2022

SINGLE CELL LINEAGE TRACING REVEALS MECHANISMS OF TUMOR INITIATION AND CHEMORESISTANCE IN SMALL CELL LUNG CANCER

Hannah Wollenzien

SINGLE CELL LINEAGE TRACING REVEALS MECHANISMS OF TUMOR INITIATION AND CHEMORESISTANCE IN SMALL CELL LUNG CANCER

By

Hannah Wollenzien

B.A., Concordia College, 2017

A Dissertation Submitted in Partial Fulfillment of
the Requirements for the Degree of Doctor of Philosophy

Division of Basic Biomedical Sciences

Basic Biomedical Sciences Program
In the Graduate School
The University of South Dakota
August 2022

The members of the committee appointed to examine the dissertation of Hannah Wollenzien find it satisfactory and that it be accepted.

Faustino

Randolph Faustino, PhD, Chairperson

Pilar de la Puente, PhD

Michael Kareta, PhD

khosrow Rezvani

Khosrow Rezvani, MD, PhD

Jill M Weimer

Jill Weimer, PhD

Paola Vermeer

Paola Vermeer, PhD

ABSTRACT

Small Cell Lung Cancer (SCLC) is a devastating disease characterized by a very low two-year survival rate and almost universal acquisition of chemoresistance. Nearly all patients have tumors driven by functional inactivation of the tumor suppressors *Rb* and *p53,* but despite the uniform origins of this tumor, not all patients are genetically or phenotypically identical. SCLC can be subtyped into four unique molecular subtypes, determined by the expression of ASCL1, NEUROD1, POU2F3, or YAP1. These subtypes are plastic, and subtype switching after chemotherapy has been documented. Without the understanding of how tumor heterogeneity arises, we cannot solve the challenge of chemoresistance in SCLC. In recent years, a powerful new tool in studying tumor heterogeneity has emerged. Genetic barcoding allows for the identification and tracking of individual tumor populations by inserting a small genetic sequence ("barcode") into the genome of tumor cells. As the cells divide, the barcode is passed on and a high-resolution lineage map is constructed. Here, genetic barcoding is used for the first time in SCLC, combined with single-cell RNA sequencing in a genetically engineered mouse model and a xenograft model of SCLC.

In the mouse model of SCLC, tumors were sequenced at early, middle, and late stages of tumor development, as well as chemoresistant tumors. While no barcodes were detected by scRNA-seq, valuable information about the process of tumor development in SCLC is observed. I identify two cellular populations ("early" and "late") that arise during tumor development. A notable difference in the two populations is the expression of genes corresponding to members of the AP-1 network. The AP-1 network was validated to be critical for tumorigenesis in SCLC.

Barcoded SCLC xenografts and chemoresistant xenografts belonging to two SCLC subtypes were generated. scRNA-seq revealed increased transcriptomic plasticity following chemotherapy treatment in SCLC-A xenografts but not SCLC-N xenografts. The Cancer Testis Antigens PAGE5 and GAGE2A were identified and validated as mediators of chemoresistance in SCLC. This work represents the first application of genetic barcoding in SCLC and identifies actionable drug targets for future development.

Dissertation advisor:________________________________

Michael Kareta, PhD

Acknowledgments

I would like to express my deepest gratitude to Dr. Michael Kareta, whose guidance and patience have been critical to my development and training. No idea is too big and no question is too small for Dr. Kareta to consider, which has led to a creative, collaborative, and highly impactful environment in which to succeed. This endeavor would not have been possible without the role of my dissertation committee, Drs. Randolph Faustino (chair), Jill Weimer, Paola Vermeer, Pilar de la Puente, and Khosrow Rezvani, without whom I would not have been successful.

I would like to extend my sincere gratitude to all members of the Kareta lab, past and present: Ethan Thompson, Ellen Voigt, Madeline Wallenberg, Kirtana Kumar, Rounak Pokaharel, and Dr. Robert Sczepaniak-Sloane, for their help in troubleshooting countless experiments, editing assistance, and never-ending moral support, as well as the number of talented and passionate interns and rotating students I have had the pleasure of working with. Many thanks to our collaborators at Sanford Research and the University of South Dakota, including Dr. Malini Mukherjee, Dr. Yohannes Tecleab, and Jared Wollman. Their contribution and collaboration have been critical.

Thanks should also go to members of my cohort, particularly Bethany Freel and Rhiannon Sears. The future of science is very bright in your capable hands.

Thank you to the administration at Sanford Research and the University of South Dakota for fostering an excellent learning and training environment. I would also like to acknowledge the funding I have received from the National Institutes of Health.

Finally, I would be remiss if I did not mention the monumental role my family and friends have played in my success as a graduate student. Jordan – thank you for your positivity, grounding, and “I told you so!” every time an experiment I was worried about was successful. Margaret – thank you for your continued support and understanding, even on my neurotic days. I hope I have made you proud. And to my parents – you’ve moved mountains for me my entire life and I would not be the person or scientist I am today without you. Thank you for always knowing exactly when to call when I needed encouragement, listening to numerous practice talks (and pretending to understand everything I said!), and believing in me even when I didn’t believe in myself.

Dedication

To my Dad. Your passion, kindness, resilience, and compassion have been a steadfast example all my life.

“Never be so clever you forget to be kind. Never be so kind you forget to be clever”

-Taylor Swift, PhD (h.c.)

Contents

Table of Figures

List of tables

Chapter 1: Introduction

Small cell lung cancer (SCLC) is a devastating disease characterized by a 5-year survival rate of only 6%[1]. The majority of patients (75%) present with extensive stage disease at diagnosis, and survival for these patients is generally under 1 year[2,3]. Given the usually extensive stage disease at diagnosis, surgical resection is rare, and the majority of patients are treated with first-line platinum agents such as cisplatin or carboplatin and etoposide, a topoisomerase, as well as PD-L1 inhibitor, regardless of PD-L1 status[4,5]. The recent addition of a PD-L1 inhibitor improved survival about two months in clinical trials[6], however advances in therapeutics are critically needed. Patients who present with brain metastasis may receive cranial radiotherapy. While initial response to chemotherapy generally seems promising, the majority of patients will rapidly acquire resistance and relapse within months (Figure 1A)[3]. After relapse, several options exist for secondary treatment, however, few patients see a benefit after a few months, and the majority then receive palliative support[5]. The therapeutic options for SCLC are rarely targeted or curative and the grim outlook for patients with SCLC has remained largely unchanged in the last 60 years[2,3].

SCLC is surprisingly uniform in genetic alterations driving the disease. Nearly all patients have functional inactivation of the tumor suppressors *RB1* (93%) and *TP53* (100%) (Figure 1C)[7]. Both critical tumor suppressors, RB is a canonical cell cycle regulator, and transcriptionally regulates the transition between G1 and S phase. The loss of RB in cancer often leads to dysregulation of the cell cycle[8]. RB has also been found to act as a transcriptional regulator of oncogenic pathways. RB deactivates pluripotency genes *SOX2* and *OCT4,* so when RB is lost in cancer, pluripotency networks are de-repressed and lead to a plastic, stem cell like state (Figure 1B)[9].

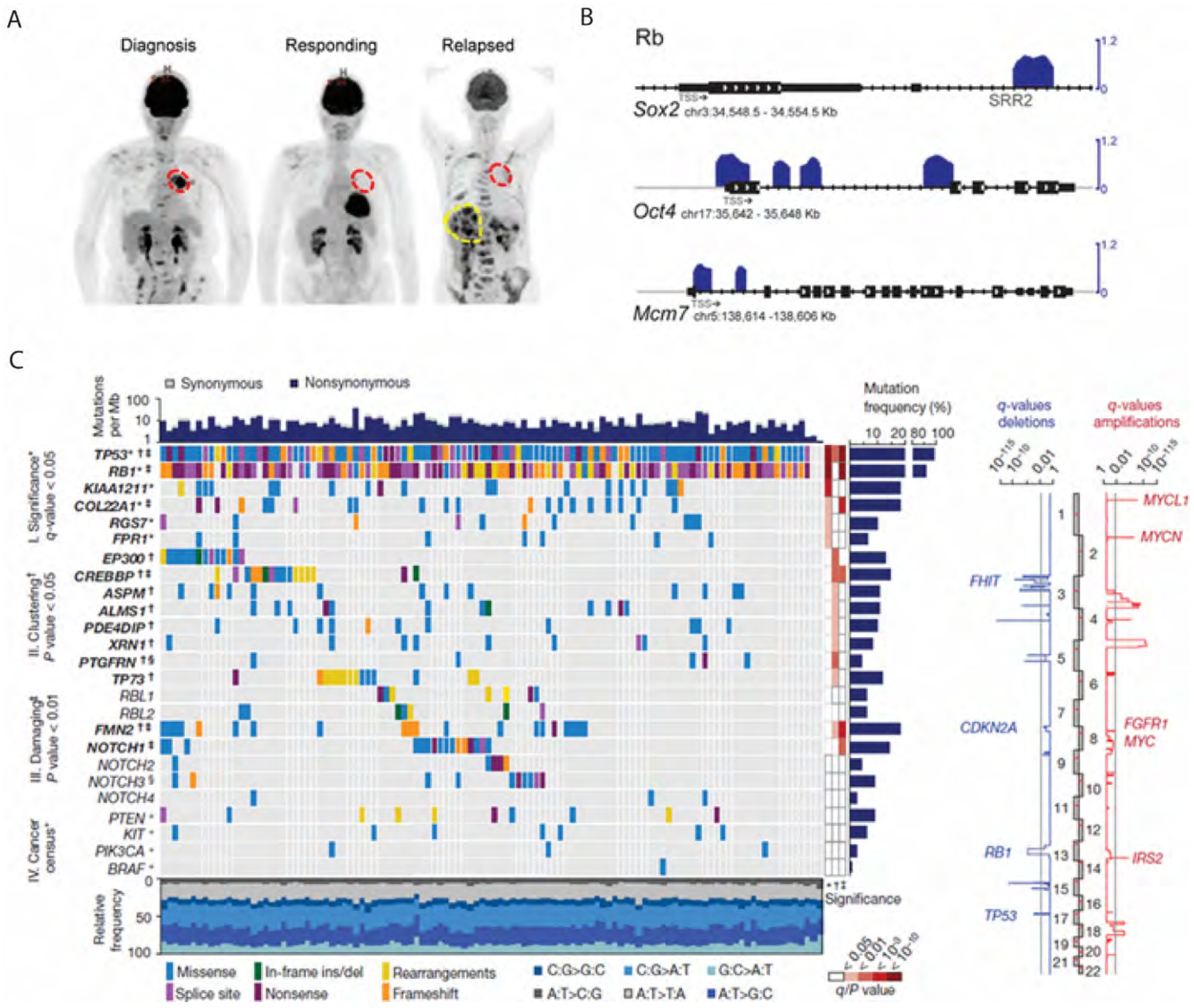


Figure 1: Clinical and genomic presentation of SCLC.

A: Radiograph of a patient with SCLC at diagnosis with a lesion circled in red. At a responding stage, the lesion is not present. At the relapsed stage, disseminated metastasis can be observed. From Stewart et al., 2020. **B**: RB binds to and activates pluripotency genes. ChIP data showing RB (blue) binding sites on the Sox2, Oct4, and Mcm7 genes. From Kareta et al., 2015.**C:** Variant allele frequency of commonly altered genes in SCLC. Gene names are indicated in rows and individual patients in columns. A colored rectangle indicates a mutation in that gene. P53 and Rb on the first and second rows, respectively, are mutated in almost 100% of patients. Other commonly altered genes indicated are NOTCH family members. **D:** Copy number alterations frequently observed in SCLC. Blue variants represent deletions in genes such as Rb and P53, and red indicates gene amplifications. Commonly amplified in this data set are MYC family genes. From George et al., 2016**.**

The expression of *SOX2* has been implicated in SCLC, where it acts as a transcription factor, and acts in an oncogenic fashion by regulating key SCLC pathways[10-12]. In addition to the de-repression of *SOX2* driven by loss of RB, *SOX2* has also been found to be amplified in around 30% of patients, and is required for tumorigenesis (Figure 2A, B, C)[10,11,13]. TP53 is implicated in the majority of human cancers. It is a transcription factor that acts in response to cellular stressors such as DNA damage, when it works to determine the cell's response. Loss of *P53* in cancer can lead to an accumulation of DNA and cellular damage and mutations[14]. Also common are alterations in *MYC.* Patients frequently have amplification of *MYC, MYCL1, or MYCN*[7]*.* The expression of *MYC* genes in SCLC are mutually exclusive, and have a role in mediating SCLC subtype and chemoresistance[7,11,13,15-18]. *NOTCH* family members are frequently implicated in SCLC pathogenesis. Around 25% of patients have genomic alterations in *NOTCH* [7]. In SCLC, *NOTCH* expression is downregulated, which allows for neuroendocrine differentiation. *NOTCH1* is epigenetically suppressed in the SCLC-A type, which allows for the activation of ASCL1[19,20]. Additionally, the NOTCH ligand DLL3 is highly expressed in SCLC, and is correlated with ASCL1 expression and subtype, with highest expression in the SCLC-A type[21,22]. Alterations in NOTCH signaling may also play a role in the plasticity between SCLC subtypes[11,16,18]. A 2017 report found that in neuroendocrine SCLC, NOTCH signaling acted as a pro-tumorigenic factor, while in non-neuroendocrine SCLC, NOTCH signaling acted as a tumor suppressive factor[20]. In MYC-driven SCLC, MYC activates NOTCH signaling, which drives the SCLC-A (neuroendocrine) to SCLC-N (non-neuroendocrine) subtype transition[16].

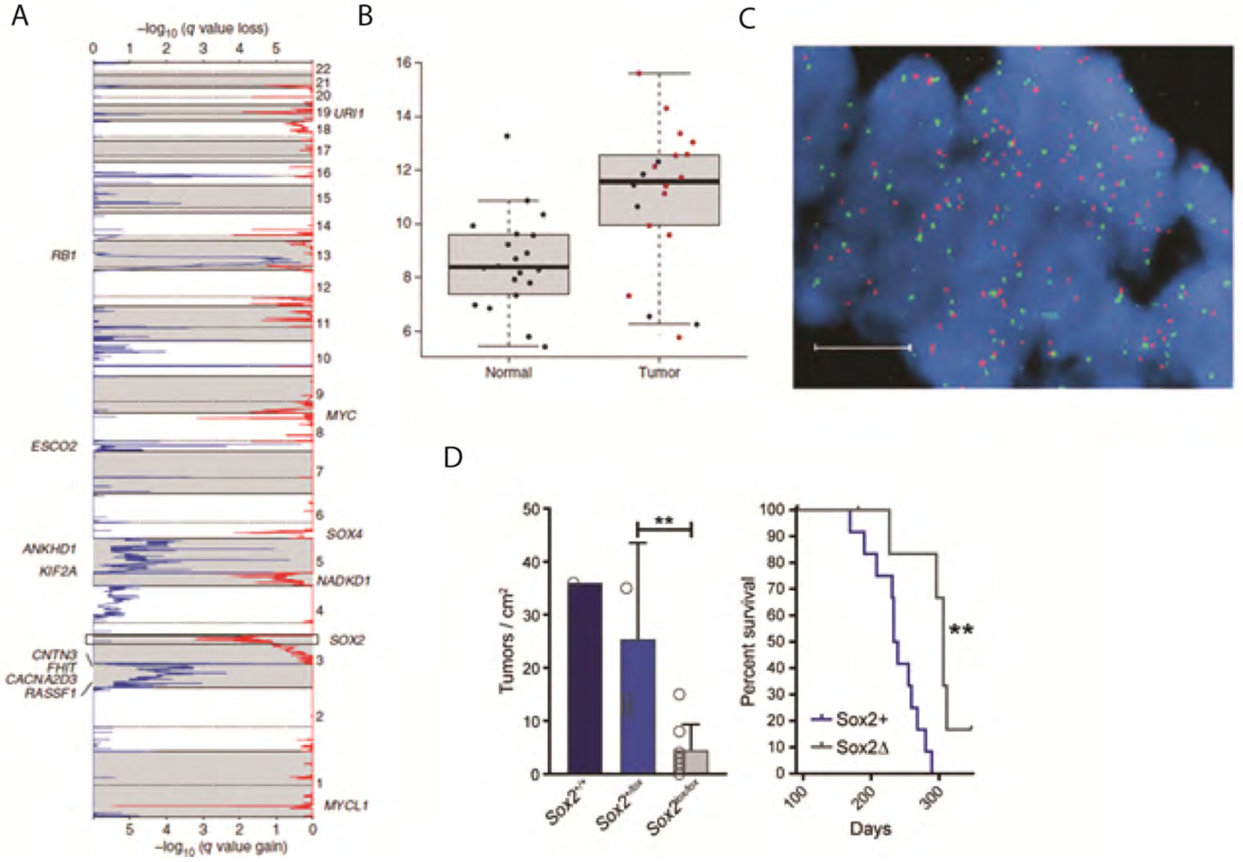


Figure 2: Sox2 is frequently amplified in SCLC and is required for tumorigenesis
A: Genetic amplification and deletions in SCLC. *SOX2* (red, boxed) is frequently amplified. **B:** Expression of *Sox2* in normal and SCLC tissues. **C:** FISH analysis of *Sox2* copy number. Red is the *Sox2* probe and green is the centromeric probe. From Rudin et al., 2013 **D:** *SOX2* is critical for tumor formation in a mouse model of SCLC. The gray bar indicates tumors from mice with biallelic deletion of *SOX2,* which form significantly fewer tumors than mice with at least one copy of *SOX2.* Mice with deletion of *SOX2* survive significantly longer than mice with expression of *SOX2.* From Voigt, Wallenberg et al., 2021.

Historically, SCLC has been thought of as being subtyped as either "classic" or "variant". Classic SCLC generally has more neuroendocrine features than variant, and they behave differently in cell culture, with the neuroendocrine-high SCLC-A lines growing more as organized spheres, and the SCLC-N subtypes growing as less organized clusters, or occasionally, adherent cell lines[23]. As genomics techniques have progressed, SCLC has been able to be subtyped based on the genomics of the tumor. Tumors can be subtyped based on expression of *ASCL1, NEUROD1, POU2F3,* and *YAP1* (Figure 3A, B, C)[24]. The subtypes differ in their molecular, histological, and phenotypic characteristics, both in culture and in the clinic. Historically called "classic" SCLC, the majority of patients with SCLC have *ASCL1*-driven disease (or SCLC-A). SCLC-A tumors generally arise from pulmonary neuroendocrine cells and display neuroendocrine features and gene expression, and are high in *MYCL, SOX2,* and *DLL3*[11,22,24]. *NEUROD1*-driven tumors (SCLC-N) are also neuroendocrine, although to a lower degree than SCLC-A. They have lower expression of *SOX2,* and express *MYC* instead of *MYCL,* and were historically classified as "variant" due to their distinct histology[11,22,24]. The *YAP1*-driven tumors (SCLC-Y) are more rare and seem to exist in the same lineage as SCLC-A and SCLC-N, as indicated by some tumors showing plasticity between the three[16]. SCLC-Y is non-neuroendocrine and occasionally has wild-type expression of RB[24]. Tumors that are SCLC-Y may have histological features that contain the more "variant" or combined cell morphology[22]. The final subtype, SCLC-P (*POU2F3* driven) is much more rare than the other three. These tumors are generally non-neuroendocrine and express more of the markers of tuft cells, suggesting SCLC-P tumors arise from a separate lineage than SCLC-A, SCLC-N, or SCLC-Y, which primarily have pulmonary neuroendocrine cells as their cell of origin[16,24]. There is ample evidence that these subtypes are not static in nature. Patient histology often demonstrates separate areas of SCLC-N and SCLC-A within the same patient sample[22,24]. Additionally, cell culture and mouse models have demonstrated subtype switching of tumors (Figure 3D, E)[11,16,18,25].

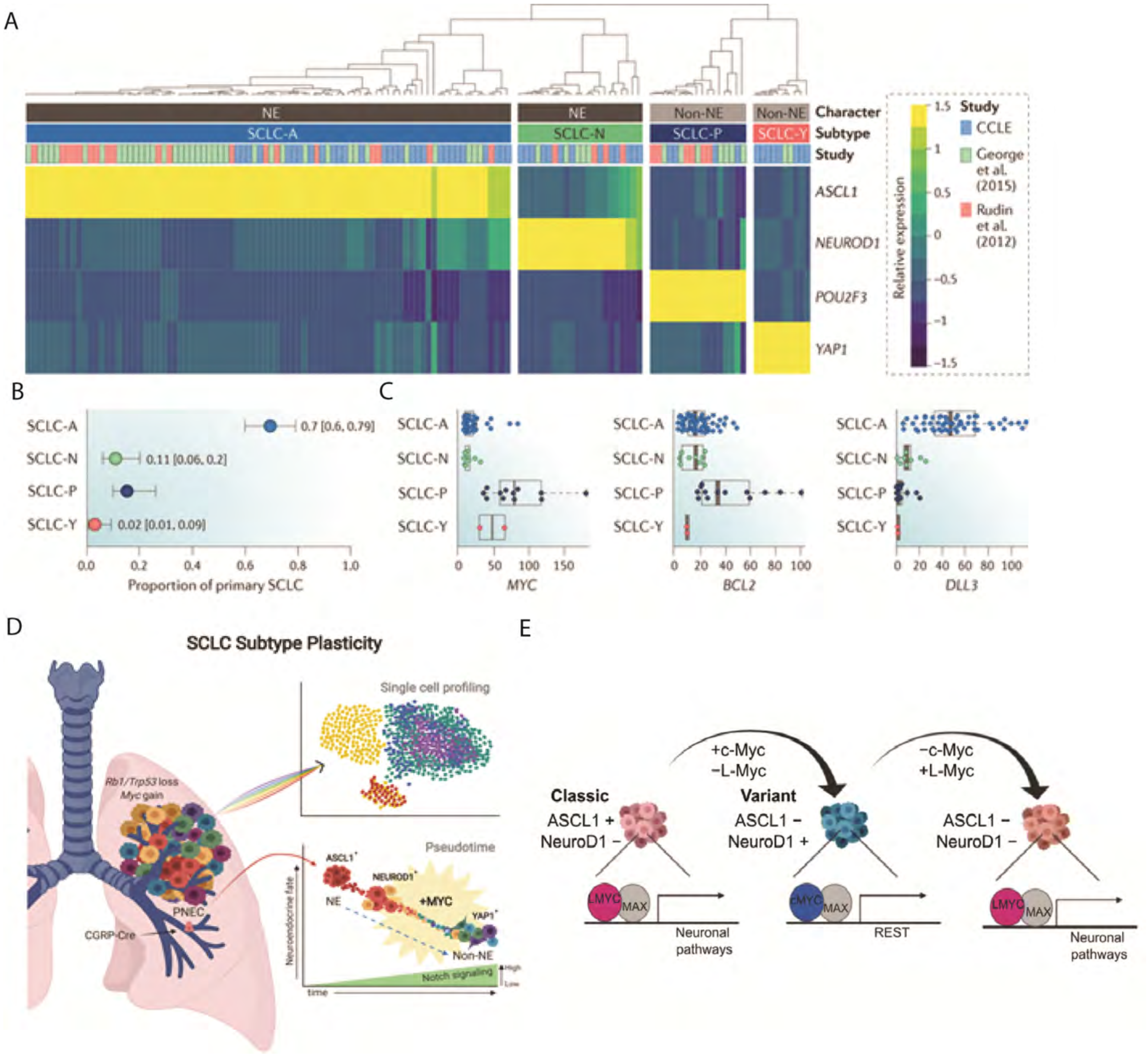


Figure 3: SCLC can be subtyped in to four plastic molecular subtypes
A: Transcriptomic data of SCLC samples and cell lines showing clustering by the expression of one of four factors: *ASCL1, NEUROD1, POU2F3,* and *YAP1.* Subytpes also stratify by neuroendocrine or non-neuroendocrine status. **B:** Proportion of patient samples belonging to each subtype. Most patients have tumors that are SCLC-A, followed by SCLC-N. **C:** Expression of SCLC associated genes *MYC, BCL2,* and *DLL3.* Expression level of these genes is dependent on subtype, indicating a distinct phenotype for each subtype. **D:** Pseudotime trajectory based on scRNA-seq of *Myc* driven SCLC demonstrates transition from SCLC-A to SCLC-N, and SCLC-Y. This coincides with a switch in *Myc* expression and upregulation of Notch family members. From Ireland et al., Cancer Cell 2020. **E:** *Myc* expression drives plasticity in SCLC subtypes from SCLC-A to SCLC-N. Notch pathways are also upregulated in this switch, and a change in neuronal pathways is observed as the subtype switches. From Patel et al., Science Advances, 2021.

Despite the almost uniform loss of *RB* and *P53,* there still may be other genetic drivers of SCLC. In order to better understand the genetic drivers of SCLC, Peifer et al (2012) did SNP array analysis, exome sequencing, and genome sequencing of tumors and cell lines. They confirmed common loss of *RB,* gain of *SOX2, FGFR1, MYCL,* and *MYCN.* The *MYC* family member amplifications were mutually exclusive. Additionally, they identified a set of likely driver genes for SCLC (*TP53, RB1, PTEN, CREBBP, EP300, SLIT2, MLL, COBL,* and *EPHA7)*[26]. To investigate the mechanism of tumor progression and identify somatic drivers of SCLC, McFadden et al (2014) performed whole exome sequencing on matched tumors and metastases from SCLC mouse models. They found frequent copy number variants (CNVs), many of which have an impact on *MycL* and the Notch pathway. Chromosome 4 genomic rearrangements were common, which may lead to *MycL* and *Nfib* amplification. These rearrangements are similar to ones seen in the human homologues in human cases of SCLC. Mutations in *Pten* and members of the *Pten* pathway were very common in their analysis. Alterations in Pten signaling may be a mechanism of tumor promotion, and loss of Pten signaling promotes tumor progression. To understand heterogeneity, they compared DNA rearrangements and point mutations in primary tumors and metastasis from individual mice, and showed that metastases had a greater number of mutations with a high allelic fraction, indicating that metastasis is a bottleneck event. In clonal analysis, most tumors had 2-5 individual tumor subclones, and in some cases, multiple metastases from different tumor subclones were seeded to the site of metastasis[27]. Using proteomic profiling, Tripathi et al sought to understand chemoresistance in SCLC Patient Derived Xenografts (PDX) and cell lines. In a chemoresistant cell line, they found significant increase in the cell surface proteome, indicated by an abundance of proteins associated with cytoskeletal reorganization and cell adhesion. They focused the analysis on the five most differentially expressed proteins: EGFR, JAG1, ITGB1, EPHA2, and MCAM. In PDX, MCAM was increased in the chemoresistant tumors, and these tumors had higher EMT markers. Knockdown of MCAM in culture did not impact EMT markers but did decrease cell proliferation and colony formation. Cells with a knockdown of MCAM were also more sensitive to chemotherapy and had an increase in pro-apoptotic proteins. MCAM overexpression in cell culture led to an increase in cell survival after chemotherapy. This is due to SOX2-dependent regulation of the PI3K/AKT pathway, which is upregulated in chemoresistant calls and acts as a regulator of MCAM. Chemoresistant cells also had a lower metabolic rate, which indicates a shift towards glycolysis in these cells[28].

Cells in a single tumor are often phenotypically different from one another and contribute differentially to the tumor dynamics, a phenomenon known as intratumoral heterogeneity (ITH)[29]. ITH is perhaps best demonstrated by the classic example of cancer stem cells (CSC), which generally make up a minority of the bulk of a tumor, yet they are the cells most directly responsible for tumor growth and maintenance[30]. ITH has an impact on response to therapy, particularly in SCLC, as demonstrated by patients' almost full response to chemotherapy, only to have extensive disease re-occur rapidly[3,31]. Clearly, there exists a population of cells are either inherently resistant to therapy, or have the plasticity to adapt and become resistant when faced with therapy[3]. Understanding ITH is critical to developing new, effective therapeutics for SCLC. Armed with the knowledge of which populations contribute to tumor dynamics, we can design intelligent therapeutics to address the most aggressive cellular populations. Understanding ITH is critical to developing and targeting new therapeutics, and ITH has been studied in many tumor types, but because of the lack of SCLC samples in the TCGA, SCLC has been excluded from large-scale analyses of ITH[32].

Tumor heterogeneity has been observed in SCLC, and in recent years, there have been several groups that have worked to understand the origins, evolution, and functional impact of heterogeneity in SCLC using genomics, transcriptomics, and proteomics. Yang et al (2018) sought to understand the origins of tumor heterogeneity depending on cell type of origin in mouse models of SCLC. They used the *$Rb^{lox/lox}$, $p53^{lox/lox}$, $p130^{lox/lox}$* (RPR2) mouse and initiated tumors using a general Cre adenovirus (CMV-Cre) and a CGRP-Cre adenovirus, which initiates tumors in just the neuroendocrine cells. The tumors from the CMV-Cre mice were generally high in Nfib, and were metastatic. Tumors from the CGRP-Cre mice required a higher concentration of virus to initiate, and to form metastatic tumors, and metastatic events were not reliant on Nfib. The ability to form metastases with and without the expression of Nfib indicates multiple metastatic pathways that the tumors could take. They used a multi-color reporter mouse bred with the RPR2 model to investigate the mechanisms of metastasis in their mouse models. The majority of metastases came from a single primary tumor and were clonal, indicating not all cells in a tumor have an equal likelihood of metastasis. In the transition from primary to metastatic lesion in the CMV-Cre mice, there were widespread changes in gene expression in genes related to neuronal differentiation and cell cycle. In the CGRP-Cre mice, there were few gene expression changes between primary and metastatic tumors. The CGRP-Cre tumors had higher expression of neuroendocrine genes, and the CMV-Cre tumors higher expression of epithelial cell markers. Overall, different cells of origin in the CGRP-Cre initiated tumors and the CMV-Cre initiated tumors led to heterogeneity in genomic profile, metastasis mechanism, and histology[33]. To understand the impact of *Myc* and *Nfib* on tumor heterogeneity and chemoresistance, Bottger et al (2019) used three common mouse models of SCLC – loss of *Rb* and *p53* (RP); loss of *Rb, p53,* with *MycL* amplification (RPM); and loss of *Rb* and *p53* with an *Nfib* overexpression (RPF).

They found heterogeneity in the histologic features of each of these three tumor models and the percentage of lesions that were bronchiolar or alveolar. *MycL* promoted lesions that were neuroendocrine and high in *Ascl1*. Regions in all models that were sensitive to chemotherapy were high in CDH1, and after chemotherapy, there was an increase in CDH1-low lesions, underlying cisplatin resistance. The proportion of the tumors that were CDH1-high and sensitive to chemotherapy differed by mouse genotype. After cisplatin treatment, there was a shift to a more epithelial signature, and changes in metabolism indicating a decrease in proliferation in response to cisplatin treatment[25]. Further investigating the link between *Myc* and tumor heterogeneity, Ireland et al (2020) used both the RPR2 model and a *Myc* overexpression mouse model (*Rb*$^{lox/lox}$*, p53*$^{lox/lox}$*, LSL-Myc*^{T58A}*;* RPM) combined with multiple timepoint scRNA-seq to evaluate *Myc* signaling in lineage fate determination. Early lesions from both models are ASCL1 high with classic neuroendocrine markers. In later lesions, the RPR2 model (generally high in *MycL*) maintained the ASCL1, high neuroendocrine subtype, and the RPM model showed a decrease in neuroendocrine markers. Pseudotime analysis reconstructed a lineage showing that *cMyc* can convert early ASCL1-high lesions to the SCLC-N or SCLC-Y subtype (Figure 3D). They validated this in primary culture of early tumor lesions from the RPM model, which start off as SCLC-A and transition to SCLC-N with high expression of non-neuroendocrine markers like NOTCH. Conversely, cells from the *MycL* high RPR2 tumors remain as SCLC-A in culture. Overexpression of *cMyc* in SCLC-A cells in culture converted cells to neuroendocrine-low SCLC-N and later to SCLC-Y. NOTCH signaling plays a role in the transition from SCLC-A to SCLC-N, and they showed that *cMyc* is directly responsible for the change in NOTCH signaling. Interestingly, if a Cre specific to AT2 or club cells was used to initiate tumors, the resulting tumors are SCLC-P, which is rarely the case in tumors resulting from neuroendocrine cells.

Therefore, the SCLC-P subtype arises from a different cell of origin than the other subtypes, and is driven by *cMyc*[16]. To study the mechanistic link of *cMyc* or *MycL* expression in determining SCLC subtype, Patel et al (2021) used transcriptomic data to associate networks with *MycL* or *cMyc* expression (Figure 3E). They found distinct transcriptomic profiles associated with either *MycL* or *cMyc*. In representative cell lines, ATAC-seq found *MycL* and *cMyc* had different DNA binding profiles, which suggests differential regulation of transcriptomes. All SCLC-N lines were associated with c-MYC accessibility at *NEUROD1* and all SCLC-A lines were associated with MYCL accessibility at *ASCL1*. Intriguingly, c-Myc or MYCL had no association with the expression of *YAP1* or *POU2F3.* Overexpression of *MYCL* in an SCLC-N cell line did not convert the line to SCLC-A, but led to an increase in neuronal genes. Depletion of *cMYC* in SCLC-N did lead to downregulation of NEUROD1. Conversely, overexpression of *cMYC* in an SCLC-A line led to a decrease in neuroendocrine markers of SCLC and more "variant" histology, and trans-differentiated an SCLC-A line to an SCLC-N lineage. Bulk RNA-seq of these cells showed an increase in epithelial genes and *cMYC* associated pathways. They also found upregulation of the Notch pathway, which has been shown to negatively regulate ASCL1 expression[18]. Taken together, *cMYC* and *MYCL* are lineage determining factors that play a direct role in activating transcriptomic networks including NOTCH and epithelial pathways that characterize the SCLC-N or SCLC-A subtypes.

Given the importance of c*Myc* and *MycL* in SCLC, Grunblatt et al (2020) wanted to evaluate the role of *Mycn* in SCLC. They developed a mouse model with deletion of *Rb* and *p53,* and overexpression of *Mycn* (RPMYCN). RPMYCN mice developed tumors faster and had a much lower median survival than the *Rb* and *p53*- loss driven mice (RP, *MycL* high). The majority of the tumors from the RPMYCN mice were of the "classical" type by histology and had ASCL1 expression by immunohistochemistry, although scattered regions of NEUROD1 and YAP1 positive cells existed. When *Mycn* is turned off in the tumors that formed in RPMYCN mice, tumors regressed, indicating that tumors that start from a *Mycn* high population are reliant on *Mycn* to continue to proliferate at as high of a rate, although they do eventually return. Both RPMYCN mice and *MYCN*-driven patient derived xenografts (PDX) were more resistant to cisplatin and etoposide treatment. To evaluate the regulation of *MYCN* on tumor dynamics, they used RNA-seq and found a number of MYC target genes to be differentially expressed including immune signaling pathways. This matches the phenotype seen in the RPMYCN mouse model, as immune cells from the *MYCN* tumor model had a significant decrease in CD3 T-cells and monocytes, as compared to cells from non-MYCN driven tumor models. Finally, using CRISPR-Cas9 sgRNA inactivation screens, they found that USP7 is responsible for maintaining MYCN stability, and when USP7 is inhibited, there were decreased levels of MYCN. USP7 inhibition also sensitized *MYCN*-driven PDX to cisplatin and etoposide[15].

To investigate tumor heterogeneity before and after chemoresistance, Stewart et al (2021) used circulating tumor cells (CTC)-derived xenografts isolated from tumor cells circulating in the blood of patients who were both chemo-naïve and chemoresistant, combined with scRNA-seq. The majority of the tumors were neuroendocrine, and most were high in ASCL1, even from chemoresistant patients. Both MYC and MYCL were activated, and some tumors had mixed expression of MYC and MYCL within a single tumor. Using the transcriptomic data, they calculated an intratumoral heterogeneity (ITH) score for each tumor and found increased ITH and an increased number of transcriptional clusters in PDX from patients who had chemoresistance. Multiple resistance pathways were upregulated within single tumors, indicating it is likely that multiple resistance pathways arise at one time. After treatment with cisplatin in

PDX, EMT score increased and ASCL1 decreased, but they did not see an increase in NEUROD1 expressing cells. This data indicates an increase in ITH after therapy and the ability for tumors to arise multiple resistance pathways within a short period of time[31]. To understand the spatial component to ITH, Rovira-Clave et al (2021) used epitope combinatorial tags combined with multiplex ion beam imaging (EpicMIBI), which allows for tracking of barcodes within the tissue in SCLC xenografts from an SCLC-N cell line (H82). EpicMIBI allows for the identification and tracking of clonal populations of cells in their spatial position combined with single-cell proteomic data. They observed heterogeneity in neuroendocrine and non-neuroendocrine states, as well as differences in epigenetic markers and vimentin expression within a single xenograft. Non-neuroendocrine cells all cluster together, and based on clonal analysis, have different growth patterns than the neuroendocrine cells. They saw that the rare cells cluster near each other, indicating that heterogeneity is not equally distributed throughout the tumor, and is instead located in subclonal "patches". A minority of patches had loss of PTEN, and these influenced the growth of their neighbor patches that still had wild-type PTEN. Xenografts grown from cell lines do indeed form heterogeneous tumors, and the clonal patches within the tumor have the ability to influence the behavior of nearby patches[34]. Given the very limited access to human specimens, Chen et al. (2021) used autopsy samples and whole-exome and transcriptome sequencing of patient samples to understand heterogeneity in human samples, particularly metastatic and immune signatures. They found a high mutational burden and high copy number variants (CNVs) present in all patients. They had access to both primary and metastatic sites for the patients, so they were able to reconstruct a clonal lineage using CNV analysis. Clonal heterogeneity was different depending on the patients but there were some signatures in common. In general anti-tumor immune markers were elevated in advanced tumors, and stratified based on tumor subtype[13].

Past work has shown that tumor heterogeneity does exist in SCLC, is plastic, and has a functional impact on tumor growth and response to chemotherapy. However, there are a number of remaining questions for SCLC tumor evolution. We have yet to investigate the very early stages of tumor formation in SCLC, to understand what the source and drivers of tumor diversity are (Figure 4). The understanding of how ITH arises are critical to designing therapeutics and targeting the most aggressive populations (Figure 4). Furthermore, the tumor evolution studies that have been done in SCLC, while incredibly useful, work on a pseudotime or retrospective perspective, which is not temporally resolved.

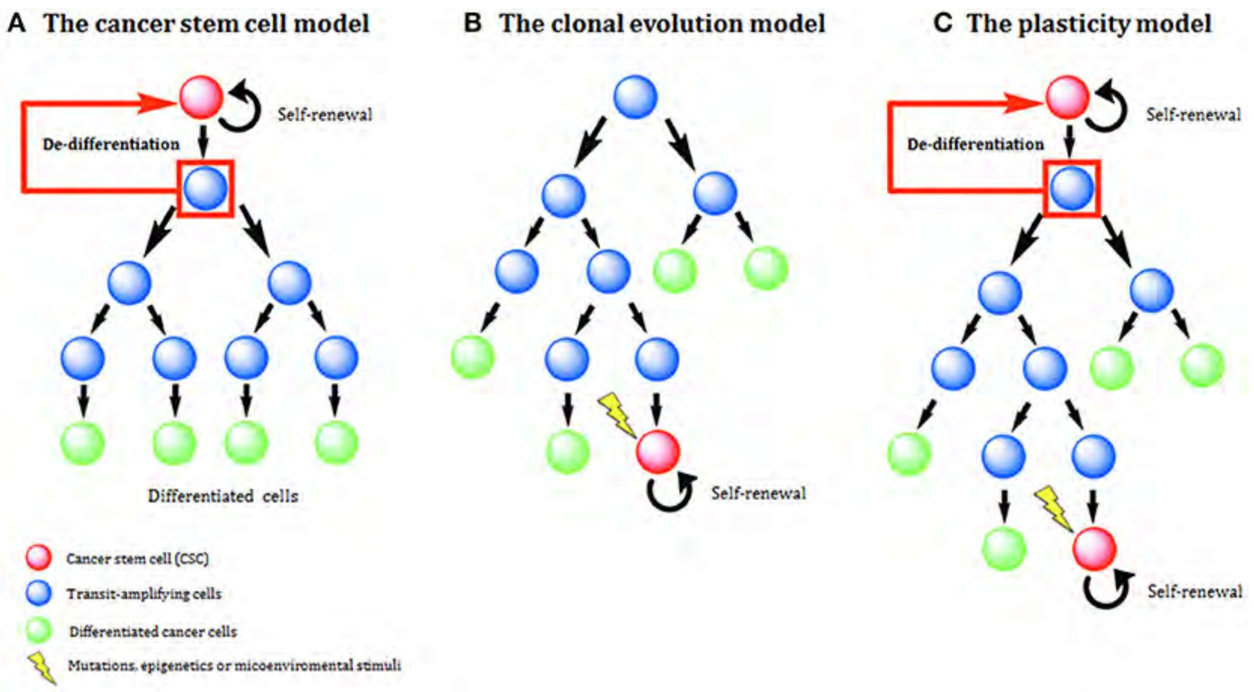


Figure 4: Theories of the development of ITH.
Understanding the origins and evolution of ITH is critical to uncovering targetable populations. This figure presents three models for the development of ITH. The cancer stem cell model (**A**) in which one self-renewing population can differentiate in to multiple clonal populations. The Clonal evolution model (**B**), has a core differentiated population and one cancer stem cell that self-maintains. **C** presents the plasticity model, in which there are multiple populations with self-renewal capacity.

Historically, tumors have been sequenced using bulk RNA sequencing methods, however, this masks the true contribution of individual populations and aggregates the signal from the entire sample[13,27]. Single-cell RNA sequencing (scRNA-seq) has radically transformed our understanding of ITH. The scRNA-seq studies that have been performed on SCLC tumors and xenografts[16,31,35] and other genetic profiling that have been used to generate pseudotime maps of ITH in response to growth and chemotherapy[13,16,18,25,27,31,34], while incredibly useful, do not allow for the pinpointing of the populations critical for these tumor dynamics. Instead, it is inferred from the populations identified in the screening. In recent years, genetic barcode lineage tracing has emerged as a novel tool to trace individual clonal populations over time[36-38]. Genetic barcode lineage tracing allows for the identification of individual cellular clones by inserting a unique piece of DNA to serve as a "barcode" in to the genome of a cell using a retrovirus or CRISPR sgRNA library[39,40]. As the cells divide, the barcode will be passed to the progeny, allowing for a high-resolution tracing of individual clonal populations (Figure 5). Originally used for tracing populations in hematopoiesis[36,41], the barcodes integrated are stable over time *in vivo*[36] and are able to be detected with scRNA-seq[41,42]. Barcoding serves as a technological advance on other lineage tracing methods such as fluorescence labeling[30]. Fluorescence labeling is limited in the number of potential fluorophores, and the tracking of mutation rates does not allow for a temporal component, since lineages must be reconstructed after heterogeneity has already developed[32,33,43].

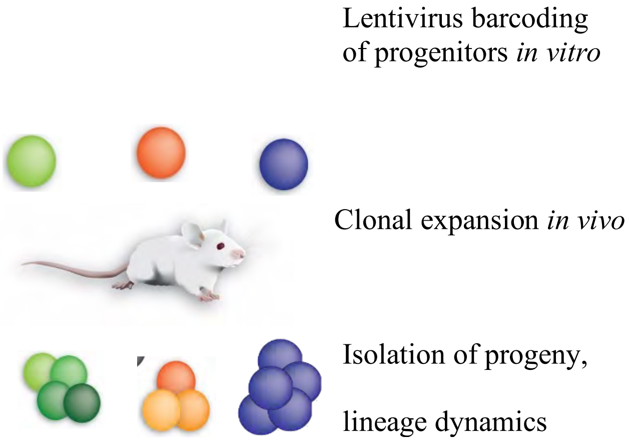


Figure 5: Overview of the barcoding system.

The use of genetic barcode lineage tracing to understand tumor heterogeneity has exploded in the last few years. It has now been used to understand cellular lineages and heterogeneity in a number of cancer types including glioblastoma, breast, non-small cell lung cancer, leukemia, and melanoma[37,40,44-50]. Often, integration of barcodes is lentiviral, often done along with a fluorescent marker like a GFP[37,44,45,48]. In glioblastoma, genetic barcoding has been used in a xenograft model to understand the cancer stem cell pool and chemoresistance. Lan et al (2017) first barcoded primary cultured glioblastoma cells and performed serial xenografts to identify the stem cell pool that is capable of tumor formation[37]. Neftel (2019) and Eyler (2020) both used lentiviral barcoding to understand tumor heterogeneity in glioblastoma after chemotherapy. Neftel identified for the first time the cellular states that simultaneously exist within a tumor, and characterized the plasticity in cell states in glioblastoma[45]. Using barcoded tumor spheres, Eyler focused on identifying lineages that were responsible for surviving targeted therapy. They were able to identify the acquisition of copy number gains in direct response to receptor tyrosine kinase inhibitor treatment in the clones that survived therapy[44]. Also using lentiviral barcoding, Emert et al (2021) combined the barcode with RNA fluorescence in situ hybridization (FISH) in melanoma cells in culture to identify cell states that lead to therapy resistance. They created two populations with identical barcodes, one that got sequenced to profile the barcodes in a treatment resistant population, and another that was molecularly profiled after using RNA FISH to isolate populations corresponding to the barcodes of the surviving clones that were sequenced. They identified multiple resistance pathways in response to therapy[48]. PRISM technology is a unique application of barcoding, which allows for the multiplexing of lenivirally barcoded cell lines from the cancer cell line encyclopedia (CCLE), can be used to assay several cell lines at once, which can then be de-multiplexed after molecular profiling. It has been used to measure therapeutic vulnerabilities and metabolic changes in hundreds of cell lines at once, which would be a tedious and near-impossible task without the ability to barcode the individual cell lines for later identification[51,52].

Another popular barcoding technique utilizes a CRISPR-Cas9-based barcode. In lung adenocarcinoma, Guernet et al (2016) used CRISPR-Cas9 to edit their gene of interest, while inserting a series of extra point mutations to serve as a heritable barcode. In this case, their barcode was detectable by qPCR, which made understanding clonal dynamics more cost-effective, but did not easily pair barcodes with large transcriptomic data[40]. Adaptable CRISPR barcodes have the ability to not only label clonal populations, but also evolve over the course of the disease to reconstruct a phylogenetic tree with higher resolution, and has been used in lung adenocarcinoma xenografts and mouse models to understand tumor evolution and metastasis[46,53]. Using a CRISPR-Cas9 approach, Rogers et al (2018) barcoded individual tumors in a mouse model of lung adenocarcinoma by barcoding founder cells during tumor initiation to understand the driver genes of tumor formation, but not necessarily subclonal lineages[54]. Recently, there has been some interest in isolating clones in real time, after identifying the barcodes of the populations of interest. ClonMapper combines barcodes and transcriptomic data that allows for the identification and recovery of populations of interest. It has been used in melanoma cells in culture, and has shown to be a powerful tool for understanding clonal dynamics *in vitro*[47].

In SCLC, genetic barcoding has been performed twice, but has not been used in combination with transcriptomic data, in a time-dependent manner, or to understand response to chemotherapy. Spatial epitope barcoding has been used in a xenograft model using one SCLC-N cell line to understand tumor architecture during tumor development[34]. Recently, a pre-print has described the use of barcoding in a mouse model of SCLC to understand tumor initiation. Cells were barcoded at the point of tumor initiation, which will allow for the identification of the alterations present in the cells able to form tumors using TUBA-seq, but does not allow for the understanding of tumor evolution[55]. To date, genetic barcoding has not been used to understand tumor evolution in a temporal manner, and has never been combined with a model of chemoresistance.

We know that ITH is key to SCLC growth and metastasis, and that SCLC tumors are highly plastic, and are able to adapt and overcome when faced with chemotherapeutic treatment. Without an in-depth understanding of ITH, we are not able to develop the most effective therapeutics for SCLC. The use of genetic barcoding and scRNA-seq in this work will allow us to view the early origins of ITH in SCLC for the first time, and identify new cellular populations and therapeutic targets to treat this disease.

Chapter 2: Methods

2.1: Ethics statement

Mice were maintained according to the guidelines set forth by the NIH and were housed in the Sanford Research Animal Research Center, accredited by AAALAC using protocols reviewed and approved by our local IACUC.

2.2: Generation and validation of barcoding libraries

2.2.1: Cloning of the barcoding libraries

To clone the retroviral barcoding library, a CAG-GFP retroviral plasmid was used as a backbone (gift from Fred Gage, Addgene plasmid # 16664 ; http://n2t.net/addgene:16664 ; RRID:Addgene_16664)[56]. A poly-A sequence was subcloned from TetO-FUW-sox2, a gift from Rudolf Jaenisch (Addgene plasmid # 20326 ; http://n2t.net/addgene:20326 ; RRID:Addgene_20326)[57], to the 3' end of the GFP at the PmeI site using InFusion Cloning (TaKaRa) and screened with PCR and restriction digests. DNA oligos containing the barcode sequence were ordered from Eurofins and annealed by heating to 95 degrees C for five minutes and allowed to cool to room temperature over the course of several hours. The CMV-GFP-polyA plasmid was digested at PmeI and HindIII (added in the polyA cloning step), and the annealed barcode was inserted by InFusion cloning (TaKaRa) at the 3' end of the polyA and 5' end of the GFP sequence. 40 of the initial colonies were screened for presence of the barcode via PCR, and four of those were further validated to contain the barcode using Sanger sequencing. All 40 colonies screened by PCR contained the barcode, and all four colonies screened by Sanger sequencing contained unique barcodes. Once presence of the barcode was confirmed in a number of colonies, the cloning product was transformed and plated to grow on five 15-cm plates of LB agar. Following overnight growth, all colonies were collected and pooled by flushing the plates with pre-warmed LB broth. Plasmid libraries were maxiprepped and the product was pooled and purified. Gamma-Retrovirus was produced by co-transfection of 293Ts with the barcoding retrovirus and retroviral packaging plasmid pCL-Amph. Retroviral supernatant was collected 48 and 72 hours after transfection and concentrated using the TaKaRa Retro-X concentrator.

To clone the AAV9-r26-GFP-BC CRISPR plasmid, a sgRNA targeted to *Rosa26* was subcloned from pU6-sgRosa26-1_CBh-Cas9-T2A-BFP, a gift from Ralf Kuehn (Addgene plasmid # 64216 ; http://n2t.net/addgene:64216 ; RRID:Addgene_64216)[58,59] and inserted in to the AAV-KPL backbone (AAV:ITR-U6-sgRNA(Kras)-U6-sgRNA(p53)-U6-sgRNA(Lkb1)-pEFS-Rluc-2A-Cre-shortPA-KrasG12D_HDRdonor-ITR (AAV-KPL), a gift from Feng Zhang (Addgene plasmid # 60224 ; http://n2t.net/addgene:60224 ; RRID:Addgene_60224))[60] at SacI and MulI using InFusion cloning (TaKaRa), and screened for presence of the insert using PCR and Sanger sequencing. Left and right homology arms to *Rosa26,* as well as a polyA signal were subcloned from pR26 CAG/GFP Dest, a gift from Ralf Kuehn (Addgene plasmid # 74281 ; http://n2t.net/addgene:74281 ; RRID:Addgene_74281)[58,59] in to the AAV9-r26 plasmid at PmlI. The barcode library, attached to a GFP was ordered from ThermoFisher's GeneArt program, and was inserted in to the AAV9-r26 plasmid at PmlI and BamHI via InFusion cloning. After the first 30 colonies were screened for insertion of the barcode by restriction digest, PCR, and Sanger sequencing, chemically competent cells were transformed with the plasmid pool and plated on ten 10-cm LB plates. After overnight growth, all plates were washed with pre-warmed LB and the pooled colonies were maxiprepped. The AAV9-r26-GFP-BC plasmid was made in to AAV9 at the University of Michigan Viral Vector Core.

2.3: Validation of barcode diversity in barcoded cells

To profile the diversity in the barcode pools, targeted amplicon sequencing was performed on the barcode region. PCR was used to amplify the barcode and add partial adaptors for Illumina sequencing. The minimal number of cycles needed to amplify the barcode region was used to minimize the risk of introducing variants, or over-saturate the sample with a limited number of barcodes that had been disproportionately amplified. Samples were sequenced on an Illumina platform at Genewiz. The targeted amplicon sequencing of the barcodes was trimmed and QC performed via CutAdapt. GREP was used to identify barcodes and export them to R Studio for analysis. In R studio, the true number of barcodes was determined by using the number of PCR cycles to backtrack diversity, combined with the PCR error rate, determined by the number of PCR errors in the constant region of the barcode. Chao2 modeling was used to estimate the number of barcodes in the pool.

2.3.1: Doubling time assays to determine fractional overlap of barcodes

In order to ensure sufficient overlap in barcodes between the "pre-growth" sample and xenografts, I sought to determine the optimal number of doublings before sufficient overlap in two independent samples was observed. 1,600,000 cells were seeded and barcoded using the CAG-GFP-BC retrovirus and a spinfection at 940xg for 2 hours. At each doubling for five doublings following spinfection, one well of cells was harvested and split in to two. The barcode region was amplified off of the cDNA, and partial adapters for Illumina-based sequencing was added. Amplicon sequencing was performed at Genewiz (South Plainfield, NJ) using an Illuminia miSeq platform. The barcodes were analyzed using a custom R script after trimming and QC via CutAdapt. To verify the results of the barcode sequencing, computational modeling was used. I simulated the same doubling time experiment 1,000 times using a custom R package.

Based on the results of the sequencing, modeling, and previously published work [41], three doublings after barcoding gives sufficient overlap between two independent samples and will be used to generate the barcoded xenografts.

2.4: Mouse protocols

2.4.1: In vivo model

The well-characterized *Rb*$^{lox/lox}$*, p53*$^{lox/lox}$*, p130*$^{lox/lox}$ SCLC mouse model [61] was bread to the *H11*$^{lox-stop-lox-Cas9}$ mouse[62] model (Jax #026816) to generate the RPR-Cas9 mouse used in this work. Tumors were initiated by intratracheal injection with Ad-CMV-Cre (Baylor Viral Vector Core) to delete the *Rb, p53,* and *p130* loci, and induce expression of Cas9. At one month intervals for five months after tumor initiation, the AAV9-r26-GFP-BC virus was delivered to separate cohorts of mice via intratracheal injection to barcode the forming tumors at the *Rosa26* locus using the CRISPR-Cas9 system. Mice were euthanized at one month intervals after their Ad-CMV-Cre injection, up to six months, or when they became moribund according to institutional IACUC guidelines. Two mice received chemotherapy at 5 mg/kg cisplatin and 10 mg/kg etoposide on day one, and 10 mg/kg etoposide on days two and three, repeated for three weeks, and were allowed to progress until they were moribund[15].

2.4.2: Xenografting of barcoded tumors

H209 and H82 SCLC cell lines were barcoded with the rCMV-GFP-BC retrovirus by spinfection at 940xg for 2 hours. Each barcoded line was allowed to double three times, to ensure overlap in barcodes in the cells sampled and cells used for xenografting. Immediately prior to xenografting, a portion of the cells were removed to generate the single-cell RNA sequencing library ("pre-growth" sample). To make the xenografts, 2500 cells were mixed in a 1:1 ratio with matrigel (Corning Life Sciences) and injected into the hind flank of NOD-SCID mice. Xenografts were measured daily after the tumors were palpable by hand in the hind flank. After xenografts reached 3 cm^3 total volume, mice were euthanized and the xenografts harvested. Tumors were dissected, dissociated, and a portion of the cells were used for single-cell RNA sequencing ("pre-chemotherapy" sample). The remainder of cells were mixed in a 1:1 ratio with matrigel and injected in to a new NOD-SCID mouse. The mice that received these serial xenografts received chemotherapy at 5 mg/kg cisplatin and 10 mg/kg etoposide on day one, and 10 mg/kg etoposide on days two and three, repeated for three weeks after tumors were palpable to generate chemoresistant xenografts [15]. When these mice reached a total tumor burden of 3 cm^3, they were euthanized, and the tumors were dissociated and used to generate single-cell RNA sequencing libraries ("post-chemotherapy" sample).

2.5: Tumor profiling using single cell RNA sequencing

2.5.1: Tissue processing and library preparation

The "pre-growth" samples from the xenograft model were prepared for scRNA-seq according to the 10X Genomics protocols. After xenografts reached a cumulative volume of 3 cm^3, mice were euthanized and tumors dissected. The MACS (Miltenyi Biotec) human tumor dissociation kit was used to digest the tumors, and cells were prepared for scRNA-seq using the 10X Genomics protocols.

After the RPR-Cas9 reached their endpoint, they were euthanized via cervical dislocation and lungs and livers were harvested. Tissues were prepared for flow cytometry according to the 10X Genomics protocols, using the MACS mouse tumor dissociation kit. After dissection, the lungs and livers were sorted for GFP+ cells, which are the barcoded population, and cells were prepared for scRNA-seq according to the 10X Genomics protocol. Tumors were microdissected from three mice by cutting out one tumor lesion, which was sequenced without undergoing

FACS to capture the stromal and microenvironment cells. A 10X Genomics Chromium Controller was used for the library preparation of all tumors.

2.6: Tumor Histology

One lobe of the lung and one lobe of the liver from each of the RPR-Cas9 mice were taken for histology to verify the presence of tumors in this sample. Samples were fixed in 4% paraformaldehyde for 15 minutes and transferred to 30% sucrose for 24-48 hours. The samples were then embedded and cryosectioned before staining. Tumors were stained for GFP and Cas9 to confirm that the barcoding system was successfully induced by the Adenovirus induction of Cas9 expression and AAV9 induction of GFP expression.

2.7: Informatics approach

The scRNA-seq data was initially filtered, trimmed and aligned via 10X Genomics CellRanger program. CutAdapt was used to extract the barcode sequences. Low quality cells were filtered out using Seurat, and clustering and psdudotime analysis was also performed in Seurat. The final data was visualized with Loupe.

2.8: Validation of candidates identified via scRNA-seq

2.8.1: Expression of PAGE5 and GAGE2A in chemotherapy treated cells

An alamar blue assay was used to determine the IC50 value of cisplatin and etoposide in H82 and H209 cell lines. Cells were seeded in 96 well plates and treated with either drug, with concentrations spanning three orders of magnitude. Cellular viability was assessed daily via Alamar Blue. The IC50 for Cisplatin was determined to be 2.876 uM. The IC50 for etoposide was determined to be 0.110 uM. These values were used for the resulting experiments. SCLC-N lines H29 and H82, and SCLC-A lines H1836 and H209 were treated with cisplatin and etoposide at the IC50 values for three days at cycles resembling the in vivo chemotherapy treatment (Cisplatin and etoposide day 1, etoposide only days 2 and 3). Cells were harvested ono days two and three and RNA was extracted following the Trizol (Ambion Biosciences) protocol, and RNA was converted to cDNA using the NEB ProtoScript Reverse Transcriptase Kit. Expression levels of *PAGE5* and *GAGE2A* were quantified via qPCR.

2.8.2: Knockdown of PAGE5 and GAGE2A and response to chemotherapy

shRNAs targeting *PAGE5* and *GAGE2A* (Table 1) were designed using pSicoligoMaker3 (Ventura lab, https://bitbucket.org/theclipper/psicoligomaker3/src/master/). shRNA oligos were cloned in to the lentiviral backbone pSicoR, a gift from Tyler Jacks (Addgene plasmid # 11579 ; http://n2t.net/addgene:11579 ; RRID:Addgene_11579) [63]. The pSicoR-shPAGE5 and pSicoR-shGAGE2A were made in to a second-generation lentivirus using pMD2.G and psPAX2 as packaging plasmids, and concentrated overnight using the TaKaRa Retro-X retroviral concentrator. SCLC-N lines H29 and H82, and SCLC-A lines H209 and H1836 were infected with pSicoR-shPAGE5 or pSicoR-shGAGE2A and sorted by GFP expression using the BD FACS Jazz. Knockdown of *PAGE5* and *GAGE2A* expression was validated in the sorted cells with qPCR. To generate a double-knockdown, the sorted cells were infected with the reciprocal virus and expression of both *PAGE5* and *GAGE2A* was assessed via qPCR. The single- and double- knockdown cells were used to generate xenografts to investigate the dependence of the chemoresistance phenotype on *PAGE5* or *GAGE2A* expression. 150,000 cells were mixed in a 1:1 ratio with GelTrex (Gibco) and injected in to the hindflank of NOD-SCID mice. Due to supply chain disruptions, a switch from Matrigel to GelTrex was necessary, however they both function the same way in providing some extracellular matrix to aid in xenograft injection. Tumors were allowed to grow until a cumulative volume of 3 mm^3 was reached, at which point mice were euthanized and the tumors kept for histology to validate the knockdown of PAGE5 and GAGE2A. Half of the mice were treated with chemotherapy at 5 mg/kg cisplatin and 10 mg/kg etoposide on day one, and 10 mg/kg etoposide on days two and three, repeated for three weeks after tumors were palpable. In culture, the shPAGE5, shGAGE2A, and double knockdown cells were treated with the IC50 value of cisplatin and viability was assessed via Annexin V and propidium iodide staining by flow cytometry (Biolegend APC Annexin V Apoptosis Detection Kit).

