepp, J.A., and Morrisey, E.E. (2019). Cellular crosstalk in the development and regeneration of the respiratory system. Nat Rev Mol Cell Bio *20*, 551–566.

Zhang, W., Girard, L., Zhang, Y.-A., Haruki, T., Papari-Zareei, M., Stastny, V., Ghayee, H.K., Pacak, K., Oliver, T.G., Minna, J.D., et al. (2018). Small cell lung cancer tumors and preclinical models display heterogeneity of neuroendocrine phenotypes. Transl Lung Cancer Res *7*, 32–49.

Zhou, J.X., and Huang, S. (2010). Understanding gene circuits at cell-fate branch points for rational cell reprogramming. Trends Genet *27*, 55—62.

Zhou, J.X., Aliyu, M.D.S., Aurell, E., and Huang, S. (2012). Quasi-potential landscape in complex multi-stable systems. J Royal Soc Interface Royal Soc *9*, 3539—53.

Zhou, J.X., Isik, Z., Xiao, C., Rubin, I., Kauffman, S.A., Schroeder, M., Huang, S., Zhou, J.X., Isik, Z., Xiao, C., et al. (2016a). Systematic drug perturbations on cancer cells reveal diverse exit paths from proliferative state. Oncotarget *7*, 7415—7425.

Zhou, J.X., Samal, A., d'Hérouël, A.F., Price, N.D., and Huang, S. (2016b). Relative stability of network states in Boolean network models of gene regulation in development. Biosystems *142*, 15–24.

Zhou, P., Wang, S., Li, T., and Nie, Q. (2021). Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics. Nat Commun *12*, 5609.

Zou, M., Toivanen, R., Mitrofanova, A., Floch, N., Hayati, S., Sun, Y., Magnen, C.L., Chester, D., Mostaghel, E.A., Califano, A., et al. (2017). Transdifferentiation as a Mechanism of Treatment Resistance in a Mouse Model of Castration-Resistant Prostate Cancer. Cancer Discov *7*, 736–749.