Inaugural dissertationfor
obtaining the doctoral degreeof
the
Combined Faculty of Mathematics, Engineering and Natural Sciences
of the
Ruprecht - Karls – University
Heidelberg

Presented by
M.Sc. Sara Chocarro Pérez de Azanza
born in: San Sebastian, Spain

Oral examination: 07-07-2023

Cell of origin and fusion variants of *Eml4-Alk* in non-small cell lung cancer

Referees: **Prof. Dr. Elmar Schiebel**

Prof. Dr. Ursula Klingmüller

Prof. Dr. Rocio Sotillo

Dr. Ana Banito

"Believe you can and you're halfway there."
- Theodore Roosevelt

Table of Contents

Table of Contents 1
Summary 5
Zusammenfassung 7
List of Tables 9
List of Figures 10
Abbreviations 13
Introduction 19
1. Lung cancer 19
1.1. Clinical staging 19
1.2. Lung cancer subtypes 21
2. ALK gene 23
3. EML4 gene 24
4. EML4-ALK 24
4.1. Lung cancer 24
4.2. Treatments options for ALK+ lung cancer 26
4.3. Fusion variants 27
4.3. Localization of the variants 28
5. Lung epithelial cell types 29
5.1. Lung cell types in homeostasis 29
5.2. Lung cell types after injury 31
6. Mouse models of NSCLC cancer and linage tracing 33
6.1. Transgenic mouse models 33
6.2. Linage-restricted mouse models 34
6.3. EML4-ALK mouse models 35
7. Cell of origin of lung adenocarcinoma 36
8. TP53 38
8.1. Role of TP53 38
8.2. p53 in EML4-ALK lung cancer 40
Aims of the thesis 42
Materials & Methods 43
Materials 43
Methods 47
Mouse lines, tamoxifen and virus injection 47

Lentiviral plasmid isolation 48
Lentivirus production and titration 49
Mouse genotyping 50
BAC DNA isolation and probe labeling 50
Fluorescence in situ hybridization (FISH) 51
DNA, RNA isolation and cDNA synthesis 52
Efficiency test 52
Tumor genotyping 53
Immunohistochemistry and immunofluorescence 53
Bulk RNA sequencing and bioinformatic analysis 55
Lung dissociation and cell sorting 55
10x single cell RNA sequencing and analysis 56
SMART-Seq RNA sequencing and analysis 57
Statistical analysis 58
Schemes and images 58
Contributions 58
Results 61
1. Characterization of the biological and molecular divergence among *Eml4-Alk* variants 61
1.1 Generation of an Eml4-Alk V3 mouse model 61
1.2. V3 induces lung adenocarcinomas 62
1.3. V3 is able to induce a few lung squamous cell carcinomas 63
1.4. V3 tumors induce a shorter survival in mice than V1 tumors 64
1.5. V1 and V3 adenoviruses generate the *Eml4-Alk* fusion with the same efficiency 65
1.6. Transcriptomic analysis of V1 and V3 tumors 66
2. Cell of origin of *Eml4-Alk* tumors 68
2.1. V3 tumors originate from club and AT2 cells 68
2.2. Club cells give rise to V1 and V3 tumors 70
2.3. AT2-originated tumors induce a shorter survival of the mice 71
2.4. Understanding the development of V1 club tumors 72
2.5. Different timepoints in tumor development show a distinct distribution of cells 73
2.6. Club cells lose their club signature and gain and AT2-like signature upon transformation 75
2.7. Club cells employ their intrinsic regeneration mechanisms to develop a tumor 76
2.8. Club cells can undergo several routes during tumor development 78
3. Tumor initiation and role of Tp53 in *Eml4-Alk* tumors 82

3.1. V1 cells undergo apoptosis early upon transformation 82
3.2. Tp53 loss is a poor prognostic factor for V1 mice 83
3.3. *Trp53* loss affects V1 and V3 tumors differently according to their origin 85
3.4. Single cell RNA sequencing of early timepoints 86
3.5. V1 cells downregulate *Trp53* levels overtime 88
Discussion 91
1. Comparison of the V1 and V3 lung tumors 91
1.1. Generation of a new *Eml4-Alk* V3 mouse model 91
1.2. Comparison of V1 and V3 tumors 92
1.2. The *Eml4-Alk* fusion can lead to lung squamous cell carcinomas 93
2. Cell of origin of Eml4-Alk lung tumors 95
2.1. V3 tumors originate from club and AT2 cells 95
2.2. Club cells employ regeneration mechanisms during tumor development 97
3. Tumor initiation driven by *Eml4-Alk* V1 99
Conclusions 101
1. Characterization of the biological and molecular divergence among *Eml4-Alk* variants 101
2. Cell of origin of *Eml4-Alk* tumors 102
3. Tumor initiation and role of Tp53 in *Eml4-Alk* tumors 103
Appendix I 105
Appendix II 107
Acknowledgements 109
References 111

Summary

Non-small cell lung cancer (NSCLC) is one of the most prevalent and lethal forms of cancer worldwide. *EML4-ALK* is a fusion gene that arises from a chromosomal translocation between the echinoderm microtubule-associated protein-like 4 (*EML4*) and the anaplastic lymphoma kinase (*ALK*) genes. This fusion gene has been found in approximately 5-7% of NSCLC cases, making it an important therapeutic target for this disease. Therefore, this thesis aimed to understand the molecular and biological mechanisms underlying *EML4-ALK* variants in NSCLC, in order to provide insights into the development and progression of this disease which could lead to more effective therapies.

To study the most prevalent *Eml4-Alk* variants in NSCLC, I made use of an available adenoviral system to induce the endogenous V1 rearrangement in mouse lungs, and generated a similar system to induce the V3 fusion gene. The results of this thesis revealed that the *Eml4-Alk* V3 rearrangement induced a diverse range of lung adenocarcinomas, including papillary lesions, solid and lepidic adenocarcinomas, as well as squamous cell carcinomas (SCC). Interestingly, SCC had not been previously observed in the *Eml4-Alk* V1 model.

Careful examination of the mice uncovered that the V3 model induced a significantly reduced overall survival rate compared to V1, correlating with a more aggressive phenotype of this variant observed in patients. This study also demonstrated that V1 cells undergo DNA damage and apoptosis early upon transformation, which could be the reason of the delayed death of V1 mice. Further results suggested that this early phenotype in V1 mice was most probably related to the Tp53 pathway, since loss of *Trp53* induced a reduction in the overall survival rate and increased the number of tumors in V1 mice but not V3 mice.

NSCLC is a heterogeneous disease with diverse molecular subtypes and one of the critical factors that contribute to the variability in NSCLC is the cell of origin. Understanding the cell of origin of NSCLC helps to identify the molecular mechanisms and signaling pathways that drive the cancer's development, progression, and response to therapy and to identify new biomarkers for early detection and prognosis. In this

thesis, I observed that V3 tumors were derived from club and alveolar type 2 (AT2) cells, consistent with previous findings in V1, and that those with an AT2 origin seem to be more aggressive. Additionally, time course single-cell RNA sequencing of the development of V1 club tumors revealed that club cells use their intrinsic regeneration mechanisms during the process of tumor development.

Taken together, these findings provide important insights into the complexity of *EML4-ALK* variants in NSCLC and the diverse ways in which these variants contribute to the development of this disease. They highlight the need for further investigation into the mechanisms underlying *EML4-ALK* variants, in order to guide the development of more effective treatments for NSCLC.

Zusammenfassung

Nicht-kleinzelliges Lungenkarzinom (NSCLC) ist weltweit eine der häufigsten und tödlichsten Krebsarten. *EML4-ALK* ist ein Fusionsgen, das aus einer chromosomalen Translokation zwischen den Genen für das Mikrotubuli-assoziierte Protein 4 (EML4) und die anaplastische Lymphomkinase (ALK) entsteht. Dieses Fusionsgen wurde in etwa 5 % der Fälle von NSCLC gefunden und ist damit ein wichtiges therapeutisches Ziel für diese Krankheit. Ziel dieser Arbeit war es daher, die molekularen und zellulären Mechanismen zu verstehen, die den *EML4-ALK*-Varianten bei nicht-kleinzelligem Lungenkrebs zugrunde liegen, um Einblicke in die Entwicklung und das Fortschreiten dieser Krankheit zu gewinnen, die zu wirksameren Therapien führen könnten.

Zur Untersuchung der häufigsten *Eml4-Alk*-Varianten bei nicht-kleinzelligem Lungenkrebs nutzte ich ein verfügbares adenovirales System, um das endogene V1-Rearrangement in Mäuselungen zu induzieren, und entwickelte ein ähnliches System zur Induktion des V3-Fusionsgens. Die Ergebnisse dieser Arbeit zeigten, dass das Eml4-Alk V3-Rearrangement eine Vielzahl von Tumortypen induziert, darunter papilläre Läsionen, solide, azinäre und lepidische Adenokarzinome sowie Plattenepithelkarzinome (SCC). Interessanterweise war SCC im *Eml4-Alk* V1-Modell bisher nicht beobachtet worden.

Eine sorgfältige Untersuchung der Mäuse ergab, dass das V3-Modell im Vergleich zu V1 eine deutlich geringere Gesamtüberlebensrate aufwies, was mit dem aggressiveren Phänotyp der Patienten korreliert. Die Studie zeigte auch, dass V1-Zellen nach der Transformation frühzeitig DNA-Schäden und Apoptose erleiden, was der Grund für den verzögerten Tod der V1-Mäuse sein könnte. Weitere Ergebnisse deuten darauf hin, dass dieser frühe Phänotyp bei V1-Mäusen höchstwahrscheinlich mit dem Tp53-Signalweg zusammenhängt, da der Verlust von *Trp53* bei V1-Mäusen, nicht aber bei V3-Mäusen, zu einer Verringerung der Gesamtüberlebensrate und einem Anstieg der Tumorzahl führte.

Der NSCLC ist eine heterogene Krankheit mit verschiedenen molekularen Subtypen, und einer der entscheidenden Faktoren, die zur Variabilität des NSCLC beitragen, ist die Ursprungszelle. Das Verständnis der Ursprungszelle von NSCLC hilft bei der Identifizierung der molekularen Mechanismen und Signalwege, die die

Entwicklung und das Fortschreiten des Krebses sowie das Ansprechen auf die Therapie steuern, und bei der Ermittlung neuer Biomarker für die Früherkennung und Prognose. In dieser Arbeit habe ich festgestellt, dass V3-Tumoren aus Keulen- und Alveolarzellen des Typs 2 (AT2) entstanden sind, was mit früheren Erkenntnissen bei V1 übereinstimmt, und dass diejenigen mit AT2-Ursprung aggressiver zu sein scheinen. Darüber hinaus ergab die zeitliche Abfolge der Einzelzell-RNA-Sequenzierung der Entwicklung von V1-Keulentumoren, dass Keulenzellen während der Tumorentwicklung ihre eigenen Regenerationsmechanismen nutzen.

Zusammengenommen bieten diese Ergebnisse wichtige Einblicke in die Komplexität der *EML4-ALK*-Varianten bei NSCLC und in die verschiedenen Arten, in denen diese Varianten zur Entwicklung dieser Krankheit beitragen. Sie machen deutlich, dass die molekularen und zellulären Mechanismen, die den *EML4-ALK*-Varianten zugrunde liegen, weiter untersucht werden müssen, um die Entwicklung wirksamerer Behandlungen für NSCLC zu ermöglichen.

List of Tables

Table 1. TMN classification of lung cancer
Table 2. Staging classification of lung cancer
Table 3. Mouse models of linage tracing and lung adenocarcinoma
Table 4. Mouse strains
Table 5. Mouse genotyping primers
Table 6. PCR primers
Table 7. Primary antibodies
Table 8. Secondary antibodies for immunofluorescence
Table 9. FACS antibodies
Table 10. Equipment
Table 11. Commercial kits
Table 12. Reagents and chemicals
Table 13. Software and algorithms
Table 14. R packages
Table 15. Nick translation reagents
Table 16. Ethanol precipitation
Table 17. Percentage of mice with at least one SCC in the indicated mouse lines.
Table 18. Marker genes used to annotate the cell types in the V1 club sequencing experiment
Table 19. Top 10 markers of cNMF activity programs
Table 20. Marker genes used in the tumor initiation experiment

List of Figures

Figure 1. Mouse lung tumor grading system
Figure 2. ALK protein
Figure 3. EML4 protein
Figure 4. *EML4-ALK* induced lung adenocarcinoma
Figure 5. Most frequent *EML4-ALK* fusion variants
Figure 6. Schematic of the human lung epithelial cell types
Figure 7. Regeneration-related transitional states
Figure 8. TP53 activation and effects
Figure 9. Titration of lentiviral particles
Figure 10. Generation of *Eml4-Alk* V3 tumors
Figure 11. V3 *Eml4-Alk* forms adenocarcinomas and squamous cell carcinomas in vivo
Figure 12. *Eml4-Alk* squamous tumors express canonical markers of squamous cell carcinomas
Figure 13. V3 induces a shorter overall survival of mice compared to V1
Figure 14. V1 and V3 adenoviruses show the same efficiency in generating the *Eml4-Alk* rearrangement *in vivo*
Figure 15. Analysis of the transcriptome of V1 and V3 tumors
Figure 16. Analysis of the differentially expressed genes between V1 and V3 tumors
Figure 17. Club cells and alveolar type II cells give rise to V3 tumors
Figure 18. Club cells are the major cell type giving rise to both V1 and V3 tumors
Figure 19. AT2-originated tumors are a risk factor for mice
Figure 20. Time course single cell RNA-sequencing experiment of V1 club tumor development
Figure 21. Cell type identification in the single cell RNA sequencing data
Figure 22. Distribution and composition of the samples
Figure 23. Club cells acquire an AT2-like identity during tumor development
Figure 24. Tumor cells show enhancement in three independent activity programs
Figure 25. Club cells can follow distinct routes towards lung adenocarcinoma
Figure 26. Comparison of club originated tumors to published AT2 tumors
Figure 27. Club tumor signatures are present in human lung adenocarcinoma
Figure 28. V1 induction results in early DNA damage and apoptosis

Figure 29. *Trp53* loss in AT2 cells or club cells reduces the survival or V1 mice but not V3 mice

Figure 30. *Trp53* loss increases the proliferation rate of V1 AT2 tumors and V3 club tumors

Figure 31. The lentiviral system labels infected cells with GFP

Figure 32. Time course single cell RNA-sequencing experiment of early timepoints

Figure 33. V1 cells downregulate *Trp53*

Figure 34. Graphical abstract of part 1

Figure 35. Graphical abstract of part 2

Figure 36. Graphical abstract of part 3

Abbreviations

A

AAH	Atypical adenomatous hyperplasia
Ad	Adenovirus
ADI	Alveolar differentiation intermediate
AEP	AT2 alveolar epithelial progenitor
ALI	Acute lung injury
ALK	Anaplastic lymphoma kinase
AM	Alveolar macrophages
AT1	Alveolar type I
AT2	Alveolar type II
ATM	Ataxia telangiectasia mutated
ATR	Ataxia-telangiectasia-mutated-and-Rad3-related kinase

B

BAC	Bacterial artificial chromosome
BADJ	Bronchioalveolar duct junction
BASC	Bronchioalveolar stem cell
BrdU	Bromodeoxyuridine

C

C57BL/6	Black 6 strain
Casp3	Cleaved caspase 3
CCSP	Club cell secretory protein
CDK	Cyclin-dependent kinase
cDNA	Complementary DNA
CHK2	Checkpoint kinase 2
CK2	Casein kinase II
CK5	Cytokeratin 5
CMV	Cytomegalovirus
cNMF	Consensus non-negative matrix factorization
CNS	Central nervous system
COPD	Chronic obstructive pulmonary disease

CreER	Cre Recombinase-modified Estrogen Receptor fusion protein
D	
DASC	Distal airway stem cells
DATP	Damage-associated transient progenitor
DEG	Differentially expressed gene
DKFZ	Deutsches Krebsforschungszentrum
DMEM	Dulbecco's Modified Eagle's Medium
DNA	Deoxyribonucleic acid
E	
EGFR	Epidermal growth factor receptor
EML4	Echinoderm microtubule-associated protein-like 4
F	
FACS	Fluorescence-activated cell sorting
FBS	Fetal Bovine Serum
FC	Fold change
FDA	Food and Drug Administration
FDR	False discovery rate
FISH	Fluorescence in situ hybridization
Fl	Floxed
FOXJ1	Forkhead Box J1
FVB	Sensitive to Friend Leukemia Virus B strain
G	
gDNA	Genomic DNA
GEMM	Genetically engineered mouse model
GFP	Green fluorescence protein
GSEA	Gene Set Enrichment Analysis
H	
H&E	Hematoxylin-Eosin
HEK	Human Embryonic Kidney cells
HELP	Hydrophobic EML protein
HOPX	Homeodomain-Only Protein

I

i.p.	Intraperitoneal
IF	Immunofluorescence
IHC	Immunohistochemistry
IM	Interstitial macrophages
IPF	Idiopathic pulmonary fibrosis

K

KP	*Kras^{G12D}; p53$^{fl/fl}$*
KRAS	Kirsten Rat Sarcoma viral oncogene
KRT5	Cytokeratin 5

L

LB	Luria-Bertani
LCLC	Large cell lung cancer
LDLa	Low-density lipoprotein receptor domain class A
LNEP	Lineage-negative epithelial progenitor
LSL	Lox-Stop-Lox
LTF	Lactotransferin
LUAD	Lung adenocarcinoma

M

MAM	Meprin, A5 protein and protein tyrosine phosphatase Mu
MAP	Microtubule associated protein
MDR-1	Multidrug resistance-1
mG	Membranous GFP
MHC	Major Histocompatibility Complex
mT	Membranous tdTomato
MUC5A	Mucin 5A
MUC5B	Mucin 5B

N

Neo	Neomycin
NES	Normalized enrichment score
NK	Natural killer
NKX2-1	NK2 Homeobox 1

NPM1	Nucleophosmin 1
NSCLC	Non-small-cell lung cancer

P

p53$^{fl/fl}$	*p53 floxed*
PAGA	Partition-based graph abstraction
PATS	Pre-alveolar type-1 transitional cell state
PBS	Phosphate Buffered Saline
PCA	Principal component analysis
PCR	Polymerase chain reaction
PDPN	Podoplanin
PFU	Plaque forming units
PNEC	Pulmonary neuroendocrine cells

R

RFP	Red fluorescence protein
RNA	Ribonucleic acid
ROS	Reactive oxygen species
rtTA	Tetracycline-controlled reverse transactivator

S

SCC	Squamous cell carcinoma
SCGB1A1	Secretoglobin Family 1A Member 1
SCLC	Small-cell lung cancer
scRNA-seq	Single cell RNA sequencing
SFTPC/SPC	Surfactant protein C
sgRNA	Single-guide RNA
SPIONs	Super-paramagnetic iron oxide nanoparticles
SSC	Saline-sodium citrate

T

TADs	Tandem transcription activation domains
TAM	Tamoxifen
TAMs	Tumor associated macrophages
TAPE	Tandem atypical beta-propeller
TCGA	The Cancer Genome Atlas

TD	Trimerization domain
TK	Tyrosine kinase
TetR	Tet repressor protein
TetO	Tet operator
TKI	Tyrosine kinase inhibitor
TM	Transmembrane domain
TME	Tumor microenvironment
TP63	Tumor protein 63
TTF-1	Thyroid transcription factor 1

U

UMAP	Uniform Manifold Approximation and Projection
UV	Ultraviolet

V

V1	Variant 1
V2	Variant 2
V3	Variant 3
V5	Variant 5

W

WD	Tryptophan-aspartic acid

Introduction

1. Lung cancer

1.1. Clinical staging

Lung cancer is one of the most commonly diagnosed cancers worldwide, with 2.2 million cases diagnosed annually, and accounting for more cancer-related deaths than any other cancer type, with almost 1.8 million deaths in 2020[1]. The overall 5-year survival rate of lung cancer patients is around 18%. The severity of this cancer type is mainly due to the late diagnosis of the disease, when tumors are very advanced and have already metastasized. The effectiveness of the treatment depends heavily on early detection, which leaves very few options since surgical resection is only possible for small, early-stage tumors. Besides tobacco smoking, long-term exposure to radon gas, pollution, asbestos, a diagnosis of chronic obstructive pulmonary disease (COPD), some occupational exposures, and genetic factors are additional risk factors for lung cancer[2–7].

Lung cancer is staged based on the TNM staging system (**Table 1**), which stands for Tumor, Node, Metastasis. The Tumor section considers the size and location of the main tumor and is divided into four groups. T0 stands for no signs of cancer. T1 describes cancers contained within the lung. T2 denotes a tumor that is either 3 to 5 cm big or that it involves the main airways. T3 refers to a tumor that is between 5 to 7 cm in size, there is presence of several nodules in the same lung lobe, or a cancer that has spread into the chest wall, pleura, or the pericardium. Finally, T4 means the cancer is larger than 7 cm, it appears in more than one lobe, or it has disseminated into other areas such as the diaphragm, heart, trachea, spinal bone, or esophagus. Node describes the number and location of lymph nodes that are affected. N1 denotes the presence of cancerous cells in the lymph nodes of the lung. N2 means there is cancer in lymph nodes in the middle of the chest or on the same side as the diseased lobe. N3 indicates the lymph nodes located in the chest opposite to the afflicted lobe or at the top of the lung are affected. Finally, Metastasis describes the occurrence of distant metastases, and only two categories exist: M0, meaning no metastasis, and M1, indicating the presence of metastasis[8,9].

Table 1. TMN classification of lung cancer. Adapted from Collins, L. et al., American Family Physician, 2007[2].

Tumor	Nodes	Metastasis
T1: <3 cm. Contained within the lung	N1: Invasion of lung lymph nodes	M0: No metastasis
T2: 3-5 cm. Invasion of pleura or main airways	N2: Invasion of chest lymph nodes on the same side as the tumor	M1: Distant organ metastasis
T3: 5-7 cm. Spread into chest wall, pleura, or pericardium	N3: Invasion of lymph nodes on the opposite side of the tumor	
T4: >7 cm. Several lobes affected and spread into diaphragm, heart, trachea, spinal bone, or esophagus		

Additionally, tumors are graded according to the histological appearance of the tumor cells. In grade 1 malignancies, cells are similar to non-malignant cells, and the tumor appears to be well differentiated, with a predominant lepidic histology. Grade 2 tumors are moderately differentiated, with an acinar or papillary pattern. Grade 3 represents poorly differentiated cancers with micropapillary, solid, or complex glandular patterns[10]. A similar grading system has been established by some researchers to classify mouse lung tumors into 4 stages (**Figure 1**). In this case, grade 1 denotes atypical adenomatous hyperplasia (AAH) or small adenomas with uniform nuclei. Grade 2 are bigger adenomas that are characterized by enlarged nuclei with prominent nucleoli, whereas grade 3 stands for adenocarcinomas with an enhanced degree of cellular pleomorphism and nuclear atypia. Finally, grade 4 malignancies are invasive adenocarcinomas with all the features of grade 3 tumors but with an increased mitotic index. This includes irregular mitoses, highly invasive stromal reaction, known as desmoplasia, and invasive borders flanking vessels, or the pleura[11].

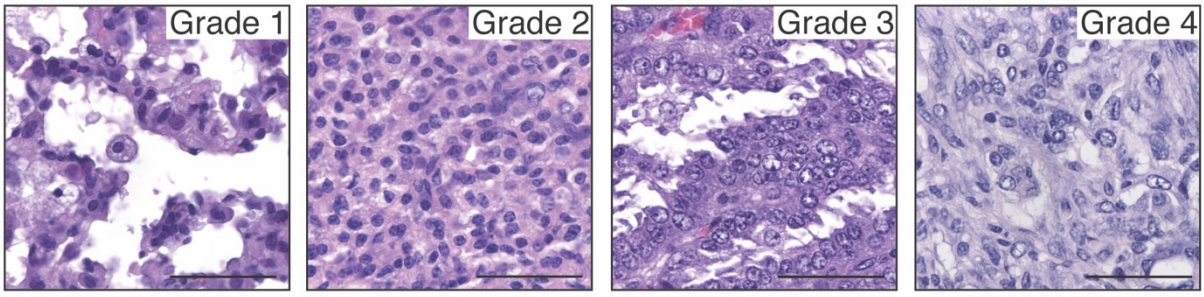

Figure 1. Mouse lung tumor grading system.
Grade 1 denotes hyperplasia with uniform nuclei. Grade 2 classifies adenomas with bigger nuclei and prominent nucleoli. Grade 3 comprises adenocarcinomas with cellular pleomorphism and nuclear atypia. Grade 4 includes all grade 3 tumors but with disorganized structure, desmoplasia, and invasive borders. Scale bar: 50 μm

The TNM and grading systems, when combined, allow for the classification of tumors into stages 1 to 4 (**Table 2**). Tumors that are small and have not spread to distant organs are categorized as stage 1. Stage 2 tumors can show different characteristics. In this group, the following malignancies are collected: tumors that are up to 5 cm and have lymph nodes near the affected lung; tumors between 5 and 7 cm with no lymph nodes affected; several tumors in one lobe; or no affected lymph nodes but the cancer has spread to the ribs, pleura, or muscle in the lung. Stage 3 comprises a wide range of tumors, but briefly, it generally describes tumors between 5 and 7 cm with or without invasion of lymph nodes and/or areas surrounding the lungs, such as ribs, pleura, muscle, peripheric nerves, or even the spinal bone. Stage 4 describes the most aggressive and advanced tumors that have already disseminated[12–14].

Table 2. Staging classification of lung cancer. Adapted from Collins, L. et al., American Family Physician, 2007[2].

Stages
Stage 1: 1-year survival 87.7% // 5-year survival 56.6% • IA (T1N0M0): <3 cm tumors contained within the lung – no affected lymph nodes • IB: (T2N0M0): 3-5 cm tumors – invasion of pleura or main airways

Stage 2: 1-year survival 73% // 5-year survival 34.1%
• IIA (T1N1M0): <3 cm tumors – invasion of lung lymph nodes
• IIB (T2N1M0 and T3N0M0):
  ○ 3-5 cm tumors – invasion of lung lymph nodes
  ○ 5-7 cm tumors with invasion of ribs, pleura, or pericardium but no affected lymph nodes

Stage 3: 1-year survival 48.7% // 5-year survival 12.6%
• IIIA (T1N2M0, T2N2M0, T3N1M0, and T3N2M0):
  ○ <5cm tumors – invasion of lymph nodes in the chest
  ○ 5-7 cm tumors spread into the lung cavity - lung or chest lymph nodes affected
• IIIB (T1-4N3M0 and T4N1-3M0):
  ○ all tumors with lymph nodes on the opposite lung affected
  ○ >7 cm tumors with lung or chest lymph nodes affected (same or opposite side as the tumors)

Stage 4 (T1-4N1-3M1): 1-year survival 19.3% // 5-year survival 2.9%. All tumors which have metastasized

1.2. Lung cancer subtypes

Lung cancers can be classified into two main histopathological groups: non-small cell lung cancer (NSCLC), which is found in 85% of lung cancer patients, and small-cell lung cancer (SCLC). SCLC is a very aggressive type of tumor that appears in heavy

smokers and has a 5-year survival rate of less than 5%. These tumors show neuroendocrine cell characteristics and are distinguished by a strong proliferative and metastatic potential[15].

According to their histological phenotype, NSCLCs can be classified into three different subtypes: large cell lung cancer (LCLC), squamous cell lung carcinoma (SCC) and lung adenocarcinoma (LUAD). The diagnosis of large cell carcinomas, the least frequent type and responsible for around 10% of NSCLC, depends on the elimination of other variables such as glandular or squamous differentiation[16]. 20% of NSCLC are SCC, and the most frequent histological markers of SCC are P63 and cytokeratin 5 (CK5)[17]. Most cases of SCC are strongly associated with smoking habits and display a high mutational burden. Although frequent genomic alterations in *TP53, FGFR1, CDKN2A, PTEN, PIK3CA, KEAP1, KMT2D,* and *NFE2L2* have been observed in these tumors, patients with SCC do not benefit from targeted therapies yet[18,19].

LUAD is the most frequent subtype of NSCLC (70-80%). The most common histological markers for LUAD diagnosis are thyroid transcription factor 1 (TTF-1), also known as NK2 Homeobox 1 (NKX2-1), cytokeratin 7, and mucin[17]. In the last decade, the development of new sequencing technologies has facilitated the identification of new mutations in driver oncogenes or tumor suppressors, which have emerged as new therapeutic targets. The most frequently found driver is Kirsten Rat Sarcoma viral oncogene (*KRAS*), which belongs to the *RAS* family. These proteins are involved in signal transduction of pathways that promote growth and survival. Point mutations in *KRAS* prevent the signaling from being revoked, resulting in constitutive signaling[20]. $KRAS^{mutant}$ tumors appear mainly in smoker patients, whereas never-smokers are associated with mutations in the Epidermal growth factor receptor (*EGFR)* or the fusion between Echinoderm microtubule-associated protein-like 4 (*EML4*) and anaplastic lymphoma kinase (*ALK*). In 1-2% of NSCLCs, ROS1 has been found to be fused to different partners in young patients[21]. Additionally, several tumor suppressors have been identified to be inactivated in LUAD, and the most prevalent one is TP53.

Histological categorization has traditionally been used to determine the course of treatment for lung cancer. Yet, given the molecular heterogeneity of lung cancer, considering genetic mutations has become crucial while making therapeutic decisions[22]. Tumors driven by EGFR mutations or EML4-ALK are currently treated with specific

targeted tyrosine kinase inhibitors (TKIs), such as afatinib, erlotinib, osimertinib, or genitinib for EGFR positive tumors, and crizotinib, alectinib, brigatinib and lorlatinib for EML4-ALK tumors[23]. KRAS has been long regarded as an undruggable oncogene because of its high affinity for GTP, few active binding sites and complex downstream signaling pathways. However, several small inhibitory molecules that are able to bind $KRAS^{G12C}$ mutation have been developed in recent years, leading to the approval of sotorasib as a treatment for $KRAS^{G12C}$ positive patients by the Food and Drug Administration (FDA)[24].

2. ALK gene

The Anaplastic Lymphoma Kinase (*ALK*) gene is located in the short arm of chromosome 2 (2p23) and it is a membrane-bound tyrosine kinase which belongs to the insulin receptor kinase superfamily. It is composed of an extracellular region, a single transmembrane-spanning domain, and a cytoplasmic catalytic domain. The extracellular region is formed by a Low-density lipoprotein receptor domain class A (LDLa) domain, which is flanked by two Meprin, A5 protein and protein tyrosine phosphatase Mu (MAM) domains, forming the ligand-binding domain (**Figure 2**)[25,26]. ALK is expressed in the central nervous system (CNS) during neonatal development and at lower levels in the CNS of adult organisms[25], and it was first discovered in anaplastic large-cell lymphoma[27]. Activation of ALK needs dimerization and autophosphorylation after the binding of its ligand, FAM150[28].

The *ALK* gene can be involved in tumorigenesis in three different ways: amplification, point mutations, or rearrangement with another gene. The rearrangement is the most common one, and more than 90 different *ALK* fusion partners have been identified, which lead to different tumor types such as lymphomas, NSCLC, breast cancer, and colorectal cancer[29]. Fusion oncogenes in which *ALK* is involved typically only retain its cytoplasmic catalytic domain, and it is the N-terminal fusion partner the one driving aberrant expression and constitutive activation of ALK[30]. Some of the most common fusion oncogenes involving ALK are Nucleophosmin 1-Anaplastic Lymphoma Kinase (*NPM1-ALK*) and *EML4-ALK*. *EML4* appears fused to ALK in 29-33% of ALK positive patients[31,32].

Figure 2. ALK protein.
Schematic of the structure of the ALK protein. The extracellular region is formed by a Low-density lipoprotein receptor domain class A (LDLa) domain which is flanked by two Meprin, A5 protein and protein tyrosine phosphatase Mu (MAM) domains. It also contains a transmembrane domain (TM) and a tyrosine kinase domain (TK), which is the intracellular part of ALK.

3. EML4 gene

The Echinoderm Microtubule-associated protein Like 4 (*EML4*) gene is located in chromosome 2p21. It belongs to the EML protein family which is part of the microtubule-associated protein (MAPs) family, and affects microtubule dynamics. Several human EML proteins have been identified. The *EML4* gene is composed of an N-terminal coiled-coil trimerization domain (TD)[33] followed by a C-terminal tandem atypical beta-propeller domain (TAPE). This TAPE domain contains a hydrophobic EML protein domain (HELP), which facilitates tubulin binding, and nine tryptophan-aspartic acid (WD) repeats that allow protein-protein interactions[26,34]. The N- and C-terminal domains are separated by a basic linker region. EML4 has been described to bind microtubules through its TD and basic region[33] and it is expressed in multiple tissues, including the lung (**Figure 3**).

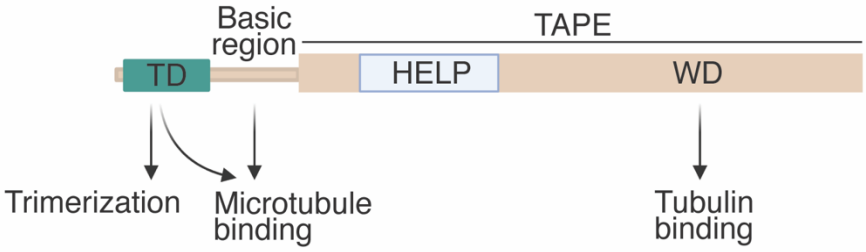

Figure 3. EML4 protein.
Schematic of the structure of the EML4 protein, formed by a trimerization domain (TD), a basic region and a TAPE domain. The TAPE domain is composed of a hydrophobic EML protein domain (HELP) and several tryptophan-aspartic acid (WD) repeats.

4. EML4-ALK

4.1. Lung cancer

The *EML4-ALK* oncogene is found in 3-7% of NSCLCs, appearing mainly in young, never-smoker patients[35] (**Figure 4A**). It is a fusion gene that happens after an inversion

of chromosome 2 (Inv(2))(p21p23)), where both genes, *EML4* and *ALK*, are located (**Figure 4B**). This rearrangement was first discovered by Soda M. et al. in 2007[36], and it has been associated with radiation exposure in some cancer types[37]. The *EML4-ALK* rearrangement leads to the internalization and constitutive activation of ALK in the lung without the need for any ligand. After the rearrangement, the tyrosine kinase domain of ALK is constantly autophosphorylated, which activates different downstream signaling pathways such as PI3K/AKT or RAS/MAPK, leading to uncontrolled proliferation and cell survival[38] (**Figure 4C**).

The diagnosis of ALK positive lung cancer was initially achieved by fluorescence in situ hybridization (FISH) and immunohistochemistry (IHC). FISH is one of the most reliable techniques to confirm the presence of EML4-ALK, but signal instability and the technical approach can sometimes make this a complicated method for diagnosis. IHC is an easy diagnostic tool; however, its specificity tends to be poor and does not help with the identification of the fusion variant. The appearance of novel techniques, such as next-generation sequencing, has also been implemented as a diagnostic tool for ALK-tumors, being the only one providing information on the EML4-ALK fusion variants[39].

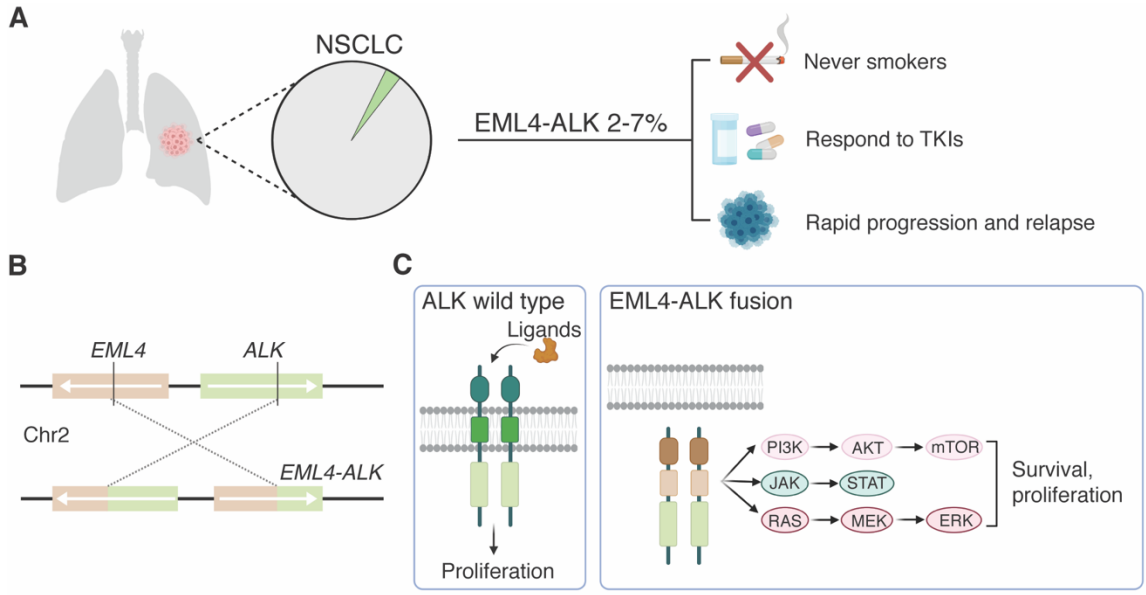

Figure 4. *EML4-ALK* induced lung adenocarcinoma.
(**A**) The EML4-ALK rearrangement appears in 2-7% of NSCLC patients, which are mainly young and never smokers. These tumors respond to tyrosine kinase inhibitors (TKIs), but patients ultimately relapse. (**B**) Schematic of the inversion that leads to the *Eml4-Alk* rearrangement. Arrows indicate the orientation of the gene expression. (**C**) Representation of the signaling cascade of EML4-ALK. Wild type ALK (left) is a transmembrane receptor which needs to be dimerized and activated by certain ligands to induce a signaling cascade resulting in cell proliferation and survival. When *ALK* is fused to *EML4* (right) the intracellular tyrosine kinase domain of ALK is internalized and constitutively active under the promoter of *EML4*, leading to the constant activation of downstream signaling pathways such as PI3K/AKT, RAS or JAK/STAT, ultimately resulting in cell survival and proliferation.

4.2. Treatments options for ALK+ lung cancer

The discovery of this oncogene led to the use of ALK TKIs as first line therapy for these patients and the design of other novel ALK TKIs. The first inhibitor that was used as a treatment for ALK positive tumors was crizotinib, which was initially developed as a c-MET inhibitor[40]. One of the main drawbacks of crizotinib is its low penetrance in the CNS, since the brain is the primary site of metastasis and therapy failure. A few years after the discovery of this fusion gene, next-generation TKIs were developed, most of them being able to penetrate the blood-brain barrier and efficiently working against resistant mutations[41–43]. The current first-line treatment for ALK positive patients is next-generation TKIs such as brigatinib, alectinib, or lorlatinib, although crizotinib is still being used. The development of all these treatments has improved the overall survival of these patients to 80 months[44], thus extensive research is needed to improve patient survival.