2.8.3: Overexpression of PAGE5 and GAGE2A

GAGE2A and *PAGE5* overexpression retroviruses were generated by amplifying the transgenes from SCLC cell lines and cloning them in to the CAG-GFP retroviral backbone. H29, H82, H1836, and H209 cells were transfected with either the rCAG-PAGE5-GFP or rCAG-GAGE2A-GFP vectors by spinfection with concentrated virus at 940xg for two hours. Transduced cells were treated with the IC50 value of cisplatin, etoposide, or cisplatin and etoposide assessed for response to chemotherapy by Annexin V and propodium iodide staining, and efficiency of overexpression assessed by qPCR.

2.8.4: Knockdown of the AP-1 pathway via overexpression of dominant-negative Jun

To inhibit the AP-1 complex, cJun was knocked down by transfection with a dominant-negative cJun construct. This is a common method for inhibiting the formation of the Jun/Fos AP-1 complex [64,65]. pMIEG3-JunDN was a kind gift from Alexander Dent (Addgene plasmid # 40350 ; http://n2t.net/addgene:40350 ; RRID:Addgene_40350) [64]. H82, H29, H1836, and H209 SCLC cell lines were transfected with pMIEG3-JunDN using Lipofectamine 3000. Upon visual GFP detection, cells were sorted using FACS for GFP expressing cells. Cells were seeded in to 6 well plates for a soft agar colony formation assay. Briefly, 0.8% Seaplaque agar (Lonza) was used as a bottom layer and 10,000 cells per well were seeded in 1.2% agar in the top layer. Plates were fed with full RPMI as needed to prevent drying out. 10 days after seeding, colonies were observed by eye and the plates were stained with 0.001% crystal violet for one hour, and plates were photographed. The number of crystal violet colonies stained was quantified with a custom CellProfilier script.

2.9: Analysis of PAGE5 and GAGE2A expression in human SCLC biopsies

2.9.1: Ethics statement

The staining and scoring, and well as storage of the data for all human specimens was approved by the Sanford Health Institutional Review Board.

2.9.2: Staining of human biopsies

Human SCLC biopsies were obtained from the Sanford Health Biobank. Slides were stained with anti-PAGE5 (Invitrogen PA5-50470) or anti-GAGE2A (Aviva Systems ARP64957-P050). Stained slides were scanned using an Apereo AT2 slide scanner. Three independent researchers viewed and scored the scanned slides based on positivity, distribution, and intensity of staining. Positivity was a binary score, with the sample earning a positive score from any singular positive cell. Distribution was scored on a 0-3 scale: 0 for 0-5% of the sample staining positively, 1 was assigned to samples 6-30% positive, 2 for samples 31-60%, and 3 for samples more than 60% positive for PAGE5 or GAGE2A. For staining intensity, a score 0-3 was assigned. 0 for samples with no PAGE5 or GAGE2A staining, 1 for samples with light staining, 2 for samples with moderate staining, and 3 for samples with intense staining.

Name	Forward	Reverse
Retroviral_PolyA	tgtacaagtaagtttaaacAAGCTTctgtgccttctagttgccagc	gaggttgattggtttccatagagcccaccgcatc
Retrivral_Barcode_Oligo	tgtacaagtaagtttaaacaagtttGTACAAGTAANNATCNNGATSSAAANNGGTNNAACNNTGTAAA	gaaggcacagAAGCTTTTACANNGTTNNACCNNTTTSSATCNNGATNNTTACTTGTACaa
AAV9_R26_gRNA	cggccgcacgcgcatgtgagggcc	ttctctgtggtgacaaaaaagcacc
AAV9_R26_LHDR	ttctcaggtaaccacgcggcaggccctcc	ccgctcggtccgcacgtgctagaaagactggagttgcagatcac
AAV9_R26_RHDR	cagtctttctagcacgtgggggatccactagttctagagc	ccgctcggtccgcacagggcatcagatcccattacaga
AAV9_BC_amp	cagtctttctagcacgtgGTGATGGTGAGCAAGGGCG	tagaactagtggatcGTACGACTTGGATCCCTCACTGG
Retroviral_miSeq	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgagctgtacaagtaagtttaaacaagtttGTACAAG	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTaccttccagggtcaaggaagg
AAV9_miSeq	ACACTCTTTCCCTACACGACGCTCTTCCGATCTgcatggacgagctgtacaagg	GACTGGAGTTCAGACGTGTGCTCTTCCGATCTaggctgatcggccgc
GAGE2A_OE_Cloning	ATTCGCTAGCGGATCGCCACCATGAGTTGGCGAGGAAGATCG	CGAGGCGGCCGGATCTTAACACTGTGATTGCTTTTCACCTTCTTCAGGC
PAGE5_OE_Cloning	ATTCGCTAGCGGATCGCCACCATGAGTGAGCATGTAACAAGATCCCA	CGAGCCGGCCGGATCCTATAGTTGCCCTTCACCTGCTT
GAGE2A_qPCR	TGAGTTGGCGAGGAAGATCG	TCCCCTTCTTCAGGTGTTGC
PAGE5_qPCR	TGATGTCAGGGAGGGGACTC	TTGGGGTCTGAACTACCTTCAA
shPAGE5_Oligo_1	ggagaaaagccttgttTGGAACCACCAACTGATAATTTCAAGAGAATTATCAGTTGGTGGTTCCTTTTTTC	ggatcctagtactcgaGAAAAAAGGAACCACCAACTGATAATTCTCTTGAAATTATCAGTTGGTGGTTCCA
shGAGE2A_Oligo_1	ggagaaaagccttgttTGCAGTTCAGTGATGAAGTTTCAAGAGAACTTCATCACTGAACTGCTTTTTTC	ggatcctagtactcgaGAAAAAAGCAGTTCAGTGATGAAGTTCTCTTGAAACTTCATCACTGAACTGCA
AAV9_BC_Oligo	GTGATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGACGGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTACGGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACCCTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAGCAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTGGTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCACAAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAACGGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCCGACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCACTACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTCCTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAGGTACAAGTAAnnATCnnGATssAAAnnGGTnnAACnnTGTAAAACGACGGCCAGTGAGGGATCCAAGTCGTAC	

Table 1: List of primers and oligos used.

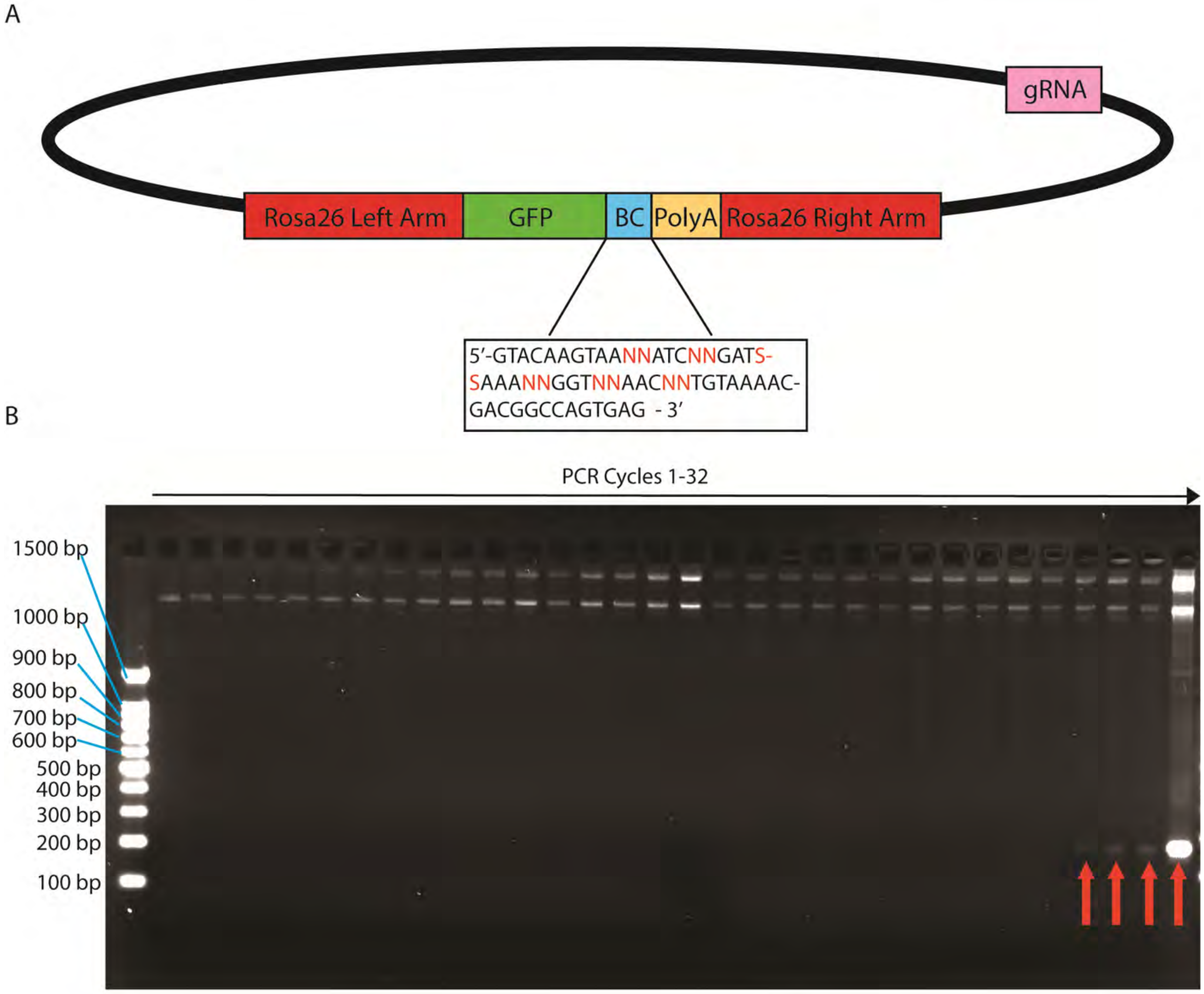
Chapter 3: Generation and Validation of Barcoded Tumors *in situ*

3.1: Validation of the barcoding AAV

The AAV9 construct was designed to deliver all components of the CRISPR-Cas9 homology directed repair (HDR) function, with the exception of Cas9, as it is already expressed in the mouse lungs upon induction with Cre adenovirus. AAV9 has adequate tropism to the lung, which is why it was selected for this aim. The AAV9 contains a guide RNA (gRNA) targeted to the *Rosa26* mouse locus, and delivers the barcode sequence and GFP between *Rosa26* homology arms (Figure 6A). The gRNA will guide Cas9 to the *Rosa26* locus, where Cas9 will make a double-stranded DNA cut. As the cell works to repair the cut, the *Rosa26* homology arms supplied by the AAV9 will be used in the homology directed repair, and the GFP-Barcode cassette will be inserted in to the genome as the repair is complete. As indicated in the methods, prior to being made in to virus, a number of colonies were screened to ensure unique barcodes in a few colonies. The plasmid was sent to the University of Michigan Viral Vector Core to be made in to AAV9.

3.2: Profiling of the barcoding AAV

To understand the diversity in barcodes, I used next-generation targeted amplicon sequencing of just the barcode region. In order to avoid introducing errors in the PCR steps of the library preparation for the amplicon sequencing, I worked to amplify the barcode insert with the lowest number of PCR cycles possible (Figure 6B). The barcodes from the plasmid pool and the AAV9 pool were sequenced via targeted amplicon sequencing at Genewiz. To ensure the most accurate representation of the true number of barcodes in the population, the number of PCR cycles determined in Figure 7B, along with an estimation of PCR error based on the error rate in the constant regions of the barcode was used to group together barcodes that may have seemed unique but were actually a result of PCR error. By using the PCR backtracking, combined with Chao2 modeling of diversity using the plasmid pool and viral pool as two unique samples, an overall diversity of around 1,300 barcodes was determined (Figure 6C).



C

Read	Estimated diversity	Lower bound	Upper bound
Read 1	1251.236	1031.620	31249.121
Read 2	1483.063	1139.552	48062.841

Figure 6: Validation of the AAV9 barcoding construct
A: The construct contains homology arms to the *Rosa26* locus flanking a GFP, the barcode sequence (inset), and a polyA sequence. This will insert the entire GFP barcoding cassette during HDR after Cas9-directed cutting at *Rosa26.* The plasmid also contains a gRNA directed to *Rosa26* to guide the Cas9 to the appropriate site to make the cut. **B**: Determination of the optimal number of PCR cycles prior to targeted amplicon sequencing. The barcode sequence was amplified from the AAV plasmid, and a small amount of sample was removed at each PCR cycle and run on an agarose gel to determine the cycle number at which a band corresponding to the size of the barcode amplicon appears. As indicated by the arrows, a band corresponding to the size of the barcode begins to appear at 29 PCR cycles, and DNA sufficient for sequencing is able to be extracted. **C**: Diversity of the barcoding vector was determined with miSeq and Chao2 modeling.

3.3: Generation and validation of the TKO-Cas9 mouse model

Our SCLC mouse model is driven by Cre-based deletion of the tumor suppressors *Rb, p53,* and *p130* (RPR model). The RPR mouse model has been used many times and is a reliable model of SCLC[11,61]. I bred it with an $H11^{lox\text{-}stop\text{-}lox\text{-}Cas9}$ mouse to generate the RPR-Cas9 ($Rb^{lox/lox}$, $p53^{lox/lox}$, $p130^{lox/lox}$, $H11^{lox\text{-}stop\text{-}lox\text{-}Cas9}$) model. The mice rapidly develop tumors after intratracheal injection with an Ad-CMV-Cre adenovirus. To determine the optimal concentration of adenovirus, I injected mice with varying concentrations of the Ad-CMV-Cre virus and stained their lungs with antibodies against both Cre (to determine viral uptake) and Cas9 (to determine functional output of Cre recombination). Antibodies against Cas9 were not very good, so I used Cre immunostaining as the threshold by which to select the concentration of Ad-CMV-Cre (Figure 7).

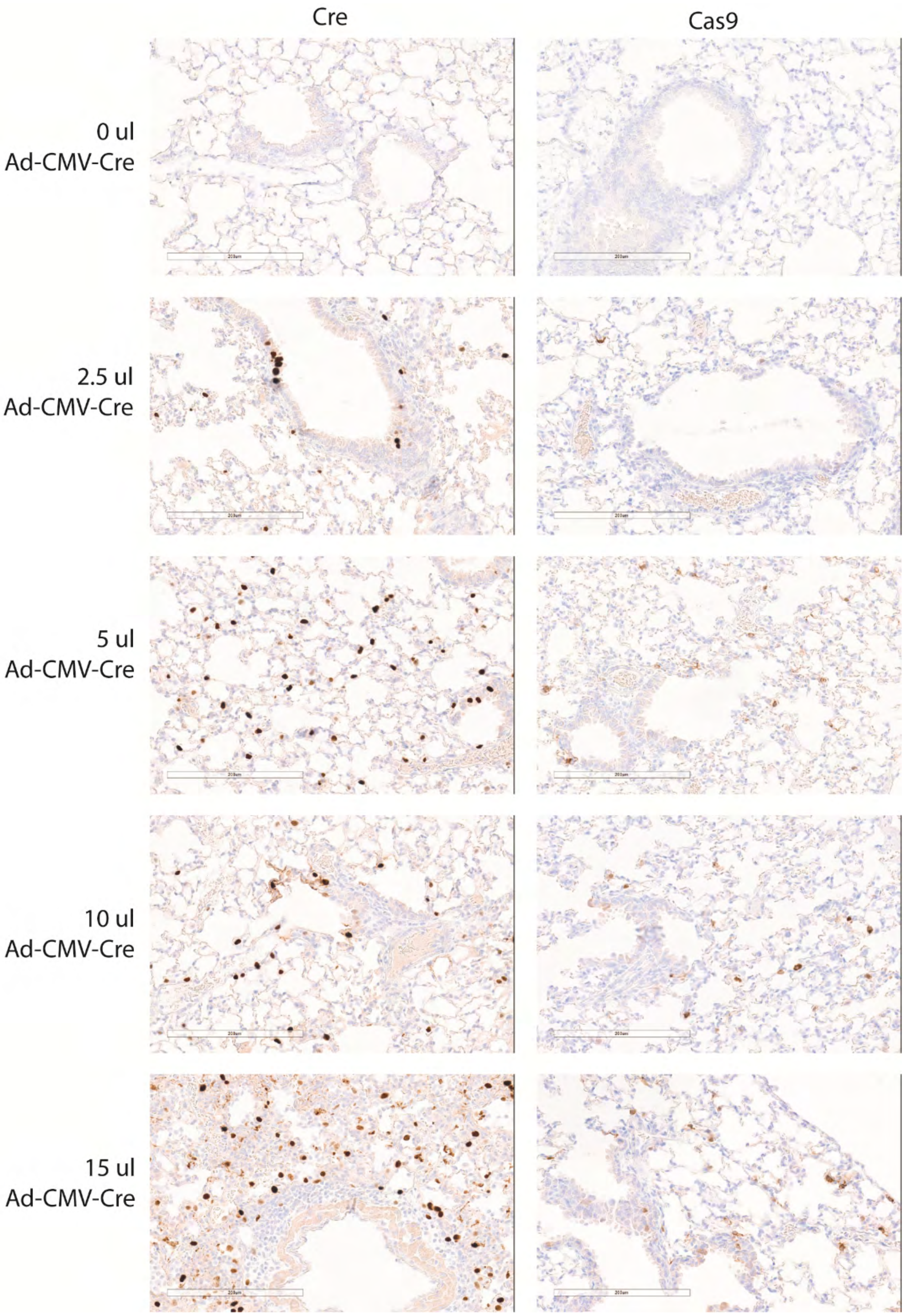
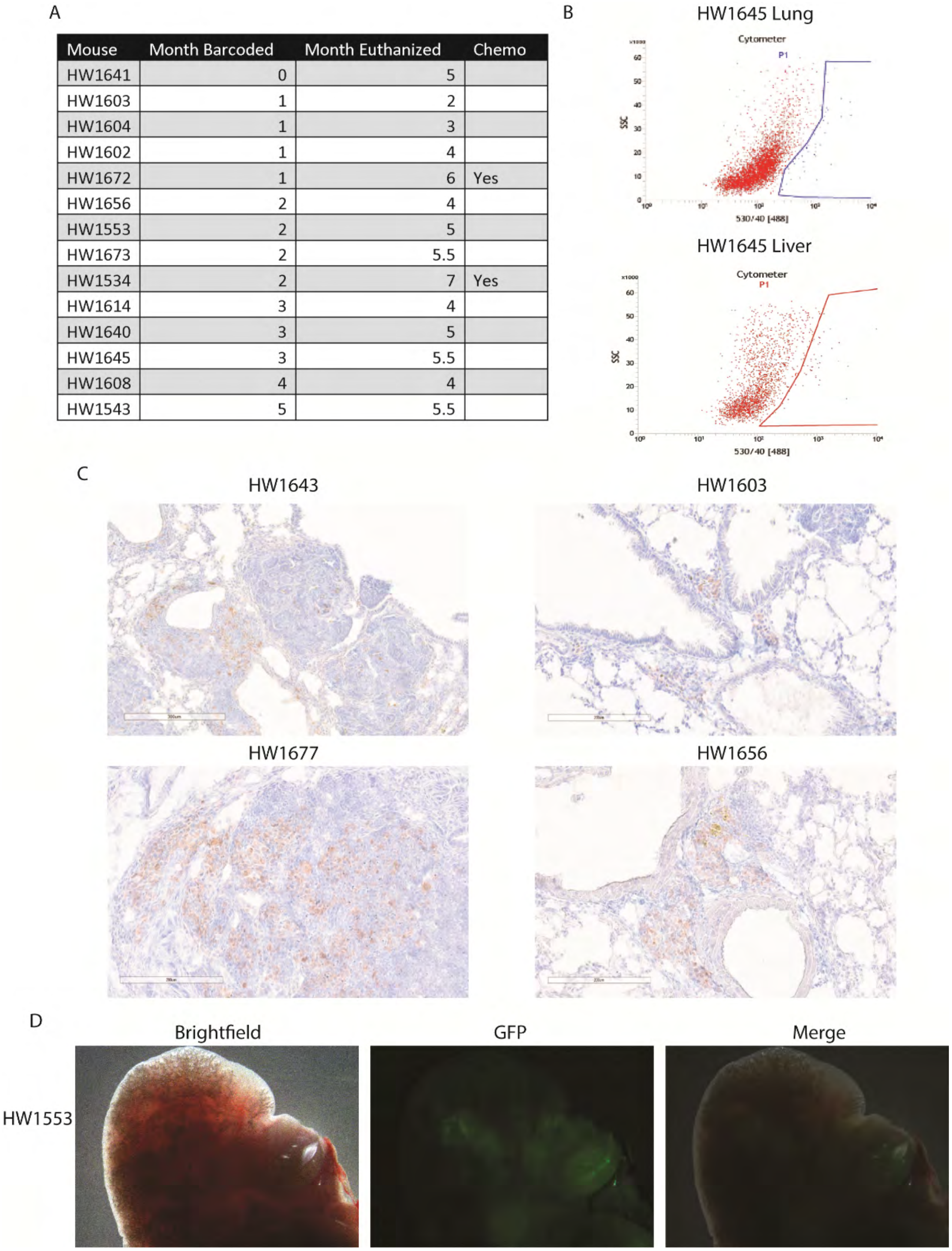


Figure 7: Lung histology to titer the Ad-CMV-Cre virus
Varying concentrations of Ad-CMV-Cre were injected intratracheally in to mouse lungs, and histology was performed to to evaluate staining for Cre (left), and the functional output Cas9 (right). 10 ul of virus is sufficient to initiate expression of Cas9 in the lungs.

Tumors generally start from a single cell and clonal diversity then develops over the course of the tumor growth. If tumors are barcoded too early, they could be barcoded at the single cell stage, and clonal diversity would be lost, as all cells within the tumor will share the same barcode. Conversely, if tumors are barcoded too late, the heterogeneity will have already formed, and clonal dynamics will not be able to be understood, as clones likely sharing the same lineage will receive different barcodes. Since the ideal timing for barcoding is unknown, I designed a matrix system for barcoding and analysis of tumors (Figure 8A). This matrix will allow us to barcode and analyze tumors from early tumor formation stages through endpoint-stage disease. By barcoding and harvesting tumors in the matrix schedule, all combinations of barcoding and harvesting are captured and I will capture the ideal timeline for tumor heterogeneity.

To validate the barcoding AAV9-r26-GFP-BC virus, mice were given Ad-CMV-Cre, followed by AAV9-r26-GFP-BC two days later, and were euthanized after another two days. Lungs were stained for GFP. After tumor initiation and barcoding, mice were allowed to progress to their scheduled endpoint, or until moribund, whichever came first. Two mice were given chemotherapy at five months post-tumor initiation, and these mice were allowed to progress to moribund (Figure 8A). All RPR-Cas9 mice that were euthanized at five months have extensive tumor burden, so this timepoint was selected for chemotherapy treatment initiation to mimic the general clinical progression, as the majority of patients present with extensive-stage disease. At their endpoint, lungs and livers were harvested from all mice, and flow cytometry was performed to isolate the GFP+ barcoded tumor cells before scRNA-seq (Figure 8B, C, D).



Mouse	Month Barcoded	Month Euthanized	Chemo
HW1641	0	5	
HW1603	1	2	
HW1604	1	3	
HW1602	1	4	
HW1672	1	6	Yes
HW1656	2	4	
HW1553	2	5	
HW1673	2	5.5	
HW1534	2	7	Yes
HW1614	3	4	
HW1640	3	5	
HW1645	3	5.5	
HW1608	4	4	
HW1543	5	5.5	

Figure 8: Validation of barcoded tumors
A: Timeline for tumor barcoding and isolation. Tumors were barcoded at one month intervals following tumor initiation, and tissues were harvested at one month intervals following barcoding. A subset of the mice received chemotherapy. **B:** Example FACS plots from a lung and liver sorted for GFP+ barcoded tumor cells. **C:** Example histology showing cells stained with an antibody against GFP in four tumors barcoded and harvested at various times. **D:** Repressive images of a lobe of the lung showing GFP+ tumor lesions under a dissecting microscope.

Chapter 4: Single-cell RNA Sequencing of SCLC Tumors Barcoded *in situ* Identifies Genetic Signatures for Tumorigenesis

4.1: Two distinct transcriptomic signatures arise during tumorigenesis

After tumors were harvested, flow cytometry was performed to isolate GFP+ cells, and they were used for scRNA-seq. Data was trimmed and aligned to the mouse genome using CellRanger. The resulting data was then used in Seurat to filter out low-quality cells, attempt to extract barcode data, and correct for cell cycle genes. Clustering was performed in Seurat to identify unique cellular populations.The scRNA-seq analysis is shown in Figures 9-14.

Upon analysis of the scRNA-seq data, no barcodes were detectable in the tumors barcoded *in situ.* Despite having tumors that are immuno-reactive to antibodies against GFP (Figure 8C), and the detection of GFP+ cells via flow cytometry (Figure 8D), the depth of scRNA-seq in this case was not sufficient to pick up reads from the GFP and barcode. Although barcodes were not detected in these samples, due to the timewise design of the animal studies, information on tumor evolution can still be gained. We observe the majority of cells sequenced are indeed SCLC cells, characterized by classic SCLC neuroendocrine markers. Other cell types that can be identified are myeloid, club, alveolar, ciliated cells, and T and B immune cells (Figure 15B). When evaluating the cell cycle composition of the cells within the tumor populations, the proportion of cells in G1 decreases with each subsequent month that tumors are allowed to form, and the percentage of cells in G2 or metaphase increases significantly after five months of tumor development (Figure 15C) We observe an "early" tumor signature that arises in the tumors isolated after two months of development, and a population of cells maintains the early signature through later tumor development, up to 5.5 months after tumor initiation. In the later months of tumor development, a "late" signature emerges, which eventually is responsible for the majority of the tumor in the longest developed tumors (Figure 15D, E). The gene signatures that characterize the early and late tumor populations differ at several gene “modules” but particularly module five (Figure 15F). Module five is notably comprised of the known SCLC regulators *Myc* and *Hes,* and also members of the AP-1 network.

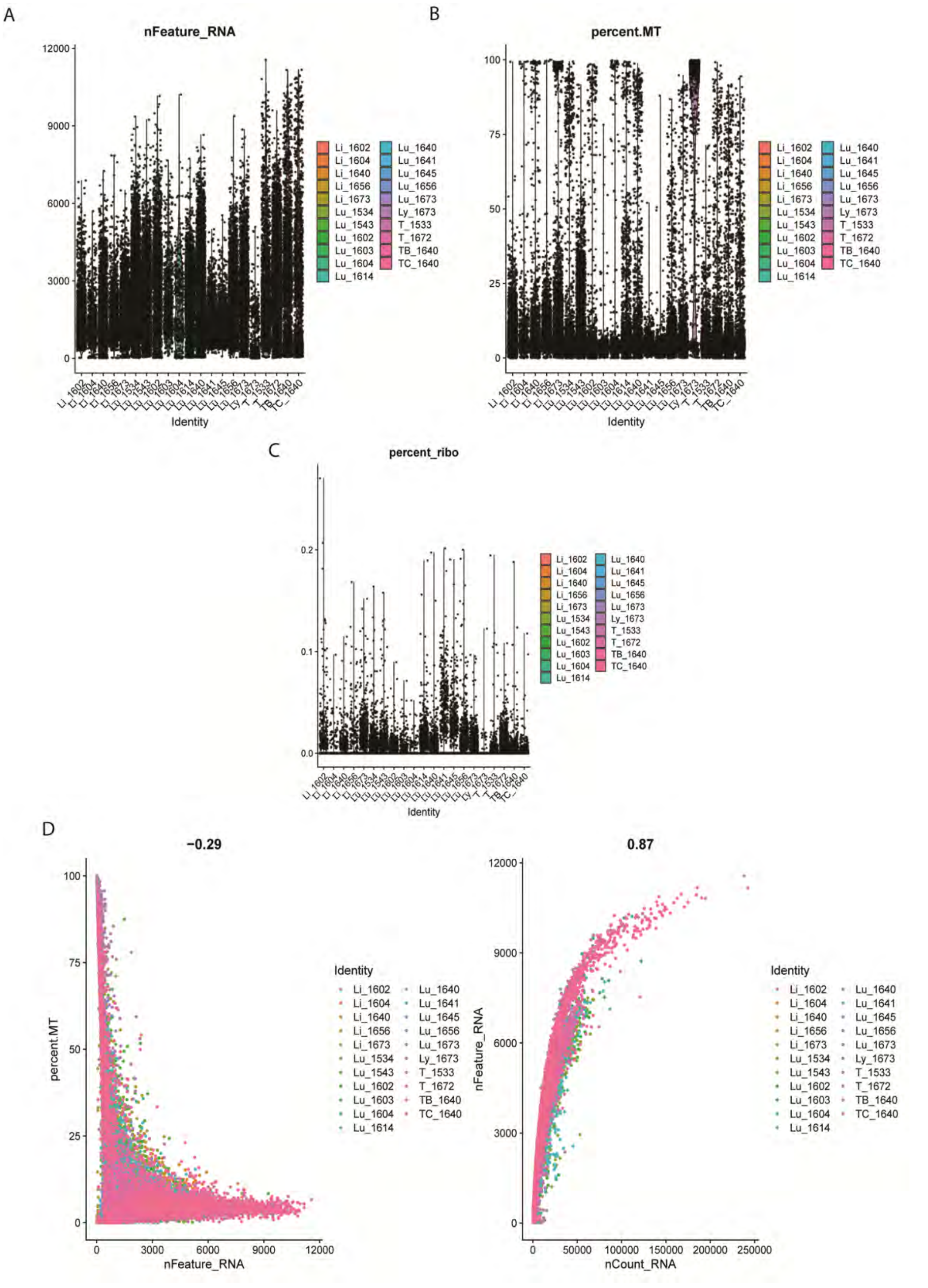


Figure 9: Pre-trimming scRNA-seq, in situ

A: Number of genes detected per sample, for all samples including lung (Lu), liver (Li) and lymph node (Ly). **B**: Percentage of reads corresponding to mitochondrial genes in all *in situ* samples. **C**: Percentage of reads corresponding to ribosomal genes. **D:** Correlation between percent mitochondrial reads and number of genes (left), or number of genes and number of features (right). Cells with a high percentage mitochondrial genes and low number of genes were filtered out as dead cells. Using the features vs reads plot, doublets were filtered.

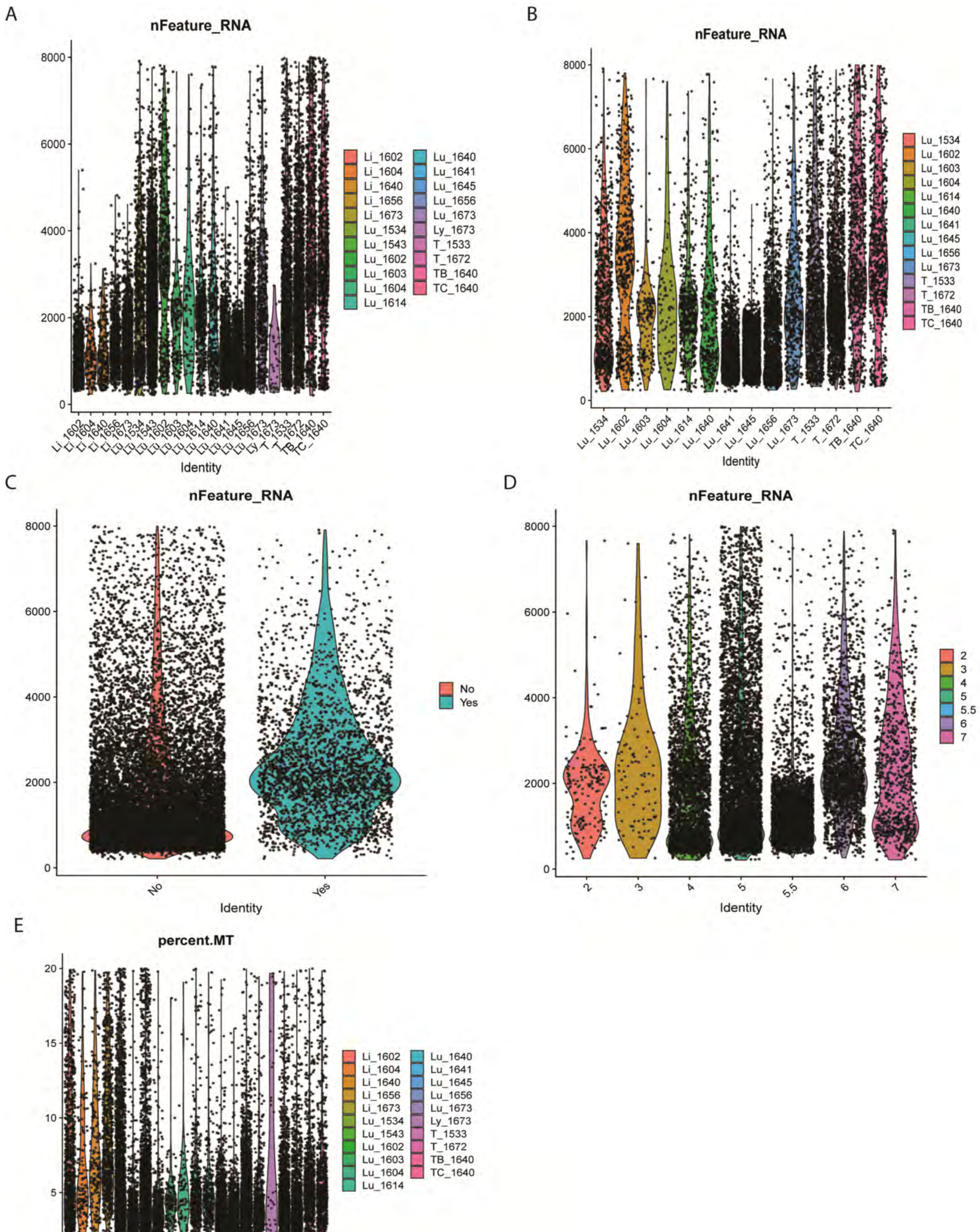


Figure 10: Post-trimming scRNA-seq, in situ

A: Number of genes sequenced per cell after trimming. **B:** Number of genes sequenced per cell after sequencing for only the lung samples. **C:** Number of genes sequenced by cell stratified by whether or not the animal received chemotherapy treatment. “No” indicates no chemotherapy treatment, and “yes” indicates chemotherapy treatment. **D:** Number of genes sequenced per cell stratified by the month of tumor harvest. The early timepoints (months 2 and 3) are lower in reads than the later timepoints. **E:** Post-trimming percentage of reads that correspond to mitochondrial genes.

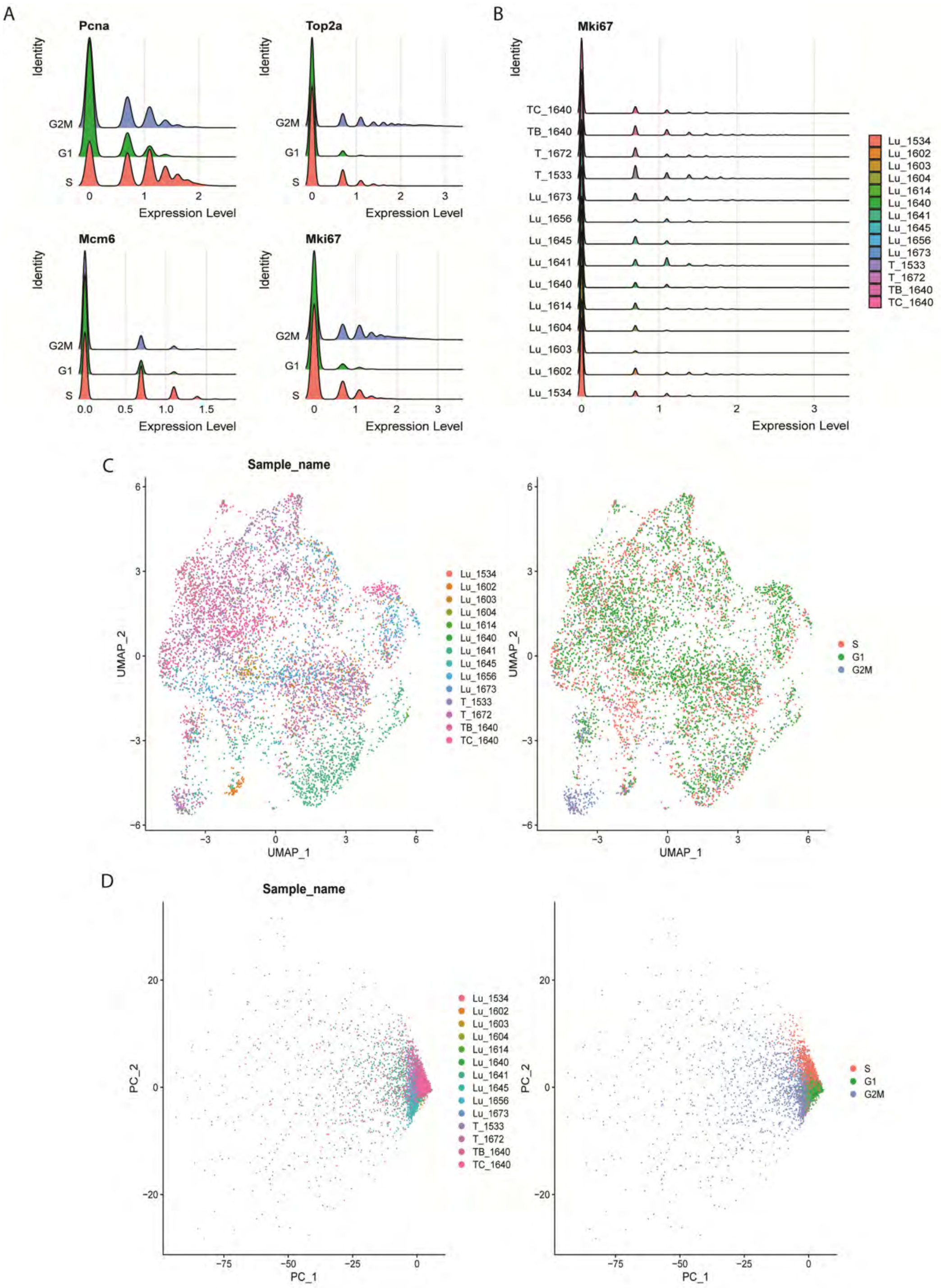


Figure 11: Cell cycle correction, in situ
A: Cell cycle marker genes pre-set by Seurat show the distribution of cells corresponding to each stage of the cell cycle. **B:** Amount of KI-67 expressed in each sample, indicating these cells are cycling. **C:** UMAP showing the distribution of lung samples (left), and the cells that are in each phase of the cell cycle (right). **D:** Cell cycle state plotted by principal component. Left – PC plot showing the distribution of the lung samples. Right – PC plot indicating which cells are in each phase of the cell cycle.

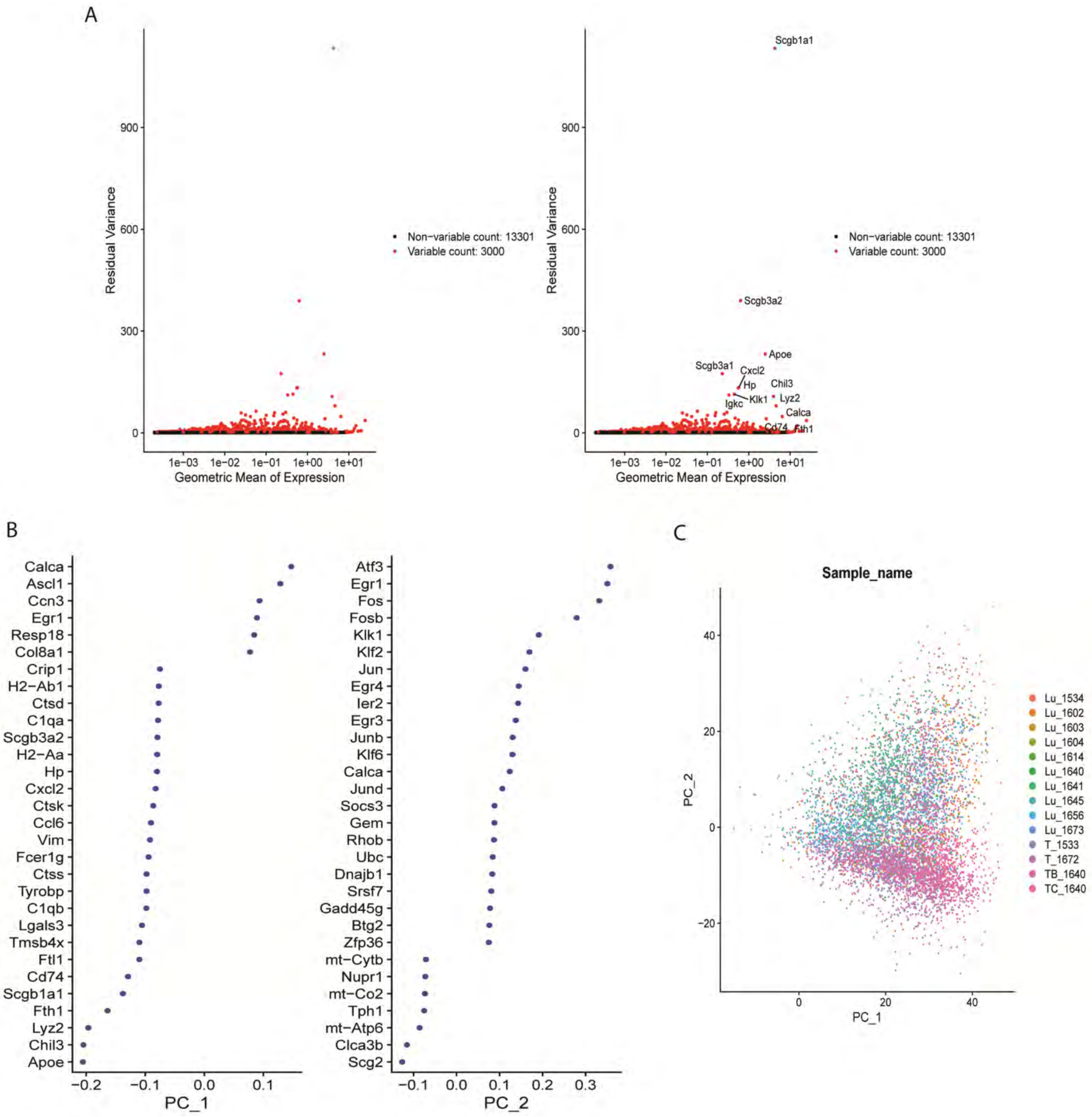


Figure 12: Stratification of cells based on two principal components, in situ
A: Top ten genes identified with Seurat analysis. **B:** DEG identified by PC-1 (left) and PC-2 (left). Top differentiated genes include Ascl1, Egr1, and members of the AP-1 network. **C:** Distribution of cells when split on two principal components.

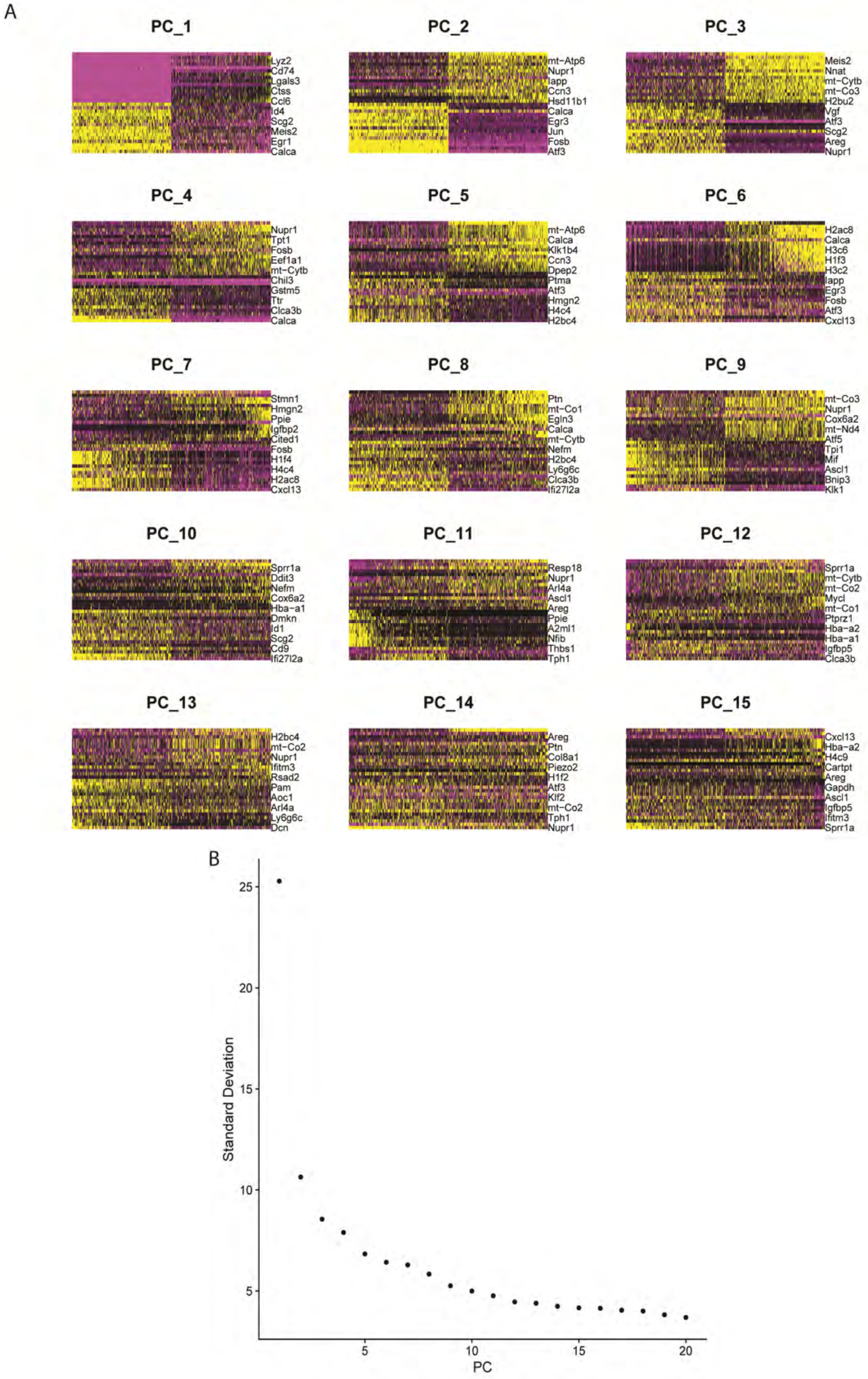


Figure 13: Increasing dimensionality leads to decreased standard deviation
A: Heatmaps of increasing dimensions. **B:** Elbow plot indicating the decrease of standard deviation with each increased dimension.

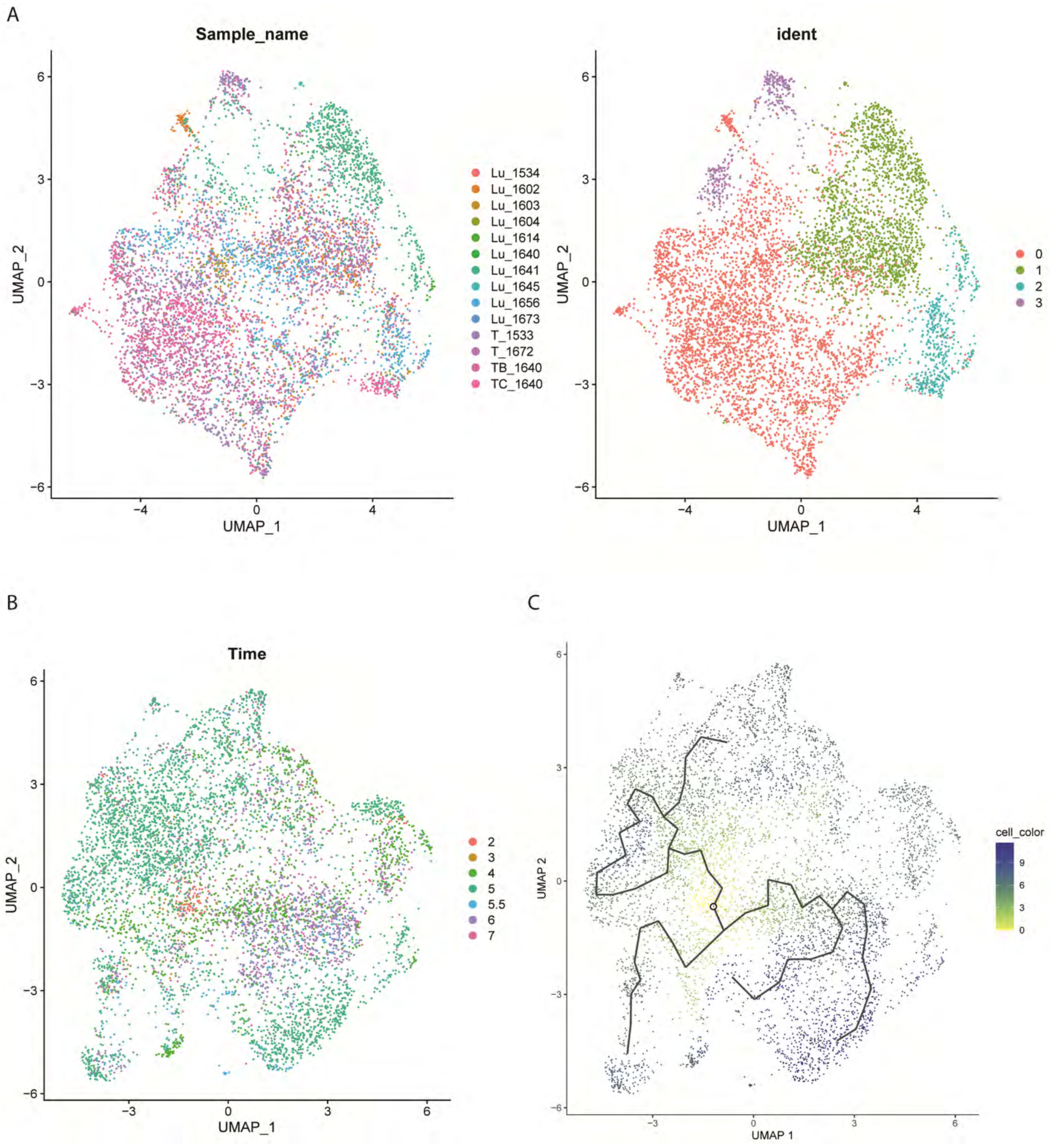


Figure 14: UMAP and clustering to generate a pseudotime trajectory, in situ samples
A: Left – UMAP of all lung samples. Left – clustering used for downstream analysis. **B:** UMAP with the distribution of cells color-coded with the time of tumor harvest in months. **C:** Pseudotime trajectory showing multiple, divergent routes of tumor evolution.

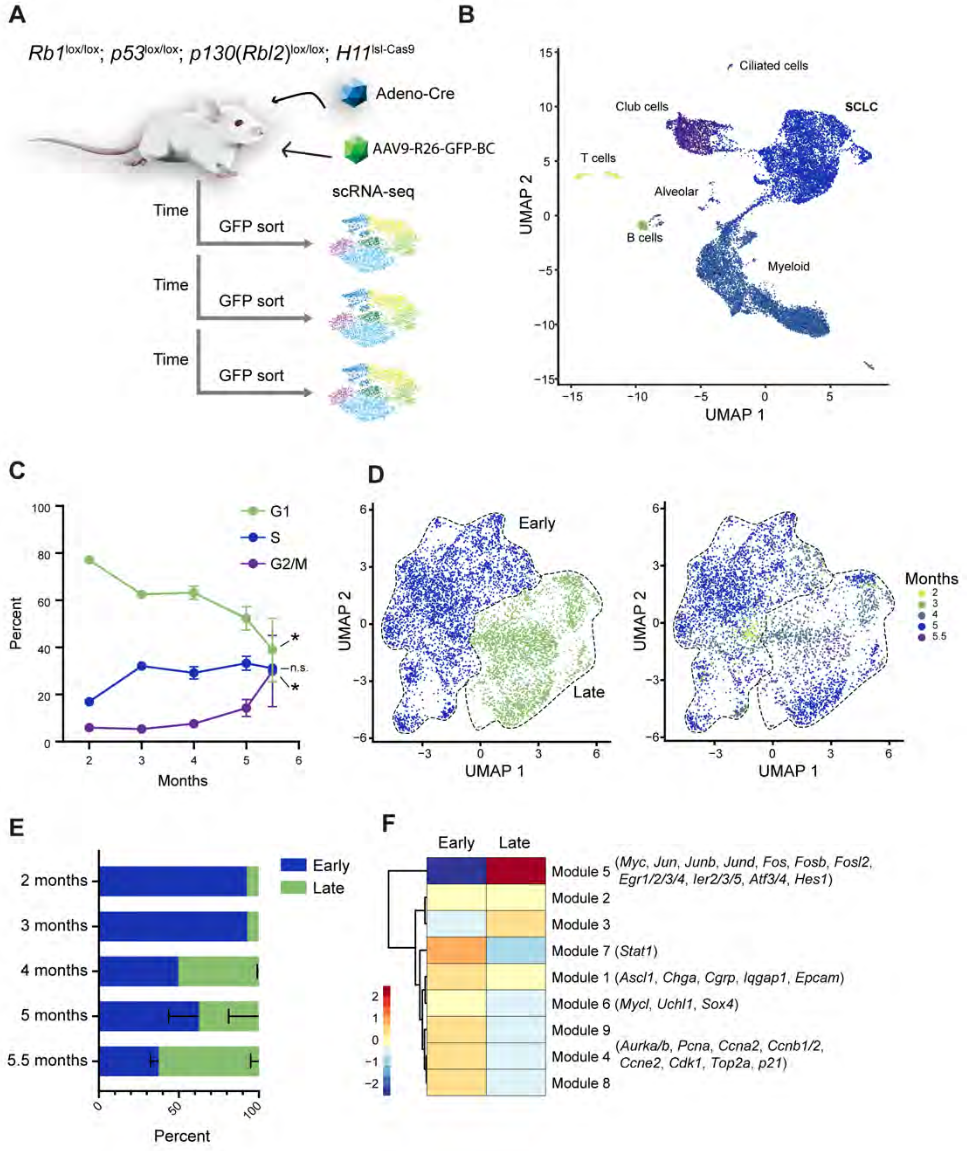


Figure 15: Single cell RNA sequencing of in situ tumors reveals two distinct tumor populations
A: Schematic overview of the *in situ* barcoding approach. Tumors are initiated with Ad-CMV-Cre, and are barcoded with the AAV9-R26-GFP-BC virus at one month intervals post-initiation. At one-month intervals following barcoding, tumors were harvested and underwent scRNA-seq. **B:** scRNA-seq detected many cell types in the lung, including SCLC tumor cells. **C:** At each month post-tumor initiation, the proportion of cells in G1 significantly decreases and the proportion of cells in G2/M significantly increases. **D:** Tumor cells stratify in to two populations – “early” and “late” (left). The early population is predominant in the tumors collected at two and three months post-initiation, while the late population arises in the tumors harvested later. **E:** Proportion of cells at each tumor harvest timepoint that correspond to the early or late clusters. Over time, the majority of the tumor is comprised of the late population. **F:** Gene modules differentiate the early and late tumor clusters. Eight gene modules can describe the transcriptomic differences between the early and late tumor clusters. Particularly of note is module 4, which contains *Myc* and *Hes,* as well as members of the AP-1 network.

4.2: The role of the AP-1 network in mediating tumorigenesis

The AP-1 network was frequently identified in the scRNA-seq data from early, middle, and late tumors, and cells high in the AP-1 network signatures were frequently found in both the "early" and "late" tumor clusters, but at particularly high levels in the late cluster (Figure 16A, B). The AP-1 network has been found to be responsible for a number of tumor hallmarks including growth, resistance to therapy, and angiogenesis, and expression has been implicated in many tumor types[65-76]. Given the strong links to tumorigenesis in other cancer types, I sought to understand the implications of AP-1 network activation in SCLC. cJun is one of the most common components of the AP-1 network, and the majority of network functions can be inhibited via knockdown of cJun using a dominant-negative Jun construct (JUNDN)[64,65,68]. Jun was found to be highly expressed in this study, particularly in the late tumor cluster (Figure 16B). I transfected the four SCLC lines (H29, H82, H209, H1836) with the JUNDN construct pMIEG3-JunDN, and did FACS to isolate the GFP+ cells that had been successfully transfected. Due to poor expression in the H209 and H1836 cell lines, only the H29 and H82 cell lines were successfully transduced and will be used for downstream crystal violet analysis.

The resulting cells were used for a soft agar colony forming assay to assess the capability of cells with disruption of AP-1 to form colonies from a single cell suspension. JunDN cells formed significantly fewer colonies than wild-type cells did (Figure 16C, D). Disruption of the AP-1 complex by knockdown of cJun inhibits colony formation, indicating that the AP-1 complex is important in tumor formation in SCLC. Validating these results in SCLC-A cell lines, which are notoriously difficult to transfect, would be a beneficial route of follow-up. The cells from the tumors treated with chemotherapy cluster predominately with the late cell cluster and are high in AP-1 network signatures (Figure 16E).

