SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS
DEPUIS 1257

CENTRE DE RECHERCHE
DES CORDELIERS

Sorbonne Université

ED 394 - Physiologie, physiopathologie et thérapeutique
Centre de Recherche des Cordeliers

A framework for characterisation of intra-tumour heterogeneity of lung cancers

Par Lara JERMAN

Thèse de doctorat de BIOINFORMATIQUE et BIOLOGIE DES SYSTÈMES

Dirigée par Aurélien DE REYNIÈS
et Magnus FONTES

Présentée et soutenue publiquement le 20 novembre 2024
devant un jury composé de :

Pr. Marco ALIFANO	President
Dr. Céline LEFEBVRE	Rapporteure
Pr. Jacques COLINGE	Rapporteur
Dr. Laetitia MARISA	Examinatrice
Pr. Magnus FONTES	Co-directeur de thèse
Pr. Aurélien DE REYNIÈS	Directeur de thèse

"Started making it. Had a breakdown. Bon appétit!"

—James Acaster, *The Great British Bake Off*

Acknowledgements

The decision to move to France to pursue my PhD here was the beginning of a special kind of adventure, one that would not be possible without the many people who have shaped it and helped me along the way. My supervisor Aurélien de Reyniès has given me this opportunity, welcomed me into his group at Cordeliers, and led me on this path with unwavering wisdom, patience and kindness, for which I am very grateful. He taught me to challenge what's in front of me, and that more complex does not necessarily mean better. My co-supervisor Magnus Fontes has given me a different outlook, and many opportunities to engage with science in a different context. He taught me that special things can happen when people come together. Thank you to both of you for your generosity - I cherish the time and work that we shared.

Every lab is made by its people, and I have spent my time here around some remarkable ones. Many things would have been much more difficult for me without my fellow "co-sufferers", Clarice and Marine, who have been showing me the way from the very beginning, and Gadea, who joined us a bit later. Thank you for all of your contributions and discussions, which make this a better work. It is also hard to imagine my time the lab without the presence of Adèle, Nicolas, Thierno, Leriem, Sophie, Laetitia, Doria, Jérôme, Hadrien, Mickael, Rosline and Bastien. Thank you all for your company, help and advice; the scientific discussions, the supernatural investigations and the memes. My time here would not have been the same without you.

My friends near and far have been a constant source of support, even across many kilometers, and it's difficult to imagine reaching the finish line without them. So thank you, Klara, for carrying so many boxes, Lidija and Domen, for the group chat debriefs, Maja for the event moral support, Katrīna and Eva for the many PhD memes sent my way, Lea and Manca for the rooftop sessions, Doris for the staircase bootcamps, Vic and Jason for the endless inspiration, and Sabina for the pep talks.

My MIDF family, too, has been with me through all the highs and the lows of the journey. Hoa, Murat, Kamila, Kristýna, Maryam, Bryan, Gabi, Bengisu, Aziza, and everyone else passing through over the years - thank you for the food, the gossip, and the laughter.

I'd also like my Unicorners friends, who never let me write alone, the medical practicioners, under whose care some of the more difficult aspects of the past three years became much more manageable, and all of my teachers, who have led, inspired and suppored me on the way here.

Finally, to my family, thank you for giving me both roots and wings.

To all of you, and too many others to be mentioned here:

Merci pour tout.

Lara Jerman
Paris, November 2024

Abstract

Lung cancer is responsible for the largest number of cancer related deaths worldwide (1.8 million in 2020 estimated by GLOBOCAN). The main histological types of lung cancer include adenocarcinomas (LUAD, 40-50%), squamous cell carcinomas (LUSC, 20-30%), small-cell lung carcinomas (SCLC, 10-15%) and large cell neuroendocrine lung carcinomas (LCNEC, ~3%). The heterogeneity of healthy and cancerous lung tissue has been described at single-cell resolution in several cell atlases. The existing atlases are limited in several ways (in particular, they are hardly usable as a reliable reference for deconvolution), and a unified atlas is still missing.

The aim of my thesis work is to better describe the intra-tumour heterogeneity (ITH) of lung cancers, by proposing a cellular atlas covering the main histological types. In building such an atlas, my objective was that it could be used as a reliable reference for deconvoluting the bulk and spatial transcriptome of lung tumours, thus allowing to properly assess their cellular composition. To achieve this goal, I first worked on three original algorithms for the analysis of single-cell and spatial transcriptome data. I then used this work to build a unified atlas of lung cancer cells by reanalysing a large number of public single-cell transcriptome data.

In the present manuscript, I start by presenting the three methodological approaches: (i) SCherlock is a marker discovery algorithm for single-cell data, that prioritises highly specific markers of cell types, that are suitable candidates for use in identification and deconvolution contexts; (ii) Spattern is an algorithm intended to delineate spatial patterns of different types (e.g. gradient, barrier) based on graph-derived properties of local neighbourhoods on simulated spatial transcriptomics data; (iii) FastCNV is a copy number variation inference algorithm for single-cell and spatial transcriptomic data with increased sensitivity and computation speed, as compared to existing methods.

Finally, I present the draft of a unified lung cancer cell atlas, encompassing the major histological types. The unified lung cancer atlas presented here is, to the best of my knowledge, the first lung cancer cell atlas covering all major lung cancer histologies. It is based on public single-cell transcriptomic datasets and has been reannotated, addressing several issues identified in the source annotations, as well as reaching a level of granularity that allows for highly specific markers of each cell type.

My work opens up the prospect of better describing the cellular communities present in lung tumours, analysing the interactions at play within them, and assessing whether they are of interest in predicting patient prognosis or response to drug treatments.

Résumé

Un cadre pour la caractérisation de l'hétérogénéité intratumorale des cancers du poumon

Le cancer du poumon est responsable du plus grand nombre de décès liés au cancer dans le monde (1,8 million en 2020, selon les estimations de GLOBOCAN). Les principaux types histologiques de cancer du poumon sont les adénocarcinomes (LUAD, 40-50%), les carcinomes épidermoïdes (LUSC, 20-30%), les carcinomes pulmonaires à petites cellules (SCLC, 10-15%) et les carcinomes pulmonaires neuroendocriniens à grandes cellules (LCNEC, ~3%). L'hétérogénéité des tissus pulmonaires sains et cancéreux a été décrite à une résolution unicellulaire dans plusieurs atlas cellulaires. Ces atlas sont limités à plusieurs égards (en particulier, ils sont difficilement utilisables comme référence fiable pour la déconvolution), et un atlas unifié des cancers du poumon fait toujours défaut.