Despite the effectiveness and specificity of these inhibitors, most EML4-ALK patients develop resistance to targeted therapy. Several mechanisms have been described and among them, gatekeeper mutations in the ALK kinase domain are the best studied. These mutations appear in 30-40% of resistant tumors, they modify the structure of the protein and interfere with the binding of the drug. The most common mutations in the ALK gene are L1196M and G1269A, involving the binding pocket of ATP. Other mutations are: C1156Y, G1202R, I1171T/N/S, S1206C/, and F1174V[45]. Second and third-line therapies have been proven to be effective for specific mutations, although resistance mutations still arise after sequential therapies. An interesting secondary resistant ALK mutation is G1202R, since it can only be inhibited by lorlatinib, which was recently approved as a first-line therapy[46]. Other known resistance mechanisms are copy number gains of the *ALK* gene, and activation of bypass signaling pathways such as EGFR, KRAS and cMET pathways[47]. Differentiation from adenocarcinoma to SCC or SCLC have also been described as resistance mechanisms in a few ALK patients[48–54].

As previously mention, ALK positive patients benefit form TKI treatments, which frequently allow to control the disease for a period of a few years. Unfortunately, most of these patients develop relapse and succumb to death. Important advancements have been made in recent years in the development of immunotherapeutic treatments, which

seem to be promising for certain types of lung cancers. Little is known about the immune landscape of EML4-ALK tumors, although it has been shown that they display an immunologically cold tumor microenvironment (TME)[55]. This finding goes in line with results in clinical trials in which generally ALK positive patients showed no benefit from immunotherapeutic treatments, alone or in combination with TKIs[56–61]. Vaccines are another promising treatment option for cancer patients and an ALK-directed vaccine could help to revert the immune cold TME of these patients to an immune-responsive phenotype[62]. In 2015 the first vaccine targeting *ALK* was developed and appeared to be effective in reducing tumor growth in EML4-ALK mouse models and extending mouse survival[63].

4.3. Fusion variants

Up to 16 different *EML4-ALK* variants have been described, and the most frequent ones are listed in **Figure 5**. All of the known variants contain the complete intracellular *ALK* tyrosine kinase domain, encoded by exons 20 to 29, and differ on the breakpoints of *EML4*. Among these variants, long variants, such as variant 1 (V1) and variant 2 (V2) contain the TD domain and a portion of the TAPE domain, whereas the short variants, apart from the TD, contain only a few residues of the TAPE domain (variant 3, V3) or lack the domain completely (variant 5, V5)[64]. V1 and V3 are the most common variants, occurring in 43% and 40% of the patients, respectively[65]. V2 is found in 10% of the patients and V5 in 2% of the patients (**Figure 5**). V1-bearing patients have been described to be significantly younger than V3 patients[66].

The importance of considering the EML4-ALK fusion variant driving the tumor to administer the proper therapy is becoming more evident over the years. Recent studies suggest the EML4-ALK variant as an important factor for therapy after showing a better survival and prognosis of V1-bearing patients[67,68]. Nevertheless, some controversy is found in the literature on which of the variants is more aggressive. In a retrospective study of stage IV EML4-ALK patients, it was observed that patients with V3-driven tumors presented more metastatic sites at diagnosis than V1 and V2 patients[69]. Lin et al. found that V3 tumors were more prone to develop ALK resistance mutations (57% vs 30% in V1 tumors), whereas under lorlatinib treatment, V3 patients showed a longer progression-free survival than V1 patients[65]. The molecular reasons for this difference between variants are unknow, but it has been suggested that the length of the proteins

might infer on the therapy response, since shorter variants seem to be more stable[70,71]. As previously mentioned, *EML4-ALK* induces mainly LUAD, although a few SCC driven by this oncogene have been identified[72,73]. *EML4-ALK* SCC have been described as relapses after histological transformation due to treatment[54] or in treatment-naïve tumors associated to short variants of the fusion gene, such as V3 or V5[74–76].

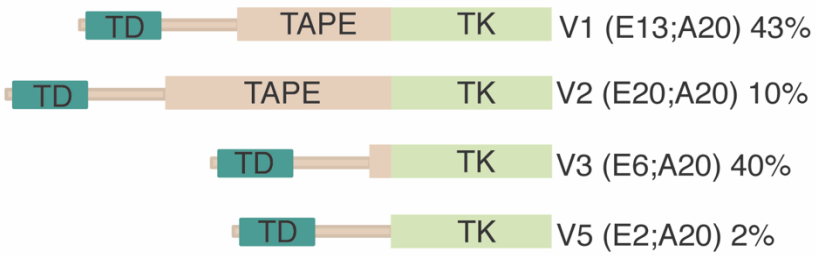

Figure 5. Most frequent *EML4-ALK* fusion variants.
Adapted from Bayliss, R. et al., Cell. Mol. Life Sci. 2016[26]. All variants contain the entire tyrosine kinase domain (TK) of *ALK* and differ on the breakpoint of *EML4*. Long variants (eg. V1 and V2) contain the trimerization domain (TD) and a partial tandem atypical propeller domain (TAPE) of *EML4*. Shorter variants lack the TAPE domain (V5) or only have a few TAPE domain residues (V3). In brackets, chromosome breaking points in *EML4* (E) and *ALK* (A) genes of the respective variants. The percentage represents the frequency of each variant.

4.3. Localization of the variants

The fusion partner of ALK determines the location of the protein. EML4 is known to stabilize microtubules and to be localized to the mitotic spindle through the TD and the basic region[77]. The localization of the EML4-ALK protein has been a controversial topic since this fusion gene was discovered. Richards M. et al. showed in 2015 that the different structures of V1 and V3 caused a difference in localization between the two variants, with V1 showing diffused cytoplasmic localization and V3 binding to microtubules, similar to the original EML4 protein[33]. They suggested that the presence of a partial TAPE domain in V1 might avoid microtubule binding, possibly through misfolding of the protein. Additionally, this localization of V3 was shown to induce microtubule stabilization, which appears to be dependent on the upregulation of Nek9 and Nek7, promoting cell migration[78]. On the other hand, more recent studies claim that EML4-ALK aggregates into granules, condensates, or foci that localize to the cytoplasm both in V1 and V3[79–81]. Additionally, these cytoplasmic forms have been described to locate the activation of downstream pathways[79,81]. Finally, in 2022, Lucken K. et al. showed that V3 localizes to both interphase microtubules and the mitotic spindle and

that the presence of V3 in microtubules increases the number of mitotic errors, increasing genomic instability[82].

5. Lung epithelial cell types

5.1. Lung cell types in homeostasis

The human lung is one of the most complex tissues since it is composed of multiple cell types that are extremely plastic. The adult lung is formed by four different compartments, including the trachea, bronchi, bronchioles, and alveoli (**Figure 6**). The alveoli are air sacs where oxygen exchange occurs. They are mainly formed by alveolar type 2 (AT2) cells and alveolar type 1 (AT1s) cells. AT1 are thin cells that cover 95% of the surface and are involved in the gas exchange. AT1 cells are characterized by the expression of marker genes such as the Homeodomain-Only Protein (*HOPX)* or Podoplanin (*PDPN*). AT2 cells are cuboidal and secrete pulmonary surfactant from their lamellar bodies, reducing alveolar surface tension. AT2 cells are marked by the expression of surfactant protein (SFTP) genes such as *SFTPA*, *SFTPB* and *SFTPC* (or SPC protein)[83].

The human airways, formed by the trachea, bronchi and bronchioles, consist mainly of club, ciliated, basal, goblet, neuroendocrine, and tuft cells. Club cells have been given credit for a number of protective functions such as regeneration of injured airways, roles in immune-modulation and toxin clearance[84], and they are characterized by the production of club cell secretory protein (CCSP), which is encoded by the Secretoglobin Family 1A Member 1 (*SCGB1A1*) gene. It is well known that club cells have the potential to self-renew and also differentiate into ciliated cells[85], but they can also de-differentiate to regenerate the basal cell population[86]. Ciliated cells contain cilia which transport mucus through the airways to maintain the pulmonary homeostasis and are known for the expression of Forkhead Box J1 (*FOXJ1*). Basal cells are defined by the expression of tumor protein 63 (TP63) and cytokeratin 5 and 14 (KRT5/14) and are known as the stem cells of the airways since they can self-renew and differentiate into most of the airway cells[87]. Goblet cells secrete mucins which are necessary for the protection of the lung from external insults and are marked by the expression of mucin 5 A and B (MUC5A and MUC5B)[88]. Pulmonary neuroendocrine cells (PNEC) are a rare epithelial cell type which is known to have neuronal and endocrine properties. They

have been proposed to play a role in airway oxygen sensing and lung response to signals such as allergens[89], and they are the cells of origin of SCLC[90,91]. Additionally, several studies suggest that PNECs are able to regenerate the airway epithelium by giving rise to club and ciliated cells after damage induced by naphthalene[92,93]. Tuft cells, also known as brush cells, form another rare epithelial cell type which have sensory and secretary characteristics. In the lung, these cells are primarily localized to the trachea[94].

For many years it was believed that these few cells were the only epithelial cell types forming the lung, but newly developed single cell technologies have opened a Pandora box on the lung research field enabling the discovery of many new cell types, linage hierarchies and cell type functions. A recently described cell type using these techniques are Ionocytes. These cells regulate airway surface physiology and are marked by the expression of *FOXI1* and *CFTR*[95].

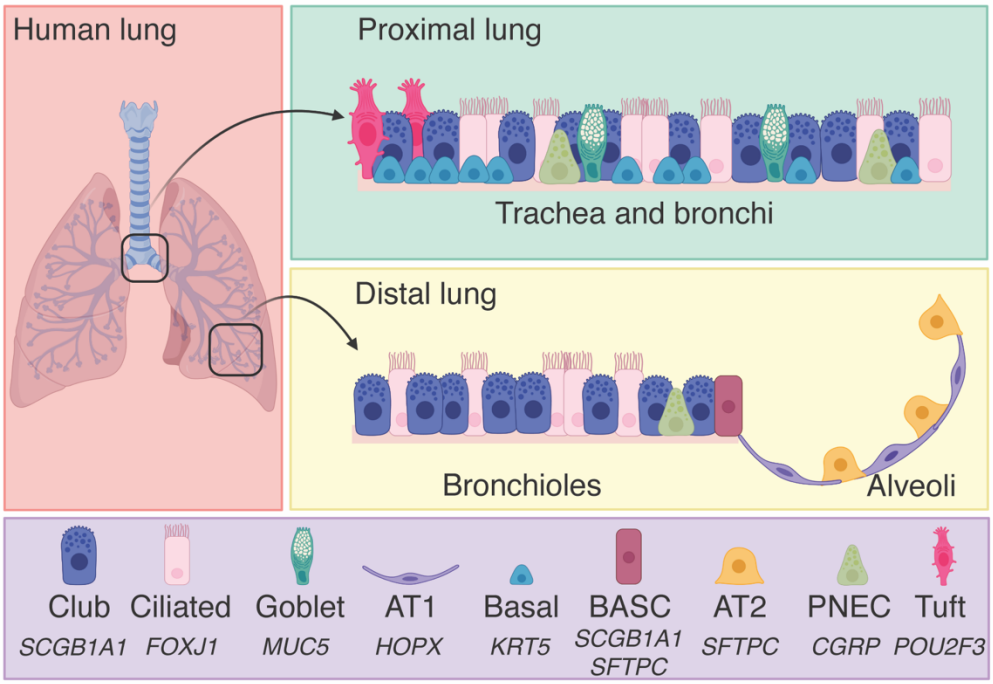

Figure 6. Schematic of the human lung epithelial cell types.
The lung is divided into proximal and distal lung. The trachea and bronchi (proximal) are composed of club cells, ciliated cells, basal cells, goblet cells, pulmonary neuroendocrine cells (PNEC) and Tuft cells. The distal bronchioles are formed by club cells, ciliated cells, PNEC and bronchioalveolar stem cells (BASCs). The alveoli contain alveolar type I (AT1) and type II (AT2) cells. The main marker gene for each cell type is indicated below the cell type name.

In the mouse, the airways open into the alveoli through the bronchioalveolar duct junction (BADJ) where bronchioalveolar stem cells (BASCs) are located. These cells express both CCSP and SPC and play a role in the regeneration of both the club and the

AT2 populations[96]. A few recent studies have been able to track BASCs using dual recombinases and have identified BASCs to also give rise to ciliated cells and AT1 cells. Seemingly, BASCs are capable of sensing which particular cell type needs to be replenished and therefore they respond differently to distinct lung injuries[97,98].

One on the main current focus in lung research is building up the human lung cell atlas. In 2020, Travalgini et al. identified gene expression profiles and locations of 15 different epithelial cell types among the 58 lung cell types described[99]. Additionally, they described 17 cell types that appear to be different between the mouse and human lung. Additionally, a recent study presented a second human lung cell atlas which included an extent number of human samples of healthy and diseased lungs[100].

5.2. Lung cell types after injury

Acute lung injury (ALI), COPD or idiopathic pulmonary fibrosis (IPF) are some of the most common diseases in which the lung tissue appears to be damaged. These diseases have been extensively reproduced in mouse models to understand the lung regeneration process undergoing during these pathologies, which has allowed to identify several new intermediate or transitional cell states.

In the alveolar region, both AT1 and AT2 cells heal the injured alveoli by self-renewal and differentiation into each other[101,102]. In addition, AT2 alveolar epithelial progenitors (AEP) which respond to Wnt have been reported to be a primary progenitor cell in the alveolar area[103]. Three recent *in vivo* studies have shown the existence of an intermediate state in the transdifferentiation of AT2s to AT1s after lung injury (**Figure 7A**). In these studies, three different nomenclatures have been used to describe this state and slight differences have been shown between them. Strunz et al. described the Krt8^{+} alveolar differentiation intermediate (ADI) state[104], which are a transitional state from both Major Histocompatibility Complex (MHC) class-II club cells (previously described as H2-K1^{high} club cells[105]) and alveolar stem cells towards AT1 regeneration. Krt8+ ADI cells express genes such as *Krt8*, *Cldn4*, *Cdnkn1a* and *Plaur* and are enriched in TGFβ and p53 signaling pathways and DNA damage response. Additionally, these cells were described to generate a network of cell-to-cell communication with stromal cells during lung regeneration. At the same time, Kobayashi et al. described a very similar cell state, appearing from AT2 to AT1 cells which was named as pre-alveolar type-1

transitional cell state (PATS)[106]. PATS also showed enrichment in TP53, TGFβ signaling, DNA-damage-response and cellular senescence and presented high expression of the Krt8 ADI markers among others. A third study published also in 2020 by Choi et al. named this cellular state as damage-associated transient progenitors (DATPs)[107]. In addition to the data on the previous two studies, DATPs appeared to be induced by IL-1β released by interstitial macrophages. They described that IL-1β is sensed by Il1r1-expressing AT2 cells which differentiate into DATP via HIF1α-mediated glycolysis. Moreover, they demonstrated that IL-1β-mediated chronic inflammation impairs AT1 differentiation, inducing the accumulation of DATPs. Moreover, it has recently been found that human AT2 cells are capable of transitioning to metaplastic Krt5$^+$ basal cells through a state which mirrors mouse Krt8$^+$ intermediate state. These findings suggest that the AT2-to-AT1 transitional state described in mice might not be restricted to the alveolar linage and that it might also be involved in the basal cell regeneration in humans[108].

Lineage-negative epithelial progenitors (LNEPs) have also been described to play a role in lung regeneration. The p63$^+$ LNEPs and the p63$^-$ LNEPs are the two subtypes that make up LNEPs. After severe damage, such as bleomycin or influenza damage, hypoxia and Notch signaling promote alveolar epithelial cells to transdifferentiate into basal-like cells co-expressing p63 and Krt5, thus, the following Notch blockage induces differentiation into AT2 and regeneration of the alveolar space[109,110]. Some groups have named these cells as distal airway stem cells (DASC) and set their location to small bronchioles (**Figure 7B**)[111,112]. DASCs have been described to be able to differentiate into AT1, AT2 and airway secretory cells upon transplantation to damaged lungs and have been located also in human lung tissues with acute lung injury[113]. The p63$^-$ LNEPs have been shown to be a subpopulation of club cells, known as H2-K1^{high} club-like lineage-negative epithelial progenitors, which expand and convert into alveolar linages after lung injury[105]. MHC I and II genes implicated in the immune response to infections, are highly expressed in these cells. They are remarkably present after influenza injury and exhibit a high number of genes that enable viral replication, which raises the intriguing possibility that the progenitor response depends on the type of injury.

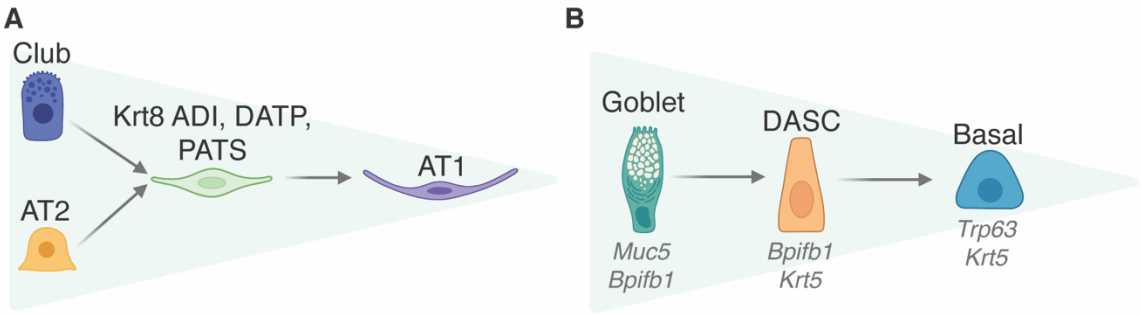

Figure 7. Regeneration-related transitional states.
(**A**) Schematic representing the transition from club or alveolar type II (AT2) cells to alveolar type I (AT1) cells through a transitional state known as Krt8+ alveolar differentiation intermediate (ADI) state, damage-associated transient progenitors (DATP) or pre-alveolar type-1 transitional cell state (PATS). (**B**) Schematic of the renewal of basal cells from goblet cells upon injury. Goblet cells undergo a transitional state known as distal airway stem cells (DASC) which expresses markers of both goblet and basal cells, before giving rise to completely differentiated basal cells.

6. Mouse models of NSCLC cancer and linage tracing

Cancer mouse models are a very valuable tool not only for reproducing human tumors and comprehending their biology but also for identifying new biomarkers for early diagnosis and studying the effects of different drug treatments. Multiple mouse models have been developed for lung cancer research (**Table 3**), ranging from the use of carcinogens to transgenic mice in which genetic abnormalities are introduced.

6.1. Transgenic mouse models

Traditional transgenic or knockout mice do not accurately mimic the development of human cancers. The generation of conditional transgenic mice has allowed researchers to better recapitulate the sporadic emergence of lung cancer. In these mice, the mutation is induced at a particular time and can even be generated only in specific cells, allowing the formation of tumors in a specific and controlled manner. The Cre/loxP system is one of the best-established technologies for conditional transgenic mice[114]. To delete tumor suppressor genes, certain exons or the entire gene are flanked by loxP sites. Expression of the Cre recombinase in certain cells or tissues will then cut those sites and enable the deletion of the desired gene. Oncogene activation can also be achieved by using Cre expression to remove a stop codon which prevents the expression of the oncogene[20]. Expression of Cre can be either constant or inducible, as in Cre Recombinase-modified Estrogen Receptor fusion protein (CreER) mouse models[115].

In these last mouse models, the Cre recombinase expression is activated after tamoxifen administration, which allows to control the specific time of tumor induction.

Another well-known system to conditionally express transgenic alleles in mice is using tetracycline-responsive regulatory expression elements. This system consists of the Tet repressor protein (TetR) and tet operator (tetO) DNA elements that control the tetracycline resistance operon of *Escherichia coli*, which fused to the VP16 transcription factor, form the tetracycline-controlled reverse transactivator (rtTA)[116]. Doxycycline administration or withdrawal switches on or off this system and therefore, the expression of the transgene.

The Lox-Stop-Lox (LSL) system together with the Cre system has extensively been used to model conditional *Kras* mutations in the mouse lung[117,118]. Usually, tumors are induced in these mice with the help of adenoviruses or lentiviruses containing Cre which are intratracheally or intranasally instilled into the mice. The Cre recombinase in these adenoviruses is typically controlled under the ubiquitous cytomegalovirus (CMV) promoter[118], or cell type specific promoters[119–121]. In addition, *Kras*-driven lung tumors are typically studied together with the concomitant loss of TP53, using the well-known $Kras^{G12D}$; $p53^{fl/fl}$ (KP) mouse model[122], which results in more aggressive tumors.

6.2. Linage-restricted mouse models

Both the doxycycline and the Cre system can be driven by cell type specific promoters, restricting the expression of the transgene to the desired cell type. Among the doxycycline system, mouse models targeting club (CCSP-rtTa) or AT2 (SPC-rtTA) cells have been developed[123]. The CCSP-rtTA model has been used to understand $EGFR^{mutant}$ or $Kras^{mutant}$-driven lung cancer[124,125].

On the other hand, using the Cre system, mouse models targeting the main epithelial cell types in the lung have been generated. Rawlins et al. generated the *Scb1a1*-CreER mouse model to understand the role of club cells in homeostasis and lung repair[85], and when crossed with LSL-eYFP mice, it was shown to label club cells and a few AT2 cells. They also generated the *Foxj1*-CreER mouse model to label ciliated cells[126]. To specifically mark AT2 cells, Chapman et al. developed the *Sftpc*-CreER[127] mouse line, whereas Jain et al. created the *Hopx-CreER* line to label AT1 cells[101]. To distinguish basal cells the *Krt5*-CreER mouse line from Indra et al. has been shown to be a valuable

model[128]. In more recent years, with the discovery of new epithelial cell types and transitional cell states, new mouse models have been developed using new markers genes, or existing mouse lines have been adapted to the lung research field. All these mouse models when combined with fluorescence reporters, they are a very useful tool for linage tracing experiments in lung homeostasis, regeneration or cancer studies.

In our group, we performed a thorough characterization of five linage tracing mouse models: *Scgb1a1*-CreER, *Sftpc*-CreER, *Krt5*-CreER, *Hopx*-CreER and *Foxj1*-CreER[129]. Unfortunately, and unexpectedly, upon tamoxifen administration, *Scgb1a1*-CreER, *Krt5*-CreER and *Hopx*-CreER appeared to be promiscuous and did not exclusively label the cell type for which they were generated. For instance, the *Scgb1a1*-CreER not only labels club cells, but also ciliated cells and a few AT2 cells. The *Hopx*-CreER mouse model labels club cells, in addition to the already known AT1 cells. Finally, the *Krt5*-CreER lines labels basal cells in the upper airways and AT1 cells in the alveoli.

Table 3. Mouse models of linage tracing and lung adenocarcinoma.

Mouse model	Inducer	Target cell type	Reference
***Scgb1a1*-CreER**	Tamoxifen	Club, AT2 and ciliated	Rawlins et al. 2009[85]
***Sftpc*-CreER**	Tamoxifen	AT2	Chapman et al. 2011[127]
***Hopx*-creER**	Tamoxifen	AT1 and club	Jain et al. 2015[101]
***Krt5*-CreER**	Tamoxifen	Basal and AT1	Indra et al. 1999[128]
***Foxj1*-CreER**	Tamoxifen	Ciliated	Rawlins et al. 2007[126]
LSL- *Kras*^{G12D}	Adeno-Cre	Lung epithelial cells	Jackson et al. 2001[118]
CCSP-rtTA	Doxycycline	Club and AT2	Perl et al. 2002[123]
SPC-rtTA	Doxycycline	AT2	Perl et al. 2002[123]
SPC-*EML4-ALK*	Not inducible	AT2	Soda et al. 2008[130]
LSL-*EML4-ALK*; *Sfptc*-CreER	Tamoxifen	AT2	Pyo et al. 2017[131]
Lenti-*Eml4-Alk*	Lentiviral instillation	Lung epithelial cells	Blasco et al. 2014[132]
Adeno-*Eml4-Alk*	Adenoviral instillation	Lung epithelial cells	Maddalo et al. 2014[133]

6.3. EML4-ALK mouse models

After the discovery of the EML4-ALK fusion, a few transgenic mouse models reproducing this oncogene have been developed (**Table 3**). The first model was generated by Soda et al. in 2008[130] and consisted on the cDNA of EML4-ALK v1 ligated with the SPC promoter to induce the overexpression of the fusion protein specifically in AT2 cells. Ten years later, an inducible model was generated by using the LSL system to

express the EML4-ALK V1[131]. LSL-EML4-ALK mice were crossed with *Sftpc*-CreER mice, and after tamoxifen administration, these mice developed lung adenocarcinomas within 1 week of induction exclusively in the lungs. In 2014, two mouse models were developed in two different studies, which generated the endogenous *Eml4-Alk* V1 rearrangement in mouse lungs, better representing the human scenario. Both of these studies made use of viral systems to deliver the CRISPR/Cas9 technology, one of them employing lentiviruses[132] and the other one using adenoviruses[133], and manage to detect lung adenocarcinomas starting at 6 weeks post-infection. It is clear that along the years, despite several variants being discovered, only V1 has been modeled in vivo, limiting *Eml4-Alk* variant related research to in vitro studies.

7. Cell of origin of lung adenocarcinoma

The cell of origin of several types of tumors has been demonstrated to impact patient outcomes and to serve as a key player in biomarker development[134]. NSCLC subtypes can develop from different regions of the lung as well as from various cells of origin, but the severe plasticity of lung epithelial cells makes the identification of the correct origin a challenging task. The first authors studying the cell of origin of LUAD assumed that the expression of specific linage markers defined the origin of these tumors[118,135]. Therefore, Johnson et al. claimed that only AT2 cells gave rise to LUAD based on the expression of SPC[135], whereas Jackson et al. suggested also a club origin due to the double positive SPC/CCSP staining of their tumors. A few years later, BASCs were described in normal mouse lungs[96] and were also proposed as cells of origin of LUAD after the observation of double-positive (SPC^+ and $CCSP^+$) cells expanding after $Kras^{mutant}$ induction.

Thanks to the development of genetically engineered mouse models (GEMMs), and new technologies such as single cell RNA sequencing (scRNA-seq), studying the cell of origin of LUAD is now a hot topic in the field. Either via the use of knock-in techniques (*Sftpc-CreER* or *Scgb1a1-CreER*)[85,136] or through the infection of lung cells with adenoviruses that drive Cre expression by cell type-specific promoters (Adeno-SPC-Cre or Adeno-CC10-Cre)[119,137], researchers have been able to induce the expression of mutant *Kras* in different cell types. Since then, multiple studies have tried to conclude which cells originate LUAD in mice and whether tumors from different origins display distinct

phenotypes and features[117,119], however, a clear and concise agreement has yet not been reached. The statement that BASCs were the cell of origin of LUAD was dismissed in later reports due to the lack of tumors appearing in the BADJ region, and the finding that only club and AT2 cells generate LUAD by inducing *Kras* mutation[136,138]. Numerous studies have suggested these two origins for LUAD, although the AT2 origin has had a greater acceptance for many years since the vast majority of tumors exhibit expression of the AT2 cell marker SPC. In addition, AT2 cells have also been shown to develop tumors driven by other oncogenes but *Kras*, such as mutant *EGFR*[124]. In addition, there is also a debate over the aggressiveness of club-originated tumors and whether club cells can develop adenocarcinomas rather than just hyperplasia or benign adenomas[117,119,136]. This assertion is now evolving and the idea that club cells might develop tumors as efficiently as AT2 cells is rapidly emerging.

In the last years, a number of studies have proven club cells to be one of the cells of origin of LUAD. In 2019, Spella et al. described that tumors induced by toxic chemicals contained in tobacco smoke arise from airway epithelial cells[139]. In line with these findings, Rosigkeit et al. described that club cells expressing $Kras^{G12D}$ and the simultaneous loss of TP53 give rise to LUAD through a progressive loss of CCSP and acquisition of SPC[140]. A few recent reports have shown that genomic alterations can distinctly contribute to the phenotype of tumors depending on their origin. In 2017, it was demonstrated that *Lkb1* loss together with $Kras^{G12D}$ mutation cause a shorter survival in mice when driving club-originated tumors compared to AT2-originated tumors[137], and Concepcion et al. found that *Smarca4* loss sensitized club cells to tumor progression[141]. Nevertheless, the biology underlying the developmental process of club tumors is not yet properly understood. In our group, we have recently proven that in addition to AT2 cells, club cells can also give rise to *Eml4-Alk* driven LUAD. We showed that club tumors despite retaining the DNA methylation pattern of their originating cell type, they undergo an epigenetic switch by which they repress the expression of certain club-specific markers and acquire an AT2-like phenotype[129], unveiling the reason for the expression of SPC in the vast majority of mouse LUADs.

Until recently, AT1 cells had barely been considered to have the capacity of forming lung tumors, since they are generally thought to be terminally differentiated and have low proliferative capabilities[142,143]. However, Minxiao et al. have recently

described that AT1 cells can also be the cell of origin of *Kras*^{G12D} tumors[144]. By inducing the *Kras* mutation in *Gramd2*-CreER mice, in which AT1 cells are labeled, they could observe that these tumors presented a distinct histopathological phenotype to the one seen in AT2 *Kras*-driven tumors and found different pathway alterations within the two origins.

The TME has extensively been shown to have an essential role in tumor development, by generating a niche for tumor cells[145]. It is known that the immune compartment of NSCLC can be very heterogenous and vary to a great extent depending on the subtype and driver mutations[137,146,147]. With this knowledge, one would hypothesize that the cell of origin most probably also influences the TME. Little has been discovered on this topic, but it has been shown that no matter the genetic alterations co-occurring with *Kras*, AT2-derived malignancies present increased immune infiltration over club tumors, which make them dependent on a pro-tumorigenic immune landscape to progress[137,147,148]. Nevertheless, further studies are required to fully unveil the complex immune landscape of LUAD with different cells of origin to favor patient benefit and the development of effective immunotherapies. To summarize, a better understanding of the development of LUAD from different cells of origin will enable the discovery of new biomarkers and therapeutic strategies which will result in better patient outcomes.

8. TP53

8.1. Role of TP53

TP53 is a tumor suppressor gene located in chromosome 17 in humans. It is known to be the major guardian of the cell, providing protection from genomic alterations related to cancer through a complex network of signaling pathways involved in regulation of the cell cycle, control of genomic stability, apoptosis, and immune response[149]. The TP53 protein binds the DNA as a tetramer and is formed by several well-known functional domain. It contains two tandem transcription activation domains (TADs) in the amino-terminal region which allow the protein to accurately induce target gene expression upon DNA damage [150]. After, the TADs, TP53 contains a proline-rich domain and the DNA-binding core domain, where most of the cancer-related mutations can be found[151].

In normal conditions, TP53 levels are determined by the degree at which it is degraded, and are maintained low throw the MDM2-TP53 binding, which drives its ubiquitination and degradation through the proteasome[152]. The regulation of TP53 through MDM2 works as a feedback loop, where the *MDM2* gene is a target of TP53. Therefore, high levels of TP53 enhance the production of MDM2, which triggers the degradation of TP53, lowering again the levels of MDM2[153]. Three main independent activator pathways of TP53 have been described: DNA damage, oncogenic signaling and insults such as ultraviolet (UV) light or drug-induced stress, all of them enhancing a wide range of protective pathways. Upon DNA damage (e.g. DNA double-strand breaks), the ataxia telangiectasia mutated (ATM) protein kinase phosphorylates the checkpoint kinase 2 (CHK2), which phosphorylates and activates TP53[154] (**Figure 8**). Oncogene-induced growth signals activate TP53 through $p19^{ARF}$. When the insults arise from drugs such as chemotherapeutics or protein kinase inhibitors, or UV, the activation of TP53 is driven by the ataxia-telangiectasia-mutated-and-Rad3-related kinase (ATR) and casein kinase II (CK2). The phosphorylation of TP53 by these proteins impairs its binding to MDM2, hence stabilizing TP53. However, the phosphorylation of TP53 is not the only mechanisms to avoid MDM2-TP53 binding. As previously mentioned, TP53 activation through oncogenic signaling induces $p19^{ARF}$, which also binds to MDM2 and prevents its binding to TP53[149].

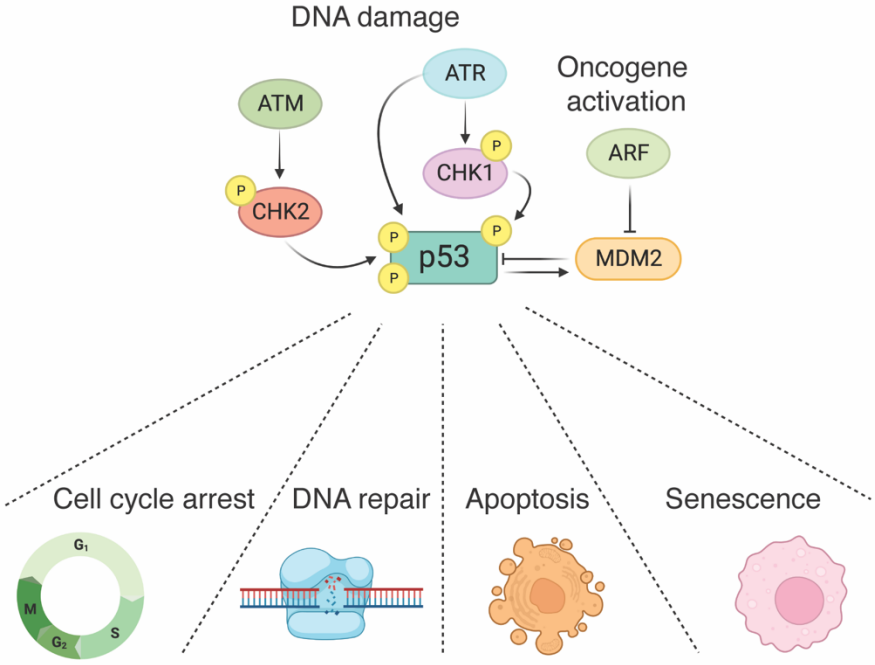

Figure 8. TP53 activation and effects.
DNA damage activates TP53 through ATM and ATR, while oncogenic signatures repress the MDM2-TP53 binding to activate TP53. Once activated, TP53 induces a network of protective signaling pathways that can result in cell cycle arrest, DNA repair, apoptosis and senescence.

Once TP53 is activated, it induces a pleotropic network of signaling pathways and the direction of this effect depends on the extent of the damage and the cell type. One of the main results of TP53 activation is the inhibition of the cell cycle through the expression of $p21^{WAF1/CIP1}$ to allow the repair of the DNA damage, resulting in cell cycle arrest. P21 is an inhibitor of cyclin-dependent kinases (CDK) encoded by *CDKN1A* and plays a role in the inhibition of G1-S and G2-M transitions in the cell cycle[155]. It has also been shown that TP53 can induce G2 arrest by upregulating the 14-3-3σ protein[156]. On the other hand, when the damage is too severe, TP53 activation can mediate cell death through apoptosis by activating proteins such as Bax, NOXA or p53AIP1[157,158], or senescence.

8.2. p53 in EML4-ALK lung cancer

TP53 is mutated in a multitude of cancers and it is considered a poor prognostic marker. The loss of function of TP53 is most frequently achieved by point mutations in hot spots or deletions in more rare cases. Missense mutations account for most of TP53 changes, which result in single residue substitutions in its DNA-binding domain, impairing the transactivation of its target genes[159]. In addition to loss of function mutations, gain of function mutations with oncogenic features have also been observed[160]. These gain of function mutations have been shown to induce the expression of certain genes such as the multidrug resistance-1 (MDR-1) gene[161]. Typically, TP53 mutations also lead to the inactivation of the remaining wild type copy of TP53 through a dominant negative effect[159].

TP53 has been described to be the most mutated gene co-occurring with EML4-ALK in lung cancer patients[162]. Recent studies have shown that despite being the most mutated gene in these patients, it only appears to be affected in 24% of the tumors. Nevertheless, TP53 mutations were shown to predict an unfavorable outcome after treatment, reinforcing the importance of studying EML4-ALK/TP53 tumors[163]. Some studies have suggested that the concurrent mutation of TP53 with EML4-ALK V3 is a factor of poor prognosis due to the enhanced metastatic potential[164]. Yet, there is certain controversy on this topic. Noh et al. described that TP53 mutations were most common in long variants (e.g. V1) than short variants (e.g. V3), whereas other studies showed opposite result, where TP53 appeared more frequently mutated in V3 tumors[66,162,165].

Due to these discrepancies, it is essential to study and understand the role of TP53 in EML4-ALK variants and the effects of the different TP53 mutations in these tumors.

Aims of the thesis

Since the discovery of the *EML4-ALK* fusion gene in lung cancer patients in 2007, this type of cancer has attracted considerable interest and is currently the subject of active investigations and discussions within the academic community. Significant research efforts have been dedicated to the development of targeted therapies for EML4-ALK lung cancer, resulting in a wealth of scientific literature and clinical studies on the subject. Nevertheless, an important knowledge gap remains with regard to the biological mechanisms and clinical implications of the various EML4-ALK fusion variants. Therefore, the aims of this thesis are:

- Characterization of the biological and molecular divergence among V1 and V3 *Eml4-Alk* fusions.
- Investigation of the cell of origin of *Eml4-Alk* tumors.
- Deciphering the tumor initiation process and role of TP53 in *Eml4-Alk* tumors.