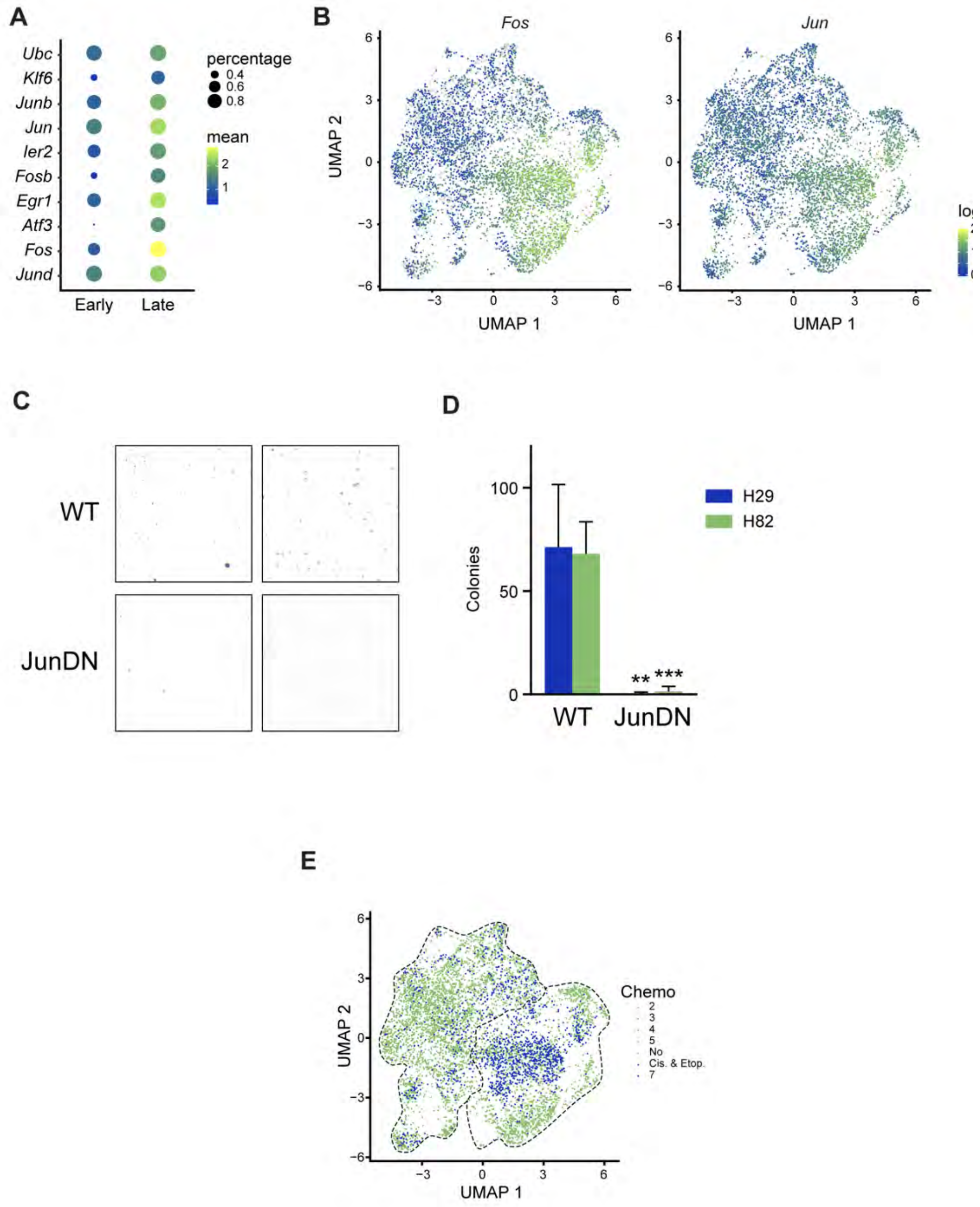
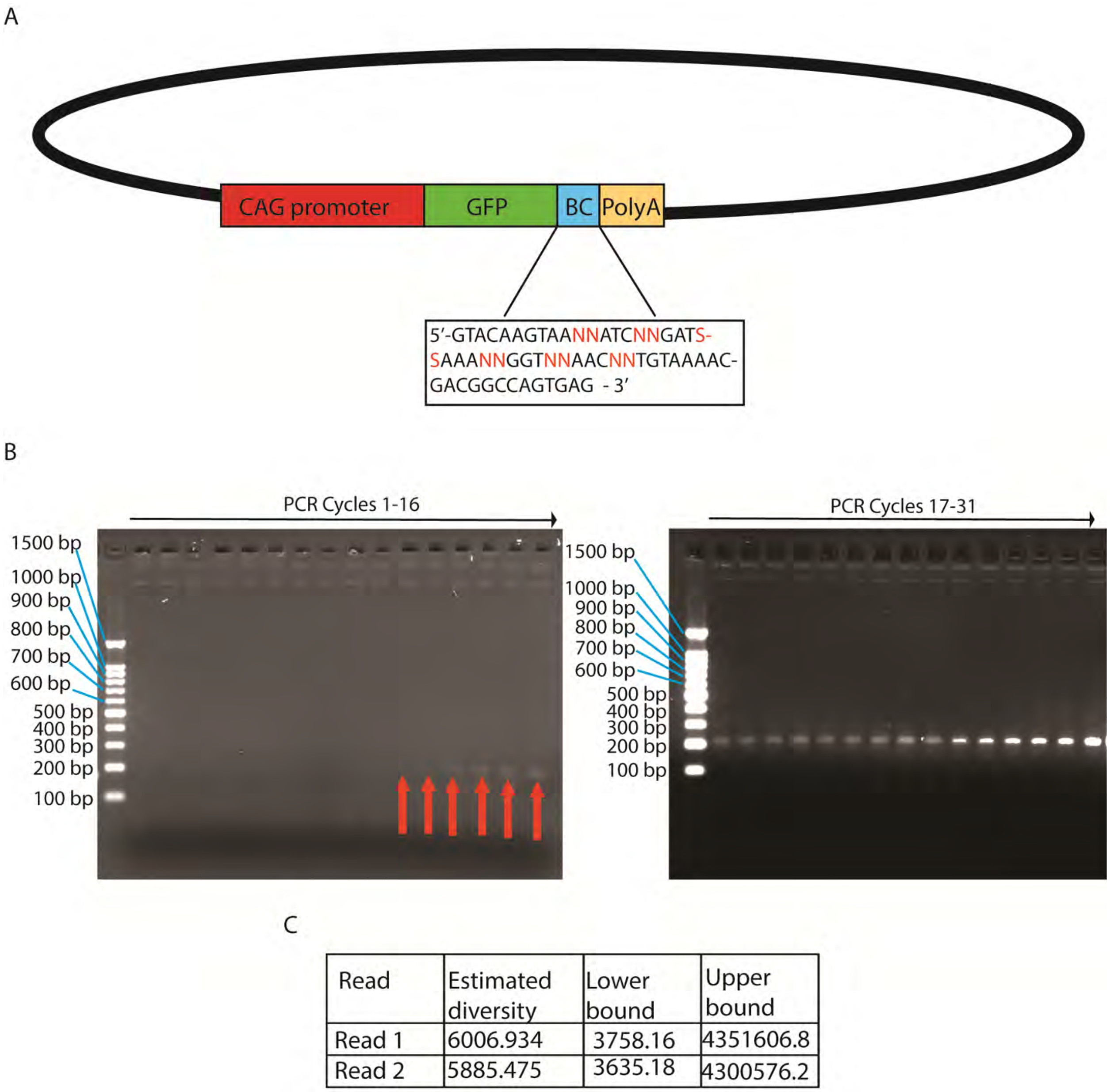


Figure 16: The AP-1 network is required for tumorigenesis.
A: Relative expression of genes belonging to the AP-1 network that are highly expressed in both the early and late tumor clusters, but to a higher degree in the late tumor cluster. **B:** UMAPs showing expression of *Fos* (left) and *Jun* (right), two critical members of the AP-1 complex. The expression of *Fos* and *Jun* is high in the sequenced tumor cells, and is particularly high in the cluster corresponding to the late population. **C:** Representative images of crystal violet staining after AP-1 inhibition due to *Jun* knockdown in SCLC cell lines that were used for a soft agar colony forming assay. The cells with AP-1 disruption formed significantly fewer colonies than the wild-type cells. **D:** Quantification of the soft agar colony forming assay shows significantly fewer colonies in the AP-1 disrupted cells. **E:** UMAP of all sequenced tumor cells, including those that received chemotherapy. The majority of the cells from the chemotherapy treated tumors cluster with the late tumor cluster.

Chapter 5: Generation and Validation of Barcoded SCLC Xenografts

5.1: Synthesis and profiling of the barcoding retrovirus

The retrovirus barcoding construct was designed to contain a CAG promoter to ensure expression regardless of the site of viral insertion, followed by a GFP, the barcode sequence, and finally a polyA sequence to give the highest chance of sequencing the barcode with the 10X Genomics 3' capture technology (Figure 17A). After cloning, the rGFP-BC plasmid was profiled for diversity of barcodes using targeted amplicon sequencing. To avoid erroneously introducing errors in the barcodes, the minimal number of PCR cycles needed to amplify the barcode was determined by removing a portion of the sample after each PCR cycle and running a gel to screen for the lowest number of cycles required to get a band that produces sufficient quantity of DNA for sequencing (Figure 17B). After determining the optimal number of PCR cycles, the plasmid pool and viral pool were sequenced to determine barcode diversity via targeted amplicon sequencing. After targeted amplicon sequencing, the same PCR error rate correction and Chao2 modeling was performed as with the AAV barcoding vector, and the estimated diversity of barcodes is roughly 6,000 unique barcodes (Figure 17C).



Read	Estimated diversity	Lower bound	Upper bound
Read 1	6006.934	3758.16	4351606.8
Read 2	5885.475	3635.18	4300576.2

Figure 17: Generation and validation of the barcoding retrovirus
A: The retroviral barcoding constructs contains a CAG promoter, GFP, and the barcode sequence with a 3' polyA tail. **B:** Determination of the optimal number of PCR cycles to amplify the barcode sequence for targeted amplicon sequencing. The barcode sequence was amplified via PCR, and a subset of the sample was removed each cycle and run on a gel to identify at which PCR cycle a band corresponding to size of the barcode amplicon would appear. As indicated by red arrows, a band is faintly visible at 11 PCR cycles, and 16 PCR cycles is sufficient to obtain DNA for sequencing. **C**: Diversity of the barcoding retrovirus was determined via miSeq and Chao2 modeling, and was determined to be around 6000 unique barcodes.

5.2: Profiling of barcode diversity across doubling times in cells

Before generating xenografts, a subset of the cells are collected to serve as a “pre-injection” sample, however, since each cell should receive a unique barcode, I want to ensure that the barcodes captured in the “pre-injection” sample match the barcodes that are in the xenografted cells, so that they are able to be traced back to their starting population using the barcode sequence. In order to ensure the barcodes captured in the pre-injection sample and the xenograft have sufficient overlap, I set up a doubling time experiment, in which cells would be barcoded, and at each doubling, the cells harvested and split in two. The barcodes of the two independent samples are then be profiled with targeted amplicon sequencing, and the overlap in barcodes in the two halves of the same initial sample quantified. In this experiment, the two halves of the initial barcoded cell pool represent the pre-injected sample and the injected xenograft. By assessing the barcode overlap over four doublings, the optimal number of doublings for sufficient overlap will be identified. For the four SCLC lines, H29, H82, H209, and H1836, the normal doubling time is not known, so I seeded a known number of cells and every 24 hours counted the number of cells in the sample to calculate the doubling time for these four lines (Figure 18A). Since only the H82 (SCLC-N) and H209 (SCLC-A) cell lines are being used to make xenografts, these are the lines that were used for the barcode overlap experiment. Cells were barcoded in culture and at each doubling, as determined in Figure 19A, one well was harvested, split in half (Figure 18B). After harvesting and splitting the cells, the RNA was extracted and the barcodes were amplified using the minimal number of PCR cycles to amplify just the barcode region, as in Figure 18B, and the barcode region was sequenced with Illumina miSeq (Figure 18C). An R script was generated to determine the percent overlap in barcodes and to model the percent overlap if the experiment was repeated 1000 times (Figure 18D). As determined by both the sequencing and the modeling, the percent overlap between two halves starts relitivley high at about 60% and increases over time before plateauing. Additionally, because one cell population will be used to make four xenografts, an R script was used to determine the percent overlap in barcodes if one sample was split in half (pre-injection sample), and the other half was split in to four (four xenografts). The overlap between two individual xenografts (Figure 18E) and the overlap between the pre-injection sample and individual xenografts (Figure 18F) was determined. From the sequencing and modeling of the barcode overlap, three doublings appears to be the most optimal to maximize barcode overlap between the pre-injection sample and the xenografts, and to minimize the barcode overlap between xenografts, so that they may serve as biological replicates. To further assure that three doublings is sufficient to observe overlap, I validated the simulated experiment by performing it using SCLC cells in culture. Cells were barcoded in culture and after three doublings, 15,500 cells were used as the "pre-injection sample" and four samples of 2,500 cells each served as a xenograft, since 15,500 cells will be sequenced pre-injection and 2,500 cells are used to make each xenograft (Figure 18B). Based on the overlap in barcodes observed in the targeted amplicon sequencing, it is clear that three doublings is sufficient to achieve overlap in barcodes between the pre-injection sample and injected xenografts (Figure 18G).

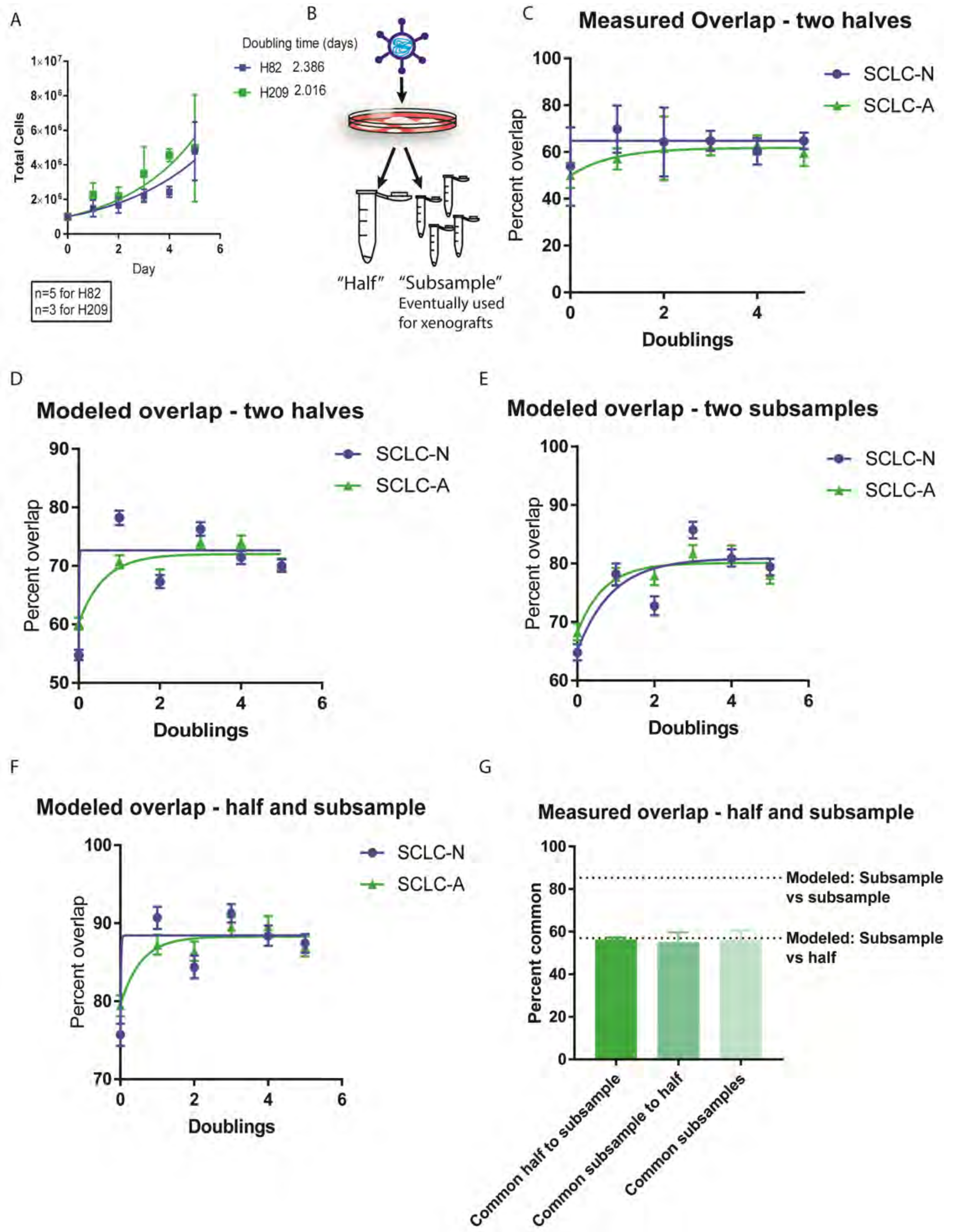


Figure 18: Sequencing and modeling of barcodes in cells infected with the barcoding retrovirus.
A: Doubling time of two SCLC cell lines, H82 (SCLC-N) and H209 (SCLC-A). Cells were seeded at a known concentration and counted daily. The doubling time is determined as 2-3 days. **B:** Schematic of the doubling time experiment. Cells in culture are barcoded and at each doubling, a well is harvested, split in half, and barcodes sequenced. If one of the halves is split in to four independent samples, they are indicated as subsamples. **C:** Overlap in barcode sequences in two halves at each doubling. **D:** Modeled overlap in the barcode sequences over doubling times. The modeling was simulated 1,000 times. **E:** Modeled overlap between two subsamples over five doublings. **F:** Modeled overlap between one subsample and the remaining half of the well. **G:** Overlap in barcodes at three doublings in the amount of cells actually used for xenografting ("subsample") and pre-injection scRNA-seq ("half").

5.3: Generation of xenografts and chemoresistant xenografts

After validating and profiling the barcodes, xenografts were generated by barcoding H209 (SCLC-A) or H82 (SCLC-N) cells in culture, allowing them to double three times, and injecting 2,500 cells per xenograft in 1:1 matrigel in to the hind flank of immunocompromised mice. After palpable, the tumors were measured daily (Figure 19A, B). As expected, the SCLC-N H82 xenografts (Figure 19B) grew much more rapidly than the SCLC-A H209 xenografts (Figure 19A). Clinically, patients that have SCLC-N tumors do more poorly, and SCLC-N subtype is most often associated with chemoresistance, so more aggressive growth behavior from the SCLC-N xenografts is logical. After the chemo-naïve tumors had reached their endpoint, they were dissected and 15,500 cells were used for scRNA-seq, while the rest were injected in to the hind flank of a new mouse as a serial xenograft, which received cisplatin and etoposide. Again, the SCLC-N xenografts grew more aggressively under chemotherapeutic pressure than the SCLC-A xenografts (Figure 19C). There was a response to chemotherapy in a subset of the SCLC-A xenografts, but all tumors ultimately regrew as chemoresistant tumors. All chemoresistant tumors were dissected and subject to scRNA-seq. Upon dissection the tumors are GFP+ under a fluorescent dissecting microscope, indicating that they are indeed barcoded, since the barcode is fused to a GFP (Figure 19D).

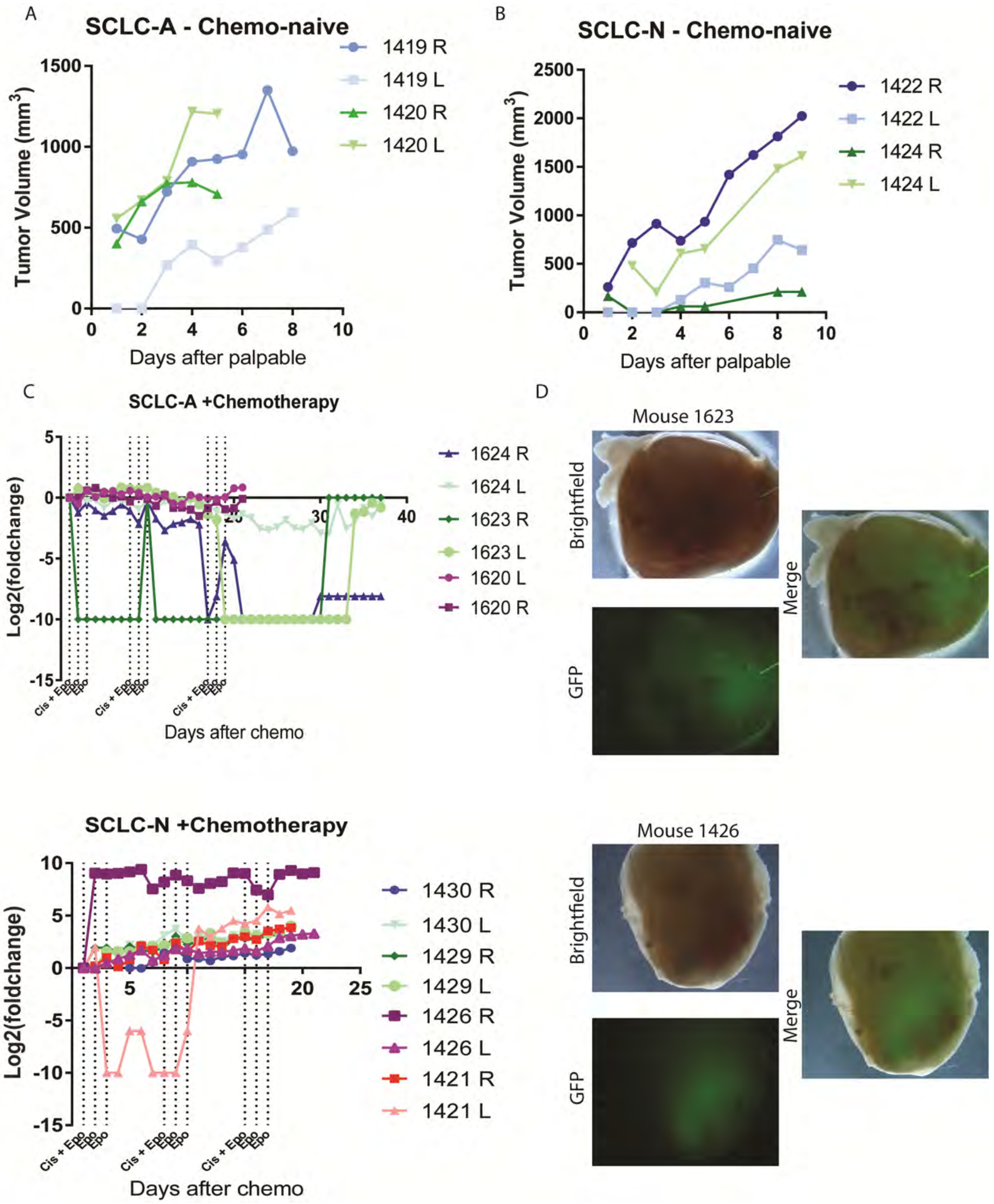


Figure 19: Generation and growth of barcoded and chemoresistant xenografts.
A: Four xenografts were generated from barcoded SCLC-A cells and were allowed to grow until the size threshold was reached. **B:** Growth curves from the four barcoded SCLC-N Xenografts. Xenografts were generated and allowed to grow until the size threshold. **C:** Xenograft growth under chemotherapy. Barcoded xenografts were serially injected in to new mice, which received three weeks of chemotherapy treatment. SCLC-A tumors (top) were somewhat responsive to chemotherapy, but eventually re-grew. SCLC-N tumors (bottom) were generally resistant to chemotherapy. **D:** Images of dissected barcoded tumors. Tumors displaying visible GFP expression were dissected, indicating some degree of barcoding in these samples.

Chapter 6: scRNA-seq of Barcoded Xenografts Reveals Increased Transcriptomic Plasticity in SCLC-A Tumors

6.1: SCLC-A tumors exhibit transcriptomic changes after chemoresistance

All xenografts, as well as the pre-injection sample underwent scRNA-seq to profile their transcriptomes on a single cell scale. scRNA-seq data was trimmed, QC performed, and mapping to the genome was performed via CellRanger. Given that the xenografts were not flow-sorted, there may have been contaminating mouse cells in the data. A mapping statistic was assigned to each cell, and it was determined that very few of the cells belong to the mouse genome (Figure 21D, Figure 27D). These cells were excluded from the resulting analysis. The data was then used in Seurat, where cells with poor quality reads were filtered out, the barcoding data was extracted, and cell cycle correction was performed. Loupe was used to visualize the final, processed data. Data analysis is shown in Figures 20-35.

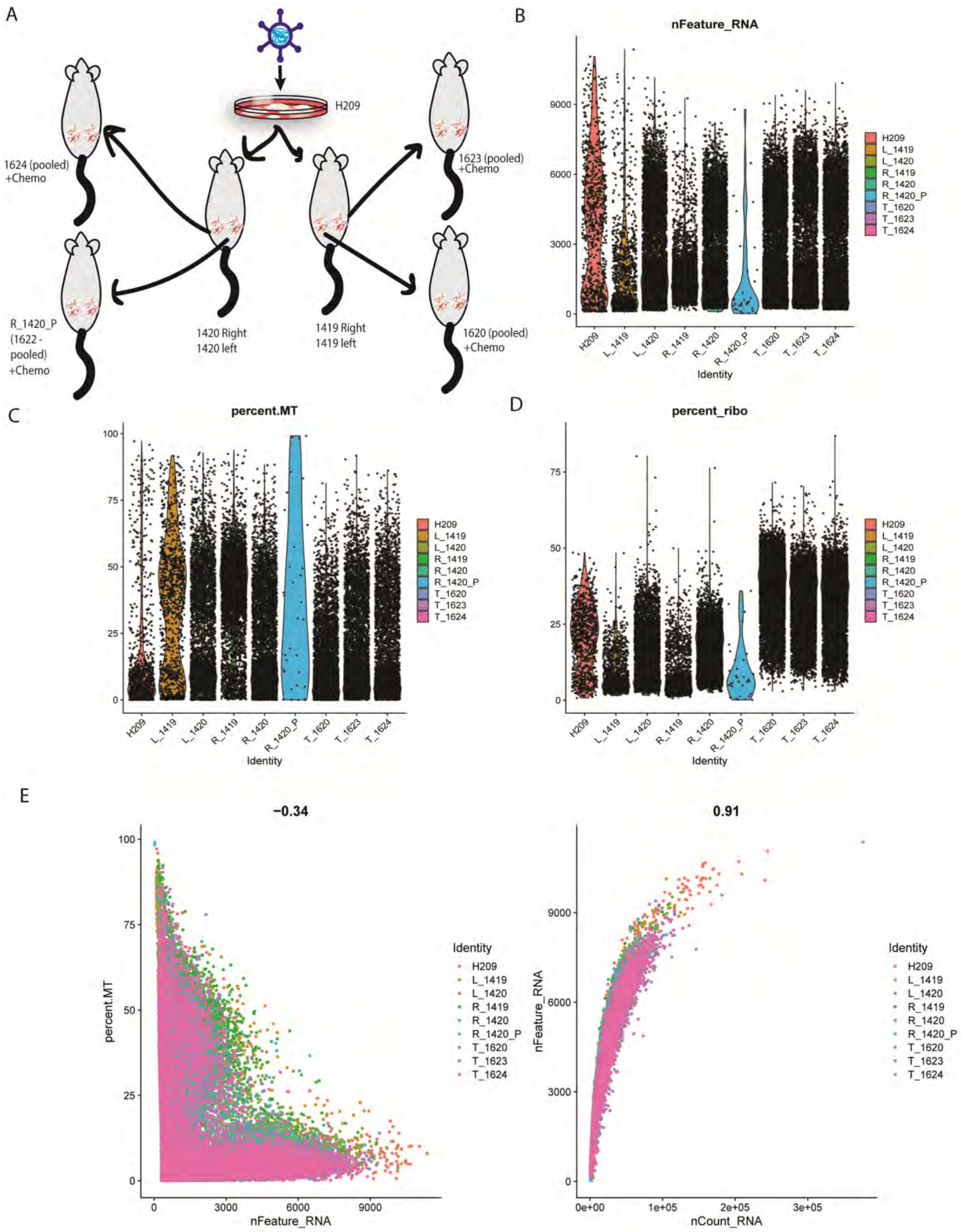


Figure 20: SCLC-A scRNA-seq samples before trimming
A: Overview of the lineages that exist in these samples. **B:** Number of genes detected per sample. Sample R_1420_P has very few genes detected and is of low quality, and should be filtered out. **C:** Percentage of reads corresponding to mitochondrial genes. **D:** Percentage of reads corresponding to ribosomal genes. **E:** Correlation between percent mitochondrial reads and number of genes (left). Cells with a high percentage mitochondrial genes and low total number of genes were filtered out as dead cells. Correlation between number of features and number of genes detected (right). Doublets are filtered out.

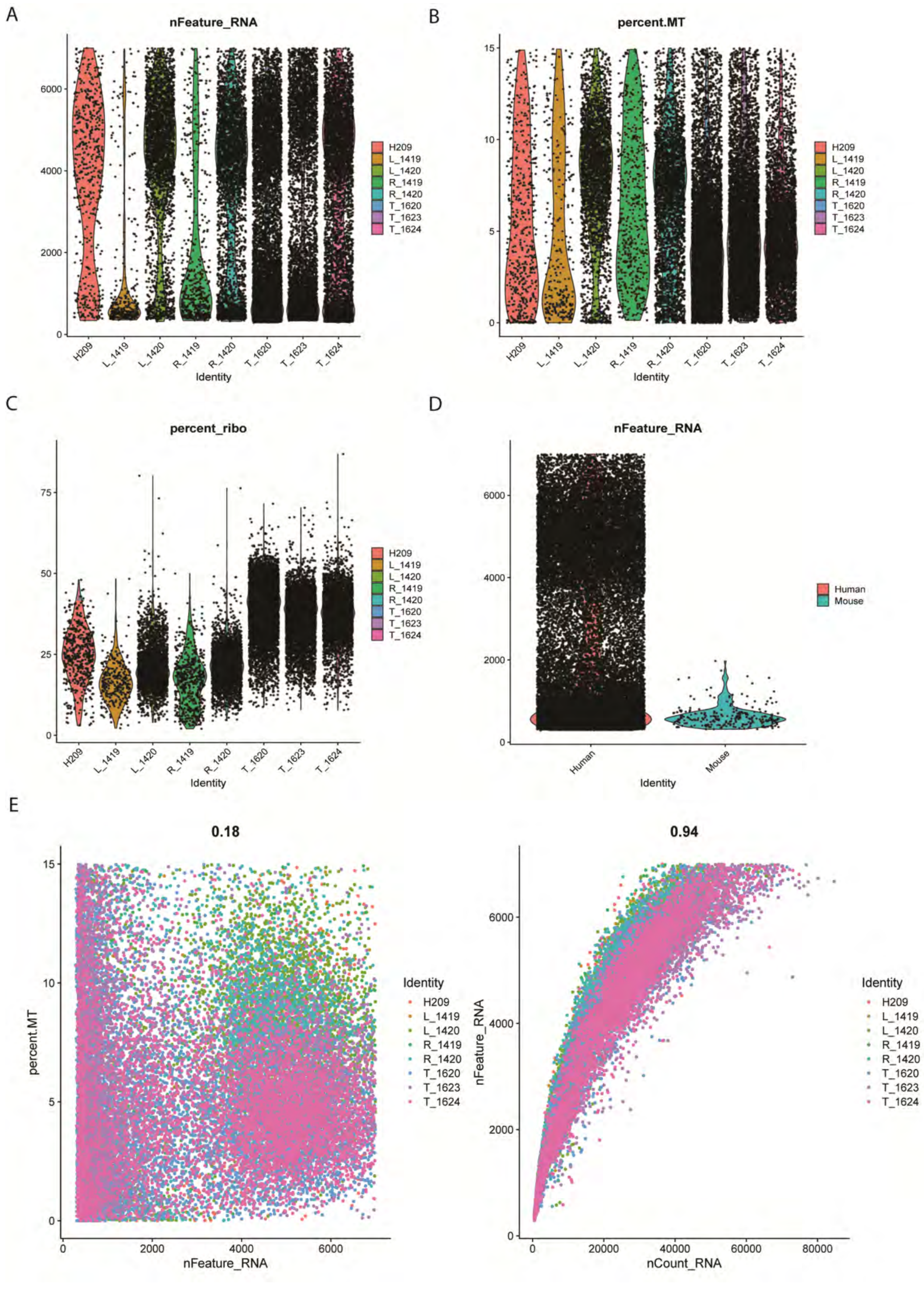


Figure 21: SCLC-A post-trimming scRNA-seq.

A: Number of genes sequenced per cell after trimming the dead cells and doublets. **B:** Percentage of reads corresponding to mitochondrial genes. **C:** Percentage of reads belonging to ribosomal genes. **D:** Number of cells that match either the mouse or human genome. The xenografts are human cells, but were injected in to mice, so it is possible that a few of the cells that were sequenced were stromal mouse cells. Very few of the reads correspond to the mouse genome, but some were still captured. **E:** Post-trimming correlation of percentage of mitochondrial reads and number of sequenced genes (left) and correlation of number of features and number of genes (right).

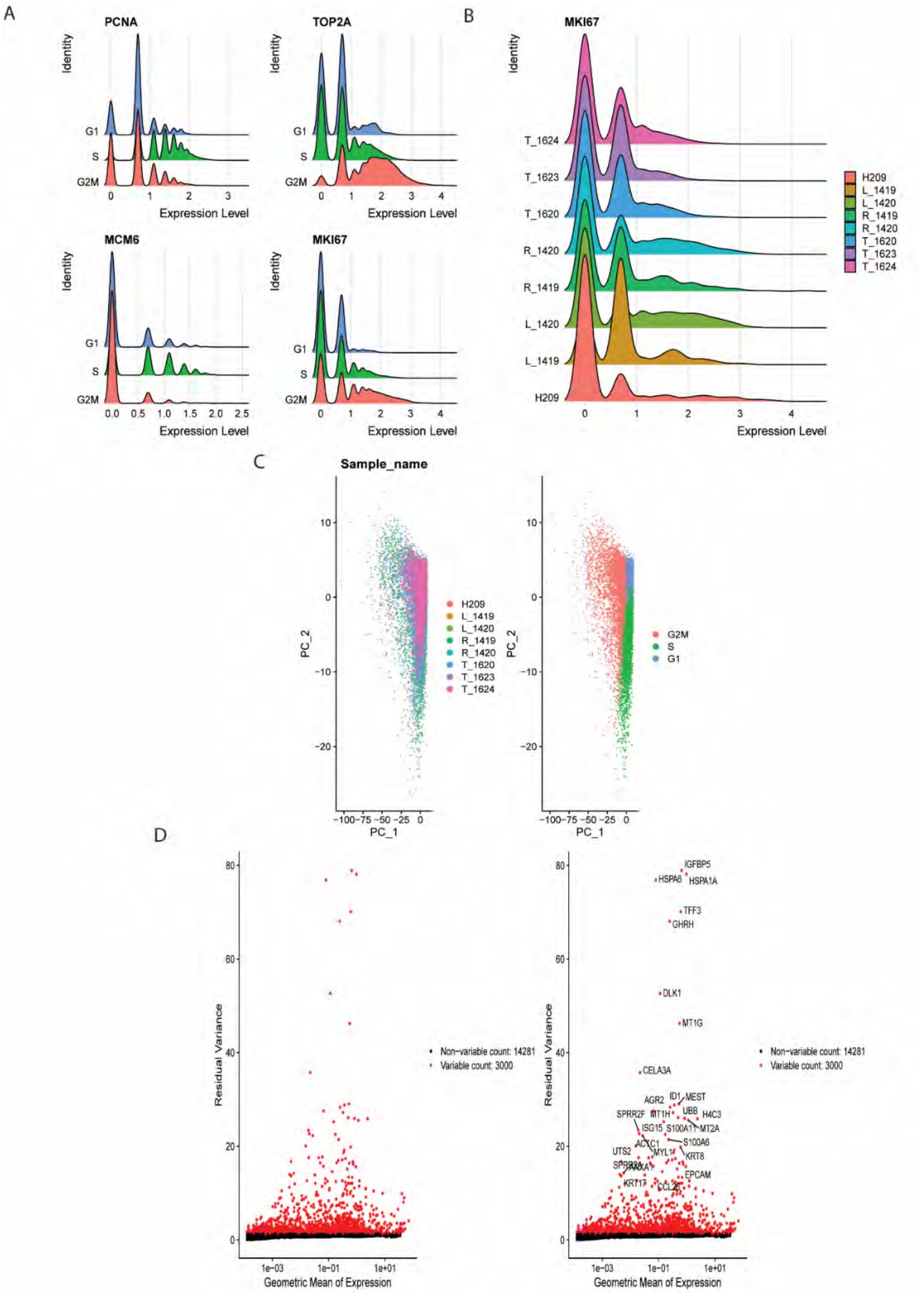


Figure 22: Cell cycle analysis of the SCLC-A scRNA-seq samples
A: Cell cycle marker genes pre-set by Seurat show the distribution of cells corresponding to each stage of the cell cycle. **B:** Amount of KI-67 expressed in each sample, indicating these cells are cycling, with the degree dependent on the sample. **C:** UMAP showing the distribution of the SCLC-A samples (left), and the cells that are in each phase of the cell cycle (right). **D:** Top 10 genes as determined by Seurat. This analysis averages all samples.

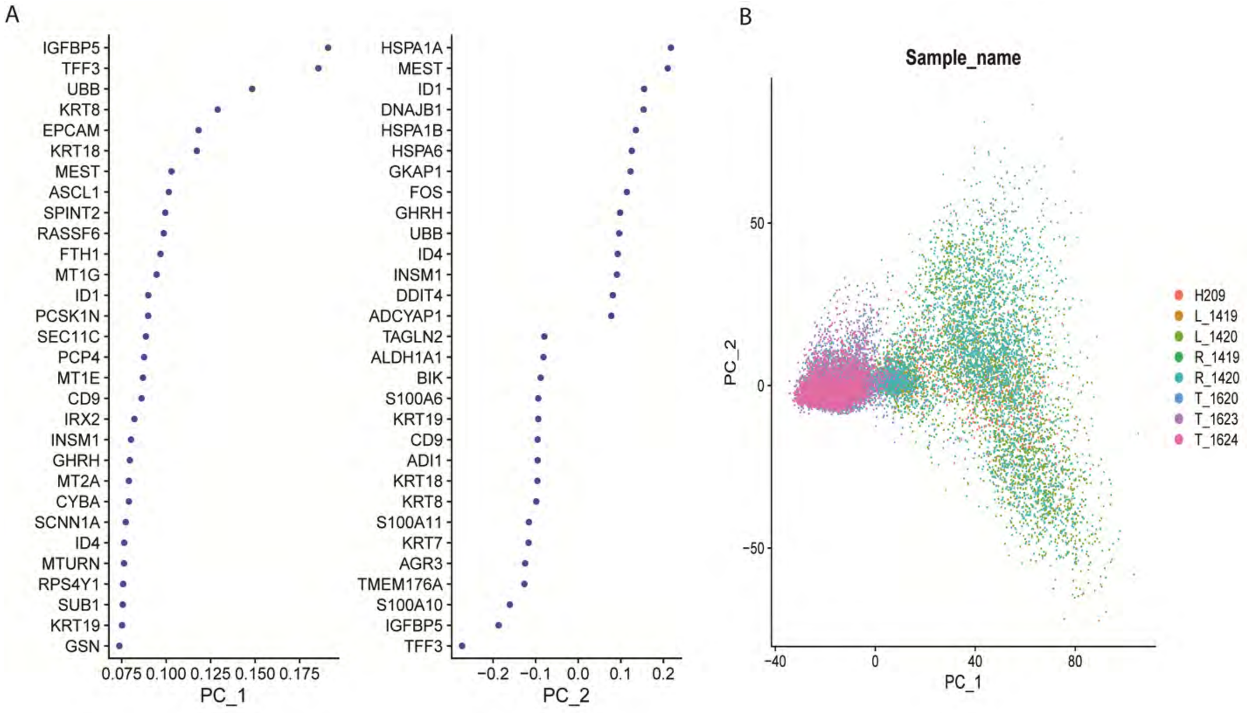


Figure 23: Stratification of the SCLC-A samples based on two principal components
A: DEG identified by PC-1 (left) and PC-2 (left). **B:** Distribution of cells when split on two principal components.

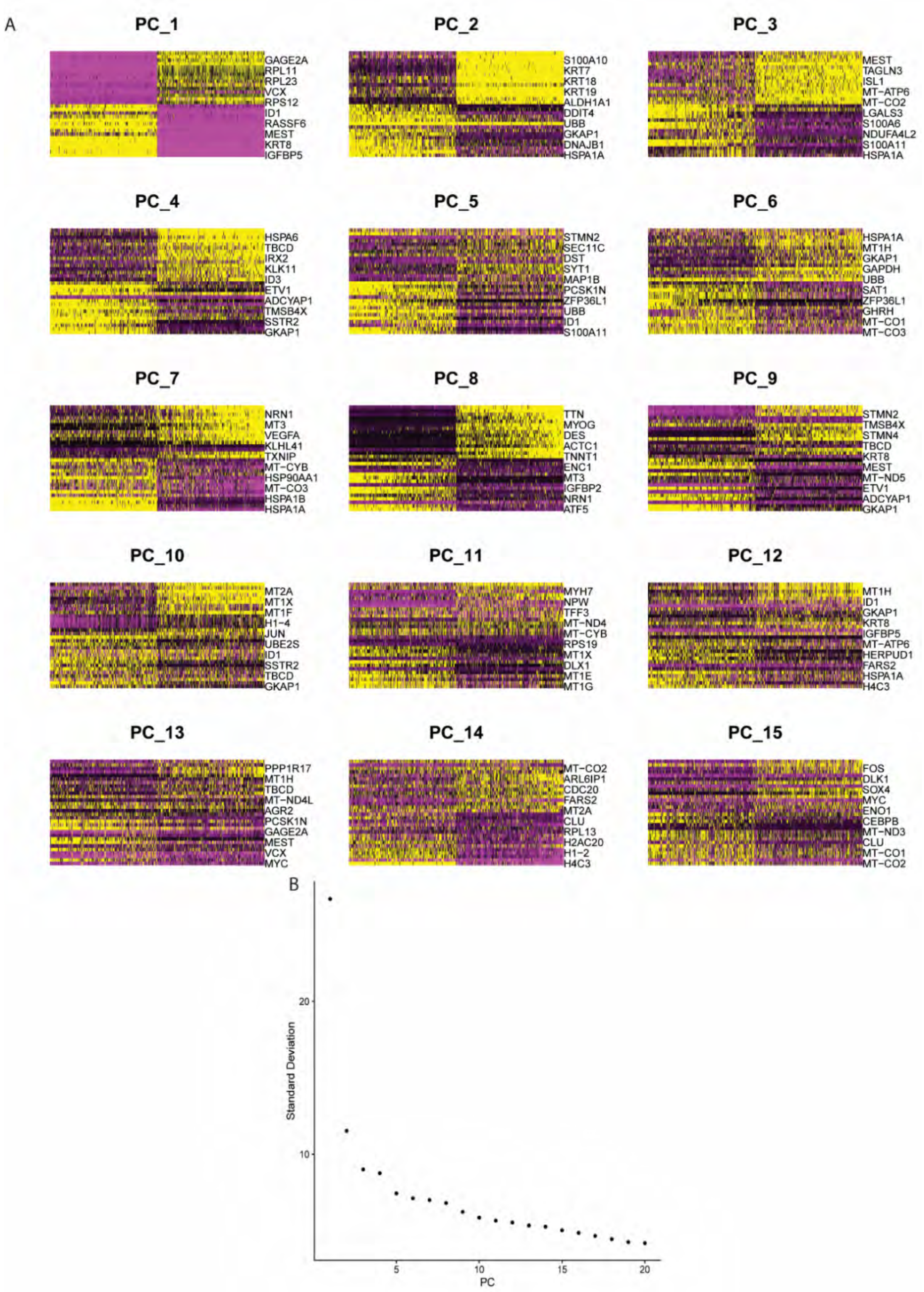


Figure 24: Increasing dimensionality leads to decreased standard deviations in the SCLC-A samples

A: Heatmaps of increasing dimensions. **B:** Elbow plot indicating the decrease of standard deviation with each increased dimension.

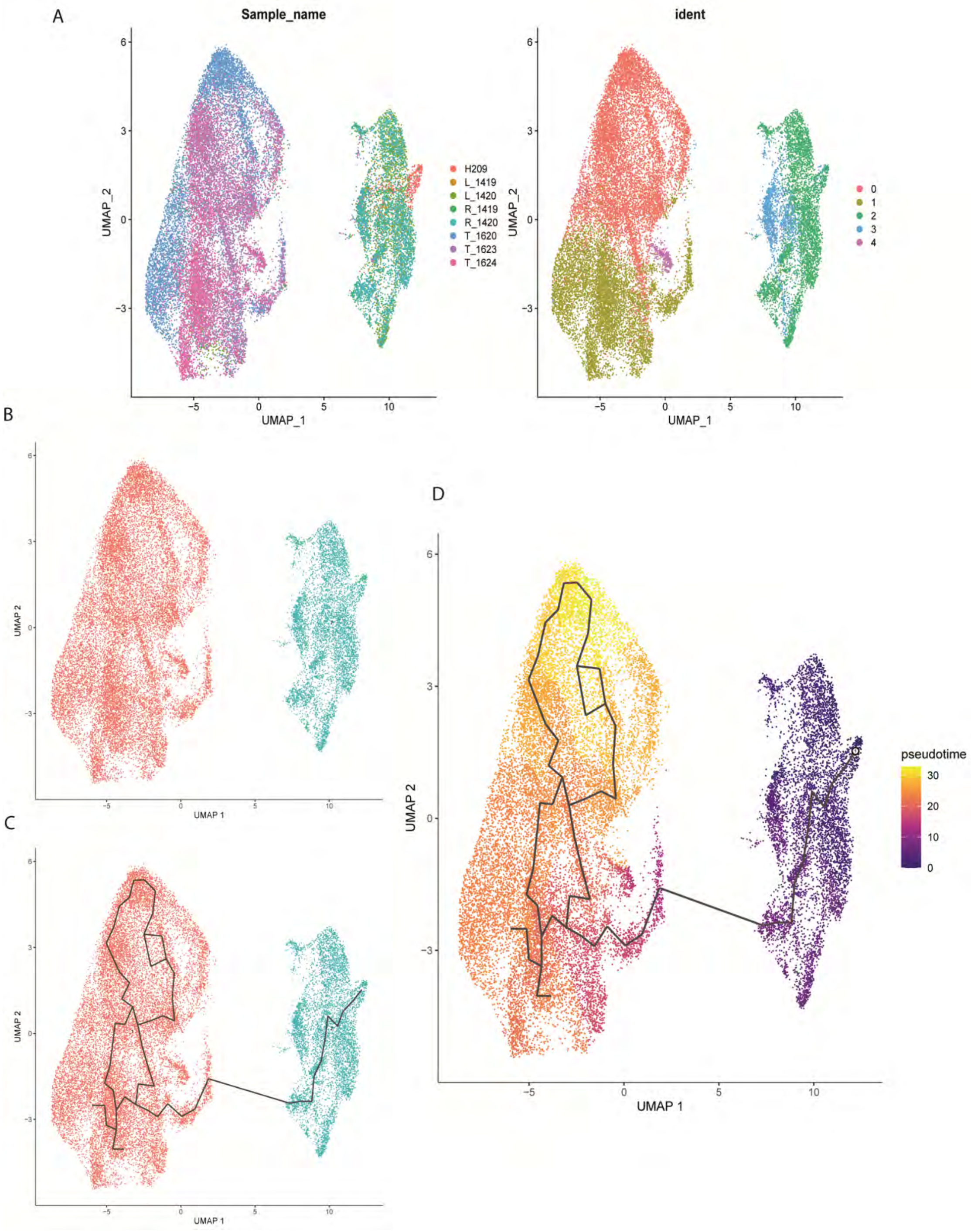


Figure 25: UMAP and clustering used to generate a pseudotime trajectory of the SCLC-A samples

A: Left – UMAP of all SCLC-A samples. Left – clustering used for downstream analysis. **B:** Seurat analysis automatically clustered the cells in to two clusters, blue (no chemotherapy) and red (chemoresistant cells). **C:** Trajectory analysis based on the clustering from B. **D:** Pseudotime trajectory using the clustering from A and accounting for the time-based resolution. The lineage starts in the cell line and continues through the chemo-naïve cells, in to the chemoresistant ones.

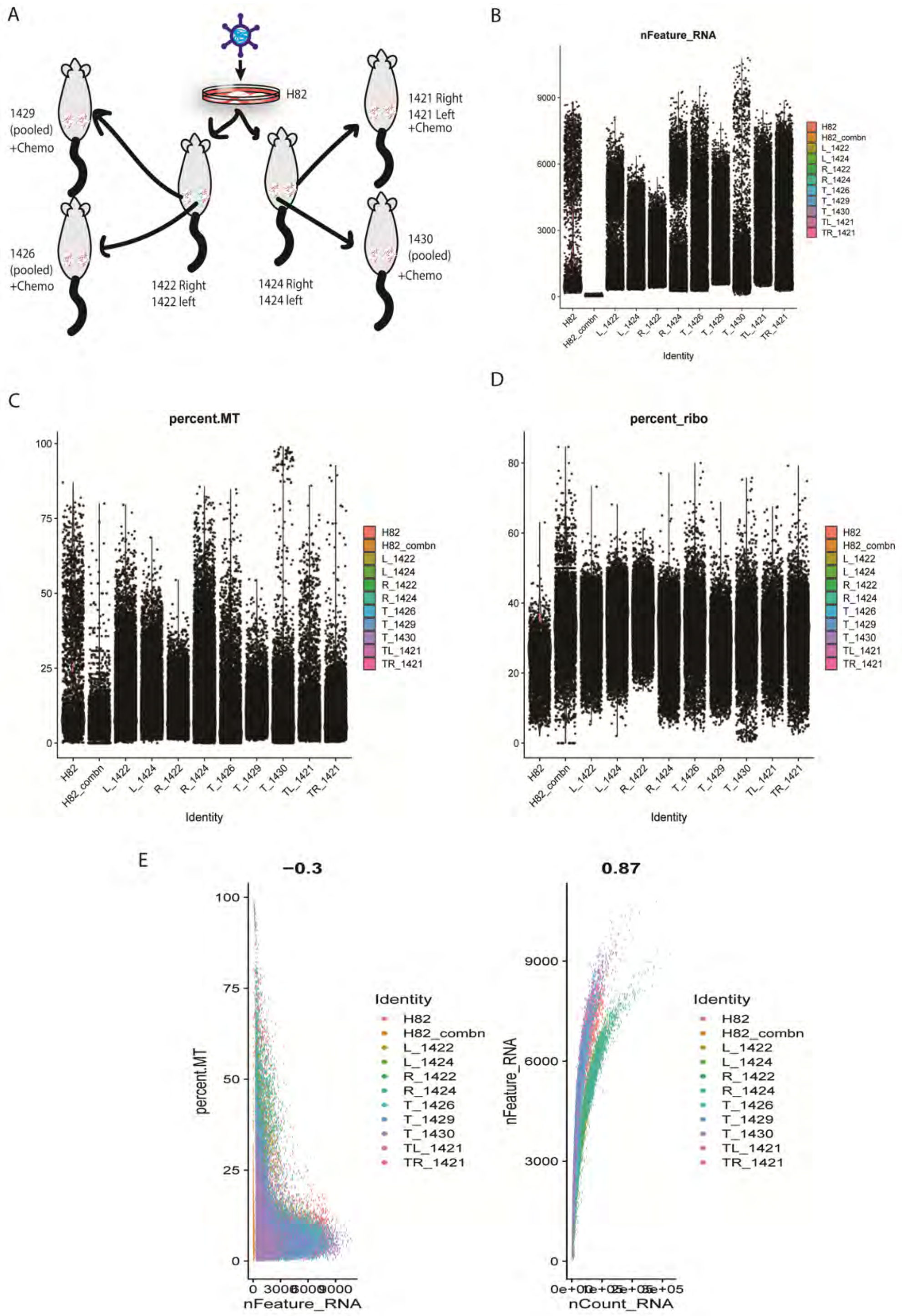


Figure 26: SCLC-N scRNA-seq results before trimming

A: Overview of the lineages that exist in these samples. **B:** Number of genes detected per sample. **C:** Percentage of reads corresponding to mitochondrial genes. **D:** Percentage of reads corresponding to ribosomal genes. **E:** Correlation between percent mitochondrial reads and number of genes (left). Cells with a high percentage mitochondrial genes and low total number of genes were filtered out as dead cells. Correlation between number of features and number of genes detected (right). Doublets are filtered out.

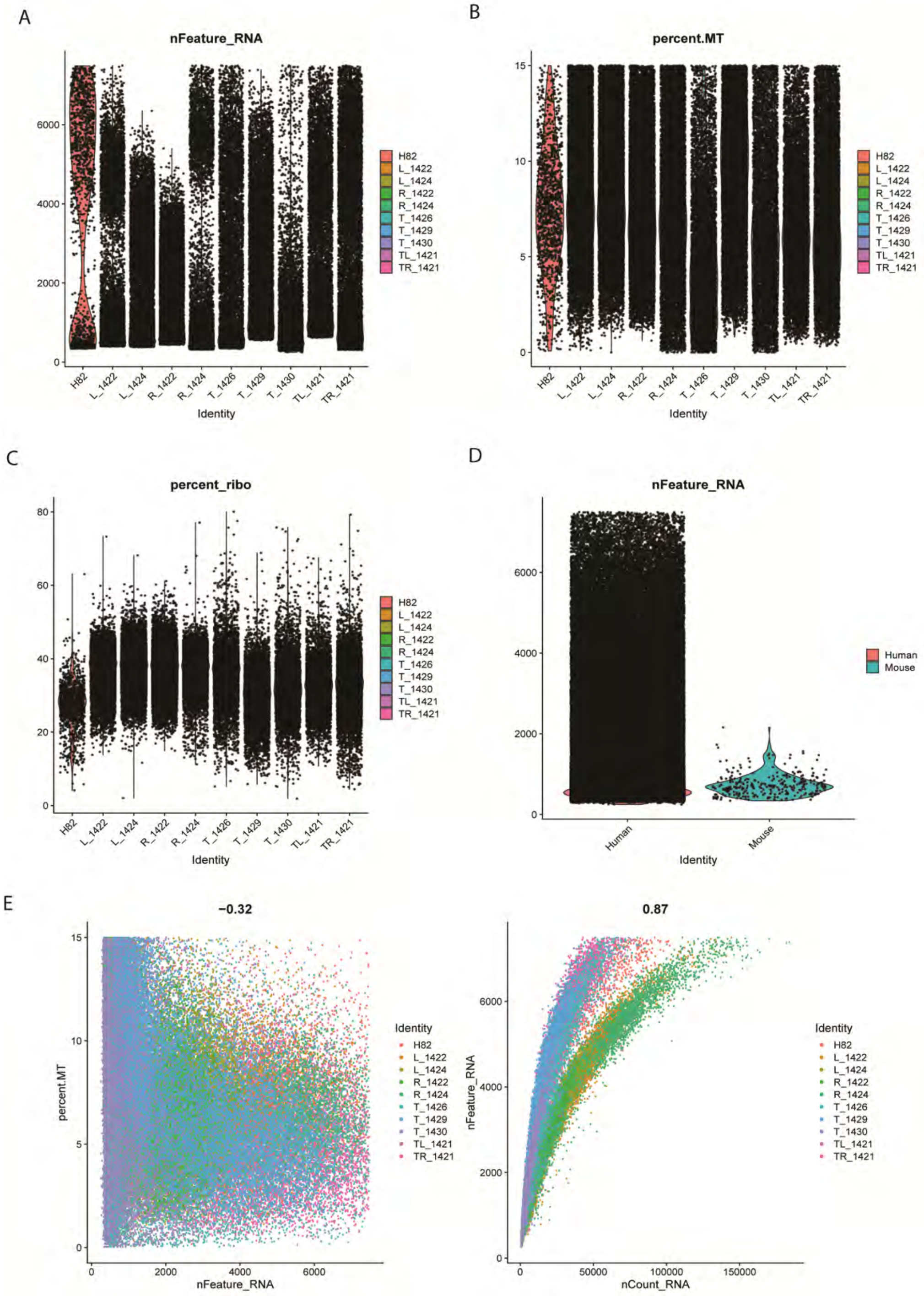


Figure 27: scRNA-seq data post-trimming for the SCLC-N samples
A: Number of genes sequenced per cell after trimming the dead cells and doublets. **B:** Percentage of reads corresponding to mitochondrial genes. **C:** Percentage of reads belonging to ribosomal genes. **D:** Number of cells that match either the mouse or human genome. The xenografts are human cells, but were injected in to mice, so it is possible that a few of the cells that were sequenced were stromal mouse cells. The vast majority of cells corresponded to the human genome, but a few mouse cells were captured. **E:** Post-trimming correlation of percentage of mitochondrial reads and number of sequenced genes (left) and correlation of number of features and number of genes (right).

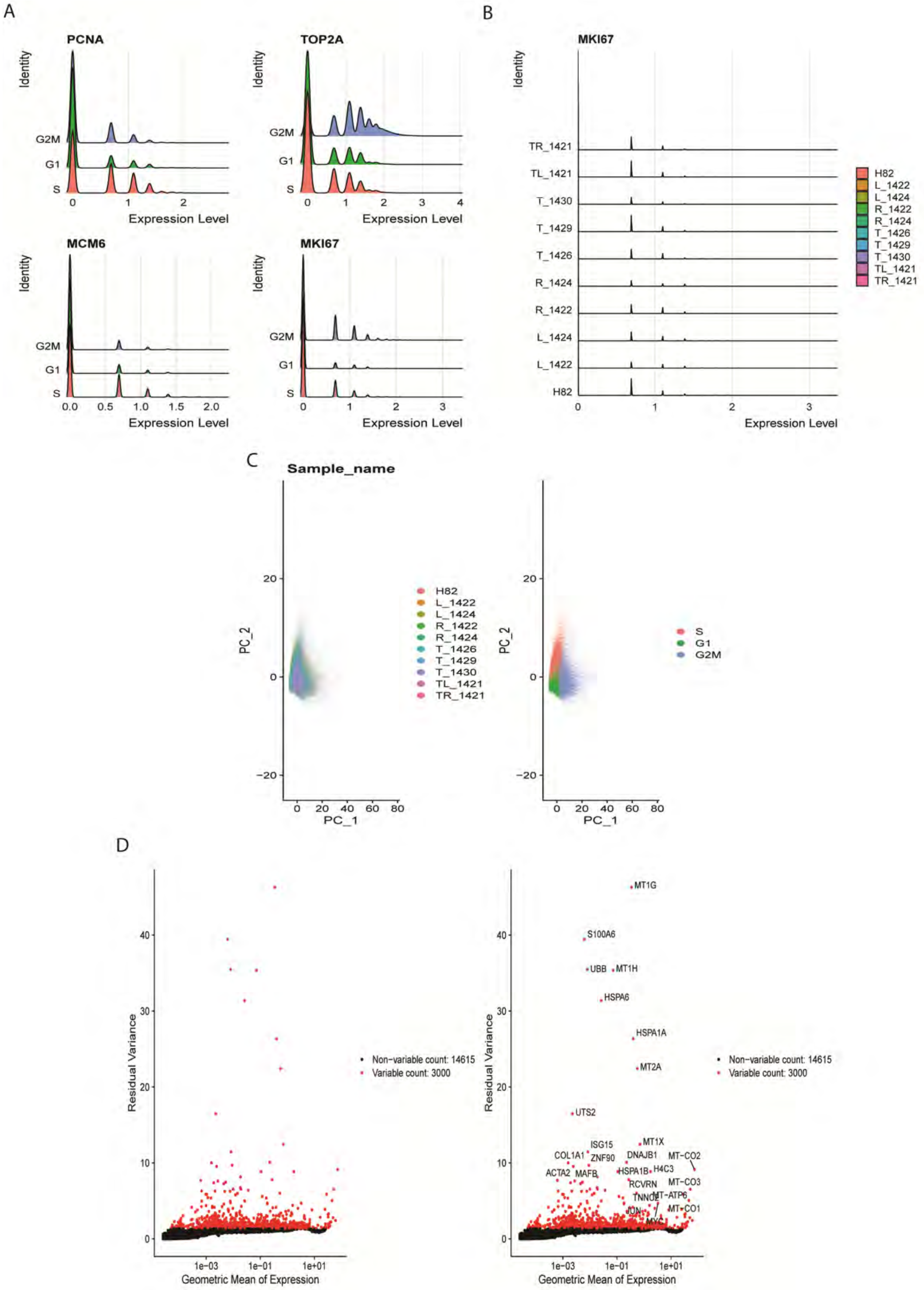


Figure 28: Cell cycle analysis of the SCLC-N scRNA-seq analysis
A: Cell cycle marker genes pre-set by Seurat show the distribution of cells corresponding to each stage of the cell cycle. **B:** Amount of KI-67 expressed in each sample, indicating these cells are cycling. **C:** UMAP showing the distribution of the SCLC-N samples (left), and the cells that are in each phase of the cell cycle (right). **D:** Top 10 genes as determined by Seurat. This analysis averages all samples.

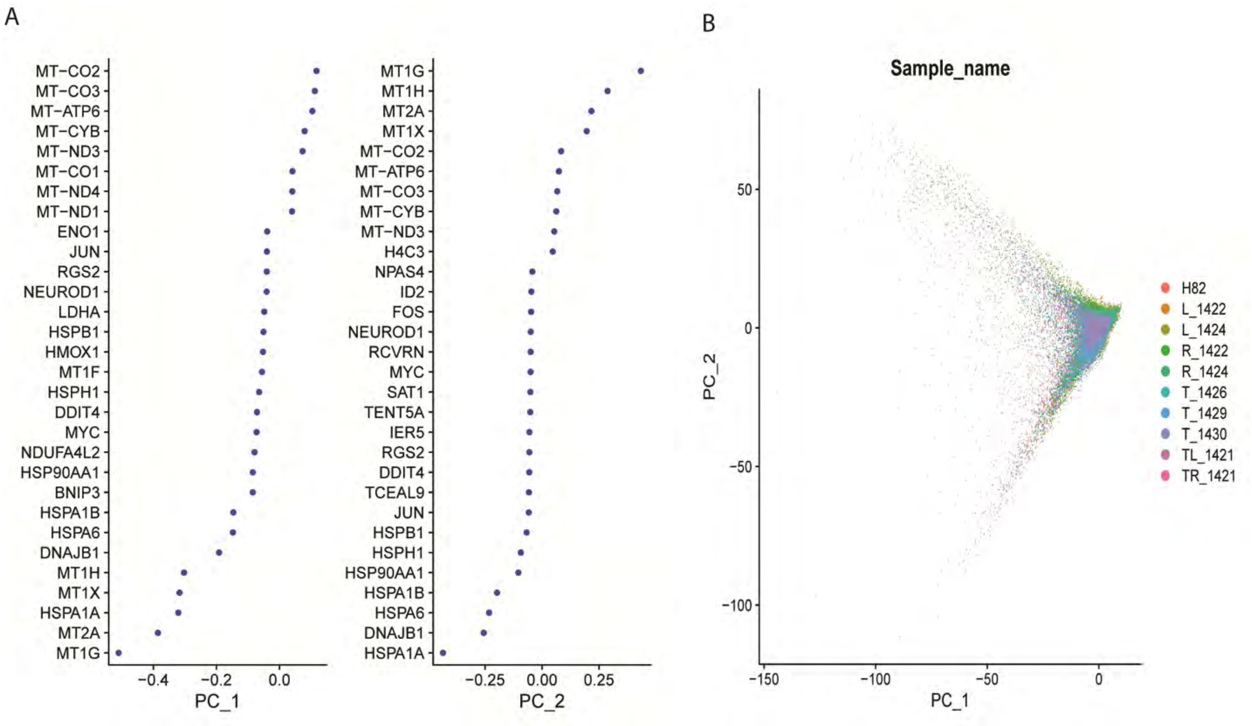


Figure 29: Stratification of the SCLC-N xenograft samples based on two principal components
A: DEG identified by PC-1 (left) and PC-2 (left). **B:** Distribution of cells when split on two principal components.