Le but de mon travail de thèse est de mieux décrire l'hétérogénéité intra-tumorale (ITH) des cancers du poumon, en proposant un atlas cellulaire couvrant les principaux types histologiques. En construisant un tel atlas, mon objectif était qu'il puisse être utilisé comme une référence fiable pour déconvoluer le transcriptome bulk et spatial des tumeurs pulmonaires, permettant ainsi d'évaluer correctement leur composition cellulaire. Pour atteindre cet objectif, j'ai d'abord travaillé sur trois algorithmes originaux pour l'analyse des données de transcriptome spatial et de cellule unique. J'ai ensuite utilisé ces travaux pour construire un atlas unifié des cellules cancéreuses du poumon, en réanalysant un grand nombre de données transciptomiques unicellulaires publiques.

Dans le présent manuscrit, je commence par présenter les trois approches méthodologiques : (i) SCherlock est un algorithme de découverte de marqueurs, pour les données unicellulaires; il donne la priorité aux marqueurs hautement spécifiques des types de cellules fournis en entrée, ces marqueurs étant alors des candidats appropriés pour une utilisation dans des contextes d'identification et de déconvolution, (ii) Spattern est un algorithme destiné à identifier des motifs spatiaux de différents types (gradient, barrière, etc.) sur la base des propriétés des voisinages locaux dérivées des graphes sur des données transcriptomiques spatiales simulées, (iii) FastCNV est un algorithme d'inférence des variations du nombre de copies pour les données transcriptomiques spatiales et unicellulaires avec une sensibilité et une vitesse de calcul accrues par rapport aux méthodes existantes.

Je présente ensuite l'ébauche d'un atlas unifié des cellules du cancer du poumon, englobant les principaux types histologiques. L'atlas unifié du cancer du poumon présenté ici est, à ma connaissance,

le premier atlas cellulaire du cancer du poumon couvrant tous les principaux types histologiques de cancer du poumon. Il est basé sur des ensembles de données transcriptomiques unicellulaires publiques et a été réannoté, ce qui a permis de résoudre plusieurs problèmes identifiés dans les annotations sources et d'atteindre un niveau de granularité permettant d'obtenir des marqueurs hautement spécifiques de chaque type de cellule.

Mes travaux ouvrent la perspective de mieux décrire les communautés cellulaires présentes dans les tumeurs pulmonaires, d'analyser les interactions qui s'y jouent et d'évaluer si elles sont intéressantes pour prédire le pronostic des patients ou la réponse aux traitements médicamenteux.

Contents

1 Introduction **19**

1.1 Cancer . . . 19

1.1.1 Hallmarks of cancer . . . 19

1.1.2 Physical traits . . . 21

1.1.3 Tumour microenvironment . . . 22

1.1.3.1 Stromal cell types . . . 22

1.1.3.2 Immune cell types . . . 23

1.1.4 Tumour heterogeneity . . . 25

1.1.4.1 Intra-tumour heterogeneity (ITH) . . . 25

1.1.4.1.1 Genetic subclonality . . . 26

1.1.4.1.2 Plasticity . . . 27

1.2 Omics in cancer research . . . 28

1.2.1 Genomics . . . 29

1.2.2 Proteomics . . . 29

1.2.3 Epigenomics . . . 29

1.2.4 Transcriptomics . . . 30

1.2.4.1 Bulk transcriptomics . . . 30

1.2.4.2 Single-cell transcriptomics . . . 31

1.2.4.3 Spatial transcriptomics . . . 34

1.2.4.4 Clinical use . . . 36

1.3 Lung cancer . . . 37

1.3.1 The human lungs . . . 37

1.3.2 Histological classification of lung cancers . . . 39

1.3.2.1 Mixed histology . . . 41

1.3.3 Environmental factors . . . 41

1.3.4 Genetic factors . . . 41

1.3.5 Standard of care . . . 42

1.3.5.1 NSCLC treatment and immunotherapy . . . 42

1.3.5.2 SCLC treatment and immunotherapy . . . 43

1.3.5.3 LCNEC treatment . . . 44

1.3.6 Molecular classifications . . . 44
1.3.6.1 Adenocarcinoma . . . 44
1.3.6.1.1 Transcriptomic classification . . . 44
1.3.6.1.2 The three transcriptomic subtypes . . . 44
1.3.6.1.3 The Cancer Genome Atlas Research Network subtype nomenclature . . . 45
1.3.6.1.4 Genetic classification . . . 45
1.3.6.1.5 EGFR, KRAS, and TP53 mutational status based classification . . . 45
1.3.6.2 Other molecular classifications . . . 46
1.3.6.2.1 DNA methylation classification . . . 46
1.3.6.2.2 Proteomic classification . . . 47
1.3.6.2.3 Multiomic classification . . . 47
1.3.6.3 Squamous cell carcinoma . . . 48
1.3.6.4 Transcriptomic classification . . . 48
1.3.6.4.1 Primitive, classical, secretory, and basal subtype . . . 48
1.3.6.4.2 Intra-tumour heterogeneity of transcriptomic subtypes . . . 49
1.3.6.5 Genetic classification . . . 49
1.3.6.5.1 The Cancer Genome Atlas Research Network study 49
1.3.6.6 Other molecular classifications . . . 50
1.3.6.6.1 DNA methylation classification . . . 50
1.3.6.6.2 Metabolomic classification . . . 50
1.3.6.6.3 Proteomic classification . . . 50
1.3.6.6.4 Multiomic classification . . . 51
1.3.6.7 SCLC . . . 51
1.3.6.7.1 Genetic classification . . . 51
1.3.6.7.2 Transcriptomic classification . . . 51
1.3.6.8 LCNEC . . . 53
1.3.6.8.1 Genetic classification . . . 53
1.3.6.8.2 Transcriptomic classification . . . 53
1.3.7 Cellular atlases . . . 54

1.3.7.1 Healthy lung . . . 55
1.3.7.2 Non-small-cell lung carcinoma . . . 59
1.3.7.3 Small-cell lung carcinoma . . . 60
1.3.8 Cell states and plasticity in lung tumours . . . 60

2 Objectives **63**
2.1 Characterisation of intra-tumour heterogeneity . . . 63
2.2 Methodological objectives . . . 63
2.2.1 Marker identification . . . 63
2.2.2 Spatial pattern classification . . . 64
2.2.3 CNV inference . . . 64
2.3 Lung cancer cell communities . . . 64