Materials & Methods

Materials

Table 4. Mouse strains

Mouse Strains	Source
FVB	DKFZ
Krt5-CreERT2	The Jackson Laboratory
mT/mG	Kindly provided by Georgios Stathopoulos
***Trp53*$^{fl/fl}$**	Kindly provided by Claudia Scholl
Rosa-LSL-Cas9-EGFP	Kindly provided by Guoliang Cui
Scgb1a1-CreERT	Kindly provided by Brigit Hogan
Sftpc-CreER-rtTA	Kindly provided by Harold Chapman

Table 5. Mouse genotyping primers

Gene	Primer name	Sequence	Product size (bp)
***Krt5*-CreER**	CreER fwr	GTGCCTGGCTAGAGATCCTG	260
	CreER rev	AGAGACTTAGGGTGCTGGA	
mT/mG	mT/mG-1 common	CTCTGCTGCCTCCTGGCTTCT	wt: 330 mutant: 250
	mT/mG-2 wt reverse	CGAGGCGGATCACAAGCAATA	
	mT/mG-3 mt reverse	TCAATGGGCGGGGGTCGTT	
***Trp53* floxed**	p53-flox (fwr)	GGTTAAACCCAGCTTGACCA	wt: 270 mutant: 390
	p53-flox (rev)	GGAGGCAGAGACAGTTGGAG	
Rosa-LSL-Cas9-EGFP	Wt reverse	CAG GAC AAC GCC CAC ACA	wt: 241 mutant: 214
	Wt foward	AAG GGA GCT GCA GTG GAG TA	
	Mut reverse	CTT CTT CTT TGG GGC CAT CT	
	Mut foward	TCC CCA TCA AGC TGA TCC	
***Scgb1a1*-CreERT**	CCSP-CreER-F	ACTCACTAT TGG GGG TGT GG	wt: 550 mutant: 245
	CCSP-CreER-wt-R	AGG CTC CTG GCT GGA ATA GT	
	CCSP-CreER-mut-R	CCA AAA GAC GGC AAT ATG GT	
***Sftpc*-CreER-rtta**	SPC 5 exon Forward	CCC AGT CCC TCT CTG AAT TTG	wt: 500 mutant: 1000
	SPC-Cre wt rev	GTT TCT ACC GAC CCT GTG AAG	
	SPC-Cre mut	CAT CGC TCG ACC AGT TTA GTT A	

Table 6. PCR primers

Gene	DNA	Primer name	Sequence
***Eml4* (V1)**	gDNA	V1.genot.f1	GTAGACAGACATCCACTGATCATACTGTG
	gDNA	V1.genot.f2	TTTTAGATGTCCTTTCTCTCCTGGGC
***Eml4* (V3)**	gDNA	V3.genot.f1	CCGTGGGACAGTCTTGTTGTCAG
	gDNA	V3.genot.f2	ACTCCCTCCAGACCTTGGAC
Alk	gDNA	Alk.genot.r2	ACTCCCTCCAGACCTTGGAC
	gDNA	Alk.rev2	ATAAGAGAGCCTGACACGAGC
***Eml4* (V3)**	cDNA	Eml4.cDNA.for.v3	CGTCCACCAACAGCTGAAAAG
Alk	cDNA	Alk.cDNA.rev	GGTCATGATGGTCGAGGTCC

Table 7. Primary antibodies

Name	Dilution	Source	Catalog number
BrdU	1:200	Abcam	ab6326
Casp3	1:200	Cell signaling	9661L
CC10	1:500	Santa Cruz	SC-9772
Cytokeratin 5	1:200	Abcam	ab53121
gamma-H2AX	1:200	Bethyl Labs	IHC-00059
GFP	1:200	Abcam	ab5450
Lactotransferin	1:500	Sigma	07-685
p40	1:200	Abcam	ab172731
p63	1:200	Sigma	ab735
ProSP-C	1:500	Millipore	AB3786
RFP	1:200	Rockland	600-401-379
TTF1	1:200	Abcam	Ab76013

Table 8. Secondary antibodies for immunofluorescence

Name	Dilution	Source	Catalog number
Alexa 488 donkey anti-goat IgG	1:500	Abcam	ab150129
Alexa 568 donkey anti-rabbit IgG	1:500	Thermofisher	A10042

Table 9. FACS antibodies

Name	Dilution	Source	Catalog number
CD11b-APC	1:200	Biolegend	101211
CD200-BV421	1:200	Biolegend	565547
CD24-BUV395	1:200	Biolegend	744471
CD31-APC	1:200	Biolegend	102409
CD31-PE	1:200	Biolegend	102408
CD45-APC	1:200	Biolegend	103111
CD45-PE	1:200	Biolegend	103105
DAPI	1 µg/ml	Life technologies	62248
EpCAM-BV711	1:200	Biolegend	118233
Propidum iodide	1 µg/ml	Life Technologies	BMS500PI
ß4-Alexa Fluor 647	1:80	Biolegend	123608
Ter-119-APC	1:200	Biolegend	116211

Table 10. Equipment

Equipment	Source
BD Canto II	BD Biosciences
BD FACSAria Fusion	BD Biosciences
HiSeq 4000	Illumina
Leica tissue processor ASP300S	Leica
LightCycler II® 480	Roche
Molecular Imager GelDoc™ XR+	Biorad
Nanodrop Lite Spectrophotometer	Life Technologies
NovaSeq 6000	Illumina
RM 2135 microtome	Leica
Thermobrite sealed humidified chamber	Abbott
Ultracentrifuge Optime XE	Beckman Coulter
ZEISS Cell Observer	Zeiss

Table 11. Commercial kits

Kit name	Source	Catalog number
AllPrep DNA/RNA Mini	Qiagen	80204
Chromium Next GEM Single Cell 3' GEM, Library & Gel Bead Kit v3.1	10x Genomics	PN-1000121
DAB Peroxidase Substrate kit	Vector Labs	SK-4100
HotStart Master Mix Kit	Qiagen	203443
Mouse on Mouse (M.O.M.) Basic Kit	Vector Labs	BMK-2202
Nick Translation kit	Abbot	07J00-001
PureLink Expi Endotoxin-Free Maxi Plasmid Purification Kit	Life Technologies	A31217
QIAamp DNA Mini Kit	Qiagen	51304
QIAamp RNeasy Mini Kit	Qiagen	74104
QIAGEN Plasmid Midi Kit	Qiagen	12143
QIAquick Gel Extraction kit	Qiagen	28706
QuantiTect Reverse Transcription Kit	Qiagen	205313
TruSeq Stranded kit	Illumina	20020596
VECTASTAIN Elite ABC kits	Vector Labs	PK-6101; PK-6105; PK-6104

Table 12. Reagents and chemicals

Reagent name	Source	Catalog number
10% formalin	Sigma	HT501128
1x Dream Taq Green Buffer	Life Technologies	B71
1X PBS	Life Technologies	14190169
Ampicillin	Sigma	A9518
BrdU	Life Technologies	000103
CD31 microbeads	Miltenyi Biotec	130-097-418
CD45 microbeads	Miltenyi Biotec	130-052-301
Chloramphenicol	Sigma	PHR1412
Dispase	Corning	354235
DNASE I, grade II	Sigma	10104159001
dNTP mix	Promega	U1240
DPX mounting media	Sigma	06522
Dulbecco's Modified Eagle Medium (DMEM)	Life Technologies	419650394
Elastase, pancreatic type I	Sigma	E1250
Eosin G	Roth	3137.2
Ethanol 96%	DKFZ	14926
Ethanol absolute 100%	DKFZ	12621
Fetal bovine serum (FBS)	Gibco	10270-106
Glycogen	Sigma	10901393001
Hematoxylin QS	Linaris	H-3404
Hydrogen Peroxide	Sigma	H1009
L-Glutamine	Gibco	25030024
LSI/WCP Hybridization buffer	Abbott	06J67-001
Mouse Cot-1 DNA	Thermo Scientific	18440-016
Mowiol	Merck Chemicals	475904
Normal donkey serum	Jackson Immuno	017-000-121
OPTI-MEM	Gibco	31985062
PBS pH 7.4	Life Technologies	10010015
pCMV-dR8.2	Addgene	8455

pCMV-VSV-G	Addgene	8454
Penicillin-streptomycin	Gibco	15070-063
Pepsin	Sigma	P6887
PeqGreen DNA dye	VWR	732-2960
ProLong Diamond Antifade Mountant - DAPI	Life Technologies	P36966
Red blood lysis buffer	Sigma	R7757
SEEBRIGHT Green 496 dUTP	Enzo	ENZ-42831
SEEBRIGHT Orange 552 dUTP	Enzo	ENZ-42842
SEEBRIGHT Red 650 dUTP	Enzo	ENZ-42522
Sodium acetate 3M ph5.2	Life Technologies	R1181
Sodium butyrate	Hölzel Diagnostika	HY-B0350A
Tamoxifen	Sigma	T5648
Triton X-100	Th. Geyer	10755706
Unmasking solution	Vector Labs	H-3300
Xylene	DKFZ	13435

Table 13. Softwares and algorithms

Tool	Source
Adobe Illustrator	Adobe
Aiforia	Aiforia.com
Benchling	Benchling.com
Biorender	Biorender.com
Fiji Software	Wayne Rasband (NIH)
FlowJo	BD biosciences
GraphPad Prism v8	GraphPad software
Kallisto v0.46.1.	Bray et al. 2016[166]
R v4.0.3	N/A
StrataQuest software	TissueGnostics
TissueFAXS technology	TissueGnostics

Table 14. R packages

Package	Reference
clusterProfiler	Yu et al. 2012[167]
DEseq2	Love et al. 2014[168]
Harmony	Korsunsky et al. 2019[169]
MuSiC	Wang et al. 2019; Fan et al. 2022[170,171]
scanpy	Wolf et al. 2018[172]
Seurat	Stuart et al. 2019[173]

Methods

Mouse lines, tamoxifen and virus injection

Mice were housed at the DKFZ (Deutsches Krebsforschungszentrum) animal facility under a constant 12-hour light-dark cycle and maintained on a standard diet with ad libitum access to food and water. Experiments were performed in accordance with institutional guidelines, with ethical approval by the Regierungspräsidium Karlsruhe, Germany, under permit numbers G265/19 and G185/17.

FVB mice were provided by DKFZ. *Trp53 floxed* (*Trp53*$^{fl/fl}$)[174] mice were supplied by Claudia Scholl (DKFZ, Heidelberg, Germany); *mT/mG* mice[175] were kindly provided by Georgios Stathopoulos (University of Patras, Greece). *Krt5-CreERT2*[128] mice were purchased from The Jackson Laboratory. *Sftpc-CreER-rtTA*[127] mice were kindly provided by Harold Chapman (University of California, San Francisco, US), and *Scgb1a1-CreERT*[85] mice were kindly provided by Brigit Hogan (Duke University, Durham, US). *Rosa-LSL-Cas9-EGFP* (LSL-Cas9-EGFP)[176] mice were kindly provided by Guoliang Cui (DKFZ, Heidelberg, Germany). *Trp53*$^{fl/fl}$ and LSL-Cas9-EGFP mice were used in homozygosity, whereas *mT/mG*, *Scg1a1-CreER, Sftpc-CreER* and *Krt5-CreER* mice were used in both homozygosity and heterozygosity. All animals were kept in FVB background, except the *Scgb1a1-CreERT*, for which both pure FVB and mixed FVB-C57BL/6 backgrounds were used, and LSL-Cas9-EGFP, which were kept in C57BL/6 background.

Adult mice aged 6-12 weeks, both female and male, were used for experiments. For cell type labeling, linage tracing mice were injected with 200 µg/kg tamoxifen (TAM) (Sigma, T5648) dissolved in corn oil by intraperitoneal injection (i.p.) for 4 consecutive days. Mice were instilled with the adenovirus 4 weeks after TAM treatment, to allow enough time for complete metabolization of the TAM. For virus instillation, mice were anesthetized by i.p. injection of 100 µg/g ketamine and 14 µg/g xylazine and intratracheally instilled with the virus. Adenoviruses were diluted in DMEM (Life Technologies, 41965039) and 2M $CaCl_2$ in a final volume of 50 µL per mouse. FVB wildtype mice received 2e8 PFU viral dose, and *Trp53*$^{fl/fl}$ or *Trp53*$^{+/+}$ mice crossed with the linage tracing mice were instilled with 1.5e8 PFU viral dose. Lentiviruses were diluted in PBS, and mice were instilled with 1x10^6 viral particles in a maximum volume

of 100 µL. The *Eml4-Alk* V1 adenovirus was designed in the Ventura lab[133] and purchased from Viraquest. The *Eml4-Alk* V3 adenovirus was design in our lab and produced by Viraquest.

Disease progression was monitored every 2 days and mice were sacrificed when they reached the established humane end point with signs of sickness such as heavy breathing, weight loss or signs of lethargy or at the indicated timepoints. For the viral efficiency test, mice were euthanized 2, 7 and 15 days after virus instillation. For proliferation detection, mice were injected with 1ml/100gr BrdU i.p. 2 hours prior to euthanizing them. All mice were euthanized by cervical dislocation. The inferior vena cava was cut and the heart was perfused with 10ml 1xPBS through the left ventricle to clear the blood from the tissues. The lungs were then collected for further analyses. For DNA or RNA isolation, tissues were snap frozen in liquid nitrogen and stored at -80°C.

Lentiviral plasmid isolation

The lentiviral *Eml4-Alk* plasmids were designed and generated by Alberto Diaz and Dr. Kalman Somogyi in our lab. For large scale plasmid production, transformed *E. coli* bacteria were grown in 200 ml Luria-Bertani (LB) media containing Ampicillin overnight at 37°C. For isolation, PureLink Expi Endotoxin-Free Maxi Plasmid Purification Kit (Life Technologies, A31217) was used following manufacturer's instructions. Briefly, bacteria were centrifuged for 10 min at 4°C. 6 ml Add 6 ml resuspension buffer with RNase A were added to the pellet to resuspend the cells. Then bacteria were lysed with 6 ml of lysis buffer and incubated for 5min at room temperature. Next, 6 ml of precipitation buffer were added and the lysate was clarified using a clarification column by centrifuging for 10 minutes. Endotoxins were removed by adding 2.5 ml of the endotoxin removal buffer and the lysate was then loaded onto a DNA-binding Column and drained using a vacuum. The column was then washed using 20 ml of wash buffer and DNA was then eluted using 15 ml endotoxin-free elution buffer, and after 1-minute centrifuge, the DNA was precipitated with isopropanol. After 30 min centrifugation at 4°C, the DNA pellet was washed with 70% ethanol and air-dried for 10 minutes. Finally, DNA was resuspended in dH_2O. Plasmid concentration was measured using NanoDrop Lite Spectrophotometer (Life Technologies, ND-LITE).

Lentivirus production and titration

For lentiviral production, 20x10^6 Lenti-X 293T Cells (Takara, 632180) were plated in a 150 mm dish. 20 hours later, cells were incubated with 5mM sodium butyrate (HDAC inhibitor) in DMEM with 10% FBS and 1% Glutamine for 5 hours. Then, for transfection, a mixture of 3 plasmids was used in 200 µl Opti-MEM I Reduced Serum Medium (Life Technologies, 31985062) per 150 mm plate: 4.95 µg pDr8.2 (Addgene, 8455), 1.9 µg VSV-G (Addgene, 8454) and 7.4 µg transfer plasmid. The plasmid mixture was then mixed with 1M PEI (previously diluted in Opti-MEM) up to a total volume of 1.2 ml per plate and incubated for 20 minutes at room temperature. The mixture was then gently added to the cells and incubated for 16-18 hours. Then, the media was replaced with collection media (DMEM with 12% FBS and 2% glutamine). Virus containing media was collected 24 and 48 hours after transfection and filtered using 45 µm filters to remove cell debris. To concentrate the virus, the collected media was loaded into Ultra-Clear 38.5 ml tubes (Beckman Coulter, 344058) and centrifuged for two hours in an ultracentrifuge at 25000 rpm at 4°C. Pellets were then resuspended in PBS pH 7.4 (Life Technologies, 10010015) and stored at -80°C.

For tittering, 200,000 Lox-Stop-Lox-YFP mouse embryonic fibroblasts were plated in 6-well plates. Duplicates were plated for control samples (no virus added), batch effect control and each of the virus to be tittered. 24 hours after plating, cells in one control well were counted and the virus was added to the rest of the corresponding wells. 48 hours after adding the virus, every well was collected and processed for flow cytometry. Cells were incubated with DAPI (1 µg/ml; Life technologies, 62248) for live/death cell exclusion. The leftover control well was used for unstained control. Cells were analyzed using a BD FACS Canto system. DAPI and YFP intensity were measured and 10,000 cells per sample were recorded (**Figure 9**). The percentage of DAPI$^-$; YFP$^+$ cells, was considered. To calculate the viral titter, the following formulas were used:

$$X = \frac{\#cells\ at\ 24h \times \%YFP{+}cells}{\mu l\ of\ virus} \quad Titter\ (viral\ particles/\mu l) = \frac{X \times 27}{\%YFP{+}cells\ with\ batch\ effect\ virus}$$

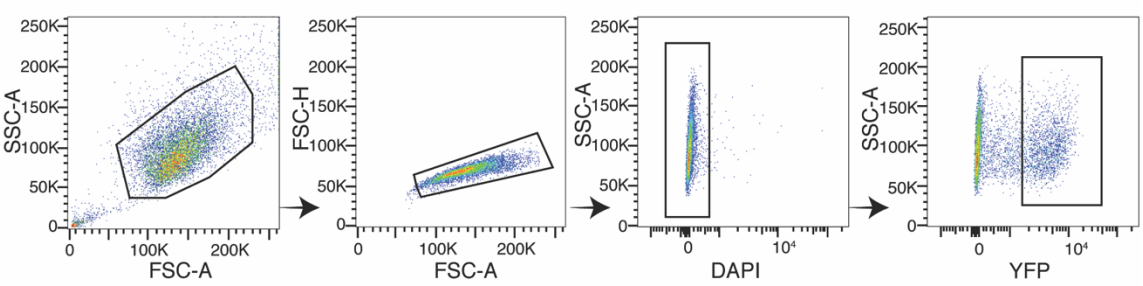

Figure 9. Titration of lentiviral particles. 10,000 Lox-Stop-Lox-YFP mouse embryonic fibroblasts were recorded. After debris (SSC-A vs FSC-A), doublets (FSC-H vs FSC-A) and dead cell exclusion (SSC-A vs DAPI), the titer was calculated based on the percentage of YFP+ cells.

Mouse genotyping

DNA was isolated from earpunches by incubating them with 100 µl of 0.05M NaOH for 1 hour at 98˚C. The reaction was neutralized with 10 µl of 1M Tris HCl pH7.5, and 1 µl of DNA was used in a total volume of 20 µl reaction. The PCR master mix was prepared as follows: 1x Dream Taq Green Buffer (Life Technologies, B71), 0.25 pmol/µl forward primer, 0.25 pmol/µl reverse primer, 200 µM dNTPs mix, 1U/20µl Taq Polymerase enzyme. The PCR program used was as follows: 95°C for 2', 30 cycles (95°C for 30'', 60°C for 30'', 72°C for 30''), 72°C for 1'. The primers used are listed in **Table 5**. The PCR products were run in a 2% agarose gel with peqGreen DNA dye at 120V and visualized under a Molecular Imager GelDoc™ XR+ (Biorad).

BAC DNA isolation and probe labeling

The following bacterial artificial chromosome (BAC) clones were used: 3' *Alk* (17qE1.3, RP23-306H20, RP23-397M18), 5' *Alk* (17qE1.3, RP23-12H17, RP23-403F20), and 5' *Eml4* (17qE4, RP23-193B15)[133], and were kindly provided by Andrea Ventura (Memorial Sloan Kettering Cancer Center, New York, US). BAC-containing bacteria were grown overnight in 200 ml of LB supplemented with Chloramphenicol at 37°C with vigorous shaking. BAC clones were isolated using the QIAGEN Plasmid Midi Kit (Qiagen, 12143). Bacteria were centrifuged at 4,000 rpm for 20 minutes and pellets were processed according to technical specifications. The DNA pellet was washed with 2 ml of 70% ethanol and centrifuged again for 10 minutes. Pellets were left to air-dry for 10 minutes and dissolved in dH_2O.

The labeling of the BAC clones was performed by using the Nick Translation kit (Abbot, 07J00-001). Briefly, the BAC DNA was mixed with the different reagents as listed

in **Table 15**. One reaction was performed for each clone. The mix was incubated at 15°C for 2 hours and the reaction was stopped by placing the mix at 70°C for 10 minutes. Then, 20 µg of Cot-1 DNA (1mg/µl stock) were added and the two clones labeling the same DNA were pooled together. Then, ethanol precipitation was performed as listed in **Table 16** and incubated on ice for 15 minutes. After a 30 minutes centrifugation at 4°C, 2 washes with 70% ethanol were performed and the pellet was left to air dry for 10 minutes. The pellet was resuspended in 15 µl of dH_2O and stored at -20°C. The probes were labeled as follows: 5' *Eml4* with orange 552 dUTP, 5' *Alk* with red 650 dUTP and 3' *Alk* with green 496 dUTP.

Table 15. Nick translation reagents

Reagent	Amount in µl
1 µg BAC DNA	X µl
dH_2O	(22.5-X) µl
0.1 mM dTTP	5 µl
0.1 mM dATP, dCTP, dGTP mix	10 µl
10x Nick translation buffer	5 µl
Nick translation enzyme	5 µl
0.2 mM colored dUTP	2.5 µl

Table 16. Ethanol precipitation. n represents number of samples.

n x 70 µl sample
n x 7 µl 3M sodium acetate
4 µl glycogen
n x 175 µl 100% Ethanol

Fluorescence in situ hybridization (FISH)

Interphase FISH was performed using a 3-color probe mix which allows the detection of the *Eml4-Alk* fusion in mouse cells. Paraffin embedded lung tissues were cut in 3 µm sections and placed onto superfrost slides (VWR International, 631-9483). Slides were deparaffinized in Xyline (two times 10min) and rehydrated through ethanol grades steps (3x 100%, 96%, 70%) for 3 minutes each and 5 minutes in dH_2O. Antigen retrieval was performed using 0,09% (v/v) unmasking solution (Linaris, H-3300) for 30min in a steamer at 70°C and slides were then rinsed in dH_2O for 5min. Sections were then incubated with 0.005% Pepsin in pre-warmed 10mM HCl for 15 minutes at 37°C and then washed 2 times with dH_2O and once with PBS for 3 minutes each. Then, sections were dehydrated through increasing grades of ethanol (70%, 85%, 100%) for 3 minutes each. For denaturation and hybridization, the probe mix was prepared as

follows: 7 μl LSI/WCP Hybridization buffer (Abbott, 06J67-001) + 1 μl of each probe per section. The 10 μl of probe mix was applied to the section, covered with a coverslip, sealed with Fixo gum and incubated in a Thermobrite sealed humidified chamber (Abbott, 07J91-020) under the following protocol: Denaturation at 76°C for 5 min, hybridization at 37°C for 20-24 hrs. Sections were then washed with 0.4x Saline-Sodium Citrate (SSC) 0.1% Tween20, followed by 0.4x SSC 0.1% Tween20 prewarmed at 74°C, 2x SSC 0.1% Tween20 and 2x SSC. Between every wash, a PBS was performed. Post hybridization washes were performed in the dark for 2 minutes each. Slides were then mounted using ProLong Diamond Antifade Mountant with DAPI (Life Technologies, P36966) and imaged using a Zeiss Cell observer. Z-stacks were taken with an interval of 0.5 μm, and image analysis was performed using Fiji.

DNA, RNA isolation and cDNA synthesis

Snap frozen tissues were powdered using a pestle and mortar, and 30 mg of tissue were used for further processing. To isolate DNA and RNA from the same tumor, the All prep DNA/RNA mini kit (Qiagen, 80204) was used. To isolate only DNA, the QIAamp DNA Mini Kit (Qiagen, 51304) was used. To isolate only RNA, the RNeasy Mini Kit (Qiagen, 74104) was used. All kits were used following manufacturer's instructions. DNA and RNA concentrations were measured using NanoDrop Lite Spectrophotometer (Life Technologies, ND-LITE). For cDNA synthesis, 1 μg of RNA was processed using the QuantiTect Rev. Transcription Kit (Qiagen, 205313) as stated by the manufacturer.

Efficiency test

To test the efficiency of V1 and V3 adenoviruses in generating the *Eml4-Alk* rearrangement, different amounts of DNA were used for each timepoint due to the technique requirements and the increasing frequency of rearrangement with time. 50 ng of DNA were used from the 2 days samples, 5 ng from the 7 days samples and 1 ng from the 15 days samples per reaction. Each reaction was performed in a total volume of 10 μl using the HotStart Master Mix Kit (Qiagen, 203443). Each reaction mix was prepared with 5 μl HotStart Taq Mix, 0.5 μl of 10 μM forward primer, 0.5 μl 10 μM reverse primer, X μl DNA and 4-X μl dH_2O. Two different *Eml4* forward primers were used for each variant to exclude higher efficiency results due to differences in primer binding efficiency. The primers used were: V1.genot.f1, V1.genot.f2, V3.genot.f1, V3.genot.f2 and

Alk.genot.r2. The primer sequences are listed in **Table 6**. The following program was applied: 95°C for 15', 38x [94°C for 30'', 61°C for 30'', 72°C for 30''], 72°C for 10'. A total of 42 reactions (24 with each forward primer) were performed per sample and a total of 6 mice were used for the 2 days timepoint and 3 mice for the 7 days and 15 days timepoints. PCR products were run on a 2% agarose gel with peqGreen DNA dye (VWR, 732-2960) at 120V and visualized with Molecular Imager GelDocTM XR+ (Biorad). Positive reactions were counted and plotted as percentage of positive reactions for each sample.

Tumor genotyping

To detect the *Eml4-Alk* V3 gDNA and cDNA, the HotStart Master Mix Kit (Qiagen, 203443) was used. To detect the gDNA, 100 ng of DNA were employed and V3.genot.f2 and Alk.rev2 primers were used. Each reaction was performed in a total volume of 20 µl and prepared with 10 µl HotStart Taq Mix, 1 µl of 10 µM primer, X µl DNA and 12-X µl dH_2O. The following program was applied: 95°C for 15', 35x [94°C for 30'', 61°C for 30'', 72°C for 30''], 72°C for 10'. For cDNA amplification, 100 ng of cDNA were used and the Eml4.cDNA.for.v3 and Alk.cDNA.rev primers were utilized. The following program was applied: 95°C for 15', 35x [94°C for 30'', 56.5°C for 30'', 72°C for 30''], 72°C for 10'.

PCR products were run on a 2% agarose gel with peqGreen DNA dye (VWR, 732-2960) at 120V, visualized with Molecular Imager GelDocTM XR+ (Biorad), and purified using QIAquick PCR purification kit (Qiagen, 28106) as indicated by the manufacturer. Sequences were confirmed by sanger sequencing performed by Mycrosynth using the following primers: V3.genot.f2, for gDNA, and Eml4.cDNA.for.v3 for cDNA. Primer sequences are listed in **Table 6**. The obtained data was analyzed using the Benchling software.

Immunohistochemistry and immunofluorescence

For histology, lungs were fixed in 10% formalin (Sigma, HT501128) for 24 hours and processed in a tissue processor (Leica, ASP300S). Tissues were then embedded in paraffin, and 3 µm sections were obtained and placed onto superfrost slides (VWR International, 631-9483) for further immunohistological staining. Slides were first deparaffinized in Xyline (two times 10min) and rehydrated through ethanol grades

steps (3x 100%, 96%, 70%) for 3 minutes each and 5 minutes in dH_2O. Antigen retrieval was performed using 0,09% (v/v) unmasking solution (Linaris, H-3300) for 30min in a steamer at 70°C and slides were then rinsed in dH_2O for 5min. Endogenous peroxidases were inactivated using 3% H_2O_2 (Sigma, H1009) for 10min. Tissues were then blocked using VECTASTAIN Elite ABC kits (Vectors Lab, PK-6101; PK-6105; PK-6104) according to the primary antibody species for 20min and incubated with primary antibodies (**Table 7**) diluted in blocking solution for 30min at room temperature or overnight at 4°C. For mouse-hosted antibodies, the Mouse on Mouse (M.O.M.) Basic Kit (Vector, BMK-2202) was used. Secondary antibody staining and biotin-streptavidin incubation were performed following manufacturer's instructions and incubated for 20min and 30 min respectively. DAB peroxidase substrate kit (Vector, SK-4100) was used for antibody detection, and hematoxylin QS (Linaris, H-3404) was used as nuclear counterstaining. After dehydration, slides were mounted using DPX mounting media (Sigma, 06522).

For immunofluorescence, after antigen retrieval, sections were blocked using a blocking solution (5% donkey serum) for 1 hour at room temperature. Then, sections were incubated with primary antibodies (diluted in blocking solution, **Table 7**) overnight at 4°C, followed by 2 washes with 0.2% Triton in PBS and 1 wash with PBS for 10 minutes each. Secondary antibodies (**Table 8**) were diluted at a 1:500 concentration in blocking solution together with DAPI (1:1000) and were incubated for 1 hour at room temperature. Next, 2 washes with 0.2% Triton in PBS and 1 wash with PBS for 10 minutes each were performed and slides were mounted using 20% Mowiol (Merck Chemicals GmbH, 475904).

For hematoxylin & eosin (H&E) staining, after rehydration, sections were incubated with hematoxylin QS (Linaris, H-3404) for nuclear counterstaining and washed 5min in dH2O. Then, tissues were incubated with eosin for 5 min for cytoplasmic counterstaining. After washing, tissues were dehydrated through ethanol grades steps (96%, 2x 100%) and incubated two times 5 minutes in Xyline. Then slides were mounted using DPX mounting media (Sigma, 06522).

Sections were visualized and scanned under a TissueFAXS system (TissueGnostic). For immunohistochemical analysis, every tumor was analyzed individually using StrataQuest software (TissueGnostic). The total number of cells in

each tumor was determined by quantifying hematoxylin positive nuclei. Afterwards, DAB positive nuclei were quantified and the percentage of DAB positive cells out of the total amount of cells was calculated. To calculate the tumor burden, hematoxylin & eosin-stained sections were used. The area of every tumor and the total lung area were measured using StrataQuest. Then the area of all tumors was summed and used to calculate the percentage of tumor area out of the total lung area. Tumor histological patterns and grading were quantified from H&E sections with an automated deep neural network created by myself and Alberto Diaz using Aiforia Technologies. Using the Aiforia platform, we trained a convolutional neural network (CNN) for semantic multi-class segmentation. The CNN underwent training to discern lung stroma and tumors, and detect and categorize tumor grades (grades 1-4) or histological patterns (solid, papillary, lepidic, acinar and squamous). We employed 97 slides from independent mice for supervised training.

Bulk RNA sequencing and bioinformatic analysis

Libraries from total RNA of single nodules were prepared with TruSeq Stranded kit for Illumina platforms using 1.2 ug starting material and the HiSeq 4000 technology (50-nucleotide single-read) was used for library sequencing.

For data analysis, the sequenced reads were aligned to the mm10 reference genome using kallisto v0.46.1[166]. The DEseq2 version 1.30 package was used to obtain the normalized counts and differentially expressed genes (DEGs)[168]. Genes were classified as significantly DEGs at an average count of ≥10 ($P < 0.05$). Pathway enrichment analysis was performed with GSEA Desktop Application v1.4[177,178] (GSEA, Broad Institute). To assess the p values, 1000 permutations for each gene set were used and corrected with the false discovery rate (FDR) method. Only pathways with an FDR-P value ≤ 0.25 were chosen as significantly enriched. For cell type deconvolution, the MuSiC package[170,171] was employed.

Lung dissociation and cell sorting

After perfusion with PBS, lungs were injected with 1 mL of digestion media (50 U/ml dispase, 250 U/ml collagenase, 5 U/ml elastase, 300 μg/ml DNAse I) through the trachea using a 20G catheter (Braun, 4269110S01) and incubated for 1 minute. Then,

the lungs were collected into 2 ml of digestion media, minced into very small pieces and incubated at room temperature on a tube rocker for 40 minutes. After the incubation period, the digestion reaction was stopped using 5 ml of DMEM supplemented with 10% FBS, 1% penicillin/streptomycin and 100 µg/ml DNAse I and the tissue was further dissociated by plungers in 10 cm petri dishes and incubated 10 minutes at room temperature in a plate shaker. The lung tissue was then mechanically dissociated by pipetting until no piece was visible and the liquid was completely homogenized. Next, the cells were filtered through a 70 µm strainer and centrifuged at 1200 rpm for 4 minutes. Red blood cells were removed by incubation with red blood lysis buffer (Sigma-Aldrich, R7757) for 5 minutes in a tube rocker. For SMART-seq2, samples were then directly stained with antibodies. For 10x single cell sequencing, after cell counting, CD45 and CD31 microbeads (Miltenyi Biotec, 130-052-301, 130-097-418) were used to remove immune and endothelial cells respectively following manufacturer's instructions. Briefly, every 10^7 cells were incubated with 10 µl of CD45 microbeads, 10 µl of CD31 microbeads and 180 µl of FACS buffer (PBS 5mM EDTA and 2% FBS) for 15 minutes at 4°C. Afterwards, cells were loaded into LS magnetic columns (Miltenyi Biotec, 130-042-401) and columns were washed three times with 3 ml of FACS buffer. Cells in the flow-through were then processed for sorting.

Single cells were resuspended at a concertation of 10^6 cells per 100 µl and incubated with antibodies for 15 minutes at 4°C. The single cell suspension was then centrifuged 5 minutes at 1200 rpm and the cell pellet was resuspended in FACS buffer. The antibodies used are listed in **Table 9**. Cells were then sorted in a BD FACSAria Fusion cell sorter.

10x single cell RNA sequencing and analysis

For the controls and early timepoints (TAM, Cas9, 2 weeks and 4 weeks), the live EpCAM$^+$; CD45$^-$; CD31$^-$; tdTomato$^-$; GFP$^+$; CD24$^-$; ß4$^+$; CD200$^+$ population was collected and for tumor samples and the 4wk-GFP sample, live EpCAM$^+$; CD45$^-$; CD31$^-$; tdTomato$^-$; GFP$^+$ population was collected. Libraries were prepared using Chromium Next GEM Single Cell 3' GEM, Library & Gel Bead Kit v3.1 (10x Genomics) according to manufacturer's instruction. Data was analyzed by our bioinformatician Dr. Reka Toth using R. Briefly, gene expression counts were acquired using Cell Ranger 3.1 count from

10× Genomics and were analyzed in R 4.0.3 using Seurat 3.2.278[173]. For quality control, the following thresholds were used: 1000–6000 for TAM, 1000–6000 for Cas9, 1000–6000 for 2 wk-1, 2000–6000 for 2 wk-2, 1500–6000 for 4 wk-1, 2000–6000 for 4-wk GFP^+, 1000–6000 for tumor. Cells with more than 10% mitochondrial genes were removed. The cell-cycle stage was calculated using the CellCycleScoring function of Seurat with a publicly available dataset[179]. Batch effect correction was performed using Harmony[169]. Cluster markers were considered when expressed in more than 25% of the cells with a log fold change (FC) of 0.25. For identifying the activity programs, the consensus non-negative matrix factorization (cNMF)[180] v.1.2 was run. The clusterProfiler package[180] was employed for overrepresentation analysis of the activity programs marker genes based on the KEGG[181] and WikiPathways[182] databases. scVelo 0.2.350 was used to model RNA velocity of individual cells, predicting their future state[183]. We used the scanpy[172] 1.7.287 package to generate the Partition-based graph abstraction (PAGA) map of single cells.

SMART-Seq RNA sequencing and analysis

Mouse lungs were processed for sorting 2 weeks, 4 weeks and 6 weeks after lentivirus instillation as previously described. Live $CD45^-$; $CD31^-$; $CD11b^-$; $Ter\text{-}119^-$; $EpCAM^+$; GFP^+ cells were collected into 384-well plates containing lysis buffer. The lysis buffer was composed of RNAse inhibitor, 10mM dNTP mix, polyT primer 10mM and 0.2% Triton. From each mouse, 192 cells were sorted into plates in a BD FACSAria Fusion cell sorter and directly frozen at -80°C. Libraries were prepared using the protocol stablished by the DKFZ Single-cell Open lab. Briefly, the plates were thaw on ice and incubated at 72°C for 3 min and then placed back on ice. The RT mix was then added and incubated at 42°C for 90 min, 70°C for 15 min and then kept at 4°C. The RT mix was composed of Maxima H-, RNAse inhibitor, 5x buffer, PEG 50%, TSO 100 µM and H_2O. Then the cDNA was amplified using Kapa High Fidelity and IS PCR primer, and the following PCR conditions were used: 98°C for 3', 21x [98°C for 20'', 67°C for 15'', 72°C for 6'], 72°C for 5'. The cDNA was then cleaned up using AmpureXP beads and ethanol. Then, 1-3 ng/µl of cDNA per cell were used for library preparation using DMF, 4x tagmentation buffer and Tn5. The plates were incubated for 10 min at 55°C. 0.1% SDS and indexes were then added together with a mix of Kapa High Fidelity and DMSO. The plates were then incubated using the following PCR protocol: 72°C for 3', 98°C for 30''

12x [98°C for 20'', 63°C for 15'', 72°C for 30''], 72°C for 3'. After library DNA clean-up, the libraries were then sequenced using the NovaSeq 6000 technology (100-nucleotide paired-end).

Data was analyzed by our bioinformatician Dr. Reka Toth. Briefly, gene expression counts were acquired by running the nf-core RNAseq 3.6 pipeline with the following parameters: --genome GRCm38 --aligner hisat2 --skip_markduplicates --pseudo_aligner salmon. The merged salmon gene counts were read into R and transformed into a Seurat object. Cells with less than 2,500 and more than 10,000 features or > 1% mitochondrial reads were removed from further analysis. The data was analyzed in R 4.2.1. using Seurat 4.3.0. After normalization and scaling, the batch effect correction was performed using Harmony based on the first 10 principal components with a lambda parameter of 2. The cell-cycle stage was calculated using the CellCycleScoring function of Seurat with a publicly available dataset[179].

Statistical analysis

Statistical analysis was performed with GraphPad Prism. p-values <0.05 were considered statistically significant. Numbers of biological replicates are represented with "n".

Schemes and images

All schemes and illustrations were generated using the BioRender Software (Biorender.com). The figures shown in the introduction, methods, results, discussion and conclusions were generated using Adobe Illustrator 2022.

Contributions

This thesis has been proofread and corrected by Prof. Dr. Rocio Sotillo. Dr. Kalman Somogyi helped in the design and cloning of the plasmids used to generate the V3 adenovirus, which was then produced by Viraquest. Alberto Diaz and Dr. Kalman Somogyi designed and generated the plasmids for the lenti-Cre, lenti-V1 and lenti-V3 viruses. Dr. Yuanyuan Chen and the team in the flow cytometry facility of the DKFZ helped sorting the cells for single cell RNA sequencing. The bioinformatic analysis of the

data from the V1 club tumor development and the tumor initiation experiments was performed by Dr. Reka Toth. The results shown in Figures 20-27 were obtained as a collective effort from Dr. Yuanyuan Chen, Dr. Reka Toth and Prof. Dr. Rocio Sotillo.

Prof. Dr. Rocio Sotillo played a crucial role in helping to define and guide the project.

Results

1. Characterization of the biological and molecular divergence among *Eml4-Alk* variants

Some of the results presented here have been submitted in a manuscript (Chocarro, S. et al., 2023). The figures and results have been adapted from the mentioned manuscript.

1.1 Generation of an Eml4-Alk V3 mouse model

Most studies aiming to understand the mechanisms underlying the development of EML4-ALK-driven lung tumors have been performed using cell lines overexpressing the EML4-ALK fusion protein[71,78,184,185]. This strategy leads to the uncontrolled expression of the fusion protein and does not resemble the reality of human tumors. Since the overexpression of EML4-ALK could lead to incorrect assumptions about the biology of these tumors, modelling the endogenous rearrangement is a better approach. To generate the *Eml4-Alk* V1 fusion gene, I took advantage of a mouse model developed by the laboratory of Andrea Ventura[133], in which intratracheal instillation of an adenovirus (V1-Ad) induces the endogenous *Eml4-Alk* V1 rearrangement exclusively in the lungs of the mice. The adenovirus contains a CRISPR-Cas9 construct formed by a Cas9 sequence and two single-guide RNAs (sg-RNA), one targeting intron 14 of the *Eml4* gene and another one targeting intron 19 of the *Alk* gene (**Figure 10A**). To generate the *Eml4-Alk* V3 gene, during my Master's Thesis, I exchanged the sg-RNA targeting intron 14 of *Eml4* for another one targeting intron 7 of *Eml4*. The V3 adenovirus (V3-Ad) was then produced by the company Viraquest, resulting in exactly the same virus as the V1-Ad but containing a different *Eml4* sg-RNA.

Once the virus was produced and tittered, I injected FVB mice with V3-Ad and waited until they developed tumors. To confirm that the induced tumors were *Eml4-Alk* V3, I performed Fluorescence In Situ Hybridization (FISH, **Figure 10B**) and genotyped single nodules at the genomic DNA (gDNA) and complementary DNA (cDNA) level (**Figure 10C and 10D**). FISH analysis demonstrated the presence of a monoallelic copy of the *Eml4-Alk* rearrangement, mimicking human tumors. The genotyping confirmed

that all the tested tumors contained the V3 *Eml4-Alk* rearrangement and that the fusion gene was expressed in the tumors.