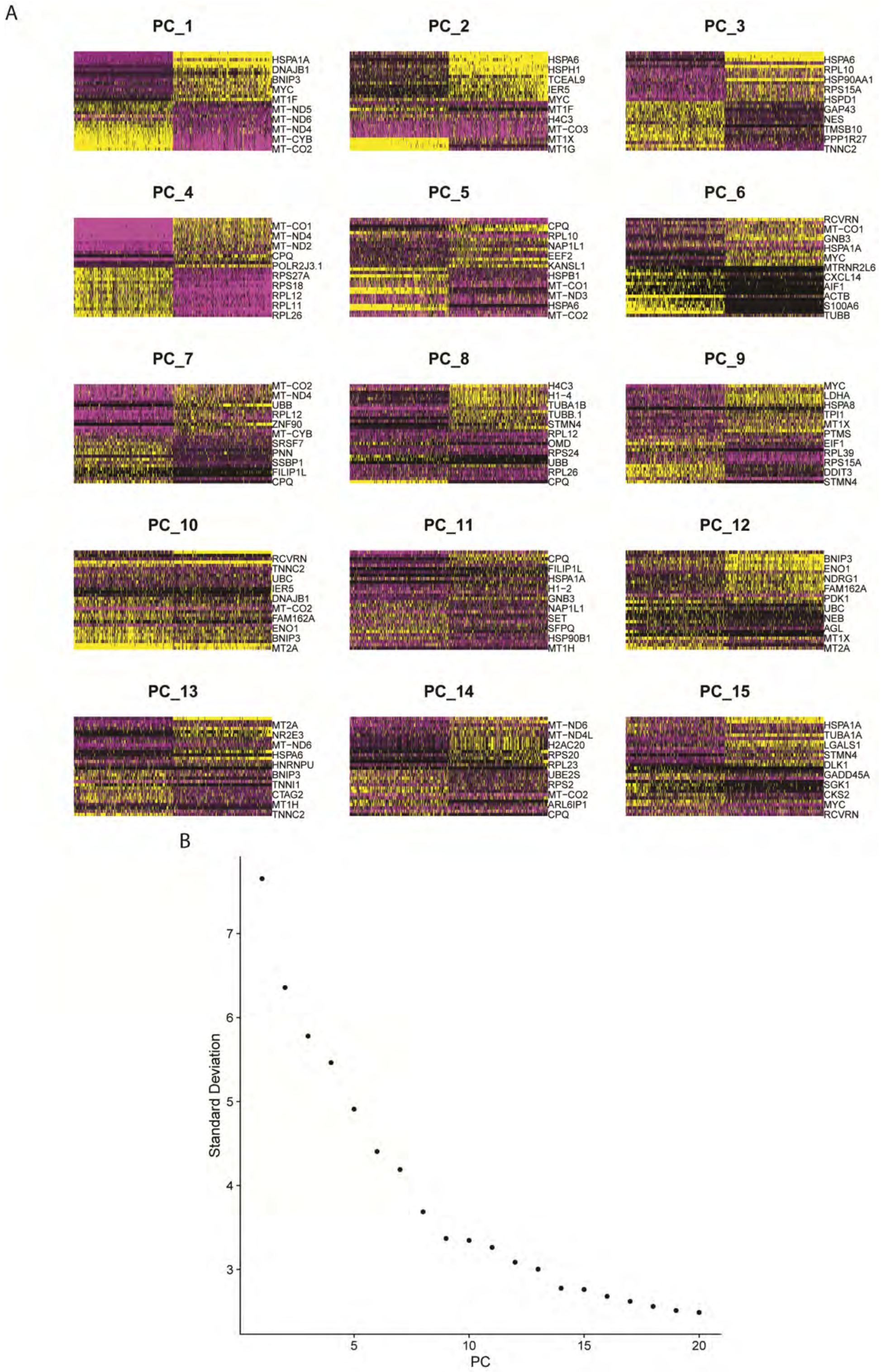


Figure 30: Increasing dimensionality leads to decreased standard deviations in the SCLC-N samples
A: Heatmaps of increasing dimensions. **B:** Elbow plot indicating the decrease of standard deviation with each increased dimension.

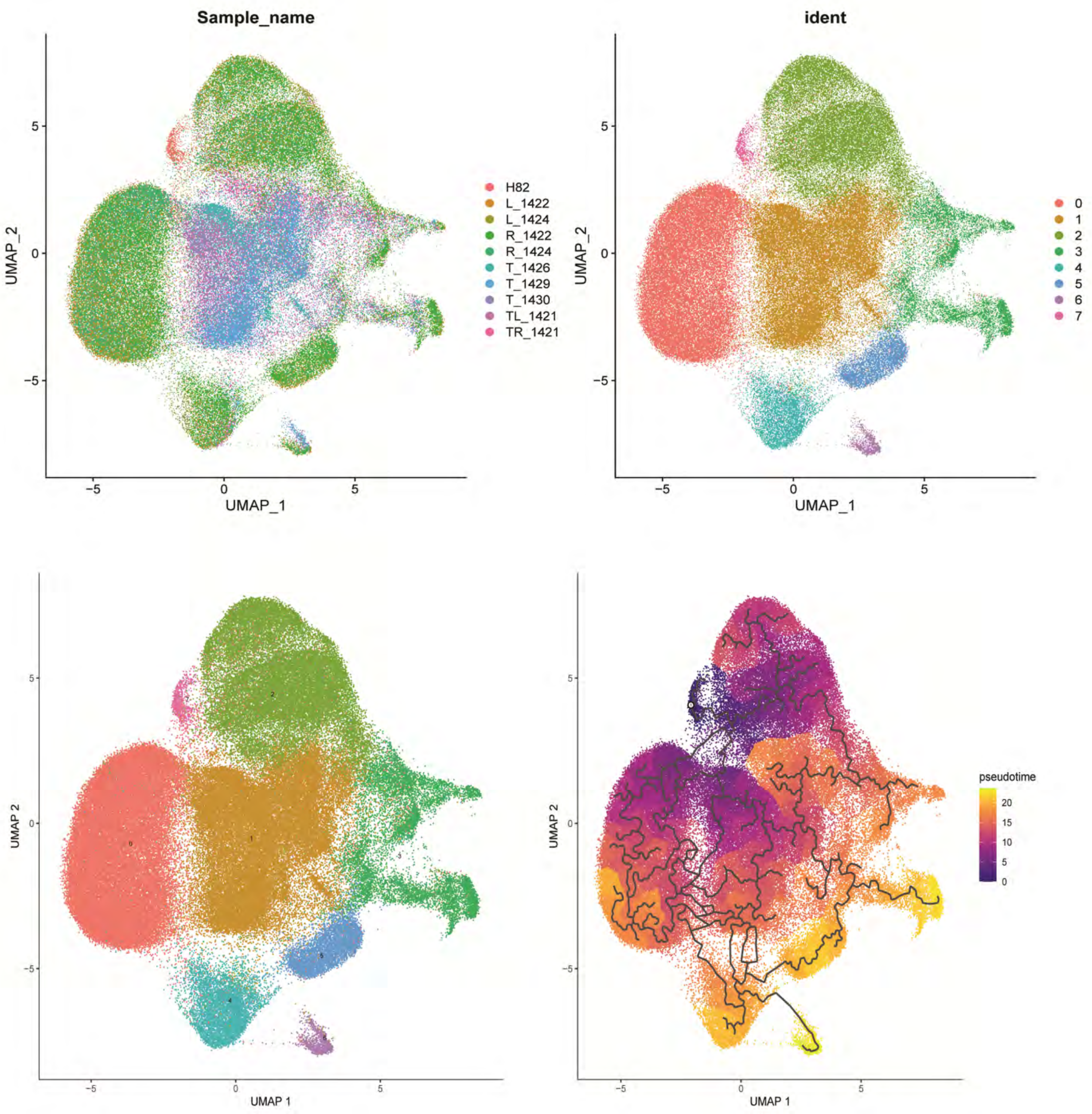


Figure 31: SCLC-N samples UMAP and clustering to generate a pseudotime trajectory.
A: Left – UMAP of all SCLC-N samples. Left – clustering used for downstream analysis. **B:** Pseudotime trajectory accounting for a time-based resolution. Left – clustering used to generate the pseudotime trajectory. Right – SCLC-N pseudotime projection. The lineage is very branched, with two main projections arising from the initial cellular population and diversifying as the xenografts grow and acquire enhanced chemoresistance.

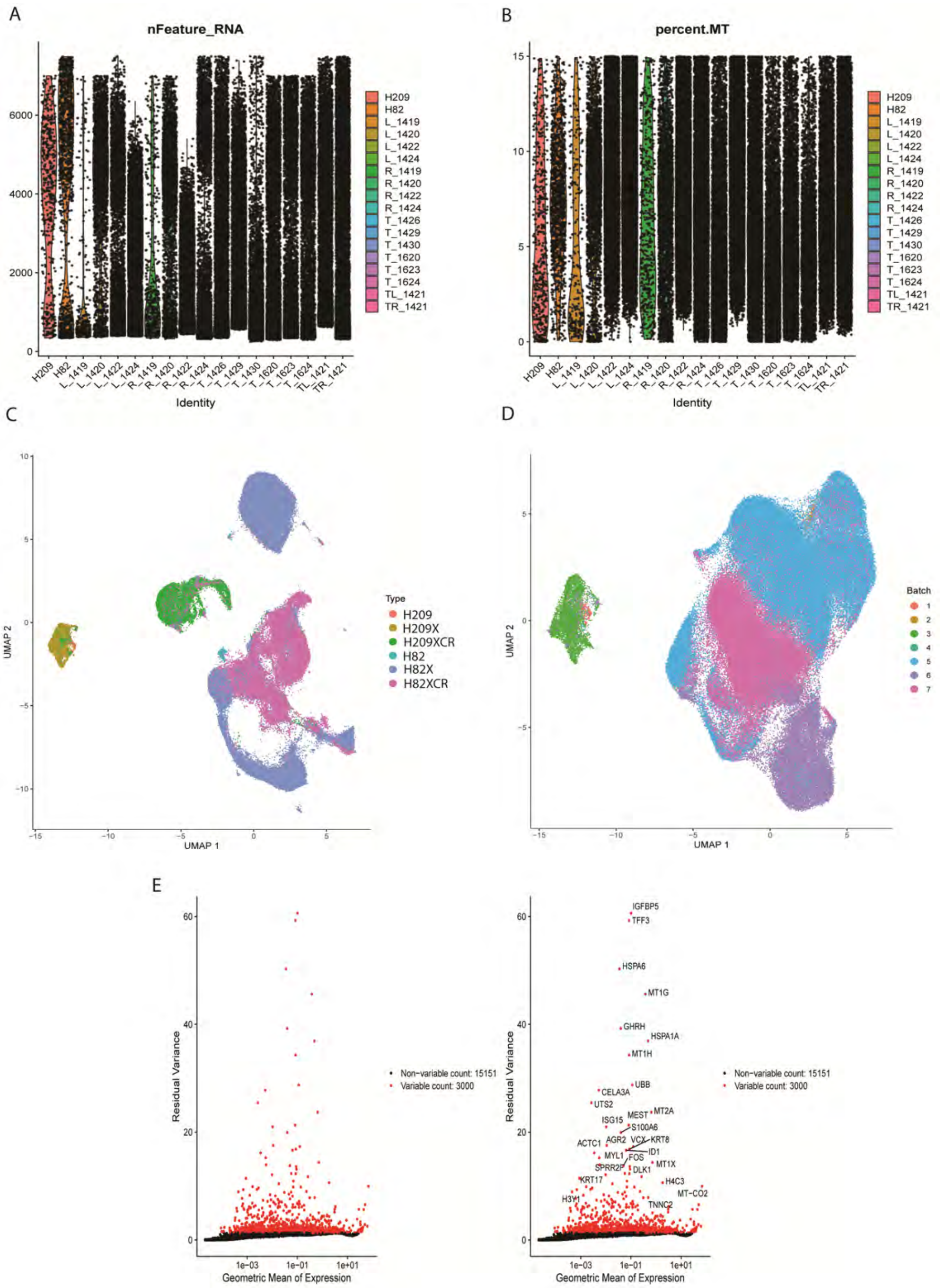


Figure 32: All xenograft scRNA-seq combined together, post-trimming.
A: Number of genes sequenced per cell for both the SCLC-A and SCLC-N cohorts. **B:** Percentage of reads corresponding to mitochondrial genes. **C:** Due to some tumors being collected at different timepoints, there is a batch effect that must be accounted for. C shows the UMAP for all samples prior to batch correction. H209 – SCLC-A cell line, H209X – SCLC-A xenograft, H209XCR – SCLC-A xenograft treated with chemotherapy, H82 – SCLC-N cell line, H82X – SCLC-N xenograft, H82 XCR – SCLC-N xenograft treated with chemotherapy. **D:** All xenograft samples UMAP after batch correction. **E:** Top 10 genes expressed as determined by Seurat.

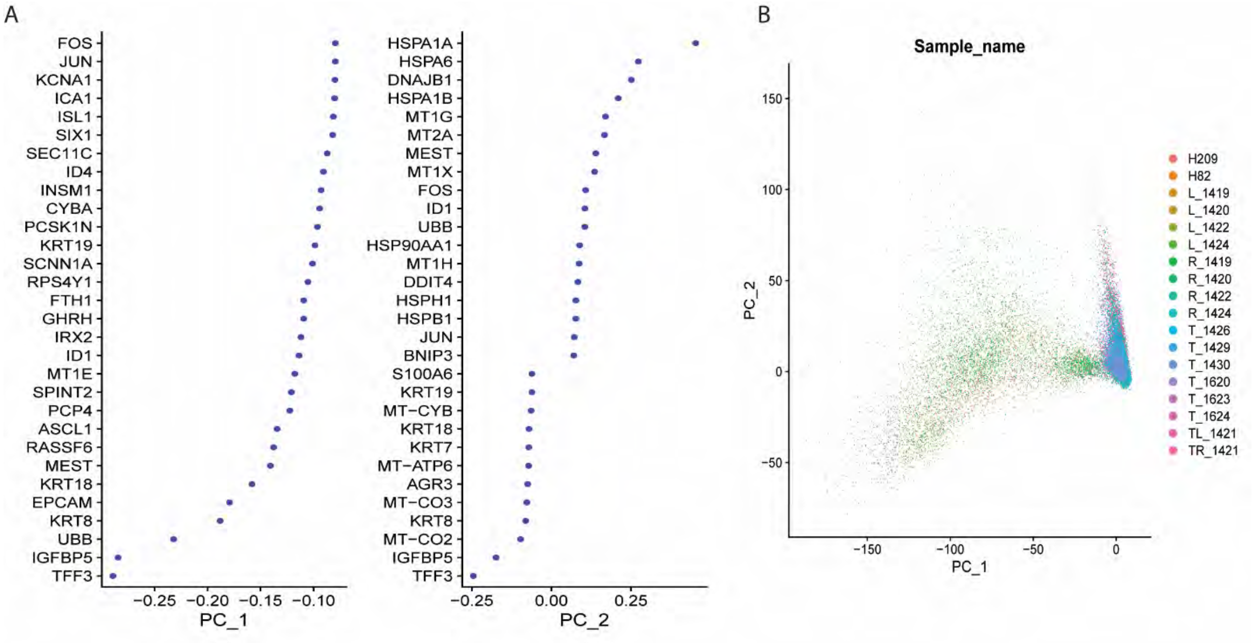


Figure 33: Distribution of all xenograft samples based on two principal components
A: Top differentially expressed genes from one (left) or two (right) principal components. **B:** Distribution of cells when two principal components are used.

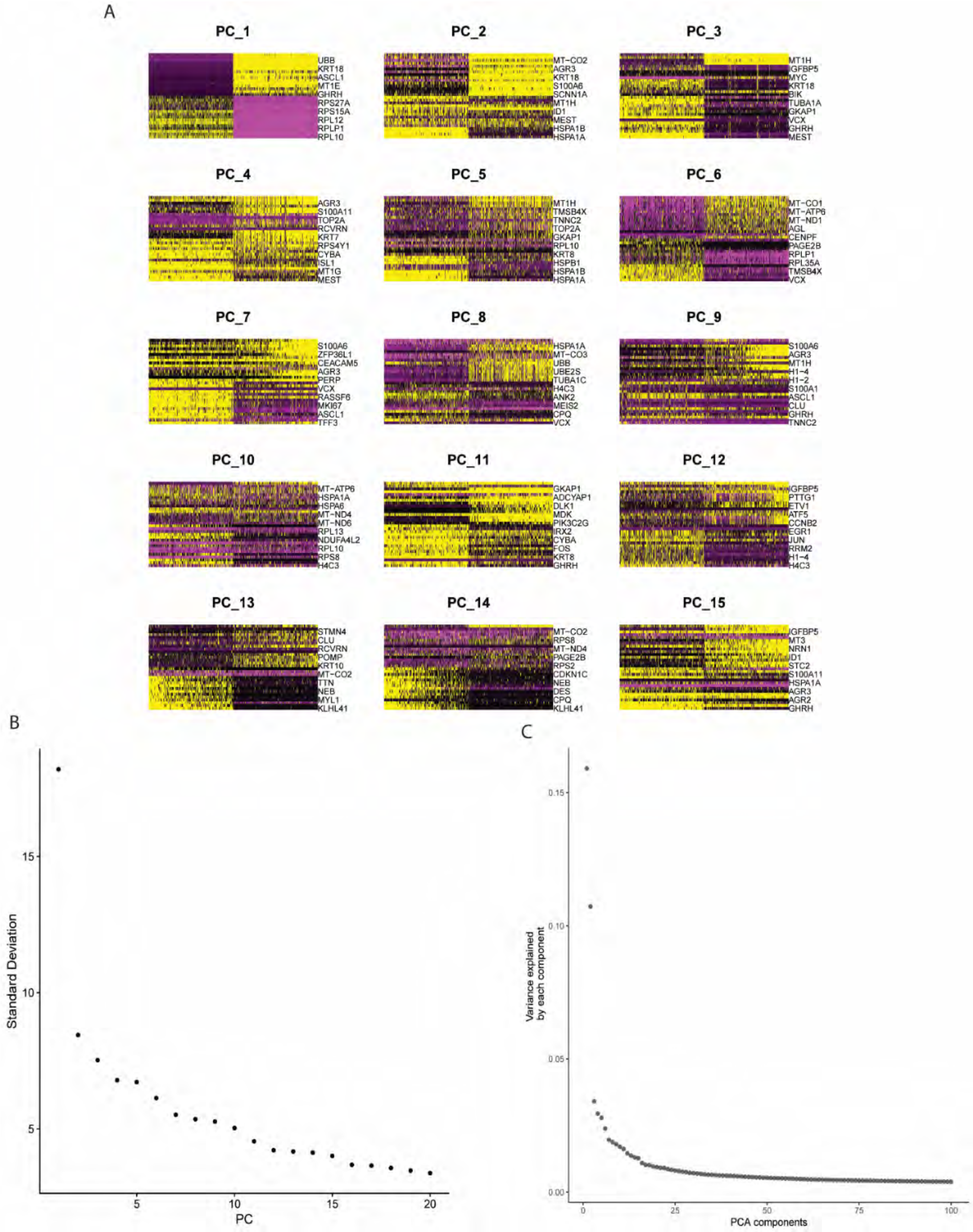


Figure 34: Increasing the number of principal components in all xenograft samples decreases standard deviation and variance.

A: Heatmaps showing DEGs when the data is stratified with increasing components. **B:** Elbow plot demonstrating the relationship between standard deviation and number of principal components. **C:** Variance in the data explained fully by components decreases with each additional component added.

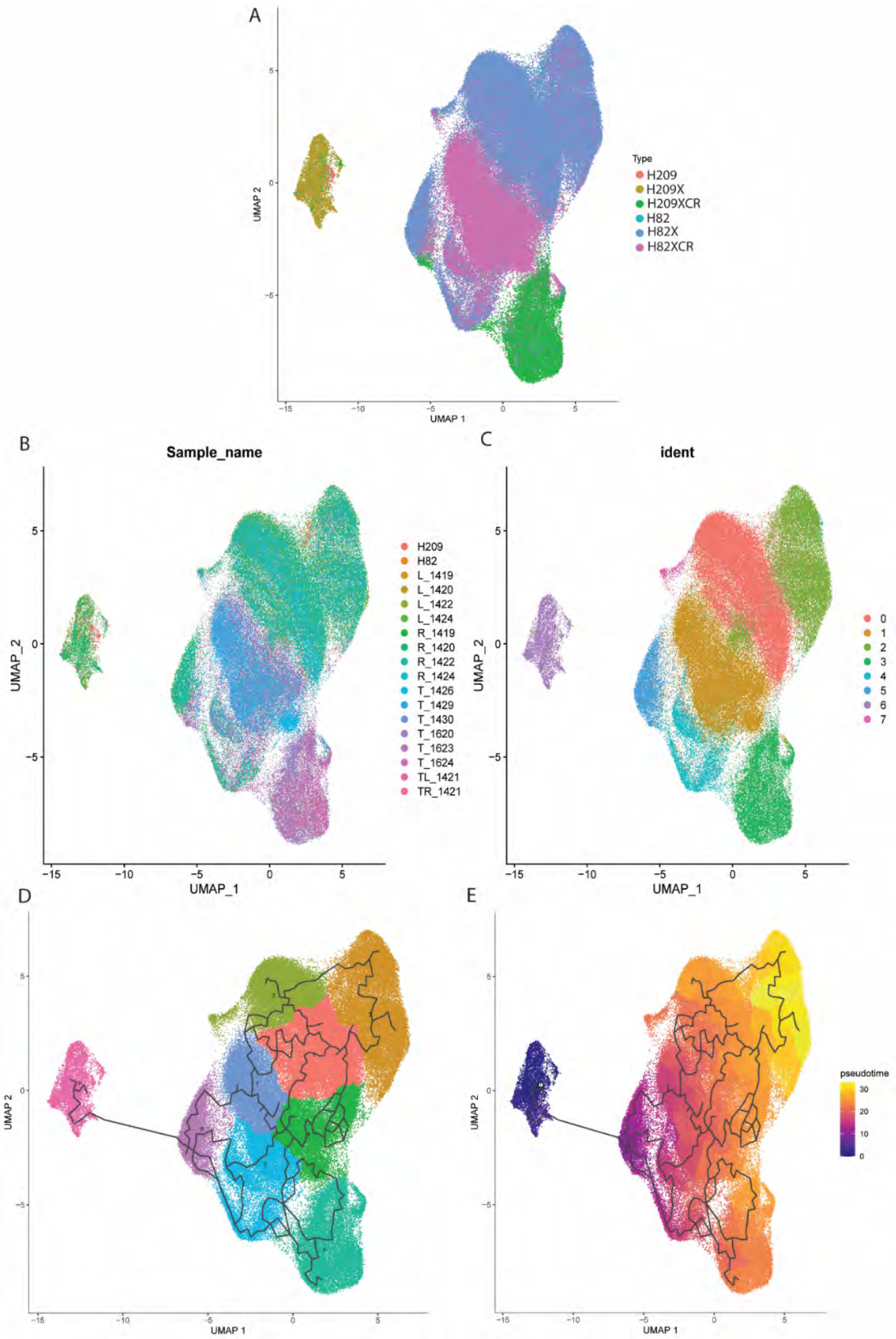


Figure 35: Clustering of all xenograft samples reveals similarities between chemoresistant SCLC-A tumors and SCLC-N tumors

A: UMAP of all xenograft samples coded by their sample type. H209 – SCLC-A cell line, H209X – SCLC-A xenograft, H209XCR – SCLC-A xenograft with chemotherapy, H82 – SCLC-N cell line, H82X – SCLC-N xenograft, H82XCR – SCLC-N xenograft treated with chemotherapy. **B:** UMAP of all xenograft samples color coded based on their sample number. **C:** Clustering used for lineage trajectory set by Seurat. **D:** Lineage trajectory of all SCLC-A and SCLC-N tumors. **E:** Pseudotime reconstruction for all SCLC-A and SCLC-N tumors combined.

There is a large transcriptomic shift between chemo-naïve tumors and chemotherapy treated tumors in SCLC-A xenografts (Figure 37A). In contrast, the SCLC-N tumors do not display a large shift after chemotherapy (Figure 37A). It is known that SCLC-N tumors can be more chemoresistant, so it makes sense that there would not be much of a transcriptomic difference between tumors treated with chemotherapy and tumors without. In contrast, SCLC-A tumors are often chemo-sensitive and a shift in gene expression is seen after chemoresistance is acquired. This is what has been observed here. The tumors that start as SCLC-A take on more of the NEUDOD1-high SCLC-N profile after chemotherapy (Figure 37B). In the SCLC-A pre-injection sample, the transcriptomes of these cells cluster mostly separately from the tumors they form, indicating either a bottleneck event, or transcriptomic shift during the event of tumor formation (Figure 36B). By utilizing the barcodes, we are able to match transcriptomes to lineage barcodes and are able to ascertain which phenomenon occurred. There are however a few cells that belong to the pre-injection sample that cluster more closely with the formed tumors. These could potentially be the tumor initiating cells that survived the bottleneck event to form the eventual tumor. Similarly, there are a handful of cells from the chemo-naïve tumors that cluster more closely with the chemoresistant tumors in the SCLC-A xenografts (Figure 36D). These could potentially be cells that are inherently chemoresistant that have the ability to give rise to a chemoresistant tumor after selection by chemotherapy. With the barcodes, we are able to track the chemoresistant cells back to the initial tumor populations to make that determination. The SCLC-N xenografts display a much lower degree of transcriptomic shift between the pre-injection sample, the chemo-naïve xenografts, and the chemoresistant xenografts (Figure 36A, C, F). Due to the known propensity for SCLC-N tumors to be chemoresistant, it is understandable that little to no transcriptomic shift after chemoresistance would occur. Still, these cells are barcoded, and we have the ability to identify the populations of cells that were able to form tumors and re-grow the tumors after chemotherapy treatment.

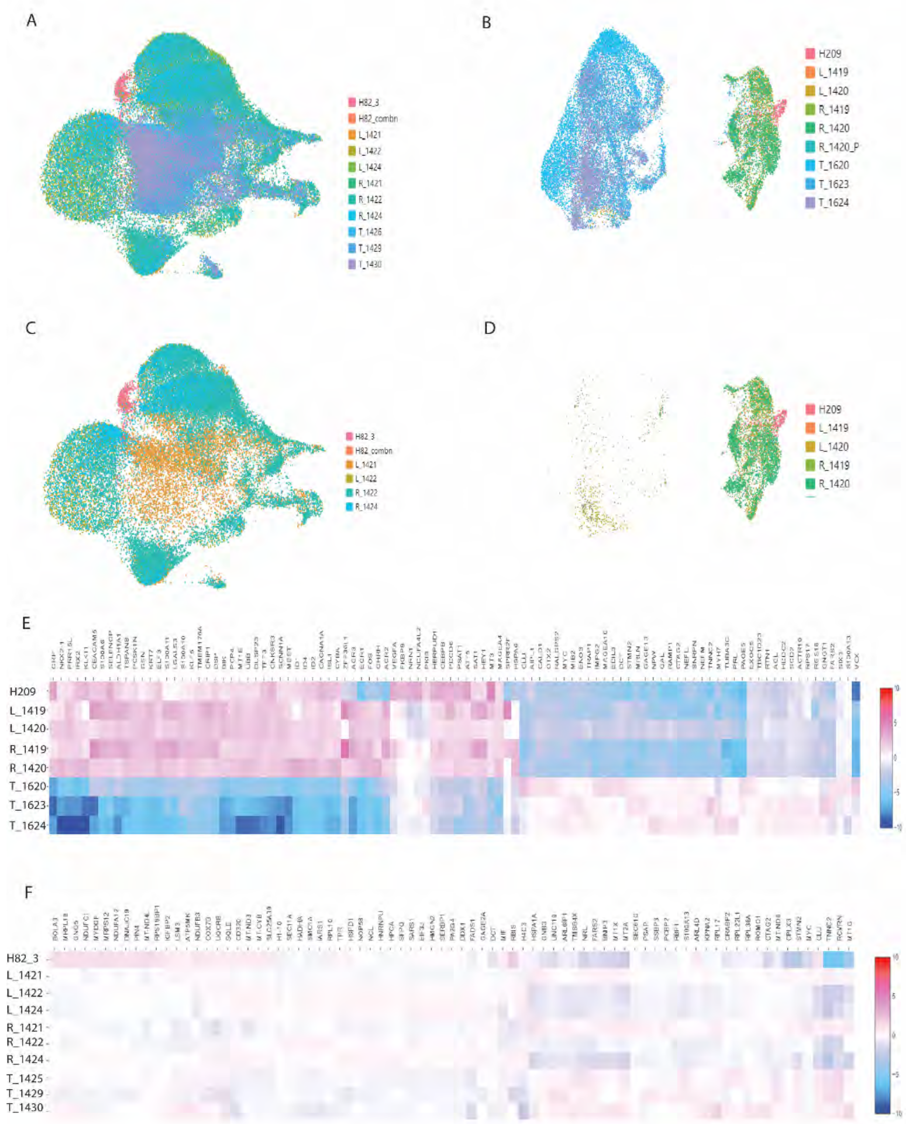


Figure 36: Transcriptomic plasticity is observed in SCLC xenografts.
A: UMAP of the SCLC-N xenograft samples. There is no striking difference in the chemotherapy treated and chemo-naïve populations. **B**: UMAP of all the SCLC-A samples. A robust transcriptomic shift is observed post-chemotherapy. **C:** UMAP of only the SCLC-N samples that did not get chemotherapy, and the pre-injection sample. **D:** UMAP of the SCLC-A samples with the samples that received chemotherapy removed. A few cells that correspond to the chemo-naïve cells cluster where the chemoresistant cells do. These could potentially be the cells with inherent chemoresistance that are responsible for seeding the chemoresistant tumor after chemotherapy. **E** Heat map showing top DEG in the SCLC-A samples. The transcriptomic difference between the chemo-naïve and chemoresistant samples is apparent. **F:** Heatmap of the top DEG for the SCLC-N xenografts. Much more transcriptomic homogeneity is observed in these samples.

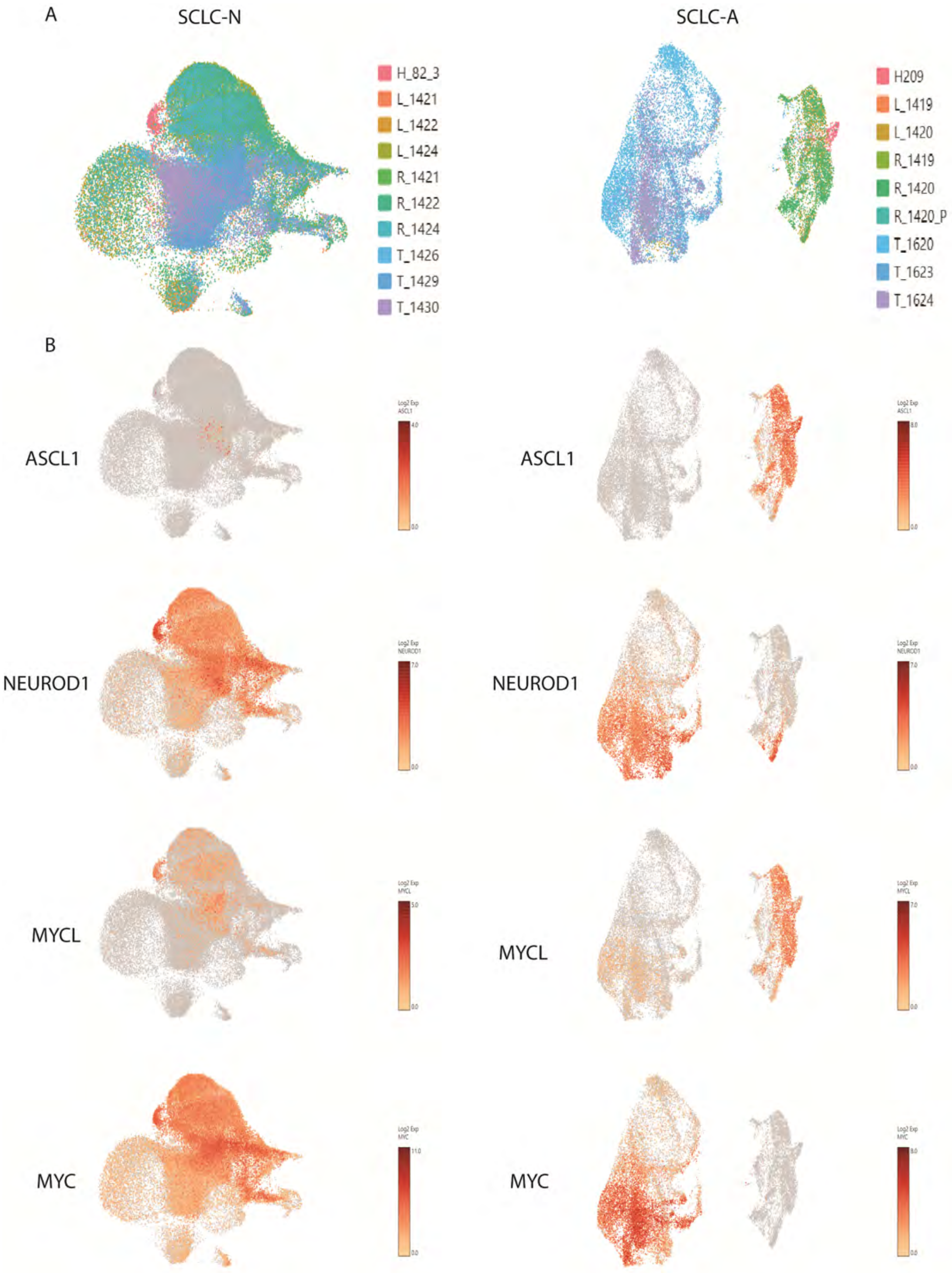


Figure 37: Expression of SCLC subtype genes show a conversion from SCLC-A to SCLC-N in chemotherapy treated tumors.

A: UMAPs of the SCLC-N (left) and SCLC-A (right) tumors color-coded by sample. The SCLC-N tumors do not show a transcriptional shift after chemoresistance, while the SCLC-A tumors demonstrate robust transcriptional changes, indicated by the leftward shift in the chemoresistant samples. **B:** UMAPs highlighting expression of SCLC subtype genes ASCL1 and NEUROD1 in SCLC-N (left) and SCLC-A (right) tumors. The SCLC-N tumors are low in ASCL1 and high in NEUROD1 regardless of treatment status, while the SCLC-A tumors shift from ASCL1 high to NEUROD1 high after chemotherapy. This corresponds with a shift from MYCL expression to MYC expression, which has been previously documented as part of this subtype switch.

Chapter 7: Cancer Testis Antigens are Mediators of Chemoresistance in SCLC

7.1: Cancer Testis Antigens PAGE5 and GAGE2A are expressed in chemoresistant populations

Cancer/Testis Antigens (CTA) are a large class of proteins almost exclusively expressed in the male germ cells and tumors. CTAs have shown promise as potentially targetable, unique cancer antigens. For this reason, they make excellent candidates for immunotherapy such as CAR-T therapy and cancer vaccines[77-83]. In addition to their role as potential cancer antigens, CT antigens also have oncogenic effects on proliferation, genomic stability, invasion, colony formation, and resistance to apoptosis[78,79,81]. In the scRNAseq, CTAs were significantly upregulated (Figure 38A, B). In particular, PAGE5 and GAGE2A were significantly upregulated after chemotherapy in SCLC-A xenografts. In the inherently chemoresistant SCLC-N xenografts, PAGE5 and GAGE2A were highly expressed in all populations. PAGE5 has been identified to be expressed in some cancers, and in melanoma was elevated as an anti-apoptotic gene in response to platinum-based chemotherapy. Expression of PAGE5 was shown to be pro-survival, and upregulated genes related to melanoma cell survival[84]. GAGE2A is another anti-apoptotic CT antigen, that seems to be related to treatment resistance in medulloblastoma[85]. I therefore sought to investigate the role of PAGE5 and GAGE2A in mediating chemoresistance in SCLC.

To evaluate the effect of chemotherapy treatment on expression of PAGE5 and GAGE2A, H29, H82, H209, and H1836 SCLC cell lines were treated with chemotherapy in culture. The IC50 value for cisplatin and etoposide treatment of cells was first determined by treating H82 or H209 cells with varying concentrations of cisplatin or etoposide, and proliferation was assessed via alamar blue assay (Figure 39A). The resulting IC50 concentration was used for the remainder of the *in vitro* chemotherapy response experiments. All four cell lines were treated with the IC50 dose of cisplatin or etoposide, and two or three days later, cells were harvested for RNA and expression of PAGE5 or GAGE2A was assessed with qPCR. Cells treated with cisplatin or etoposide demonstrate higher expression of PAGE5 and GAGE2A than cells not treated with chemotherapy (Figure 38C), indicating cells increase the expression of CTAs in response to chemotherapy treatment in culture. In order to understand how chemotherapy treatment impacts CTA expression, xenografts using the four SCLC lines were generated in NSG mice. After tumors were palpable, some mice were treated with chemotherapy. When the mice had reached their endpoint, tumors were dissected and stained for expression of PAGE5 and GAGE2A. Tumors generated from SCLC-N cell lines (H29 and H82) stained positively for expression of PAGE5 and GAGE2A (Figure 38D). Tumors from mice that received SCLC-A tumors (H209 and H1836 cell lines had almost no expression of PAGE5 or GAGE2A until after chemotherapy treatment (Figure 38D). SCLC-A cells express CTAs at a very low level prior to chemotherapy treatment, while SCLC-N cells and tumors express CTAs at a moderate level, which is increased upon treatment with chemotherapy. Universally, chemotherapy treatment of SCLC cells in culture leads to an upregulation of PAGE5 and GAGE2A expression, and treatment of xenografts with chemotherapy also increases the expression of PAGE5 and GAGE2A (Figure 38). The association between chemotherapy treatment and CTA expression suggests a role of CTAs in mediating response to chemotherapy in SCLC.

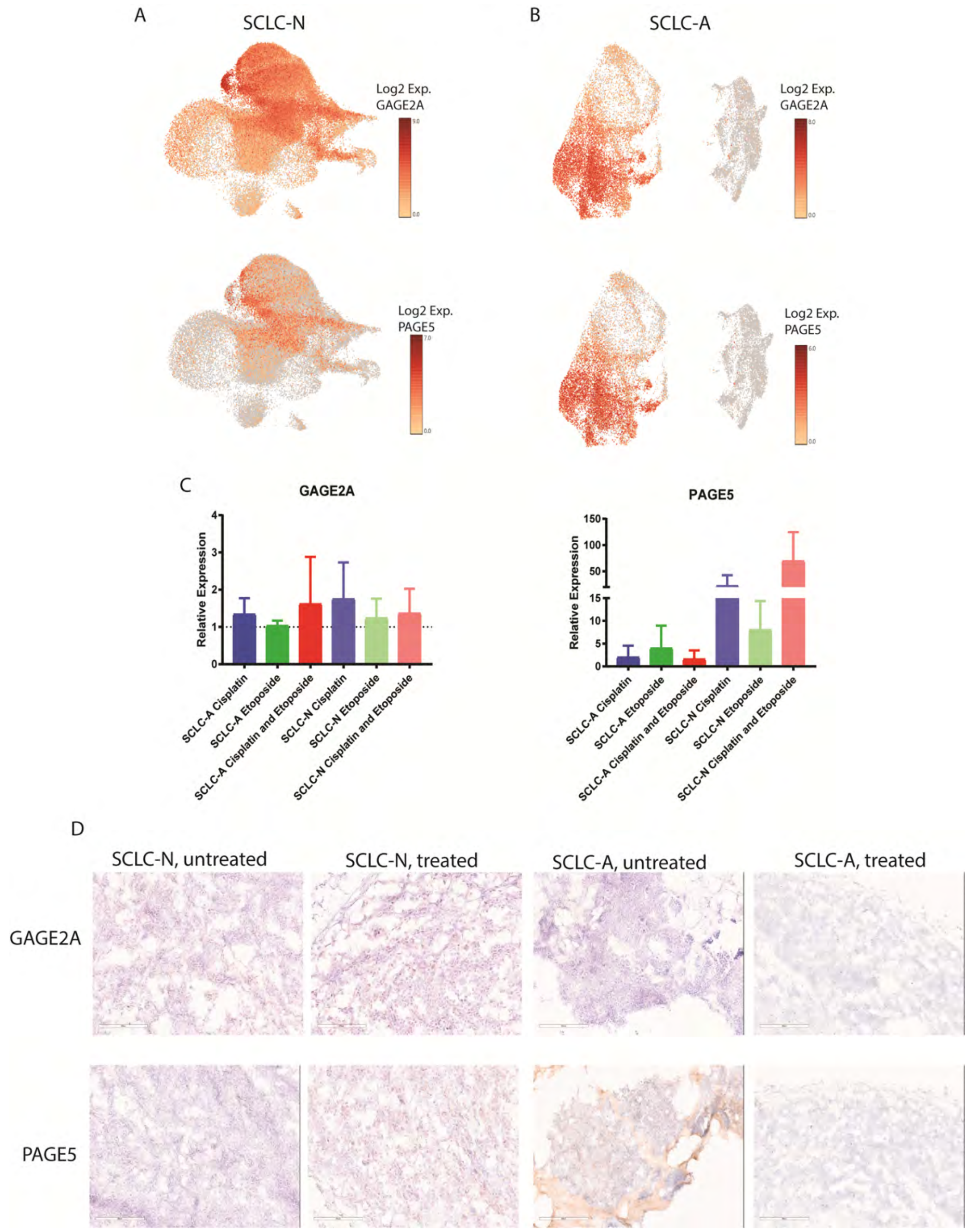


Figure 38: Cancer testis antigen expression is increased after chemotherapy in SCLC.
A: UMAP of the SCLC-N xenografts demonstrating robust GAGE2A (top) and PAGE5 (bottom) expression in all sequenced cells. **B:** SCLC-A UMAP showing increased GAGE2A (top) and PAGE5 (bottom) expression only after chemotherapy treatment. **C:** GAGE2A (left) and PAGE5 (right) expression increase after treatment with chemotherapy in culture. SCLC cell lines were treated with cisplatin, etoposide, or combination therapy, and the level of *GAGE2A* or *PAGE5* expression was assessed via qPCR. A marked increase in expression is observed. **D:** SCLC xenografts treated with chemotherapy show increased immunostaining of GAGE2A and PAGE5. SCLC-N xenografts (left) have existing expression of these CTAs, but the SCLC-A xenografts have immunoreactivity only after chemotherapy treatment.

7.2: Overexpression of CTAs drives chemoresistance in culture

To investigate the effect of PAGE5 and GAGE2A expression on the response to chemotherapy, PAGE5 or GAGE2A cDNA was overexpressed via retroviral expression in H29, H82, H209, or H1836 cell lines (Figure 39B). The cells were treated with the IC50 doses of cisplatin or etoposide alone or in combination for two days, and cell death was assessed using an Annexin V and Propidium Iodide flow cytometry assay. Cells with an overexpression of PAGE5 or GAGE2A were significantly more resistant to cell death caused by cisplatin or etoposide treatment (Figure 39C). An increase in expression of CTAs in SCLC can confer resistance to chemotherapy.

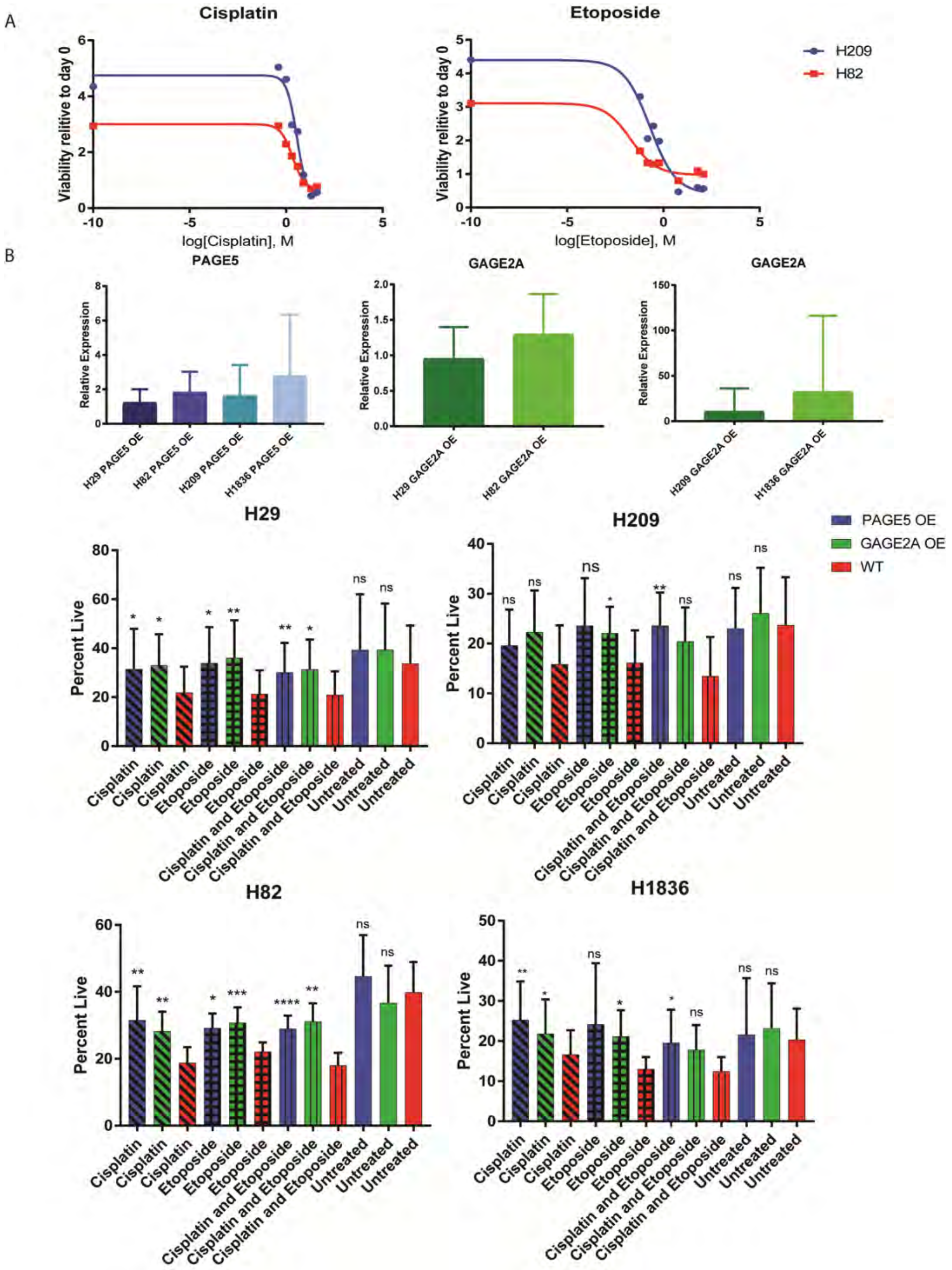


Figure 39: Overexpression of PAGE5 and GAGE2A leads to chemoresistance in SCLC cells in culture

A: The IC50 value of cisplatin and etoposide was determined via Alamar Blue assay. **B**: PAGE5 or GAGE2A were overexpressed via retroviral transduction and the overexpression of PAGE5 (left), or GAGE2A (center, right) was validated via qPCR. **C**: Cells with PAGE5 (blue) or GAGE2A (green) overexpression were treated with chemotherapy in culture and the percentage live cells was assessed with Annexin V and Propidium Iodide staining. Overexpression of either PAGE5 or GAGE2A lead to chemoresistance in these cell lines. Paired t-tests were used to determine statistical differences in percentage live cells between the wild-type and the overexpression groups. N=5 for H29, n=6 for H82, n=7 for H1836, and n=9 for H209. ns = p>0.05, * = p≤0.05, ** = p≤0.01, *** = p≤0.001, **** = p≤0.0001.

7.3: Knockdown of CTAs sensitizes SCLC cells to chemotherapy in culture and xenografts

In order to further investigate the role of CTA expression in mediating chemoresistance, PAGE5 and GAGE2A were knocked down by retroviral expression of shRNAs in four H29, H82, H209, and H1836 cell lines. To ensure a pure population of cells with PAGE5 or GAGE2A knocked down, FACS was performed to isolate the populations containing the knockdown construct. The expression of PAGE5 or GAGE2A was assessed with qPCR (Figure 40A). These single knockdown cells were treated with cisplatin, etoposide, or combination therapy for two days in culture and the percentage of dead cells was assessed with Annexin V and Propidium Iodide flow cytometry assay (Figure 40B). Interestingly, knockdown of PAGE5 or GAGE2A alone was not sufficient to broadly confer resistance to chemotherapy in culture. Given that both PAGE5 and GAGE2A have been shown to be involved in resistance to cisplatin in the literature[84,86], I investigated the impact of a dual PAGE5 and GAGE2A knockdown on response to cisplatin. I used the single knockdown shPAGE5 or shGAGE2A SCLC cell lines and added the reciprocal shGAGE2A or shPAGE5 lentivirus in saturating concentrations. Since both the shPAGE5 and shGAGE2A lentiviruses use GFP as a reporter, FACS could not be performed after the second viral infection, so the sorted, pure population received the second virus in a high dose to maximize the number of cells that get infected with the second virus. After two days in culture, RNA was harvested and expression of PAGE5 and GAGE2A was assessed via qPCR (Figure 40C). The qPCR data shows that knocking down PAGE5 or GAGE2A reciprocally was successful and the resulting cell lines had decreased expression of both PAGE5 and GAGE2A. These double knockdown cell lines were treated with the IC50 dose of cisplatin for two days, and proportion of dead cells was assessed with the Annexin V and Propidium Iodide flow cytometry assay (Figure 40D). Cells with both PAGE5 and GAGE2A knocked down were more sensitive to cell death caused by cisplatin (Figure 40D). Without cisplatin treatment, there was no difference in the amount of dead cells in the double knockdown cells, indicating a role of CTAs in resisting cell death only following treatment with chemotherapy, but the expression of PAGE5 or GAGE2A is not required for cell survival in the absence of chemotherapy. Given that both PAGE5 and GAGE2A had to be knocked down to sensitize cells to chemotherapy treatment, but only one had to be overexpressed, the expression of only one of these two CTAs is sufficient to drive chemoresistance in SCLC cells in culture.

To investigate the impact of PAGE5 and GAGE2A knockdown *in vivo,* the double knockdown cell lines were injected as xenografts in to the hind flank of immunocompromised mice. When tumors were measurable, cisplatin and etoposide were given. The growth of tumors was tracked over time, and all tumors were collected for histology when they reached euthanasia criteria. These animal studies are ongoing, as the final growth curves are still being generated. Preliminarily, animals that received tumors with shGAGE2A and shPAGE5 were much more responsive to chemotherapy. Some of the mice demonstrated complete response to chemotherapy, where xenografts were not palpable anymore. This lasted as long as six weeks in one animal, and there are currently three with complete responses.

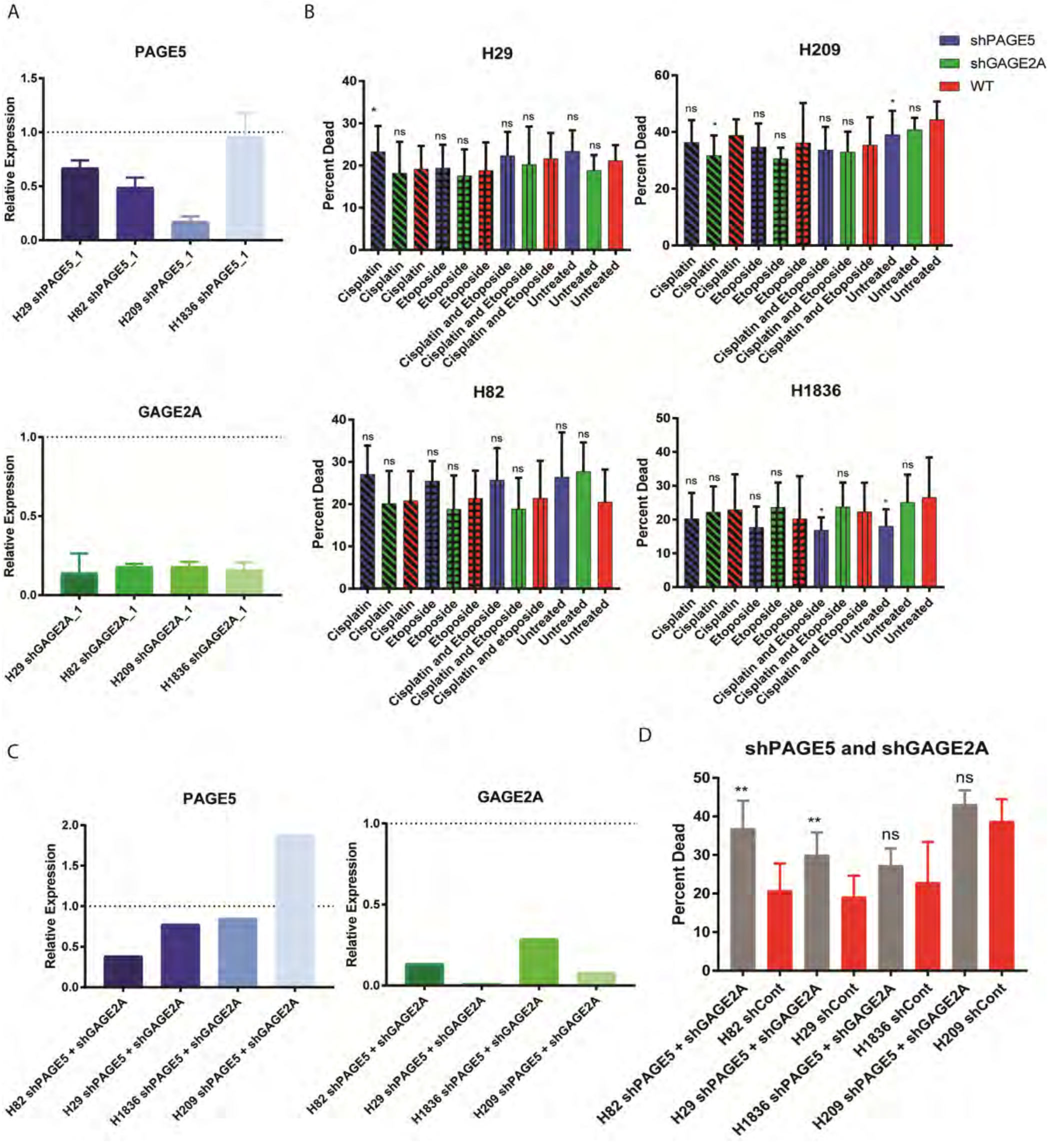


Figure 40: Knockdown of PAGE5 and GAGE2A confer sensitivity to chemotherapy in SCLC cells in culture.

A: FACS plots during the isolation of GFP+ cells containing the shPAGE5, shGAGE2A, or shControl constructs. **B**: qPCR to validate the knockdown of PAGE5 (top) or GAGE2A (bottom). **C**: Annexin V and Propidium iodide assay after chemotherapy treatment of the shPAGE5 (blue), shGAGE2A (green), or shControl (red). Cells that are dead stain positively for both Annexin V and propidium iodide. In general, a single knockdown of PAGE5 or GAGE2A is not sufficient to confer chemo-sensitivity. A paired t-test was used to compare the death of the knockdown cells versus the wild-type cells. N = 7 for all cell types. **D**: The reciprocal construct was added to the single knockdown cells and qPCR was used to confirm knockdown of both PAGE and GAGE2A. E: Annexin V and propidium iodide assay of the double knockdown cells treated with cisplatin show that knockdown of both PAGE5 and GAGE2A in SCLC-N cells does sensitize cells to death induced by cisplatin. Unpaired t-tests were used to compare the wild-type to the double knockdown cells. N = 5 for all double knockdown cells, and n = 7 for all wild-type cells. ns = p>0.05, * = p≤0.05, ** = p≤0.01, *** = p≤0.001, **** = p≤0.0001.

7.4: CTAs are signatures of chemoresistance in patient samples

To understand the clinical implications of PAGE5 and GAGE2A expression, I obtained 29 human SCLC biopsies from the Sanford Health BioBank. The tumor sections were stained with antibodies against PAGE5 or GAGE2A, and slides were scanned with an Apero Slide Scanner. The samples were quantified based on a binary of any staining, and the overall intensity and distribution were scored on 0-3 scale. Two additional researchers quantified the staining. The decision to score by human and not by software was made for a number of reasons. First, many of the sections had patches of blood, which could be confused by software as being the same color as the DAB staining. Secondly, given that these are patient biopsies, not all areas of the tissue are tumor, as areas of surrounding healthy lung are often captured in the biopsy. I wanted to ensure the scoring only evaluated the tumor areas, and not healthy lung, which is more straightforward to train a human than a computer on. Additionally, these antibodies produce variable levels of background staining, depending on tissue processing, which is difficult to account for via software programs. All researchers that scored tumors were trained in the same way and were provided a scoring guide with representative images. All scorers also used the same computer with identical screen settings to further decrease the chance for variability. Representative images are shown in Figure 41A. The majority of the samples (almost 90%, Figure 41B) had some positivity towards PAGE5. Around 40% of tumors (Figure 41B) demonstrated positivity towards GAGE2A. There was some degree of heterogeneity in the scores of the tumors (Figure 41C), indicating perhaps a role of tumor subtype or treatment status. We were not able to obtain meta data about patient treatment or response, but pairing that data to the histology would be an exciting avenue of follow-up.