3 Results **65**
3.1 SCherlock algorithm - robust cell type markers identification . . . 65
3.1.1 Introduction . . . 65
3.1.2 Methods . . . 66
3.1.2.1 Algorithm overview . . . 66
3.1.2.2 Theoretical distribution parameter estimation . . . 66
3.1.2.2.1 Notation . . . 66
3.1.2.2.2 Binomial distribution fitting . . . 67
3.1.2.3 Theoretical distribution scoring . . . 67
3.1.2.3.1 Multiplication factor introduction . . . 69
3.1.2.3.2 Cross-cell type correction . . . 69
3.1.2.3.3 Aggregation, ordering and filtering . . . 70
3.1.2.4 Empirical distribution scoring . . . 70
3.1.2.5 Metamarker construction . . . 70
3.1.3 Results . . . 71
3.1.3.1 Specificity of markers . . . 71
3.1.3.2 Contextuality of markers . . . 75
3.1.3.3 Detection of markers with low levels of expression . . . 77
3.1.4 Discussion . . . 79

3.1.5 Supplementary results . . . 79
3.2 Spattern algorithm - spatial pattern classification . . . 86
3.2.1 Introduction . . . 86
3.2.2 Methods . . . 86
3.2.2.1 Spatial data generator . . . 86
3.2.2.1.1 Uniform . . . 87
3.2.2.1.2 Gradient . . . 87
3.2.2.1.3 Structure . . . 89
3.2.2.1.4 Barrier . . . 89
3.2.2.1.5 Examples of generated data . . . 89
3.2.2.2 Spatial pattern classifier . . . 92
3.2.2.2.1 Feature type 1: Number of large neighbours . . . 92
3.2.2.2.2 Feature type 2: Expression-based random walks . . . 92
3.2.2.2.3 Decision tree classifier . . . 93
3.2.2.3 Results . . . 94
3.2.2.3.1 Classifier 1 ($N = 1000$ transcripts) . . . 94
3.2.2.3.2 Classifier 2 ($N = 100$ transcripts) . . . 95
3.2.3 Discussion . . . 96
3.2.4 Supplementary results . . . 97
3.3 FastCNV algorithm - transcriptome-based CNV inference . . . 102
3.3.1 Introduction . . . 102
3.3.2 Methods . . . 102
3.3.2.1 Reference assignment . . . 102
3.3.2.2 Amplification step . . . 102
3.3.2.3 CNV calling . . . 103
3.3.3 Results . . . 103
3.3.4 Discussion . . . 107
3.4 Unified lung cancer cell atlas . . . 108
3.4.1 Introduction . . . 108
3.4.2 Methods . . . 108
3.4.2.1 Atlas composition . . . 108
3.4.2.1.1 SCLC data . . . 108

3.4.2.1.2 LUAD and LUSC data . . . 109
3.4.2.1.3 Neuroendocrine NSCLC data . . . 110
3.4.2.1.4 Dataset split . . . 111
3.4.2.2 Atlas analysis . . . 111
3.4.2.2.1 Data integration . . . 111
3.4.2.2.2 CNV analysis . . . 111
3.4.2.2.3 Marker analysis . . . 111
3.4.2.2.4 Reannotation . . . 112
3.4.3 Results . . . 112
3.4.3.1 The mesenchymal subset . . . 115
3.4.3.2 The myeloid subset . . . 117
3.4.3.3 The lymphoid subset . . . 118
3.4.3.4 The epithelial subset . . . 120
3.4.3.5 Marker analysis . . . 122
3.4.4 Discussion . . . 126
3.4.5 Supplementary results . . . 127

4 Discussion **139**
4.1 Limitations . . . 140
4.2 Conclusion . . . 141
4.3 Perspectives . . . 142

Bibliography **143**

List of abbreviations

CAF Cancer-associated fibroblast

CNV Copy number variation

DC Dendritic cell

ITH Intra-tumour heterogeneity

LCNEC Large-cell neuroendocrine carcinoma (of the lung)

LUAD Lung adenocarcinoma

LUSC Lung squamous cell carcinoma

MDSC Myeloid-derived suppressor cell

MHC Major histocompatibility complex

NGS Next-generation sequencing

NK Natural killer

NSCLC Non-small-cell lung carcinoma

OS Overall survival

PFS Progression-free survival

ROI Region of interest

SCLC Small-cell lung carcinoma

SMG Submucosal gland

TAM Tumour-associated macrophage

TLS Tertiary lymphoid structure

TME Tumour microenvironment

UMAP Uniform manifold approximation and projection

1 Introduction

The first chapter is an introduction to cancer and its defining properties in general, and lung cancer more specifically. I briefly overview its histological classification, risk factors and standard of care, as well as molecular heterogeneity and cellular composition. On the more technical side, I introduce the field of transcriptomics, its recent advances and its relationship to cancer research.

1.1 Cancer

Cancer is a complex group of diseases, characterized by abnormal cell proliferation and the potential of the growing mass to invade neighbouring tissue or spread to other body parts. It is the second leading cause of death worldwide, responsible for nearly 10 million deaths annually[1]. Though individual cancer types and cases can have very different mechanisms, origins, characteristics, prognoses and treatment options, they are joined by certain common aspects and cellular properties.

1.1.1 Hallmarks of cancer

At the core of cancer lies gradual dysregulation of growth. As human cells move from normal to malignant tumoural states, it is necessary for them to acquire a specific set of behaviours in order to overcome the mechanisms of the cell set in place to keep growth in check:

- Sustaining proliferative signalling

  Growth promoting factors in healthy tissue are tightly regulated. Cancer cells thrive by deregulating these signals, ensuring continuous activation of pathways. They may produce the necessary growth factors themselves, elicit them from the stromal cells, or circumvent them completely and become ligand independent[2].

- Evading growth suppressors

  There are many tumour suppressor genes and mechanisms capable of slowing down or stopping unexpected cell growth. Cancers are able to progress by genetic or epigenetic inactivation (loss) of these suppressors[3]. The famous tumour suppressor *TP53* is the most mutated gene across all human cancers[4].

- Resisting cell death

  Cancers employ different strategies to avoid apoptosis that would usually be triggered by the stress and imbalances of cancer cell activity. Loss of *TP53* and its DNA-damage sensor function is the most common one, but there is also a variety of other aberrations/dysregulations of the pathways involved[2].

- Enabling replicative immortality

  Most normal cells are only able to divide a limited number of times. Eventually, when telomeres, the protective structures at the end of chromosomes, become short enough, they either enter a non-proliferative state of senescence, or face cell death. Cancer cells are able to overcome these limits by maintaining telomere length and replicate indefinitely[2].

- Inducing or accessing vasculature

  Like normal tissue, tumours require supply of nutrients and oxygen, and exit routes for metabolites and carbon dioxide. The needs of rapidly growing tissue are even greater, so vasculature formation or access is crucial. Typically, angiogenesis is induced, and the resulting vessel distorted and abnormally formed.[2]

- Activating invasion and metastasis

  Cancer cells leave their original sites, spread through the bloodstream and form metastases in secondary tissue. Epithelial–mesenchymal transition is a key step for initiation of this process - the cells lose their initial epithelial context and polarity, and become migrant[5].