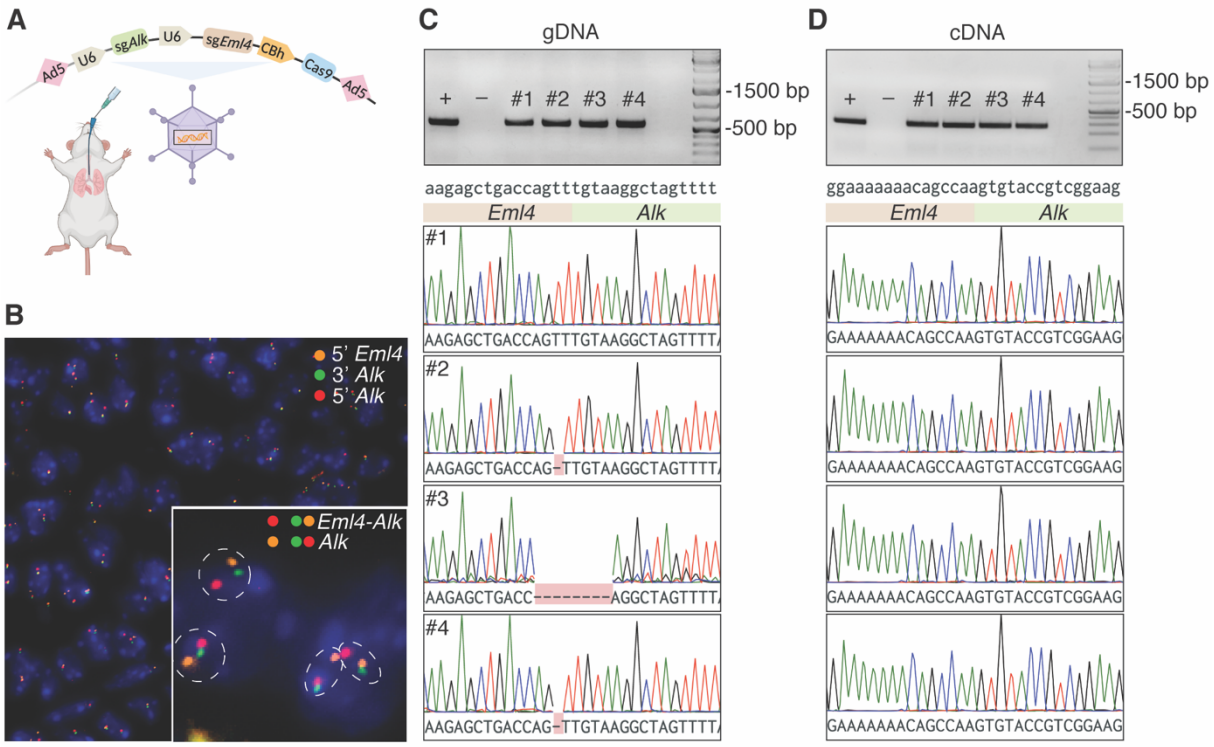

Figure 10. Generation of *Eml4-Alk* V3 tumors.
(**A**) Schematic of adenoviral plasmid and intratracheal instillation. (**B**) Break-apart interphase Fluorescence In Situ Hybridization (FISH) on paraffin sections of V3 lung tumors using three probes: an orange probe targeting 5' *Eml4*, a green probe targeting 3' *Alk*, and a red probe targeting 5' *Alk*. The green and orange probes together and the red probe far apart show the presence of *Eml4-Alk*. The green and red probes together and the orange probe far apart show wildtype *Alk*. In the magnification, two independent cells are displayed. The dashed circles show independent chromosomes. (**C, D**) Detection by PCR of the *Eml4-Alk* V3 fusion gene at the genomic DNA level (gDNA, **C**) and complementary DNA (cDNA, **D**) in microdissected tumors. PCR products were verified by Sanger sequencing. Below, the chromatograms of the *Eml4-Alk* V3 gDNA or cDNA sequences of the four tumors tested are displayed.

1.2. V3 induces lung adenocarcinomas

It has been previously published that *Eml4-Alk* V1-induced mouse tumors are LUADs with a papillary or acinar structure[133]. Careful histopathological examination of my tumors uncovered that V3 not only formed papillary lesions but also induced solid, and lepidic adenocarcinomas. In addition, a few squamous cell carcinomas (SCC) were also found (**Figure 11A**), which had never been reported in the V1 model. Quantification of the LUAD histological patterns revealed that the most prevalent one was the papillary pattern (~60%), followed by the solid (~36%) and lepidic (~4%) patterns (**Figure 11B**). Adenocarcinoma nodules stained positive for Nkx2-1 (pneumocyte marker) and negative for Krt5 (basal cell marker), consistent with the adenocarcinoma classification,

and were also positive for SPC (alveolar type II marker, AT2) and negative for CCSP (club cell marker) (**Figure 11C**), which suggested an AT2-like phenotype of these tumors. Quantification of the number of LUAD and SCC depicted that 1.1% of the tumors (14 out of 1275 tumors) were SCC and that 23% of the mice harbored at least one SCC.

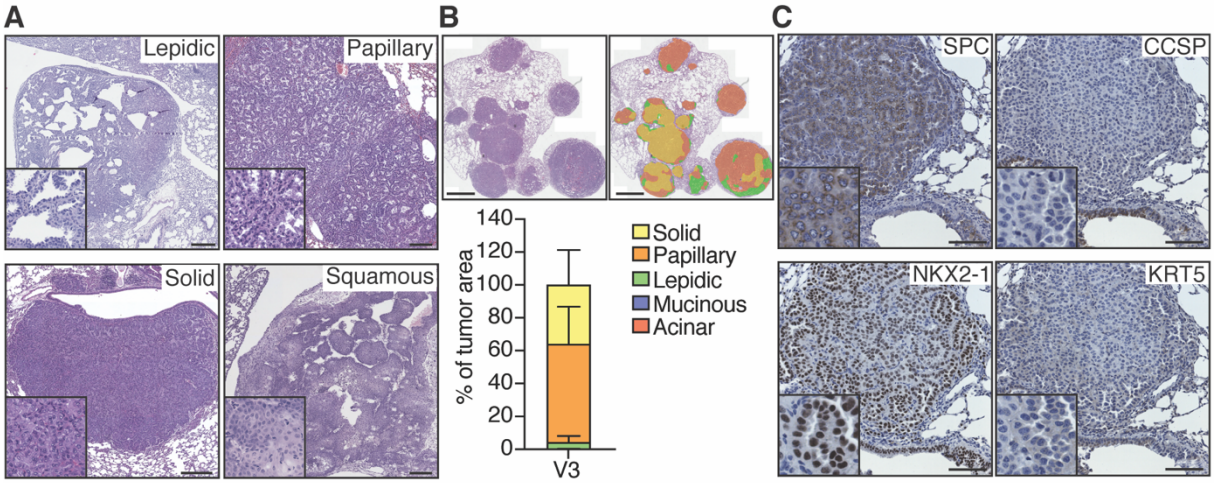

Figure 11. V3 *Eml4-Alk* forms adenocarcinomas and squamous cell carcinomas *in vivo*.
(**A**) Hematoxylin and eosin staining of V3 tumors. Representative images of lepidic, papillary, solid, and squamous tumors. Scale bar: 400 μm. (**B**) Representative H&E image (top left) and representative image of Aiforia pseudocolored LUAD histological patterns (top right) represented as solid (yellow), papillary (orange), lepidic (green), mucinous (blue) and acinar (red). Scale bars: 2mm. Quantification of the percentage of each LUAD histological pattern among the whole tumor area (bottom). (**C**) Immunohistochemistry of the indicated markers in an adenocarcinoma. SPC: marker of alveolar type II cells; CCSP: marker of club cells; NKX2-1: marker of adenocarcinoma; KRT5: marker of squamous cell carcinoma.

1.3. V3 is able to induce a few lung squamous cell carcinomas

FISH analysis confirmed that the SCC found in V3 mice contained the *Eml4-Alk* rearrangement. Surprisingly, several SCC appeared to have multiple copies of the rearrangement, and all of them clustered together (**Figure 12A**). All SCC were positive for Krt5, p63 and p40, and negative for Nkx2-1, SPC and CCSP (**Figure 12B**), in agreement with a squamous architecture. When performing the histopathological characterization of SCC tumors, I found a few nodules which presented both adenocarcinoma and squamous carcinoma features (adeno-squamous, **Figure 12C**), suggesting that SCC tumors could arise from previously existing LUAD after a transdifferentiation, rather than initially originating as SCC. Additionally, some SCC showed a certain level of positivity for CCSP (**Figure 12D**), which suggested that these tumors might originate from club cells.

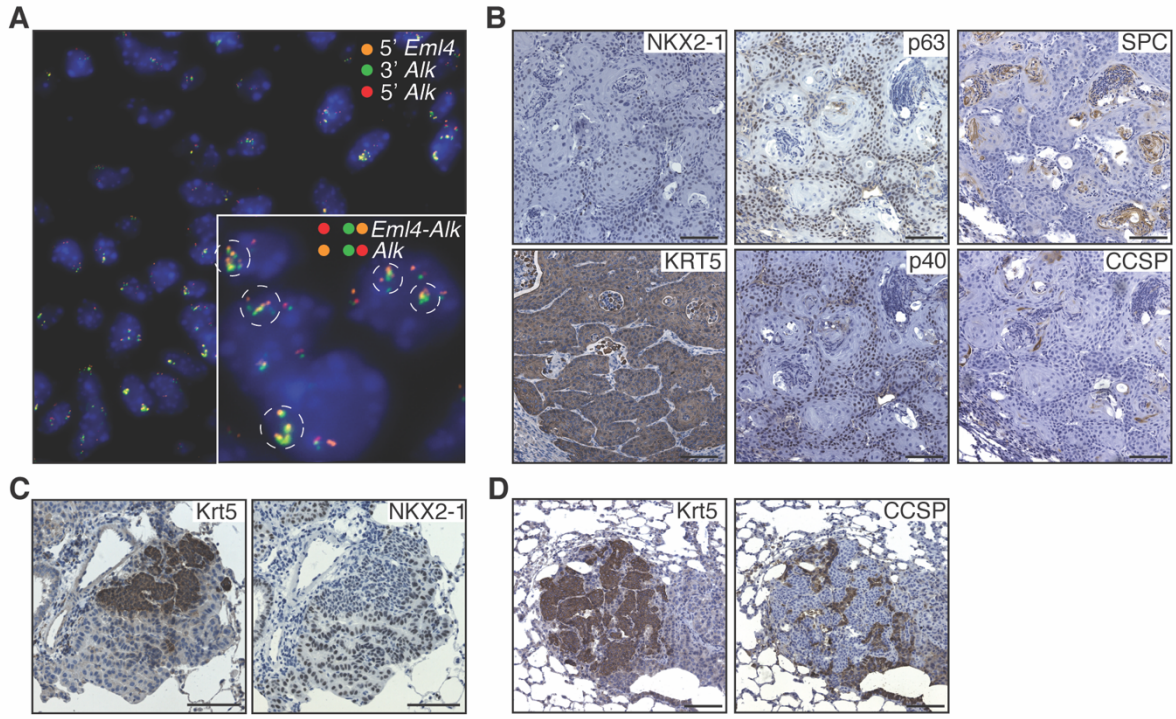

Figure 12. *Eml4-Alk* squamous tumors express canonical markers of squamous cell carcinomas. (**A**) Break-apart interphase Fluorescence In Situ Hybridization (FISH) on paraffin sections of a V3 squamous tumor using three probes: an orange probe targeting 5' *Eml4*, a green probe targeting 3' *Alk* and a red probe targeting 5' *Alk*. The green and orange probes together and the red probe far apart, show the presence of *Eml4-Alk*. The green and red probes together and the orange probe far apart show wild type *Alk*. The dashed circles in the magnification show *Eml4-Alk* clustered copies. (**B**) Immunohistochemistry of the indicated markers in a squamous cell carcinoma. NKX2-1: marker of adenocarcinoma; KRT5: marker of squamous cell carcinoma; SPC: marker of alveolar type II cells; CCSP: marker of club cells; p40 and p63: markers of squamous cell carcinoma. (**C**) Immunohistochemistry in an adenosquamous tumor. (**D**) Immunohistochemistry in a squamous tumor expressing CCSP. (**B-D**) Scale bar 100 μm.

1.4. V3 tumors induce a shorter survival in mice than V1 tumors

Several studies have identified V3 as a more aggressive variant than V1, since patients harboring a V3-driven tumor show increased number of metastases at diagnosis and respond worse to treatment[68,69,186]. Therefore, to understand the development of these tumors, I injected FVB mice with either V1-Ad or V3-Ad and followed them over time until humane endpoint. I observed that V3 mice had a significantly reduced overall survival rate (median: 46 days) compared to V1 mice (median: 126 days) (**Figure 13A**), which correlated with the more aggressive phenotype observed in the clinics. At endpoint, V1 and V3 mice showed no differences in tumor burden or number of nodules, suggesting that the prolonged survival of V1 mice was not due to a smaller number of tumors being induced (**Figure 13B and 13C**). Quantification of the tumor grade following the grading system shown in Figure 1, revealed no significant differences between V1 and V3 tumors (**Figure 13D**). Since the median

survival of V3 mice was 46 days, I compared the lungs of V1 and V3 mice 46 days post-adenovirus instillation. While V3 mice had a tumor burden of ~40%, V1 mice presented only ~1% of tumor burden (**Figure 13E**). These results correlated with a high number of tumors found in V3, whereas tumors in V1 mice at that timepoint were barely observed, similar to the findings in the original V1 publication[133]. (**Figure 13F**). Altogether, these results suggested that V3 tumors developed earlier than V1 tumors upon induction of the *Eml4-Alk* fusion.

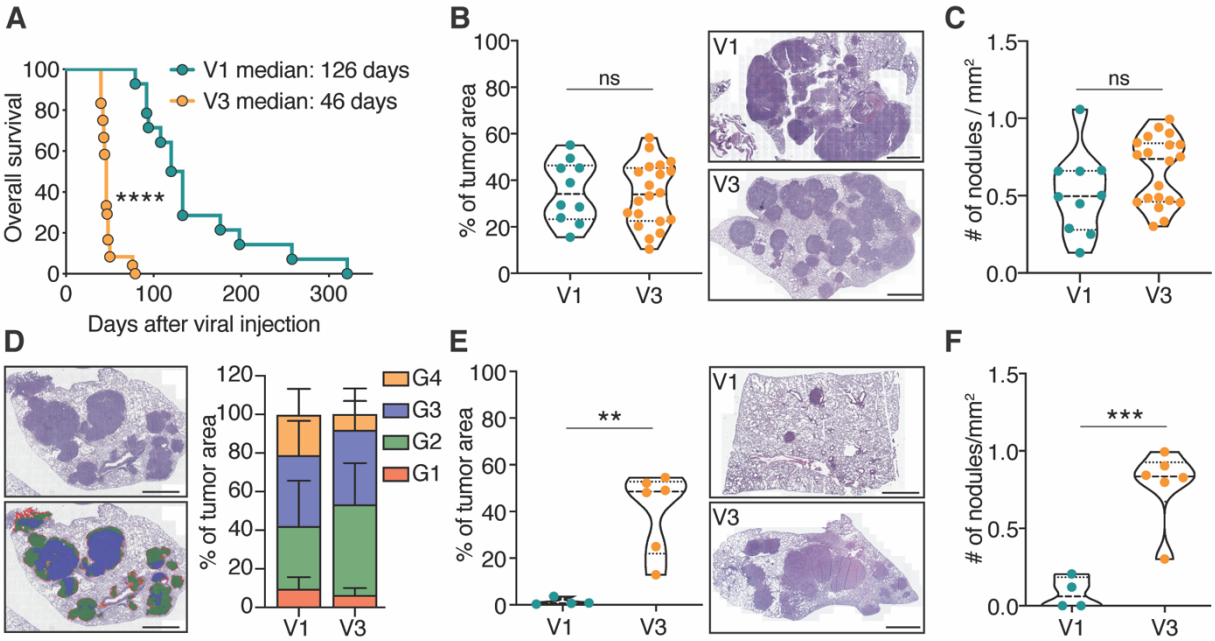

Figure 13. V3 induces a shorter overall survival of mice compared to V1.
(**A**) Overall survival of FVB mice harboring V1 or V3 tumors. V1 median survival of 126 days (n=14) and V3 median survival of 46 days (n=24). **** p<0.0001, Log-rank (Mantel-Cox) test. (**B**) Percentage of tumor area at humane endpoint. On the right, representative images of lungs from V1 and V3 mice. Scale bar: 2mm. (**C**) Number (#) of nodules per mm² of lung at humane endpoint. (**B, C**) V1 n=10, V3 n=20. (**D**) Representative H&E image (top left) and representative image of Aiforia pseudocolored tumor grading (bottom left) represented as Grade 1 (red), Grade 2 (green), Grade 3 (blue), and Grade 4 (orange). Scale bars: 2mm. Quantification of the percentage of each grade among the whole tumor area (right). (**E**) Percentage of tumor area 46 days post-adenovirus instillation. On the right, representative images of lungs from V1 and V3 mice. Scale bar: 2mm. (**F**) Number (#) of nodules per mm² of lung 46 days post-adenovirus instillation. (**E, F**) V1 n=4, V3 n=6. (**B-F**) **p=0.0013, ***p=0.0007, ns (not significant); Unpaired t-test. For all, except D, each dot represents a mouse.

1.5. V1 and V3 adenoviruses generate the *Eml4-Alk* fusion with the same efficiency

CRISPR/Cas9 genome editing technology has been shown to have off-targets[187] and different efficiencies depending on the sgRNA employed. To address if the difference in survival was due to a distinct efficiency of the different *Eml4* sgRNAs in the two adenoviruses, I analyzed the frequency of the *Eml4-Alk* rearrangement very early after virus injection (2, 7 and 15 days) by designing a digital PCR-like technique (**Figure 14A**).

The digital PCR is used to determine the precise concentration of a target nucleic acid molecule in a sample. It allows to quantify rare mutations and copy number variations. This technique entails the division of a sample into several very small droplets, followed by the individual PCR amplification of each partitioned sample. Therefore, to detect the *Eml4-Alk* rearrangement, I used the lowest possible amount of DNA per reaction and performed 42 PCR reactions per sample, to then quantify in a comparable way the percentage of V1 and V3 positive reactions. I observed that the frequency of *Eml4-Alk* was not significantly different between the two variants at any of the three timepoints (**Figure 14B**), suggesting that the shorter survival of V3 mice must be a phenotype of the variant and not a consequence of the CRISPR/Cas9 system. Nevertheless, at 7 and 15 days, a slight increase in V3 was observed, which could be due to a faster growth of V3 cells.

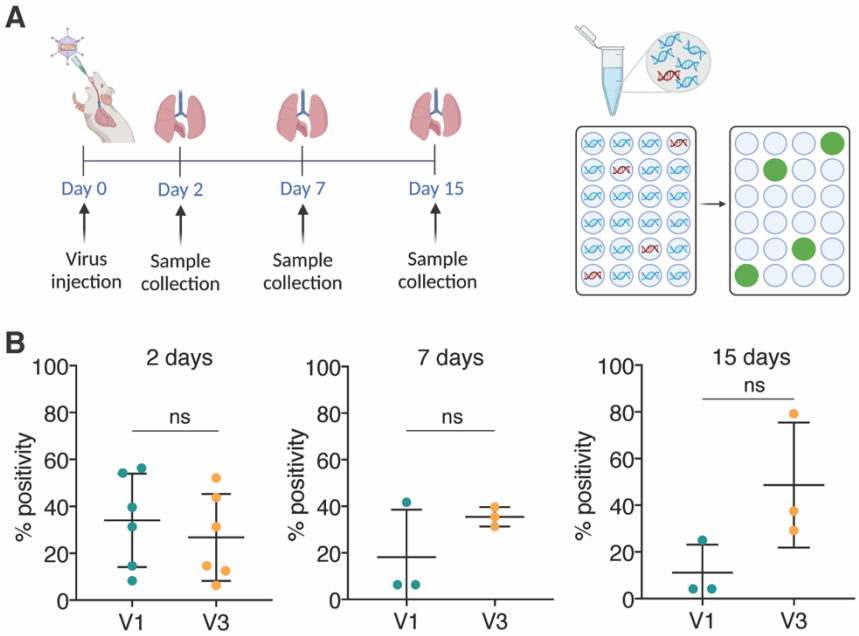

Figure 14. V1 and V3 adenoviruses show the same efficiency in generating the *Eml4-Alk* rearrangement *in vivo*.
(**A**) Schematic of the experimental design for quantifying the efficiency of V1 and V3 adenoviruses in generating the *Eml4-Alk* fusion gene (left) and representation of the digital-PCR-like technique employed to quantify *Eml4-Alk*. Blue DNA sequences represent normal DNA, whereas red sequences represent the Eml4-Alk sequences. The green circles represent the positive PCR reactions. (**B**) *Eml4-Alk* rearrangement 2, 7 and 15 days after virus instillation, measured as percentage of positive *Eml4-Alk* PCR reactions. 50 ng, 5 ng and 1ng of DNA were used per reaction for the 2 days (n=6), 7 days (n=3) and 15 days (n=3) samples, respectively. ns, not significant, p=0.5269 (2 days), p=0.2234 (7 days), p=0.0914 (15 days) Unpaired t test. SD was plotted.

1.6. Transcriptomic analysis of V1 and V3 tumors

Very little is known about the mechanisms driving the different phenotypes between V1 and V3 tumors. Therefore, together with Alberto Diaz, a PhD student in the

lab, we analyzed several tumor nodules by bulk RNA sequencing. We isolated single nodules and sequenced 8 V1 and 6 V3 tumors from independent mice. Principal component analysis (PCA) showed that while V1 tumors clustered together, V3 tumors had a more spread distribution; thus, the variance between V1 and V3 tumors was not very big (**Figure 15A**). Notably, unsupervised clustering of the samples based on the normalized counts demonstrated a high intertumoral heterogeneity in both variants, which resulted in the tumors clustering independently of the *Eml4-Alk* variant they harbored (**Figure 15B**). This heterogeneity observed could be due to the fact that entire tumor nodules were sequenced and not only sorted tumor cells. I therefore, deconvoluted the data according to mouse lung cell types, and examined what was the contribution of different immune cell types to each sample. I observed a heterogeneous composition of the immune compartment among the tumors with no variant-specific pattern (**Figure 15C**), reinforcing the idea that this could be the reason for the enhanced variability.

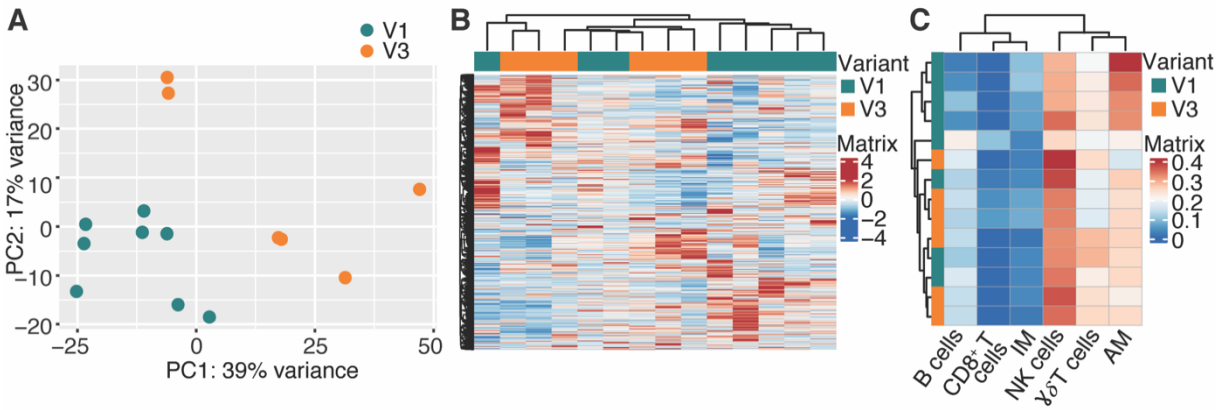

Figure 15. Analysis of the transcriptome of V1 and V3 tumors.
(**A**) Principal component analysis (PCA) plot showing the variance between V1 and V3 tumor samples. Each dot represents an individual tumor. (**B**) Heatmap depicting the clustering of the samples according to the normalized counts. Data scaled by row. (**C**) RNA deconvolution showing the immune cell composition of the samples based on the Aging Lung Atlas[188]. B cells, CD8+ T cells, interstitial macrophages (IM), natural killer (NK) cells, delta-gamma T cells and alveolar macrophages (AM) showed certain enrichment.

Nevertheless, 2,233 genes were found to be differentially expressed between V1 and V3 tumors (**Figure 16A and 16B**). *Alk* appeared to be one of the differentially expressed genes, showing an increase in V1 tumors (**Figure 16C**). To study the signaling pathways that could be involved in each of the variants, I performed Gene Set Enrichment Analysis (GSEA) using the differentially expressed genes. This analysis did not suggest very conclusive results since very few pathways or gene sets were significantly enriched. The only hallmark gene set enriched in V1 was allograft rejection, while V3 showed enrichment in the MYC targets and E2F targets hallmarks (**Figure 16D**

and 16E). These results did not help understand the mechanisms behind the variant phenotypes, although they might indicate a more aggressive phenotype in V3 tumors.

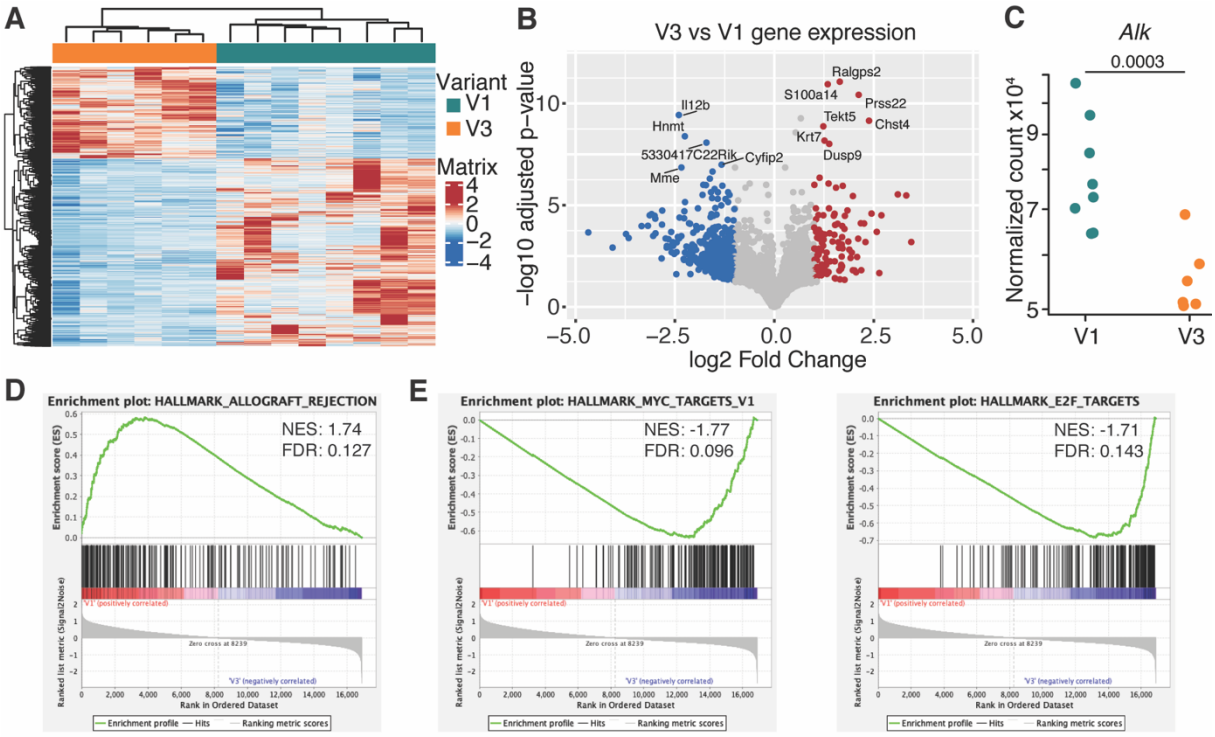

Figure 16. Analysis of the differentially expressed genes between V1 and V3 tumors.
(**A**) Heatmap displaying the expression of the differentially expressed genes between V1 and V3 tumors. Data scaled by row. Genes were considered significant when padj<0.05. (**B**) Volcano plot displaying the differentially expressed genes in V3 compared to V1 by log2 fold change and -log10 padj. Genes with log2 fold change <-1 or >1 are colored. Red represents upregulation in V3 and blue represents downregulation in V1. The names of the top genes are displayed. (**C**) Normalized counts (x10^4) of *Alk* in V1 and V3 tumors. Wald test. (**D, E**) Gene Set Enrichment Analysis of hallmarks enriched in V1 (D) or V3 (E) tumors. NES: normalized enrichment score. FDR: false discovery rate. For all, V1 n=8; V3 n=6.

2. Cell of origin of *Eml4-Alk* tumors

2.1. V3 tumors originate from club and AT2 cells

The cell of origin of LUAD has garnered significant attention and has emerged as a highly-discussed area of investigation in contemporary research, since studies in other cancer types suggest its importance in tumor development and response to therapy. Additionally, a different cell of origin of V1 and V3 tumors could be a major reason for the different overall survival rates. Dr. Yuanyuan Chen in our laboratory investigated the cell of origin of *Eml4-Alk* V1 tumors and discovered that club and AT2 cells can give rise to these tumors, but not ciliated, AT1 or basal cells[129]. To explore whether club and AT2 cells also contribute to V3 tumors, I employed linage tracing mouse models to label these cell types; *Scgb1a1*-CreER[85] for club cells and *Sftpc*-CreER[127] for AT2 cells.

Additionally, since I observed squamous tumors in V3 mice, I used the *Krt5*-CreER line[128] to label basal cells (**Figure 17A**), a well-known cell of origin of SCC[189]. Since AT1 and ciliated cells were not observed to form tumors in the V1 model or any other known model during the time of the project, the *Hopx*-CreER[190] (AT1 cells) and the *Foxj1*-CreER[126] (ciliated cells) mouse lines were not tested for V3. To be able to trace each cell type, all the models were combined with the mT/mG reporter mouse strain[175]. In these models, the expression of membranous tdTomato (mT) switches to the expression of membranous GFP (mG) by Cre-mediated recombination after tamoxifen (TAM) injection, labeling only the specified cell types.

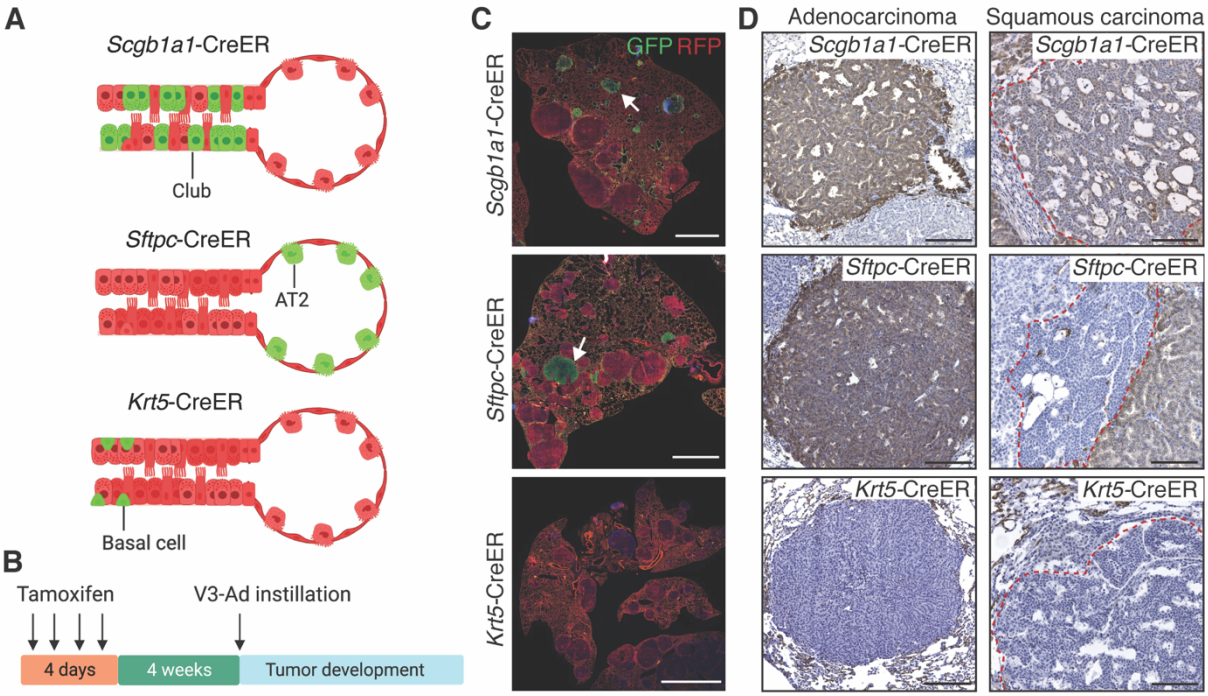

Figure 17. Club cells and alveolar type II cells give rise to V3 tumors.
(**A**) Schematic of the linage tracing mouse models used to label specific cell types with GFP after tamoxifen administration: *Scgb1a1*-CreER labels club cells in the airways, *Sftpc*-CreER labels AT2 cells in the alveoli and *Krt5*-CreER labels basal cells in the airways. (**B**) Schematic of the experimental design. Adult mice were treated with tamoxifen for 4 days and infected with the adenovirus (V3-Ad) 4 weeks later. (**C**) Immunofluorescent staining of GFP and RFP of V3 lungs in the indicated mouse lines. The arrows show examples of GFP⁺ tumors. Scale bar: 2 mm. (**D**) Immunohistochemical staining of GFP of V3 tumors in the different mouse models. Examples of adenocarcinoma (left) and squamous cell carcinoma (right).

I injected these mice with TAM for 4 consecutive days and waited 4 weeks before adenovirus instillation to allow complete labelling and metabolization of the TAM. Then, I waited for mice to develop tumors and euthanized the them at humane endpoint (**Figure 17B**). I observed that club and AT2 cells gave rise to V3 tumors, in line with previous results in V1, since GFP⁺ tumors were found in both models. Moreover, I found no GFP⁺ tumors in the *Krt5*-CreER model, suggesting that basal cells are not the cell of

origin of V3 tumors. Interestingly, none of the SCC observed in the *Krt5*-CreER or *Sftpc*-CreER mice were labeled with GFP, implying that SCC were not arising from basal cells nor from AT2 cells. In contrast, a few SCC were GFP-labeled in the *Scgb1a1*-CreER model, suggesting that club cells could be the cell of origin of these tumors (**Figure 17C**).

2.2. Club cells give rise to V1 and V3 tumors

Using the mouse models described above, I quantified the contribution of each cell type to the percentage of GFP⁺ tumors. I found that in both V1 and V3, ~50% of the tumors were positive for GFP in the *Scgb1a1*-CreER model, whereas ~38% and ~28% of the tumors were positive in the *Sftpc*-CreER mouse model in V1 and V3, respectively (**Figure 18A**). This difference between V1 and V3 in the *Sftpc*-CreER mouse model indicated that a third cell type might be involved in the initiation of V3 tumors, which had not been identified in V1 yet. The fact that in both V1 and V3, club cells are the main cell of origin suggested that the different aggressiveness between the variants is not due to differences in the cell of origin of these tumors. Additionally, no differences were observed in the tumor sizes between the different origins in either of the variants (**Figure 18B**).

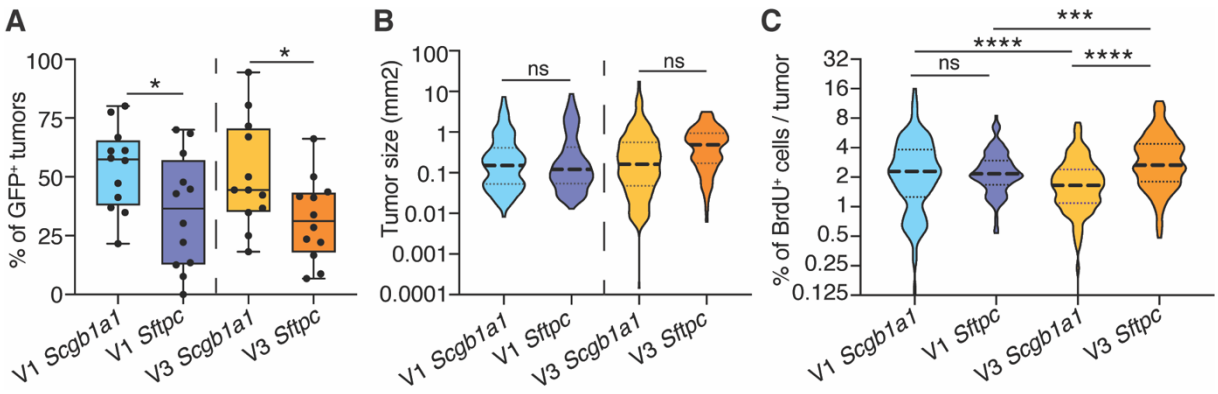

Figure 18. Club cells are the major cell type giving rise to both V1 and V3 tumors.
(**A**) Percentage of GFP⁺ tumors in the indicated mouse models infected with V1 or V3 adenovirus. n= 12 mice per group. *p<0.05, Unpaired t-test. (**B**) Size of GFP positive tumors in the indicated mice. V1 *Scgb1a1*-CreER n=232; V1 *Sftpc*-CreER n=216; V3 *Scgb1a1*-CreER n=350; V3 *Sftpc*-CreER n=244 ns (not significant), Unpaired t-test. (**C**) Proliferation of V1 and V3 GFP+ tumors with the indicated origins measured as percentage of BrdU⁺ cells per tumor. V1 *Scgb1a1*-CreER n=124; V1 *Sftpc*-CreER n=153; V3 *Scgb1a1*-CreER n=219; V3 *Sftpc*-CreER n=121. ***p<0.0005, ****p<0.0001, ns (not significant), Ordinary one-way ANOVA.

To measure the proliferation rate of these tumors, I injected the mice with Bromodeoxyuridine (BrdU), a nucleoside analog that is incorporated by proliferating cells during S phase, 2 hours prior to sacrificing them. Among V3 tumors, those originating from AT2 cells proliferate more than club originated tumors, whereas there

was no significant difference in V1 tumors. When comparing the proliferation rate between origins, among AT2 tumors, V3 proliferated more than V1, although I observed the opposite finding in club tumors, where V1 proliferated more than V3 (**Figure 18C**). These results suggest that the cell of origin might impact the proliferation rate of the tumors differently depending on the driver oncogene.

2.3. AT2-originated tumors induce a shorter survival of the mice

A big variability between the mice could be observed in the percentage of GFP⁺ tumors in both the *Sftpc*-CreER and *Scgb1a1*-CreER models. Therefore, to understand the aggressiveness related to the cell of origin, I stratified mice into groups with low or high percentage of GFP⁺ tumors depending on their deviation from the mean percentage in each condition (**Figure 19A**).