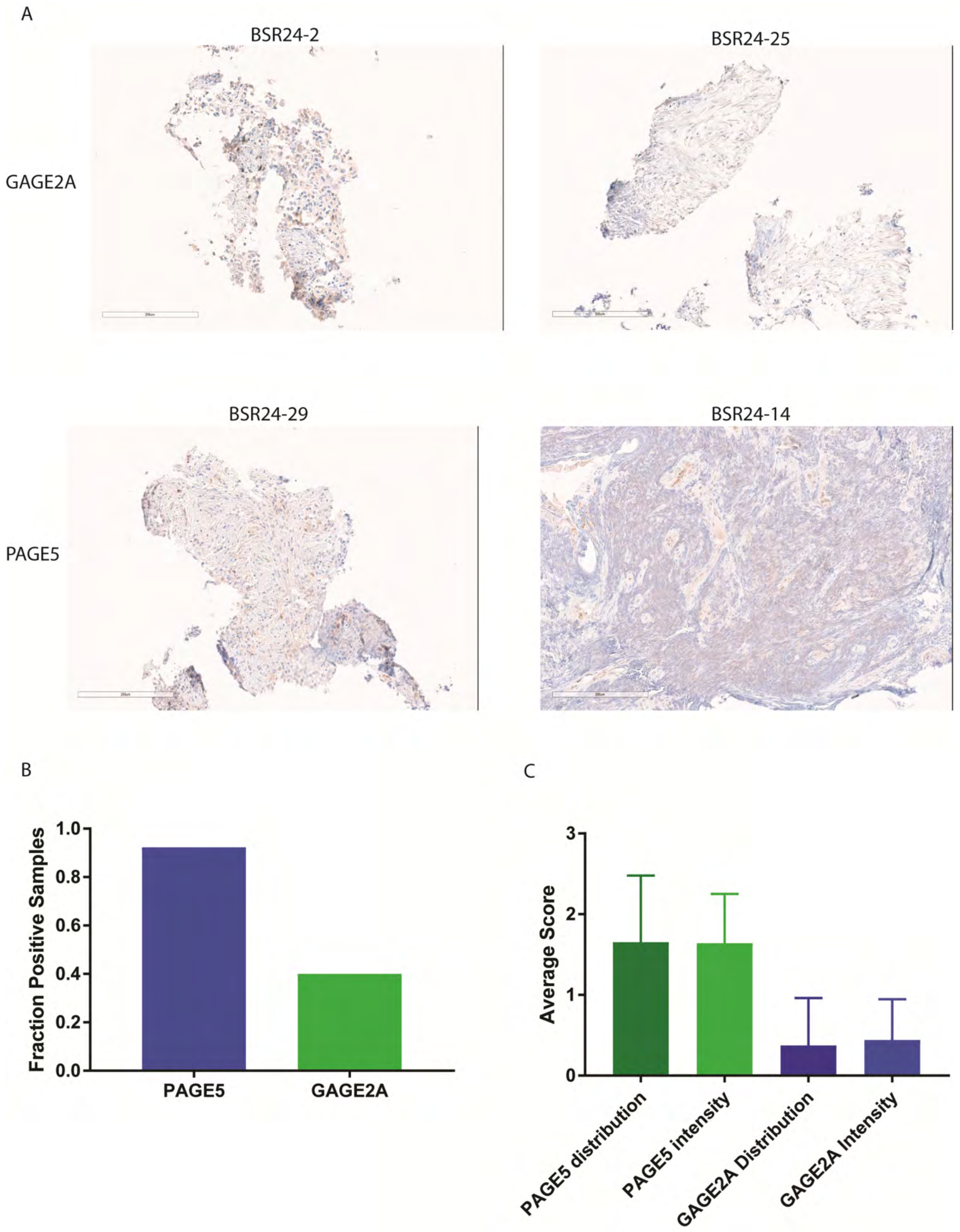


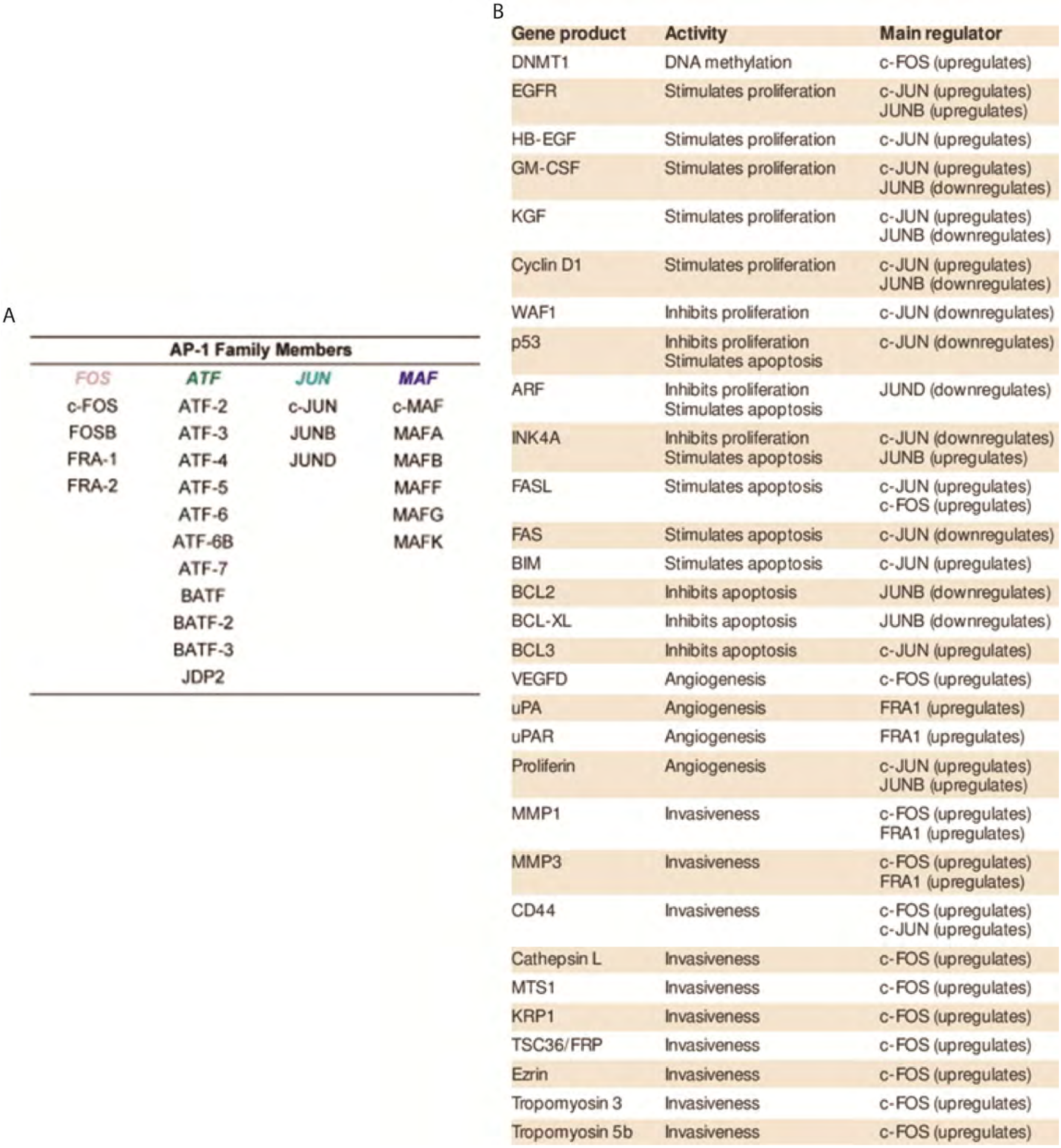
Figure 41: Staining and quantification of CTA expression in human SCLC samples
A: Representative images of the human SCLC biopsies stained with antibodies against PAGE5 (left) or GAGE2A (right). The stain appears as a rust-color DAB staining. **B:** Fraction of tumor samples that had any positive staining for PAGE5 or GAGE2A. **C:** Quantification of the distribution and intensity of the PAGE5 and GAGE2A staining. Scores range from 0 (no staining) to 3 (very widespread or strong).

Chapter 8: Discussion

In this work, I have developed the *in situ* barcoding model in SCLC. The tumors were successfully edited, and GFP+ cells were able to be isolated via FACS. Upon analysis of the scRNA-seq data, no barcodes were detectable. The scRNA-seq sequencing depth was not sufficient to detect the barcodes via scRNA-seq. Previous work has generally detected barcodes via DNA sequencing, which is then used to complement the information gained from scRNA-seq[40,46]. Future work in the *in situ* genetic barcoding field would benefit from insertion of the barcode at a more highly expressed gene than *Rosa26,* or by insertion of a strong promoter to drive expression of the barcode at a rate high enough to be detected with scRNA-seq. Alternately, an endogenous "barcode" such as the V(D)J region could be used to tag cells, and the 10X Genomics feature capture technology could be used to readily detect the barcodes. Despite the absence of the barcode in this model, there are still many insights that can be gathered from this data. Since tumors were sequenced at one month intervals from early after tumor initiation until very late stage tumor burden, we are able to evaluate the transcriptomic makeup of the tumors throughout the course of the disease and reconstruct a lineage hierarchy.

The *in situ* tumors largely display markers of the ASCL1-high SCLC-A subtype, which is expected from the RPR mouse model. At the onset of tumor development, the majority of the tumor cells share a common "early tumor" transcriptomic profile that is higher in stem-like and neuroendocrine markers. As the tumors progress, the proportion of tumor cells that belong to the early group decreases, and in its place are cells that belong to a second population of "late tumor" cells. These are characterized by having less cancer stem cell characteristics and more highly express members of the AP-1 family. The late tumor population is maintained as the primary cellular population through the latest tumor time-point. The identification of the early tumor population sheds light on the populations responsible for the earliest stages of neoplastic transformation. The late tumor population identifies the cells that maintain the tumor population over long periods of time.

One of the most highly expressed networks in the *in situ* dataset, particularly in the late tumor cluster, was the AP-1 family. The AP-1 (Activator Protein 1) complex is a powerful transcriptional controller comprised of members of the FOS, JUN and ATF families (Figure 42A)[64]. In development and differentiation in iPSCs, it plays a role in chromatin accessibility and helps to select enhancers to activate cell-specific networks in fibroblast differentiation[87,88]. It has been implicated in many tumor types and is responsible for tumor hallmarks like growth, metastasis, resistance to therapy, angiogenesis, activation of senescence pathways, cell cycle dysregulation, and inflammation, but the mechanisms by which the AP-1 network impacts tumor growth are often tumor type specific (Figure 42B)[65-76]. While the AP-1 network generally acts in an oncogenic fashion, sometimes family members can act as tumor suppressors. This is somewhat cell type or cancer type dependent and depends on which Jun protein is expressed most highly. cJun is most commonly associated with tumor progression and cell cycle dysregulation, while JunB is generally anti-proliferative, and can even drive expression of tumor suppressors[72], and JunD can act as either oncogenic or suppressive, depending on tumor type[67,73]. KRAS driven lung-adenocarcinoma further illustrates the duality of JunD and cJun, where JunD acts as a pro-tumorigenic factor in response to loss of cJun[73]. Also in lung adenocarcinoma, pharmacologic inhibition of AP-1 reduced metastatic formation but not tumorigenesis of an *ex vivo* metastasis model[89]. In lung adenocarcinoma xenografts, pharmacologic inhibition of an upstream activator of AP-1 signaling reduced both tumor proliferation and metastasis[90]. In melanoma, the AP-1 network has a role in the maintenance of cellular plasticity and heterogeneity by mediating cell state[91,92]. AP-1 family members can confer resistance to both MAPK inhibitors and BRAF inhibitors via activation of c-Jun[91,93]. Jun family members have differential roles in melanoma. Knockdown of c-Jun leads to cell cycle arrest and apoptosis, while knockdown of JunB leads to an increase in proliferation and tumorigenesis, due to an increase in cJun expression. The combination knockdown leads to apoptosis, indicating that JunB only acts as a tumor promoter in melanoma when c-Jun is knocked out[94]. cJun has also been implicated in liver cancer, where it acts independently of p53 to maintain cell survival in tumor initiation[95]. Similarly, in breast cancer, Levels of AP-1 family members were found to be significantly higher in cancer than in adjacent non-tumor tissues, and patients with high levels of cJun had a worse outcome[96,97]. In breast cancer cell culture and xenografts, knockdown of the AP-1 network by a dominant-negative cJun (DNJun) led to a decrease in proliferation overall and an inhibition of proliferation in response to growth factors[65,98]. In this context, the AP-1 network acts as a regulator of the cell cycle by regulating the expression of cyclins and CDKs, and drives progression by activation of pro-inflammatory cytokines[68,97]. In Prostate cancer expression of Fos and JunB are protective against advanced disease, but they are often lost as the tumor progresses, allowing for the upregulation of cJun, which then drives tumor progression in late stage[99]. Despite the links to tumorigenesis and resistance to therapy in other tumors, the role of the AP-1 network has yet to be evaluated in SCLC.



A

AP-1 Family Members			
FOS	ATF	JUN	MAF
c-FOS	ATF-2	c-JUN	c-MAF
FOSB	ATF-3	JUNB	MAFA
FRA-1	ATF-4	JUND	MAFB
FRA-2	ATF-5		MAFF
	ATF-6		MAFG
	ATF-6B		MAFK
	ATF-7		
	BATF		
	BATF-2		
	BATF-3		
	JDP2		

B

Gene product	Activity	Main regulator
DNMT1	DNA methylation	c-FOS (upregulates)
EGFR	Stimulates proliferation	c-JUN (upregulates) JUNB (upregulates)
HB-EGF	Stimulates proliferation	c-JUN (upregulates)
GM-CSF	Stimulates proliferation	c-JUN (upregulates) JUNB (downregulates)
KGF	Stimulates proliferation	c-JUN (upregulates) JUNB (downregulates)
Cyclin D1	Stimulates proliferation	c-JUN (upregulates) JUNB (downregulates)
WAF1	Inhibits proliferation	c-JUN (downregulates)
p53	Inhibits proliferation Stimulates apoptosis	c-JUN (downregulates)
ARF	Inhibits proliferation Stimulates apoptosis	JUND (downregulates)
INK4A	Inhibits proliferation Stimulates apoptosis	c-JUN (downregulates) JUNB (upregulates)
FASL	Stimulates apoptosis	c-JUN (upregulates) c-FOS (upregulates)
FAS	Stimulates apoptosis	c-JUN (downregulates)
BIM	Stimulates apoptosis	c-JUN (upregulates)
BCL2	Inhibits apoptosis	JUNB (downregulates)
BCL-XL	Inhibits apoptosis	JUNB (downregulates)
BCL3	Inhibits apoptosis	c-JUN (upregulates)
VEGFD	Angiogenesis	c-FOS (upregulates)
uPA	Angiogenesis	FRA1 (upregulates)
uPAR	Angiogenesis	FRA1 (upregulates)
Proliferin	Angiogenesis	c-JUN (upregulates) JUNB (upregulates)
MMP1	Invasiveness	c-FOS (upregulates) FRA1 (upregulates)
MMP3	Invasiveness	c-FOS (upregulates) FRA1 (upregulates)
CD44	Invasiveness	c-FOS (upregulates) c-JUN (upregulates)
Cathepsin L	Invasiveness	c-FOS (upregulates)
MTS1	Invasiveness	c-FOS (upregulates)
KRP1	Invasiveness	c-FOS (upregulates)
TSC36/FRP	Invasiveness	c-FOS (upregulates)
Ezrin	Invasiveness	c-FOS (upregulates)
Tropomyosin 3	Invasiveness	c-FOS (upregulates)
Tropomyosin 5b	Invasiveness	c-FOS (upregulates)

Figure 42: AP-1 complex members
A: The AP-1 complex is comprised of members of the FOS, ATF, JUN, or MAF families that complex together to activate transcriptional networks. Modified from Garces de Los Fayos Alonso et al., 2018. **B:** Target genes of the AP-1 complex and their role in cancer hallmarks. A number of gene targets of AP-1 ("gene product") are regulated by members of AP-1 ("main regulator"). The target genes have been found to play a role in a myriad of cancer hallmarks, but these seem to be tumor type-specific. Modified from Eferl and Wagner, 2003.

In this work, scRNA-seq identified members of the AP-1 network as playing a role in tumorigenesis of SCLC. Inhibition of the AP-1 network by transfection with a dominant-negative cJun construct significantly inhibited colony formation of SCLC cells in a clonogeneic assay. The AP-1 network is required for tumor maintenance in SCLC, and future work should warrant investigation *in vivo* of the impact of AP-1 inhibition in SCLC. The bioinformatics analysis also identified AP-1 as a potential mediator of chemoresistance in SCLC. This is also an important avenue for future follow-up.

A retroviral barcoding system has been generated to understand tumor dynamics and heterogeneity over time and under chemotherapeutic pressure in SCLC xenografts. Prior to the generation of the xenografts, extensive validation of the barcoding system was performed. The true diversity of barcodes was modeled based on data from targeted amplicon sequencing. The modeling is a significant advance on current barcoding reports, as estimating the true diversity is more informative than sequencing alone. Additional validation prior to xenografting revealed three doublings is optimal to ensure overlap in barcode populations between the pre-injection sample and the injected xenografts, and three doubling lowers the amount of overlap between two individual xenografts. Minimal overlap in barcodes between two xenografts is relevant because they may then serve as biological replicates. If the same population of cells is detected in two xenografts and they share a barcode, it may be that those cells shared a common lineage in the cells barcoded in culture and were simply injected in to the xenografts as clones. However, if the same population of cells is detected in two separate xenografts that have unique barcodes, there can be some confidence that that lineage arose independently in those two unique xenografts and it may represent a tumor-relevant biologic phenomenon. Understanding the overlap in barcodes by first performing the doubling time experiment led to confidence that the barcoding system was truly in place prior to xenografting.

Barcoded xenografts were generated and serially injected as barcoded xenografts that received chemoresistance. All tumors, as well as the initial cellular population were profiled with scRNA-seq, and barcodes were detected in all samples, although not in every cell. The reason barcodes were detectable from the xenografted tumors, but not the *in situ* barcoded tumors is likely due to the strong CAG promoter inserted with the GFP and barcode in to the cells used for xenografting. Additionally, the barcode sequence for the xenografts was inserted by use of a retrovirus, instead of CRISPR-Cas9. The efficiency of vial insertion is much higher than that of CRISPR-Cas9, leading to the increased insertion of barcodes in to the cells, and thereby increasing the likelihood of detecting barcodes by scRNA-seq. After chemotherapy, there is a broad transcriptomic shift in the SCLC-A xenografts. The SCLC-A barcoded xenografts begin as tumors that are very high in the expression of *ASCL1*, the marker for the SCLC-A subtype. As expected, the SCLC-N xenografts are very high in expression of *NEUROD1.* As the chemoresistant tumors develop, the SCLC-A tumors shift from high expression of *ASCL1* to *NEUROD1*, and the SCLC-N tumors remain consistently high in *NEUROD1.* SCLC-N tumors are often more chemoresistant, so the increased expression of *NEUROD1* in chemoresistant tumors is logical. SCLC tumors have been documented to have changes in molecular subtypes, particularly a shift from SCLC-A to SCLC-N, coinciding with a change in MYC signaling, which has been documented to play a role in the transition of SCLC subtypes[16,18]. In these tumors, a change in MYC signaling was indeed observed in the post-chemotherapeutic shift from SCLC-A to SCLC-N. Despite the broad transcriptomic shift observed pre-and post-chemotherapy in the SCLC-A tumors, there are a handful of cells from the pre-chemotherapy tumors that cluster more closely with the chemoresistant tumors. It is not known in SCLC if chemoresistant is inherent or induced. Inherent chemoresistance would result from a subset of tumor cells that already exist as cells with the inherent ability to be resistant to chemotherapy, which are then selected for upon treatment. Induced chemoresistance is a result of a subset of transcriptomically plastic cells that after chemotherapeutic pressure upregulate networks responsible for chemoresistance and develop a chemoresistant tumor. Without the barcoding system, it would be difficult to ascertain the mechanism leading to chemoresistance in SCLC. By utilizing the barcode system, the barcodes from the cells that form chemoresistant tumors can be tracked to cells in the pre-chemotherapy tumors, and their transcriptomes can be evaluated for changes that may have led to the development of chemoresistant tumors. If the transcriptomes from cells with matching barcodes are the same in the pre- and post-chemotherapy samples, that would indicate inherent chemoresistance in SCLC. If the transcriptomes of populations with matching barcodes change after chemotherapy, there would be evidence of induced chemoresistance after chemotherapy. In the SCLC-N chemoresistant tumors, there is a much more subtle transcriptomic change from pre- to post-chemotherapy. This is consistent with reports of SCLC-N tumors being more chemoresistant[16]. By pairing cellular populations that share the same barcodes pre- and post-chemotherapy in the SCLC-N tumors as well, we can ascertain whether the chemoresistance in SCLC-N tumors is inherent, or if there are subtle changes that lead to induced chemoresistance.

Similarly, there is a transcriptomic shift between the pre-injection samples and the resulting tumor, particularly in the SCLC-A tumors, indicating a bottleneck event that allowed for a subset of cells to form the resultant tumor. By matching cells that contain the same barcode in the pre-injection and pre-chemotherapy samples, we can evaluate the requirements for cells that are able to generate tumors. In the SCLC-N tumors, there is a smaller transcriptomic shift between the pre-injection sample and the formed tumor, which may indicate increased propensity for these cells to form tumors.

One of the most commonly upregulated families in the chemoresistant tumors was the cancer testis antigen (CTA) family. Cancer/Testis Antigens (CTA) are a large class of proteins almost exclusively expressed in the male germ cells and tumors. CTAs have shown promise as potentially targetable, unique cancer antigens. For this reason, they make excellent candidates for immunotherapy such as CAR-T therapy and cancer vaccines[77-83,100]. CTAs have also been shown to be diagnostic and prognostic in many cancers, although the particular CTA with prognostic or diagnostic value seems to be cancer-type specific[83,101]. In addition to their role as potential cancer antigens, CT antigens also have oncogenic effects on proliferation, genomic stability, invasion, colony formation, and resistance to apoptosis, and are associated with cancer stem cells (Figure 42A)[78,79,81,100,102-104]. In melanoma and synovial cell carcinoma, a CAR-T targeted to the CTA NY-ESO-1 led to complete response in a subset of patients, and is an ongoing avenue for investigation of new therapeutics (Figure 42B)[80,105]. Another study found the expression of CTAs to be drivers of breast cancer by increasing the HIF, WNT, and TGFbeta pathways[100]. In blood samples from patients with non-small cell lung cancer, the concentration of CTAs were significantly higher than in patients without cancer, and a panel of CTAs may serve as a blood-based diagnostic or screening test[101]. Very little work has investigated CTAs in SCLC. NY-ESO-1 has been found to be decreased in the blood of patients with SCLC, and would serve as an independent diagnostic indicator[106]. Another CTA, NOLA4, has been found to be significantly expressed in SCLC cell lines and serum from patients with SCLC, although the functional impact has not been evaluated[107]. A clinical trial used cell lysate from a large cell lung cancer cell line that is high in expression of CTAs as a cancer vaccine in patients with lung cancers or thoracic metastasis. Two of the 24 patients in this study had SCLC. Patients developed antibodies against the CTAs NY-ESO-1, GAGE7, and MAGE-C2. They also observed a decrease in the number of regulatory T cells and a decrease in the expression of PD-L1 on tumor-infiltrating immune cells in patients who got the cell lysate vaccine. A follow-up clinical trial of this study is ongoing to evaluate the utility of a cell lysate vaccine in patients with lung cancer[108]. PAGE5 (CT16) has been identified to be expressed in some cancers, and in melanoma was elevated as an anti-apoptotic gene in response to platinum-based chemotherapy. Expression of PAGE5 was shown to be pro-survival, and upregulated genes related to melanoma cell survival in response to chemotherapy[84]. GAGE2A is another anti-apoptotic CTA, that seems to be related to treatment resistance in medulloblastoma[85,86]. CTAs have been identified just once in SCLC, but have not been investigated[109]. While the study of CTAs in SCLC has been very limited, they have shown progress as prognostic, diagnostic, and therapeutic targets in other cancers, and are a promising avenue of exploration in SCLC.

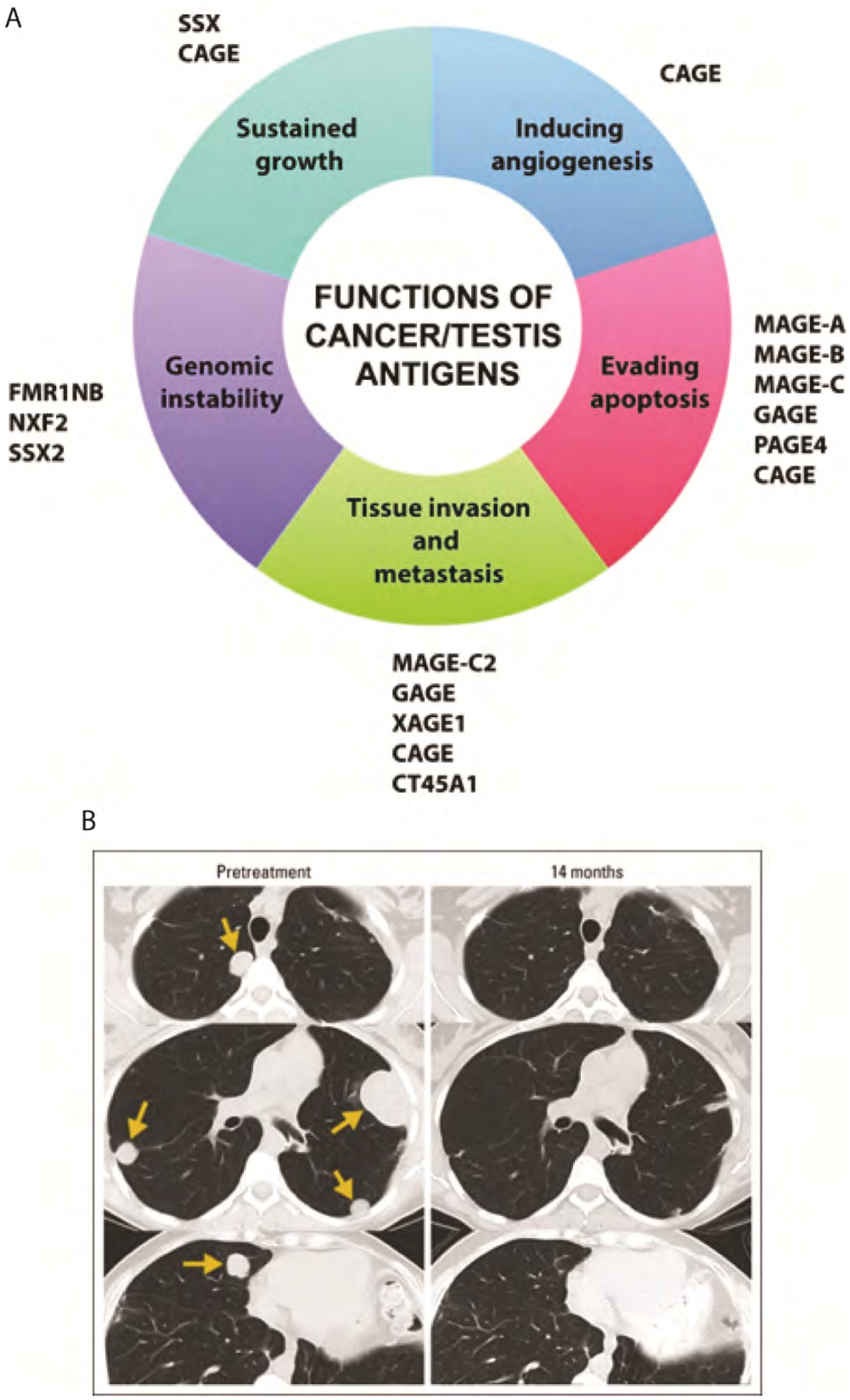


Figure 43: Cancer Testis Antigens have oncogenic functions.
A: Multiple CTAs can impact tumorigenesis, and the affect is CTA and cancer type-dependent. From Gjerstorff et al., Oncotarget, 2015. **B:** CT scan of a patient with lung metastases (arrowheads) from synovial cell carcinoma treated with NY-ESO-1-targeted T cells. At 14 months after treatment, a dramatic response to therapy can be observed by noting the absence of lung metastases. From Robbins et al., Journal of Clinical Oncology, 2011.

The CTAs PAGE5 and GAGE2A were highly expressed in the SCLC-A chemoresistant tumors, and were highly expressed in all of the SCLC-N tumors. I investigated the role of PAGE5 and GAGE2A in mediating resistance to chemotherapy in SCLC. Treatment of SCLC-A and SCLC-N cell lines in culture leads to robust upregulation of PAGE5 and GAGE2A in just a couple of days, which indicates that they may play a role in mediating cellular response to chemotherapy. Overexpression of PAGE5 or GAGE2A in SCLC cell lines leads to a decrease in chemotherapy-induced cell death. Expression of PAGE5 or GAGE2A in SCLC cell lines is sufficient to confer chemoresistance in culture. I knocked down PAGE5 or GAGE2A with an shRNA construct and found that knockdown of either PAGE5 or GAGE2A does not have any impact on response to chemotherapy. However, when both PAGE5 and GAGE2A are knocked down via shRNA, cells are significantly more sensitive to chemotherapy-induced cell death. PAGE5 and GAGE2A in SCLC act as mediators of chemoresistance, where expression of only one is sufficient to confer resistance to chemotherapy, but inhibition of both is required to sensitize cells to chemotherapy. To test the impact of PAGE5 and GAGE2A in conferring chemoresistance *in vivo,* xenografts using the shRNA knocked-down cell lines were generated. Many CTAs, including PAGE5 and GAGE2A, do not have a homologue in mice, so transgenic mouse models are not able to be used to investigate them, and it may be a reason that CTAs have not been identified in previous SCLC studies. Xenografts generated from cells with PAGE5 and GAGE2A knockdown seem to have a durable response to chemotherapy, indicating a role for PAGE5 and GAGE2A *in vivo* as well. Human SCLC biopsies stain positive for PAGE5 and GAGE2A, which further solidifies the association of CTAs with

In this dissertation, I have developed the genetic barcode lineage tracing system in SCLC. For the first time, we have a glimpse in to the events that lead to tumor initiation, clonal diversity, and chemoresistance in SCLC. The bioinformatics pipeline developed in this work represents a significant advance for the analysis of genetic barcode lineage tracing studies not only for uses in cancer, but also in other situations where clonal analysis is critical, such as developmental biology. I have described two distinct populations of cells that arise during tumor formation *in situ* that are maintained throughout the life of the tumor, while the relative proportion of cells in each cluster change during tumor progression. The AP-1 family was significantly upregulated in the late cluster, and has been validated as being critical for tumor initiation in SCLC. In xenograft studies, the clonal diversity of two subtypes of SCLC, SCLC-A and SCLC-N. The SCLC-A tumors exhibit more plasticity and transcriptomic shifts after chemotherapy than the SCLC-N. I have identified and validated the cancer testis antigens PAGE5 and GAGE2A as being mediators of chemoresistance in SCLC. Given the clinical success of other CTA-targeted therapies, this represents a promising avenue towards a new therapy for SCLC.

SCLC is a devastating disease, and very little progress has been made in generating truly targeted therapeutics. The expansion of the barcoding technology in this dissertation has allowed us to examine ITH in SCLC with unprecedented resolution for the first time. Identifying the AP-1 network as being responsible for tumorigenesis has provided knowledge of the critical early days of tumor formation in SCLC. Two targetable antigens, PAGE5 and GAGE2A have been identified in SCLC for the first time. Their role in mediating chemoresistance could be mitigated in the future with therapeutic antibodies, CAR-T cells, or even cancer vaccines. This work has contributed to both the basic and translational science and represents a significant advance towards a cure for this terrible disease.

Chapter 9: References

1 Drapkin, B. J. & Rudin, C. M. Advances in Small-Cell Lung Cancer (SCLC) Translational Research. *Cold Spring Harb Perspect Med* **11**, doi:10.1101/cshperspect.a038240 (2021).

2 Byers, L. A. & Rudin, C. M. Small cell lung cancer: where do we go from here? *Cancer* **121**, 664-672, doi:10.1002/cncr.29098 (2015).

3 Rudin, C. M., Brambilla, E., Faivre-Finn, C. & Sage, J. Small-cell lung cancer. *Nat Rev Dis Primers* **7**, 3, doi:10.1038/s41572-020-00235-0 (2021).

4 Kalemkerian, G. P. *et al.* NCCN Guidelines Insights: Small Cell Lung Cancer, Version 2.2018. *J Natl Compr Canc Netw* **16**, 1171-1182, doi:10.6004/jnccn.2018.0079 (2018).

5 Ganti, A. K. P. *et al.* Small Cell Lung Cancer, Version 2.2022, NCCN Clinical Practice Guidelines in Oncology. *Journal of the National Comprehensive Cancer Network* **19**, 1441-1464, doi:10.6004/jnccn.2021.0058 (2021).

6 Rudin, C. M. *et al.* Pembrolizumab or Placebo Plus Etoposide and Platinum as First-Line Therapy for Extensive-Stage Small-Cell Lung Cancer: Randomized, Double-Blind, Phase III KEYNOTE-604 Study. *J Clin Oncol* **38**, 2369-2379, doi:10.1200/jco.20.00793 (2020).

7 George, J. *et al.* Comprehensive genomic profiles of small cell lung cancer. *Nature* **524**, 47-53, doi:10.1038/nature14664 (2015).

8 Giacinti, C. & Giordano, A. RB and cell cycle progression. *Oncogene* **25**, 5220-5227, doi:10.1038/sj.onc.1209615 (2006).

9 Kareta, M. S. *et al.* Inhibition of pluripotency networks by the Rb tumor suppressor restricts reprogramming and tumorigenesis. *Cell Stem Cell* **16**, 39-50, doi:10.1016/j.stem.2014.10.019 (2015).

10 Rudin, C. M. *et al.* Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat Genet* **44**, 1111-1116, doi:10.1038/ng.2405 (2012).

11 Voigt, E. *et al.* Sox2 Is an Oncogenic Driver of Small-Cell Lung Cancer and Promotes the Classic Neuroendocrine Subtype. *Molecular Cancer Research* **19**, 2015-2025, doi:10.1158/1541-7786.Mcr-20-1006 (2021).

12 Wollenzien, H., Voigt, E. & Kareta, M. S. Somatic Pluripotent Genes in Tissue Repair, Developmental Disease, and Cancer. *SPG Biomed* **1**, doi:10.32392/biomed.18 (2018).

13 Chen, H. Z. *et al.* Genomic and Transcriptomic Characterization of Relapsed SCLC Through Rapid Research Autopsy. *JTO Clin Res Rep* **2**, 100164, doi:10.1016/j.jtocrr.2021.100164 (2021).

14 Zilfou, J. T. & Lowe, S. W. Tumor suppressive functions of p53. *Cold Spring Harb Perspect Biol* **1**, a001883, doi:10.1101/cshperspect.a001883 (2009).

15 Grunblatt, E. *et al.* MYCN drives chemoresistance in small cell lung cancer while USP7 inhibition can restore chemosensitivity. *Genes Dev* **34**, 1210-1226, doi:10.1101/gad.340133.120 (2020).

16 Ireland, A. S. *et al.* MYC Drives Temporal Evolution of Small Cell Lung Cancer Subtypes by Reprogramming Neuroendocrine Fate. *Cancer Cell* **38**, 60-78.e12, doi:https://doi.org/10.1016/j.ccell.2020.05.001 (2020).

17 Jahchan, N. S. *et al.* Identification and Targeting of Long-Term Tumor-Propagating Cells in Small Cell Lung Cancer. *Cell Rep* **16**, 644-656, doi:10.1016/j.celrep.2016.06.021 (2016).

18 Patel, A. S. *et al.* Prototypical oncogene family Myc defines unappreciated distinct lineage states of small cell lung cancer. *Science Advances* **7**, eabc2578, doi:10.1126/sciadv.abc2578 (2021).

19 Augert, A. *et al.* Targeting NOTCH activation in small cell lung cancer through LSD1 inhibition. *Science Signaling* **12**, eaau2922, doi:doi:10.1126/scisignal.aau2922 (2019).

20 Lim, J. S. *et al.* Intratumoural heterogeneity generated by Notch signalling promotes small-cell lung cancer. *Nature* **545**, 360-364, doi:10.1038/nature22323 (2017).

21 Leonetti, A. *et al.* Notch pathway in small-cell lung cancer: from preclinical evidence to therapeutic challenges. *Cell Oncol (Dordr)* **42**, 261-273, doi:10.1007/s13402-019-00441-3 (2019).

22 Baine, M. K. *et al.* SCLC Subtypes Defined by ASCL1, NEUROD1, POU2F3, and YAP1: A Comprehensive Immunohistochemical and Histopathologic Characterization. *J Thorac Oncol* **15**, 1823-1835, doi:10.1016/j.jtho.2020.09.009 (2020).

23 Gazdar, A. F., Carney, D. N., Nau, M. M. & Minna, J. D. 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 (1985).

24 Rudin, C. M. *et al.* Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nat Rev Cancer* **19**, 289-297, doi:10.1038/s41568-019-0133-9 (2019).

25 Böttger, F. *et al.* Tumor Heterogeneity Underlies Differential Cisplatin Sensitivity in Mouse Models of Small-Cell Lung Cancer. *Cell Rep* **27**, 3345-3358.e3344, doi:10.1016/j.celrep.2019.05.057 (2019).

26 Peifer, M. *et al.* Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat Genet* **44**, 1104-1110, doi:10.1038/ng.2396 (2012).

27 McFadden, D. G. *et al.* Genetic and clonal dissection of murine small cell lung carcinoma progression by genome sequencing. *Cell* **156**, 1298-1311, doi:10.1016/j.cell.2014.02.031 (2014).

28 Tripathi, S. C. *et al.* MCAM Mediates Chemoresistance in Small-Cell Lung Cancer via the PI3K/AKT/SOX2 Signaling Pathway. *Cancer Res* **77**, 4414-4425, doi:10.1158/0008-5472.Can-16-2874 (2017).

29 Ramón, Y. C. S. *et al.* Clinical implications of intratumor heterogeneity: challenges and opportunities. *J Mol Med (Berl)* **98**, 161-177, doi:10.1007/s00109-020-01874-2 (2020).

30 Terraneo, N., Jacob, F., Dubrovska, A. & Grünberg, J. Novel Therapeutic Strategies for Ovarian Cancer Stem Cells. *Frontiers in oncology* **10**, doi:10.3389/fonc.2020.00319 (2020).

31 Stewart, C. A. *et al.* Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. *Nat Cancer* **1**, 423-436, doi:10.1038/s43018-019-0020-z (2020).

32 Dentro, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239-2254.e2239, doi:10.1016/j.cell.2021.03.009 (2021).

33 Yang, D. *et al.* Intertumoral Heterogeneity in SCLC Is Influenced by the Cell Type of Origin. *Cancer Discov* **8**, 1316-1331, doi:10.1158/2159-8290.Cd-17-0987 (2018).

34 Rovira-Clavé, X. *et al.* Spatial epitope barcoding reveals subclonal tumor patch behaviors. *bioRxiv*, 2021.2006.2029.449991, doi:10.1101/2021.06.29.449991 (2021).

35 Chan, J. M. *et al.* Signatures of plasticity, metastasis, and immunosuppression in an atlas of human small cell lung cancer. *Cancer Cell* **39**, 1479-1496.e1418, doi:https://doi.org/10.1016/j.ccell.2021.09.008 (2021).

36 Gerrits, A. *et al.* Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* **115**, 2610-2618, doi:10.1182/blood-2009-06-229757 (2010).

37 Lan, X. *et al.* Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. *Nature* **549**, 227-232, doi:10.1038/nature23666 (2017).

38 Morgan, D., Jost, T. A., De Santiago, C. & Brock, A. Applications of high-resolution clone tracking technologies in cancer. *Curr Opin Biomed Eng* **19**, doi:10.1016/j.cobme.2021.100317 (2021).

39 Gardner, A., Morgan, D., Al'Khafaji, A. & Brock, A. Functionalized Lineage Tracing for the Study and Manipulation of Heterogeneous Cell Populations. *Methods Mol Biol* **2394**, 109-131, doi:10.1007/978-1-0716-1811-0_8 (2022).

40 Guernet, A. *et al.* CRISPR-Barcoding for Intratumor Genetic Heterogeneity Modeling and Functional Analysis of Oncogenic Driver Mutations. *Mol Cell* **63**, 526-538, doi:10.1016/j.molcel.2016.06.017 (2016).

41 Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science* **367**, doi:10.1126/science.aaw3381 (2020).

42 Hurley, K. *et al.* Reconstructed Single-Cell Fate Trajectories Define Lineage Plasticity Windows during Differentiation of Human PSC-Derived Distal Lung Progenitors. *Cell Stem Cell* **26**, 593-608.e598, doi:10.1016/j.stem.2019.12.009 (2020).

43 Ludwig, L. S. *et al.* Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics. *Cell* **176**, 1325-1339.e1322, doi:10.1016/j.cell.2019.01.022 (2019).

44 Eyler, C. E. *et al.* Single-cell lineage analysis reveals genetic and epigenetic interplay in glioblastoma drug resistance. *Genome Biology* **21**, 174, doi:10.1186/s13059-020-02085-1 (2020).

45 Neftel, C. *et al.* An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma. *Cell* **178**, 835-849.e821, doi:10.1016/j.cell.2019.06.024 (2019).

46 Quinn, J. J. *et al.* Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science* **371**, doi:10.1126/science.abc1944 (2021).

47 Gutierrez, C. *et al.* Multifunctional barcoding with ClonMapper enables high-resolution study of clonal dynamics during tumor evolution and treatment. *Nat Cancer* **2**, 758-772, doi:10.1038/s43018-021-00222-8 (2021).

48 Emert, B. L. *et al.* Variability within rare cell states enables multiple paths toward drug resistance. *Nat Biotechnol* **39**, 865-876, doi:10.1038/s41587-021-00837-3 (2021).

49 Echeverria, G. V. *et al.* Resistance to neoadjuvant chemotherapy in triple-negative breast cancer mediated by a reversible drug-tolerant state. *Sci Transl Med* **11**, doi:10.1126/scitranslmed.aav0936 (2019).

50 Ben-David, U. *et al.* Genetic and transcriptional evolution alters cancer cell line drug response. *Nature* **560**, 325-330, doi:10.1038/s41586-018-0409-3 (2018).

51 Li, H. *et al.* The landscape of cancer cell line metabolism. *Nat Med* **25**, 850-860, doi:10.1038/s41591-019-0404-8 (2019).

52 Yu, C. *et al.* High-throughput identification of genotype-specific cancer vulnerabilities in mixtures of barcoded tumor cell lines. *Nat Biotechnol* **34**, 419-423, doi:10.1038/nbt.3460 (2016).

53 Yang, D. *et al.* Lineage Recording Reveals the Phylodynamics, Plasticity and Paths of Tumor Evolution. *bioRxiv*, 2021.2010.2012.464111, doi:10.1101/2021.10.12.464111 (2021).

54 Rogers, Z. N. *et al.* Mapping the in vivo fitness landscape of lung adenocarcinoma tumor suppression in mice. *Nat Genet* **50**, 483-486, doi:10.1038/s41588-018-0083-2 (2018).

55 Lee, M. C. *et al.* A multiplexed in vivo approach to identify driver genes in small cell lung cancer. *bioRxiv*, 2022.2003.2028.485708, doi:10.1101/2022.03.28.485708 (2022).

56 Zhao, C., Teng, E. M., Summers, R. G., Jr., Ming, G. L. & Gage, F. H. Distinct morphological stages of dentate granule neuron maturation in the adult mouse hippocampus. *J Neurosci* **26**, 3-11, doi:10.1523/jneurosci.3648-05.2006 (2006).

57 Brambrink, T. *et al.* Sequential expression of pluripotency markers during direct reprogramming of mouse somatic cells. *Cell Stem Cell* **2**, 151-159, doi:10.1016/j.stem.2008.01.004 (2008).

58 Chu, V. T. *et al.* Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nat Biotechnol* **33**, 543-548, doi:10.1038/nbt.3198 (2015).

59 Chu, V. T. *et al.* Efficient generation of Rosa26 knock-in mice using CRISPR/Cas9 in C57BL/6 zygotes. *BMC Biotechnol* **16**, 4, doi:10.1186/s12896-016-0234-4 (2016).

60 Platt, R. J. *et al.* CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell* **159**, 440-455, doi:10.1016/j.cell.2014.09.014 (2014).

61 Schaffer, B. E. *et al.* Loss of p130 accelerates tumor development in a mouse model for human small-cell lung carcinoma. *Cancer Res* **70**, 3877-3883, doi:10.1158/0008-5472.can-09-4228 (2010).

62 Chiou, S. H. *et al.* Pancreatic cancer modeling using retrograde viral vector delivery and in vivo CRISPR/Cas9-mediated somatic genome editing. *Genes Dev* **29**, 1576-1585, doi:10.1101/gad.264861.115 (2015).

63 Ventura, A. *et al.* Cre-lox-regulated conditional RNA interference from transgenes. *Proc Natl Acad Sci U S A* **101**, 10380-10385, doi:10.1073/pnas.0403954101 (2004).

64 Wang, Z. Y. *et al.* Regulation of IL-10 gene expression in Th2 cells by Jun proteins. *J Immunol* **174**, 2098-2105, doi:10.4049/jimmunol.174.4.2098 (2005).

65 Liu, Y. *et al.* Inhibition of AP-1 transcription factor causes blockade of multiple signal transduction pathways and inhibits breast cancer growth. *Oncogene* **21**, 7680-7689, doi:10.1038/sj.onc.1205883 (2002).

66 Wu, Z., Nicoll, M. & Ingham, R. J. AP-1 family transcription factors: a diverse family of proteins that regulate varied cellular activities in classical hodgkin lymphoma and ALK+ ALCL. *Experimental Hematology & Oncology* **10**, 4, doi:10.1186/s40164-020-00197-9 (2021).

67 Eferl, R. & Wagner, E. F. AP-1: a double-edged sword in tumorigenesis. *Nature Reviews Cancer* **3**, 859-868, doi:10.1038/nrc1209 (2003).

68 Liu, Y. *et al.* AP-1 blockade in breast cancer cells causes cell cycle arrest by suppressing G1 cyclin expression and reducing cyclin-dependent kinase activity. *Oncogene* **23**, 8238-8246, doi:10.1038/sj.onc.1207889 (2004).

69 Lopez-Bergami, P., Lau, E. & Ronai, Z. Emerging roles of ATF2 and the dynamic AP1 network in cancer. *Nat Rev Cancer* **10**, 65-76, doi:10.1038/nrc2681 (2010).

70 Martínez-Zamudio, R. I. *et al.* AP-1 imprints a reversible transcriptional programme of senescent cells. *Nature Cell Biology* **22**, 842-855, doi:10.1038/s41556-020-0529-5 (2020).

71 Garces de Los Fayos Alonso, I. *et al.* The Role of Activator Protein-1 (AP-1) Family Members in CD30-Positive Lymphomas. *Cancers (Basel)* **10**, doi:10.3390/cancers10040093 (2018).

72 Shaulian, E. & Karin, M. AP-1 in cell proliferation and survival. *Oncogene* **20**, 2390-2400, doi:10.1038/sj.onc.1204383 (2001).

73 Ruiz, E. J. *et al.* JunD, not c-Jun, is the AP-1 transcription factor required for Ras-induced lung cancer. *JCI Insight* **6**, doi:10.1172/jci.insight.124985 (2021).

74 Orlando, K. A. *et al.* Re-expression of SMARCA4/BRG1 in small cell carcinoma of ovary, hypercalcemic type (SCCOHT) promotes an epithelial-like gene signature through an AP-1-dependent mechanism. *Elife* **9**, doi:10.7554/eLife.59073 (2020).

75 Inoue, Y. *et al.* Extracellular signal-regulated kinase mediates chromatin rewiring and lineage transformation in lung cancer. *Elife* **10**, doi:10.7554/eLife.66524 (2021).

76 Tyagi, A. *et al.* Cervical cancer stem cells manifest radioresistance: Association with upregulated AP-1 activity. *Sci Rep* **7**, 4781, doi:10.1038/s41598-017-05162-x (2017).

77 Al-Khadairi, G. & Decock, J. Cancer Testis Antigens and Immunotherapy: Where Do We Stand in the Targeting of PRAME? *Cancers (Basel)* **11**, doi:10.3390/cancers11070984 (2019).

78 Gjerstorff, M. F., Andersen, M. H. & Ditzel, H. J. Oncogenic cancer/testis antigens: prime candidates for immunotherapy. *Oncotarget* **6**, 15772-15787, doi:10.18632/oncotarget.4694 (2015).

79 Jakobsen, M. K. & Gjerstorff, M. F. CAR T-Cell Cancer Therapy Targeting Surface Cancer/Testis Antigens. *Frontiers in Immunology* **11**, doi:10.3389/fimmu.2020.01568 (2020).

80 Thomas, R. *et al.* NY-ESO-1 Based Immunotherapy of Cancer: Current Perspectives. *Frontiers in Immunology* **9**, doi:10.3389/fimmu.2018.00947 (2018).

81 Wei, X. *et al.* Cancer-Testis Antigen Peptide Vaccine for Cancer Immunotherapy: Progress and Prospects. *Transl Oncol* **12**, 733-738, doi:10.1016/j.tranon.2019.02.008 (2019).

82 Scanlan, M. J., Gure, A. O., Jungbluth, A. A., Old, L. J. & Chen, Y. T. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. *Immunol Rev* **188**, 22-32, doi:10.1034/j.1600-065x.2002.18803.x (2002).

83 Yao, J. *et al.* Tumor subtype-specific cancer-testis antigens as potential biomarkers and immunotherapeutic targets for cancers. *Cancer Immunol Res* **2**, 371-379, doi:10.1158/2326-6066.Cir-13-0088 (2014).

84 Nylund, C. *et al.* Melanoma-associated cancer-testis antigen 16 (CT16) regulates the expression of apoptotic and antiapoptotic genes and promotes cell survival. *PLoS One* **7**, e45382, doi:10.1371/journal.pone.0045382 (2012).

85 Kasuga, C. *et al.* Expression of MAGE and GAGE genes in medulloblastoma and modulation of resistance to chemotherapy. Laboratory investigation. *J Neurosurg Pediatr* **1**, 305-313, doi:10.3171/ped/2008/1/4/305 (2008).

86 Cilensek, Z. M., Yehiely, F., Kular, R. K. & Deiss, L. P. A member of the GAGE family of tumor antigens is an anti-apoptotic gene that confers resistance to Fas/CD95/APO-1, Interferon-gamma, taxol and gamma-irradiation. *Cancer Biol Ther* **1**, 380-387 (2002).

87 Vierbuchen, T. *et al.* AP-1 Transcription Factors and the BAF Complex Mediate Signal-Dependent Enhancer Selection. *Mol Cell* **68**, 1067-1082.e1012, doi:10.1016/j.molcel.2017.11.026 (2017).

88 Madrigal, P. & Alasoo, K. AP-1 Takes Centre Stage in Enhancer Chromatin Dynamics. *Trends in Cell Biology* **28**, 509-511, doi:https://doi.org/10.1016/j.tcb.2018.04.009 (2018).

89 Mishra, D. K. & Kim, M. P. SR 11302, an AP-1 Inhibitor, Reduces Metastatic Lesion Formation in Ex Vivo 4D Lung Cancer Model. *Cancer Microenvironment* **10**, 95-103, doi:10.1007/s12307-017-0202-0 (2017).

90 Lee, Y. S. *et al.* A small molecule targeting CHI3L1 inhibits lung metastasis by blocking IL-13Rα2-mediated JNK-AP-1 signals. *Mol Oncol* **16**, 508-526, doi:10.1002/1878-0261.13138 (2022).

91 Comandante-Lou, N., Baumann, D. G. & Fallahi-Sichani, M. AP-1 transcription factor network explains diverse patterns of cellular plasticity in melanoma. *bioRxiv*, 2021.2012.2006.471514, doi:10.1101/2021.12.06.471514 (2021).

92 Riesenberg, S. *et al.* MITF and c-Jun antagonism interconnects melanoma dedifferentiation with pro-inflammatory cytokine responsiveness and myeloid cell recruitment. *Nat Commun* **6**, 8755, doi:10.1038/ncomms9755 (2015).

93 Ramsdale, R. *et al.* The transcription cofactor c-JUN mediates phenotype switching and BRAF inhibitor resistance in melanoma. *Science Signaling* **8**, ra82-ra82, doi:doi:10.1126/scisignal.aab1111 (2015).

94 Gurzov, E. N., Bakiri, L., Alfaro, J. M., Wagner, E. F. & Izquierdo, M. Targeting c-Jun and JunB proteins as potential anticancer cell therapy. *Oncogene* **27**, 641-652, doi:10.1038/sj.onc.1210690 (2008).

95 Min, L. *et al.* Liver cancer initiation is controlled by AP-1 through SIRT6-dependent inhibition of survivin. *Nat Cell Biol* **14**, 1203-1211, doi:10.1038/ncb2590 (2012).

96 Kharman-Biz, A. *et al.* Expression of activator protein-1 (AP-1) family members in breast cancer. *BMC Cancer* **13**, 441, doi:10.1186/1471-2407-13-441 (2013).

97 Qiao, Y. *et al.* AP-1 Is a Key Regulator of Proinflammatory Cytokine TNF&#x3b1;-mediated Triple-negative Breast Cancer Progression *. *Journal of Biological Chemistry* **291**, 5068-5079, doi:10.1074/jbc.M115.702571 (2016).

98 Ibrahim, S. A. E. *et al.* The role of AP-1 in self-sufficient proliferation and migration of cancer cells and its potential impact on an autocrine/paracrine loop. *Oncotarget* **9**, 34259-34278, doi:10.18632/oncotarget.26047 (2018).

99 Riedel, M. *et al.* Targeting AP-1 transcription factors by CRISPR in the prostate. *Oncotarget* **12** (2021).

100 Maxfield, K. E. *et al.* Comprehensive functional characterization of cancer–testis antigens defines obligate participation in multiple hallmarks of cancer. *Nature Communications* **6**, 8840, doi:10.1038/ncomms9840 (2015).

101 Zhang, R., Ma, L., Li, W., Zhou, S. & Xu, S. Diagnostic value of multiple tumor-associated autoantibodies in lung cancer. *Onco Targets Ther* **12**, 457-469, doi:10.2147/ott.S187734 (2019).

102 Gordeeva, O. Cancer-testis antigens: Unique cancer stem cell biomarkers and targets for cancer therapy. *Semin Cancer Biol* **53**, 75-89, doi:10.1016/j.semcancer.2018.08.006 (2018).

103 Yin, B. *et al.* MAGE-A3 is highly expressed in a cancer stem cell-like side population of bladder cancer cells. *Int J Clin Exp Pathol* **7**, 2934-2941 (2014).

104 Taguchi, A. *et al.* A search for novel cancer/testis antigens in lung cancer identifies VCX/Y genes, expanding the repertoire of potential immunotherapeutic targets. *Cancer Res* **74**, 4694-4705, doi:10.1158/0008-5472.Can-13-3725 (2014).

105 Robbins, P. F. *et al.* Tumor regression in patients with metastatic synovial cell sarcoma and melanoma using genetically engineered lymphocytes reactive with NY-ESO-1. *J Clin Oncol* **29**, 917-924, doi:10.1200/jco.2010.32.2537 (2011).

106 Yang, J., Jiao, S., Kang, J., Li, R. & Zhang, G. Application of serum NY-ESO-1 antibody assay for early SCLC diagnosis. *Int J Clin Exp Pathol* **8**, 14959-14964 (2015).

107 Kim, Y. R. *et al.* Cancer Testis Antigen, NOL4, Is an Immunogenic Antigen Specifically Expressed in Small-Cell Lung Cancer. *Curr Oncol* **28**, 1927-1937, doi:10.3390/curroncol28030179 (2021).

108 Zhang, M. *et al.* Randomized phase II trial of a first-in-human cancer cell lysate vaccine in patients with thoracic malignancies. *Transl Lung Cancer Res* **10**, 3079-3092, doi:10.21037/tlcr-21-1 (2021).

109 Gardner, E. E. *et al.* Chemosensitive Relapse in Small Cell Lung Cancer Proceeds through an EZH2-SLFN11 Axis. *Cancer Cell* **31**, 286-299, doi:10.1016/j.ccell.2017.01.006 (2017).

Appendix 1

SPG Biomed, 2018, doi: 10.32392/biomed.18

Somatic Pluripotent Genes in Tissue Repair, Developmental Disease, and Cancer

Hannah Wollenzien[1,2,]*, Ellen Voigt[1,]*, Michael S. Kareta[1,2,3,4]

[1]Genetics and Genomics Group, Cellular Therapies and Stem Cell Biology Group, and the Cancer Biology and Immunotherapies Group, Sanford Research, 2301 East 60th Street North, Sioux Falls, SD 57104, USA. [2]Division of Basic Biomedical Sciences, Sanford School of Medicine, University of South Dakota, 414 E. Clark St. Vermillion, SD 57069, USA. [3]Department of Pediatrics, Sanford School of Medicine, 1400 W. 22nd St., Sioux Falls, SD 57105, USA. [4]Department of Chemistry and Biochemistry, South Dakota State University, 1175 Medary Ave, Brookings, SD 57006, USA.

* Equal Contribution

Embryonic stem cells possess the ability to differentiate into all cell types of the body. This pliable developmental state is achieved by the function of a series of pluripotency factors, classically identified as *OCT4*, *SOX2*, and *NANOG*. These pluripotency factors are responsible for activating the larger pluripotency networks and the self-renewal programs which give ES cells their unique characteristics. However, during differentiation pluripotency networks become downregulated as cells achieve greater lineage specification and exit the cell cycle. Typically the repression of pluripotency is viewed as a positive factor to ensure the fidelity of cellular identity by restricting cellular pliancy. Consistent with this view, the expression of pluripotency factors is greatly restricted in somatic cells. However, there are examples whereby cells either maintain or reactivate pluripotency factors to preserve the increased potential for the healing of wounds or tissue homeostasis. Additionally there are many examples where these pluripotency factors become reactivated in a variety of human pathologies, particularly cancer. In this review, we will summarize the somatic repression of pluripotency factors, their role in tissue homeostasis and wound repair, and the human diseases that are associated with pluripotency factor misregulation with an emphasis on their role in the etiology of multiple cancers.

THE CORE PLURIPOTENCY NETWORK

Pluripotency factors regulate a host of biological processes essential to establishing the embryonic state. Of these, three factors, *SOX2*, *OCT4*, and *NANOG*, have been identified as the three core factors regulating cellular pluripotency[1-3]. Beginning in the early embryo, *SOX2*, *OCT4*, and *NANOG* are expressed in the inner cell mass (ICM) of the developing blastocyst and are required for the maintenance of pluripotency, and upon embryonic differentiation these factors are downregulated[4-8]. These core factors are so vital for the maintenance of a pluripotent state that they have now become part of the standard reprogramming cocktail for the generation of induced pluripotent stem (iPS) cells[9]. *OCT4* and *SOX2*, along with *c-MYC* and *KLF4*, are crucial for the generation of iPS cells, and the gene expression profile of these iPS cells is nearly identical to that of embryonic stem (ES) cells, illustrating their importance for maintaining the stem cell phenotype[10]. Indeed the ability of these reprogramming factors has in part given them the designation of master regulators, where they can activate target genes even when epigenetically repressed[11,12]. Therefore, due to the powerful transcriptional effects of these pluripotency genes, they must be subject to rigorous regulation throughout development to restrict their activation and allow for proper development.

SILENCING OF PLURIPOTENCY IN THE SOMA

Given that pluripotency is restricted to the ICM of the blastocyst a mechanism of silencing in somatic tissues should exist. It has been found that in ES cells the core pluripotency genes are marked by the activating histone modification histone H3 lysine 4 trimethylation (H3K4me3), and then during differentiation this mark is replaced by the silencing histone 3 lysine 27 trimethylation (H3K27me3) mark[13,14]. Concurrent with this regulation of histone methylation, there is a clear correlation of DNA methylation on the epigenetic regulation of the core pluripotency genes. The DNA at the promoters of the core genes are typically unmethylated in the embryonic state, however, they become rapidly methylated during differentiation, although there are some cases where *Sox2* evades DNA methylation[14,15]. This regulation is mediated in part by the activity of both DNMT activity in ES cells and the DNA demethylases such as TET1. *Oct4* specifically is methylated both at enhancer and promoter regions during the differentiation process and is dependent on DNMT3a and DNMT1 for this methylation[16]. When *Tet1* is downregulated, the *Nanog* promoter becomes methylated and it is subsequently silenced[17]. TET proteins including TET1 and TET2, and the DNMT3 family are crucial for methylating DNA during differentiation and silencing of pluripotent genes. In a study evaluating the epigenome of differentiated and ES cells, the DNA cytosine methylation in ES cells was mostly in a non-CpG context. These marks were associated with gene bodies and were greatly depleted as cells differentiated. The reduced non-CpG methylation was associated with lower transcriptional activity of developmentally relevant genes in differentiated cells, indicating that non-CpG DNA cytosine methylation might be key for the regulation of developmental genes[18]. Pluripotency genes may also be regulated by miRNAs. It was found that *let-7* miRNAs suppress self-renewal in ES cells and their downregulation was able to de-differentiate somatic cells to iPS cells. *Let-7* miRNAs are able to directly downregulate *Oct4*, *Sox2*, and *Nanog* and likely contribute to the stability of the differentiated state[19].

TISSUE HOMEOSTASIS AND WOUND HEALING

Pluripotency networks are not only crucial for the differentiation and organogenesis of embryonic tissues, but there is increasing evidence that tissue homeostasis and regeneration could involve the temporary acquisition of pluripotent gene networks. To maintain these tissues rare populations of adult stem cells actively dividing and differentiating[20,21]. In particular, *Sox2*, *Oct4*, and *Nanog* are involved in maintaining the plasticity of these adult stem cells.

Sox2 in Homeostasis and Wound Healing

Sox2 remains expressed in many adult tissues including the sperm cells, cervix, gut, esophagus, trachea, bronchiolar epithelium, the brain and sensory cells like the retina and taste buds[22,23]. These *Sox2*$^+$ cells originate from *Sox2*$^+$ progenitors and are essential for the maintenance of these tissues[22]. *Sox2*$^+$ cells have also been found in the adult brain in sites such as the white matter, cerebellum, and the hippocampus[24-26]. In the hippocampus, *Sox2* is required for the maintenance of neural stem cells during adulthood[26]. Beyond maintenance of the adult brain, *Sox2* expression has been shown to be upregulated in response to invasive brain injuries by activation of Notch and Sonic hedgehog signaling [27,28]. Sox2 is also required for the maintenance of many types of neuroendocrine cells throughout the body[29-31].