- Reprogramming cellular metabolism

  Sustained and elevated growth requires increased metabolism rates, as well as metabolites nurturing the desired environment niche, and cancer cells undergo specific reprogramming to this purpose. In contrast to normal cells, for example, they tend to shift towards aerobic glycolysis for energy generation (i.e. the Warburg effect)[6].

- Avoiding immune destruction

  Cancer cells resort to different evasion mechanisms to disrupt normal immune function. They, for example, avoid destruction by cytotoxic $CD8^+$ T cells by lowering MHC-I expression, but maintain some level of it to avoid destruction by NK cells.

These behaviours are collectively known as the hallmarks of cancer[7]. As our understanding of cancer deepens, hallmarks continue to emerge and become further established in our conception of it.

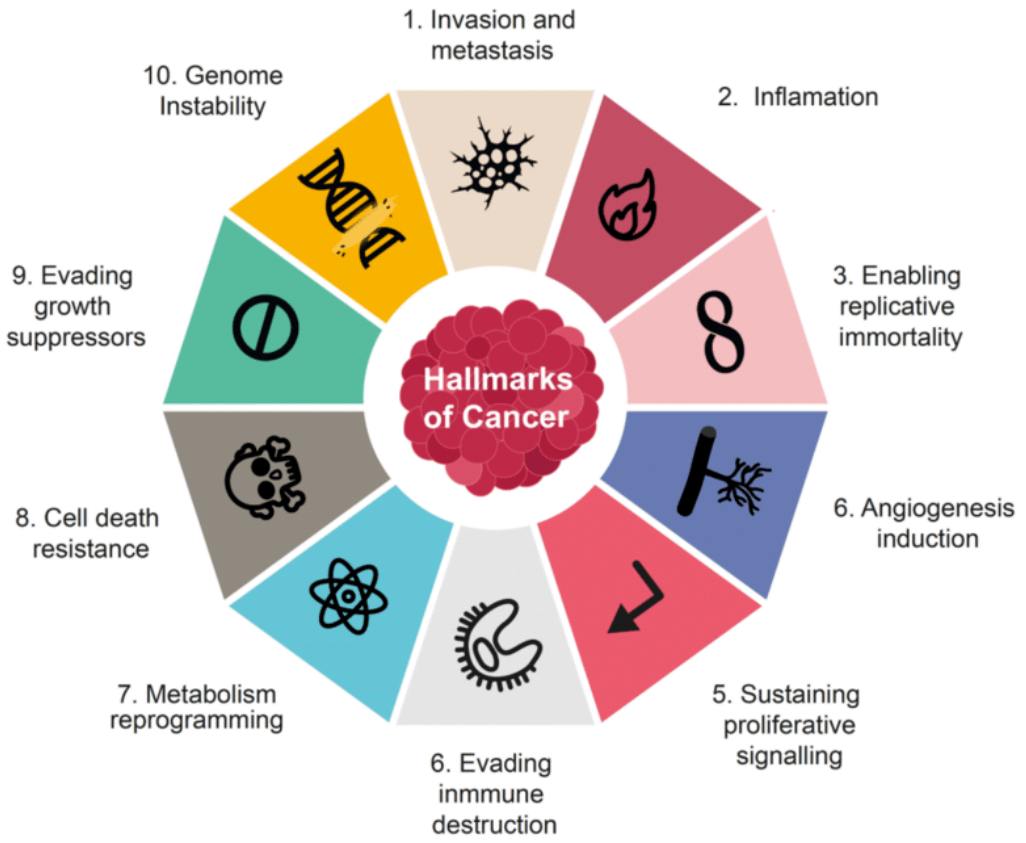

Figure 1: The established hallmarks of cancer. Source: Prieto García (2022)[8] adapted from Hanahan and Weinberg (2021)[2].

They are brought about by the so-called enabling characteristics: genome instability and mutation are necessary for the acquisition of several hallmark properties; while tumour-promoting inflammation delivers bioactive molecules to the tumour microenvironment, supplying the growth factors, survival factors, and other molecules key for execution of hallmark programs.

1.1.2 Physical traits

Further physical properties of cancer contribute to its growth and development[9].

Elevated solid stress, the force contained in the solid and elastic elements of cells and extracellular matrix, is caused by increased tissue volume (due to cell proliferation and infiltration), normal tissue displacement, matrix component swelling and cell contractions. Such stress can compress and collapse vessels, contribute to hypoxia, and prevent delivery of treatments.

Interstitial fluid pressure is caused by any abnormalities in vessel structure, including as a consequence of solid stress. Such fluid stress pressure drives interstitial flow in the tumour margin, putting extravascular cells under stress, as well as driving the interstitial fluid from the tumour into the surrounding tissue, promoting invasion by transporting growth factors and cancer cells.

Tissue stiffness has been used both as a diagnostic and as a prognostic factor. In multiple cancer types, malignant tumours have been shown to be stiffer than benign ones[10–12]. It is primarily caused by increased deposition and linking of extracellular matrix, and promotes tumour progression and invasion.

Disturbed matrix architecture and cell geometry can alter cell-matrix and cell-cell associations and initiate morphogenesis. An example of an invasive morphogenic switch of cell geometry is the epithelial-to-mesenchymal transition. Cell geometry can also influence the shape of the nucleus and indirectly control gene expression.

1.1.3 Tumour microenvironment

Tumours do not exist in isolation from external circumstances. They are embedded into their respective environments, which shape them, and which they shape in return. The tumour microenvironment (TME) is a complex system, composed of cancer cells, the surrounding stromal tissue, and the extracellular matrix, in which they are embedded. In the following subsections, I briefly review the cellular makeup of the TME and their respective functions in the context of cancer, summarized in Figure 2.

1.1.3.1 Stromal cell types

Vasculature plays an important role in delivering key components and oxygen to the growing cell mass, and is itself constantly developing to cover the ever-growing needs, with the process of angiogenesis being very typical of cancer sites. Nonetheless, a hypoxic environment can frequently be observed. Vessels are lined by endothelial cells (ECs), while blood microvessels are encompassed by pericytes.

Cancer-associated fibroblasts (CAF) are one of the central components of the TME. They are versatile and plastic, and aid tumour progression through extensive interaction with both cancer and other cells. They have a perpetually high capacity for extracellular matrix remodelling[15]. Other cells contributing to tumor stroma include cancer-associated adipocytes mesenchymal stromal cells, both involved in tumour promotion.