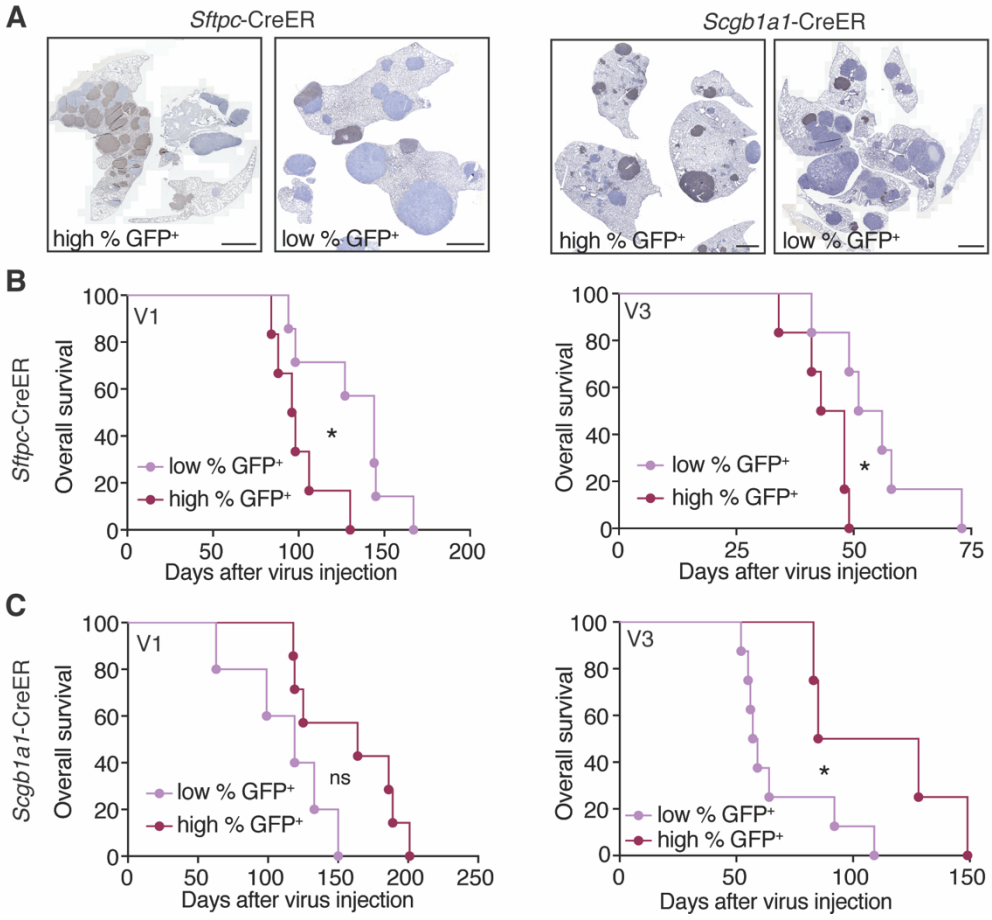

Figure 19. AT2-originated tumors are a risk factor for mice.
(**A**) Representative images of GFP immunohistochemical staining of *Scgb1a1*-CreER and *Sftpc*-CreER mice. Lungs with high and low percentage of GFP⁺ tumors are shown and were stratified according on their deviation from the mean percentage. (**B**) Overall survival of *Sftpc*-CreER V1 (left) and V3 (right) mice segregated into high and low percentage of GFP⁺ tumors. V1 high % n= 5; V1 low % n= 7; V3 high % n= 6; V3 low % n= 6. * p<0.05, Log-rank (Mantel-Cox) test. (**C**) Overall survival of *Scgb1a1*-CreER V1 (left) and V3 (right) mice segregated into high and low percentage of GFP⁺ tumors. V1 high % n= 7; V1 low % n= 5; V3 high % n= 4; V3 low % n= 8. * p<0.05, ns (not significant), Log-rank (Mantel-Cox) test.

In the *Sftpc*-CreER mouse model, the overall survival of those mice with a higher percentage of GFP$^+$ tumors, was significantly reduced in both V1 and V3 (**Figure 19B**), suggesting that the presence of more AT2 tumors is a risk factor for the mice. On the contrary, among the *Scgb1a1*-CreER mice, those with a higher percentage of GFP$^+$ tumors survived longer (**Figure 19C**), reinforcing the previous observation. These results suggest that AT2 tumors might be more aggressive than club tumors when compared among mice carrying the same oncogene.

2.4. Understanding the development of V1 club tumors

Part of the results described in this chapter have been already published: "Club cells employ regeneration mechanisms during lung tumorigenesis", to which I contributed as a co-first author[129]. Some figures and text have been adapted from the mentioned publication. The data presented here was a collective effort from Dr. Yuanyuan Chen, Dr. Reka Toth and several collaborators.

Results from our lab pointed out that club cells experience an epigenetic switch by which they transdifferentiate into AT2-like cells upon *Eml4-Alk* V1 transformation[129]. The general lack of knowledge on the molecular mechanisms driving club-originated tumors triggered us to analyze these tumors in detail. Therefore, together with Yuanyuan and Reka, we focused on the mechanisms underlying the development of V1 club-originated tumors. We performed a time course experiment using *Scgb1a1*-CreER mice (**Figure 20A**) in which we collected club cells (CD45$^-$, CD31$^-$, tdTomato$^-$, EpCam$^+$, GFP$^+$, CD24$^-$, ß4$^+$, CD200$^+$) from control mice (only TAM or TAM + Cas9-Ad), and from V1-Ad injected mice, 2 and 4 weeks after induction (**Figure 20B**). We also collected GFP$^+$ cells (CD45$^-$, CD31$^-$, tdTomato$^-$, EpCam$^+$, GFP$^+$) from tumors at humane endpoint and from a 4-weeks V1-Ad (4wk GFP) to exclude the contribution of the labelled AT2 cells to the observed phenotype (**Figure 20C**). We then performed 10x single-cell RNA sequencing (scRNA-seq) of all these cells.

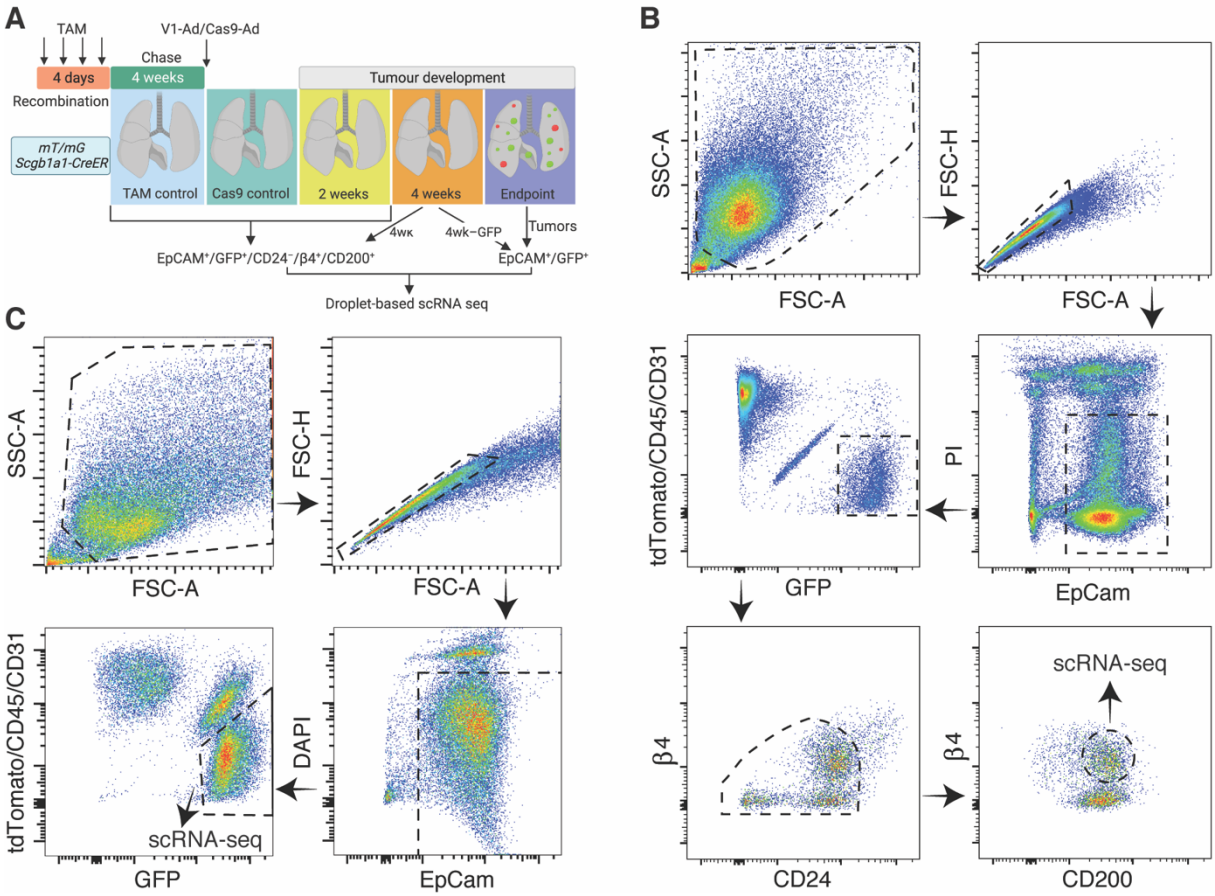

Figure 20. Time course single cell RNA-sequencing experiment of V1 club tumor development. (**A**) Schematic of the experimental design. *Scgb1a1-CreER* mice were treated with tamoxifen (TAM) for 4 consecutive days. Then Cas9 or V1 adenovirus (Ad) were intratracheally instilled 4 weeks after TAM. The indicated timepoints were collected and sorted as specified. (**B**) Sorting strategy to acquire GFP⁺ club cells from the TAM and Cas9 controls, 2 weeks and 4 weeks samples for further single-cell RNA-sequencing (scRNA-seq). (**C**) Sorting strategy to obtain total GFP⁺ cells from the 4 weeks-GFP and Tumor samples for further scRNA-seq. (**B, C**) The gates used to choose the subpopulations from the parental cells are shown by dashed lines. Arrows indicate the direction of the sorting strategy.

2.5. Different timepoints in tumor development show a distinct distribution of cells

We captured 29,277 high-quality cells, which we annotated according to their expression of well-known cell type-specific markers and signatures[104,105,120,188,191] (**Appendix I Table 18**). The majority of the cells expressed marker genes such as *Scgb1a1* or *Scgb3a2*, and were therefore annotated as club cells. All these cells arose from the controls, 2 weeks and 4 weeks samples, suggesting that these cells were normal cells (**Figure 21A and 21B**). Importantly, we found a group of cells mainly coming from the 4 weeks-GFP sample, in which GFP⁺ normal AT2 cells had not been excluded. These cells highly expressed AT2 markers and were annotated as AT2 cells. Interestingly, a small group of cells expressed markers of both AT1 and AT2 cells and were exclusively

found in the Tumor sample, and thus were entitled AT1/AT2-like tumor cells. Furthermore, we captured a recently described progenitor cell type, named H2-K1^{high} club-like progenitors[105], which express *H2-K1* and *AW112010* and were present mainly in the 2 weeks and 4 weeks samples. In addition, we found a group of proliferating cells, mostly present in the 2 weeks, 4 weeks, and tumor samples, and a group of cells present almost exclusively in the tumor sample. 14 independent clusters were discovered by unsupervised clustering (**Figure 21C**).

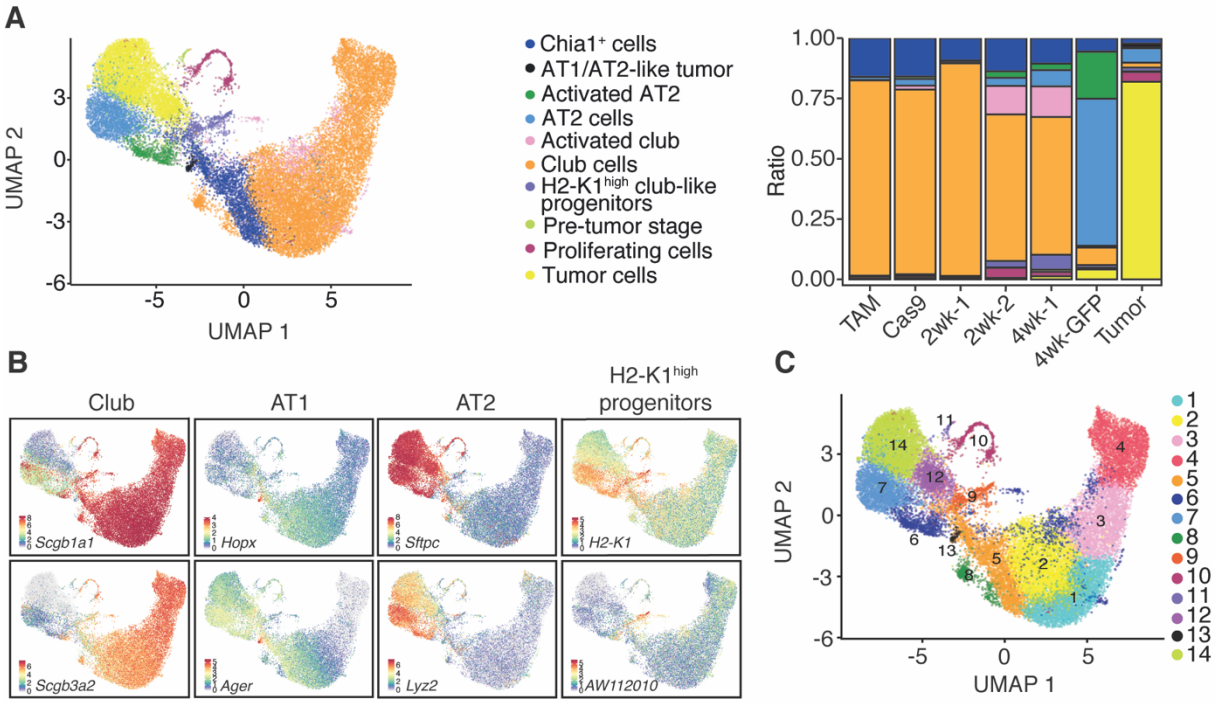

Figure 21. Cell type identification in the single cell RNA sequencing data.
(**A**) UMAP embedding colored by cell type annotation (right) and barplot showing the cell type composition of each sample (left). (**B**) UMAPs displaying expression levels of the most common cell type markers: *Scgb1a1*, *Scgb3a2* (club); *Hopx*, *Ager* (AT1); *Sftpc*, *Lyz2* (AT2) and *H2-k1*, *AW112010* (H2-k1^{high} club-like progenitors). Colors represent the normalized expression levels. (**C**) UMAP colored by clusters obtained by unsupervised nearest neighbor modularity optimization-based clustering.

We observed a distinctive distribution of the samples throughout cancer progression (**Figure 22A**), with the tumor sample clustering almost completely apart from the rest of the samples. Clusters 1 to 5 together with 8 were mostly formed by cells from the control and 2 weeks samples (22-75%) (**Figure 22B**), suggesting that these cells were most probably normal cells. Clusters 6 and 9-11 were primarily formed (>55%) by cells from the 2 weeks and 4 weeks samples. Moreover, most of the cells in cluster 7 (74%) appeared to be from the 4 weeks GFP sample, and the ones in clusters 12, 13 and 14 were mainly derived from the tumor sample (>80%).

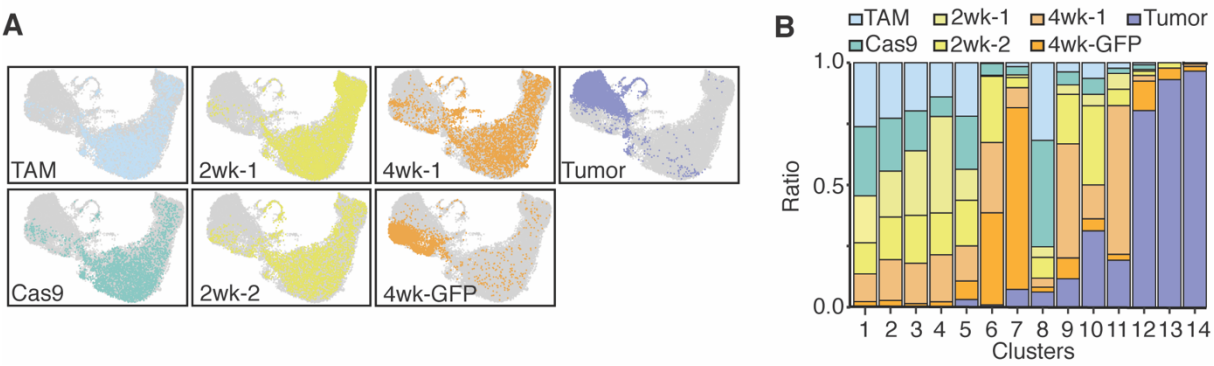

Figure 22. Distribution and composition of the samples.
(**A**) UMAPs presenting the distribution of cells from the different samples. (**B**) Sample composition of each individual cluster shown as relative contribution normalized to the total number of cells in each cluster.

2.6. Club cells lose their club signature and gain and AT2-like signature upon transformation

We calculated signature scores for the different clusters and found that cluster 1-5 and 8 showed high and low scores of the club and AT2 signatures, respectively (**Figure 23A**), reinforcing that these cells were most probably normal club cells. Nevertheless, cluster 4 also displayed enrichment in the goblet signature, indicating that cells from this cluster were most probably goblet-like club cells. Cluster 7 denoted the highest score in the AT2 signature, confirming that these cells were the normal AT2 cells not excluded in the sorting.

Interestingly, we identified a tumor signature that was mainly enriched in clusters 12 and 14 (**Figure 23B and 23C**). To prove that the cells denominated as tumor cells were indeed tumor cells, we examined the expression levels of *Alk* in the different clusters and discovered that, besides the tumor clusters 12, 13 and 14, cluster 11 also expressed *Alk* (**Figure 23D**). Hence, we further investigated cluster 11 and found that it presented high levels of certain markers of basal cells, such as *Trp63* (**Figure 23E**), in addition to higher levels of the club signature compared to the tumor clusters, yet lower than normal club cell clusters (**Figure 23A**). Besides, cluster 11 did not display very high levels of the tumor markers, suggesting this cluster as a pre-tumor stage cluster. Altogether, our data nicely recapitulated the developmental process of V1 club tumors and proved a gradual loss of club cell identity and a boost of AT2 markers.

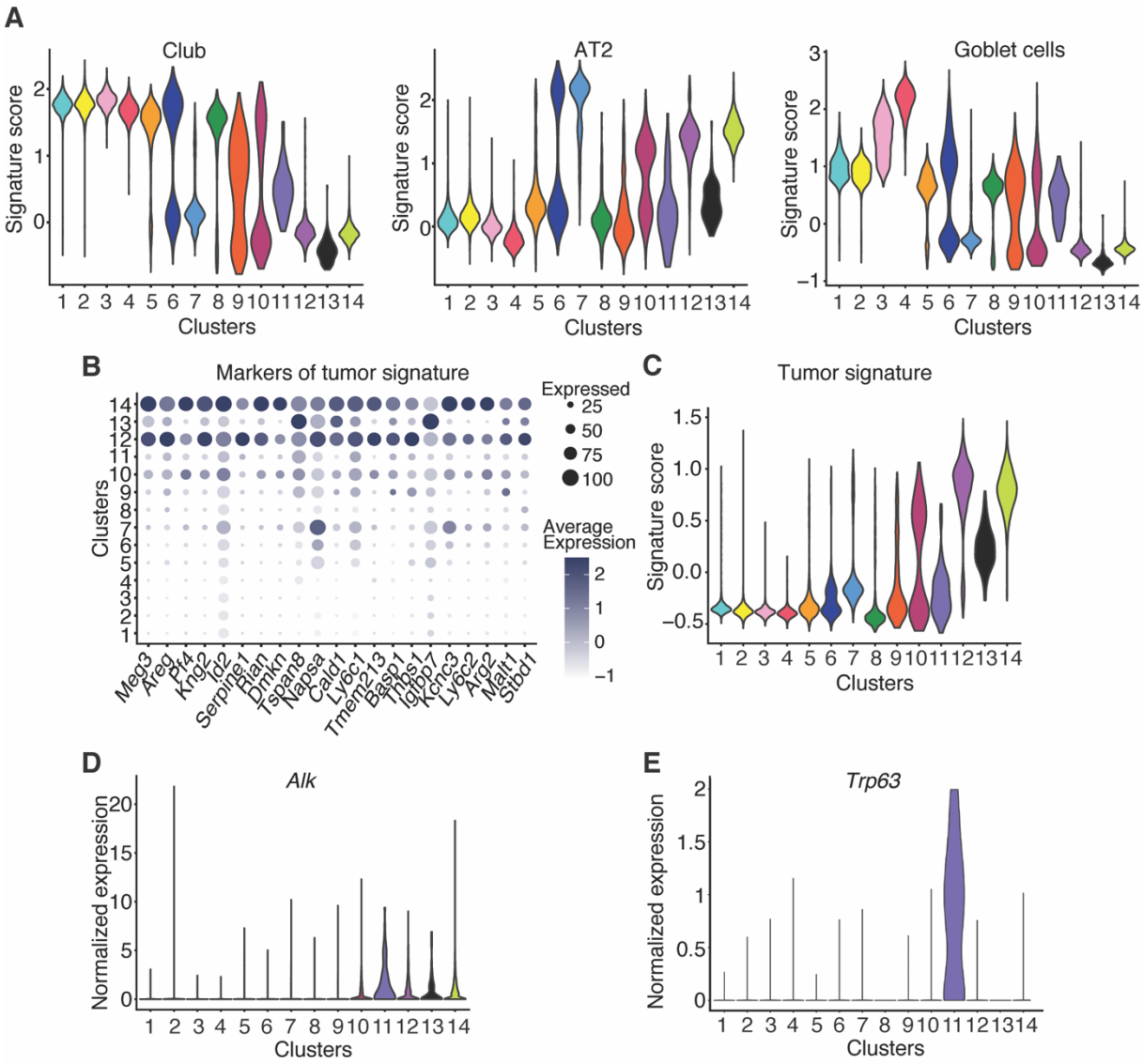

Figure 23. Club cells acquire an AT2-like identity during tumor development.
(**A**) Violin plots depicting the signature scores of club cell (left), AT2 cell (middle) and goblet cell signatures by the described clusters. The signatures were obtained from the study of Strunz et al[104]. (**B**) Dot plot representing the expression of the markers in the tumor signature for each of the identified clusters. The size of the dots represents the percentage of cells that express the gene and the color displays the average expression of the gene. (**C**) Violin plot showing the score of the tumor signature for each cluster. (**D, E**) Violin plots representing the normalized expression of *Alk* (**D**) and *Trp63* (**E**) for each cluster.

2.7. Club cells employ their intrinsic regeneration mechanisms to develop a tumor

To better comprehend the transitional cell states during tumor development, we employed consensus non-negative matrix factorization (cNMF) to identify gene expression activity programs. We could distinguish nine independent activity programs (**Figure 24A**), two of them presenting features of normal club and AT2 cells, and another one containing markers of activated club cells, which was mainly represented

by the goblet-like club cells (cluster 4). In addition, we depicted two immune-related programs, which we annotated as the interferon program and the immune activation/inflammation program. These two programs were exclusively correlating with clusters 6 and 9 (**Figure 24B**), the second being the H2-K1^{high} club-like progenitors. This finding was not surprising since H2-K1^{high} club-like progenitors have been shown to have an immune responsive phenotype to repopulate the lung upon injury[105].

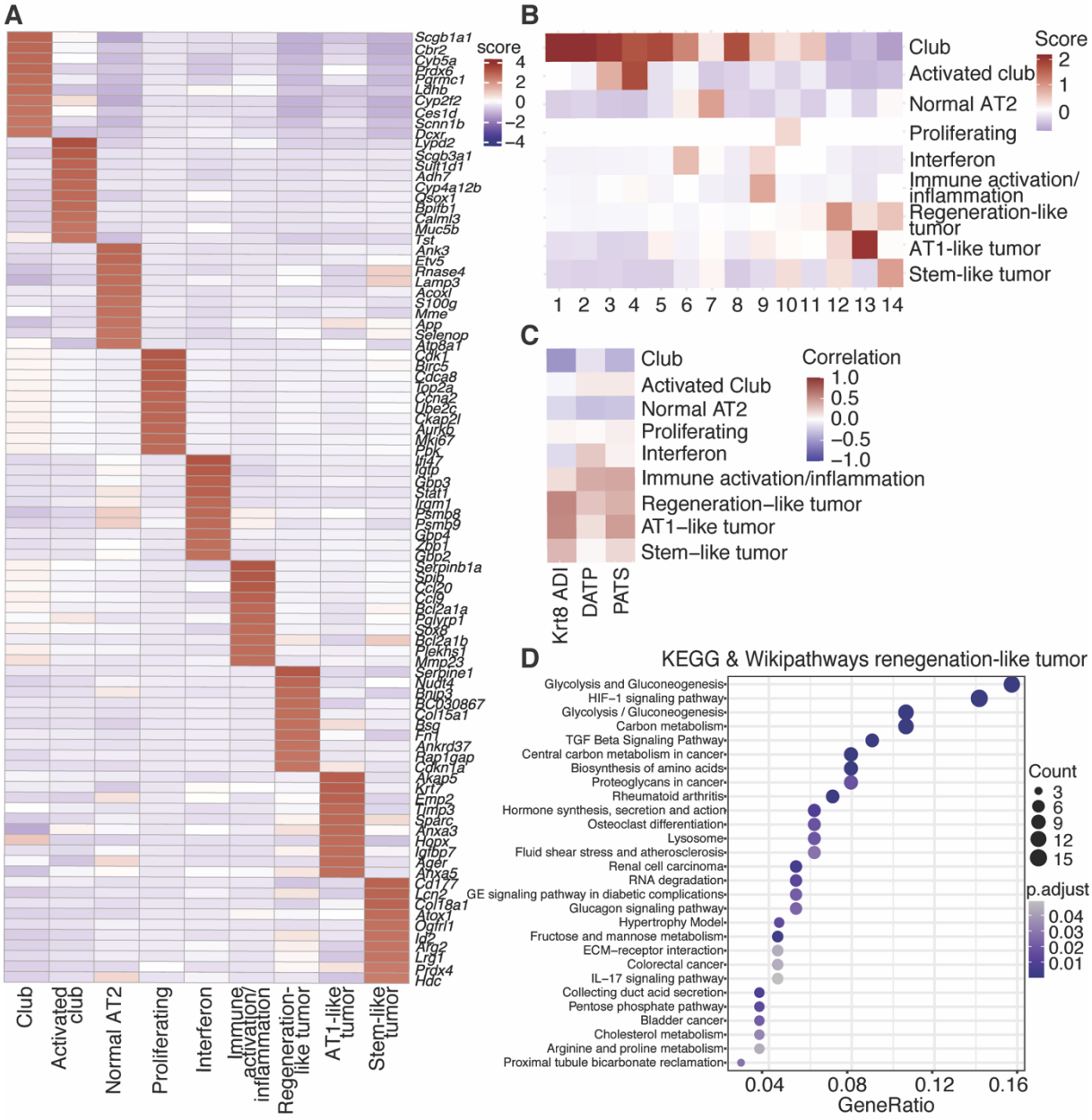

Figure 24. Tumor cells show enhancement in three independent activity programs.
(**A**) Heatmap depicting the top 10 genes involved in each identified activity program by cNMF. (**B**) Heatmap showing the module score of each cluster for every activity program. The top 10 genes of each program were used to calculate the scores. (**C**) Heatmap showing the correlation between the 9 activity programs and the published signatures for Krt8 ADI[104], DATP[107] and PATS[106] cell states. (**D**) Dot plot showing the pathway enrichment analysis based on KEGG and WikiPathways for the regeneration-like tumor program. The number of genes within the top 200 genes is represented by the size of the dots shows. The level of significance is shown by the color.

Surprisingly, we discovered three programs that were exclusive to the tumor clusters, and each one was mostly represented by one of the tumor clusters. Among the two main tumor clusters, cluster 12 displayed the highest enrichment in a program characterized by the expression of markers genes of transient cellular states that have been described in lung regeneration (Krt8 ADI, DATP and PATS[104,106,107]). Moreover, this program showed high activity in HIF1 and IL-17 signaling, TGF-β pathway and metabolic pathways such as glycolysis; pathways proven to be active in these transitional states, thus we annotated it as regeneration-like tumor program (**Figure 24C and 24D**). Another program showed features of AT1 cells and was mainly composed of cells from cluster 13, although cluster 12 also showed certain enrichment in this program. Similar to the regeneration-like tumor program, our AT1-like tumor program showed a positive correlation to regeneration signatures. The last tumor program that we identified was mostly represented by cluster 14, and was named as stem-like tumor program due to the enhanced expression of stem cell markers such as *Id2* and *Sox9* together with AT2 markers such as *Sftpc*, *Lamp3* and *Lcn2* (**Appendix I Table 19**). Altogether, these results suggested that club-originated tumors resemble lung regeneration signatures.

2.8. Club cells can undergo several routes during tumor development

Next, to unveil the developmental trajectory of normal club cells towards tumorigenic cells, we employed RNA velocity and partition-based graph abstraction (PAGA) analysis. RNA velocity foretells the cell's fate by assessing the ratio of spliced and unspliced RNA[192] (**Figure 25A**). On the other hand, PAGA creates a map with weighted edges connecting the nodes exemplifying the connectedness between clusters[193] (**Figure 25B**). The combination of both allowed us to direct the edges generated by PAGA[183], and therefore to understand the cell trajectories in an unbiased way. Since cluster 6 seemed to be composed of two different cell types, activated club and activated AT2 (**Figure 21A and Figure 23A**), we split it into clusters 6a and 6b to obtain a cleaner trajectory analysis.

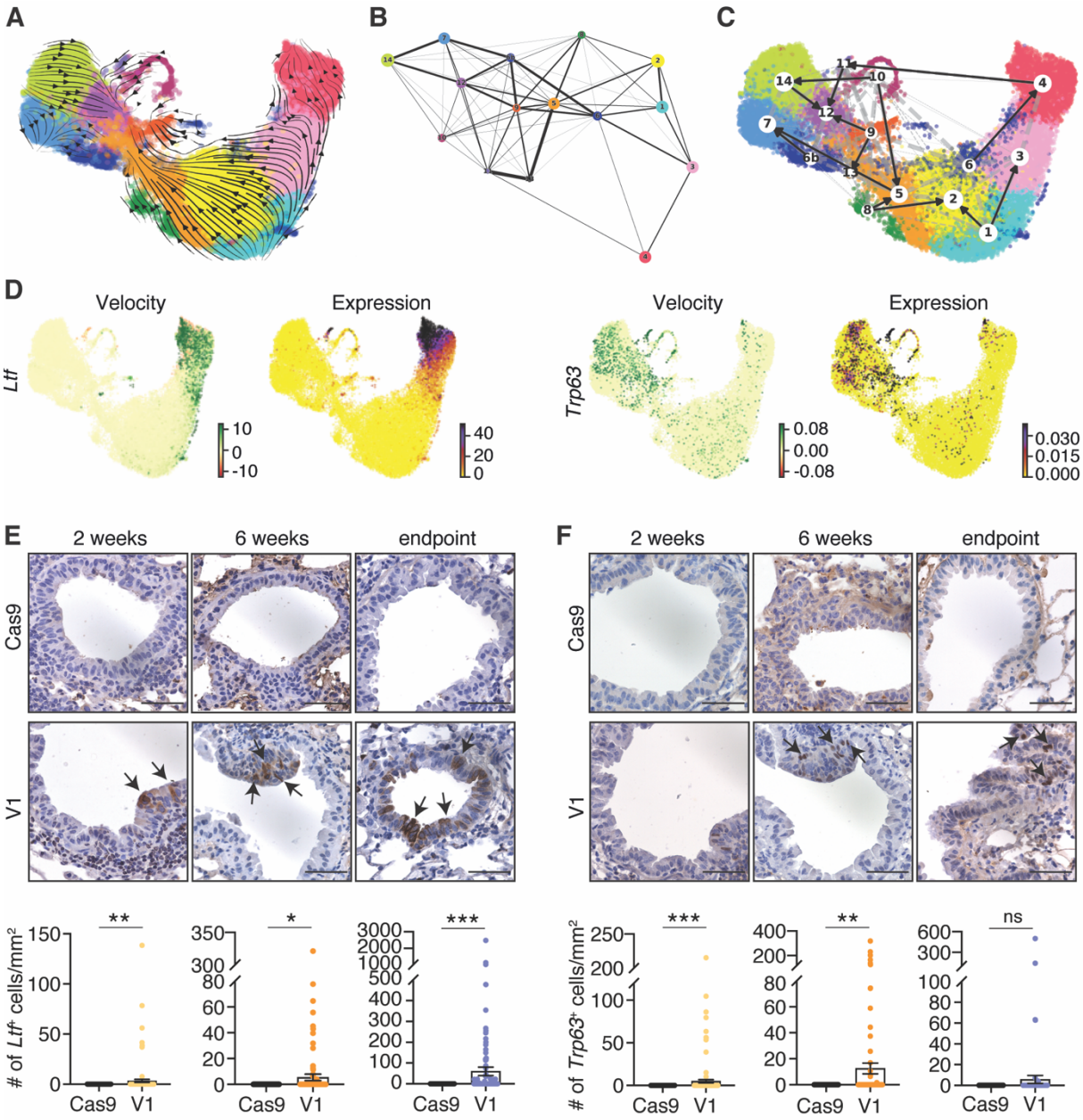

Figure 25. Club cells can follow distinct routes towards lung adenocarcinoma.
(**A**) RNA velocity analysis embedded into the UMAP and showing the direction of cellular transitions. Colors represent the original clusters. (**B**) Partition-based graph abstraction (PAGA) displaying the connectivity between clusters. Clusters are connected by weighted edges which indicate the strength of the connection. (**C**) UMAP showing the combined RNA velocity and PAGA analysis. RNA velocity was used to determine the direction of the arrows, whereas PAGA was used to create connections between clusters. (**D**) RNA velocity and gene expression levels of *Ltf* (left) and *Trp63* (right). In the RNA velocity plots, the green color represents high velocity, thus the gene will be upregulated, whereas red denotes the gene will get downregulated. (**E, F**) Immunohistochemistry of *Ltf* (**E**) and *Trp63* (**F**) in Cas9 (control) and V1 mice at different timepoints. Scale bar 50μm. Below, the corresponding quantifications represented as number (#) of positive cells per mm^2. Each dot represents a bronchus. *p<0.05, **p<0.005, ***p<0.0005, ns = not significant. Mann–Whitney test. The SEM is displayed. n = 3 mice per group except Cas9 2 weeks and V1 4 weeks (n = 4) and Cas9 endpoint (n = 2).

Careful examination of the results uncovered a complex transition from normal to tumor cells. One of the most striking routes towards tumor cells was originated from cluster 4, composed of goblet-like club cells, which then transitioned to cluster 11, the

pre-tumor stage, before it developed into the tumor cluster 12 and partially 13 (**Figure 25C**). This route was marked by the early expression of goblet markers such as *Ltf, Bpifa1, Bpifb1* and *Reg3g* in cluster 4 followed by cluster 11 which maintained the expression of *Ltf*, and gained expression of basal-like markers such as *Trp63* (**Figure 25D**). This transition mimics a recently described cell state known as the goblet-basal intermediate state, which appears upon lung regeneration[194], but had never been described to arise during lung tumor development. In addition, this trajectory and the clusters involved in it had a strong correlation to the regeneration-like program. Moreover, clusters 12 and 13 could also be originated from progenitor-like cells in cluster 9, whereas cluster 14 seemed to arise exclusively from proliferating cells in cluster 10, and could further transition into cluster 12.

To validate our findings, we stained lung tissues from control mice (Cas9-Ad), and V1 mice at early (2 weeks), mid (6 weeks), and late stages of tumor development (endpoint) for *Ltf* and *Trp63* and observed that indeed, these markers were present in early lesions in the bronchi of V1 mice but not in Cas9 mice (**Figure 25E and 25F**). Altogether, our results emphasize the complexity of club originated tumors and reinforce the involvement of intrinsic lung regeneration signals and pathways in this process.

To verify that our findings were not general tumor signatures, we integrated our 10x scRNA-seq data with 10x data from *Kras* mutant LUAD samples from Yang et al.[195], and also compared our signatures to those described by Marjanovic et al., both obtained from *Kras* mutant AT2 originated tumors. From the integration with Yang's et al. data, we could confirm that our samples clustered independently to AT2 originated tumors (**Figure 26A and 26B**) and that except for our cluster 13 (AT1-like tumor), which had certain similarity to their signatures, all of our clusters appeared to have no correlation to Yang's signatures (**Figure 26C**). Additionally, we observed that our AT1-like tumor program correlated to a certain extent with the AT1/AT2 signature described by Marjanovic et al. Similarly, the Stem-like tumor program showed correlation to their Embryonic liver-like signature. Altogether, these results suggested that club and AT2 tumors may converge in similar states during tumor evolution (**Figure 26D**).