Similarly, *Sox2* expressing cells are present in other non-neural or neuroendocrine tissues in the adult as well. A population of *Sox2* expressing cells is found in the adult pituitary and help it regenerate in response to injury[32-35]. There are similar mechanisms throughout the body including the trachea and the intestinal crypts where *Sox2* expressing cells maintain and repair these tissues[36,37]. Furthermore, Sox2 is required for osteoblast function and self-renewal[38]. Therefore there is a significant role for *SOX2* in the development and maintenance of many tissues outside of the embryonic state.

Oct4 and Nanog in Homeostasis and Wound Healing

Mainly *Oct4,* sometimes in combination with *Nanog,* has been shown to be expressed in a variety of adult tissues, most commonly seen in hematopoietic and mesenchymal progenitors found in the bone marrow[39-43]. *Oct4* is also found in a wide variety of other progenitors in different body tissues, yet *Oct4* expression is not required for tissue homeostasis in the same way as *Sox2*[44]. The one exception is the need for *Oct4* expression for the viability of adult germ cells[45,46].

Although *Oct4* itself may not be required for tissue regeneration like *Sox2*, small populations of cells in the body that exhibit stem-ness population have been seen. A population of cells called very small embryonic-like cells (VSELs) has been discovered in many adult tissues that do express *Oct4* and *Nanog* and are able to differentiate into all the germ layers but not self-renew[47,48]. It is unknown if these VSELs play a role in tissue homeostasis in contrast to other *Oct4*$^{+}$ progenitor cells in the adult[48].

ABERRANT PLURIPOTENCY FACTOR EXPRESSION IN DEVELOPMENTAL DISEASE

Due to the importance of the core pluripotency factors in the establishment of ES and iPS cells, it is no surprise that mutations in these factors can cause developmental diseases. As *Sox2* remains expressed past the blastocyst stage and into organogenesis, mutations in the gene can cause a multitude of developmental defects (Table 1)[23,49]. In contrast, *Oct4* and *Nanog* are largely not expressed after the early stages of development, but they do contribute to the viability of germ cells[50-53]. In the past two decades, scientists have attributed many developmental problems to misregulation of these core factors, predominantly *SOX2*.

The Role of Sox2 in Early Development

The transcription factor *Sox2* is necessary for development from the earliest stages after conception. It has been shown that most *Sox2*$^{-/-}$ zygotes arrest as morulas, although a few can survive to become blastocysts where they fail at implantation[7,54]. In the blastocyst stage, *Sox2* is expressed as the earliest marker of the inner cell mass, and the trophectoderm[54,55] *Sox2* continues to be expressed in the extraembryonic endoderm as well as the primitive ectoderm [7]. As the germ layers are formed, *Sox2* is upregulated in cells that choose the neural ectoderm fate and suppresses the formation of mesoderm [56].

Sox2 in Neural/Sensory System Development and Disease

Sox2 is present in the neuroectoderm from early stages, and remains expressed in neural stem cells to promote survival in the central and peripheral nervous system[57,58]. In early development, the brain forms normally without *Sox2* and no defects are seen at midgestation in the mouse[59]. However, mutations in *Sox2* do cause defects in postnatal mouse development in the telencephalon, particularly in the hippocampus dentate gyrus through misregulation of sonic hedgehog signaling[26,60]. In later fetal development, *Sox2* is strongly expressed in the thalamus and hypothalamus[60,61]. It is no surprise that mutations in *Sox2* have been known to affect the formation of the hypothalamus-pituitary system, by causing hypoplasia of the anterior pituitary and gonadotrophin deficiency, resulting in fertility deficiencies[60,62]. Mutations in *Sox2* have also been shown to affect eye development, causing anophthalmia or microthalmia[63-65]. These defects are caused by misregulation of differentiation in the optic cup by disruption of *Notch1* signaling and *Pax6* function which are both orchestrated by *Sox2* function[63,64,66]. Other sensory systems are affected as well including the development of the cochlea and regulation of WNT signaling to form taste buds[65,67,68]. *Sox2* mutations can result in these defects occurring together: coloboma, heart malformation, atresia of the choanae, retarded growth and development, and genital and ear abnormalities or (CHARGE) syndrome as a result[65].

Sox2 in Gut, Lung, Kidney System Development and Disease

Sox2 is involved in the development of other organs, such as the gut where it is essential for anterior and posterior patterning and guides the tissue towards a gastric fate over an intestinal identity[69-71]. In the development of the foregut, *Sox2* is expressed to form the trachea, esophagus, and the esophageal epithelium[70,72]. If *Sox2* is mutated, this can sometimes result in anophthalmia,-esophageal-genital syndrome (AEG) where the formation of the esophagus and trachea is abnormal and these structures fail to separate[63,73]. Once the lungs have formed, *Sox2* is essential for normal lung branching and the maintenance of lung progenitor cells[29,74]. *Sox2* mutations have also been implicated in chronic kidney disease[65,75].

Oct4 in Early Development

Oct4 is present throughout the morula, expressed highly in the inner cell mass, and promotes differentiation into primitive endoderm[76-78]. As the blastocyst differentiates into the germ layers, *Oct4* specifies mesoderm while suppressing neural ectoderm[56,79,80].

Oct4 in System Development and Disease

Although *Oct4* plays an important role in early development, it is silenced in embryonic stem cells, and not expressed in the development of the organs with the exception of the primordial germ cells[44,52,81]. Oct4 is necessary for the switch from the pluripotent stem cells to the germ cells, thus problems with this mechanism can lead to infertility[51-53]. Although mutations in *Oct4* itself do not cause any developmental diseases directly, the misregulation of many of Oct4's binding partners is associated with diseases[82].

The Role of Nanog in Development and Developmental Disease

Nanog appears in the late morula, the blastocyst and is expressed in the inner cell mass[83]. *Nanog*$^{-/-}$ blastocysts cannot survive, although after implantation *Nanog* becomes downregulated[83,84]. *Nanog* is commonly expressed temporally and spatially with *Oct4*[83,85,86]. Similarly to *Oct4*, *Nanog* is not expressed in the tissues after early development except for the primordial germ cells where it necessary for the PGCs to mature on the germ ridge[50].

ABERRANT PLURIPOTENCY FACTOR EXPRESSION IN CANCER

The role of Sox2 in Cancer

In recent years, much work has begun to elucidate the role and association of *Sox2* in cancer in a vast array of human and mouse tumor types (Table 2). In a chemically-induced model of mouse squamous cell carcinoma, *Sox2* enriched cells were the tumor propagating cells, and conditional deletion of *Sox2* decreased tumor formation and led to regression in existing tumors[87]. *Sox2* expression was required for tumorigenicity of mouse osteosarcoma and knockout of *Sox2* decreases the cancer stem cell-like phenotype seen in *Sox2*$^{+}$ osteosarcoma cells[88]. In both human and mouse bladder cancer, *Sox2* is overexpressed in pre-neoplastic and neoplastic tumors, where the knockout of *Sox2* decreased tumor invasiveness[89]. Given that *Sox2* is a pluripotency gene, it is unsurprising that expression of *Sox2* in human glioblastoma multiforme (GBM) cells was able to direct differentiation in to a stem-like state capable of tumor propagation[90]. Also in human glioma and glioblastomas, *Sox2* expression had a positive correlation with tumor grade. In this cohort, *Sox2* expression was highest in hypercellular areas with highly proliferative cells[91]. A separate study verifies these results by showing that *Sox2* is decreased in more differentiated GBM samples, and overexpression of *Sox2* in cell culture leads to increased proliferation and stemness[92]. GBM cells in culture are dependent on *Sox2* to proliferate and form colonies and knockout of *Sox2* reduced these phenomena[93]. Human ER-positive breast cancer cells in culture that were resistant to tamoxifen therapy had high levels of *Sox2*. In fact, in a cohort of patients with ER-positive breast cancer, *Sox2* was more highly expressed in those who were not responsive to treatment, compared to patients whose cancer was responsive to treatment. In a larger patient set, *Sox2* expression was found to be prognostic of poor overall and disease-free survival[94]. Despite the high expression of *Sox2* seen in patients with ER-positive breast cancer in Piva *et al.*, patients with sporadic, basal-like breast cancer in a separate cohort had an inverse relationship between *Sox2* expression and ER expression[95]. In two cervical cancer lines, *Sox2* was overexpressed and marked a subset of stem-like cells[96]. *Sox2* has also been found to be upregulated in liquid tumors such as ALD-positive large-cell lymphoma, in which *Sox2* expression imparts a more “plastic” phenotype[97]. Finally, *Sox2* has been implicated in the switch to androgen resistance and involves the function of the tumor suppressors Rb1 and p53[98].

In addition to the studies linking Sox2 expression to cancer phenotypes, a number of studies have shown an association between expression of *Sox2* and clinical outcome. In breast cancer, it was suggestive that *Sox2* expression could be a biomarker of resistance to therapy, as well as poor overall and disease free survival[94]. *Sox2* expression in head and neck squamous cell carcinoma was associated with tumor recurrence and poor prognosis[99]. In tongue squamous cell carcinoma, *Sox2* expression is significantly associated with tumor stage, cell differentiation, and metastasis[100]. A large study of patients with gastric cancer who had undergone surgical resection of the tumor found that *Sox2* positivity was correlated with invasion depth, lymph node metastasis or invasion, and that the prognosis of patients with *Sox2* positive cancers was significantly worse than the prognosis of patients who had *Sox2* negative cancers[101]. In a study of non small-cell lung cancer samples, *Sox2* was significantly overexpressed in cancer cells, and not in preneoplastic or healthy tissues, although no correlation with histopathological data was seen in this study[102]. Interestingly, in synovial sarcoma, *Sox2* was expressed at relatively low levels and had no correlation to clinicopathological data[103].

Much work has shown that *Sox2* is expressed in a wide array of cancers. However, the exact molecular mechanism of *Sox2* activation in cancer is unknown, although there are several hypotheses for how *Sox2* drives tumor dynamics. A 2014 study by Justilien *et al* found an overexpression of *SOX2* by way of amplification of chromosome 3q26 in five human lung cancer cell lines. Mechanistically, it was found that PKCi, which is also amplified on chromosome 3q26, phosphorylates SOX2, which regulates SOX2 binding to hedgehog acyl transferase (HHAT). HHAT is crucial for hedgehog ligand binding and activation by SOX2 binding leads to downstream hedgehog activation. In the lung cancer lines studied, the expression of *Sox2*, HHAT, and PKCi were all required for the formation of oncospheres and proliferation in culture[104]. Chromosomal amplification of SOX2 has also been implicated in small cell lung cancer (SCLC)[105]. The means of *Sox2* upregulation may be varied and tissue-specific, however, as *Sox2* can be directly repressed by RB1, loss of *Rb1* function is often a driver mutation for many tumors associated with *Sox2* upregulation [98,105-108]. However, not all cases of Sox2 upregulation are connected to *Rb1* function. In contrast to small cell lung cancer, lung squamous cell carcinomas are not strongly associated with *Rb1* mutation, yet *Sox2* is clearly associated with their growth and maintenance[109]. In mouse and human skin squamous cell carcinoma overexpression of *Sox2* is observed and was found to be associated with activating histone marks, when it should be associated with repressive marks in healthy tissue[87]. One study in human ALK-positive large-cell lymphoma cells found that *Sox2* overexpression, along with doxorubicin-resistance and more aggressive growth, was triggered by oxidative stress caused by hydrogen peroxide[97]. In cervical cancer with upregulations of epidermal growth factor (EGF) receptor, knockdown of the EGF/PI3K pathway reduced expression of *Sox2*, suggesting that this pathway may play a role in the upregulation of *Sox2* in cervical cancer. Also in this study, it was found that expression of *miR-181a-2-3p* and *let-7i-5p* was able to downregulate *Sox2* expression, alluding to a dual role of miRNA and EGF receptor in mediating *Sox2* levels[96].

The downstream targets of Sox2 activity are also varied and likely tumor-specific. The *Sox2* target YAP, a member of the Hippo pathway, was found to be activated in a mouse osteosarcoma model and was a direct driver of initiation and proliferation of the cancer[88]. Another pathway implicated with *Sox2* in cancer is *Wnt/ -Catenin*. In human breast cancer and ALK-positive large-cell lymphoma, higher levels of *Sox2* expression led to higher *Wnt* signaling, which was associated with resistance to tamoxifen in breast cancer and doxorubicin in lymphoma and could propagate the cancer stem cell phenotype[94,97]. Overexpression of *Myc* was found to be associated with the same *Sox2*/*Wnt/ -Catenin* signaling axis in lymphoma[97]. A study in tongue squamous cell carcinoma also found *Sox2*-dependent overactivation of the *Wnt/ -Catenin* pathway, which was associated with epithelial-to-mesenchymal transition (EMT)[100]. In head and neck squamous cell carcinoma, *Sox2* directly promoted cancer proliferation by upregulation of cyclin B1 and increase in SNAIL expression, which is associated with EMT, necessary for metastasis[99]. An alternate mechanism found in head and neck squamous cell carcinoma is through *Sox2* mediated expression of *AFF4*, which is a core component of the super elongation complex. *AFF4* levels changed in parallel with *Sox2*, and knockout of *AFF4* led to decreased proliferation, migration, and invasion of cells, as well as decreased aldehyde dehydrogenase activity, important for tumor initiation[110]. *Sox2* may not only exert its tumorigenic properties via upregulation of cancer progressing pathways, but it also appears to have a role in the downregulation of tumor suppressors. In GBM cells, Sox2 expression downregulates the tumor suppressors BEX1 and BEX2, however this effect is likely indirect as there are no SOX2 binding domains in either BEX protein[93].

In addition to multiple signaling pathways, *Sox2* could exhibit its oncogenic effects by regulation of microRNA expression. Sequencing of GBM cells, showed that that *miR-145*, *miR-143*, *miR-253-5p*, and *miR-462* expression levels were responsive to *Sox2* levels. The implications of some miRNA expression in cancer has yet to be established but *miR-145* is thought to target *Sox2* to downregulate its expression, so overexpression of *Sox2* combined with downregulation of *miR-145* could potentiate the tumorigenic effect of *Sox2*[93]. In a separate study on breast cancer cell lines and patient samples resistant to Adriamycin therapy, low *miR-129-5p* expression was correlated with treatment resistance and a more aggressive phenotype in culture. Given that *miR-129-5p* binds directly to *Sox2*, when levels of *miR-129-5p* were high, levels of *Sox2* decreased and sensitized the cancer cells to treatment[111].

While it is clear that *Sox2* is upregulated in a number of tumor types, and is likely correlated with clinical phenotype, more work is needed to determine how *Sox2* affects cancer phenotypes. Given the variety of pathways and associations with *Sox2* in cancer, it is possible that the exact mechanism will be tumor or tissue-of-origin specific.

Oct4 in Cancer

Given the reprogramming power of *Oct4*, it also warrants investigation in a cancer setting. When using *Oct4* to reprogram fibroblasts to iPS cells, the methylation pattern in early reprogramming resembles that of cancer cells, and these cells were able to form teratomas with malignant characteristics in xenografts[112]. In somatic tissues of adult mice, expression of *Oct4* was sufficient to drive epithelial growths, which are dependent on *Oct4* for proliferation. In the intestines of these animals, *Oct4* expression inhibits differentiation of progenitor cells and reverts them to an embryonic-like phenotype[113]. In lung adenocarcinoma cells, *Oct4* is significantly elevated and is associated with expression of the stem cell marker CD133, as well as increased drug resistance and a higher propensity for EMT[114]. In non-small cell lung cancer with an activating epidermal growth factor receptor (EGFR) mutation, *Oct4* was also associated with treatment resistance and expression of CD133[115]. When using human tumor-derived cell cultures of lung adenocarcinoma and bronchioloalveolar carcinoma it was found that, when compared to healthy tissue, only lung adenocarcinoma expressed higher levels of *Oct4*[116]. In an analysis of human prostate cancer lines, *Oct4* was highly expressed in a subset of cells that were highly clonogenic and resistant to treatment with both docetaxel and gamma-radiation. These cells were CD133+, exhibited a stem-like state in culture, and formed highly aggressive tumors in mice[117].

Laboratory based studies of *Oct4* in cancer have clearly indicated it is an important factor, as have studies evaluating clinical correlates. Along with the expression of *Nanog* and the EMT marker Slug, *Oct4* expression marks high-grade lung adenocarcinomas and is associated with a worse prognosis for patients[114]. Also in lung adenocarcinoma, *Oct4* was upregulated and correlated with decreased differentiation, decreased survival, and increased tumor stage with worse clinical outcomes than *Oct4*-negative lung adenocarcinoma[118]. Oct4 was found to be highly expressed in EGFR-mutant non-small cell lung cancer and may be a marker for treatment resistance in these patients[115]. In a separate study of human non-small cell lung cancer, *Oct4* expression was associated with poor differentiation, and poor prognosis in patients who underwent surgical resection[102] *Oct4* was also overexpressed in ovarian cancer samples and was correlated with histological grade[119].

Oct4 has been shown to activate a number of downstream pathways when implicated in cancer. *Oct4* overexpression in mice that resulted in epithelial growths showed increased -catenin signaling in these cells[113]. In human ovarian cancer, follicle stimulating hormone (FSH) has previously been shown to inhibit apoptosis, and has found to be dependent on the presence of

Oct4. *Oct4*-mediated expression of FSH, leading to apoptotic inhibition also increased the expansion of ovarian stem-like cancer cells and upregulated the expression of other cancer-relevant genes like *Notch*, *Sox2*, and *Nanog*[120]. *Oct4* was also found to regulate the rate of apoptosis in breast cancer by a different mechanism. In breast cancer cell lines, *Oct4* expression regulated the expression of *p16INK4a*, *p14ARF*, *Bcl-2/Bax*, and *p53*, which may collectively lead to *Oct4*-mediated cell cycle progression and decreased rates of apoptosis[121].

It is possible that *Oct4* may also exert its effect through regulation of long non-coding RNAs (lncRNA). In a study of human lung cancer samples, multiple lncRNAs were found to be direct transcriptional targets of *Oct4*. The most relevant of these were nuclear paraspeckle assembly transcript 1 (*NEAT1*), metastasis-associated lung adenocarcinoma transcript 1 (*MALAT1*), and urothelial carcinoma-associated 1 (*UCA1*). *NEAT1* or *MALAT1* overexpression led to cancer cell proliferation, migration, and invasion, and knocking down *NEAT1* or *MALAT1* decreased cancer cell growth and motility. These lncRNAs were so important to cancer progression that co-expression of both, along with *Oct4* was predictive of poor prognosis in lung cancer patients[122].

Ordinarily, *Oct4* would be regulated by degradation by the ubiquitin proteasomal system (UPS), however, it is clear that in a cancer state, there is some level of misregulation that occurs. In healthy tissues this is mediated by OCT4 binding with CAV-1, a scaffolding protein, which allows for the degradation of OCT4 via UPS. In human lung cancer cells, nitric oxide (NO) facilitates the phosphorylation of CAV-1 by AKT, which subsequently does not allow OCT4 to complex with it, and therefore OCT4 does not get degraded via UPS. It is possible that the NO upregulation seen in many cancers is causal for increased levels of OCT4[123]. Another potential mechanism for regulation was shown in lung adenocarcinoma cells. Here, BEX4, was more highly expressed in cancer samples than in healthy tissue, and was shown to positively regulate the expression of *Oct4* and was required for proliferation of these cells. Interestingly, BEX4 expression was regulated by mTOR activation and suggests a role for an *mTOR*/*BEX4*/*Oct4* cascade in lung adenocarcinoma[124].

Intriguingly, one study has uncovered a positive role for *Oct4* overexpression in cancer. In a large study of gastric cancer patients who underwent surgical resection, tumors that were *Oct4* negative correlated with invasion depth and lymph node metastasis or invasion. In this study, *Oct4* negative patients had significantly worse outcomes than patients whose tumors were *Oct4* positive. The authors suggested that *Oct4* might suppress tumorigenesis, but in light of the strong links to *Oct4* expression and poor outcomes, it is probable that the positive effect of *Oct4* expression observed in this study is specific to gastric cancer[101].

Nanog in cancer

Unsurprisingly, the third pluripotency factor covered in the scope of this review, *Nanog*, has also been implicated in cancer. In a mouse model of mammary cancer, *Nanog* signaling accelerated tumor growth and caused tumors to be highly metastatic[125]. *Nanog* is overexpressed in human colorectal carcinoma cells, and it was found that these cells in culture have a high propensity towards a stem-like state. Human colorectal carcinoma cells readily form spheroids in culture and expression levels of *Nanog* increase greatly as the spheroids form. Inhibition of *Nanog* in this model decreased proliferation and G2-cell cycle related protein activation[120]. Studies with *Nanog* positive human hepatocellular carcinoma cells in culture demonstrated that *Nanog* positive cells readily differentiate into a wide variety of cancer cells, indicating the stemness of *Nanog* positive cells. These cells are highly invasive and metastatic, as well as resistant to chemotherapy[126]. In lung adenocarcinoma, *Nanog* was highly expressed and overexpression increased the CD133+ population in culture, as well as increasing drug resistance and EMT[114].

Like *Sox2* and *Oct4*, *Nanog* also has strong clinical correlates. In hepatocellular carcinoma, expression of *Nanog* was correlated with a worse clinical outcome[126]. In a study of human colorectal carcinoma cases, *NANOG* expression was associated with liver metastasis, which could make *NANOG* a marker of liver metastasis in colorectal carcinoma[120,127]. Also in colorectal cancer, *NANOG* was more highly expressed in $CD133^+$ tumor cells than in $CD133^-$ tumor cells, and expression was related to tumor grade, lymph node metastasis, and tumor stage using the TNM (tumor extent, node invasion, presence of metastasis) staging system[127]. *NANOG* expression in human cervical cancer was associated with immune evasion and was found to be positively correlated with outcome and disease stage[128]. A study evaluating *Nanog* expression in lip squamous cell carcinoma, actinic cheilitis, and normal lip epithelium found that *Nanog* was more highly expressed in the pre-cancerous actinic chelitis, and in lip squamous cell carcinoma, when compared to normal epithelium. It is therefore possible that *Nanog* has a role in the switch from healthy to precancerous, to cancerous tissue[129]. Patients with gastric adenocarcinoma had a higher expression of *NANOG* in their excised tumors than in healthy tissue. Additionally, the expression of *NANOG* was correlated with tumor stage, lymph node status, extent of infiltration, differentiation, and poor prognosis[130]. A large meta-analysis of gastrointestinal luminal cancer found that *NANOG* expression was associated with patient gender, depth of infiltration differentiation, TNM stage, and poor overall and disease-free survival, which implicates *NANOG* as a potential biomarker for gastrointestinal luminal cancer[131]. Surprisingly, a tissue microarray of human esophageal squamous cell carcinoma samples showed that increased expression of *NANOG* was associated with favorable prognosis and response to cisplatin[132]. Given the above evidence, it is puzzling that *NANOG* was favorable in these cases. Nevertheless, it is possible that the effect of Nanog is tumor or tissue specific and warrants further research.

Nanog may exert its tumorigenic effects via a variety of downstream targets. In a transgenic mouse model of breast cancer, overexpression of *Nanog* alone is not sufficient to induce cancer. Instead, when in combination with an upregulation of *Wnt-1*, *Nanog* was able to promote the growth of highly metastatic tumors. In this model, *Nanog* was found to be associated with the expression of a number of tumor-relevant genes, including EMT markers and PDGFRa, which can also drive tumorigenesis, angiogenesis, and metastasis, corroborating *Nanog's* effect in breast cancer[125]. In *Nanog*-positive hepatocellular carcinoma cells, insulin-like growth factor 2 (IGF2) and insulin-like growth factor receptor (IGF1R) were upregulated, and their levels sensitive to changes in *Nanog* expression. *Nanog* expression levels decreased when *IGF1R* was knocked out, indicating the presence of some sort of feedback loop along this signaling axis[126]. In a study of a variety of human cancer types, *NANOG* in precancerous and cancerous cervical tissue was related to the expression of *TCL1A* and phosphorylated AKT, which have a role in promoting chemotherapy resistance and immune evasion[128]. Interestingly, in human colorectal carcinoma, *NANOG* knockdown decreased expression of *SOX2* and *OCT4*, indicating that the probable feedback loop between these three factors is relevant in cancer as well[119].

How *Nanog* becomes expressed in cancers is still largely unknown, but some work has been done to elucidate a mechanism. In ovarian cancer, both *NANOG* and the androgen receptor (AR) are highly expressed. Given that the androgen 5a-dihydrostestosterone (DHT) activated *NANOG* transcription, it is possible that AR induces *NANOG* transcription. Cells given DHT had higher tumorigenesis, proliferation, migration, and colony and sphere formation, all phenotypes observed in *NANOG*-high cancers[133].

SUMMARY AND OUTSTANDING QUESTIONS

While a great body of work has elucidated the role of the pluripotency factors in ES and iPS cells and there still requires a larger analysis of the roles of the core pluripotency network outside of pluripotency. While some pluripotency factors have been studied outside of pluripotency, such as the role of *Sox2* in neural stem cells, there still remains to be a rigorous comparison of downstream network activation in non-pluripotent tissues compared to pluripotent stem cells. The information gathered to date indicate that there are indeed different roles for pluripotent genes in postnatal cell types that are different from their roles ES cells. This brings forth an interesting question: if exogenous pluripotent gene expression (such as those expressed during iPS reprogramming) are able to reprogram a cell to an ES-like state, why is it that adult cells that express core pluripotency genes not also mimic ES cells? Is it truly the combination of recombination factors alone, or are there some cell type-specific effects that modulate the network of genes activated by these pluripotency genes in contrast to the standard definition of a master regulator[134]? Furthermore, what contribution could these cell type-specific effects have on the outcome of pluripotency factor expression in a cell, whether it be a normal response to injury or a pathological response such as the formation of a tumor? This requires a greater understanding of the upstream regulators of the pluripotency factors in somatic tissues to understand these seemingly diverse roles of regeneration or disease.

The function of the core pluripotency genes, *SOX2*, *OCT4*, and *NANOG*, is of vast importance in understanding early development, embryonic stem cell function, and cellular reprogramming of iPS cells. However, their roles are not limited to early development. They are responsible for the maintenance of many adult tissues, and their regeneration after wounding. Importantly, they are key to understanding multiple pathologies, including cancer. An understanding of how these genes work outside of a pluripotent context will be critical to guiding new therapies to the clinic to treat many pathologies and to perhaps enhance wound regeneration.

ACKNOWLEDGEMENTS

This work was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103620.

Table 1 – Developmental disorders linked to pluripotency factor misregulation

Disorder	Pluripotency Factors	Gene Networks	References
Anothplamia,-esophageal-genital syndrome	Sox2	Notch, Pax6	63-65,73
CHARGE syndrome	Sox2	Chd7	65
Cochlear malformation	Sox2	Wnt	67
Chronic kidney disease	Sox2		65,75
Dentate gyrus hypoplasia	Sox2	Shh	26
Hypogonadotropic hypogonadism	Sox2	Wnt	60
Taste sensory defects	Sox2	Wnt	68

Table 2 – Cancers linked to pluripotency factor misregulation

Cancer	Pluripotency Factors	Roles	Gene Networks	References
Bladder cancer	SOX2	Tumor Invasiveness		89
Breast cancer	SOX2, OCT4, NANOG	Tamoxifen resistance, poor survival, proliferation, metastasis	Wnt/B-Catenin, miR-129-5p, p16INK4a, p14ARF, Bcl-2/Bax, p53	94,95,111,121,125
Cervical cancer	SOX2, NANOG	Maintain CSCs, immune evasion	EGF/PI3K, miR-181a-2-3p, let-7i-5p, AKT	96,128
Colorectal cancer	NANOG	Maintain CSCs, proliferation, drug resistance		120,127
Esophageal cancer	SOX2	Differentiation		108
Gastric cancer	SOX2, OCT4, NANOG	Tumor grade, metastasis, poor survival	IGF2, IGF2R	101,130-132

Glioblastoma	SOX2	Maintain CSCs, tumor propagation, proliferation, dedifferentiation	BEX1 and BEX2, miR-145, miR-143, miR-253, miR-462	90-93
Hepatocellular carcinoma	NANOG	Maintain CSCs, metastasis,		126
Lung cancer (non-small cell)	SOX2, OCT4, NANOG	Tumor progression, drug resistance, EMT, poor survival	CD133, EGFR, NEAT1, MALAT1	102,114-116,118,122-124 109
Lung cancer (small cell)	SOX2	Proliferation		105
Lung cancer (squamous)	SOX2	Proliferation	Hedgehog	104
Lymphoma	SOX2	Dedifferentiation	Wnt/B-Catenin, Myc	97
Osteosarcoma	SOX2	Maintain CSCs, proliferation	Hippo/YAP	88
Ovarian cancer	OCT4, NANOG	Tumor progression, proliferation	FSH, Notch, AR	119,133
Prostate cancer	SOX2, OCT4	Maintain CSCs, drug resistance	AR, CD133	98,117
Squamous cell carcinoma (head and neck)	SOX2, NANOG	Tumor reoccurrence, poor survival, metastasis	Wnt/B-Catenin, Cyclin B, SNAIL, AFF4	99,100,110,129
Squamous cell carcinoma (skin)	SOX2	Maintain CSCs, tumor propagation, proliferation, survival, adhesion, invasion and paraneoplastic syndrome	Epigenetic regulation	87

REFERENCES

1. Heng JC, Ng HH. Transcriptional regulation in embryonic stem cells. *Advances in experimental medicine and biology.* 2010;695:76-91.
2. Chambers I, Tomlinson SR. The transcriptional foundation of pluripotency. *Development.* 2009;136(14):2311-2322.
3. Young RA. Control of the embryonic stem cell state. *Cell.* 2011;144(6):940-954.
4. Nakai-Futatsugi Y, Niwa H. Transcription factor network in embryonic stem cells: heterogeneity under the stringency. *Biological & pharmaceutical bulletin.* 2013;36(2):166-170.
5. Hart AH, Hartley L, Ibrahim M, Robb L. Identification, cloning and expression analysis of the pluripotency promoting Nanog genes in mouse and human. *Developmental dynamics : an official publication of the American Association of Anatomists.* 2004;230(1):187-198.
6. Nichols J, Zevnik B, Anastassiadis K, et al. Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4. *Cell.* 1998;95(3):379-391.
7. Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, Lovell-Badge R. Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes Dev.* 2003;17(1):126-140.
8. Nishiyama A, Sharov AA, Piao Y, et al. Systematic repression of transcription factors reveals limited patterns of gene expression changes in ES cells. *Scientific reports.* 2013;3:1390.
9. Takahashi K, Yamanaka S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell.* 2006;126(4):663-676.
10. Cai Y, Dai X, Zhang Q, Dai Z. Gene expression of OCT4, SOX2, KLF4 and MYC (OSKM) induced pluripotent stem cells: identification for potential mechanisms. *Diagnostic pathology.* 2015;10:35.
11. Iwafuchi-Doi M, Zaret KS. Pioneer transcription factors in cell reprogramming. *Genes Dev.* 2014;28(24):2679-2692.
12. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev.* 2011;25(21):2227-2241.

13. Bernstein BE, Mikkelsen TS, Xie X, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell.* 2006;125(2):315-326.
14. Mikkelsen TS, Ku M, Jaffe DB, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature.* 2007;448(7153):553-560.
15. Fouse SD, Shen Y, Pellegrini M, et al. Promoter CpG methylation contributes to ES cell gene regulation in parallel with Oct4/Nanog, PcG complex, and histone H3 K4/K27 trimethylation. *Cell Stem Cell.* 2008;2(2):160-169.
16. Athanasiadou R, de Sousa D, Myant K, Merusi C, Stancheva I, Bird A. Targeting of de novo DNA methylation throughout the Oct-4 gene regulatory region in differentiating embryonic stem cells. *PLoS One.* 2010;5(4):e9937.
17. Ito S, D'Alessio AC, Taranova OV, Hong K, Sowers LC, Zhang Y. Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. *Nature.* 2010;466(7310):1129-1133.
18. Lister R, Pelizzola M, Dowen RH, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature.* 2009;462(7271):315-322.
19. Melton C, Judson RL, Blelloch R. Opposing microRNA families regulate self-renewal in mouse embryonic stem cells. *Nature.* 2010;463(7281):621-626.
20. Wabik A, Jones PH. Switching roles: the functional plasticity of adult tissue stem cells. *EMBO J.* 2015;34(9):1164-1179.
21. Roy S, Gascard P, Dumont N, et al. Rare somatic cells from human breast tissue exhibit extensive lineage plasticity. *Proc Natl Acad Sci U S A.* 2013;110(12):4598-4603.
22. Arnold K, Sarkar A, Yram MA, et al. Sox2(+) adult stem and progenitor cells are important for tissue regeneration and survival of mice. *Cell Stem Cell.* 2011;9(4):317-329.
23. Driessens G, Blanpain C. Long live sox2: sox2 lasts a lifetime. *Cell stem cell.* 2011;9(4):283-284.
24. Oliver-De La Cruz J, Carrion-Navarro J, Garcia-Romero N, et al. SOX2+ cell population from normal human brain white matter is able to generate mature oligodendrocytes. *PloS one.* 2014;9(6):e99253.
25. Ahlfeld J, Filser S, Schmidt F, et al. Neurogenesis from Sox2 expressing cells in the adult cerebellar cortex. *Scientific reports.* 2017;7(1):6137.
26. Favaro R, Valotta M, Ferri AL, et al. Hippocampal development and neural stem cell maintenance require Sox2-dependent regulation of Shh. *Nature neuroscience.* 2009;12(10):1248-1256.
27. Bani-Yaghoub M, Tremblay RG, Lei JX, et al. Role of Sox2 in the development of the mouse neocortex. *Developmental biology.* 2006;295(1):52-66.
28. Sirko S, Behrendt G, Johansson PA, et al. Reactive glia in the injured brain acquire stem cell properties in response to sonic hedgehog. [corrected]. *Cell stem cell.* 2013;12(4):426-439.
29. Gontan C, de Munck A, Vermeij M, Grosveld F, Tibboel D, Rottier R. Sox2 is important for two crucial processes in lung development: branching morphogenesis and epithelial cell differentiation. *Developmental biology.* 2008;317(1):296-309.
30. Wilson ME, Yang KY, Kalousova A, et al. The HMG box transcription factor Sox4 contributes to the development of the endocrine pancreas. *Diabetes.* 2005;54(12):3402-3409.
31. Yu X, Cates JM, Morrissey C, et al. SOX2 expression in the developing, adult, as well as, diseased prostate. *Prostate Cancer Prostatic Dis.* 2014;17(4):301-309.
32. Fauquier T, Rizzoti K, Dattani M, Lovell-Badge R, Robinson IC. SOX2-expressing progenitor cells generate all of the major cell types in the adult mouse pituitary gland. *Proc Natl Acad Sci U S A.* 2008;105(8):2907-2912.
33. Andoniadou CL, Matsushima D, Mousavy Gharavy SN, et al. Sox2(+) stem/progenitor cells in the adult mouse pituitary support organ homeostasis and have tumor-inducing potential. *Cell Stem Cell.* 2013;13(4):433-445.
34. Fu Q, Gremeaux L, Luque RM, et al. The adult pituitary shows stem/progenitor cell activation in response to injury and is capable of regeneration. *Endocrinology.* 2012;153(7):3224-3235.

35. Gremeaux L, Fu Q, Chen J, Vankelecom H. Activated phenotype of the pituitary stem/progenitor cell compartment during the early-postnatal maturation phase of the gland. *Stem cells and development.* 2012;21(5):801-813.
36. Kuzmichev AN, Kim SK, D'Alessio AC, et al. Sox2 acts through Sox21 to regulate transcription in pluripotent and differentiated cells. *Current biology : CB.* 2012;22(18):1705-1710.
37. Que J, Luo X, Schwartz RJ, Hogan BL. Multiple roles for Sox2 in the developing and adult mouse trachea. *Development (Cambridge, England).* 2009;136(11):1899-1907.
38. Basu-Roy U, Ambrosetti D, Favaro R, Nicolis SK, Mansukhani A, Basilico C. The transcription factor Sox2 is required for osteoblast self-renewal. *Cell death and differentiation.* 2010;17(8):1345-1353.
39. Jiang Y, Jahagirdar BN, Reinhardt RL, et al. Pluripotency of mesenchymal stem cells derived from adult marrow. *Nature.* 2002;418(6893):41-49.
40. Pochampally RR, Smith JR, Ylostalo J, Prockop DJ. Serum deprivation of human marrow stromal cells (hMSCs) selects for a subpopulation of early progenitor cells with enhanced expression of OCT-4 and other embryonic genes. *Blood.* 2004;103(5):1647-1652.
41. Yannarelli G, Pacienza N, Montanari S, Santa-Cruz D, Viswanathan S, Keating A. OCT4 expression mediates partial cardiomyocyte reprogramming of mesenchymal stromal cells. *PloS one.* 2017;12(12):e0189131.
42. Anjos-Afonso F, Bonnet D. Nonhematopoietic/endothelial SSEA-1+ cells define the most primitive progenitors in the adult murine bone marrow mesenchymal compartment. *Blood.* 2007;109(3):1298-1306.
43. Kuroda Y, Kitada M, Wakao S, et al. Unique multipotent cells in adult human mesenchymal cell populations. *Proceedings of the National Academy of Sciences of the United States of America.* 2010;107(19):8639-8643.
44. Lengner CJ, Camargo FD, Hochedlinger K, et al. Oct4 expression is not required for mouse somatic stem cell self-renewal. *Cell stem cell.* 2007;1(4):403-415.
45. Ohbo K, Yoshida S, Ohmura M, et al. Identification and characterization of stem cells in prepubertal spermatogenesis in mice. *Developmental biology.* 2003;258(1):209-225.
46. Dann CT, Alvarado AL, Molyneux LA, Denard BS, Garbers DL, Porteus MH. Spermatogonial stem cell self-renewal requires OCT4, a factor downregulated during retinoic acid-induced differentiation. *Stem cells (Dayton, Ohio).* 2008;26(11):2928-2937.
47. Kucia M, Reca R, Campbell FR, et al. A population of very small embryonic-like (VSEL) CXCR4(+)SSEA-1(+)Oct-4+ stem cells identified in adult bone marrow. *Leukemia.* 2006;20(5):857-869.
48. Kassmer SH, Krause DS. Very small embryonic-like cells: biology and function of these potential endogenous pluripotent stem cells in adult tissues. *Molecular reproduction and development.* 2013;80(8):677-690.
49. Jaenisch R, Young R. Stem cells, the molecular circuitry of pluripotency and nuclear reprogramming. *Cell.* 2008;132(4):567-582.
50. Chambers I, Silva J, Colby D, et al. Nanog safeguards pluripotency and mediates germline development. *Nature.* 2007;450(7173):1230-1234.
51. Fang F, Angulo B, Xia N, et al. A PAX5-OCT4-PRDM1 developmental switch specifies human primordial germ cells. *Nature cell biology.* 2018;20(6):655-665.
52. Kehler J, Tolkunova E, Koschorz B, et al. Oct4 is required for primordial germ cell survival. *EMBO reports.* 2004;5(11):1078-1083.
53. Yeom YI, Fuhrmann G, Ovitt CE, et al. Germline regulatory element of Oct-4 specific for the totipotent cycle of embryonal cells. *Development (Cambridge, England).* 1996;122(3):881-894.
54. Keramari M, Razavi J, Ingman KA, et al. Sox2 is essential for formation of trophectoderm in the preimplantation embryo. *PloS one.* 2010;5(11):e13952.
55. Guo G, Huss M, Tong GQ, et al. Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. *Developmental cell.* 2010;18(4):675-685.

56. Thomson M, Liu SJ, Zou LN, Smith Z, Meissner A, Ramanathan S. Pluripotency factors in embryonic stem cells regulate differentiation into germ layers. *Cell.* 2011;145(6):875-889.

57. Wegner M, Stolt CC. From stem cells to neurons and glia: a Soxist's view of neural development. *Trends Neurosci.* 2005;28(11):583-588.

58. Mandalos NP, Remboutsika E. Sox2: To crest or not to crest? *Seminars in cell & developmental biology.* 2017;63:43-49.

59. Ferri A, Favaro R, Beccari L, et al. Sox2 is required for embryonic development of the ventral telencephalon through the activation of the ventral determinants Nkx2.1 and Shh. *Development (Cambridge, England).* 2013;140(6):1250-1261.

60. Kelberman D, de Castro SC, Huang S, et al. SOX2 plays a critical role in the pituitary, forebrain, and eye during human embryonic development. *The Journal of clinical endocrinology and metabolism.* 2008;93(5):1865-1873.

61. Sisodiya SM, Ragge NK, Cavalleri GL, et al. Role of SOX2 mutations in human hippocampal malformations and epilepsy. *Epilepsia.* 2006;47(3):534-542.

62. Kelberman D, Rizzoti K, Avilion A, et al. Mutations within Sox2/SOX2 are associated with abnormalities in the hypothalamo-pituitary-gonadal axis in mice and humans. *The Journal of clinical investigation.* 2006;116(9):2442-2455.

63. Hagstrom SA, Pauer GJ, Reid J, et al. SOX2 mutation causes anophthalmia, hearing loss, and brain anomalies. *American journal of medical genetics Part A.* 2005;138a(2):95-98.

64. Taranova OV, Magness ST, Fagan BM, et al. SOX2 is a dose-dependent regulator of retinal neural progenitor competence. *Genes Dev.* 2006;20(9):1187-1202.

65. Engelen E, Akinci U, Bryne JC, et al. Sox2 cooperates with Chd7 to regulate genes that are mutated in human syndromes. *Nature genetics.* 2011;43(6):607-611.

66. Matsushima D, Heavner W, Pevny LH. Combinatorial regulation of optic cup progenitor cell fate by SOX2 and PAX6. *Development (Cambridge, England).* 2011;138(3):443-454.

67. Kiernan AE, Pelling AL, Leung KK, et al. Sox2 is required for sensory organ development in the mammalian inner ear. *Nature.* 2005;434(7036):1031-1035.

68. Okubo T, Pevny LH, Hogan BL. Sox2 is required for development of taste bud sensory cells. *Genes & development.* 2006;20(19):2654-2659.

69. Raghoebir L, Bakker ER, Mills JC, et al. SOX2 redirects the developmental fate of the intestinal epithelium toward a premature gastric phenotype. *Journal of molecular cell biology.* 2012;4(6):377-385.

70. Que J, Okubo T, Goldenring JR, et al. Multiple dose-dependent roles for Sox2 in the patterning and differentiation of anterior foregut endoderm. *Development (Cambridge, England).* 2007;134(13):2521-2531.

71. Raghoebir L, Biermann K, Buscop-van Kempen M, et al. Disturbed balance between SOX2 and CDX2 in human vitelline duct anomalies and intestinal duplications. *Virchows Archiv : an international journal of pathology.* 2013;462(5):515-522.

72. Ishii Y, Rex M, Scotting PJ, Yasugi S. Region-specific expression of chicken Sox2 in the developing gut and lung epithelium: regulation by epithelial-mesenchymal interactions. *Developmental dynamics : an official publication of the American Association of Anatomists.* 1998;213(4):464-475.

73. Williamson KA, Hever AM, Rainger J, et al. Mutations in SOX2 cause anophthalmia-esophageal-genital (AEG) syndrome. *Human molecular genetics.* 2006;15(9):1413-1422.

74. Volckaert T, De Langhe SP. Wnt and FGF mediated epithelial-mesenchymal crosstalk during lung development. *Developmental dynamics : an official publication of the American Association of Anatomists.* 2015;244(3):342-366.

75. Genheimer CW, Ilagan RM, Spencer T, et al. Molecular characterization of the regenerative response induced by intrarenal transplantation of selected renal cells in a rodent model of chronic kidney disease. *Cells, tissues, organs.* 2012;196(4):374-384.

76. Rosner MH, Vigano MA, Ozato K, et al. A POU-domain transcription factor in early stem cells and germ cells of the mammalian embryo. *Nature.* 1990;345(6277):686-692.

77. Palmieri SL, Peter W, Hess H, Scholer HR. Oct-4 transcription factor is differentially expressed in the mouse embryo during establishment of the first two extraembryonic cell lineages involved in implantation. *Developmental biology.* 1994;166(1):259-267.

78. Wu G, Scholer HR. Role of Oct4 in the early embryo development. *Cell Regen (Lond).* 2014;3(1):7.

79. Loh KM, Lim B. A precarious balance: pluripotency factors as lineage specifiers. *Cell Stem Cell.* 2011;8(4):363-369.

80. Niwa H. How is pluripotency determined and maintained? *Development (Cambridge, England).* 2007;134(4):635-646.

81. Feldman N, Gerson A, Fang J, et al. G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nature cell biology.* 2006;8(2):188-194.

82. Pardo M, Lang B, Yu L, et al. An expanded Oct4 interaction network: implications for stem cell biology, development, and disease. *Cell stem cell.* 2010;6(4):382-395.

83. Chambers I, Colby D, Robertson M, et al. Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells. *Cell.* 2003;113(5):643-655.

84. Mitsui K, Tokuzawa Y, Itoh H, et al. The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells. *Cell.* 2003;113(5):631-642.

85. Boyer LA, Lee TI, Cole MF, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell.* 2005;122(6):947-956.

86. Navarro P, Festuccia N, Colby D, et al. OCT4/SOX2-independent Nanog autorepression modulates heterogeneous Nanog gene expression in mouse ES cells. *EMBO J.* 2012;31(24):4547-4562.

87. Boumahdi S, Driessens G, Lapouge G, et al. SOX2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma. *Nature.* 2014;511(7508):246-250.

88. Maurizi G, Verma N, Gadi A, Mansukhani A, Basilico C. Sox2 is required for tumor development and cancer cell proliferation in osteosarcoma. *Oncogene.* 2018.

89. Zhu F, Qian W, Zhang H, et al. SOX2 Is a Marker for Stem-like Tumor Cells in Bladder Cancer. *Stem cell reports.* 2017;9(2):429-437.

90. Suva ML, Rheinbay E, Gillespie SM, et al. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell.* 2014;157(3):580-594.

91. Annovazzi L, Mellai M, Caldera V, Valente G, Schiffer D. SOX2 expression and amplification in gliomas and glioma cell lines. *Cancer genomics & proteomics.* 2011;8(3):139-147.

92. Fiscon G, Conte F, Licursi V, Nasi S, Paci P. Computational identification of specific genes for glioblastoma stem-like cells identity. *Scientific reports.* 2018;8(1):7769.

93. Fang X, Yoon JG, Li L, et al. The SOX2 response program in glioblastoma multiforme: an integrated ChIP-seq, expression microarray, and microRNA analysis. *BMC genomics.* 2011;12:11.

94. Piva M, Domenici G, Iriondo O, et al. Sox2 promotes tamoxifen resistance in breast cancer cells. *EMBO molecular medicine.* 2014;6(1):66-79.

95. Rodriguez-Pinilla SM, Sarrio D, Moreno-Bueno G, et al. Sox2: a possible driver of the basal-like phenotype in sporadic breast cancer. *Mod Pathol.* 2007;20(4):474-481.

96. Chhabra R. let-7i-5p, miR-181a-2-3p and EGF/PI3K/SOX2 axis coordinate to maintain cancer stem cell population in cervical cancer. *Scientific reports.* 2018;8(1):7840.

97. Wu C, Gupta N, Huang YH, et al. Oxidative stress enhances tumorigenicity and stem-like features via the activation of the Wnt/beta-catenin/MYC/Sox2 axis in ALK-positive anaplastic large-cell lymphoma. *BMC Cancer.* 2018;18(1):361.

98. Mu P, Zhang Z, Benelli M, et al. SOX2 promotes lineage plasticity and antiandrogen resistance in TP53- and RB1-deficient prostate cancer. *Science.* 2017;355(6320):84-88.

99. Lee SH, Oh SY, Do SI, et al. SOX2 regulates self-renewal and tumorigenicity of stem-like cells of head and neck squamous cell carcinoma. *British journal of cancer.* 2014;111(11):2122-2130.

100. Liu X, Qiao B, Zhao T, Hu F, Lam AK, Tao Q. Sox2 promotes tumor aggressiveness and epithelialmesenchymal transition in tongue squamous cell carcinoma. *International journal of molecular medicine.* 2018;42(3):1418-1426.

101. Matsuoka J, Yashiro M, Sakurai K, et al. Role of the stemness factors sox2, oct3/4, and nanog in gastric carcinoma. *The Journal of surgical research.* 2012;174(1):130-135.

102. Li X, Wang J, Xu Z, et al. Expression of Sox2 and Oct4 and their clinical significance in human non-small-cell lung cancer. *International journal of molecular sciences.* 2012;13(6):7663-7675.

103. Zayed H, Petersen I. Stem cell transcription factor SOX2 in synovial sarcoma and other soft tissue tumors. *Pathology, research and practice.* 2018;214(7):1000-1007.

104. Justilien V, Walsh MP, Ali SA, Thompson EA, Murray NR, Fields AP. The PRKCI and SOX2 oncogenes are coamplified and cooperate to activate Hedgehog signaling in lung squamous cell carcinoma. *Cancer Cell.* 2014;25(2):139-151.

105. Rudin CM, Durinck S, Stawiski EW, et al. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat Genet.* 2012;44(10):1111-1116.

106. Kareta MS, Gorges LL, Hafeez S, et al. Inhibition of pluripotency networks by the rb tumor suppressor restricts reprogramming and tumorigenesis. *Cell Stem Cell.* 2015;16(1):39-50.

107. Scholz RB, Kabisch H, Weber B, Roser K, Delling G, Winkler K. Studies of the RB1 gene and the p53 gene in human osteosarcomas. *Pediatr Hematol Oncol.* 1992;9(2):125-137.

108. Ishida H, Kasajima A, Kamei T, et al. SOX2 and Rb1 in esophageal small-cell carcinoma: their possible involvement in pathogenesis. *Mod Pathol.* 2017;30(5):660-671.

109. Mukhopadhyay A, Berrett KC, Kc U, et al. Sox2 cooperates with Lkb1 loss in a mouse model of squamous cell lung cancer. *Cell reports.* 2014;8(1):40-49.

110. Deng P, Wang J, Zhang X, et al. AFF4 promotes tumorigenesis and tumor-initiation capacity of head and neck squamous cell carcinoma cells by regulating SOX2. *Carcinogenesis.* 2018;39(7):937-947.

111. Zeng H, Wang L, Wang J, et al. microRNA-129-5p suppresses Adriamycin resistance in breast cancer by targeting SOX2. *Archives of biochemistry and biophysics.* 2018;651:52-60.

112. Ohm JE, Mali P, Van Neste L, et al. Cancer-related epigenome changes associated with reprogramming to induced pluripotent stem cells. *Cancer Res.* 2010;70(19):7662-7673.

113. Hochedlinger K, Yamada Y, Beard C, Jaenisch R. Ectopic expression of Oct-4 blocks progenitor-cell differentiation and causes dysplasia in epithelial tissues. *Cell.* 2005;121(3):465-477.

114. Chiou SH, Wang ML, Chou YT, et al. Coexpression of Oct4 and Nanog enhances malignancy in lung adenocarcinoma by inducing cancer stem cell-like properties and epithelial-mesenchymal transdifferentiation. *Cancer Res.* 2010;70(24):10433-10444.

115. Kobayashi I, Takahashi F, Nurwidya F, et al. Oct4 plays a crucial role in the maintenance of gefitinib-resistant lung cancer stem cells. *Biochemical and biophysical research communications.* 2016;473(1):125-132.

116. Karoubi G, Gugger M, Schmid R, Dutly A. OCT4 expression in human non-small cell lung cancer: implications for therapeutic intervention. *Interactive cardiovascular and thoracic surgery.* 2009;8(4):393-397.

117. Kanwal R, Shukla S, Walker E, Gupta S. Acquisition of tumorigenic potential and therapeutic resistance in CD133+ subpopulation of prostate cancer cells exhibiting stem-cell like characteristics. *Cancer letters.* 2018;430:25-33.

118. Zhang X, Han B, Huang J, et al. Prognostic significance of OCT4 expression in adenocarcinoma of the lung. *Japanese journal of clinical oncology.* 2010;40(10):961-966.

119. Zhang Z, Zhu Y, Lai Y, et al. Follicle-stimulating hormone inhibits apoptosis in ovarian cancer cells by regulating the OCT4 stem cell signaling pathway. *International journal of oncology.* 2013;43(4):1194-1204.

120. Zhang J, Espinoza LA, Kinders RJ, et al. NANOG modulates stemness in human colorectal cancer. *Oncogene.* 2013;32(37):4397-4405.
121. Meng L, Hu H, Zhi H, et al. OCT4B regulates p53 and p16 pathway genes to prevent apoptosis of breast cancer cells. *Oncology letters.* 2018;16(1):522-528.
122. Jen J, Tang YA, Lu YH, Lin CC, Lai WW, Wang YC. Oct4 transcriptionally regulates the expression of long non-coding RNAs NEAT1 and MALAT1 to promote lung cancer progression. *Molecular cancer.* 2017;16(1):104.
123. Maiuthed A, Bhummaphan N, Luanpitpong S, et al. Nitric oxide promotes cancer cell dedifferentiation by disrupting an Oct4: caveolin-1 complex: A new regulatory mechanism for cancer stem cell formation. *J Biol Chem.* 2018.
124. Zhao Z, Li J, Tan F, Gao S, He J. mTOR up-regulation of BEX4 promotes lung adenocarcinoma cell proliferation by potentiating OCT4. *Biochemical and biophysical research communications.* 2018;500(2):302-309.
125. Lu X, Mazur SJ, Lin T, Appella E, Xu Y. The pluripotency factor nanog promotes breast cancer tumorigenesis and metastasis. *Oncogene.* 2014;33(20):2655-2664.
126. Shan J, Shen J, Liu L, et al. Nanog regulates self-renewal of cancer stem cells through the insulin-like growth factor pathway in human hepatocellular carcinoma. *Hepatology (Baltimore, Md).* 2012;56(3):1004-1014.
127. Xu F, Dai C, Zhang R, Zhao Y, Peng S, Jia C. Nanog: a potential biomarker for liver metastasis of colorectal cancer. *Digestive diseases and sciences.* 2012;57(9):2340-2346.
128. Noh KH, Kim BW, Song KH, et al. Nanog signaling in cancer promotes stem-like phenotype and immune evasion. *J Clin Invest.* 2012;122(11):4077-4093.
129. Scotti FM, Mitt VC, Vieira DS, Biz MT, Castro RG, Modolo F. Expression of stem cell markers Nanog and Nestin in lip squamous cell carcinoma and actinic cheilitis. *Oral diseases.* 2018.
130. Lin T, Ding YQ, Li JM. Overexpression of Nanog protein is associated with poor prognosis in gastric adenocarcinoma. *Medical oncology (Northwood, London, England).* 2012;29(2):878-885.
131. Liang C, Zhao T, Ge H, et al. The clinicopathological and prognostic value of Nanog in human gastrointestinal luminal cancer: A meta-analysis. *International journal of surgery (London, England).* 2018;53:193-200.
132. Shimada Y, Okumura T, Sekine S, et al. Expression analysis of iPS cell - inductive genes in esophageal squamous cell carcinoma by tissue microarray. *Anticancer research.* 2012;32(12):5507-5514.
133. Ling K, Jiang L, Liang S, et al. Nanog interaction with the androgen receptor signaling axis induce ovarian cancer stem cell regulation: studies based on the CRISPR/Cas9 system. *Journal of ovarian research.* 2018;11(1):36.
134. Chan SS, Kyba M. What is a Master Regulator? *J Stem Cell Res Ther.* 2013;3.

Appendix 2

Molecular Cancer Research, 2021, doi: 10.1158/1541-7786.MCR-20-1006

Sox2 is an oncogenic driver of small cell lung cancer and promotes the classic neuroendocrine subtype

Ellen Voigt[1,2,†], Madeline Wallenburg[1,2,†], Hannah Wollenzien[1,2,3], Ethan Thompson[1,2], Kirtana Kumar[1,2], Joshua Feiner[4], Moira McNally[1,2], Hunter Friesen[1,2], Malini Mukherjee[5], Yohannes Afeworki[5], Michael S. Kareta[1,2,3,5,6,7,*]

[12]Cancer Biology and Immunotherapies Group and the Genetics & Genomics Group, Sanford Research, Sioux Falls, South Dakota, USA. [3]Division of Basic Biomedical Sciences, University of South Dakota, Vermillion, South Dakota, USA. [4]Dakota Wesleyan University, Mitchel, South Dakota, USA. [5]Functional Genomics & Bioinformatics Core, Sanford Research, Sioux Falls, SD, USA [6]Department of Pediatrics, Sanford School of Medicine, Sioux Falls, South Dakota, USA. [7]Department of Biochemistry, South Dakota State University, Brookings, South Dakota, USA.

[†] Equal Contribution
* Correspondence may be addressed to M.K. (michael.kareta@sanfordhealth.org)

Running Title
Role of SOX2 in Small Cell Lung Cancer

Disclosures of Potential Conflicts of Interest
The authors declare no potential conflicts of interest.

Abstract

Although many cancer prognoses have improved in the past fifty years due to advancements in treatments, there has been little improvement in therapies for small cell lung cancer (SCLC). One promising avenue to improve treatment for SCLC is to understand its underlying genetic alterations that drive its formation, growth, and cellular heterogeneity. *RB*-loss is one key driver of SCLC, and *RB*-loss has been associated with an increase in pluripotency factors such as *SOX2*. *SOX2* is highly expressed and amplified in SCLC and has been associated with SCLC growth. Using a genetically engineered mouse model, we have shown that *Sox2* is required for efficient SCLC formation. Furthermore, genome-scale binding assays have indicated that *SOX2*

can regulate key SCLC pathways such as *NEUROD1*, and *MYC*. This data suggests that *SOX2* can be associate with the switch of SCLC from an *ASCL1* subtype to a *NEUROD1* subtype. Understanding this genetic switch is key to understanding such processes as SCLC progression, cellular heterogeneity, and treatment resistance.

Implication Statement

Understanding the molecular mechanisms of SCLC initiation and development are key to opening new potential therapeutic options for this devastating disease.

Introduction

Small cell lung cancer (SCLC) is a devastating disease with markedly low survival rates, rapid metastasis, and almost invariable resistance to therapy. Patients who are stricken by this disease face a 6% two-year survival rate, while most will succumb less than a year after diagnosis (1, 2). Despite this alarming statistic, the standard of care for treating SCLC has remained essentially the same for the past 40 years and few innovations have been approved for this disease. First line treatments still rely primarily on platinum-based chemotherapy that often leads to treatment refractory tumors and poor patient outcomes (3-5). Recently immunotherapy options have been available for SCLC; however, while the results have been encouraging in select individuals, the patient responses have been generally poor (6). Therefore, in the pursuit of new therapies for SCLC, we have sought to understand the genetic factors underlying SCLC dynamics.

On a genetic level, SCLC is both rather simple and complex. It is simple in that the genetic drivers of SCLC are relatively clear. Patients have an almost invariable loss of the tumor suppressors *p53* (*TP53*) and *RB1* (*RB*) (7-9). Intriguingly, established SCLC can be genetically complex considering that, even with almost identical driver mutations, SCLC can be subdivided into four main subtypes defined by the function of key genetic regulators, *ASCL1*, *NEUROD1*, *POU2F3*, and *YAP1* (10-13, reviewed in: 14). Critically linked to the regulatory networks of the *ASCL1* (SCLC-A) and the *NEUROD1* (SCLC-N) subtypes is the role of the MYC family of oncogenes. MYC (cMYC) is highly expressed and a determining factor for the SCLC-N subtype (15). MYCL (L-Myc) rather, is predominantly expressed in SCLC-A, and is key to SCLC-A growth (7, 11, 16, 17). While MYC family regulation is important to SCLC growth and development (18), how MYC family members are regulated in SCLC is currently unclear (19).

The question of how a tumor with such homogenous driver mutations (*RB1*- and *p53*-loss) can lead to the diversity of genetic heterogeneity observed in SCLC remains unanswered. One clue to address this question can be found in the nature of the initiating mutations themselves. Beyond its role in regulating the G1/S checkpoint, RB also plays a multitude of roles in regulating gene expression (20-22). One of the genes regulated by RB is the transcription factor *SOX2* (23). Known primarily as a pluripotency factor, *SOX2* is also a key master regulator of neural and neuroendocrine cell types (24-28). As a master regulator, *SOX2* influences cell identity early and widely in cell fate decisions. Indeed, SOX2 is commonly amplified in SCLC (7). Pulmonary neuroendocrine cells are the predominant cell of origin for SCLC (29), therefore it is possible that *SOX2* upregulation in neuroendocrine cells following *RB1*-loss induces stem or progenitor genetic networks that help to drive oncogenesis. To that end, we generated a conditional knockout mouse in which we could perturb *Sox2* activity in a well-characterized SCLC mouse model to assess the consequence of *Sox2*-loss on SCLC formation. Combined with a genome-wide investigation into *SOX2* transcriptional regulation in SCLC, we observed that

SOX2 is indeed required for SCLC formation and regulates key genetic regulators of SCLC including *NEUROD1* and members of the *MYC* family.