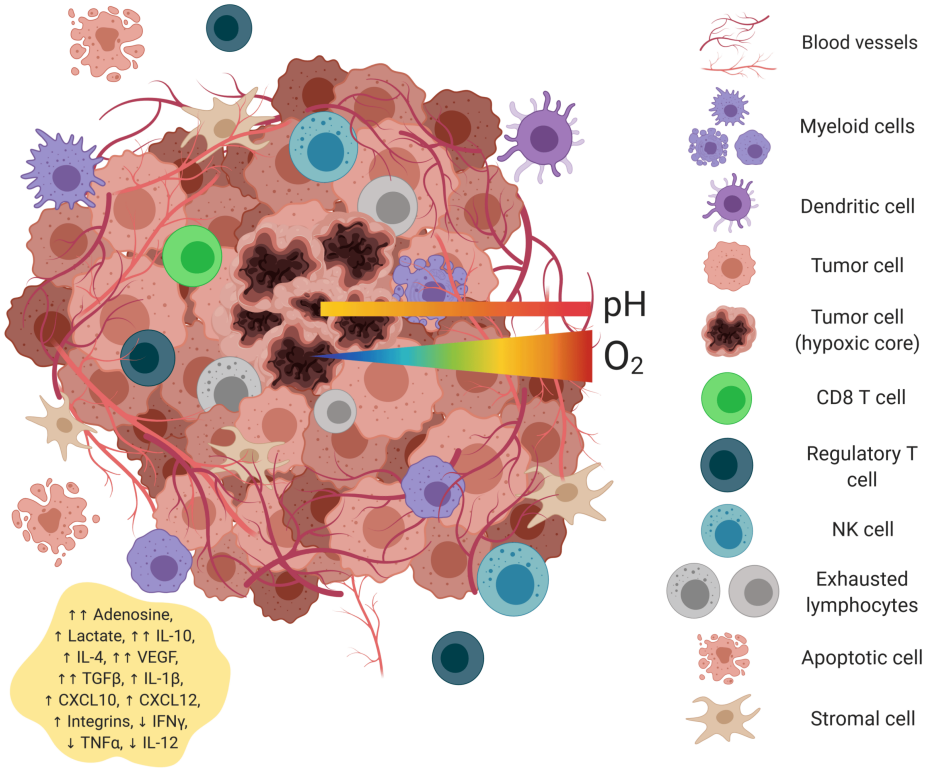

Figure 2: A simplified schematic of the tumour microenvironment and its components. Source: Piñeiro Fernández et al. (2019).[13]

1.1.3.2 Immune cell types

A variety of immune cells of adaptive and innate immunity is typically present in the tumour. They can either promote or suppress cancer growth. The latter role mainly falls on the adaptive immune system. A high concentration of tumour infiltrating lymphocytes (including T cells, B cells and NK cells) is generally associated with a good prognosis[16]. A schematic representation of anti and pro-tumoural microenvironments is presented in Figure 3.

$CD8^+$ T cells recognise cancer cells by binding to the antigens presented by MHC class I via TCR and destroy them employing their cytotoxic function. As a consequence of the chronic antigen stimulation in tumours, they often end up in a dysfunction exhaustion state[17].

$CD4^+$ helper T cells influence the immune cells around them. On the one hand, the T_h1 subtype assists the anti-tumoural cytotoxic activity of $CD8^+$ T cells and B cells, as well as targets cancer cells directly by IFNγ and TNF-α production. T_h2, however, secretes anti-inflammation mediators. TGF-β, produced by tumour cells, myeloid-derived suppressor cells (MDSCs) and tumour

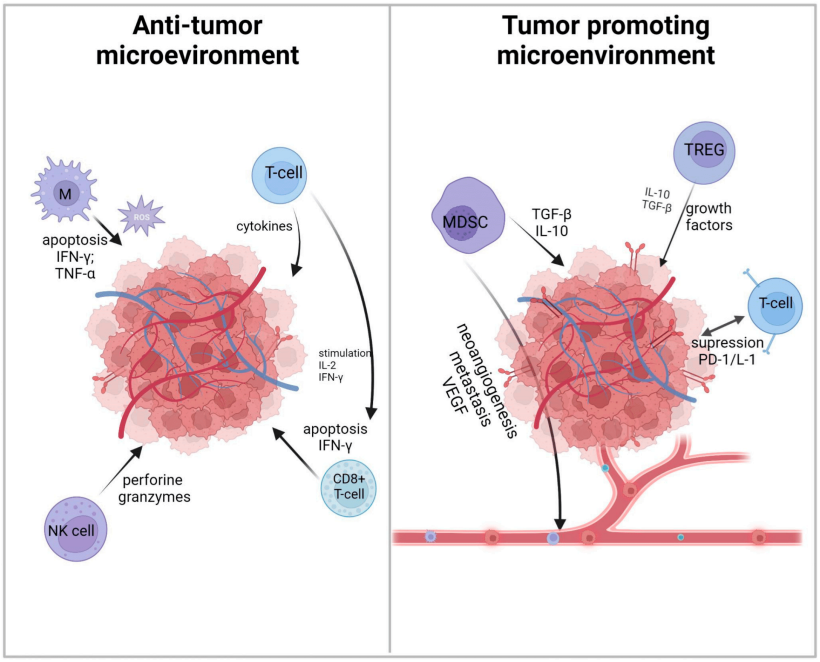

Figure 3: A simplified schematic of the tumour microenvironment immune cell component tumour suppressing and tumour promoting activity. Source: Chmiel et al. (2023).[14]

associated macrophages (TAMs), leads to conversion of $CD4^+$ T cells into T_{reg}s[18].

T_{reg} cells are highly immunosuppressive. They are necessary to maintain self-tolerance to healthy tissue, but in cancer they repress the anti-tumoural immune response and thus play a tumour promoting role[19].

B-cells mediate humoral immune response via antibody cytotoxicity and complement activation. They can infiltrate the tumour as individual cells, or reside in tertiary lymphoid structures (TLSs). Both the presence of B cells and TLS structures are considered a good prognostic factor in most cancers, including lung carcinoma[20,21].

Natural killer (NK) cells are cytotoxic lymphocytes of the innate immune system, able to kill cancer cells that lack MHC class I expression. Their infiltration is generally associated with a favourable outcome[22].

Tumour-associated macrophages (TAMs) are diverse in their function. The M1 population favours immune stimulation and is associated with good prognosis, while the M2 subtype downregulates adaptive immune response and promotes angiogenesis and is associated with worse prognosis[23].

Myeloid-derived suppressor cells (MDSCs) are a diverse population of potent suppressor cells, able to suppress lymphocytic anti-tumour response. They are considered tumour promoting[24].

Dendritic cells (DCs) play an antigen presenting role and induce T-cell mediated response to the antigens released into the microenvironment via MHC-I and MHC-II. As part of the innate system, they produce cytokines and growth factors in response to danger signals. The levels of DCs appear decreased in cancer patients[25]

1.1.4 Tumour heterogeneity

Tumour heterogeneity can refer to inter-tumour heterogeneity, the differences between tumours from distinct patients, or to intra-tumour heterogeneity, the differences that can be observed between various compartments (or cells) of the tumour of one individual.