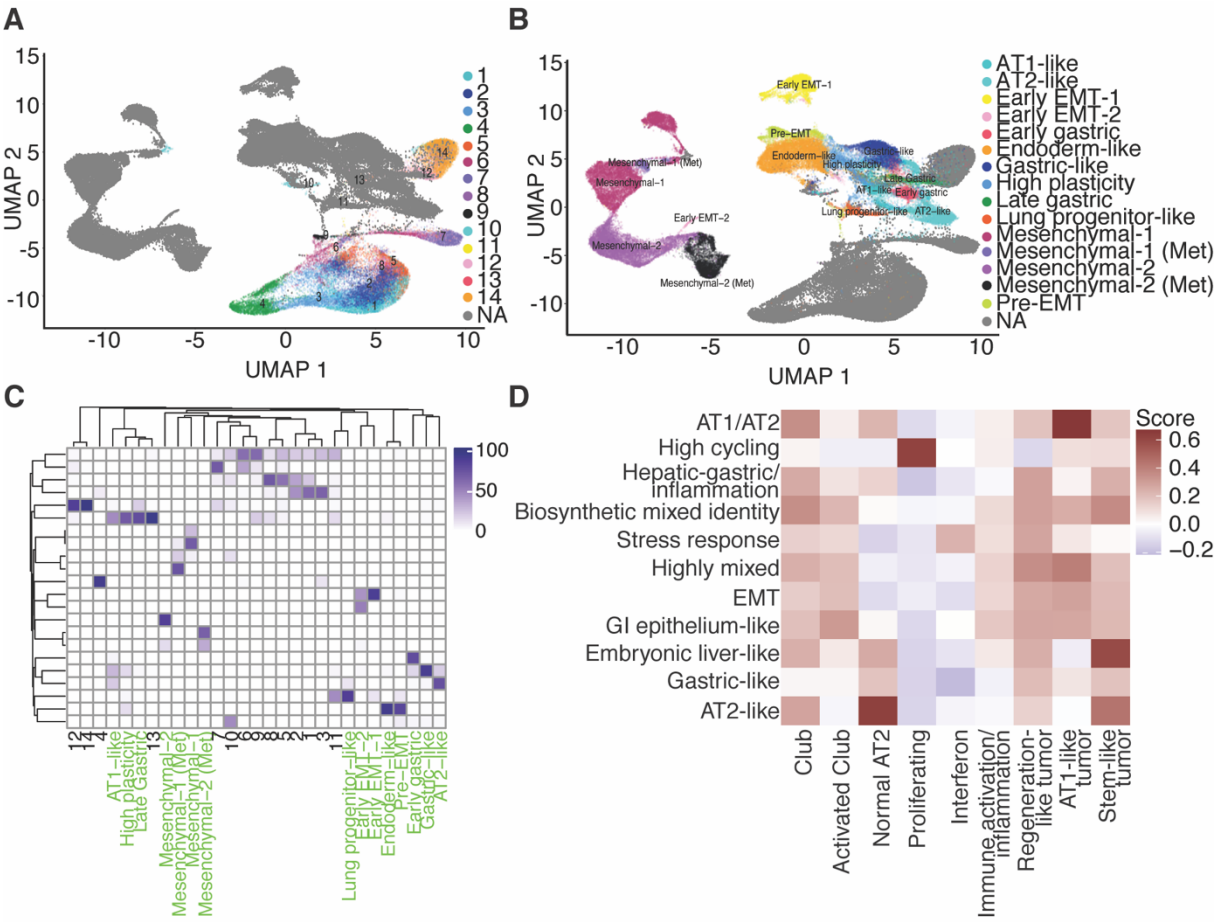

Figure 26. Comparison of club originated tumors to published AT2 tumors.
(**A, B**) UMAP embedding depicting the integration of our data and the AT2 tumor data from Yang et al., 2022[195]. (**A**) The cells are colored based on our clusters or (**B**) the clusters from Yang et al. Gray was used to represent cell from the other study. (**C**) Correlation-based heatmap comparing the integration clusters to our original clusters and the clusters from Yang et al. The rows represent the integrated clusters and columns are the original clusters from both studies. Clusters from the Yang et al. study are colored in green. (**D**) Heatmap showing the correlation score between our activity programs (x axis) and the cNMF modules identified in the Marjanovic et al. 2020[120] study (y axis).

Finally, to investigate the human relevance of our signatures, we compared our tumor programs with a published scRNA-seq dataset of human LUAD[196], and found that the three programs were present in human lung tumors (**Figure 27A**). Additionally, we made use of the RNAseq and survival data of LUAD patients from The Cancer Genome Atlas (TCGA) database[197] and calculated the hazard ratio of our identified activity programs. We observed that our three tumor programs were indicative of poor prognosis, with the regeneration-like tumor program being the most pronounced one (**Figure 27B**).

Altogether, our results suggest that the regeneration mechanisms imprinted in club cells drive tumor development in these cells upon transformation, yet different routes can be followed towards the endpoint tumors. Moreover, we described that one

of the main trajectories followed initiates from goblet-like club cells, goes through a pre-tumor stage that mirrors goblet-basal intermediate cells, and ends up developing into heterogenous tumors (**Figure 27C**).

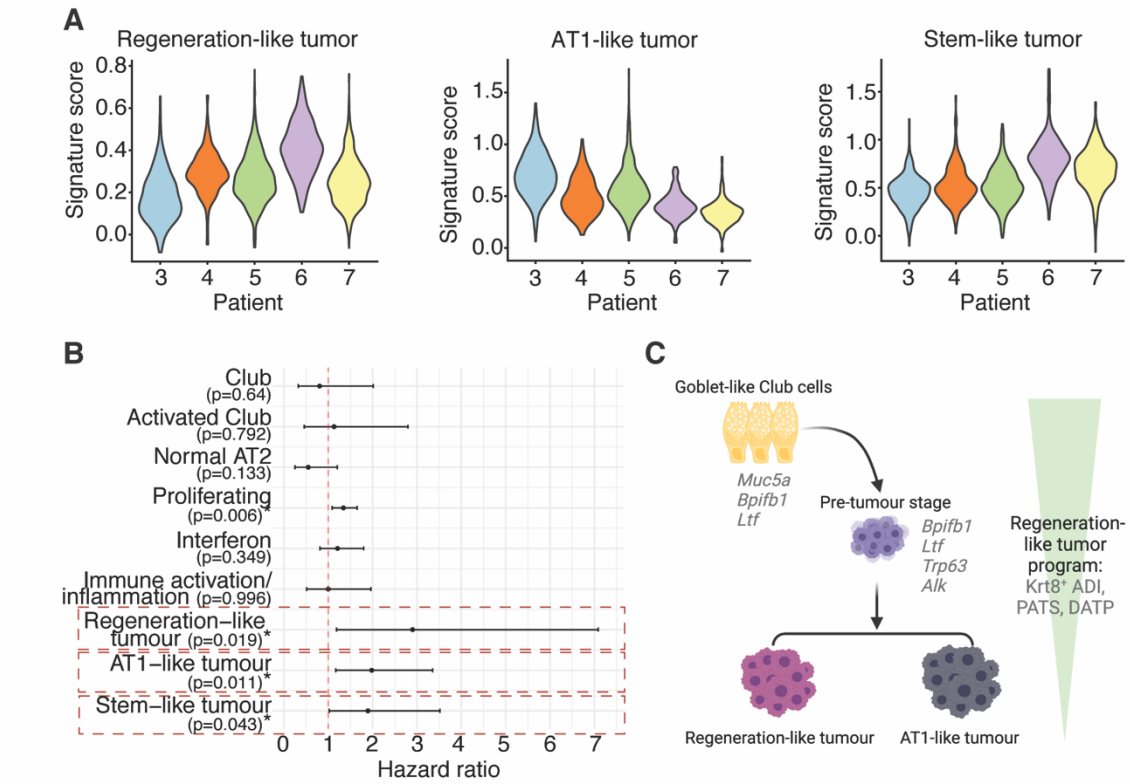

Figure 27. Club tumor signatures are present in human lung adenocarcinoma.
(**A**) Violin plots showing the score for our identified tumor programs in human LUAD patients from a single cell RNA sequencing dataset[196]. (**B**) Forest plot displaying the hazard ratio of the 9 activity programs based on a Cox proportional hazards model in TCGA LUAD patients (n=533).

3. Tumor initiation and role of Tp53 in *Eml4-Alk* tumors

3.1. V1 cells undergo apoptosis early upon transformation

Once I had investigated the cell of origin of V1 tumors and since I had already seen that V3 tumors also originate from club cells, similar to V1 I still wondered why there was such a significant difference between the survival of V1 and V3 mice. I hypothesized that there could be a difference in tumor initiation between the variants, which resulted in the fast and early death of V3 mice. Therefore, I examined samples 7 days after virus injection by immunohistochemical analysis and measured the proliferation rate and the levels of apoptosis and DNA damage in the cells. I observed no differences in the number of proliferating cells between the two variants (**Figure 28A**), whereas V1 presented a significant increase in the number of apoptotic cells, measured

by cleaved-caspase 3 (Casp3) (**Figure 28B**), and a higher number of cells positive for γ-H2AX, a well-known marker of DNA damage (**Figure 28C**).

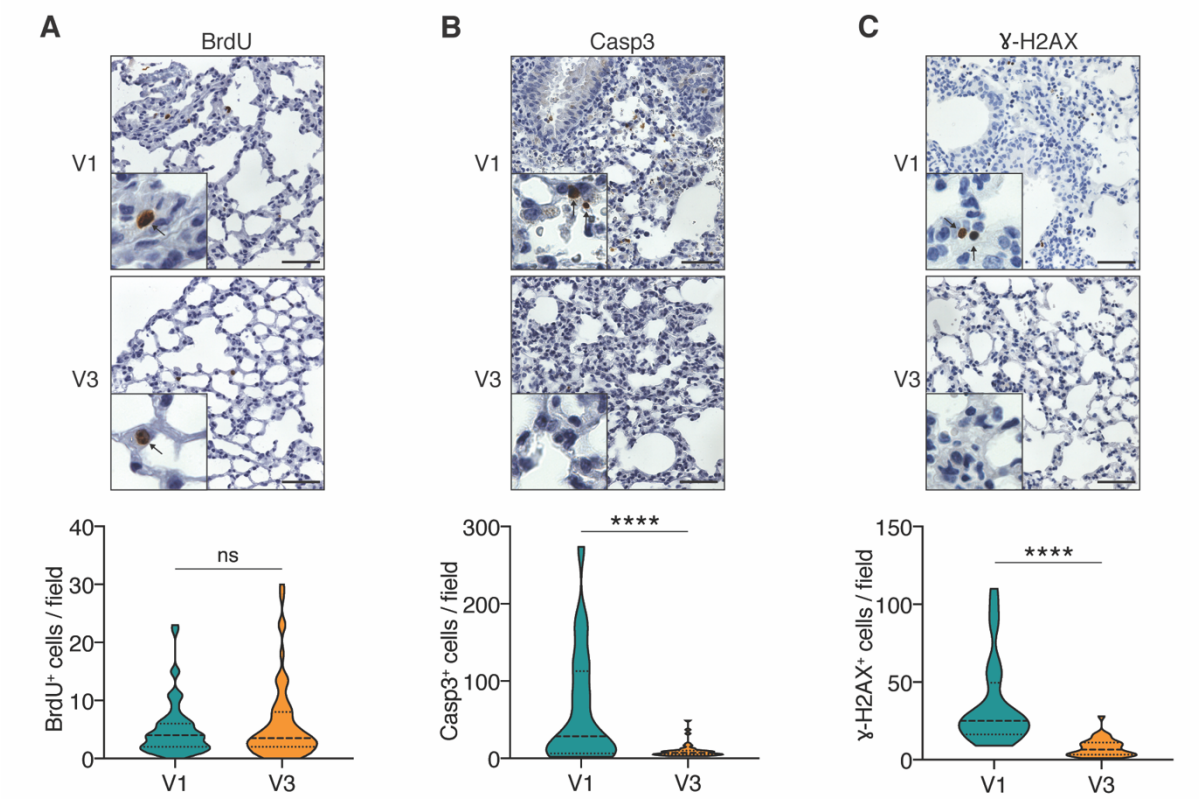

Figure 28. V1 induction results in early DNA damage and apoptosis.
Immunohistochemistry images and quantification of proliferation (**A**), apoptosis (**B**) and DNA damage (**C**) in lung cells represented as number of positive cells per field analyzed for BrdU, Casp3 and γ-H2AX, respectively. Mice were sampled 7 days after V1 or V3 adenovirus instillation. 16 fields of 1mm^2 were analyzed per animal. n=3 per group. ns, not significant, p=0.3229 (**A**) and p<0.0001 (**B, C**), Unpaired t test. Scale bar for all 50 μm.

3.2. Tp53 loss is a poor prognostic factor for V1 mice

Trp53 has been described to be the most commonly mutated gene in EML4-ALK patients, being more prevalent in long variants, such as V1, than short variants, like V3[162]. One of the triggers of the activation of the Tp53 pathway is DNA damage, resulting in the induction of apoptosis. Therefore, I hypothesized that the generation of V1 might result in early activation of Tp53 and cell death, which may result in delayed tumor initiation. In addition, it has been described that senescence in AT2 cells can be induced through the activation of the Tp53-p21-Rb pathway in fibrotic lung disease[198]. To elucidate whether the p53 pathway activation is responsible for the delay of V1 tumors and whether this is cell type dependent, I induced V1 and V3 in a *Trp53* conditional background (*Trp53*floxed mice, *Trp53*$^{fl/fl}$), together with linage-tracing mouse lines

(*Scgb1a1*-CreER;mT/mG and *Sftpc*-CreER;mT/mG). In these mice, TAM injection not only labels the specific cells with GFP but also deletes the *Trp53* gene.

Analysis of these mice revealed a reduction in the survival of V1 *Trp53*$^{fl/fl}$ mice in both the *Sftpc*-CreER and the *Scgb1a1*-CreER lines compared to *Trp53* wild type (*Trp53*$^{+/+}$) mice (**Figure 29A and 29B**). These results suggest a strong cooperation between the loss of *Trp53* in combination with the V1 rearrangement, regardless of the cell type. In contrast, the survival of V3 mice was not affected by the loss of *Trp53* in either of the cell types (**Figure 29C and 29D**). Since Cre expression labels with GFP and deletes *Trp53* in the same cell, I employed the GFP positive staining to distinguish *Trp53*-null tumors (*Trp53*$^{\Delta/\Delta}$). The survival results were further confirmed by an increase in the percentage of V1 GFP^{+} tumors in both mouse lines upon the deletion of *Trp53*, suggesting that the loss of *Trp53* may facilitate the initiation of V1 tumors, whereas it does not affect the tumor initiation of V3 (**Figure 29E and 29F**).

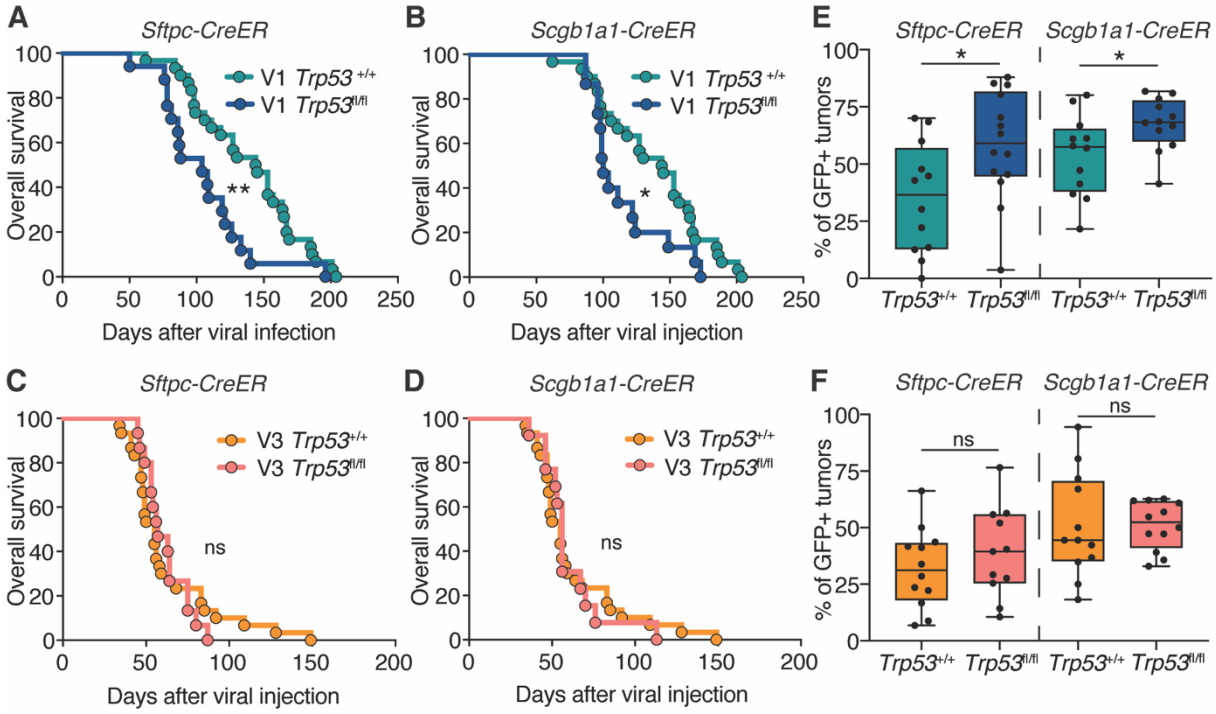

Figure 29. *Trp53* loss in AT2 cells or club cells reduces the survival or V1 mice but not V3 mice. (**A, B**) Overall survival of V1; *Trp53*$^{+/+}$ compared to *Trp53*$^{fl/fl}$ *Sftpc-CreER* mice (A) or *Trp53*$^{fl/fl}$ *Scgb1a1-CreER* mice. *Trp53*$^{+/+}$ n=30; *Trp53*$^{fl/fl}$ *Sftpc-CreER* n=16; *Trp53*$^{fl/fl}$ *Scgb1a1-CreER* n=12. **p=0.0231, *p=0.0104 Log-rank (Mantel-Cox) test. (**C, D**) Overall survival of V3 *Trp53*$^{+/+}$ compared to *Trp53*$^{fl/fl}$ *Sftpc-CreER* mice (A) or *Trp53*$^{fl/fl}$ *Scgb1a1-CreER* mice. *Trp53*$^{+/+}$ n=30; *Trp53*$^{fl/fl}$ *Sftpc-CreER* n=13; *Trp53*$^{fl/fl}$ *Scgb1a1-CreER* n=13. Not significant (ns) Log-rank (Mantel-Cox) test. (**E, F**) Percentage of GFP^{+} tumors in *Trp53*$^{+/+}$ and *Trp53*$^{fl/fl}$ *Sftpc-CreER* mice and *Scgb1a1-CreER* V1 mice (E) or V3 mice (F). *p<0.05, ns (not significant); Unpaired t-test.

3.3. *Trp53* loss affects V1 and V3 tumors differently according to their origin.

Tp53 is well known for its implications in cell cycle regulation and cell proliferation[149]. Therefore, to understand whether the loss of *Trp53* results in an increase in the proliferation rate of the tumors, I measured the number of BrdU positive cells in the GFP$^+$ tumors (*Trp53*$^{\Delta/\Delta}$) of the independent mouse models. I observed that V1 *Trp53*$^{\Delta/\Delta}$ tumors originating from AT2 cells showed an increase in proliferation compared to their *Trp53*$^{+/+}$ control, while no changes were observed in club originated tumors (**Figure 30A and 30B**). This result suggested that Tp53 most likely plays a role in the proliferation of V1 AT2 tumors, but not in those arising from club cells. On the contrary, although *Trp53* loss did not affect the proliferation of V3 AT2 tumors, nodules originating from club cells displayed an increase in their proliferation rate upon *Trp53* loss (**Figure 30C and 30D**).

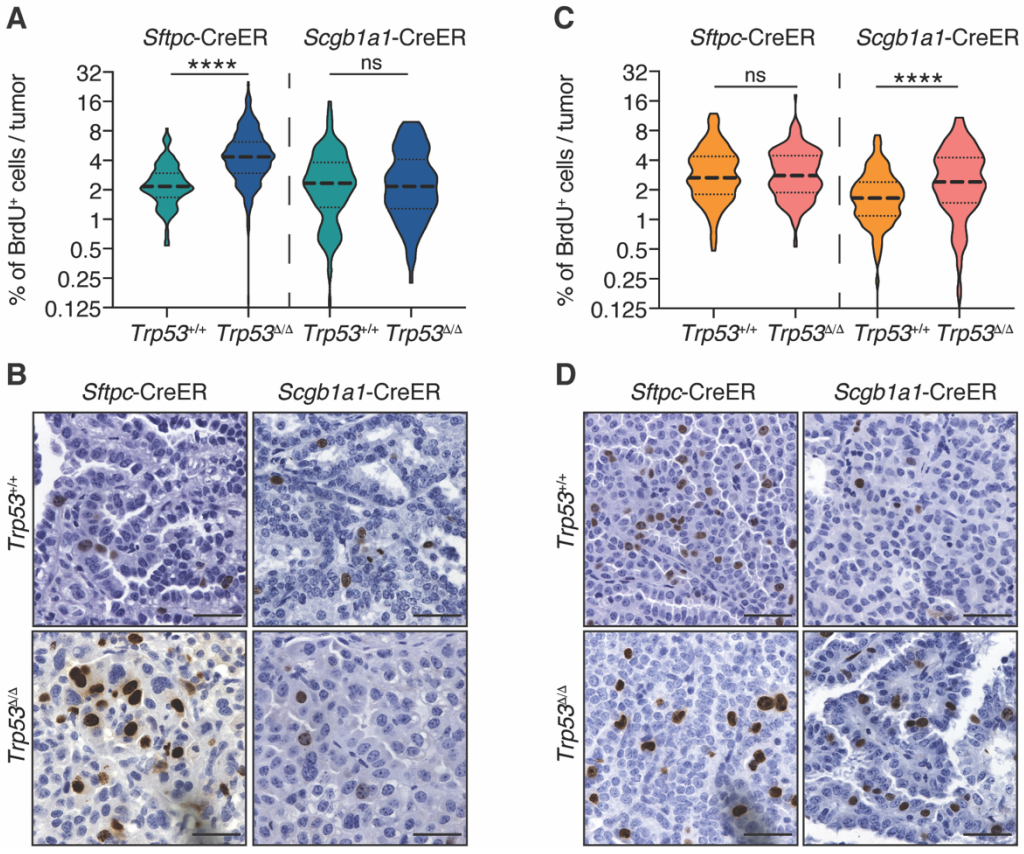

Figure 30. *Trp53* loss increases the proliferation rate of V1 AT2 tumors and V3 club tumors. (**A**) Quantification of the proliferation rate of V1 tumors in the indicated mouse models. Only GFP positive tumors were considered for both Trp53 wildtype (*Trp53*$^{+/+}$) and Trp53 null (*Trp53*$^{\Delta/\Delta}$) tumors. (**B**) Representative images of BrdU immunohistochemistry in V1 tumors. Scale bar: 40 μm. (**C**) Quantification of the proliferation rate of V3 tumors in the indicated mouse models. Only GFP positive tumors were considered for both Trp53 wildtype (*Trp53*$^{+/+}$) and Trp53 null (*Trp53*$^{\Delta/\Delta}$) tumors. (**D**) Representative images of BrdU immunohistochemistry in V3 tumors. Scale bar: 40 μm. (**A, C**) ****p<0.0001, ns = not significant, Unpaired t-test.

Interestingly, I observed that the loss of *Trp53* in club cells resulted in an increase in the number of squamous tumors in both variants (**Table 17**). Among the *Scgb1a1*-CreER mice, 38% of the V1 *Trp53*$^{fl/fl}$ harbored at least one SCC (5 out of 13 mice), compared to 0% in the *Trp53*$^{+/+}$ mice (0 out of 12 mice). Similarly, 46% of the V3 *Trp53*$^{fl/fl}$ mice contained at least one SCC (6 out of 13 mice), while only 21% of the *Trp53*$^{+/+}$ mice did (3 out of 14 mice). Altogether, these results suggest that the loss of *Trp53* in club cells facilitates the development of *Eml4-Alk* squamous cell carcinomas in both V1 and V3.

Table 17. Percentage of mice with at least one SCC in the indicated mouse lines.

Mouse line	V1 *Trp53*$^{+/+}$	V1 *Trp53*$^{fl/fl}$	V3 *Trp53*$^{+/+}$	V3 *Trp53*$^{fl/fl}$
FVB	7%		23%	
***Scgb1a1*-CreER**	0%	38%	21%	46%
***Sftpc*-CreER**	0%	7%	5%	18%

3.4. Single cell RNA sequencing of early timepoints

To unveil the mechanisms that were driving the early DNA damage and death in V1 cells, I analyzed samples at early stages by scRNA-seq. The induction of the *Eml4-Alk* rearrangement by adenoviral infection, does not allow the identification of positive cells and hence, these cells cannot be detected by sorting. Therefore, I used a lentiviral system to be able to enrich for the population of infected cells, which should contain the *Eml4-Alk* rearrangement. This system was designed by Alberto Diaz and Dr. Kalman Somogyi in the lab, for which they cloned the original sgRNAs targeting Eml4 and Alk into lentiviral plasmids. In addition, these plasmids contain Cre recombinase, that after injection into Rosa26-Lox-Stop-Lox-Cas9-EGFP (LSL-Cas9-EGFP) mice, results in the expression of Cas9 and EGFP (**Figure 31A**). LSL-Cas9-EGFP mice also express the neomycin (*Neo*) gene; therefore, a lentivirus containing only Cre and a sgRNA targeting *Neo* (lenti-Cre) was used as control, since the injection of lenti-Cre would create a double strand break in the *Neo* gene but would not induce the *Eml4-Alk* rearrangement.

Intratracheal instillation of this the lentiviral system, results in a slight delay in tumor development in both V1 and V3 to an equal extent (data from Alberto Diaz), with V3 mice showing a survival rate of around 12 weeks and V1 mice of around 36 weeks.

Additionally, I observed that V3 mice already harbored fully developed tumors at 6 weeks. Therefore, to capture tumor initiation, I collected the lungs of V1 mice 2 weeks (n=1), 4 weeks (n=3) and 6 weeks (n=3) after lentivirus instillation, but only 2 weeks (n=1) and 4 weeks (n=3) of V3 mice. To consider the immune response induced by the virus instillation, I collected control samples from lenti-Cre mice at 2 weeks (n=1) and 6 weeks (n=3) (**Figure 31B**). I sorted live CD45⁻, CD31⁻, CD11b⁻, Ter-119⁻, EpCam⁺, GFP⁺ cells (**Figure 31C**) and performed plate-based SMART-seq2 sequencing. The bioinformatic analysis was then performed by Dr. Reka Toth.

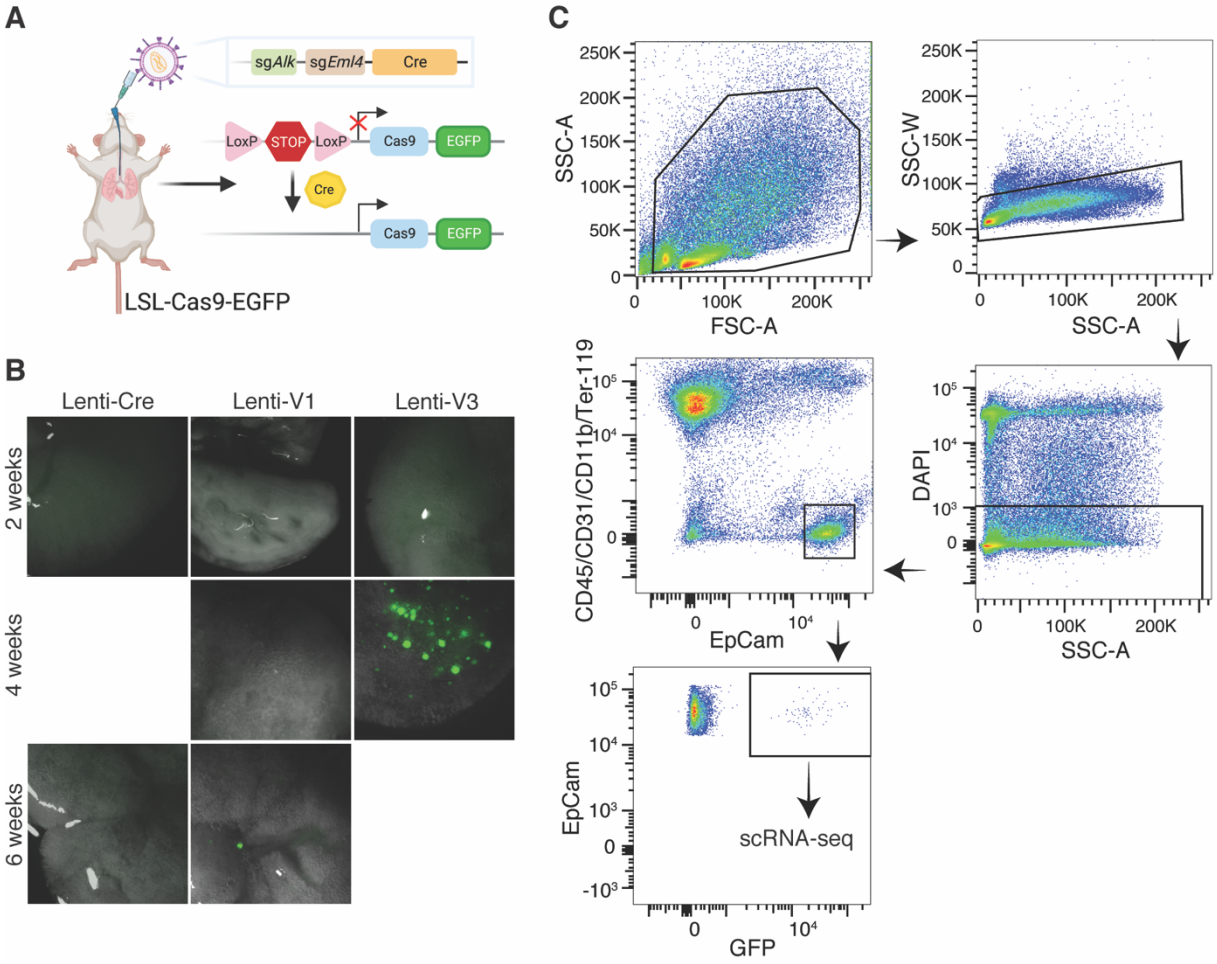

Figure 31.The lentiviral system labels infected cells with GFP.
(**A**) Schematic representation of the lentivirus system and Rosa26-Lox-Stop-Lox-Cas9-EGFP (LSL-Cas9-EGFP) mice. (**B**) Representative fluorescence images of lungs injected with lenti-Cre, lenti-V1 or lenti-V3 and collected at the indicated time points. Scale bar: 2mm. (**C**) Sorting strategy to acquire GFP⁺ epithelial cells from the LSL-Cas9-EGFP lung samples for further scRNA-seq.

2,574 high-quality cells were captured, which clustered into 8 independent clusters (**Figure 32A**). Clusters 3 and 7 were the main clusters containing *Alk*-expressing cells (**Figure 32B**), suggesting that these cluster contained the transformed and more advanced cells. To further understand the data, I checked which samples

formed each of these clusters (**Figure 32C**). Cluster 0 was composed majorly of cells from the 4wk and 6wk V1 samples, in addition to cells from the 4wk-V3 and 6wk-cre samples. Cluster 1 comprised predominantly cells from the 2wk samples, whereas cluster 3 and 7 were mainly formed by cells in the 4wk-V3 samples and a few from the 6wk-samples. The rest of the clusters contained a mix of cells from all the groups. Regarding the sample composition, the cells in all the 2wk samples were predominantly included in cluster 1 (**Figure 32D**), which together with the cluster composition, suggested that this cluster might be related to the initial immune response to the viral system. Interestingly, the rest of the groups were primarily formed by cluster 0, except for the 4wk-V3 group which showed an enrichment in cluster 3.

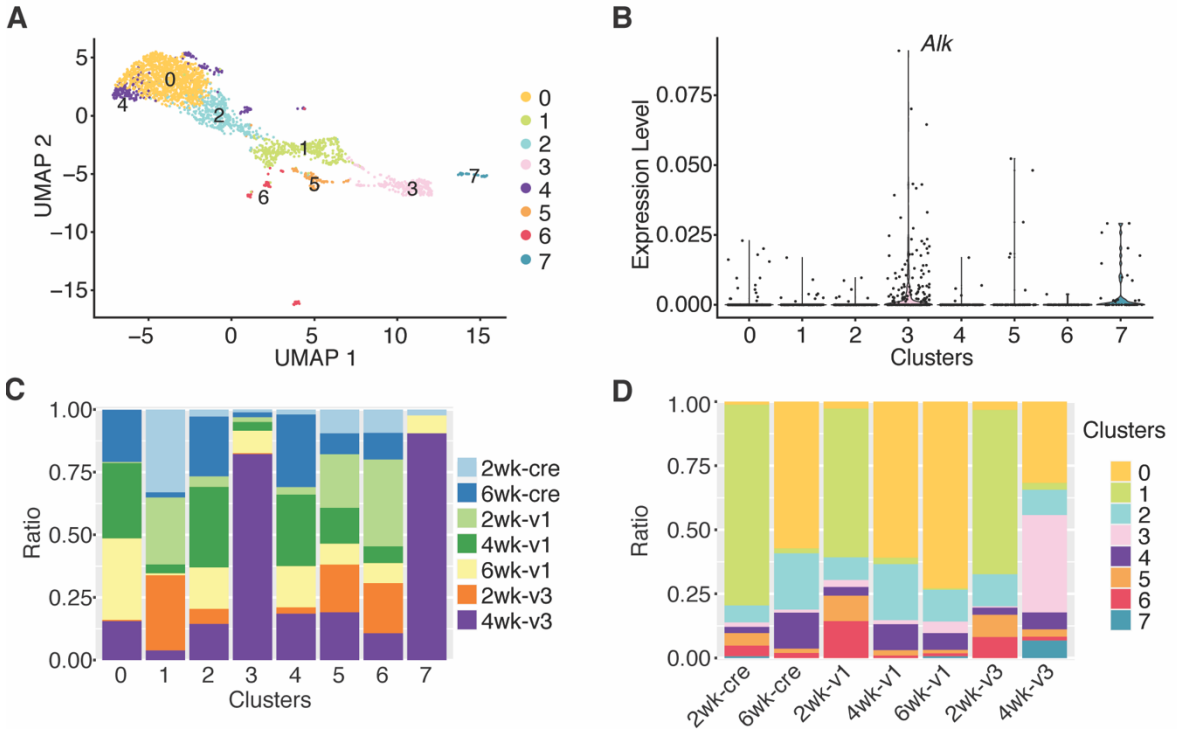

Figure 32. Time course single cell RNA-sequencing experiment of early timepoints.
(**A**) UMAP embedding by clusters obtained by unsupervised nearest neighbor modularity optimization-based clustering. (**B**) Violin plot representing the normalized expression of *Alk* for each cluster. (**C**) Cluster composition colored by sample group shown as relative contribution normalized to the total number of cells in each cluster. (**D**) Sample composition colored by cluster shown as relative contribution normalized to the total number of cells in each group.

3.5. V1 cells downregulate *Trp53* levels overtime

Experiments in our lab showed that the used lentiviruses mainly infect AT2 cells in the mouse lung, and barely infect any other cell type. Therefore, to confirm that the sequenced cells were indeed AT2 cells, and to check which cells showed signs of transformation, I checked their enrichment for published cell type markers[104], in

addition to the tumor signature that we identified from the V1 club originated tumors[129] (**Appendix I Table 20**). All clusters showed high enrichment in the AT2 signature, although clusters 3, 7 and, to a less extent, 5 showed a decrease in this signature (**Figure 33A**). It has been shown that AT2 cells loss their linage signature upon transformation[120]. These results could indicate that these clusters contained transformed cells. Moreover, I observed almost no enrichment in the club signature, confirming that all the cells sequenced were AT2 cells. Surprisingly, cluster 5 showed the highest levels of the AT1 signature, suggesting that these cells have both an AT2 and AT1 phenotype. Furthermore, cluster 3 was marked by the expression of the tumor signature, confirming that these cells were already transformed, whereas cluster 7 was characterized by the expression of the proliferation signature (**Figure 33B**).

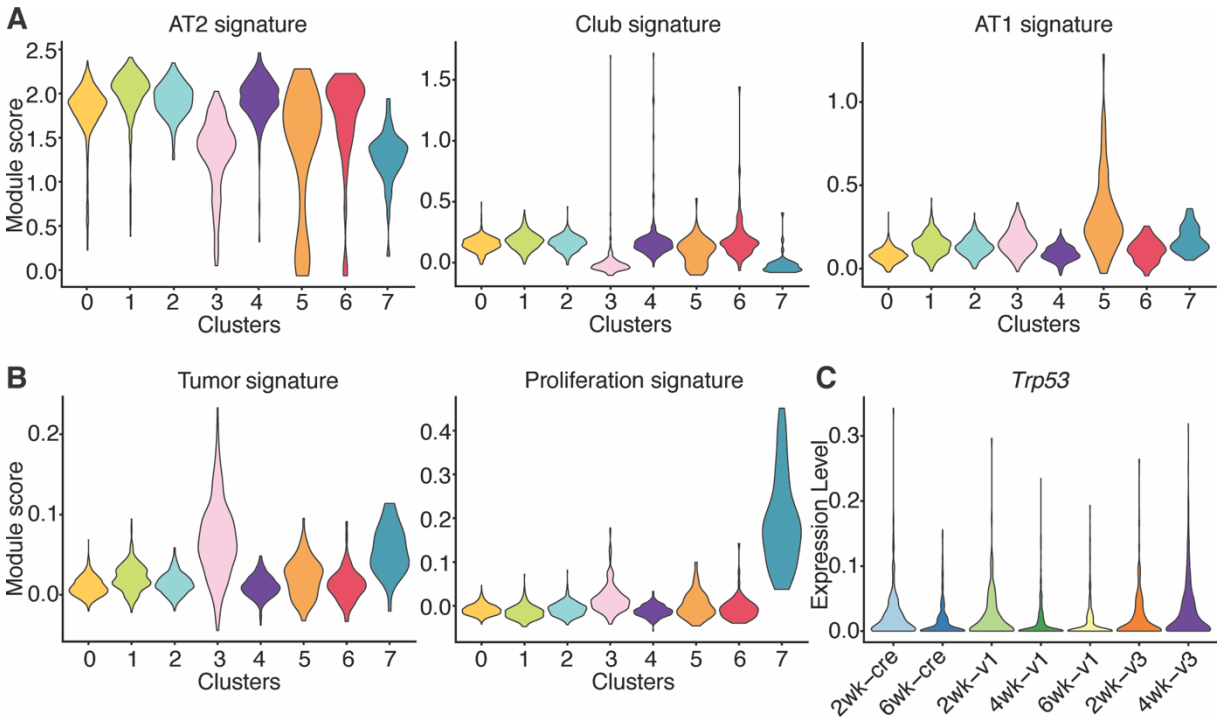

Figure 33. V1 cells downregulate *Trp53*
(**A**) Violin plots depicting the signature scores of the AT2, club and AT1 signatures by the described clusters. The signatures were obtained from the study of Strunz et al.[104]. (**B**) Violin plots depicting the signature scores of our Tumor signature[129] and the proliferation signature by the described clusters. The signatures were obtained from the study of Strunz et al. (**C**) Violin plot showing the expression levels of *Trp53* at the different time points.

Unfortunately, I did not identify a V1-specific cluster that would allow me to understand how these cells overcome the apoptotic process they suffer early upon transformation. Therefore, rather than analyzing the data by clusters, I focused on the different sample groups. Since I had observed that the loss of *Trp53* shortens the survival of V1 mice and increases the number of V1 tumors, I examined the levels of

Trp53. I noted that among the V1 samples, there was a slight decrease of the levels of *Trp53* over time, which did not occur in the V3 samples (**Figure 33C**). Altogether, these results suggest that p53 may be delaying tumor initiation in V1, and that these cells may need to downregulate *Trp53* to be able to start proliferating and forming a tumor.

Discussion

1. Comparison of the V1 and V3 lung tumors

1.1. Generation of a new *Eml4-Alk* V3 mouse model

EML4-ALK is a fusion protein that results from a specific genetic abnormality found in a subset of NSCLC patients, and it plays a crucial role in the development and progression of NSCLC[36]. Since the discovery that this oncogene can result in several different variants, very little effort has been put into understanding the differences among them. A few clinical studies have suggested that for instance V1 and V3 might have different tumorigenic and metastatic potential, or even behave differently upon treatment[65,68,69,186], but no systematic analysis has ever been performed. Therefore, it is essential to study this oncogene to identify potential new therapeutic drugs for the treatment of this type of NSCLC. In several studies, overexpression of the EML4-ALK protein has been used to analyze the protein's effect in cells or tissues[71,130,131,184]. However, the overexpression of EML4-ALK does not resemble the situation of human lung tumors[199]. The normal expression level of EML4-ALK in human LUAD is crucial for its function in the disease, and artificially elevating its expression level in research experiments may lead to false conclusions about its role in tumor development and therapeutic resistance. Therefore, in the study presented, the CRISPR/Cas9 system was used to induce the endogenous *Eml4-Alk* V1 and V3 in mouse lungs. Prior to this work, there was no available V3 mouse model, since only V1 *in vivo* models had been developed[130–133]. The generation of the V3 mouse model provides a valuable resource for investigating the pathogenesis of this variant and opens up new avenues for understanding the underlying tumorigenic mechanisms compared to V1. In addition, with this new mouse model, we can now start to comprehend the impact of the tumor microenvironment in the tumors driven by different *Eml4-Alk* variants allowing the development of better immunotherapies for ALK$^+$ patients, which do not respond to the current lines of immunotherapeutic treatments[55].