Materials and Methods

Ethics statement

Mice were maintained according to the guidelines set forth by the NIH and were housed in the Sanford Research Animal Research Center, accredited by AAALAC using protocols reviewed and approved by our local IACUC.

SCLC mouse tumor initiation

We modeled SCLC in the *Rb1*$^{lox/lox}$, *p53*$^{lox/lox}$, *p130*$^{lox/lox}$, *Rosa*^{luc} (RPR2) mouse line (30), which readily develop SCLC after a few months, and added *Sox2* $^{+/+}$,$^{+/lox}$, or $^{lox/lox}$ alleles (Jackson Laboratories Stock #013093)(31). To study SCLC tumor initiation, we injected Cre-recombinase adenovirus (Ad5-CMV-Cre, Baylor Vector Development Lab, 0.91 L of a 5x10^{12} pt/mL viral preparation used per mouse) into the mouse lungs by intratracheal intubation to excise the lox-flanked genes (32). The mice were assigned to either a six-month cohort, a three-month cohort, or the survival curve. Mouse lungs, livers, and any other metastases were harvested for immunohistochemistry. Tumors were screened in a blinded manner by an independent pathologist.

SCLC lung and liver immunohistochemistry

The Sanford Research Histology & Imaging Core performed the immunohistochemistry for this study. The mouse lungs, livers and tumors were stained with H&E, for SOX2 (Abcam ab92494, 1:100), calcitonin gene related peptide (CGRP, Sigma C8198, 1:2,000), anti-phospho-histone H3 (pH3, EMD Millipore 06-570, 1:500), cleaved caspase 3 (CC3, Cell Signaling 9664, 1:100), ki67 (Biocare CRM325, 1:100), ASCL1 (Abcam ab74065, 1:500), and MYC (c-MYC, Invitrogen

MA1-980, 1:100) (Supplemental Table S4). To computationally assess tumor burden and feature characteristics, we digitized each slide using an Aperio VERSA slide scanner. The five images from each sample (H&E and SOX2, CGRP, ki67, pH3, and CC3 IHC) were registered using the Register Virtual Stack Slices Plugin in FIJI/ImageJ (33). We then used CellProfiler (34) to count the tumors and features. The H&E staining was used to identify tumors, then the intensity of IHC staining for the markers SOX2, CGRP, ki67, pH3, and CC3 was determined for the corresponding tumor areas in the other virtual slide images. Registration and CellProfiler scripts are available on the Kareta Lab website (https://research.sanfordhealth.org/researchers-and-labs/kareta-lab).

SCLC cell lines

We used the murine SCLC cell lines KP1 and KP3 (*Rb1*$^{\text{lox/lox}}$; *p53*$^{\text{lox/lox}}$) and the human SCLC lines NJH29 (H29), NCI-H82 (H82), NCI-H1836 (H1836), and NCI-H209 (H209) (30, 35). The cells were maintained in suspension and cultured in RPMI with 10% bovine growth serum and penicillin/streptomycin. All cell lines regularly tested negative for mycoplasma contamination.

Lentiviral transduction and cell assays

We made the lentivirus for the shRNA-mediated knockdown using the packaging plasmids VSVG, pMDL, and RSV in 293T cells, transfecting them with PEI with a nearly 90% transduction rate. Resulting lentivirus was concentrated using Lenti-X Concentrator (Takara Bio, Inc.) and titered for reproducible transductions. Controls consisted of an empty pSicoR vector or a pSicoR vector containing a shRNA to *Luciferase* (23). Transduced cells were selected for by culture with Puromycin for 5 days. We measured cellular viability after *SOX2* knock down with an alamar blue assay, and the levels of apoptosis with Annexin V staining combined with flow cytometry. qPCR was used to confirm the knock down of *Sox2* in the cells. Cas9-mediated knockdown of SOX2 was achieved by cloning a SOX2 gRNA sequence (ATTATAAATACCGGCCCCGG) into the TLCV2 inducible lentiviral Cas9 vector (36), which was packaged in to lentivirus using the methods above. Transfection was achieved using Lipofectamine 3000 (ThermoFisher Scientific) according to the manufacturer's protocol. To enhance transfection efficiency, after adding the transfection mix the cells were processed according to a modified spinfection protocol where they were centrifuged at 940 xg for 2 hours at room temperature. Mock controls were Lipofectamine-treated and spinfected cells that were processed the same but without the presence of the DNA vector. Due to high transfection efficiencies (typically greater than 70%), cells were neither selected nor sorted to minimize stress.

Chromatin Immunoprecipitation (ChIP) and CUT&RUN Assays

In preparation for HA-RB1   CDK chromatin immunoprecipitation, cells were transfected with pCMV-HA-hRb1-delta-CDK (Addgene, #58906) using Lipofectamine 3000 (ThermoFisher). ChIP for HA-RB1   CDK was performed as previously described (23) with several additional optimizations (37). The alternative swelling buffer was used for cell lysis. Chromatin was sonicated using a ME220 (Covaris, Inc.). ChIP-grade Protein AG magnetic beads (Pierce) were pre-blocked with BSA and salmon sperm DNA for 15 minutes on a rotating platform at 4^{o}C. The chromatin was pre-cleared before being diluted and incubated with an anti-HA antibody (Sigma H6908, 4   g) for immunoprecipitation. The antibody-chromatin complexes were incubated with blocked beads for 2 hours at 4^{o}C on a rotating platform prior to washing two times each with low-salt, high-salt, and LiCl wash buffers.

CUT&RUN assays were carried out according to the protocol (Version 3) published by Janssens and Henikoff (38) which is based on the original protocol developed by Skene *et al*. (39), using the CUTANA^{TM} pAG-MNase (EpiCypher), and concanavalin-A coated beads (BioMag Plus #86057). The optional high-calcium/low-salt conditions were included to prevent premature chromatin release after digestion. Both ChIP and CUT&RUN assays were performed using SOX2 antibodies from both EMD/Millipore (17-656) and R&D Systems (AF2018). ChIP and CUT&RUN libraries were analyzed on an Agilent Bioanalyzer System by the Sanford Research Functional Genomics & Biochemistry Core and sequenced at the Sanford Burnham Prebys Genomics Core. Both ChIP and CUT&RUN reads were aligned to the hg38 genome build using Bowtie 2 version 2.3.4.3 (40) and peaks called using MACS2 version 2.1.2 (41). As described by the authors of CUT&RUN, the top 99.5th percentile of peaks after sorting by q-values (including peaks with the same q-value at cutoff) were selected for further analysis (39). HOMER was used for heatmap generation and motif enrichment (42), Diffbind was used for differential peak identification and PCA visualization (43), and Ingenuity Pathway Analysis for network analysis (QIAGEN Inc.). Weighted gene co-expression network analysis was performed using the WGCNA package from Bioconductor (44). RNA-seq data was analyzed using DESeq2 (45).

Results

***Sox2* is critical for SCLC tumor initiation**

To investigate if *Sox2* is required for the formation of SCLC, we bred a mouse line containing a conditional *Sox2* allele (*Sox2*$^{\text{lox/lox}}$) to the RPR2 [*Rb1*$^{\text{lox/lox}}$; *p53*$^{\text{lox/lox}}$; *Rbl2*(*p130*)$^{\text{lox/lox}}$] mouse model of SCLC (Fig. 1A) (29, 30, 46, 47). With the addition of the conditional *Sox2* allele, we therefore named this line RPR2S. Tumors from RPR2 mice display all the common hallmarks of human SCLC, mainly the same histological characteristics as scored by an independent pathologist, rapid metastasis, and chemoresistance (30, 47, 48). To overcome the dramatic effects of global *Rb1*- and *p53*-loss in the mouse, we localized Cre-mediated recombination by an intratracheal instillation of a Cre-expressing adenovirus (Adeno-CMV-Cre-GFP) to target recombination specifically to the lung epithelium (49). As expected, we observed early lesions around 3-months, with a robust tumor burden 6-months after Cre-recombination (30).

By utilizing a breeding strategy that generates all three allelic combinations of *Sox2*: *Sox2*$^{+/+}$, *Sox2*$^{+/\mathrm{lox}}$, and *Sox2*$^{\mathrm{lox/lox}}$ (Supplemental Table S1), we were able to query if one or both alleles of *Sox2* are involved in SCLC formation. Three and six months after Adeno-Cre tumor initiation, *Rb1*$^{\mathrm{lox/lox}}$; *p53*$^{\mathrm{lox/lox}}$; *p130*$^{\mathrm{lox/lox}}$ mice showed a sizeable number of tumor foci displaying the histological characteristics of SCLC. However, the RPR2S mice had a nearly complete loss of SCLC foci observed at the same timepoint (Fig. 1B). To fully characterize these tumors and ensure complete *Sox2* loss in the RPR2S mice, we optimized immunohistochemistry staining and an unbiased image analysis pipeline using ImageJ and CellProfiler (34) resulting in a thorough statistical analysis of the number and marker expression in the RPR2 tumors compared to the few RPR2S tumors (Figs. 1C and 1D, Supplemental Fig. S1). The RPR2 tumors showed typical SCLC histology including high *Cgrp* expression, indicative of a neuroendocrine tumor type, and highly proliferative cells as indicated by ki67 and phospho-Histone H3 (pH3) staining (50) (Supplemental Fig. S1). At 6 months, there were a handful of very small tumors observed in the *Rb1*$^{\mathrm{lox/lox}}$; *p53*$^{\mathrm{lox/lox}}$; *p130*$^{\mathrm{lox/lox}}$; *Sox2* $^{\mathrm{lox/lox}}$ mice (Fig. 1D and Supplemental Fig. S1), although a sizeable number of these showed immunoreactivity to SOX2 antibodies, indicating that they are the result of incomplete Cre function. However, a small minority of SCLC tumors can initiate without *Sox2*, indicating that *Sox2* activity may not be absolutely necessary in some SCLC tumors or tumor subtypes. However, those tumors that grew even when *Sox2* was deleted were markedly smaller in size than the *Sox2*$^{+}$ tumors (Supplemental Fig. S1D). Importantly, we observed a significant lengthening of the lifespan of the *Rb1*$^{\text{lox/lox}}$; *p53*$^{\text{lox/lox}}$; *p130*$^{\text{lox/lox}}$; *Sox2*$^{\text{lox/lox}}$ mice (Fig. 1E), compared to *Sox2*-expressing controls.

***SOX2* is required for the growth of established SCLC lines**

The results indicating *Sox2* function in the initiation of SCLC tumors in mice led us to investigate if *Sox2* is required in established tumors. We utilized shRNA-mediated knockdown to reduce *SOX2* expression in both mouse and human SCLC cell lines. We were able to achieve a ~60-90% knockdown of *SOX2* by RT-qPCR (Supplemental Fig. S2A). We observed that knockdown of *SOX2* in both mouse and human cell lines significantly reduces the growth of these cells in culture compared to mock-transduced cells (Fig. 2A), similar to a previously reported *SOX2* knockdown in human SCLC cell lines (7). Concurrent with a loss of cellular viability, we observed an increase in the number of apoptotic cells upon *SOX2* knockdown (Supplemental Fig. S2B). As *RB1*-loss is one of the primary genetic drivers of SCLC (7, 9, 47), and the RB protein can bind to and repress the *Sox2* locus in fibroblasts (23), we set out to investigate if RB is capable of repressing *SOX2* in SCLC to indicate if RB-loss in SCLC could be the driver of *SOX2* upregulation. To this end, we overexpressed an *RB1* transgene in human SCLC cell lines in which the CDK phosphorylation sites have been mutated (*RB1 CDK*) to render RB resistant to CDK inactivation (51). Overexpression of *RB1 CDK* greatly reduced the viability of human SCLC cell lines (Fig. 2B)(52). Furthermore, overexpression of RB1 resulted in the repression of *Sox2* (Fig. 2C). By chromatin immunoprecipitation (ChIP) we tested if RB1 CDK binds to the promoter or the two known proximal SOX2 enhancers, *SRR1* and *SRR2* (53). Indeed, we do observe significant enrichment of RB1 CDK-bound regions at the *SOX2* promoter and the downstream *SRR2* enhancer (Fig. 2D). Finally, overexpression of *SOX2-t2a-GFP* rescued the repression of RB1 CDK growth-inhibited SCLC cell lines (Fig. 2E, Supplemental Fig S2C). Together these data confirm that *SOX2* is required for SCLC tumor growth and that *SOX2* expression is most likely a consequence of *RB1*-loss.

***SOX2* regulates key SCLC pathways**

To observe the genomic localization of SOX2 in human SCLC cell lines, we performed both ChIP and Cleavage Under Targets and Release Using Nuclease (CUT&RUN) (39) using the endogenous SOX2 from both H1836 and H29 cells. While SOX2 ChIP allowed for broad localization studies, we found that SOX2 CUT&RUN was much more sensitive for comparative genomic localization studies due to the lack of chemical crosslinking and the release of SOX2-bound DNA due to SOX2 antibody:ProteinA/G:MNase complexes rather than sonication. We observed a very similar localization of SOX2 in both cell lines (Fig. 3A, Supplemental Fig. S3, Supplemental Table S3). Unbiased motif enrichment of the SOX2 peaks identified an HMG binding domain as the most highly enriched motif (Fig. 3B). The HMG domain is the DNA-binding domain of the SOX family of proteins therefore, the presence of HMG motifs validates the specificity of the SOX2 localization (54). As expected for a neuroendocrine tumor, and with the known role of *SOX2* in the regulation of neurogenesis (55, 56), the top ontology terms for the SOX2 adjacent genes were related to neural development and function (Fig. 3C). To assess if the binding topology of SOX2 in SCLC is similar to other *SOX2*-expressing cells, we compared the binding similarity by read counts for SOX2 datasets from human embryonic stem (ES) cells, induced pluripotent stem (iPS) cells, neural stem cells (NSCs), and glioblastoma (57-60). We observe that SOX2 binding in SCLC is distinct from both NSCs and pluripotent cells (ES and iPS cells). The closest binding profile to SCLC was glioblastoma therefore the function of SOX2 in cancer may be distinct from its role in normal cellular development (Fig. 3D).

The genes that are bound by SOX2 appear to show a biphasic distribution of high- or low-expression, indicating that they are either upregulated or repressed by SOX2 (Fig. 3E,F, Supplemental Fig. S4). Indeed, SOX2 can either repress or transactivate target genes based upon the cofactors recruited (54, 61, 62), and it appears these two roles of SOX2 are maintained in SCLC. To better describe the genetic networks that are regulated by *SOX2* in SCLC we performed a weighted gene co-expression network analysis (WGCNA) to identify the gene networks co-expressed with *SOX2* using the SCLC cell lines in the Cancer Cell Line Encyclopedia (63, 64). The WCGNA analysis identified multiple modules that are co-expressed with *SOX2* (Supplemental Fig. S5, S6). The most highly upregulated module with *SOX2* contained *ASCL1*, a known regulator of classic SCLC (10, 11, 14)(Fig. 3G). The most downregulated module identified contained a MYC network, which is associated with the variant state of SCLC (15) (Fig. 3H).

***SOX2* regulates SCLC-subtype specific specification**

To further investigate the result that high levels of *SOX2* favors *ASCL1* gene modules and is anti-correlated with *MYC* gene modules (Fig. 3G) we investigated if SOX2 expression favors the ASCL1 SCLC subtype. We performed unbiased clustering of CCLE SCLC cell lines based on their expression of *ASCL1*, *NEUROD1*, *YAP1*, *POU2F3*, *MYC*, and *MYCL* (Fig. 4A). The cell lines generally clustered by subtype and *SOX2* specifically clustered with the *ASCL1* subtype. As it is unclear if the ASCL1-SOX2 module (Fig. 3G) is due to direct SOX2 regulation of ASCL1 or a correlation due to high SOX2 levels in the SCLC-A subtype cell lines (Supplemental Fig. S6), we set out to determine if the regulation of the SCLC subtype-specific factors *ASCL1* and *NEUROD1* is directly regulated by SOX2. Overexpression of *SOX2-t2a-GFP* in two SCLC-A (H1836 & H209) and two SCLC-N (H29 & H82) cell lines does not appear to perturb *ASCL1* levels, but does result in significant downregulation of *NEUROD1* (Fig. 4B, Supplemental Fig. S7). We observed similar changes at the protein level, although levels of NEUROD1 were marked lower in H1836 and H209 cells (Fig 4D). We then used an inducible Cas9-mediated knockdown of *SOX2* rather than shRNA-mediated knockdown to observe the rapid effects of target gene expression after Cas9 induction, which results in significant SOX2 knockdown (Supplemental Fig. S8A). In contrast to *SOX2* overexpression, we observed a significant upregulation of *NEUROD1* (Fig. 4C, Supplemental Fig. S8A). To test if regulation of *NEUROD1* by SOX2 is direct we performed ChIP of SOX2. We observed significant binding of SOX2 at the *NEUROD1* and *MYC* promoters (Fig. 4E, Supplemental Fig. S8B). Therefore, it appears that SOX2 does not directly regulate *ASCL1*; however, it is associated with the progression of SCLC tumors to the *NEUROD1* state.

***SOX2* directly regulates *MYC* and *MYCL* in the ASCL1 and NEUROD1 SCLC Subtypes**

With the observation that SOX2 potentially regulates MYC networks in SCLC (Fig. 3H), we investigated if SOX2 directly regulates the MYC family in SCLC. We observed binding of SOX2 to both *MYC* and *MYCL* in SCLC from the CUT&RUN data (Figs. 3A and 5A). Both *MYC* and *MYCL* are expressed in SCLC, with *MYCL* predominantly expressed in the SCLC-A subtype and *MYC* expressed in the SCLC-N subtype (14-16). Interestingly, SOX2 appears bound at *MYCL* in H1836 cells, which are of the SCLC-A subtype and is bound at *MYC* in H29 cells, which are of the SCLC-N subtype (Supplemental Fig S7)(65). This is consistent with a role for SOX2 to activate these genes in their respective SCLC subtype. Overexpression of *SOX2* in both SCLC-A and SCLC-N cells further supports a role for SOX2 in the regulation of *MYC* and *MYCL*. When *SOX2-t2a-GFP* is transfected into the SCLC-A cell lines H1836 and H209, we observed a downregulation of *MYC* at both the mRNA (Fig. 5B) and protein levels (Fig. 5C and

5D). Rather, in the SCLC-N lines H29 and H82, there is significant downregulation of *MYCL* upon *SOX2* overexpression (Fig. 5B) and an apparent, but not significant increase in the protein levels of MYC (P=0.0724), perhaps not reaching significance due to the already elevated levels of MYC in these cell lines (Supplemental Fig S7). This indicates that overexpression of *SOX2*, in contrast to normal levels of expression (Fig. 5A), is repressive at either *MYC* or *MYCL* yet still favoring *MYCL* expression in the SCLC-A subtype and *MYC* expression in SCLC-N. We tested for either ASCL1*,* NEUROD1 or MYC expression in the tumors from the RPR2S mice, and observed that *Sox2*$^{+}$ tumors display high ASCL1 and low NEUROD1/MYC staining, which is expected as the RPR2 mice predominantly form tumors of the SCLC-A subtype (15). However, the few *Sox2*$^{\text{lox/lox}}$ tumors showed reduced ASCL1 staining and increased NEUROD1/MYC immunoreactivity (Fig. 5E). ASCL1, NEUROD1, and MYC staining showed nuclear localization consistent with SCLC cells and not infiltrating cells (Supplemental Fig. S9). Blinded scoring of the tumors as either ASCL1^{+} NEUROD1^{+}, or MYC^{+} showed a significant increase in the number of NEUROD1^{+}/MYC^{+} tumors from the *Sox2*$^{\text{lox/lox}}$ mice (Fig. 5F, Supplemental Table S2). Therefore, it appears that SOX2 favors the formation of an SCLC-A subtype.

Discussion

There have been a few indications that *SOX2* may be a key factor in SCLC, however its role in SCLC has so far been obscure. Rudin and colleagues showed that *SOX2* is amplified in ~27% of SCLC patients and cell lines, and that knockdown of *SOX2* can impair growth of SCLC cell lines (7). We have previously shown that *RB1*-loss, one of the two driver mutations required for SCLC initiation, can result in *SOX2* upregulation (23). *SOX2* has been observed to be misregulated in various cancers of the epithelium (66). As SCLC is a cancer that rises from the lung epithelium, predominantly from pulmonary neuroendocrine cells which themselves express

SOX2 during development, it seemed reasonable that *SOX2* may indeed be a driver of SCLC (29, 67). However, the role for *SOX2* in SCLC initiation and its mechanism in SCLC was unclear.

To that end, we generated a genetically engineered mouse model of SCLC based on the RPR2 [*Rb1*$^{lox/lox}$; *p53*$^{lox/lox}$; *Rbl2*(*p130*)$^{lox/lox}$] line, where we introduced a conditional *Sox2*$^{lox/lox}$ allele (named the RPR2S line). We observed that deletion of *Sox2* in these mice greatly hampers the formation of SCLC tumors. The requirement of SOX2 in SCLC formation was not completely penetrant, however, as there were a handful of small tumors that developed in the absence of *Sox2*. These tumors had properties similar to the SCLC-N subtype as they showed low levels of ASCL1 and high NEUROD1 and MYC. Therefore, SOX2 may be required primarily for SCLC-A type tumors, which are the primary subtype of the RPR2 line, and that any escapees were able to activate *Neurod1* subtype networks to compensate and/or bypass the *Ascl1* state.

To assess the function of SOX2, we assessed its genomic localization and observed that SOX2 primarily binds to genes involved in neurogenesis, where neural gene signatures are commonly found in SCLC (10, 11). Intriguingly, the genes bound by SOX2 did not strictly overlap with SOX2 binding profiles in either pluripotent cells (ES and iPS cells) or NSCs. Rather the SOX2 binding profile was most similar to glioblastoma multiforme, indicating that SOX2 may share a more common function amongst cancer than its well-studied functions in development. This is perhaps unexpected as SOX2 has been described as a pioneer factor that is able to bind its target DNA sequences regardless of any regional heterochromatin, and therefore should be able to regulate target sequences in a wide assortment of donor cells (68). Rather we observe that the cellular context does impart some level of regulation on the broader SOX2 network. This is particularly relevant considering that SCLC can arise from a few different cell types on the lung epithelium and can influence the resulting SCLC subtype (18, 29, 69). It is possible that the few NEUROD1^{+}/MYC^{+} lesions observed in the *Sox2*$^{\text{lox/lox}}$ mice are a result of tumors initiating from a non-neuroendocrine lineage. Finally, what cell-type specific factors may be constraining SOX2 function will be of particular importance towards understanding SOX2 regulation in SCLC, and potentially provide novel avenues for therapeutic targeting SCLC, and perhaps other SOX2-driven cancers.

We observed two regulator networks that correlate with *SOX2* expression in SCLC. The first is ASCL1 that is required for SCLC formation in the RPR2 mouse model, and indeed is localized at *SOX2* indicating a direct role in SOX2 regulation (11). Consistent with ASCL1 lying upstream of *SOX2* in established SCLC cell lines, we observe that neither overexpression nor knockdown of *SOX2* alters *ASCL1* expression. This prompts the question of how *ASCL1* can lie upstream of *SOX2* if *SOX2* upregulation is a direct consequence of *RB1*-loss, one of the two SCLC driver mutations. It could be that *RB1*-loss promotes the derepression of *SOX2*, but *ASCL1* activity is required for full *SOX2* transactivation and subsequent tumor development. Intriguingly, ASCL1 and SOX2 have been found at similar enhancer regions (70), therefore the regulation of these two factors may not be strictly linear. Further investigation into the genetic networks at play in early SCLC tumors will be required to address these questions.

With the potential link between *SOX2* activity and *ASCL1*, we also investigated the other neuroendocrine SCLC subtype specific factor, *NEUROD1*. SOX2 has been found to regulate *Neurod1* in neural progenitor cells, where it functions to maintain an epigenetically permissive state at the *Neurod1* promoter (71). Conversely, in neural stem cells of the adult hippocampus, it was observed that SOX2 binds to the *Neurod1* promoter and silences *Neurod1* expression (72). In SCLC, we observe that SOX2 overexpression leads to *NEUROD1* silencing, while basal levels of SOX2 appear to be associated with activation or attenuation of the levels of activated *NEUROD1*. This regulation appears direct as we observe SOX2 bound at the *NEUROD1* promoter by ChIP, although binding at *NEUROD1* was unclear in the CUT&RUN data. It is possible that these two techniques may recognize different SOX2 protein complexes due to their differing methods to assess DNA localization. As *MYC* is a target of NEUROD1 (11), SOX2 loss could then promote a maintenance of the SCLC-N subtype network.

We also uncovered a role of SOX2 in the regulation of *MYC* and *MYCL* in SCLC. We observe that endogenous levels of SOX2 appear associated with activation as SOX2 was found at *MYCL* in SCLC-A subtype cell lines while it was bound at *MYC* in SCLC-N cell lines. Yet, in contrast we observe that overexpression of SOX2 enhanced repression of *MYC* and *MYCL* in SCLC-A and SCLC-N, respectively. As was shown for SOX2 in embryonic stem cells (61), we also observe that SOX2 can be associated with both gene activation and gene silencing. The alternating functions of SOX2 of both gene activation or repression most likely reflect differing SOX2 protein complexes that are assembled in a context-specific manner, with tight stoichiometric regulation of the endogenous activating complex so that overexpressed SOX2 favors the formation of a more promiscuous repressive complex. Further investigation into the SOX2 protein interactome in SCLC and specifically in different SCLC subtypes will be required to delineate the mechanistic function of SOX2 on different gene targets. SOX2, while typically oncogenic in the lung (73, 74), can indeed act as a tumor suppressor when overexpressed in multiple cancer types (75) indicating cell-type specific roles. Consequently, SOX2 may possess differing functions, either favoring transcriptional activation or silencing in different cells within a single SCLC tumor, or tumors that arise from alternative cells of origin as SCLC is indeed a heterogeneous tumor comprised of multiple cell types responsible for tumor propagation and treatment resistance (19, 76-79). Further investigation into the mechanism of SOX2 activity in these different cell types may shed additional light on the development of SCLC heterogeneity and treatment resistance.

Together we have illustrated that *SOX2* is strongly favorable to SCLC formation in the RPR2 SCLC mouse model. SOX2 serves to regulate *NEUROD1* expression and is associated with the switch from *MYCL* to *MYC* expression, although further investigation into its regulatory mechanisms of this switch are required. ASCL1 is the predominant network controlling SCLC activity in the early tumor; however, during tumor progression there is a switch to the *NEUROD1* state, driven in part by MYC and is linked with poorer patient outcomes (18, 79). Our data indicates that *SOX2* is associated with this process by the concurrent regulation of *NEUROD1*, *MYC*, and *MYCL*. Understanding the genetic networks that underlie this switch during SCLC tumor progression will add to the explanation of such processes as treatment resistance, and ultimately lead to improved therapies to treat this devastating disease.

Author's Contributions

Conception and design: E. Voigt, M. Wallenburg, H. Wollenzien, M. Kareta

Development of methodology: E. Voigt, M. Wallenburg, H. Wollenzien, E. Thompson, M. Kareta

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): E. Voigt, M. Wallenburg, H. Wollenzien, K. Kumar, E. Thompson, J. Feiner, M. McNally, H. Friesen, M. Mukherjee, M. Kareta

Analysis and interpretation of data (e.g. statistical analysis, biostatistics, computational analysis): E. Voigt, M. Wallenburg, H. Wollenzien, E. Thompson, J. Feiner, M. McNally, H. Friesen, M. Mukherjee, Y. Afeworki, M. Kareta

Writing, review, and/or revision of the manuscript: E. Voigt, M. Wallenburg, H. Wollenzien, M. Kareta

Administrative, technical, or material support (i.e. reporting or organizing data, constructing databases): M. Kareta

Study supervision: M. Kareta

Acknowledgements

We would like to acknowledge the NIH NIGMS Center for Pediatric Research 5P20GM103620 and the NIH NCI/NIGMS grant R01CA233661 for research support (to M.K.). M. M. was supported by NIH NICHD grant R25HD097633. The Sanford Research Pathology and Flow Cytometry Cores are supported by the NIGMS Center for Cancer Research P20GM103548. H.W. is thankful to the University of South Dakota-Neuroscience, Nanotechnology and Networks (USD-N3) Program, supported by a grant from the National Science Foundation Research Traineeship program DGE-1633213. We are grateful to Ryan Askeland M.D. (Sanford Health Pathology Clinic) for the unbiased and blinded characterization of SCLC tissue samples. KP1, KP3, and NJH29 cells were a kind gift from Julien Sage. *Sox2* hairpin vectors were a kind gift from Alejandro Sweet-Cordero. pCMV-HA-hRb1-delta-CDK was a gift from Steven Dowdy (Addgene plasmid #58906), SOX2-t2A-GFP was a gift from Jennifer Mitchell (Addgene plasmid #127537), and TLCV2 was a gift from Adam Karpf (Addgene plasmid #87360).

References

1. Gazdar AF, Bunn PA, Minna JD. Small-cell lung cancer: what we know, what we need to know and the path forward. Nat Rev Cancer. 2017;17(12):725-37. doi: 10.1038/nrc.2017.87. PubMed PMID: 29077690.
2. Byers LA, Rudin CM. Small cell lung cancer: where do we go from here? Cancer. 2015;121(5):664-72. doi: 10.1002/cncr.29098. PubMed PMID: 25336398; PMCID: PMC5497465.

3. Pietanza MC, Byers LA, Minna JD, Rudin CM. Small cell lung cancer: will recent progress lead to improved outcomes? Clin Cancer Res. 2015;21(10):2244-55. doi: 10.1158/1078-0432.CCR-14-2958. PubMed PMID: 25979931; PMCID: PMC4497796.

4. Sandler AB. Chemotherapy for small cell lung cancer. Semin Oncol. 2003;30(1):9-25. doi: 10.1053/sonc.2003.50012. PubMed PMID: 12635086.

5. Berns A. The therapy escapes of small-cell lung cancer. Nature Cancer. 2020;1(4):374-5. doi: 10.1038/s43018-020-0058-y.

6. Wu Y, Liu Y, Sun C, Wang H, Zhao S, Li W, Chen B, Wang L, Ye L, He Y, Zhou C. Immunotherapy as a treatment for small cell lung cancer: a case report and brief review. Transl Lung Cancer Res. 2020;9(2):393-400. doi: 10.21037/tlcr.2020.03.20. PubMed PMID: 32420081; PMCID: PMC7225158.

7. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, Bergbower EA, Guan Y, Shin J, Guillory J, Rivers CS, Foo CK, Bhatt D, Stinson J, Gnad F, Haverty PM, Gentleman R, Chaudhuri S, Janakiraman V, Jaiswal BS, Parikh C, Yuan W, Zhang Z, Koeppen H, Wu TD, Stern HM, Yauch RL, Huffman KE, Paskulin DD, Illei PB, Varella-Garcia M, Gazdar AF, de Sauvage FJ, Bourgon R, Minna JD, Brock MV, Seshagiri S. Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet. 2012;44(10):1111-6. Epub 2012/09/04. doi: ng.2405 [pii] 10.1038/ng.2405. PubMed PMID: 22941189.

8. Peifer M, Fernandez-Cuesta L, Sos ML, George J, Seidel D, Kasper LH, Plenker D, Leenders F, Sun R, Zander T, Menon R, Koker M, Dahmen I, Muller C, Di Cerbo V, Schildhaus HU, Altmuller J, Baessmann I, Becker C, de Wilde B, Vandesompele J, Bohm D, Ansen S, Gabler F, Wilkening I, Heynck S, Heuckmann JM, Lu X, Carter SL, Cibulskis K, Banerji S, Getz G, Park KS, Rauh D, Grutter C, Fischer M, Pasqualucci L, Wright G, Wainer Z, Russell P, Petersen I, Chen Y, Stoelben E, Ludwig C, Schnabel P, Hoffmann H, Muley T, Brockmann M, Engel-Riedel W, Muscarella LA, Fazio VM, Groen H, Timens W, Sietsma H, Thunnissen E, Smit E, Heideman DA, Snijders PJ, Cappuzzo F, Ligorio C, Damiani S, Field J, Solberg S, Brustugun OT, Lund-Iversen M, Sanger J, Clement JH, Soltermann A, Moch H, Weder W, Solomon B, Soria JC, Validire P, Besse B, Brambilla E, Brambilla C, Lantuejoul S, Lorimier P, Schneider PM, Hallek M, Pao W, Meyerson M, Sage J, Shendure J, Schneider R, Buttner R, Wolf J, Nurnberg P, Perner S, Heukamp LC, Brindle PK, Haas S, Thomas RK. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. Nat Genet. 2012;44(10):1104-10. Epub 2012/09/04. doi: ng.2396 [pii] 10.1038/ng.2396. PubMed PMID: 22941188.

9. George J, Lim JS, Jang SJ, Cun Y, Ozretic L, Kong G, Leenders F, Lu X, Fernandez-Cuesta L, Bosco G, Muller C, Dahmen I, Jahchan NS, Park KS, Yang D, Karnezis AN, Vaka D, Torres A, Wang MS, Korbel JO, Menon R, Chun SM, Kim D, Wilkerson M, Hayes N, Engelmann D, Putzer B, Bos M, Michels S, Vlasic I, Seidel D, Pinther B, Schaub P, Becker C, Altmuller J, Yokota J, Kohno T, Iwakawa R, Tsuta K, Noguchi M, Muley T, Hoffmann H, Schnabel PA, Petersen I, Chen Y, Soltermann A, Tischler V, Choi CM, Kim YH, Massion PP, Zou Y, Jovanovic D, Kontic M, Wright GM, Russell PA, Solomon B, Koch I, Lindner M, Muscarella LA, la Torre A, Field JK, Jakopovic M, Knezevic J, Castanos-Velez E, Roz L, Pastorino U, Brustugun OT, Lund-Iversen M, Thunnissen E, Kohler J, Schuler M, Botling J, Sandelin M, Sanchez-Cespedes M, Salvesen HB, Achter V, Lang U, Bogus M, Schneider PM, Zander T, Ansen S, Hallek M, Wolf J, Vingron M, Yatabe Y, Travis WD, Nurnberg P, Reinhardt C, Perner S, Heukamp L, Buttner R, Haas SA, Brambilla E, Peifer M, Sage J, Thomas RK. Comprehensive genomic profiles of small cell lung cancer. Nature. 2015;524(7563):47-53. doi: 10.1038/nature14664. PubMed PMID: 26168399.

10. Augustyn A, Borromeo M, Wang T, Fujimoto J, Shao C, Dospoy PD, Lee V, Tan C, Sullivan JP, Larsen JE, Girard L, Behrens C, Wistuba, II, Xie Y, Cobb MH, Gazdar AF, Johnson JE, Minna JD. ASCL1 is a lineage oncogene providing therapeutic targets for high-grade neuroendocrine lung

cancers. Proc Natl Acad Sci U S A. 2014;111(41):14788-93. doi: 10.1073/pnas.1410419111. PubMed PMID: 25267614; PMCID: 4205603.

11. Borromeo MD, Savage TK, Kollipara RK, He M, Augustyn A, Osborne JK, Girard L, Minna JD, Gazdar AF, Cobb MH, Johnson JE. ASCL1 and NEUROD1 Reveal Heterogeneity in Pulmonary Neuroendocrine Tumors and Regulate Distinct Genetic Programs. Cell reports. 2016;16(5):1259-72. doi: 10.1016/j.celrep.2016.06.081. PubMed PMID: 27452466; PMCID: PMC4972690.

12. McColl K, Wildey G, Sakre N, Lipka MB, Behtaj M, Kresak A, Chen Y, Yang M, Velcheti V, Fu P, Dowlati A. Reciprocal expression of INSM1 and YAP1 defines subgroups in small cell lung cancer. Oncotarget. 2017;8(43):73745-56. doi: 10.18632/oncotarget.20572. PubMed PMID: 29088741; PMCID: PMC5650296.

13. Huang YH, Klingbeil O, He XY, Wu XS, Arun G, Lu B, Somerville TDD, Milazzo JP, Wilkinson JE, Demerdash OE, Spector DL, Egeblad M, Shi J, Vakoc CR. POU2F3 is a master regulator of a tuft cell-like variant of small cell lung cancer. Genes Dev. 2018;32(13-14):915-28. doi: 10.1101/gad.314815.118. PubMed PMID: 29945888; PMCID: PMC6075037.

14. Rudin CM, Poirier JT, Byers LA, Dive C, Dowlati A, George J, Heymach JV, Johnson JE, Lehman JM, MacPherson D, Massion PP, Minna JD, Oliver TG, Quaranta V, Sage J, Thomas RK, Vakoc CR, Gazdar AF. Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. Nat Rev Cancer. 2019;19(5):289-97. doi: 10.1038/s41568-019-0133-9. PubMed PMID: 30926931.

15. Mollaoglu G, Guthrie MR, Bohm S, Bragelmann J, Can I, Ballieu PM, Marx A, George J, Heinen C, Chalishazar MD, Cheng H, Ireland AS, Denning KE, Mukhopadhyay A, Vahrenkamp JM, Berrett KC, Mosbruger TL, Wang J, Kohan JL, Salama ME, Witt BL, Peifer M, Thomas RK, Gertz J, Johnson JE, Gazdar AF, Wechsler-Reya RJ, Sos ML, Oliver TG. MYC Drives Progression of Small Cell Lung Cancer to a Variant Neuroendocrine Subtype with Vulnerability to Aurora Kinase Inhibition. Cancer Cell. 2017;31(2):270-85. doi: 10.1016/j.ccell.2016.12.005. PubMed PMID: 28089889; PMCID: PMC5310991.

16. Kim DW, Wu N, Kim YC, Cheng PF, Basom R, Kim D, Dunn CT, Lee AY, Kim K, Lee CS, Singh A, Gazdar AF, Harris CR, Eisenman RN, Park KS, MacPherson D. Genetic requirement for Mycl and efficacy of RNA Pol I inhibition in mouse models of small cell lung cancer. Genes Dev. 2016;30(11):1289-99. doi: 10.1101/gad.279307.116. PubMed PMID: 27298335; PMCID: PMC4911928.

17. Semenova EA, Kwon MC, Monkhorst K, Song JY, Bhaskaran R, Krijgsman O, Kuilman T, Peters D, Buikhuisen WA, Smit EF, Pritchard C, Cozijnsen M, van der Vliet J, Zevenhoven J, Lambooij JP, Proost N, van Montfort E, Velds A, Huijbers IJ, Berns A. Transcription Factor NFIB Is a Driver of Small Cell Lung Cancer Progression in Mice and Marks Metastatic Disease in Patients. Cell reports. 2016;16(3):631-43. doi: 10.1016/j.celrep.2016.06.020. PubMed PMID: 27373156; PMCID: PMC4956617.

18. Ireland AS, Micinski AM, Kastner DW, Guo B, Wait SJ, Spainhower KB, Conley CC, Chen OS, Guthrie MR, Soltero D, Qiao Y, Huang X, Tarapcsak S, Devarakonda S, Chalishazar MD, Gertz J, Moser JC, Marth G, Puri S, Witt BL, Spike BT, Oliver TG. MYC Drives Temporal Evolution of Small Cell Lung Cancer Subtypes by Reprogramming Neuroendocrine Fate. Cancer Cell. 2020;38(1):60-78 e12. doi: 10.1016/j.ccell.2020.05.001. PubMed PMID: 32473656; PMCID: PMC7393942.

19. Shue YT, Lim JS, Sage J. Tumor heterogeneity in small cell lung cancer defined and investigated in pre-clinical mouse models. Transl Lung Cancer Res. 2018;7(1):21-31. doi: 10.21037/tlcr.2018.01.15. PubMed PMID: 29535910; PMCID: PMC5835592.

20. Burkhart DL, Sage J. Cellular mechanisms of tumour suppression by the retinoblastoma gene. Nat Rev Cancer. 2008;8(9):671-82. Epub 2008/07/25. doi: nrc2399 [pii] 10.1038/nrc2399. PubMed PMID: 18650841.

21. Chinnam M, Goodrich DW. RB1, development, and cancer. Current topics in developmental biology. 2011;94:129-69. doi: 10.1016/B978-0-12-380916-2.00005-X. PubMed PMID: 21295686; PMCID: 3691055.

22. Dyson NJ. RB1: a prototype tumor suppressor and an enigma. Genes Dev. 2016;30(13):1492-502. doi: 10.1101/gad.282145.116. PubMed PMID: 27401552; PMCID: PMC4949322.

23. Kareta MS, Gorges LL, Hafeez S, Benayoun BA, Marro S, Zmoos AF, Cecchini MJ, Spacek D, Batista LF, O'Brien M, Ng YH, Ang CE, Vaka D, Artandi SE, Dick FA, Brunet A, Sage J, Wernig M. Inhibition of pluripotency networks by the rb tumor suppressor restricts reprogramming and tumorigenesis. Cell Stem Cell. 2015;16(1):39-50. doi: 10.1016/j.stem.2014.10.019. PubMed PMID: 25467916.

24. Abdelalim EM, Emara MM, Kolatkar PR. The SOX transcription factors as key players in pluripotent stem cells. Stem Cells Dev. 2014;23(22):2687-99. Epub 2014/08/16. doi: 10.1089/scd.2014.0297. PubMed PMID: 25127330.

25. Arnold K, Sarkar A, Yram MA, Polo JM, Bronson R, Sengupta S, Seandel M, Geijsen N, Hochedlinger K. Sox2(+) adult stem and progenitor cells are important for tissue regeneration and survival of mice. Cell Stem Cell. 2011;9(4):317-29. doi: 10.1016/j.stem.2011.09.001. PubMed PMID: 21982232; PMCID: PMC3538360.

26. Avilion AA, Nicolis SK, Pevny LH, Perez L, Vivian N, Lovell-Badge R. Multipotent cell lineages in early mouse development depend on SOX2 function. Genes Dev. 2003;17(1):126-40. doi: 10.1101/gad.224503. PubMed PMID: 12514105; PMCID: 195970.

27. Driessens G, Blanpain C. Long live sox2: sox2 lasts a lifetime. Cell Stem Cell. 2011;9(4):283-4. Epub 2011/10/11. doi: 10.1016/j.stem.2011.09.007. PubMed PMID: 21982223.

28. Ellis P, Fagan BM, Magness ST, Hutton S, Taranova O, Hayashi S, McMahon A, Rao M, Pevny L. SOX2, a persistent marker for multipotential neural stem cells derived from embryonic stem cells, the embryo or the adult. Dev Neurosci. 2004;26(2-4):148-65. doi: 10.1159/000082134. PubMed PMID: 15711057.

29. Sutherland KD, Proost N, Brouns I, Adriaensen D, Song JY, Berns A. Cell of origin of small cell lung cancer: inactivation of Trp53 and Rb1 in distinct cell types of adult mouse lung. Cancer Cell. 2011;19(6):754-64. doi: 10.1016/j.ccr.2011.04.019. PubMed PMID: 21665149.

30. Schaffer BE, Park KS, Yiu G, Conklin JF, Lin C, Burkhart DL, Karnezis AN, Sweet-Cordero EA, Sage J. Loss of p130 accelerates tumor development in a mouse model for human small-cell lung carcinoma. Cancer Res. 2010;70(10):3877-83. doi: 10.1158/0008-5472.CAN-09-4228. PubMed PMID: 20406986; PMCID: PMC2873158.

31. Shaham O, Smith AN, Robinson ML, Taketo MM, Lang RA, Ashery-Padan R. Pax6 is essential for lens fiber cell differentiation. Development. 2009;136(15):2567-78. doi: 10.1242/dev.032888. PubMed PMID: 19570848; PMCID: PMC2709063.

32. DuPage M, Dooley AL, Jacks T. Conditional mouse lung cancer models using adenoviral or lentiviral delivery of Cre recombinase. Nat Protoc. 2009;4(7):1064-72. doi: 10.1038/nprot.2009.95. PubMed PMID: 19561589; PMCID: PMC2757265.

33. Cardona A, Arganda-Carreras I, Saalfeld A. Register Virtual Stack Slices. Available from: https://imagej.net/Register_Virtual_Stack_Slices.

34. Kamentsky L, Jones TR, Fraser A, Bray MA, Logan DJ, Madden KL, Ljosa V, Rueden C, Eliceiri KW, Carpenter AE. Improved structure, function and compatibility for CellProfiler: modular high-

throughput image analysis software. Bioinformatics. 2011;27(8):1179-80. doi: 10.1093/bioinformatics/btr095. PubMed PMID: 21349861; PMCID: PMC3072555.

35. Jahchan NS, Dudley JT, Mazur PK, Flores N, Yang D, Palmerton A, Zmoos AF, Vaka D, Tran KQ, Zhou M, Krasinska K, Riess JW, Neal JW, Khatri P, Park KS, Butte AJ, Sage J. A drug repositioning approach identifies tricyclic antidepressants as inhibitors of small cell lung cancer and other neuroendocrine tumors. Cancer Discov. 2013;3(12):1364-77. doi: 10.1158/2159-8290.CD-13-0183. PubMed PMID: 24078773; PMCID: PMC3864571.

36. Tan YS, Sansanaphongpricha K, Xie Y, Donnelly CR, Luo X, Heath BR, Zhao X, Bellile E, Hu H, Chen H, Polverini PJ, Chen Q, Young S, Carey TE, Nor JE, Ferris RL, Wolf GT, Sun D, Lei YL. Mitigating SOX2-potentiated Immune Escape of Head and Neck Squamous Cell Carcinoma with a STING-inducing Nanosatellite Vaccine. Clin Cancer Res. 2018;24(17):4242-55. doi: 10.1158/1078-0432.CCR-17-2807. PubMed PMID: 29769207; PMCID: PMC6125216.

37. Burkhart DL, Wirt SE, Zmoos AF, Kareta MS, Sage J. Tandem E2F binding sites in the promoter of the p107 cell cycle regulator control p107 expression and its cellular functions. PLoS Genet. 2010;6(6):e1001003. Epub 2010/06/30. doi: 10.1371/journal.pgen.1001003. PubMed PMID: 20585628; PMCID: 2891812.

38. Janssens D, Henikoff, S. CUT&RUN: Targeted in situ genome-wide profiling with high efficiency for low cell numbers. protocols.io 2019. Available from: https://dx.doi.org/10.17504/protocols.io.zcpf2vn.

39. Skene PJ, Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife. 2017;6. doi: 10.7554/eLife.21856. PubMed PMID: 28079019; PMCID: PMC5310842.

40. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357-9. doi: 10.1038/nmeth.1923. PubMed PMID: 22388286; PMCID: 3322381.

41. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9(9):R137. Epub 2008/09/19. doi: gb-2008-9-9-r137 [pii] 10.1186/gb-2008-9-9-r137. PubMed PMID: 18798982; PMCID: 2592715.

42. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell. 2010;38(4):576-89. doi: 10.1016/j.molcel.2010.05.004. PubMed PMID: 20513432; PMCID: 2898526.

43. Ross-Innes CS, Stark R, Teschendorff AE, Holmes KA, Ali HR, Dunning MJ, Brown GD, Gojis O, Ellis IO, Green AR, Ali S, Chin SF, Palmieri C, Caldas C, Carroll JS. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature. 2012;481(7381):389-93. doi: 10.1038/nature10730. PubMed PMID: 22217937; PMCID: 3272464.

44. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. doi: 10.1186/1471-2105-9-559. PubMed PMID: 19114008; PMCID: PMC2631488.

45. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. doi: 10.1186/s13059-014-0550-8. PubMed PMID: 25516281; PMCID: PMC4302049.

46. Gouyer V, Gazzeri S, Bolon I, Drevet C, Brambilla C, Brambilla E. Mechanism of retinoblastoma gene inactivation in the spectrum of neuroendocrine lung tumors. American journal of respiratory cell and molecular biology. 1998;18(2):188-96. doi: 10.1165/ajrcmb.18.2.3008. PubMed PMID: 9476905.

47. Meuwissen R, Linn SC, Linnoila RI, Zevenhoven J, Mooi WJ, Berns A. Induction of small cell lung cancer by somatic inactivation of both Trp53 and Rb1 in a conditional mouse model. Cancer Cell. 2003;4(3):181-9. Epub 2003/10/03. doi: S1535610803002204 [pii]. PubMed PMID: 14522252.

48. Park KS, Liang MC, Raiser DM, Zamponi R, Roach RR, Curtis SJ, Walton Z, Schaffer BE, Roake CM, Zmoos AF, Kriegel C, Wong KK, Sage J, Kim CF. Characterization of the cell of origin for small cell lung cancer. Cell Cycle. 2011;10(16):2806-15. Epub 2011/08/09. doi: 17012 [pii]. PubMed PMID: 21822053.

49. Gierut JJ, Jacks TE, Haigis KM. In vivo delivery of lenti-Cre or adeno-Cre into mice using intranasal instillation. Cold Spring Harb Protoc. 2014;2014(3):307-9. doi: 10.1101/pdb.prot073445. PubMed PMID: 24591689; PMCID: PMC4169259.

50. Drivsholm L, Paloheimo LI, Osterlind K. Chromogranin A, a significant prognostic factor in small cell lung cancer. British journal of cancer. 1999;81(4):667-71. doi: 10.1038/sj.bjc.6690745. PubMed PMID: 10574253; PMCID: PMC2362890.

51. Narasimha AM, Kaulich M, Shapiro GS, Choi YJ, Sicinski P, Dowdy SF. Cyclin D activates the Rb tumor suppressor by mono-phosphorylation. Elife. 2014;3. doi: 10.7554/eLife.02872. PubMed PMID: 24876129; PMCID: PMC4076869.

52. Nikitin AY, Juarez-Perez MI, Li S, Huang L, Lee WH. RB-mediated suppression of spontaneous multiple neuroendocrine neoplasia and lung metastases in Rb+/- mice. Proc Natl Acad Sci U S A. 1999;96(7):3916-21. PubMed PMID: 10097138; PMCID: PMC22395.

53. Tomioka M, Nishimoto M, Miyagi S, Katayanagi T, Fukui N, Niwa H, Muramatsu M, Okuda A. Identification of Sox-2 regulatory region which is under the control of Oct-3/4-Sox-2 complex. Nucleic Acids Res. 2002;30(14):3202-13. Epub 2002/07/24. PubMed PMID: 12136102; PMCID: 135755.

54. Kamachi Y, Kondoh H. Sox proteins: regulators of cell fate specification and differentiation. Development. 2013;140(20):4129-44. doi: 10.1242/dev.091793. PubMed PMID: 24086078.

55. Suh H, Consiglio A, Ray J, Sawai T, D'Amour KA, Gage FH. In vivo fate analysis reveals the multipotent and self-renewal capacities of Sox2+ neural stem cells in the adult hippocampus. Cell Stem Cell. 2007;1(5):515-28. doi: 10.1016/j.stem.2007.09.002. PubMed PMID: 18371391; PMCID: PMC2185820.

56. Bennett L, Yang M, Enikolopov G, Iacovitti L. Circumventricular organs: a novel site of neural stem cells in the adult brain. Mol Cell Neurosci. 2009;41(3):337-47. doi: 10.1016/j.mcn.2009.04.007. PubMed PMID: 19409493; PMCID: PMC2697272.

57. Zhou C, Yang X, Sun Y, Yu H, Zhang Y, Jin Y. Comprehensive profiling reveals mechanisms of SOX2-mediated cell fate specification in human ESCs and NPCs. Cell Res. 2016;26(2):171-89. doi: 10.1038/cr.2016.15. PubMed PMID: 26809499; PMCID: PMC4746607.

58. Narayan S, Bryant G, Shah S, Berrozpe G, Ptashne M. OCT4 and SOX2 Work as Transcriptional Activators in Reprogramming Human Fibroblasts. Cell reports. 2017;20(7):1585-96. doi: 10.1016/j.celrep.2017.07.071. PubMed PMID: 28813671; PMCID: PMC5648000.

59. Ng SY, Bogu GK, Soh BS, Stanton LW. The long noncoding RNA RMST interacts with SOX2 to regulate neurogenesis. Mol Cell. 2013;51(3):349-59. doi: 10.1016/j.molcel.2013.07.017. PubMed PMID: 23932716.

60. Fang X, Yoon JG, Li L, Yu W, Shao J, Hua D, Zheng S, Hood L, Goodlett DR, Foltz G, Lin B. The SOX2 response program in glioblastoma multiforme: an integrated ChIP-seq, expression microarray, and microRNA analysis. BMC genomics. 2011;12:11. doi: 10.1186/1471-2164-12-11. PubMed PMID: 21211035; PMCID: PMC3022822.

61. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, Guenther MG, Kumar RM, Murray HL, Jenner RG, Gifford DK, Melton DA, Jaenisch R, Young RA. Core transcriptional

regulatory circuitry in human embryonic stem cells. Cell. 2005;122(6):947-56. Epub 2005/09/13. doi: S0092-8674(05)00825-1 [pii] 10.1016/j.cell.2005.08.020. PubMed PMID: 16153702; PMCID: 3006442.

62. Boyer LA, Plath K, Zeitlinger J, Brambrink T, Medeiros LA, Lee TI, Levine SS, Wernig M, Tajonar A, Ray MK, Bell GW, Otte AP, Vidal M, Gifford DK, Young RA, Jaenisch R. Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature. 2006;441(7091):349-53. Epub 2006/04/21. doi: nature04733 [pii] 10.1038/nature04733. PubMed PMID: 16625203.

63. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article17. doi: 10.2202/1544-6115.1128. PubMed PMID: 16646834.

64. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, Reddy A, Liu M, Murray L, Berger MF, Monahan JE, Morais P, Meltzer J, Korejwa A, Jane-Valbuena J, Mapa FA, Thibault J, Bric-Furlong E, Raman P, Shipway A, Engels IH, Cheng J, Yu GK, Yu J, Aspesi P, Jr., de Silva M, Jagtap K, Jones MD, Wang L, Hatton C, Palescandolo E, Gupta S, Mahan S, Sougnez C, Onofrio RC, Liefeld T, MacConaill L, Winckler W, Reich M, Li N, Mesirov JP, Gabriel SB, Getz G, Ardlie K, Chan V, Myer VE, Weber BL, Porter J, Warmuth M, Finan P, Harris JL, Meyerson M, Golub TR, Morrissey MP, Sellers WR, Schlegel R, Garraway LA. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603-7. doi: 10.1038/nature11003. PubMed PMID: 22460905; PMCID: PMC3320027.

65. Coles GL, Cristea S, Webber JT, Levin RS, Moss SM, He A, Sangodkar J, Hwang YC, Arand J, Drainas AP, Mooney NA, Demeter J, Spradlin JN, Mauch B, Le V, Shue YT, Ko JH, Lee MC, Kong C, Nomura DK, Ohlmeyer M, Swaney DL, Krogan NJ, Jackson PK, Narla G, Gordan JD, Shokat KM, Sage J. Unbiased Proteomic Profiling Uncovers a Targetable GNAS/PKA/PP2A Axis in Small Cell Lung Cancer Stem Cells. Cancer Cell. 2020;38(1):129-43 e7. doi: 10.1016/j.ccell.2020.05.003. PubMed PMID: 32531271; PMCID: PMC7363571.

66. Novak D, Huser L, Elton JJ, Umansky V, Altevogt P, Utikal J. SOX2 in development and cancer biology. Semin Cancer Biol. 2019. doi: 10.1016/j.semcancer.2019.08.007. PubMed PMID: 31412296.

67. Gontan C, de Munck A, Vermeij M, Grosveld F, Tibboel D, Rottier R. Sox2 is important for two crucial processes in lung development: branching morphogenesis and epithelial cell differentiation. Developmental biology. 2008;317(1):296-309. doi: 10.1016/j.ydbio.2008.02.035. PubMed PMID: 18374910.

68. Soufi A, Donahue G, Zaret KS. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. Cell. 2012;151(5):994-1004. doi: 10.1016/j.cell.2012.09.045. PubMed PMID: 23159369; PMCID: 3508134.

69. Yang D, Denny SK, Greenside PG, Chaikovsky AC, Brady JJ, Ouadah Y, Granja JM, Jahchan NS, Lim JS, Kwok S, Kong CS, Berghoff AS, Schmitt A, Reinhardt HC, Park KS, Preusser M, Kundaje A, Greenleaf WJ, Sage J, Winslow MM. Intertumoral Heterogeneity in SCLC Is Influenced by the Cell Type of Origin. Cancer Discov. 2018;8(10):1316-31. doi: 10.1158/2159-8290.CD-17-0987. PubMed PMID: 30228179; PMCID: PMC6195211.

70. Christensen CL, Kwiatkowski N, Abraham BJ, Carretero J, Al-Shahrour F, Zhang T, Chipumuro E, Herter-Sprie GS, Akbay EA, Altabef A, Zhang J, Shimamura T, Capelletti M, Reibel JB, Cavanaugh JD, Gao P, Liu Y, Michaelsen SR, Poulsen HS, Aref AR, Barbie DA, Bradner JE, George RE, Gray NS, Young RA, Wong KK. Targeting transcriptional addictions in small cell lung cancer with a covalent CDK7 inhibitor. Cancer Cell. 2014;26(6):909-22. doi: 10.1016/j.ccell.2014.10.019. PubMed PMID: 25490451; PMCID: PMC4261156.

71. Amador-Arjona A, Cimadamore F, Huang CT, Wright R, Lewis S, Gage FH, Terskikh AV. SOX2 primes the epigenetic landscape in neural precursors enabling proper gene activation during

hippocampal neurogenesis. Proc Natl Acad Sci U S A. 2015;112(15):E1936-45. doi: 10.1073/pnas.1421480112. PubMed PMID: 25825708; PMCID: PMC4403144.

72. Kuwabara T, Hsieh J, Muotri A, Yeo G, Warashina M, Lie DC, Moore L, Nakashima K, Asashima M, Gage FH. Wnt-mediated activation of NeuroD1 and retro-elements during adult neurogenesis. Nature neuroscience. 2009;12(9):1097-105. doi: 10.1038/nn.2360. PubMed PMID: 19701198; PMCID: PMC2764260.

73. Ferone G, Song JY, Sutherland KD, Bhaskaran R, Monkhorst K, Lambooij JP, Proost N, Gargiulo G, Berns A. SOX2 Is the Determining Oncogenic Switch in Promoting Lung Squamous Cell Carcinoma from Different Cells of Origin. Cancer Cell. 2016;30(4):519-32. doi: 10.1016/j.ccell.2016.09.001. PubMed PMID: 27728803; PMCID: PMC5065004.

74. Lu Y, Futtner C, Rock JR, Xu X, Whitworth W, Hogan BL, Onaitis MW. Evidence that SOX2 overexpression is oncogenic in the lung. PLoS One. 2010;5(6):e11022. Epub 2010/06/16. doi: 10.1371/journal.pone.0011022. PubMed PMID: 20548776; PMCID: 2883553.

75. Metz EP, Wuebben EL, Wilder PJ, Cox JL, Datta K, Coulter D, Rizzino A. Tumor quiescence: elevating SOX2 in diverse tumor cell types downregulates a broad spectrum of the cell cycle machinery and inhibits tumor growth. BMC Cancer. 2020;20(1):941. doi: 10.1186/s12885-020-07370-7. PubMed PMID: 32998722; PMCID: PMC7528478.

76. Jahchan NS, Lim JS, Bola B, Morris K, Seitz G, Tran KQ, Xu L, Trapani F, Morrow CJ, Cristea S, Coles GL, Yang D, Vaka D, Kareta MS, George J, Mazur PK, Nguyen T, Anderson WC, Dylla SJ, Blackhall F, Peifer M, Dive C, Sage J. Identification and Targeting of Long-Term Tumor-Propagating Cells in Small Cell Lung Cancer. Cell reports. 2016;16(3):644-56. doi: 10.1016/j.celrep.2016.06.021. PubMed PMID: 27373157; PMCID: PMC4956576.

77. Stewart CA, Gay CM, Xi Y, Sivajothi S, Sivakamasundari V, Fujimoto J, Bolisetty M, Hartsfield PM, Balasubramaniyan V, Chalishazar MD, Moran C, Kalhor N, Stewart J, Tran H, Swisher SG, Roth JA, Zhang J, de Groot J, Glisson B, Oliver TG, Heymach JV, Wistuba I, Robson P, Wang J, Byers LA. Single-cell analyses reveal increased intratumoral heterogeneity after the onset of therapy resistance in small-cell lung cancer. Nature Cancer. 2020;1(4):423-36. doi: 10.1038/s43018-019-0020-z.

78. Simpson KL, Stoney R, Frese KK, Simms N, Rowe W, Pearce SP, Humphrey S, Booth L, Morgan D, Dynowski M, Trapani F, Catozzi A, Revill M, Helps T, Galvin M, Girard L, Nonaka D, Carter L, Krebs MG, Cook N, Carter M, Priest L, Kerr A, Gazdar AF, Blackhall F, Dive C. A biobank of small cell lung cancer CDX models elucidates inter- and intratumoral phenotypic heterogeneity. Nature Cancer. 2020;1(4):437-51. doi: 10.1038/s43018-020-0046-2.