The development path, accumulation of damage, and exact behaviour of each cancer instance is virtually unique. Any two tumours differ in their genetic, epigenetic and microenvironmental properties - even primary and metastatic tumours within the same patient. However, numerous molecular and cellular modifications can converge to similar phenotypes (e.g. the inactivation of a tumour suppressor gene can arise from an array of molecular events), which allow to define prototypical paths to cancer development and progression. Thus, the study of tumour heterogeneity does not seek to accumulate a collection of unique rarities, but rather to identify oncogenic scenarios that can be used to classify as many tumours as possible.

In affluent countries, the clinical management of individual cases can depend on these differences, as specific therapies are only effective in patients with vulnerabilities they are attempting to exploit, or with conditions conductive to the response they are attempting to elicit[26].

1.1.4.1 Intra-tumour heterogeneity (ITH)

Heterogeneity is not limited to distinct tumour cases. Individual tumours themselves are far from being uniform lumps of mass - they display internal heterogeneity at both genetic (e.g. point mutations, chromosomal structural aberrations) non-genetic (e.g. DNA methylation, chromatin modifications) level, along with their varied and dynamic mircoenvironments[27]. Sources of heterogeneity are summarized in Figure 4. In the following subsections, I address genetic subclonality and plasticity in further detail, as these pertain to results obtained through the course of this thesis.

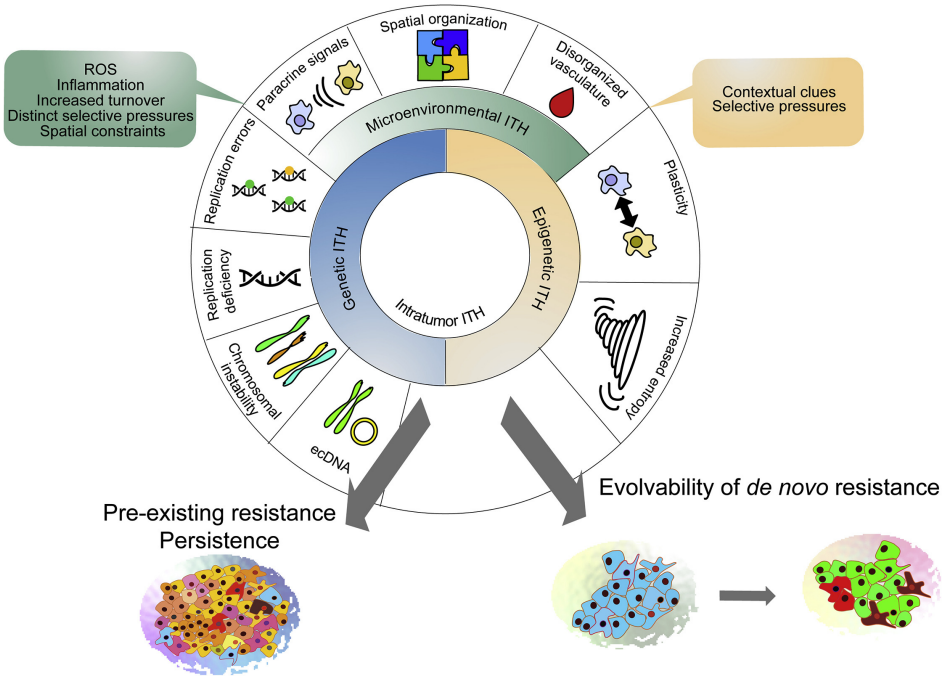

Figure 4: Sources of intra-tumour heterogeneity. Source: Marusyk et al. (2020).[27]

1.1.4.1.1 Genetic subclonality

Tumour cell populations are subject to clonal evolution, through the process of tumour formation as well as progression and metastasis[28]. As mutations are introduced into the population pool, some of them (so-called *driver* mutations) present a selective advantage — the mutated cell is fitter, better suited to the environment, survives longer, multiplies faster, ... As a consequence, the clones of that cell expand in the population[29].

This process continues past the commitment point of a tumour (its most recent common ancestor - MRCA) as mutations accumulate, genomic instability increases and further aberrations are acquired. These are no longer present in the entire tumour cell population, only subpopulations, and are thus classed as subclonal[30].

Such genetic heterogeneity makes the tumour more adaptable and resistant to harsh or changing environmental conditions, including treatments. While a portion or even the majority of the cancer cells might be eradicated, an unaffected subpopulation survives and pushes tumour progression into a new direction, resistant to the previous treatment[31,32].

Such a process also assists tumour spread through facilitating adaptation to the altered conditions after metastasising to a new site.

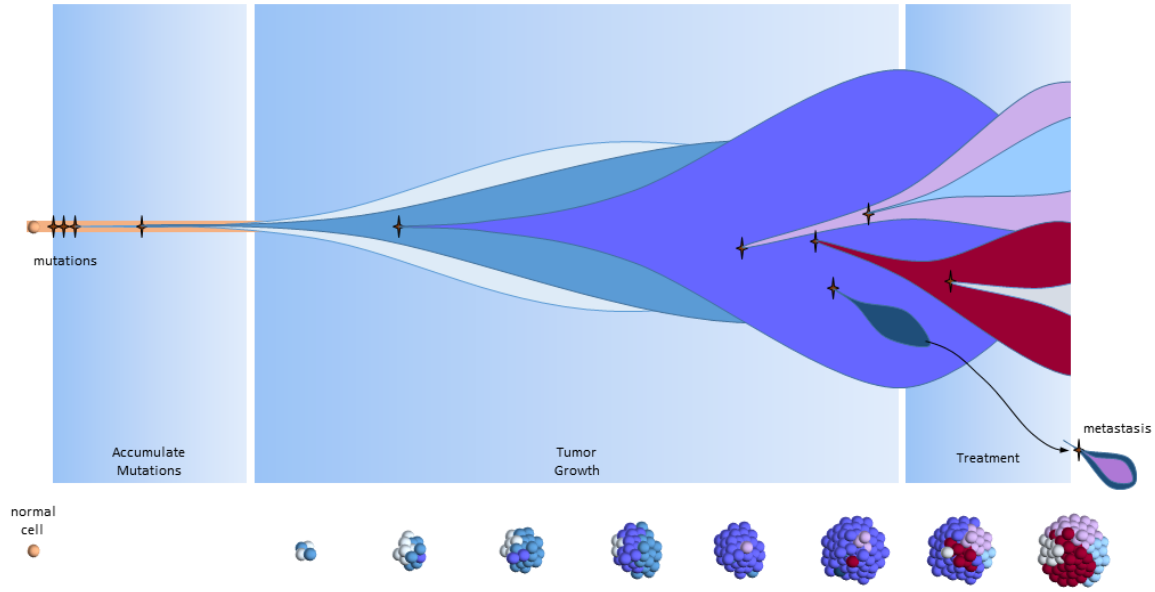

Figure 5: Intra-tumour heterogeneity of a tumour through progression under the influence of clonal evolution. A normal cell accumulates mutations and passes the commitment point of a tumour. Mutations acquired after that lead to subclonal populations. Some of them might be particularly fit to resist treatment or metastasise to secondary sites. Source: Dessert (2015).[33]