Both adenoviruses employed in the study were identical except for the sgRNA targeting *Eml4*, which differed between the two viruses. It is well known that the efficiency of the CRISPR/Cas9 system can be influenced by several factors, including the

choice of the sgRNA used to direct the Cas9 nuclease to its target site in the genome. Doench et al. found that the efficiency of the system can be influenced by factors such as the length and composition of the sgRNA[200]. Similarly, Wang et al. demonstrated the effectiveness of gene targeting in human embryonic stem cells, which was dependent on the sgRNA design[201].

In light of this knowledge, it is important to carefully consider the design and validation of sgRNAs when using the CRISPR/Cas9 system for gene editing applications. Specifically, I assessed the efficiency of both V1 and V3 adenoviruses in inducing *Eml4-Alk* to determine if the sgRNA used in each virus affects the success of the gene editing process. The lack of working antibodies to detect the mouse *ALK* protein and the complexity of the FISH analysis made it difficult to quantify the number of *ALK* positive cells early upon transformation. Therefore, I elaborated a digital PCR-like technique which allowed me to compare the *Eml4-Alk* rearrangement at very early timepoints. No significant differences were observed between V1 and V3 using this technique, which suggested that the efficiency of both adenoviruses in generating the fusion gene is very similar. These results allowed the comparison of both variants in tumor formation and overall survival along the rest of the thesis. Nevertheless, it would be interesting for future experiments to test additional sgRNAs for both variants, which would reinforce the results presented in this thesis.

1.2. Comparison of V1 and V3 tumors

Lung adenocarcinoma is characterized by diverse histological patterns. These patterns are important for diagnosis, prognosis, and treatment planning. In particular, the solid, papillary, and lepidic patterns have been identified as the most common patterns seen in LUAD[10]. It is common to observe different histological subtypes mixed within a single tumor, and it has been described that the proportion of subtypes such as solid or papillary can affect patient prognosis[202]. The solid pattern is associated with a higher risk of recurrence and a worse prognosis compared to other patterns, and the lepidic histology is associated with a better prognosis[203], particularly when it is the predominant pattern. I my study, I observed that V3 tumors showed predominantly a papillary pattern, although a high percentage of tumor showed a solid pattern. Maddalo et al. described that their V1 tumors presented papillary and acinar patterns[133]. These differences could be due to a different phenotype between V1 and V3, or to the fact that

my analysis was performed at a later timepoint, in which tumors might have had time to develop towards other patterns. Therefore, it would be of interest to analyze and compare the histological patterns of V1 and V3 under the same conditions. Altogether, these data correlate with the reduced survival observed in V3 mice. Although discrepancies can be found in the literature, the results of my thesis point towards those studies in which V3 is suggested to be a more aggressive variant[68,69,186], and therefore, it should be considered as a poor prognostic factor in these patients.

Analysis of the bulk transcriptome of V1 and V3 tumors did not result in conclusive findings about the aggressiveness of V3 tumors. Some studies in the literature associated this aggressiveness to a better stability of the V3 protein, due to its shorter sequence and conformation[71,204,205]. A stronger ALK signaling has also been described in V3 cells compared to V1 cells after the observation of higher phosphorylation of the ALK protein, which could be the cause of more potent tumors[79,204,205]. Nevertheless, most of these studies were performed by overexpressing the EML4-ALK variants and contradictory studies have been reported in most of the cases. Therefore, the mechanisms and pathway that distinct V1 and V3 tumors still remain elusive.

1.2. The *Eml4-Alk* fusion can lead to lung squamous cell carcinomas

The occurrence of the EML4-ALK rearrangement in the lung is typically associated with LUAD and the rarity of treatment-naïve SCCs harboring the EML4-ALK fusion in patient samples is noteworthy. Previous studies have reported ALK positivity in only a small percentage of SCCs (1-2%)[74,76,206–208]. In my study, I observed that *Eml4-Alk* V3 can indeed give rise to SCCs, which had not been previously reported in the V1 mouse models[133]. Interestingly, I found that many of these squamous tumors exhibited amplification of the *Eml4-Alk* fusion gene, which suggests that increased copy numbers of *Eml4-Alk* may play a role in driving SCC development. This finding is of particular significance, as SCC is a subtype of lung cancer that is generally resistant to targeted therapies directed against oncogenic drivers, and to my knowledge, SCC are not routinely tested for the presence of the EML4-ALK fusion. Therefore, these results reinforce that, although few patients with SCC could be EML4-ALK positive, they could benefit from the current standards of treatment for ALK positive patients if tested for this fusion. Although the effectiveness of the current TKIs on *Eml4-Alk* SCCs have not

been tested in this thesis due to the scarcity of these tumors in the *Trp53* wt background, one could think of employing *Trp53* knockout to test this hypothesis. Additionally, it would be of great interest to generate mouse models which exclusively drive *Eml4-Alk* SCC and investigate the mechanisms driving this tumor type and their therapy response to the available TKIs. These experiments would be important to stablish treatment regimens for treatment naïve EML4-ALK SCC patients and for those which develop SCC as a resistance mechanism to TKIs.

Importantly, I identified club cells as the cell of origin of *Eml4-Alk* squamous cell carcinomas, since these tumors were only found to be labelled in the *Scgb1a1*-CreER mouse model. These findings are consistent with a previous study in which CC10+ cells were found to be the predominant origin of adenosquamous carcinomas driven by $Kras^{G12D}$ and *Lkb1* loss[137]. Moreover, I observed that the loss of *Trp53* in club cells enhanced the appearance of tumors with squamous histology, also in the V1 mouse model. This suggests that both V1 and V3 might have the potential to induce SCCs, although V1's capability for this might be diminished by tumor suppressors such as Tp53. Altogether, these results demonstrate that airway epithelial cells can serve as the cell of origin for SCCs, and that the loss of tumor suppressor genes such as *Trp53* can drive tumorigenesis in these cells. Therefore, this study provides new insights into the role of *Eml4-Alk* fusions in the development of squamous cell carcinomas, which had barely been previously described. However, my results do not unveil whether the loss of *Trp53* is essential to drive the development of *Eml4-Alk* SCC, or if it only facilitates the emergence of this cancer subtype. Considering that *Trp53* is the most common mutation found in EML4-ALK patients it is important to understand what effects does the loss of *Trp53* have of the treatment response of these tumors.

In conclusion, my study expands our understanding of the role of *Eml4-Alk* in the development of lung cancer, demonstrating its potential to drive the growth of SCCs. My findings also highlight the importance of considering the cellular origins of lung cancer subtypes, and suggest that club cells may be a potential source of SCCs driven by *Eml4-Alk*. Further studies are needed to fully elucidate the molecular mechanisms underlying *Eml4-Alk*-driven SCC development and to develop effective targeted therapies for this aggressive subtype of lung cancer.

2. Cell of origin of Eml4-Alk lung tumors

2.1. V3 tumors originate from club and AT2 cells

The cell of origin of NSCLC has been extensively studied in recent years, and it is now widely accepted that this disease arises from the transformation of different cell types in the lung, including club, basal, and AT2 cells, among others[209]. The cell of origin might play a crucial role in determining the molecular and genetic alterations that contribute to the development and progression of NSCLC, as well as the response of the tumor to therapy. In this context, the results of this thesis provide important insights into the role of the cell of origin in the development and progression of NSCLC with *EML4-ALK* fusion gene variants. Here, I employed different Cre recombinase-driven models to investigate the cellular origins of V3 tumors, which revealed that these malignancies can be derived from club and AT2 cells, but not basal cells, consistent with previous findings in V1[129]. Studying this matter can be challenging, especially when using lineage tracing mouse models since they can be promiscuous and label unintended cell types, which can lead to inaccurate conclusions about the cell of origin of a given tumor. The dose and regimen scheme of the TAM treatment in these mice clearly impacts the labelling results[11,140]. In our lab, we have described the promiscuity of these mouse lines[129], and in particular, we showed that the *Scgb1a1*-CreER mice label club, AT2 and ciliated cells, which could mask the real cell of origin of the GFP$^+$ tumors found in this model. Experiments in our lab not shown in this thesis, have demonstrated that *Scgb1a1*-CreER mice treated with a lower dose of TAM, which exclusively labels club cells, still harbored GFP$^+$ tumors and the percentage of these was not affected, confirming that club cells are the cell of origin of the GFP$^+$ tumors found in this model.

Furthermore, very little is known about how exactly the cell of origin of NSCLC impacts the phenotype and behavior of the tumors. In addition, the clinical significance of the cell of origin in terms of patient prognosis also remains elusive. Therefore, I used the linage tracing mice to understand the role of the cell of origin in the survival and proliferation of *Eml4-Alk* NSCLC. I found that in *Sftpc*-CreER mice, those with a higher percentage of AT2-derived tumors had a significantly shorter survival than those with fewer AT2-derived tumors. In contrast, I observed the opposite trend in *Scgb1a1*-CreER mice, where a higher percentage of club cell-derived tumors was associated with longer

survival. These results were consistent for both V1 and V3 and suggest that the cell of origin can have a significant impact on the phenotype and prognosis of *Eml4-Alk* NSCLC.

One aspect that was also not considered, is the fact that SCC is accepted to be a more aggressive disease when compared to LUAD. In my analysis I did not further consider that club originated tumors might become SCC and hence increase the aggressiveness. In any case the amount of SCC tumors might not have impinged in the results presented.

Tumors derived from AT2 cells may be more aggressive and have a poorer prognosis than those derived from club cells. These findings are consistent with previous studies suggesting that AT2 cells may be more susceptible to oncogenic transformation due to their high proliferative capacity and increased susceptibility to DNA damage[107,119,136]. However, the aggressiveness of the tumors originated from distinct cell types may be influenced by the combination of genomic alterations in several oncogenes and tumor suppressor genes. For instance, a recent study described that the loss of *Smarca4* exclusively on club cells in the KP model induces malignant and aggressive tumors, whereas it only induced low-grade tumors when absent in AT2 cells[141]. Since the effort and the number of mice needed to study each of these individual genomic alterations would not be possible to achieve, one could think about alternative approaches as the one used by the laboratory of Monte Winslow[210–212] where inactivation of several tumor suppressor genes can be combined to test this hypothesis.

Nevertheless, I found that in the V3 model, AT2-derived tumors had a higher proliferation rate than club cell-derived tumors, but I did not observe this difference in the V1 model. This again suggests that the impact of the cell of origin on tumor proliferation may vary depending on the specific genetic alterations present in the tumor. Further studies aimed at understanding the molecular mechanisms underlying these differences and using distinct oncogenes may provide valuable insights into the development of targeted therapies for the distinct cells of origin.

Through the creation of a niche for tumor cells, the TME has repeatedly been demonstrated to play a crucial role in tumor growth[145], thus understanding the perturbations in the TME according to the cancer originating cell type will be an important part of the cell of origin studies in the near future. Some of the few

investigations on this matter have described that AT2-derived tumors show a higher infiltration of $CD45^+$ cells compared to those originated from club cells[137,147,148]. Yet, a complete mapping of the TME landscape of the different origins and its impact in treatment response is still lacking in lung adenocarcinoma development. Investigating the TME with the mouse models used in this thesis is a difficult task due to the ubiquitous initiation of the tumors. To be able to examine this, it would be important to develop mouse models in which the *Eml4-Alk* endogenous rearrangement can be induced exclusively in the desired cell types. Modifying the adenoviruses employed in this thesis to express the sgRNAs under cell type specific promoters could be a useful approach. Further single cell transcriptomic analysis would then help unveil the different microenvironment populations. Importantly, this approach would allow to investigate the therapy response of tumors from different origins.

2.2. Club cells employ regeneration mechanisms during tumor development

To further understand the development of club-originated tumors, we performed a time course scRNA-seq experiment. By increasing their own population and transdifferentiating into AT2 and AT1 cells, club cells are known to rebuild the lung epithelium[85,104]. This great adaptability of club cells was also found during *Eml4-Alk* carcinogenesis in our scRNA-seq, where we observed that the V1 oncogenic stimulus rewired the club cells' transcriptome to evolve to an alternate destiny program, similar to the one identified in lung regeneration. Several cellular phases that had previously only been demonstrated to be transitory conditions during lung regeneration after damage, were discovered in our scRNA-seq study. For instance, we saw that goblet-like club cells, which were characterized by the expression of *Ltf* and *Bpifb1*, differentiated into a pre-tumor stage expressing *Alk*, which was marked by the presence of *Ltf*, together with basal cell markers such as *Krt5*, and *Trp63*. During lung regeneration, a transitional cell state has been described to originate from goblet cells (expressing *Bpifb1*) and dedifferentiate into terminal Basal cells (*Krt5, Trp63*)[194]. Therefore, experiments performed in this thesis, suggested that club cells resemble this condition during early phases of cancer development.

Analysis of the scRNA-seq also unveiled a set of tumor-associated activity programs which were characterized by the expression of markers also seen in intermediary stages between club or AT2 cells toward AT1 cells during lung

regeneration. More specifically, the regeneration-like tumor module that we identified revealed activation of the pathways reported as being crucial for AT1 differentiation during lung injury and regeneration, including the HIF1 signaling, TGF-beta pathway, IL-17 signaling, and glycolysis[104,106,107]. The fingerprints of Krt8ADI and PATS, two cellular states discovered in lung regeneration, were also highly similar to one of the routes that club cells take towards the tumor clusters. Our findings support a connection between lung cancer and lung regeneration, and they emphasize the significance of lung cell plasticity during carcinogenesis. It is interesting to note that the club cells' tumorigenic pathways identified correlated with a few patterns that have previously been described in tumors from an AT2 origin. In a *Kras*-driven LUAD model, it was described that AT2 cells lose their cellular identity during progression towards a metastatic state[120,121] and additionally, some of these cells upregulate club-related genes, such as *Scgb1a1*[195]. Similarly, we observed that club cells downregulate their cell type-specific transcriptional signature and gradually acquire features of AT2 cells. In addition, these studies showed that AT2-derived tumors progress through an embryonic-like state and an AT1/AT2 mixed state that were also present in some of our club-derived tumor cells. This shows that even though lung tumors can develop from a variety of cell types, some stages of carcinogenesis might be very similar. The similarity between some of these stages in club and AT2-originated tumors and the promiscuity in the labelling of the *Scgb1a1*-CreER model could imply that our club tumors were not be purely club-derived nodules. However, the trajectory analysis did not predict any contribution of the GFP^+ labelled AT2 cells to the tumor clusters, suggesting that our tumor cells were indeed originating from club cells.

In summary, we discovered that club cells take advantage of the signaling pathways that regularly control healthy lung regeneration and growth, and we outlined the strategies they take to achieve a highly heterogeneous state via several key tumorigenic pathways. It is of great relevance to continue investigating which features are similar of different between tumors from distinct cells of origin and how do those features impact the tumor behavior. In addition, the vast majority of the cell of origin studies have been performed using *in vivo* mouse models, and it is essential that this research area starts to move towards identifying markers that could distinguish the cell of origin of human LUAD. It is likely, that when more thorough and advanced molecular

profiling of both normal and cancer cells are carried out, the question of the cell-of-origin of human LUAD will become more solvable.

3. Tumor initiation driven by *Eml4-Alk* V1

The development of tumors is a complex process that involves a series of genetic alterations. One of the key players in this process is the *Trp53* tumor suppressor gene, which is known to play a crucial role in regulating the cell cycle and preventing the formation of cancerous cells[149]. The findings of this study provide important insights into the mechanisms underlying the initial development of V1 tumors. I found that V1 induces DNA damage and apoptosis very early upon induction, which presumably slows down tumor formation. However, in the absence of *Trp53*, tumors developed more rapidly in these mice, and tumor latency was significantly reduced. Additionally, scRNA-seq unveiled the downregulation of *Trp53* in V1 cells across a span of weeks, which may contribute to the development of these tumors. Altogether, my results suggest that *Trp53* prevents the formation of V1 cancerous cells, and that inactivation of the Tp53 pathway reverts this phenotype. These results are in line to those in which *Kras* has been described to induce cellular senescence early upon induction due to its strong oncogenic potential[213], a phenotype that is reversed after the inactivation of Tp53. This may imply that V1 could be an extremely potent oncogene, that upon activation induces Tp53, causing early death of the cells due to excessive oncogenic signaling. Therefore, after inactivation of Tp53, cells can cope with this signaling, do not die by apoptosis and as a result, mice develop tumors earlier. This hypothesis would suggest that the early death of V3 mice might only be due to an earlier and easier initiation of V3 tumors, implying that V3 would be in fact a less potent oncogene compared with V1.

Nevertheless, the scRNA-seq experiment to unveil the tumor initiation process of V1 and V3 did not result in very conclusive findings. Results obtained in our lab after performing the experiment showed that the lentiviral system used can be very variable in tumor latency, which could have influenced my results, since tumor development might have been more delayed than expected in these mice. Moreover, very few cells were captured, which also made the identification of significant pathways and activity programs challenging. Despite these facts, our bioinformatician was able to uncover

several activity programs which are currently being analyzed and which could shed light into the process of early death of V1 cells.

In summary, further studies are needed to fully elucidate the mechanisms underlying the role of *Trp53* in the process of V1 tumor development, and to develop better therapeutic approaches for these patients. However, our study provides a strong foundation for future research aimed at improving the diagnosis and treatment of V1 tumors.

Conclusions

1. Characterization of the biological and molecular divergence among *Eml4-Alk* variants

Conclusions of this part of the thesis:

- Histological similarity/difference between V1 and V3 (solid/grading).
- *Eml4-Alk* V3 has the potential to generate adenocarcinomas and squamous cell carcinomas.
- *Eml4-Alk* V3 mice show a shorter overall survival than V1 mice.

To summarize, the findings of this thesis indicate that the *Eml4-Alk* V3 fusion gene possesses the capability to generate adenocarcinomas and squamous cell carcinomas. Moreover, the research also indicates that V3 mice exhibit a decreased overall survival as compared to V1 mice. These results suggest that V1 and V3 have a substantially different influence on the pathogenesis and progression of cancer. The implications of these findings open the door to understanding the mechanisms underlying the development and progression of this cancer type. Nevertheless, further research is required to elucidate the specific mechanisms underlying the association between the *Eml4-Alk* V3 fusion gene and the development of more aggressive tumors. The generation of this V3 mouse model provides a powerful resource for preclinical testing of novel therapeutic strategies, which could ultimately lead to improved patient outcomes.

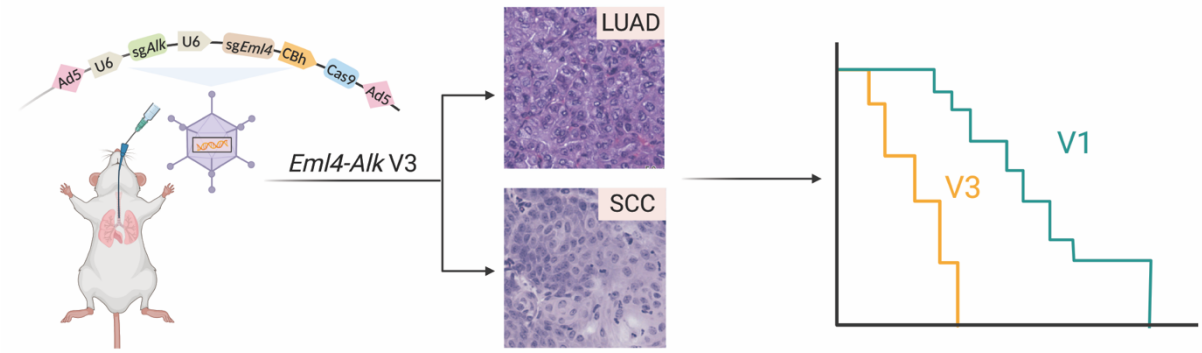

Figure 34. Graphical abstract part 1.
Eml4-Alk V3 induces lung adenocarcinomas (LUAD) and squamous cell carcinomas (SCC) and a shorter overall survival compared to V1 mice.

2. Cell of origin of *Eml4-Alk* tumors

Conclusions of this part of the thesis:

- Club and AT2 cells can give rise to *Eml4-Alk* V3 tumors.
- Tumors arising from AT2 cells cause a reduced lifespan compared to club-originated tumors.
- Club-derived malignancies lose their linage identity and acquire an AT2-like transcriptional pattern.
- Club cells exploit their intrinsic regeneration mechanisms in the tumorigenesis process.
- Club cells employ several routes towards fully developed tumor cells.
- Club-derived tumors are enriched in activity programs distinct to those of AT2-originated tumors.
- The signature of club tumors is present in human lung adenocarcinomas.

The present study provides valuable insights into the cell of origin of *Eml4-Alk* tumors, derived from club and AT2 cells in the lung. The findings suggest that club cells exploit their intrinsic regeneration mechanisms to drive tumorigenesis through several routes, and that tumors arising from club cells lose their lineage identity and acquire an AT2-like transcriptional pattern. Moreover, the study shows that club-derived tumors exhibit distinct activity programs compared to AT2-originated tumors. These findings may have important implications for the development of novel therapeutic strategies for lung cancer.

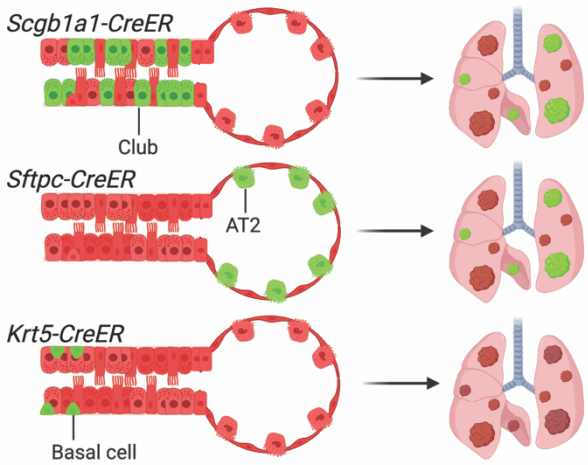

Figure 35. Graphical abstract part 2.
Both club cells and AT2 cells can be the cell of origin of *Eml4-Alk* V3 tumors

3. Tumor initiation and role of Tp53 in *Eml4-Alk* tumors

Conclusions of this part of the thesis:

- V1 induces DNA damage and results in Tp53 activation and apoptosis early upon induction.
- *Trp53* loss accelerates the death of V1 mice.
- *Trp53* loss enhances tumor formation in V1 but not V3 mice.
- V1 cells downregulate *Trp53* overtime.

The results of this part of the thesis highlight the crucial role of the *Trp53* tumor suppressor gene in the development of V1 tumors. The findings suggest that V1 induces DNA damage, which results in the activation of p53 and apoptosis, delaying tumor formation. The loss of *Trp53* leads to increased tumor formation in V1 mice and accelerated death. Additionally, downregulation of *Trp53* in V1 cells may contribute to the development of these tumors.

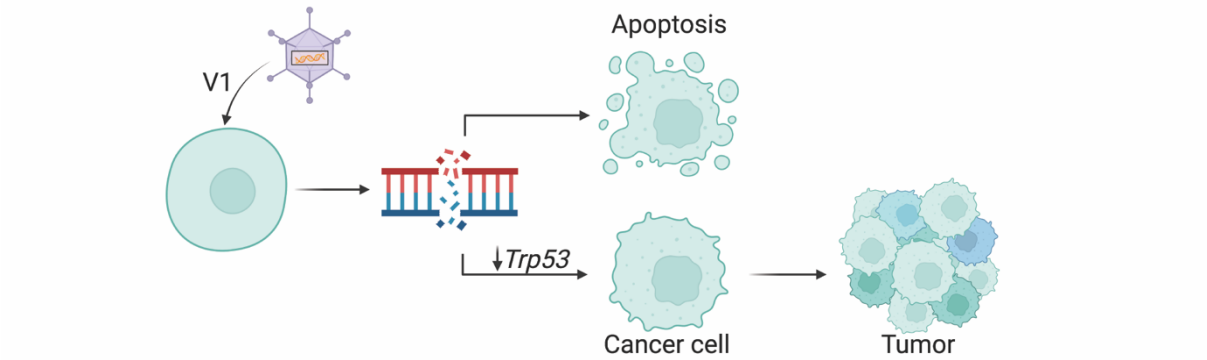

Figure 36. Graphical abstract of part 3.
Eml4-Alk V1 causes DNA damage and apoptosis early upon induction, delaying tumor formation, and downregulation of *Trp53* is required to promote tumorigenesis.

Appendix I

Table 18. Marker genes used to annotate the cell types in the V1 club sequencing experiment.

Club	AT2	AT1	Goblet	Basal	Proliferation	Mixed AT1-AT2	Activated club	Club progenitor	Activated AT2
Scgb1a1	*Sftpc*	*Ager*	*Scgb3a1*	*Krt15*	*Hmgb2*	*Clic3*	*Cst3*	*Cd74*	*Npc2*
Cyp2f2	*Sftpa1*	*Cldn18*	*Scgb3a2*	*Rpl13*	*D17H6S56E-5*	*Ager*	*H2-Ab1*	*Cd14*	*Lcn2*
Hp	*Slc34a2*	*Sparc*	*Bpifb1*	*Aqp3*	*Mki67*	*Col4a3*	*H2-Eb1*	*H2-K1*	*Sftpd*
Cbr2	*Sftpb*	*Emp2*	*Reg3g*	*Dapl1*	*Top2a*	*Dpysl2*	*Psap*	*AW112010*	*Sftpb*
Selenbp1	*Lyz2*	*Akap5*	*Lypd2*	*Ifitm3*	*Ptma*	*Mmp11*	*Ctss*		*Cxcl15*
Mgst1	*Napsa*	*Timp3*	*Tff2*	*Cyp2f2*	*Smc4*	*Akap5*	*H2-Aa*		*Chi3l1*
Fmo2	*Hc*	*Clic5*	*Pigr*	*Bpifa1*	*H2afz*	*Rtkn2*	*Laptm5*		*Tmsb4x*
Scgb3a2	*Sftpd*	*Igfbp7*	*Scgb1a1*	*Rps5*	*Tubb5*	*Hopx*	*Actb*		*Lyz2*
Retnla	*Cxcl15*	*Vegfa*	*Gsto1*	*Abi3bp*	*Prc1*	*Radil*	*Alox5ap*		*Sftpa1*
Alas1	*Scd1*	*Msn*	*Ltf*	*Tmem176b*	*Stmn1*	*Sec14l3*	*Srgn*		*Lamp3*
Pigr	*Lamp3*	*Ndnf*	*Bpifa1*	*F3*	*Tuba1b*		*Ccl17*		*Napsa*
Trf	*Fasn*	*Myl12a*	*Cp*	*Rps14*	*Hmgb1*		*Lsp1*		*Scd1*
Cp	*Lpcat1*	*Anxa5*	*Wfdc2*	*Rps8*	*Hmgn2*		*Scgb3a1*		*Sftpc*
Prdx6	*S100g*	*Aqp5*	*Mgst1*	*Rps6*	*Wbp5*		*S100a4*		*Elovl1*
mt-Cytb	*Malat1*	*Lmo7*	*Cyp2f2*	*Dcn*	*Cenpf*				*Il33*
mt-Co1	*Npc2*	*Krt7*	*Retnla*	*Ifitm1*	*Ncl*				*Sfta2*
Gsta3	*Rnase4*	*Cryab*	*Muc5b*	*Rplp0*	*2700094K13Rik*				*H2afj*
Por	*Abca3*	*Malat1*	*Sult1d1*	*Rps3*	*Spc24*				*Lrg1*
Ppap2b	*Chi3l1*	*Anxa3*	*Qsox1*	*Fxyd3*	*Kif23*				*Ly6e*
Aldh1a1	*Elovl1*	*Clic3*	*Fmo3*	*Fmo2*	*Birc5*				*Slc34a2*
Ldhb	*Ctsc*	*Icam1*	*Cxcl17*	*Rpl18a*	*Smc2*				*Rps19*
Ces1d	*Lgi3*	*Serpinb9*	*Aldh1a1*	*Rpl5*	*2810417H13Rik*				*S100g*
Mettl7a1	*Egfl6*	*Anxa1*	*Trf*	*Rps18*	*Ctsh*				*Cd36*
mt-Nd2	*Ctsh*	*Tpm1*	*Clca3*	*Krt19*	*Ranbp1*				*Slpi*
B430010I23Rik	*Ager*	*Ctsh*	*Chad*	*S100a6*	*Lig1*				*Egfl6*
Cldn10	*Gpr116*	*App*	*Agr2*	*Sult1d1*	*Hmmr*				*Atp5e*
Cyb5	*Sepp1*	*Myl6*	*Adh7*	*Rpl12*	*Ybx1*				*Acot7*
Fmo3	*Ptprf*	*Sdpr*	*Fxyd3*	*Rps9*	*Cenpe*				*Ptgs1*
Cyp4b1	*Nrp1*	*Hopx*	*Cbr2*	*Gsto1*	*Rangap1*				*Acsl4*
Gstm2	*Sfta2*	*Serpinb6b*	*Gsta4*	*Rpl28*	*Tpx2*				*Dram1*

Table 19. Top 10 markers of cNMF activity programs

Club	Activated club	Normal AT2	Proliferating	Interferon	Immune activation	Regeneration-like	AT1-like	Stem-like
Scgb1a1	*Lypd2*	*Ank3*	*Cdk1*	*Ifi47*	*Serpinb1a*	*Serpine1*	*Akap5*	*Cd177*
Cbr2	*Scgb3a1*	*Etv5*	*Birc5*	*Igtp*	*Spib*	*Nudt4*	*Krt7*	*Lcn2*
Cyb5a	*Sult1d1*	*Rnase4*	*Cdca8*	*Gbp3*	*Ccl20*	*Bnip3*	*Emp2*	*Col18a1*
Prdx6	*Adh7*	*Lamp3*	*Top2a*	*Stat1*	*Ccl9*	*BC030867*	*Timp3*	*Atox1*
Pgrmc1	*Cyp4a12b*	*Acoxl*	*Ccna2*	*Irgm1*	*Bcl2a1a*	*Col15a1*	*Sparc*	*Ogfrl1*
Ldhb	*Qsox1*	*S100g*	*Ube2c*	*Psmb8*	*Pglyrp1*	*Bsg*	*Anxa3*	*Id2*
Cyp2f2	*Bpifb1*	*Mme*	*Ckap2l*	*Psmb9*	*Sox8*	*Fn1*	*Hopx*	*Arg2*
Ces1d	*Calml3*	*App*	*Aurkb*	*Gbp4*	*Bcl2a1b*	*Ankrd37*	*Igfbp7*	*Lrg1*
Scnn1b	*Muc5b*	*Selenop*	*Mki67*	*Zbp1*	*Plekhs1*	*Rap1gap*	*Ager*	*Prdx4*
Dcxr	*Tst*	*Atp8a1*	*Pbk*	*Gbp2*	*Mmp23*	*Cdkn1a*	*Anxa5*	*Hdc*
Cldn10	*Fxyd3*	*Acot1*	*Plk1*	*Ifitm3*	*Mfge8*	*Sstr2*	*Fbln5*	*Ctsd*
Gsta3	*Aldh3a1*	*Sfta2*	*Cdca3*	*Psmb10*	*Msln*	*Bnip3l*	*Crip2*	*Chil1*
Gstm1	*Oit1*	*Slc34a2*	*Stmn1*	*Tap1*	*Il4i1*	*Egln1*	*Msn*	*Rabgap1l*
Hp	*Cxcl17*	*Pi4k2b*	*Nusap1*	*Gbp8*	*Bcl2a1d*	*Degs1*	*Myl12a*	*Nme2*
Cisd1	*Chad*	*Adk*	*Prc1*	*Gbp5*	*Flt3l*	*Steap4*	*Cldn18*	*Cxcl15*
Gstm2	*Kcne3*	*Rab27b*	*Ccnb1*	*Bst2*	*AW112010*	*Tpi1*	*Ndnf*	*Cldn2*
Scgb1c1	*Tspan13*	*Usp8*	*Tpx2*	*Gbp6*	*Lbhd2*	*Areg*	*Itgb6*	*Pf4*
Gsta2	*Reg3g*	*Mid1ip1*	*Racgap1*	*Rtp4*	*Tnfrsf4*	*Ccdc186*	*Dpysl2*	*Snx7*
Aldh1a7	*Agr2*	*Lyz2*	*Hmgb2*	*Tap2*	*Tnfaip2*	*Flt1*	*Rtkn2*	*Dmkn*
Alas1	*Tff2*	*Cd36*	*Anln*	*Wars*	*Csf2rb2*	*Aldoa*	*Hs2st1*	*Bex2*
Ftl1	*Scgb3a2*	*Ppp1r14c*	*Kif11*	*Psme2*	*Rac2*	*Slc4a4*	*Qk*	*Noxo1*
Gstk1	*Anxa8*	*Car8*	*Knl1*	*Gbp9*	*Gpr35*	*Ptges*	*Col4a3*	*Cltrn*
Sec14l3	*Gsto1*	*Abcd3*	*Pclaf*	*Ly6i*	*Relt*	*Thbs1*	*X2200002D01Rik*	*Sh3bgrl3*
Fmo2	*Lgr6*	*Acot7*	*Mis18bp1*	*Iigp1*	*Cfp*	*Lipa*	*Clic5*	*X1110008P14Rik*
Col23a1	*Cldn22*	*Sftpa1*	*Hmmr*	*Isg15*	*Pglyrp2*	*Inhbb*	*Pdgfa*	*Cd302*

Table 20. Marker genes used in the tumor initiation experiment

Club	AT2	AT1	Proliferation	Tumor
Scgb1a1	*Sftpc*	*Ager*	*Hmgb2*	*Meg3*
Cyp2f2	*Sftpa1*	*Cldn18*	*D17H6S56E-5*	*Areg*
Hp	*Slc34a2*	*Sparc*	*Mki67*	*Pf4*
Cbr2	*Sftpb*	*Emp2*	*Top2a*	*Kng2*
Selenbp1	*Lyz2*	*Akap5*	*Ptma*	*Id2*
Mgst1	*Napsa*	*Timp3*	*Smc4*	*Serpine1*
Fmo2	*Hc*	*Clic5*	*H2afz*	*Cdkn1a*
Scgb3a2	*Sftpd*	*Igfbp7*	*Tubb5*	*Rian*
Retnla	*Cxcl15*	*Vegfa*	*Prc1*	*Dmkn*
Alas1	*Scd1*	*Msn*	*Stmn1*	*Tspan8*

Appendix II

Combination therapy of crizotinib and super-paramagnetic iron oxide nanoparticles reprograms the tumor microenvironment and reduces lung cancer relapse

Natalie K. Horvat*, Sara Chocarro*, Oriana Marques, Tobias Bauer, Ryuyen Qiu, Alberto Diaz-Jimenez, Barbara Helm, Yuanyuan Chen, Stefan Sawall, Richard Sparla, Ursula Klingmüller, Matthias Barz, Matthias W. Hentze*, Rocío Sotillo*, Martina U. Muckenthaler*

*equal contribution

During my PhD thesis I collaborated with Dr. Natalie Horvat, from the lab of Prof. Martina Muckenthaler, in a project which has resulted in the manuscript: "Combination therapy of crizotinib and super-paramagnetic iron oxide nanoparticles reprograms the tumor microenvironment and reduces lung cancer relapse", to which I am co-first author and which is currently being submitted for publication.

Tumor-associated macrophages (TAMs) promote lung cancer progression, but can be engineered to have a pro-inflammatory tumor suppressive phenotype. Such reprogramming can be accomplished by iron loading of TAMs, which is associated with a better prognosis for individuals with lung cancer. In this study, we applied super-paramagnetic iron oxide nanoparticles (SPIONs)[214] to TAMs and demonstrated that their administration induced the release of cytokines and reactive oxygen species with pro-tumoricidal action. We further demonstrated that SPIONs administered to *Eml4-Alk* V1 tumor bearing-mice greatly altered the immunosuppressive lung tumor microenvironment in a way that promoted the recruitment of $CD3^+CD8^+$ T cells.

In this project I mainly contributed to the *in vivo* experiments, in which we characterized the effect of the SPION treatment as a monotherapy or in combination with crizotinib, a targeted therapy against ALK. In these experiments we examined the tumor growth and histological characteristics of SPION-treated compared to non-treated mice, and evaluated the tumor microenvironment landscape of these mice. We concluded that SPIONs are a promising therapy for lung cancer patients as an adjuvant treatment, which could help improve the current standards of care.

Acknowledgements

I would like to express my gratitude to my supervisor, Rocio Sotillo, for her invaluable guidance, support, and encouragement throughout my PhD studies. Her insights and feedback have been very instrumental in shaping my research and professional development.

I also extend my heartfelt appreciation to my TAC committee members, Prof. Elmar Schiebel, Prof. Ursula Klingmüller, and MD. PhD. Georgios Stathopoulos, for their time, expertise, and constructive feedback. Their thoughtful comments and suggestions during the TAC meetings have greatly enriched my work. Moreover, I want to thank Dr. Ana Banito for agreeing to participate in my PhD examination committee.

I wish to express my gratitude to the DKFZ Core Facilities, especially the microscopy, flow cytometry, animal facility and Genomics and Proteomics teams for their exceptional assistance and contribution to the improvement of the quality of my work. In addition, I appreciate the help of the whole DKFZ Single-Cell Sequencing Open Lab for helping me in designing and conducting the scRNA-Seq experiments.

My collaborator, Natalie Horvat, deserves a special mention. Together, we worked on a project outside of my thesis, and her insights, enthusiasm, and expertise were invaluable in pushing the project forward. Working with her was a very enriching experience, and I am grateful we got to work and enjoy the project together. Undoubtedly, the best experience was that our collaboration turned into a friendship that I deeply cherish.

I would like to thank my lab mates for their help and collaboration. Their enthusiasm and dedication have made my PhD journey more stimulating. Thank you for all the feedback and discussions during the lab meetings and casual conversations in the lab. They have for sure helped me move forward with my project it the most difficult moments.

My family and boyfriend have been a constant source of love, encouragement, and inspiration. I am deeply grateful to them for their support, especially during the challenging times. Their love and encouragement have given me the strength and

motivation to pursue my dreams and overcome obstacles along the way. I am particularly grateful to my parents, who have always been my role models and source of inspiration. To my sister, who has always been by my side. Her unwavering support and encouragement have been a mental relief for me during the most challenging times. She has been a great listener, a trusted advisor, and a wonderful friend throughout my PhD. And to Manuel, to whom I am indebted for his patience, love and understanding. Gracias por ayudarme a disfrutar de cada pequeño detalle y hacer felices mis peores días. Os quiero.