79. Lim JS, Ibaseta A, Fischer MM, Cancilla B, O'Young G, Cristea S, Luca VC, Yang D, Jahchan NS, Hamard C, Antoine M, Wislez M, Kong C, Cain J, Liu YW, Kapoun AM, Garcia KC, Hoey T, Murriel CL, Sage J. Intratumoural heterogeneity generated by Notch signalling promotes small-cell lung cancer. Nature. 2017;545(7654):360-4. doi: 10.1038/nature22323. PubMed PMID: 28489825; PMCID: PMC5776014.

Figure Legends

Figure 1 Sox2 is required for SCLC formation. **A,** Genetically engineered mouse model for the study of *Sox2* in SCLC. **B,** Representative H&E stained lung sections from *Rb1*$^{\text{lox/lox}}$; *p53*$^{\text{lox/lox}}$; *p130*$^{\text{lox/lox}}$; *Sox2*$^{+/+}$ (left), *Rb1*$^{\text{lox/lox}}$; *p53*$^{\text{lox/lox}}$; *p130*$^{\text{lox/lox}}$; *Sox2*$^{+/\text{lox}}$ (middle), and *Rb1*$^{\text{lox/lox}}$; *p53*$^{\text{lox/lox}}$; *p130*$^{\text{lox/lox}}$; *Sox2*$^{\text{lox/lox}}$ (right) mice, 6 months after Cre recombination. **C,** Number of tumors as indicated by H&E staining 3 months after Cre recombination. **D,** Number of tumors as indicated by H&E staining 6 months after Cre delivery. Numbers of mice used in C-D can be found in Supplemental Table S1. **E,** Kaplan-Meier survival curve of SOX2 WT mice (*Sox2*$^{+/\text{lox}}$) compared to *Sox2*$^{\text{lox/lox}}$ mice. Violin plots show median (white dot), interquartile range (box) and the continuous distribution of the data; significance for all panels determined by a two-tailed t-test where * = $P<0.05$, ** = $P<0.01$, *** = $P<0.01$.

Figure 2 RB represses SOX2 and is required for SCLC. **A,** Using 3 hairpins designed to murine Sox2, (shSox2-1,2,&4) and one designed to human SOX2 (shSOX2-5) we tested the effect on cellular proliferation by an Alamar Blue assay in KP1, KP3, H29, and H82 cell lines. **B,** Alamar Blue assays of H29 and H1836 cells after transfection with *RB1 CDK*. **C,** Expression of *Rb1* and *Sox2* measured by qPCR after transduction of Adeno-Rb1 virus in KP1 and KP3 cells. **D,** ChIP of HA-RB CDK or mock transfected cells (H29 and H1836). Regions tested for ChIP enrichment by qPCR are the *SOX2* proximal promoter (PP), the *SRR1* and *SRR2* enhancers of *SOX2*, *MCM3* promoter as a positive control and *ACTB* promoter as a negative control. **E,** Alamar Blue assay on day 4 to determine the proliferation of H29, H82, H1836, and H209 cells after transfection with *RB1 CDK*, and or *SOX2-t2a-GFP*. Proliferation is plotted as the fold

change compared to a mock-transfected control. Individual values are notated by grey circles. Bar graphs show mean and SEM, significance for all panels determined by a two-tailed t-test where * = P<0.05, ** = P<0.01, *** = P<0.01.

Figure 3 SOX2 regulates key SCLC pathways. **A,** SOX2 CUT&RUN heatmap from H1836 and H29 cell lines. Each row represents the normalized read counts at all peaks identified for SOX2 binding. **B,** *De novo* motif identification discovers an HMG domain as the most prominent motif in the SOX2 peaks. **C,** Top ten GO terms enriched at the genes associated with the SOX2 peaks. **D,** PCA plot of other human SOX2 genome binding profiles. Studies include samples from induced pluripotent stem (iPS) cells, embryonic stem (ES) cells, neural stem cells (NSCs) and iPS-derived NSCs, and glioblastoma (GBM). Datasets include GSE69479, GSE81900, GSE49405, GSE23839 (57-60). **E,** Density plot of the log(fpkm) values of all genes associated with a SOX2 peak. **F,** Number of genes in (**E**) that are predicted to be part of the low- or high-expression group after Gaussian mixed model clustering (Supplemental Fig. S4). **G & H,** WGCNA identified two networks that include *ASCL1* and *MYC*. Color scale reflects the relative expression of each gene in the network from the expression profiles available in the CCLE.

Figure 4 SOX2 partially regulates *NEUROD1* **A,** Heatmap of the log transformed fpkm values from SCLC cell lines from CCLE. Cell lines are clustered independently from *SOX2* expression. **B,** Transfection of H1836 and H209 (SCLC-A) and H29 and H82 (SCLC-N) with *SOX-t2a-GFP*. qPCR of *ASCL1* and *NEUROD1* are shown. **C,** qPCR of *ASCL1* and *NEUROD1* are shown upon Cas9-mediated knockdown of *SOX2*. **D,** Western blots for NEUROD1, SOX2, and TUBULIN

after SOX2 overexpression or a mock transfected as a control. **E,** Quantitation of NEUROD1 protein levels as assessed by western blotting as in (**E**), n=3. **F,** ChIP of SOX2 or an IgG control assessed by qPCR at *SOX2*, *NEUROD1*, *ASCL1*, *MYC*, and *ACTB* as a negative control. Bar graphs show mean and SEM, significance for all panels determined by a two-tailed t-test where * = P<0.05, ** = P<0.01, *** = P<0.01.

Figure 5 SOX2 is a regulator of *MYC* and *MYCL* in SCLC. **A,** Gene track showing SOX2 CUT&RUN at the *MYC and MYCL* loci in H1836 and H29 cells. Blue/green track represents the normalized read maps across the loci, the black bars under the track represent regions where significant peaks were called. **B,** qPCR of *MYC* and *MYCL* in H1836 and H209 (SCLC-A) and H29 and H82 (SCLC-N) plotted as a log$_2$ ratio of *SOX2* overexpressed cells to the mock control. Values greater than one indicate higher expression upon *SOX2* overexpression. Significance determined by a two-tailed t-test **C,** Quantitation of MYC protein levels as assessed by western blotting as in (**D**), n=3. **D,** Western blot of MYC, SOX2, and TUBULIN after *SOX2* overexpression or mock transfected cells as a control. **E,** Immunohistochemistry of ASCL1, NEUROD1, and MYC in murine SCLC tumors. Representative tumors shown. Scale bar = 100 m. **F,** Quantification of tumors scored as ASCL1$^+$, NEUROD1$^+$, or MYC$^+$ expressing in (**E**). Number of tumors and their staining classifications are notated in Supplemental Table S2. Significance assessed by ANOVA. Bar graphs show mean and SEM, significance identified where * = P<0.05, ** = P<0.01, *** = P<0.01.

Figure 1

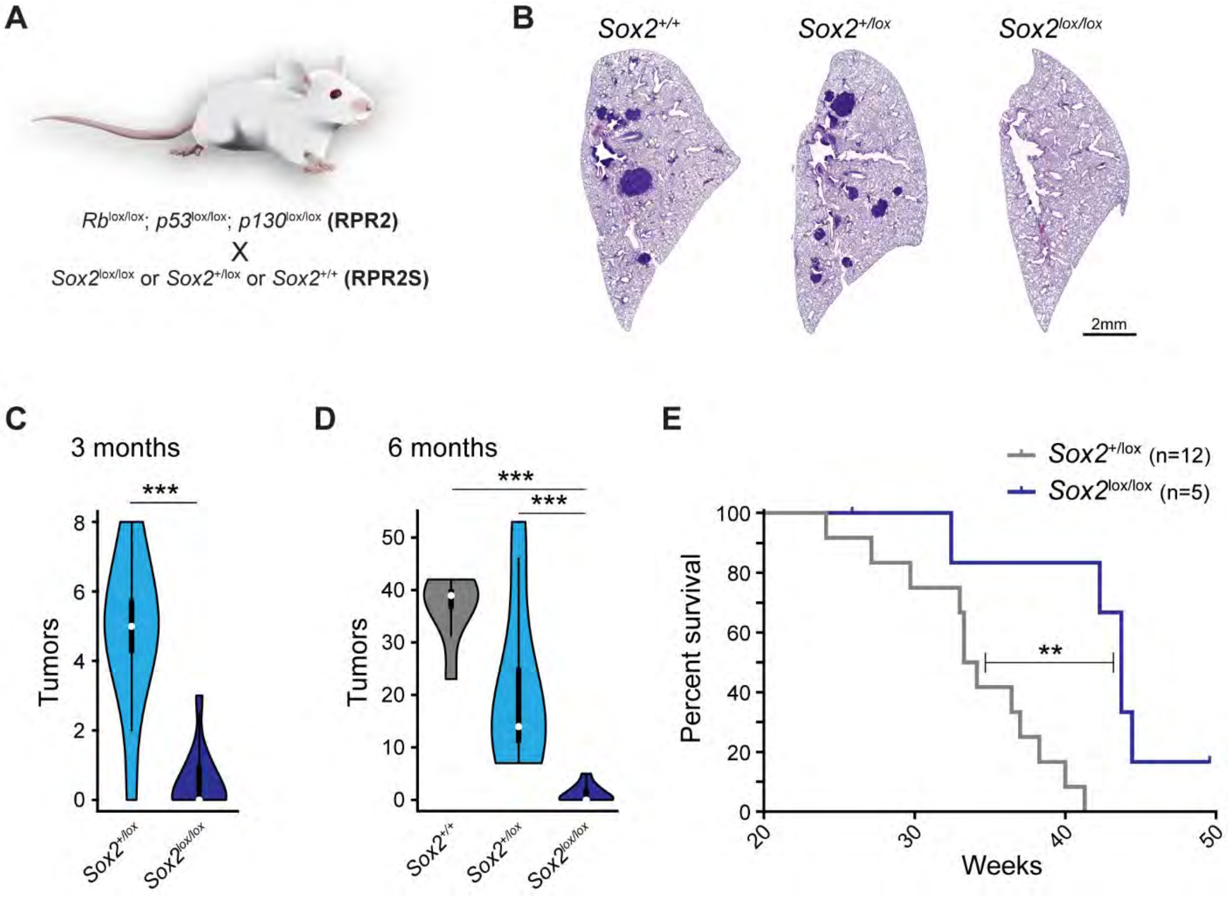

A
$Rb^{lox/lox}$; $p53^{lox/lox}$; $p130^{lox/lox}$ (RPR2)
X
$Sox2^{lox/lox}$ or $Sox2^{+/lox}$ or $Sox2^{+/+}$ (RPR2S)
B
$Sox2^{+/+}$
$Sox2^{+/lox}$
$Sox2^{lox/lox}$
2mm
C
3 months
***
Tumors
0
2
4
6
8
$Sox2^{+/lox}$
$Sox2^{lox/lox}$
D
6 months
***
***
Tumors
0
10
20
30
40
50
$Sox2^{+/+}$
$Sox2^{+/lox}$
$Sox2^{lox/lox}$
E
$Sox2^{+/lox}$ (n=12)
$Sox2^{lox/lox}$ (n=5)
Percent survival
0
20
40
60
80
100
**
20
30
40
50
Weeks

Figure 2

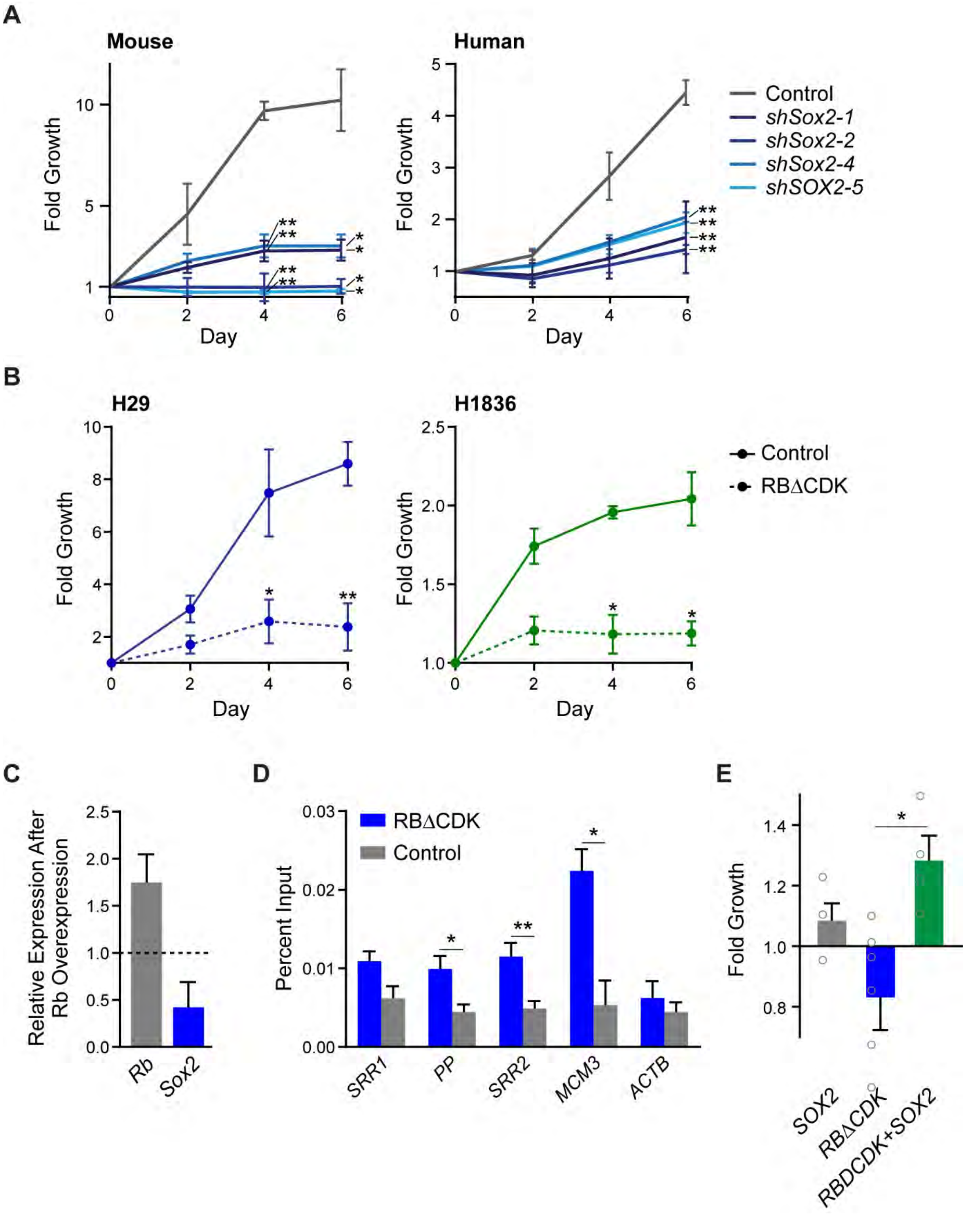

A
Mouse
Fold Growth
10
5
1
0
2
4
6
Day
**
**
**
**
*
*
*
*
Human
Fold Growth
5
4
3
2
1
0
2
4
6
Day
**
**
**
**
Control
shSox2-1
shSox2-2
shSox2-4
shSOX2-5
B
H29
Fold Growth
10
8
6
4
2
0
2
4
6
Day
*
**
H1836
Fold Growth
2.5
2.0
1.5
1.0
0
2
4
6
Day
*
*
Control
RBΔCDK
C
Relative Expression After Rb Overexpression
2.5
2.0
1.5
1.0
0.5
0.0
Rb
Sox2
D
Percent Input
0.03
0.02
0.01
0.00
RBΔCDK
Control
*
**
*
SRR1
PP
SRR2
MCM3
ACTB
E
Fold Growth
1.4
1.2
1.0
0.8
*
SOX2
RBΔCDK
RBDCDK+SOX2

Figure 3

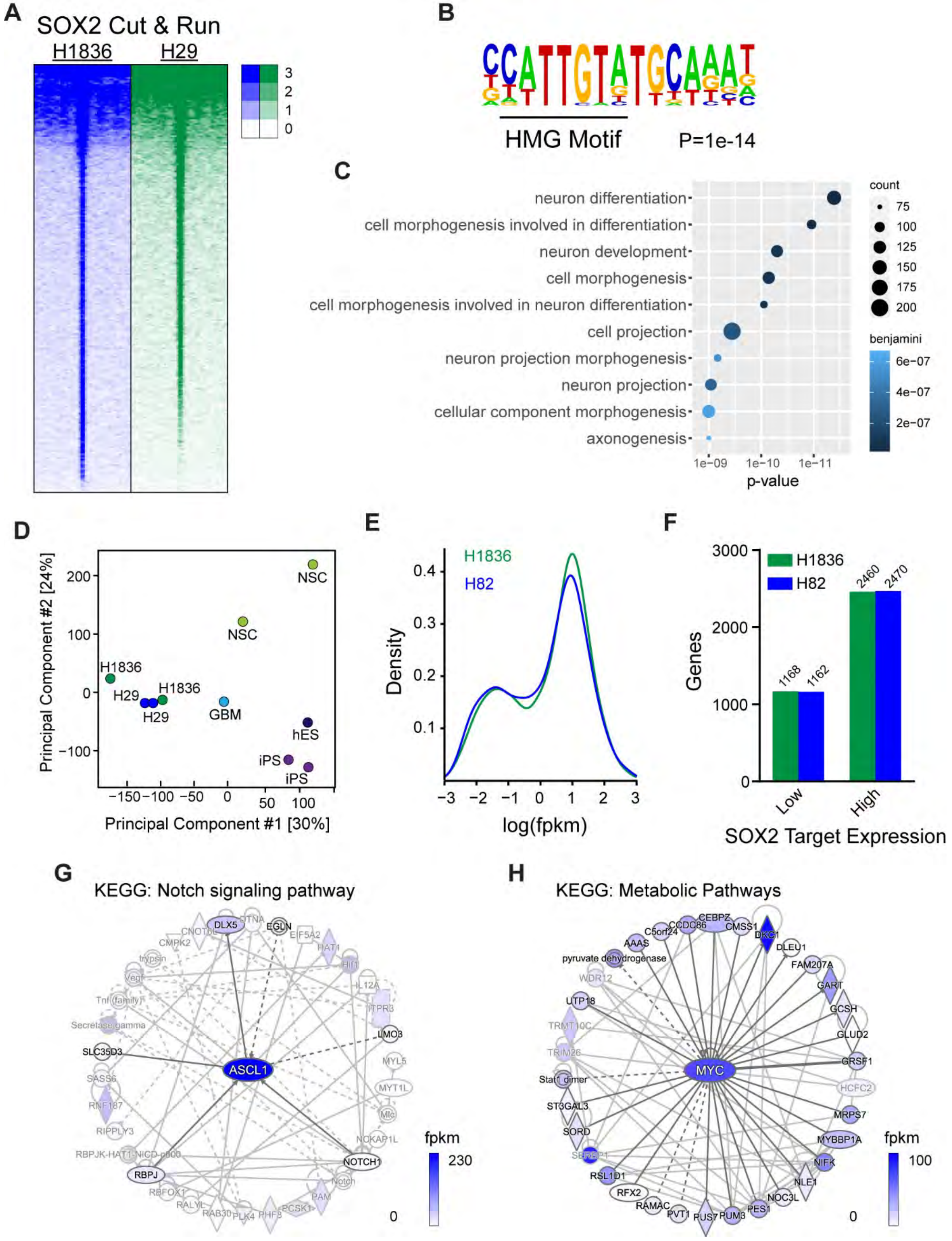

A
SOX2 Cut & Run
H1836
H29
3
2
1
0
B
HMG Motif
P=1e-14
C
neuron differentiation
cell morphogenesis involved in differentiation
neuron development
cell morphogenesis
cell morphogenesis involved in neuron differentiation
cell projection
neuron projection morphogenesis
neuron projection
cellular component morphogenesis
axonogenesis
1e-09
1e-10
1e-11
p-value
count
75
100
125
150
175
200
benjamini
6e-07
4e-07
2e-07
D
Principal Component #2 [24%]
Principal Component #1 [30%]
200
100
0
−100
−150
−100
−50
0
50
100
NSC
NSC
H1836
H29
H1836
H29
GBM
hES
iPS
iPS
E
H1836
H82
Density
0.4
0.3
0.2
0.1
−3
−2
−1
0
1
2
3
log(fpkm)
F
H1836
H82
Genes
3000
2000
1000
0
1168
1162
2460
2470
Low
High
SOX2 Target Expression
G
KEGG: Notch signaling pathway
DLX5
DTNA
CNOT6L
EGLN
EIF5A2
CMPK2
HAT1
trypsin
Hif1
Vegf
IL12A
Tnf (family)
ITPR3
Secretase gamma
LMO3
SLC35D3
ASCL1
MYL5
SASS6
MYT1L
RNF187
Mlc
RIPPLY3
NCKAP1L
RBPJK-HAT1-NICD-p300
NOTCH1
RBPJ
Notch
RBFOX1
PAM
RALYL
PCSK1
RAB30
PLK4
PHF8
fpkm
230
0
H
KEGG: Metabolic Pathways
CEBPZ
C5orf24
CCDC86
CMSS1
AAAS
DKC1
pyruvate dehydrogenase
DLEU1
FAM207A
WDR12
GART
UTP18
GCSH
TRMT10C
GLUD2
TRIM26
GRSF1
MYC
Stat1 dimer
HCFC2
ST3GAL3
MRPS7
SORD
MYBBP1A
SERBP1
NIFK
RSL1D1
NLE1
RFX2
NOC3L
RAMAC
PES1
PVT1
PUS7
PUM3
fpkm
100
0

Figure 4

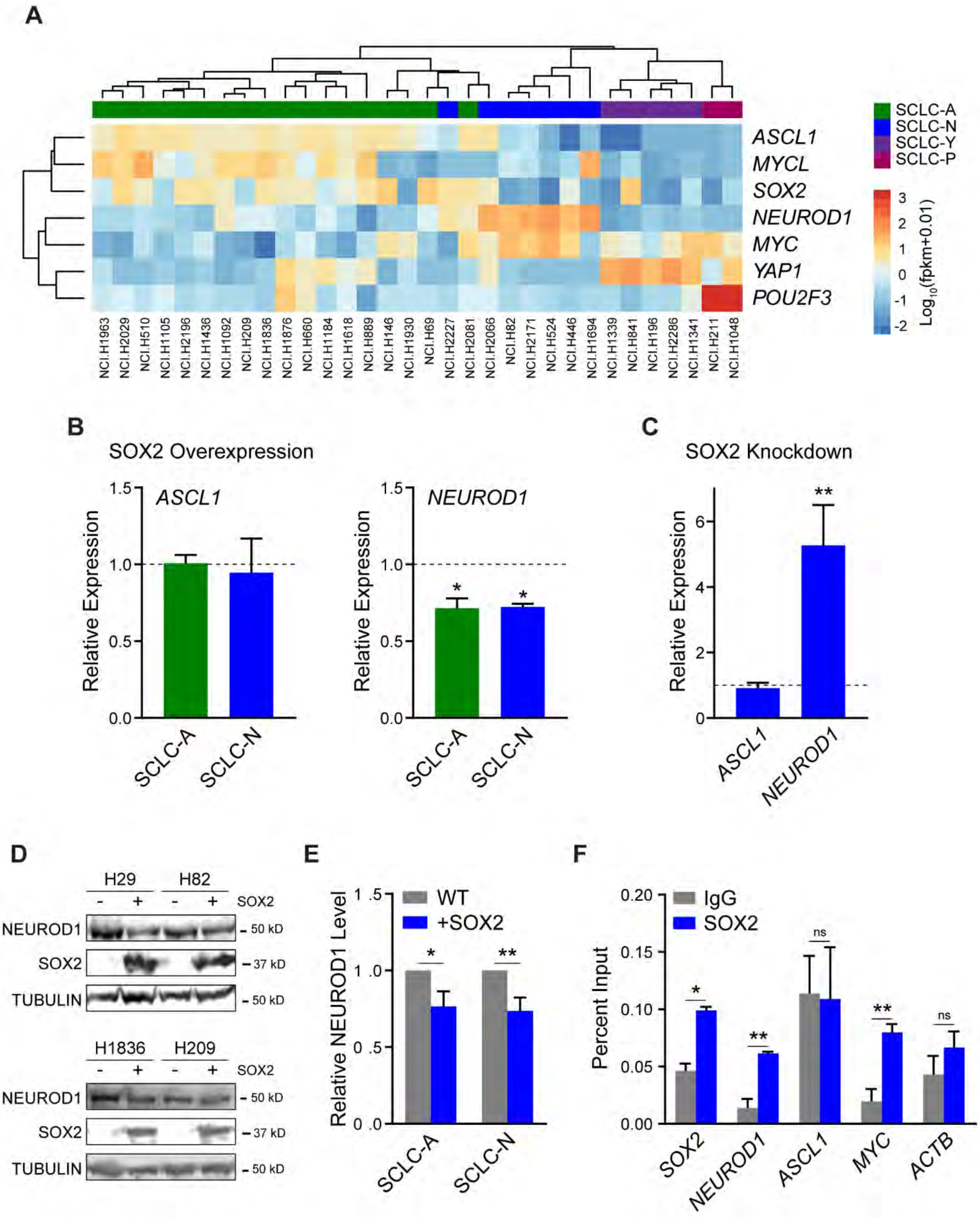
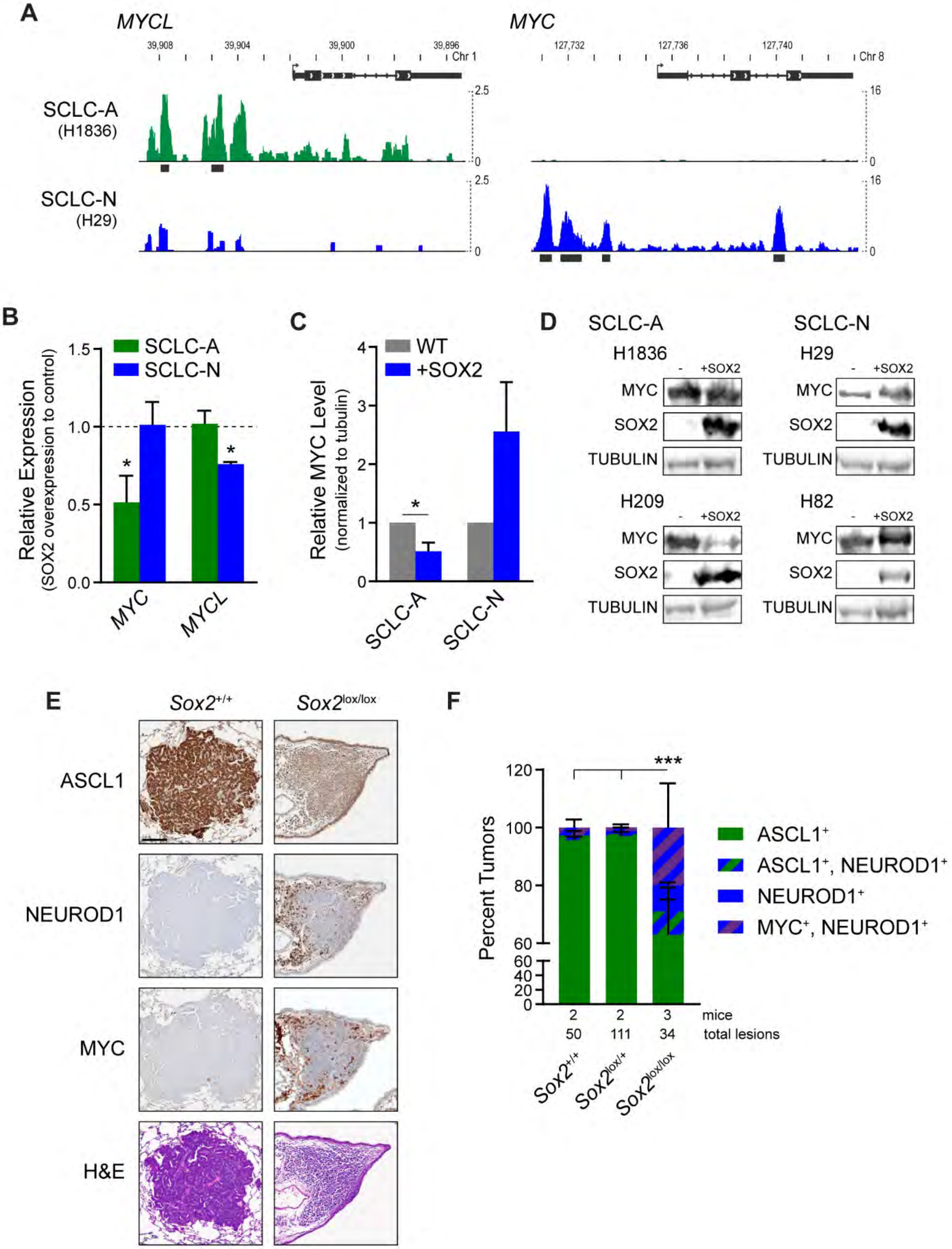

A
SCLC-A
SCLC-N
SCLC-Y
SCLC-P
3
2
1
0
-1
-2
$\mathrm{Log_{10}(fpkm+0.01)}$
ASCL1
MYCL
SOX2
NEUROD1
MYC
YAP1
POU2F3
NCI.H1963
NCI.H2029
NCI.H510
NCI.H1105
NCI.H2196
NCI.H1436
NCI.H1092
NCI.H209
NCI.H1836
NCI.H1876
NCI.H660
NCI.H1184
NCI.H1618
NCI.H889
NCI.H146
NCI.H1930
NCI.H69
NCI.H2227
NCI.H2081
NCI.H2066
NCI.H82
NCI.H2171
NCI.H524
NCI.H446
NCI.H1694
NCI.H1339
NCI.H841
NCI.H196
NCI.H2286
NCI.H1341
NCI.H211
NCI.H1048
B
SOX2 Overexpression
ASCL1
Relative Expression
1.5
1.0
0.5
0.0
SCLC-A
SCLC-N
NEUROD1
Relative Expression
1.5
1.0
0.5
0.0
*
*
SCLC-A
SCLC-N
C
SOX2 Knockdown
Relative Expression
6
4
2
0
**
ASCL1
NEUROD1
D
H29
H82
-
+
-
+
SOX2
NEUROD1
SOX2
TUBULIN
50 kD
37 kD
50 kD
H1836
H209
-
+
-
+
SOX2
NEUROD1
SOX2
TUBULIN
50 kD
37 kD
50 kD
E
Relative NEUROD1 Level
1.5
1.0
0.5
0.0
WT
+SOX2
*
**
SCLC-A
SCLC-N
F
Percent Input
0.20
0.15
0.10
0.05
0.00
IgG
SOX2
*
**
ns
**
ns
SOX2
NEUROD1
ASCL1
MYC
ACTB

Figure 5


A
MYCL
39,908
39,904
39,900
39,896
Chr 1
MYC
127,732
127,736
127,740
Chr 8
SCLC-A
(H1836)
2.5
0
16
0
SCLC-N
(H29)
2.5
0
16
0
B
Relative Expression
(SOX2 overexpression to control)
SCLC-A
SCLC-N
1.5
1.0
0.5
0.0
*
*
MYC
MYCL
C
Relative MYC Level
(normalized to tubulin)
WT
+SOX2
4
3
2
1
0
*
SCLC-A
SCLC-N
D
SCLC-A
SCLC-N
H1836
H29
-
+SOX2
MYC
SOX2
TUBULIN
H209
H82
E
Sox2+/+
Sox2lox/lox
ASCL1
NEUROD1
MYC
H&E
F
Percent Tumors
120
100
80
60
60
40
20
0
***
ASCL1+
ASCL1+, NEUROD1+
NEUROD1+
MYC+, NEUROD1+
2
2
3
mice
50
111
34
total lesions
Sox2+/+
Sox2lox/+
Sox2lox/lox

Appendix 3

Methods in Molecular Biology, 2021, doi: 10.1007/978-1-0716-7_12

Transcriptional Profiling During Neural Conversion

Yohannes Tecleab[1], Hannah Wollenzien[2,3], Michael S. Kareta[1-6]

[1]The Functional Genomics & Bioinformatics Core, [2]Division of Basic Biomedical Sciences, Sanford School of Medicine, University of South Dakota, 414 E. Clark St. Vermillion, SD 57069, USA [3]Genetics and Genomics Group, [4]Cellular Therapies and Stem Cell Biology Group, Sanford Research, 2301 East 60th Street North, Sioux Falls, SD 57104, USA.. [5]Department of Pediatrics, Sanford School of Medicine, 1400 W. 22nd St., Sioux Falls, SD 57105, USA. [6]Department of Chemistry and Biochemistry, South Dakota State University, 1175 Medary Ave, Brookings, SD 57006, USA.

ABSTRACT

The processes that underlie neuronal conversion ultimately involve a reorganization of transcriptional networks to establish a neuronal cell fate. As such, transcriptional profiling is a key component towards understanding this process. In this chapter, we will discuss methods of elucidating transcriptional networks during neuronal reprogramming, and considerations that should be incorporated in experimental design.

KEY WORDS Neuronal conversion, Induced neurons, Transcriptional profiling, RNA-seq

1 INTRODUCTION

Cellular reprogramming is a powerful process that harnesses the potential of the genome to alter a cell's identity. Typically, the direct conversion process involves the transfer of cDNAs, mRNAs, or proteins that harbor master regulator activities, or small molecules that influence master regulator function to a differentiated cell type to drive the conversion of that cell to one of a different lineage, without going through a pluripotent intermediate. This approach utilizes the existing genetic material in a cell to lead to transcriptomic changes that drive cell fate conversion. One of the first and best-studied systems for understanding direct conversion is the formation of induced neuronal (iN) cells. Reported in 2010 from the lab of Marius Wernig, the first account of fully functional iN cells being induced from mouse fibroblasts used a three-factor reprogramming cocktail [1]. Since then, a number of groups have recapitulated this process, with slightly different combinations of factors and culture conditions using both human and mouse differentiated cells as source material [2-6,1,7-9]. Generally, reprogramming of iN cells is defined and characterized by a number of morphological, molecular, and functional parameters [7], however given that the factors supplied to drive reprogramming are often transcriptional regulators such as *Ascl1,* transcriptional profiling should serve as a powerful tool for the understanding of transcriptional states of these cells [10,11,5,6]. Transcriptomic profiling of reprogrammed cells allows for the understanding of the complex transcriptomic changes that occur during cell fate switching, aids in understanding lineage hierarchies, and identifies transcriptional targets of the reprogramming factors.

A powerful player in the field of transcriptomic profiling, RNA sequencing served for many years as the foremost technique in understanding gene expression at a tissue level. In this method of bulk transcriptomics, the expression of all RNAs in the cellular population is sequenced and used to generate an expression profile of the sample, allowing for the evaluation of transcriptional networks that are activated or repressed during the reprogramming process [12]. Bulk RNA sequencing at various timepoints throughout the reprogramming process has been used to uncover lineage pathways that emerge as these cells are reprogrammed [13]. The transcriptomic data gained from bulk RNA sequencing can be used to determine in an unbiased manner the fate of reprogrammed cells and illuminate the intermediate states that were traversed during reprogramming [10,14,15].

While bulk RNA sequencing is an important tool for understanding the transcriptome of iN cells, it does not account for the heterogeneity of cellular populations that occurs in tissues or during the reprogramming process. Bulk approaches average the contribution of the transcriptome of all the cells, which can lead to the masking of rare populations within the sample. An emergent powerhouse in transcriptomic analysis, single cell RNA sequencing (scRNA-seq) instead evaluates the transcriptome of each individual cell in a population, and using complex bioinformatics tools, can stratify individual populations of cells within a sample to determine a more complete picture of the state of these cells. Using scRNA-seq, we have the ability to evaluate differential gene expression within a sample, draw lineage maps, and identify rare or novel populations [16-18]. scRNA-seq has been used in the field of iN differentiation, and has helped to identify the transcriptional networks that "prime" a cell for differentiation, and create a comprehensive profile of transcriptional reprogramming states [19]. Using scRNA-seq to uncover the lineage path and single-cell transcriptomes regulated at various timepoints throughout reprogramming has led to the comprehensive characterization of clonal populations and heterogeneity present during iN reprogramming [5]. While scRNA-sequencing is a powerful tool for transcriptional profiling, due to its high sensitivity, special considerations must be taken to ensure analysis faithfully recapitulates the biological phenotypes [16,18,20].

This protocol will describe the general methods and considerations that should be considered in the transcriptomic analysis of iN transcriptome analysis. Due to the wide variety of methods to prepare samples for transcriptomic analysis, and since many of these involve the use of proprietary kits with established protocols, we will instead focus on the sample design considerations and downstream data analysis in regards to understanding the changes in the transcriptome during neuronal reprogramming.

2 MATERIALS

1. Basic RNA-extraction methods such as Trizol to isolate RNA for bulk RNA-seq or scRNA-seq library preparation platforms such as a 10X Chromium Controller (10X Genomics, Pleasanton, CA, USA) or a Fluidigm C1 system (Fluidigm, South San Francisco, CA, USA).
2. Large computational resources that run on a UNIX environment for most command-line software. An installation of R is required for many computational tools and can be run in a Unix, Windows, or iOS environment [21].

3 METHODS

3.1 Bulk RNA-seq of reprogrammed cells

Study designs in cell reprogramming typically involve time course sampling to monitor changes in expression profile as cell differentiation progresses. A minimum of three replicate samples from each time point is required for statistically reasonable results. For studies involving bulk RNAseq, differential gene expression analysis is typically followed by downstream pathway enrichment analysis.

Bioinformatics analysis pipeline is depicted in Figure 1 and the details of the steps involved are provided below.

a. **Checking the quality of sequencing**. Raw reads from sequencer are typically in the form of fastq format and come in general with adapter sequences clipped. Nevertheless, it is good practice to first do quality assessment of the reads and check for presence of adapter contamination. The most commonly used tool for quality is FASTQC [22] and for trimming Trimmomatic [23].
b. **Alignment to reference genome**. There are several widely used packages for alignment of reads to the reference genome, including Hisat2 [24], STAR [25].
c. **Read counting**. Reads mapping to genomic regions (genes) are counted using several commonly used packages like featurecounts in Rsubread package [26] and HTseq-count [27]. STAR aligner has "quantoMode" option in the mapping stage that counts reads.
d. **Differential gene expression**. A full statistical model of expression in relation to condition and time points and their interaction (Expression ~ Condition + Time + Time*Condition) and a reduced model (Expression ~ Condition + Time) are run in DESeq2 [28]. Then a likelihood ratio test between the full and reduced models evaluates changes in expression between conditions at any time points beyond the first. This can be followed by clustering analysis to identify groups of genes that share a similar expression profile in time. Other packages that are frequently used and have capabilities for the above two mentioned tests are EdgeR [29] and limma [30]. An alternative approach is to

model expression as a continuous function of time to find genes that show significant difference in their pattern or to cluster genes based on their similarity of expression trajectory in time. The R packages maSigPro [31] and ImpulseDE2 [32] perform well for identifying significant changes in time and between conditions. The python package DPGP (Dirichlet process Gaussian process mixture model) can be used to identify groups of genes with similar temporal trajectory [33].

e. **Gene set enrichment**. Pathway analysis in RNAseq data is primarily done using the hypergeomeric test, evaluating overrepresentation of pathways given the differentially expressed genes. However, there is length bias in the probability of being differentially expressed (DE), which is inherent in RNAseq data. This combined with the difference in length distribution of genes within pathways leads to bias in enrichment analysis. The result is that pathways that are disproportionately composed of large genes will more likely be declared enriched in all conditions making the analysis lack specificity to the condition being investigated. By controlling for gene length the package GOseq [34] yields enrichment results that are more relevant to the study at hand. Commonly used and publicly available gene sets include the Gene Ontology Consortium [35,36], The Kyoto Encyclopedia of Genes and Genomes (KEGG) [37], and a collection of gene sets at the Broad Institute [38].

3.2 Single cell transcriptomic analysis of neuronal conversion

Singe cell RNAseq data suffers from high technical variability that arises during sample processing. This can be minimized by ensuring that multiple biological replicates are mixed in each batch [39]. Cells within batches are later assigned to samples based on unique genotypes for example using Demuxlet [40]. scRNAseq analysis pipeline is different from the bulk RNAseq in the preprocessing stages and in the statistical analysis of the expression data. scRNAseq data is typically multiplexed with unique identifiers provided for each cell and sometimes for individual transcripts. Preprocessing steps are thus required to assign reads to their respective unique cells. Subsequent steps are similar to bulk RNAseq analysis. Statistical analysis of scRNA data initially is a classification problem with the objective being to find unique groupings of the cells based on their similar expressions. Once groupings are identified, differential expression tests between groups are used to find unique markers for each unique cell group.

Bioinformatics analysis pipeline is depicted in Figure 2 and the details of the steps involved are provided below.

a. **Demultiplexing of reads**. Many sequencers supply individual fastq files for each sample/cell. In some cases, one fastq file containing all cell barcodes and sample mapping is provided. In this this case, demultiplexing tools like Sabre [41], UmiTools [42], and zUmi [43] are used to allocate reads to their respective samples, strip the barcode and UMI from the reads, and create individual fastq files for each cell.
b. **Alignment to reference genome**. Unlike bulk RNAseq, scRNAseq data set comprises thousands of cells. As result, the mapping step requires fast algorithms to get results in reasonable time. STAR and Kallisto are the two fast aligners that are routinely used in scRNAseq data analysis [25,44].
c. **Quality control of cells**. Indicators of poor quality cells include prevalence of large number of mitochondrial genes and very low or very large numbers of genes [45]. Prevalence of mitochondrial genes is an indicator that the cell was not viable. Very low

number of detected genes implies that the starting material is an empty droplet. While large number of genes implies multiple cells in a droplet. To retain good quality cells based on the above criteria, it is best practice to explore each dataset and find outlier cells to remove from the dataset. The R package Seurat [46] package has procedures to explore and visualize the distribution of the number of genes and percentage mitochondrial genes. These can be used to set cutoff points relevant to the data at hand. Good quality cells also tend to have higher mapping rate, lower number of duplicates and lower ERCC spike-in to exonic read counts ratio [45]. Since these metrics are likely to differ from study to study, a reasonable approach is to explore their distribution (e.g distribution of number of unique mappers) across all cells and remove cells that are outliers [e.g. see 47].

d. **Normalization**. Choice of normalization along with the library preparation method has the biggest effect on the downstream analysis of differential expression in scRNAseq data [48]. Normalization methods developed for bulk RNAseq data are usually not appropriate for scRNAseq data [49]. Methods developed for scRNAseq include BASiCS [50], GRM [51], SCnorm [52], and Scran [53]. Using simulated data, Vallejos *et al* (2017) and Vieth *et al* (2019) evaluated bulk-based and scRNA-specific methods and found that Scran and SCnorm provided a better normalization with stable number of highly variable genes for clustering. The widely used scRNA analysis package Seurat has built-in log-normalization methods. For robust results, data is first normalized using methods specific to scRNA such as Scran or SCnorm. The normalized data can then be fed in to popular scRNA tools such as Seurat.

e. **Clustering of cells.** Unsupervised clustering methods and dimension reduction techniques are combined to partition cells in to distinct groups based on distance. Among the dimension reduction techniques typically used are Principal Components Analysis (PCA), tSNE, and UMAP. These also are critical in data visualizations on lower number of dimensions for assessing cell clusters.

f. **Marker gene identification and labeling**. Once the unique cells are identified, tests are conducted to find marker genes that are differentially expressed in each group in contrast to the others. The AllMarkers function in Seurat (R package) conducts DE test and has options for different types of tests including Negative binomial as in DEseq2. To label the cell types, marker genes from each cell type are compared with markers from known cells. Where available, correlation analysis between the identified cell types and gene expression data from known cell types or from bulk RNAseq data from tissues enriched with cells of interest can be used as a strong indicator of cell identity. One can also conduct pathway enrichment analysis for each identified cell type to determine their putative functions. Confidence on the identity of the cell types can be strengthened if evidence from presence of markers, high correlation, and presence of cell specific pathways is combined.

g. **Ordering of cells based on expression trajectory**. There is a large selection of methods for ordering of cells based on the progression of their expression [54]. Choice of which methods to use depends on the kind of trajectory one is looking for, i.e. linear, cyclic, or tree. In their extensive evaluation of 29 different methods, Saelens *et al* (2019), found that methods perform well in correctly detecting trajectories they were originally designed for. Based on their results, the authors provide a practical guideline for choosing an appropriate tool. For example reCAT [55], outperforms other methods when the underlying trajectory is a cycle. While Monocle DDRTree [56-58], Slingshot [59] and TSCAN [60] perform well if the underlying trajectory is a more complex branching tree. Slingshot and TSCAN perform well when the underlying trajectory is a bifurcation. For discovering linear trajectories SCORPIOUS [61] performs better. Besides finding the trajectory some of these tools provide functions to find genes that cause bifurication (Slingshot) or show marked changes along the trajectory eg. Monocle DDrTree, Slingshot, and TSCAN. A recently developed tool tradSeq [62], can be used downstream of the above packages to detect differential expression of genes along a lineage or between lineages. Using general additive models, TradeSeq fits gene expression as a continuous function of psuedotime which affords flexibility in identifying marker genes at different points.

h. **Elucidating differentially regulated genetic networks.** Pathways enrichment analysis in scRNA is conducted similarly as in bulk RNA-seq data using the package GOseq [34]. Pathway enrichment analysis on marker genes of each unique cell type can help highlight differences in function between cells. Similarly, pathways enriched in genes that cause bifurcation events or genes that show significant changes along lineages can be identified. In these cases, these biological processes are the putative causes of the branching events or the differentiation of the cells.

ACKNOWLEDGEMENT

This work was supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103620 (to MK and YT) and by the National Institutes of Health grant number

R01CA233661 supported by the National Cancer Institute and the National Institute of General Medical Sciences (to MK).

REFERENCES

1. Vierbuchen T, Ostermeier A, Pang ZP, Kokubu Y, Sudhof TC, Wernig M (2010) Direct conversion of fibroblasts to functional neurons by defined factors. Nature 463 (7284):1035-1041. doi:nature08797 [pii] 10.1038/nature08797

2. Ambasudhan R, Talantova M, Coleman R, Yuan X, Zhu S, Lipton SA, Ding S (2011) Direct reprogramming of adult human fibroblasts to functional neurons under defined conditions. Cell Stem Cell 9 (2):113-118. doi:10.1016/j.stem.2011.07.002

3. Marro S, Pang ZP, Yang N, Tsai MC, Qu K, Chang HY, Sudhof TC, Wernig M (2011) Direct lineage conversion of terminally differentiated hepatocytes to functional neurons. Cell Stem Cell 9 (4):374-382. doi:10.1016/j.stem.2011.09.002

4. Pang ZP, Yang N, Vierbuchen T, Ostermeier A, Fuentes DR, Yang TQ, Citri A, Sebastiano V, Marro S, Sudhof TC, Wernig M (2011) Induction of human neuronal cells by defined transcription factors. Nature 476 (7359):220-223. doi:10.1038/nature10202

5. Treutlein B, Lee QY, Camp JG, Mall M, Koh W, Shariati SA, Sim S, Neff NF, Skotheim JM, Wernig M, Quake SR (2016) Dissecting direct reprogramming from fibroblast to neuron using single-cell RNA-seq. Nature 534 (7607):391-395. doi:10.1038/nature18323

6. Tsunemoto R, Lee S, Szucs A, Chubukov P, Sokolova I, Blanchard JW, Eade KT, Bruggemann J, Wu C, Torkamani A, Sanna PP, Baldwin KK (2018) Diverse reprogramming codes for neuronal identity. Nature 557 (7705):375-380. doi:10.1038/s41586-018-0103-5

7. Yang N, Ng YH, Pang ZP, Sudhof TC, Wernig M (2011) Induced neuronal cells: how to make and define a neuron. Cell Stem Cell 9 (6):517-525. doi:10.1016/j.stem.2011.11.015

8. Yoo AS, Sun AX, Li L, Shcheglovitov A, Portmann T, Li Y, Lee-Messer C, Dolmetsch RE, Tsien RW, Crabtree GR (2011) MicroRNA-mediated conversion of human fibroblasts to neurons. Nature 476 (7359):228-231. doi:10.1038/nature10323

9. Chanda S, Ang CE, Davila J, Pak C, Mall M, Lee QY, Ahlenius H, Jung SW, Sudhof TC, Wernig M (2014) Generation of induced neuronal cells by the single reprogramming factor ASCL1. Stem cell reports 3 (2):282-296. doi:10.1016/j.stemcr.2014.05.020

10. Lin M, Lachman HM, Zheng D (2016) Transcriptomics analysis of iPSC-derived neurons and modeling of neuropsychiatric disorders. Mol Cell Neurosci 73:32-42. doi:10.1016/j.mcn.2015.11.009

11. Tekin H, Simmons S, Cummings B, Gao L, Adiconis X, Hession CC, Ghoshal A, Dionne D, Choudhury SR, Yesilyurt V, Sanjana NE, Shi X, Lu C, Heidenreich M, Pan JQ, Levin JZ, Zhang F (2018) Effects of 3D culturing conditions on the transcriptomic profile of stem-cell-derived neurons. Nat Biomed Eng 2 (7):540-554. doi:10.1038/s41551-018-0219-9

12. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10 (1):57-63. doi:10.1038/nrg2484

13. Wapinski OL, Vierbuchen T, Qu K, Lee QY, Chanda S, Fuentes DR, Giresi PG, Ng YH, Marro S, Neff NF, Drechsel D, Martynoga B, Castro DS, Webb AE, Sudhof TC, Brunet A, Guillemot F, Chang HY, Wernig M (2013) Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons. Cell 155 (3):621-635. doi:10.1016/j.cell.2013.09.028

14. Hjelm BE, Salhia B, Kurdoglu A, Szelinger S, Reiman RA, Sue LI, Beach TG, Huentelman MJ, Craig DW (2013) In vitro-differentiated neural cell cultures progress towards donor-identical brain tissue. Hum Mol Genet 22 (17):3534-3546. doi:10.1093/hmg/ddt208

15. Stein JL, de la Torre-Ubieta L, Tian Y, Parikshak NN, Hernandez IA, Marchetto MC, Baker DK, Lu D, Hinman CR, Lowe JK, Wexler EM, Muotri AR, Gage FH, Kosik KS, Geschwind DH (2014) A quantitative framework to evaluate modeling of cortical development by neural stem cells. Neuron 83 (1):69-86. doi:10.1016/j.neuron.2014.05.035

16. Kulkarni A, Anderson AG, Merullo DP, Konopka G (2019) Beyond bulk: a review of single cell transcriptomics methodologies and applications. Curr Opin Biotechnol 58:129-136. doi:10.1016/j.copbio.2019.03.001

17. Stark R, Grzelak M, Hadfield J (2019) RNA sequencing: the teenage years. Nat Rev Genet 20 (11):631-656. doi:10.1038/s41576-019-0150-2

18. Hwang B, Lee JH, Bang D (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 50 (8):96. doi:10.1038/s12276-018-0071-8

19. Nguyen QH, Lukowski SW, Chiu HS, Senabouth A, Bruxner TJC, Christ AN, Palpant NJ, Powell JE (2018) Single-cell RNA-seq of human induced pluripotent stem cells reveals cellular heterogeneity and cell state transitions between subpopulations. Genome Res 28 (7):1053-1066. doi:10.1101/gr.223925.117

20. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, Leonhardt H, Heyn H, Hellmann I, Enard W (2017) Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell 65 (4):631-643 e634. doi:10.1016/j.molcel.2017.01.023

21. R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria,. https://www.R-project.org/.

22. Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. Babraham Institute. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

23. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (15):2114-2120. doi:10.1093/bioinformatics/btu170

24. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nature methods 12 (4):357-360. doi:10.1038/nmeth.3317

25. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29 (1):15-21. doi:10.1093/bioinformatics/bts635

26. Liao Y, Smyth GK, Shi W (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res 47 (8):e47. doi:10.1093/nar/gkz114

27. Anders S, Pyl PT, Huber W (2015) HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31 (2):166-169. doi:10.1093/bioinformatics/btu638

28. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15 (12):550. doi:10.1186/s13059-014-0550-8

29. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26 (1):139-140. doi:10.1093/bioinformatics/btp616

30. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43 (7):e47. doi:10.1093/nar/gkv007

31. Nueda MJ, Tarazona S, Conesa A (2014) Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. Bioinformatics 30 (18):2598-2602. doi:10.1093/bioinformatics/btu333

32. Fischer D (2019) ImpulseDE2: Differential expression analysis of longitudinal count data sets. R package version 1110

33. McDowell IC, Manandhar D, Vockley CM, Schmid AK, Reddy TE, Engelhardt BE (2018) Clustering gene expression time series data using an infinite Gaussian process mixture model. PLoS computational biology 14 (1):e1005896. doi:10.1371/journal.pcbi.1005896

34. Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11 (2):R14. doi:10.1186/gb-2010-11-2-r14

35. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25 (1):25-29. doi:10.1038/75556

36. Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32 (Supplemental 1):D258-D261

37. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28 (1):27-30. doi:10.1093/nar/28.1.27

38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102 (43):15545-15550. doi:0506580102 [pii] 10.1073/pnas.0506580102

39. Hicks SC, Townes FW, Teng M, Irizarry RA (2018) Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics 19 (4):562-578. doi:10.1093/biostatistics/kxx053

40. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, Wan E, Wong S, Byrnes L, Lanata CM, Gate RE, Mostafavi S, Marson A, Zaitlen N, Criswell LA, Ye CJ (2018) Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol 36 (1):89-94. doi:10.1038/nbt.4042

41. Nowosad J, Stepinski T (2018) Spatial association between regionalizations using the information-theoretical V-measure. International Journal of Geographical Information Science 32 (12):2386-2401. doi:10.1080/13658816.2018.1511794

42. Smith T, Heger A, Sudbery I (2017) UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res 27 (3):491-499. doi:10.1101/gr.209601.116

43. Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I (2018) zUMIs - A fast and flexible pipeline to process RNA sequencing data with UMIs. bioRxiv:153940. doi:10.1101/153940

44. Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34 (5):525-527. doi:10.1038/nbt.3519

45. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, Teichmann SA (2016) Classification of low quality cells from single-cell RNA-seq data. Genome Biol 17:29. doi:10.1186/s13059-016-0888-1

46. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, 3rd, Hao Y, Stoeckius M, Smibert P, Satija R (2019) Comprehensive Integration of Single-Cell Data. Cell 177 (7):1888-1902 e1821. doi:10.1016/j.cell.2019.05.031

47. Kumar RM, Cahan P, Shalek AK, Satija R, DaleyKeyser A, Li H, Zhang J, Pardee K, Gennert D, Trombetta JJ, Ferrante TC, Regev A, Daley GQ, Collins JJ (2014) Deconstructing transcriptional heterogeneity in pluripotent stem cells. Nature 516 (7529):56-61. doi:10.1038/nature13920

48. Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I (2019) A systematic evaluation of single cell RNA-seq analysis pipelines. Nature communications 10 (1):4667. doi:10.1038/s41467-019-12266-7

49. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC (2017) Normalizing single-cell RNA sequencing data: challenges and opportunities. Nature methods 14 (6):565-571. doi:10.1038/nmeth.4292

50. Vallejos CA, Marioni JC, Richardson S (2015) BASiCS: Bayesian Analysis of Single-Cell Sequencing Data. PLoS computational biology 11 (6):e1004333. doi:10.1371/journal.pcbi.1004333

51. Ding B, Zheng L, Zhu Y, Li N, Jia H, Ai R, Wildberg A, Wang W (2015) Normalization and noise reduction for single cell RNA-seq experiments. Bioinformatics 31 (13):2225-2227. doi:10.1093/bioinformatics/btv122

52. Bacher R, Chu LF, Leng N, Gasch AP, Thomson JA, Stewart RM, Newton M, Kendziorski C (2017) SCnorm: robust normalization of single-cell RNA-seq data. Nature methods 14 (6):584-586. doi:10.1038/nmeth.4263

53. Lun AT, McCarthy DJ, Marioni JC (2016) A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res 5:2122. doi:10.12688/f1000research.9501.2

54. Saelens W, Cannoodt R, Todorov H, Saeys Y (2019) A comparison of single-cell trajectory inference methods. Nat Biotechnol 37 (5):547-554. doi:10.1038/s41587-019-0071-9

55. Liu Z, Lou H, Xie K, Wang H, Chen N, Aparicio OM, Zhang MQ, Jiang R, Chen T (2017) Reconstructing cell cycle pseudo time-series via single-cell transcriptome data. Nature communications 8 (1):22. doi:10.1038/s41467-017-00039-z

56. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen TS, Rinn JL (2014) The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol 32 (4):381-386. doi:10.1038/nbt.2859

57. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C (2017) Single-cell mRNA quantification and differential analysis with Census. Nature methods 14 (3):309-315. doi:10.1038/nmeth.4150

58. Qiu X, Mao Q, Tang Y, Wang L, Chawla R, Pliner HA, Trapnell C (2017) Reversed graph embedding resolves complex single-cell trajectories. Nature methods 14 (10):979-982. doi:10.1038/nmeth.4402

59. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S (2018) Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. BMC genomics 19 (1):477. doi:10.1186/s12864-018-4772-0

60. Ji Z, Ji H (2016) TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res 44 (13):e117. doi:10.1093/nar/gkw430

61. Cannoodt R, Saelens W, Sichien D, Tavernier S, Janssens S, Guilliams M, Lambrecht B, Preter KD, Saeys Y (2016) SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. bioRxiv:079509. doi:10.1101/079509

62. Van den Berge K, Roux de Bezieux H, Street K, Saelens W, Cannoodt R, Saeys Y, Dudoit S, Clement L (2020) Trajectory-based differential expression analysis for single-cell sequencing data. Nature communications 11 (1):1201. doi:10.1038/s41467-020-14766-3

Figure 1 Analysis pipeline for bulk RNA-seq

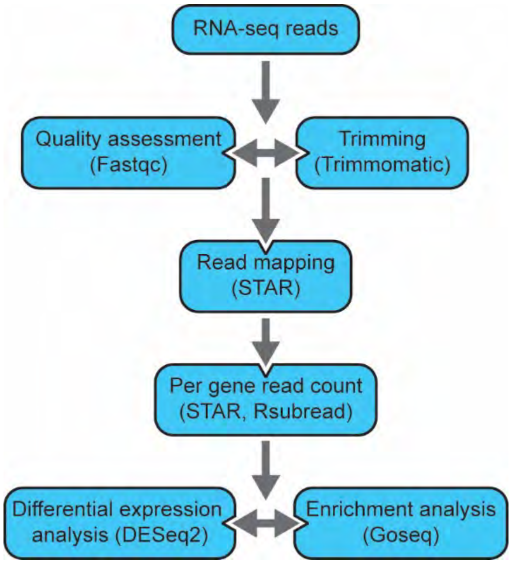
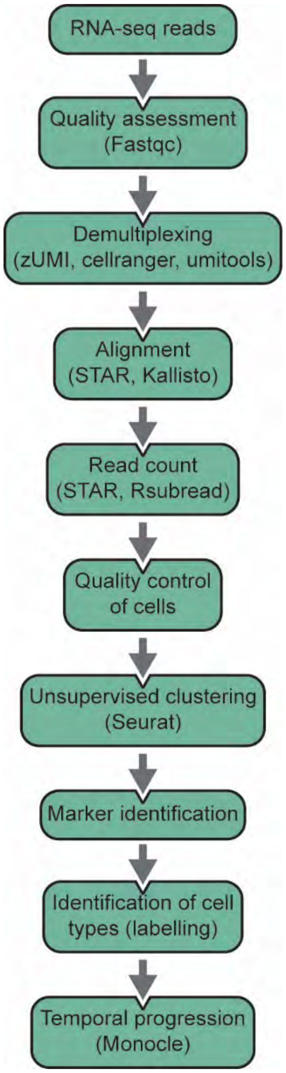

RNA-seq reads
Quality assessment (Fastqc)
Trimming (Trimmomatic)
Read mapping (STAR)
Per gene read count (STAR, Rsubread)
Differential expression analysis (DESeq2)
Enrichment analysis (Goseq)

Figure 2 Analysis pipeline for scRNA-seq


RNA-seq reads
Quality assessment (Fastqc)
Demultiplexing (zUMI, cellranger, umitools)
Alignment (STAR, Kallisto)
Read count (STAR, Rsubread)
Quality control of cells
Unsupervised clustering (Seurat)
Marker identification
Identification of cell types (labelling)
Temporal progression (Monocle)