1.1.4.1.2 Plasticity

Cellular plasticity is the ability of cells to be reprogrammed and take on a new identity or phenotype in response to internal or external factors. Though plasticity is often associated with stem cells, it is not exclusive to them. Differentiated cells can undergo dedifferentiation and return to a pluripotent state; or transition onto a different differentiated state via transdifferentiation[34].

Dedifferentiation, as well as blockage of regular progression into differentiation, facilitate tumour development and progression by locking cells into proliferative progenitor states, rather than allowing them to pass and stay in typically non-proliferative fully differentiated states.

Transdifferentiation allows cells to switch into a different lineage as a response to changes in gene expression levels and the TME. It is an additional source of intra-tumour heterogeneity and increases the ability of the tumour to thrive and adapt to circumstances. Plasticity can play a role in the initial tumour genesis, malignant progression, and/or treatment response.

A prominent example of cellular plasticity, existing in physiological conditions and reused in can-

cer, is epithelial–mesenchymal transition (EMT), a process through which epithelial cells lose their polarity and adhesive properties, and become migratory and invasive mesenchymal stem cells. This process enables tumour invasion and initiates the process of metastasis[2].

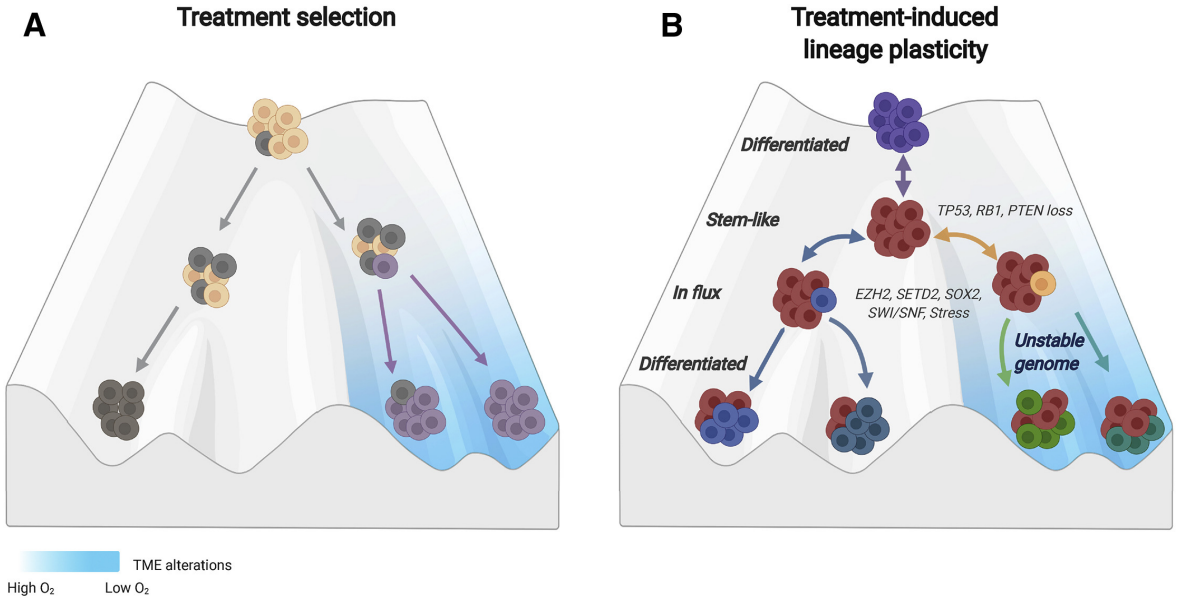

Figure 6: Treatment resistance in the cases of pre-existing subclonal ITH and treatment induced lineage plasticity. Source: Rubin et al. (2020).[35]

1.2 Omics in cancer research

Omic technologies enable quantification of various types of molecules at a large-scale. The first generation of these technologies, which appeared at the end of the 1990s, was based on DNA chips (microarrays). Nowadays, they are mostly based on sequencing techniques and are typically used to measure RNA expression (transcriptomics), DNA sequences (genomics), DNA methylation (methylomics), DNA bound by certain proteins (epigenomics). They also include techniques to measure protein expression (proteomics) or metabolites (metabolomics). Over the last two decades, several major cancer research programs (The Cancer Genome Atlas (TCGA) in the USA, The International Cancer Genome Consortium (ICGC) at the international level, Cartes d'Identité des Tumeurs (CIT) in France) have used omics extensively to study cancer heterogeneity and to identify oncogenic mechanisms. These efforts produced multi-omic profiles of thousands of tumours, corresponding to dozens of cancer types, and allowed major progresses in the characterisation of inter-tumour heterogeneity, through the identification of molecular subtypes,

and a better understanding of the underlying biology. This work has been initially performed on bulk tissues. Then, thanks to the Human Cell Atlas international research program (launched in 2016), single-cell omic technologies (in particular single-cell RNA sequencing), have become very popular in cancer research, to identify cell types and characterise intra-tumour heterogeneity.

1.2.1 Genomics

Genomics has come a long way since the days of the Human Genome Project. Today, either whole genome sequencing (WGS), or the more accessible whole exome sequencing (WES), limited to the coding regions of the genome, can be performed with the use of high-throughput DNA sequencing technologies[36]. The tumour sample is evaluated in comparison to paired normal tissue (typically blood) in order to separate somatic events from germline ones, and single-nucleotide variants (SNVs), fusions, rearrangements, and copy number alterations (CNAs) are identified. (The CNAs are estimated, rather than determined directly.) Single-cell DNA sequencing is currently limited to a few targets, as measuring the WGS or WES in single cells is yet very challenging with existing techniques.

1.2.2 Proteomics

The study of proteins at a large scale is marked by a much larger molecular diversity. Obtaining proteomics data in its true sense (at the scale of the proteome) is more difficult than genomic data, as direct sequencing is not possible, and recognition of individual products is thus necessary. As a result, the current methods only measure a fraction of the proteome; they are based either on microarray technology, or mass spectrometry, with the latter being predominant, despite limited sensitivity[37,38]. Recent technological advancements (e.g. Olink[39] - enabling screenings of panels of up to several thousands of preselected proteins at once) promise further possibilites. In addition to these bulk proteome analysis techniques, the expression of membrane proteins (in limited numbers) can be measured on a single-cell scale using techniques such as fluorescence activated sorting (FACS), but given the limited number of molecules analysed, this case is not considered a proteome.