Finally, I would like to express my sincere appreciation and love to all my friends, both here in Heidelberg and in San Sebastian, for their encouragement, and moral support. Their friendship has been a precious gift that has helped me navigate the ups and downs of the PhD journey. Special mention to Lorena Salgueiro, how has helped me during these 5 years and has been by my side no matter what.

References

1. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *Ca Cancer J Clin* 71, 209–249 (2021).

2. Collins, L. G., Haines, C., Perkel, R. & Enck, R. E. Lung cancer: diagnosis and management. *Am Fam Physician* 75, 56–63 (2007).

3. Attanoos, R. L. Asbestos-Related Lung Disease. *Surg Pathology Clin* 3, 109–127 (2010).

4. Molina, J. R., Yang, P., Cassivi, S. D., Schild, S. E. & Adjei, A. A. Non-Small Cell Lung Cancer: Epidemiology, Risk Factors, Treatment, and Survivorship. *Mayo Clin Proc* 83, 584–594 (2008).

5. Castelletti, N. *et al.* Risk of lung adenocarcinoma from smoking and radiation arises in distinct molecular pathways. *Carcinogenesis* 40, 1240–1250 (2019).

6. Brenner, D. J. *et al.* Cancer risks attributable to low doses of ionizing radiation: Assessing what we really know. *Proc National Acad Sci* 100, 13761–13766 (2003).

7. Moore, J. X., Akinyemiju, T. & Wang, H. E. Pollution and regional variations of lung cancer mortality in the United States. *Cancer Epidemiol* 49, 118–127 (2017).

8. Lababede, O. & Meziane, M. A. The 8th Edition of TNM Staging of Lung Cancer: Reference Chart and Diagrams. *Oncol* 23, 844–848 (2018).

9. Rami-Porta, R. Staging Manual in Thoracic Oncology, 2nd Edition. *North Fort Myers, FL: Editorial Rx Press* (2016).

10. Nicholson, A. G. *et al.* The 2021 WHO Classification of Lung Tumors: Impact of Advances Since 2015. *J Thorac Oncol* 17, 362–387 (2022).

11. DuPage, M., Dooley, A. L. & Jacks, T. Conditional mouse lung cancer models using adenoviral or lentiviral delivery of Cre recombinase. *Nat Protoc* 4, 1064–1072 (2009).

12. Brierley, ames D., Gospodarowicz, M. K. & Wittekind, C. TNM Classification of Malignant Tumours*, 8th edition*. (2016).

13. Amin, M. B. *et al.* Cancer Staging Manual. Springer Cham (2016).

14. Barletta, J. A., Yeap, B. Y. & Chirieac, L. R. Prognostic significance of grading in lung adenocarcinoma. *Cancer* 116, 659–69 (2009).

15. Rekhtman, N. Neuroendocrine Tumors of the Lung: An Update. *Arch Pathol Lab Med* 134, 1628–1638 (2010).

16. Bubendorf, L., Lantuejoul, S., Langen, A. J. de & Thunnissen, E. Nonsmall cell lung carcinoma: diagnostic difficulties in small biopsies and cytological specimens. *European Respir Rev* 26, 170007 (2017).

17. Kargi, A., Gurel, D. & Tuna, B. The Diagnostic Value of TTF-1, CK 5/6, and p63 Immunostaining in Classification of Lung Carcinomas. *Applied Immunohistochemistry & Molecular Morphology* 15, 415–420 (2016).

18. Pan, Y., Han, H., Labbe, K. E., Zhang, H. & Wong, K.-K. Recent advances in preclinical models for lung squamous cell carcinoma. *Oncogene* 40, 2817–2829 (2021).

19. Niu, Z., Jin, R., Zhang, Y. & Li, H. Signaling pathways and targeted therapies in lung squamous cell carcinoma: mechanisms and clinical trials. *Signal Transduct Target Ther* 7, 353 (2022).

20. Meuwissen, R. & Berns, A. Mouse models for human lung cancer. *Gene Dev* 19, 643–664 (2005).

21. Li, S. *et al.* Comprehensive Characterization of Oncogenic Drivers in Asian Lung Adenocarcinoma. *J Thorac Oncol* 11, 2129–2140 (2016).

22. Travis, W. D. Pathology of Lung Cancer. *Clin Chest Med* 32, 669–692 (2011).

23. Yuan, M., Huang, L.-L., Chen, J.-H., Wu, J. & Xu, Q. The emerging treatment landscape of targeted therapy in non-small-cell lung cancer. *Signal Transduct Target Ther* 4, 61 (2019).

24. Spagnuolo, A., Maione, P. & Gridelli, C. The treatment of advanced non-small cell lung cancer harboring KRAS mutation: a new class of drugs for an old target—a narrative review. *Transl Lung Cancer Res* 0, 0–0 (2022).

25. Iwahara, T. *et al.* Molecular characterization of ALK, a receptor tyrosine kinase expressed specifically in the nervous system. *Oncogene* 14, 439–449 (1997).

26. Bayliss, R., Choi, J., Fennell, D. A., Fry, A. M. & Richards, M. W. Molecular mechanisms that underpin EML4-ALK driven cancers and their response to targeted drugs. *Cell Mol Life Sci* 73, 1209–1224 (2016).

27. Morris, S. W. *et al.* Fusion of a Kinase Gene, ALK, to a Nucleolar Protein Gene, NPM, in Non-Hodgkin's Lymphoma. *Science* 263, 1281–1284 (1994).

28. Reshetnyak, A. V. *et al.* Augmentor α and β (FAM150) are ligands of the receptor tyrosine kinases ALK and LTK: Hierarchy and specificity of ligand–receptor interactions. *Proc National Acad Sci* 112, 15862–15867 (2015).

29. Schonherr, C., Hallberg, B. & Palmer, R. Anaplastic Lymphoma Kinase in Human Cancer. *Crit Rev Oncogenesis* 17, 123–143 (2012).

30. Ducray, S. P., Natarajan, K., Garland, G. D., Turner, S. D. & Egger, G. The Transcriptional Roles of ALK Fusion Proteins in Tumorigenesis. *Cancers* 11, 1074 (2019).

31. Chiarle, R., Voena, C., Ambrogio, C., Piva, R. & Inghirami, G. The anaplastic lymphoma kinase in the pathogenesis of cancer. *Nat Rev Cancer* 8, 11–23 (2008).

32. Chia, P. L., Mitchell, P., Dobrovic, A. & John, T. Prevalence and natural history of ALK positive non-small-cell lung cancer and the clinical impact of targeted therapy with ALK inhibitors. *Clin Epidemiology* 6, 423–432 (2014).

33. Richards, M. W. *et al.* Microtubule association of EML proteins and the EML4-ALK variant 3 oncoprotein require an N-terminal trimerization domain. *The Biochemical journal* 467, 529–536 (2015).

34. Eichenmüller, B., Everley, P., Palange, J., Lepley, D. & Suprenant, K. A. The Human EMAP-like Protein-70 (ELP70) Is a Microtubule Destabilizer That Localizes to the Mitotic Apparatus*. *J Biol Chem* 277, 1301–1309 (2002).

35. Shaw, A. T. *et al.* Clinical Features and Outcome of Patients With Non–Small-Cell Lung Cancer Who Harbor EML4-ALK. *J Clin Oncol* 27, 4247–4253 (2009).

36. Soda, M. *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. *Nature* 448, 561–566 (2007).

37. Efanov, A. A. *et al.* Investigation of the Relationship Between Radiation Dose and Gene Mutations and Fusions in Post-Chernobyl Thyroid Cancer. *Jnci J National Cancer Inst* 110, 371–378 (2017).

38. Lin, J. J., Riely, G. J. & Shaw, A. T. Targeting ALK: Precision Medicine Takes on Drug Resistance. *Cancer Discov* 7, 137–155 (2017).

39. Lu, S. *et al.* Comparison of EML4-ALK fusion gene positive rate in different detection methods and samples of non-small cell lung cancer. *J Cancer* 11, 1525–1531 (2020).

40. Zou, H. Y. *et al.* An Orally Available Small-Molecule Inhibitor of c-Met, PF-2341066, Exhibits Cytoreductive Antitumor Efficacy through Antiproliferative and Antiangiogenic Mechanisms. *Cancer Res* 67, 4408–4417 (2007).

41. Kim, D.-W. *et al.* Activity and safety of ceritinib in patients with ALK-rearranged non-small-cell lung cancer (ASCEND-1): updated results from the multicentre, open-label, phase 1 trial. *Lancet Oncol* 17, 452–463 (2016).

42. Ou, S.-H. I. *et al.* Alectinib in Crizotinib-Refractory ALK- Rearranged Non–Small-Cell Lung Cancer: A Phase II Global Study. *J Clin Oncol* 34, 661–668 (2016).

43. Shaw, A. T. *et al.* Alectinib in ALK-positive, crizotinib-resistant, non-small-cell lung cancer: a single-group, multicentre, phase 2 trial. *Lancet Oncol* 17, 234–242 (2016).

44. Pacheco, J. M. *et al.* Natural History and Factors Associated with Overall Survival in Stage IV ALK-Rearranged Non–Small Cell Lung Cancer. *J Thorac Oncol* 14, 691–700 (2019).

45. Gainor, J. F. *et al.* Molecular Mechanisms of Resistance to First- and Second-Generation ALK Inhibitors in ALK-Rearranged Lung Cancer. *Cancer discovery* 6, 1118–1133 (2016).

46. Zou, H. Y. *et al.* PF-06463922, an ALK/ROS1 Inhibitor, Overcomes Resistance to First and Second Generation ALK Inhibitors in Preclinical Models. *Cancer Cell* 28, 70–81 (2015).

47. Pan, Y., Deng, C., Qiu, Z., Cao, C. & Wu, F. The Resistance Mechanisms and Treatment Strategies for ALK-Rearranged Non-Small Cell Lung Cancer. *Frontiers Oncol* 11, 713530 (2021).

48. Takegawa, N. *et al.* Transformation of ALK rearrangement-positive adenocarcinoma to small-cell lung cancer in association with acquired resistance to alectinib. *Ann Oncol* 27, 953–955 (2016).

49. Fujita, S., Masago, K., Katakami, N. & Yatabe, Y. Transformation to SCLC after Treatment with the ALK Inhibitor Alectinib. *J Thorac Oncol* 11, e67–e72 (2016).

50. Cha, Y. J., Cho, B. C., Kim, H. R., Lee, H.-J. & Shim, H. S. A Case of ALK-Rearranged Adenocarcinoma with Small Cell Carcinoma-Like Transformation and Resistance to Crizotinib. *J Thorac Oncol* 11, e55–e58 (2016).

51. Wang, F., Qin, J., Xie, F., Wu, Q. & Lu, H. Transformation of EML4-ALK fusion-positive adenocarcinoma into squamous cell carcinoma in association with acquired resistance to crizotinib. *Lung Cancer* 140, 118–120 (2020).

52. Gong, J. *et al.* Squamous Cell Transformation of Primary Lung Adenocarcinoma in a Patient With EML4-ALK Fusion Variant 5 Refractory to ALK Inhibitors. *J Natl Compr Canc Ne* 17, 297–301 (2019).

53. Kaiho, T., Nakajima, T., Iwasawa, S., Yonemori, Y. & Yoshino, I. ALK Rearrangement Adenocarcinoma with Histological Transformation to Squamous Cell Carcinoma Resistant to Alectinib and Ceritinib. *Oncotargets Ther* 13, 1557–1560 (2020).

54. Ball, M. *et al.* Histological and Molecular Plasticity of ALK-positive Non-Small-Cell Lung Cancer under Targeted Therapy - a Case Report. *Mol Case Stud* 8, mcs.a006156 (2022).

55. Schenk, E. L. Narrative review: immunotherapy in anaplastic lymphoma kinase (ALK)+ lung cancer—current status and future directions. *Transl Lung Cancer Res* 12, 322–336 (2023).

56. Bylicki, O. *et al.* Efficacy and safety of programmed cell-death-protein-1 and its ligand inhibitors in pretreated patients with epidermal growth-factor receptor-mutated or

anaplastic lymphoma kinase-translocated lung adenocarcinoma. *Medicine* 99, e18726 (2020).

57. Heo, J. Y. *et al.* The efficacy of immune checkpoint inhibitors in anaplastic lymphoma kinase-positive non-small cell lung cancer. *Thorac Cancer* 10, 2117–2123 (2019).

58. Mazieres, J. *et al.* Immune checkpoint inhibitors for patients with advanced lung cancer and oncogenic driver alterations: results from the IMMUNOTARGET registry. *Ann Oncol* 30, 1321–1328 (2019).

59. Felip, E. *et al.* Ceritinib plus Nivolumab in Patients with Advanced ALK-Rearranged Non–Small Cell Lung Cancer: Results of an Open-Label, Multicenter, Phase 1B Study. *J Thorac Oncol* 15, 392–403 (2020).

60. Patel, S. P. *et al.* Phase Ib Study of Crizotinib plus Pembrolizumab in Patients with Previously Untreated Advanced Non-Small Cell Lung Cancer with ALK Translocation. *Oncol* 25, 562-e1012 (2020).

61. Chalmers, A. W. *et al.* Phase I Trial of Targeted EGFR or ALK Therapy with Ipilimumab in Metastatic NSCLC with Long-Term Follow-Up. *Target Oncol* 14, 417–421 (2019).

62. Sankar, K., Nagrath, S. & Ramnath, N. Immunotherapy for ALK-Rearranged Non-Small Cell Lung Cancer: Challenges Inform Promising Approaches. *Cancers* 13, 1476 (2021).

63. Voena, C. *et al.* Efficacy of a Cancer Vaccine against ALK-Rearranged Lung Tumors. *Cancer Immunol Res* 3, 1333–1343 (2015).

64. Sabir, S. R., Yeoh, S., Jackson, G. & Bayliss, R. EML4-ALK Variants: Biological and Molecular Properties, and the Implications for Patients. *Cancers* 9, (2017).

65. Lin, J. J. *et al.* Impact of EML4-ALK Variant on Resistance Mechanisms and Clinical Outcomes in ALK-Positive Lung Cancer. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* 36, 1199–1206 (2018).

66. Zhou, X. *et al.* Molecular and clinical analysis of Chinese patients with anaplastic lymphoma kinase (ALK)-rearranged non-small cell lung cancer. *Cancer Sci* 110, 3382–3390 (2019).

67. Camidge, D. R. *et al.* Correlation of baseline molecular and clinical variables with ALK inhibitor efficacy in ALTA-1L. *J Clin Oncol* 38, 9517–9517 (2020).

68. Yoshida, T. *et al.* Differential Crizotinib Response Duration Among ALK Fusion Variants in ALK-Positive Non–Small-Cell Lung Cancer. *J Clin Oncol* 34, 3383–3389 (2016).

69. Christopoulos, P. *et al.* EML4-ALK fusion variant V3 is a high-risk feature conferring accelerated metastatic spread, early treatment failure and worse overall survival in ALK+ non-small cell lung cancer. *International journal of cancer* 142, 2589–2598 (2018).

70. Richards, M. W. *et al.* Crystal structure of EML1 reveals the basis for Hsp90 dependence of oncogenic EML4-ALK by disruption of an atypical β-propeller domain. *Proc National Acad Sci* 111, 5195–5200 (2014).

71. Heuckmann, J. M. *et al.* Differential protein stability and ALK inhibitor sensitivity of EML4-ALK fusion variants. *Clinical cancer research: an official journal of the American Association for Cancer Research* 18, 4682–4690 (2012).

72. Lewis, W. E. *et al.* Brief Report: Efficacy of Targeted Inhibitors in Metastatic Lung Squamous Cell Carcinoma with EGFR or ALK Alterations. *Jto Clin Res Reports* 2, 100237 (2021).

73. Watanabe, J. *et al.* Clinical features of squamous cell lung cancer with anaplastic lymphoma kinase (ALK)-rearrangement: a retrospective analysis and review. *Oncotarget* 9, 24000–24013 (2018).

74. Tamiya, A., Shimizu, S. & Atagi, S. A Case of Squamous Cell Carcinoma Harboring an EML4-ALK Rearrangement that Was Unsuccessfully Treated with the ALK Inhibitor Alectinib. *J Thorac Oncol* 10, e74–e75 (2015).

75. Liang, L. *et al.* Efficacy and Drug Resistance Analysis of ALK Inhibitors in Combination with Stereotactic Body Radiation Therapy for Treating Lung Squamous Carcinoma Patient Harboring EML4-ALK Rearrangement. *Oncotargets Ther* 14, 5385–5389 (2021).

76. Shiihara, J. *et al.* Dramatic response to alectinib in a patient with ALK-rearranged squamous cell lung cancer. *Thorac Cancer* 12, 2420–2423 (2021).

77. Fry, A. M., O'Regan, L., Montgomery, J., Adib, R. & Bayliss, R. EML proteins in microtubule regulation and human disease. *Biochemical Society transactions* 44, 1281–1288 (2016).

78. O'Regan, L. *et al.* EML4–ALK V3 oncogenic fusion proteins promote microtubule stabilization and accelerated migration through NEK9 and NEK7. *J Cell Sci* 133, jcs241505 (2020).

79. Sampson, J., Richards, M. W., Choi, J., Fry, A. M. & Bayliss, R. Phase-separated foci of EML4-ALK facilitate signalling and depend upon an active kinase conformation. *Embo Rep* 22, e53693 (2021).

80. Qin, Z. *et al.* Phase separation of EML4–ALK in firing downstream signaling and promoting lung tumorigenesis. *Cell Discov* 7, 33 (2021).

81. Tulpule, A. *et al.* Kinase-mediated RAS signaling via membraneless cytoplasmic protein granules. *Cell* 184, 2649-2664.e18 (2021).

82. Lucken, K. *et al.* EML4-ALK Variant 3 Promotes Mitotic Errors and Spindle Assembly Checkpoint Deficiency Leading to Increased Microtubule Poison Sensitivity. *Mol Cancer Res* 20, 854–866 (2022).

83. Liebler, J. M. *et al.* Combinations of differentiation markers distinguish subpopulations of alveolar epithelial cells in adult lung. *Am J Physiol-lung C* 310, L114–L120 (2016).

84. Barnes, P. J. Club Cells, Their Secretory Protein, and COPD. *Chest* 147, 1447–1448 (2015).

85. Rawlins, E. L. *et al.* The role of Scgb1a1+ Clara cells in the long-term maintenance and repair of lung airway, but not alveolar, epithelium. *Cell stem cell* 4, 525–534 (2009).

86. Tata, P. R. *et al.* Dedifferentiation of committed epithelial cells into stem cells in vivo. *Nature* 503, 218–223 (2013).

87. Ruysseveldt, E., Martens, K. & Steelant, B. Airway Basal Cells, Protectors of Epithelial Walls in Health and Respiratory Diseases. *Frontiers Allergy* 2, 787128 (2021).

88. Whitsett, J. A. Airway Epithelial Differentiation and Mucociliary Clearance. *Ann Am Thorac Soc* 15, S143–S148 (2018).

89. Mou, H. *et al.* Airway basal stem cells generate distinct subpopulations of PNECs. *Cell Reports* 35, 109011 (2021).

90. Sutherland, K. D. *et al.* Cell of Origin of Small Cell Lung Cancer: Inactivation of Trp53 and Rb1 in Distinct Cell Types of Adult Mouse Lung. *Cancer Cell* 19, 754–764 (2011).

91. Park, K.-S. *et al.* Characterization of the cell of origin for small cell lung cancer. *Cell Cycle* 10, 2806–2815 (2014).

92. Reynolds, S. D., Giangreco, A., Power, J. H. T. & Stripp, B. R. Neuroepithelial Bodies of Pulmonary Airways Serve as a Reservoir of Progenitor Cells Capable of Epithelial Regeneration. *Am J Pathology* 156, 269–278 (2000).

93. Song, H. *et al.* Functional characterization of pulmonary neuroendocrine cells in lung development, injury, and tumorigenesis. *Proc National Acad Sci* 109, 17531–17536 (2012).

94. Barr, J. *et al.* Injury-induced pulmonary tuft cells are heterogenous, arise independent of key Type 2 cytokines, and are dispensable for dysplastic repair. *Elife* 11, e78074 (2022).

95. Montoro, D. T. *et al.* A revised airway epithelial hierarchy includes CFTR-expressing ionocytes. *Nature* 560, 319–324 (2018).

96. Kim, C. F. B. *et al.* Identification of Bronchioalveolar Stem Cells in Normal Lung and Lung Cancer. *Cell* 121, 823–835 (2005).

97. Liu, Q. *et al.* Lung regeneration by multipotent stem cells residing at the bronchioalveolar-duct junction. *Nat Genet* 51, 728–738 (2019).

98. Salwig, I. *et al.* Bronchioalveolar stem cells are a main source for regeneration of distal lung epithelia in vivo. *The EMBO journal* 38, (2019).

99. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* 587, 619–625 (2020).

100. Sikkema, L. *et al.* An integrated cell atlas of the human lung in health and disease. *Biorxiv* 2022.03.10.483747 (2022).

101. Jain, R. *et al.* Plasticity of Hopx(+) type I alveolar cells to regenerate type II cells in the lung. *Nature Communications* 6, 6727 (2015).

102. Barkauskas, C. E. *et al.* Type 2 alveolar cells are stem cells in adult lung. *The Journal of clinical investigation* 123, 3025–3036 (2013).

103. Zacharias, W. J. *et al.* Regeneration of the lung alveolus by an evolutionarily conserved epithelial progenitor. *Nature* 555, 251–255 (2018).

104. Strunz, M. *et al.* Alveolar regeneration through a Krt8+ transitional stem cell state that persists in human lung fibrosis. *Nature Communications* 11, 3559 (2020).

105. Kathiriya, J. J., Brumwell, A. N., Jackson, J. R., Tang, X. & Chapman, H. A. Distinct Airway Epithelial Stem Cells Hide among Club Cells but Mobilize to Promote Alveolar Regeneration. *Cell stem cell* 26, 346-358.e4 (2020).

106. Kobayashi, Y. *et al.* Persistence of a regeneration-associated, transitional alveolar epithelial cell state in pulmonary fibrosis. *Nature cell biology* 22, 934–946 (2020).

107. Choi, J. *et al.* Inflammatory Signals Induce AT2 Cell-Derived Damage-Associated Transient Progenitors that Mediate Alveolar Regeneration. *Cell stem cell* 27, 366-382.e7 (2020).

108. Kathiriya, J. J. *et al.* Human alveolar type 2 epithelium transdifferentiates into metaplastic KRT5+ basal cells. *Nat Cell Biol* 24, 10–23 (2022).

109. Vaughan, A. E. *et al.* Lineage-negative Progenitors Mobilize to Regenerate Lung Epithelium after Major Injury. *Nature* 517, 621–625 (2015).

110. Xi, Y. *et al.* Local lung hypoxia determines epithelial fate decisions during alveolar regeneration. *Nature cell biology* 19, 904–914 (2017).

111. Kumar, P. A. *et al.* Distal Airway Stem Cells Yield Alveoli In Vitro and during Lung Regeneration following H1N1 Influenza Infection. *Cell* 147, 525–538 (2011).

112. Zuo, W. *et al.* p63(+)Krt5(+) distal airway stem cells are essential for lung regeneration. *Nature* 517, 616–620 (2015).

113. Taylor, M. S. *et al.* A Conserved Distal Lung Regenerative Pathway in Acute Lung Injury. *Am J Pathology* 188, 1149–1160 (2018).

114. Nagy, A. Cre recombinase: the universal reagent for genome tailoring. *Genesis New York N Y 2000* 26, 99–109 (2000).

115. Feil, R., Wagner, J., Metzger, D. & Chambon, P. Regulation of Cre Recombinase Activity by Mutated Estrogen Receptor Ligand-Binding Domains. *Biochem Bioph Res Co* 237, 752–757 (1997).

116. Das, A. T., Tenenbaum, L. & Berkhout, B. Tet-On Systems For Doxycycline-inducible Gene Expression. *Curr Gene Ther* 16, 156–167 (2016).

117. Mainardi, S. *et al.* Identification of cancer initiating cells in K-Ras driven lung adenocarcinoma. *Proceedings of the National Academy of Sciences of the United States of America* 111, 255–260 (2014).

118. Jackson, E. L. *et al.* Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras. *Gene Dev* 15, 3243–3248 (2001).

119. Sutherland, K. D. *et al.* Multiple cells-of-origin of mutant K-Ras-induced mouse lung adenocarcinoma. *Proceedings of the National Academy of Sciences of the United States of America* 111, 4952–4957 (2014).

120. Marjanovic, N. D. *et al.* Emergence of a High-Plasticity Cell State during Lung Cancer Evolution. *Cancer Cell* 38, 229-246.e13 (2020).

121. LaFave, L. M. *et al.* Epigenomic State Transitions Characterize Tumor Progression in Mouse Lung Adenocarcinoma. *Cancer Cell* 38, 212-228.e13 (2020).

122. Jackson, E. L. *et al.* The differential effects of mutant p53 alleles on advanced murine lung cancer. *Cancer research* 65, 10280–10288 (2005).

123. Perl, A.-K. T., Tichelaar, J. W. & Whitsett, J. A. Conditional gene expression in the respiratory epithelium of the mouse. *Transgenic Res* 11, 21–29 (2002).

124. Politi, K. *et al.* Lung adenocarcinomas induced in mice by mutant EGF receptors foundin human lung cancers respondto a tyrosine kinase inhibitor orto down-regulation of the receptors. *Gene Dev* 20, 1496–1510 (2006).

125. Fisher, G. H. *et al.* Induction and apoptotic regression of lung adenocarcinomas by regulation of a K-Ras transgene in the presence and absence of tumor suppressor genes. *Gene Dev* 15, 3249–3262 (2001).

126. Rawlins, E. L., Ostrowski, L. E., Randell, S. H. & Hogan, B. L. M. Lung development and repair: Contribution of the ciliated lineage. *Proc National Acad Sci* 104, 410–417 (2007).

127. Chapman, H. A. *et al.* Integrin α6β4 identifies an adult distal lung epithelial population with regenerative potential in mice. *J Clin Invest* 121, 2855–2862 (2011).

128. Indra, A. K. *et al.* Temporally-controlled site-specific mutagenesis in the basal layer of the epidermis: comparison of the recombinase activity of the tamoxifen-inducible Cre-ER(T) and Cre-ER(T2) recombinases. *Nucleic acids research* 27, 4324–4327 (1999).

129. Chen, Y. *et al.* Club cells employ regeneration mechanisms during lung tumorigenesis. *Nat Commun* 13, 4557 (2022).

130. Soda, M. *et al.* A mouse model for EML4-ALK-positive lung cancer. *Proc National Acad Sci* 105, 19893–19897 (2008).

131. Pyo, K. H. *et al.* Establishment of a Conditional Transgenic Mouse Model Recapitulating EML4-ALK–Positive Human Non–Small Cell Lung Cancer. *J Thorac Oncol* 12, 491–500 (2017).

132. Blasco, R. B. *et al.* Simple and Rapid In Vivo Generation of Chromosomal Rearrangements using CRISPR/Cas9 Technology. *Cell Reports* 9, 1219–1227 (2014).

133. Maddalo, D. *et al.* In vivo engineering of oncogenic chromosomal rearrangements with the CRISPR/Cas9 system. *Nature* 1–15 (2014).

134. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* 173, 291-304.e6 (2018).

135. Johnson, L. *et al.* Somatic activation of the K-ras oncogene causes early onset lung cancer in mice. *Nature* 410, 1111–1116 (2001).

136. Xu, X. *et al.* Evidence for type II cells as cells of origin of K-Ras-induced distal lung adenocarcinoma. *Proceedings of the National Academy of Sciences of the United States of America* 109, 4910–4915 (2012).

137. Nagaraj, A. S. *et al.* Cell of Origin Links Histotype Spectrum to Immune Microenvironment Diversity in Non-small-Cell Lung Cancer Driven by Mutant Kras and Loss of Lkb1. *Cell Reports* 18, 673–684 (2017).

138. Lin, C. *et al.* Alveolar Type II Cells Possess the Capability of Initiating Lung Tumor Development. *Plos One* 7, e53817 (2012).

139. Spella, M. *et al.* Club cells form lung adenocarcinomas and maintain the alveoli of adult mice. *eLife* 8, (2019).

140. Rosigkeit, S. *et al.* Definitive evidence for Club cells as progenitors for mutant Kras/Trp53-deficient lung cancer. *Int. J. Cancer* 149, 1670–1682 (2021).

141. Concepcion, C. P. *et al.* SMARCA4 inactivation promotes lineage-specific transformation and early metastatic features in the lung. *Cancer Discov* candisc.0248.2021 (2021).

142. Williams, M. C. ALVEOLAR TYPE I CELLS: Molecular Phenotype and Development. *Annu Rev Physiol* 65, 669–695 (2003).

143. Brody, J. S. & Williams, M. C. Pulmonary Alveolar Epithelial Cell Differentiation. *Annu Rev Physiol* 54, 351–371 (1992).

144. Yang, M. *et al.* Alveolar epithelial type 1 cells serve as a cell of origin for lung adenocarcinoma with distinct molecular and phenotypic presentation. *Biorxiv* 2022.10.29.514334 (2022).

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

146. Kargl, J. *et al.* Neutrophils dominate the immune cell composition in non-small cell lung cancer. *Nat Commun* 8, 14381 (2017).

147. Skoulidis, F. *et al.* Co-occurring Genomic Alterations Define Major Subsets of KRAS-Mutant Lung Adenocarcinoma with Distinct Biology, Immune Profiles, and Therapeutic Vulnerabilities. *Cancer Discov* 5, 860–877 (2015).

148. Best, S. A. *et al.* Distinct initiating events underpin the immune and metabolic heterogeneity of KRAS-mutant lung adenocarcinoma. *Nat Commun* 10, 4190 (2019).

149. Marei, H. E. *et al.* p53 signaling in cancer progression and therapy. *Cancer Cell Int* 21, 703 (2021).

150. Hafner, A., Bulyk, M. L., Jambhekar, A. & Lahav, G. The multiple mechanisms that regulate p53 activity and cell fate. *Nat Rev Mol Cell Bio* 20, 199–210 (2019).

151. Pavletich, N. P., Chambers, K. A. & Pabo, C. O. The DNA-binding domain of p53 contains the four conserved regions and the major mutation hot spots. *Gene Dev* 7, 2556–2564 (1993).

152. Momand, J., Wu, H.-H. & Dasgupta, G. MDM2 — master regulator of the p53 tumor suppressor protein. *Gene* 242, 15–29 (2000).

153. Juven-Gershon, T. & Oren, M. Mdm2: The Ups and Downs. *Mol Med* 5, 71–83 (1999).

154. Hirao, A. *et al.* DNA Damage-Induced Activation of p53 by the Checkpoint Kinase Chk2. *Science* 287, 1824–1827 (2000).

155. Fischer, M. Census and evaluation of p53 target genes. *Oncogene* 36, 3943–3956 (2017).

156. Laronga, C., Yang, H.-Y., Neal, C. & Lee, M.-H. Association of the Cyclin-dependent Kinases and 14-3-3 Sigma Negatively Regulates Cell Cycle Progression*. *J Biol Chem* 275, 23106–23112 (2000).

157. Oda, E. *et al.* Noxa, a BH3-Only Member of the Bcl-2 Family and Candidate Mediator of p53-Induced Apoptosis. *Science* 288, 1053–1058 (2000).

158. Oda, K. *et al.* p53AIP1, a Potential Mediator of p53-Dependent Apoptosis, and Its Regulation by Ser-46-Phosphorylated p53. *Cell* 102, 849–862 (2000).

159. Willis, A., Jung, E. J., Wakefield, T. & Chen, X. Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. *Oncogene* 23, 2330–2338 (2004).

160. Muller, P. A. J. & Vousden, K. H. p53 mutations in cancer. *Nat Cell Biol* 15, 2–8 (2013).

161. Dittmer, D. *et al.* Gain of function mutations in p53. *Nat Genet* 4, 42–46 (1993).

162. Noh, K.-W. *et al.* Molecular breakdown: a comprehensive view of anaplastic lymphoma kinase (ALK)-rearranged non-small cell lung cancer. *The Journal of pathology* 243, 307–319 (2017).

163. Kron, A. *et al.* Impact of TP53 mutation status on systemic treatment outcome in ALK-rearranged non-small-cell lung cancer. *Ann Oncol* 29, 2068–2075 (2018).

164. Christopoulos, P. *et al.* Identification of a highly lethal V3+TP53+ subset in ALK+ lung adenocarcinoma. *Int. J. Cancer* 144, 190–199 (2019).

165. Tao, H. *et al.* Distribution of EML4-ALK fusion variants and clinical outcomes in patients with resected non-small cell lung cancer. *Lung Cancer* 149, 154–161 (2020).

166. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34, 525–527 (2016).

167. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omics J Integr Biology* 16, 284–287 (2012).

168. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15, 550 (2014).

169. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 16, 1289–1296 (2019).

170. Fan, J. *et al.* MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. *Brief Bioinform* 23, (2022).

171. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat Commun* 10, 380 (2019).

172. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol* 19, 15 (2018).

173. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* 177, 1888-1902.e21 (2019).

174. Marino, S., Vooijs, M., Gulden, H. van der, Jonkers, J. & Berns, A. Induction of medulloblastomas in p53-null mutant mice by somatic inactivation of Rb in the external granular layer cells of the cerebellum. *Gene Dev* 14, 994–1004 (2000).

175. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* 45, 593–605 (2007).

176. Platt, R. J. *et al.* CRISPR-Cas9 Knockin Mice for Genome Editing and Cancer Modeling. *Cell* 159, 440–455 (2014).

177. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc National Acad Sci* 102, 15545–15550 (2005).

178. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 34, 267–273 (2003).

179. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* 161, 1202–1214 (2015).

180. Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *eLife* 8, (2019).

181. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res* 49, gkaa970- (2020).

182. Martens, M. *et al.* WikiPathways: connecting communities. *Nucleic Acids Res* 49, gkaa1024- (2020).

183. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* 38, 1408–1414 (2020).

184. Takezawa, K., Okamoto, I., Nishio, K., Jänne, P. A. & Nakagawa, K. Role of ERK-BIM and STAT3-survivin signaling pathways in ALK inhibitor-induced apoptosis in EML4-ALK-positive lung cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* 17, 2140–2148 (2011).

185. Miyanaga, A. *et al.* EML4-ALK induces cellular senescence in mortal normal human cells and promotes anchorage-independent growth in hTERT-transduced normal human cells. *Bmc Cancer* 21, 310 (2021).

186. Camidge, D. R. *et al.* Brigatinib Versus Crizotinib in Advanced ALK Inhibitor–Naive ALK-Positive Non–Small Cell Lung Cancer: Second Interim Analysis of the Phase III ALTA-1L Trial. *J Clin Oncol* 38, 3592–3603 (2020).

187. Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S. & Yang, S.-H. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Mol Ther Nucleic Acids* 4, e264 (2015).

188. Angelidis, I. *et al.* An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature Communications* 10, 963–17 (2019).

189. Ferone, G. *et al.* SOX2 Is the Determining Oncogenic Switch in Promoting Lung Squamous Cell Carcinoma from Different Cells of Origin. *Cancer Cell* 30, 519–532 (2016).

190. Takeda, N. *et al.* Interconversion Between Intestinal Stem Cell Populations in Distinct Niches. *Science* 334, 1420–1424 (2011).

191. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* 172, 1091-1107.e17 (2018).

192. Manno, G. L. *et al.* RNA velocity of single cells. *Nature* 560, 494–498 (2018).

193. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology* 20, 59 (2019).

194. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nat Methods* 19, 159–170 (2022).

195. Yang, D. *et al.* Lineage tracing reveals the phylodynamics, plasticity, and paths of tumor evolution. *Cell* (2022).

196. Zilionis, R. *et al.* Single-Cell Transcriptomics of Human and Mouse Lung Cancers Reveals Conserved Myeloid Populations across Individuals and Species. *Immunity* 50, 1317-1334.e10 (2019).

197. Collisson, E. A. *et al.* Comprehensive molecular profiling of lung adenocarcinoma. *Nature* 511, 543–550 (2014).

198. Jiang, C. *et al.* Serpine 1 induces alveolar type II cell senescence through activating p53-p21-Rb pathway in fibrotic lung disease. *Aging Cell* 16, 1114–1124 (2017).

199. Martelli, M. P. *et al.* EML4-ALK Rearrangement in Non-Small Cell Lung Cancer and Non-Tumor Lung Tissues. *Am J Pathology* 174, 661–670 (2009).

200. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* 34, 184–191 (2016).

201. Wang, H. *et al.* One-Step Generation of Mice Carrying Mutations in Multiple Genes by CRISPR/Cas-Mediated Genome Engineering. *Cell* 153, 910–918 (2013).

202. Moreira, A. L. *et al.* A Grading System for Invasive Pulmonary Adenocarcinoma: A Proposal From the International Association for the Study of Lung Cancer Pathology Committee. *J Thorac Oncol* 15, 1599–1610 (2020).

203. Nguyen, T. T. *et al.* A lepidic gene signature predicts patient prognosis and sensitivity to immunotherapy in lung adenocarcinoma. *Genome Med* 14, 5 (2022).

204. Childress, M. A. *et al.* ALK Fusion Partners Impact Response to ALK Inhibition: Differential Effects on Sensitivity, Cellular Phenotypes, and Biochemical Properties. *Mol Cancer Res Mcr* 16, 1724–1736 (2018).

205. Chuang, T.-P. *et al.* ALK fusion NSCLC oncogenes promote survival and inhibit NK cell responses via SERPINB4 expression. *P Natl Acad Sci Usa* 120, e2216479120 (2023).

206. Wang, J. *et al.* Detection of ALK protein expression in lung squamous cell carcinomas by immunohistochemistry. *J Exp Clin Canc Res* 33, 109 (2014).

207. Caliò, A. *et al.* ALK/EML4 Fusion Gene May Be Found in Pure Squamous Carcinoma of the Lung. *J Thorac Oncol* 9, 729–732 (2014).

208. Takanashi, Y. *et al.* ALK-rearranged squamous cell carcinoma of the lung. *Respirol Case Reports* 3, 105–107 (2015).

209. Ferone, G., Lee, M. C., Sage, J. & Berns, A. Cells of origin of lung cancers: lessons from mouse studies. *Genes & development* 34, 1017–1032 (2020).

210. Foggetti, G. *et al.* Genetic Determinants of EGFR-Driven Lung Cancer Growth and Therapeutic Response In Vivo. *Cancer Discov* 11, 1736–1753 (2021).

211. Cai, H. *et al.* A Functional Taxonomy of Tumor Suppression in Oncogenic KRAS–Driven Lung Cancer. *Cancer Discov* 11, 1754–1773 (2021).

212. Yousefi, M. *et al.* Combinatorial Inactivation of Tumor Suppressors Efficiently Initiates Lung Adenocarcinoma with Therapeutic Vulnerabilities. *Cancer Res* 82, 1589–1602 (2022).

213. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras Provokes Premature Cell Senescence Associated with Accumulation of p53 and p16INK4a. *Cell* 88, 593–602 (1997).

214. Bauer, T. A. *et al.* Core Cross-Linked Polymeric Micelles for Specific Iron Delivery: Inducing Sterile Inflammation in Macrophages. *Adv Healthc Mater* 2100385 (2021).