1.2.3 Epigenomics

Epigenomics encompasses the study of a variety of different chemical modifications on molecules that play a role in genetic processes, such as DNA or histones. These reversible epigenetic modi-

fications often serve a regulatory purpose and are a common mechanism of plasticity and differentiation[40]. The most common approaches to obtain epigenomic data include the study of DNA methylation (bisulfite sequencing), chromatin accessibility (ATAC-seq[41]), and histone modification (ChIP-seq[42]). Single-cell measurements are not common, except for the increasingly popular single-cell ATAC-seq technique.

1.2.4 Transcriptomics

At the core of cell functionality is what we know as the central dogma of molecular biology: genes are transcribed from DNA to RNA and then translated into proteins. Other than any newly acquired mutations, gene sequences stored in the DNA are static. As previously stated, proteins are much more difficult to comprehensively quantify on a large scale across the board of all the different molecules, as they cannot simply be sequenced[38].

It is thus the more accessible measurement of the transcriptome, the totality of RNA transcripts present in the sample at a given point in time, that has gained popularity as a snapshot of the current state of the tissue being studied. It is an imperfect measurement - transcription levels, for example, only give a rough estimate of what ends up actually translated into protein[43]. Ideally, we would be free to combine information from multiple omic modalities (e.g. transcriptomics, proteomics, epigenomics) at every step.

Yet even transcriptomics alone has had a lot of success in cancer research[44]. Transcriptomic profiles appear to be more closely related to the observed patient phenotypes than their respective genomic and epigenomic profiles[45–47]. The comprehensive quantification of gene expression levels offers opportunities for comparison between contrasting conditions, patient differences, tissue types (e.g. healthy/tumoural), developmental stages, and clinical outcomes. Genes, identified as differentially expressed in such comparisons, give insights into the potential functions, pathways and mechanisms of action involved. Further analysis might involve association of results with knowledge contained in existing resources such as ontologies, constructing new signatures of interest that could be useful in other context, or identifying potential targets to be studied via additional modalities, performed in a more limited scope (e.g. immunofluorescence staining of specific proteins)[48].

1.2.4.1 Bulk transcriptomics

The classic approach to transcriptomics has been to analyse and quantify the gene transcripts for the entire tissue sample together. The two prevalent techniques are microarrays[49], based on hy-

bridisation of transcripts to complementary probes and fluorescence output measurement, and RNA-seq[50], based on reverse transcription of RNA into cDNA and high-throughput sequencing. The latter is more in use nowadays[51].

By definition, the gene expression measured from a bulk tissue sample is a mixture of gene expression for all cells present in the sample. The aggregation into a singular output inherent to this method effectively obfuscates the many sources of transcriptomic variety coming into play, even more so when it comes to studying cancerous tissue.

The expression signals can already differ at the level of tumour cell subpopulations as a result of intra-tumour heterogeneity (ITH). Then there are other factors: the tumour microenvironment component, the variable level of tissue infiltration, the proportion of normal and cancerous tissue in the sample. None of these considerations can be addressed directly at this resolution, though it can be attempted statistically by separating signal sources by the use of various proposed deconvolution methods[52–54].

Regardless of its limitations, bulk transcriptomics is a technology that has been around for longer, and remains much more accessible (technically and financially speaking) than its higher resolution counterparts. Consequently, many more community resources have already been developed, and a huge number of datasets have been made publicly available. There have been some notable consortium efforts such as The Cancer Genome Atlas (TCGA)[55], which is a key source of reference bulk RNA-seq datasets, covering no less than 33 of the most common cancer types. Numerous other, organ specific efforts have yielded datasets with focus on a specific type of cancer (e.g. the TRACERx consortium lung cancer study[56]).

Furthermore, the extensive spread of this technology in basic and clinical setting means that there is all kind of individual study data deposited publicly or semi-publicly in databases such as Gene Expression Omnibus[57] or The European Genome-phenome Archive[58], or elsewhere as downloadable files. Samples of tissue in different states of development, health and disease can often be identified, and even data on rare types of cancer not covered by the TCGA such as the pulmonary large cell neuroendocrine carcinoma (LCNEC) can be found[59]. The availability of such readily available data remains an invaluable research resource.

1.2.4.2 Single-cell transcriptomics

Thanks to technological advances, the study of transcriptomics at a single-cell resolution has become possible and has been gaining momentum over the past years. The general steps involved

are cell separation and isolation, lysis of the membrane, amplification of transcripts via reverse transcription into cDNA, and quantification of expression levels. Currently, the most used approach for single-cell sequencing is based on microfluidics technologies. In this case, each cell is co-encapsulated in an oil droplet with a bead charged with oligonucleotides and a solution with all the needed reagents; the droplet then serves as a reaction chamber, as illustrated in Figure 7. Sometimes, the cells from the cell suspension used as input for this microfluidics device are first selected using fluorescence activated sorting (FACS) and selections rules based on a panel of proteins.[60,61]. The reaction acting in each droplet allows obtaining a cDNA library, with individual tags for each cell, prior to sequencing.

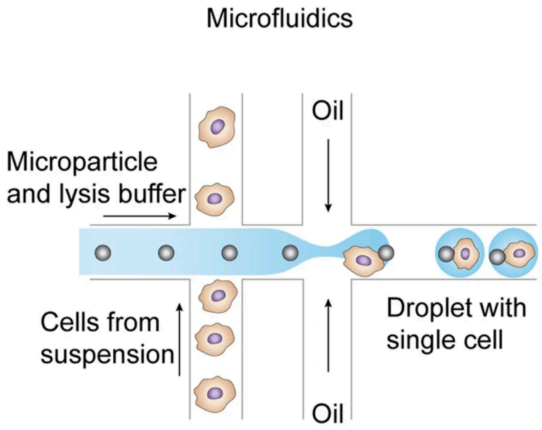

Figure 7: Cell separation by microfluidics. Source: Hwang et al. (2024).[62]

The increased resolution enables the study and cataloguing of cell populations in healthy and tumour tissue and its surrounding microenvironment. Cell types and states can be identified and investigated, with new ones frequently reported.

The Human Cell Atlas[63] is a large consortium effort to chart the many different cells that constitute all the organs of the human body, but many other efforts of various sizes and research targets have been going on too. Numerous and diverse methodological advancements and applications of interests driven by this body of work have also been emerging in the scientific community.

In cancer research, this approach is key for the study of intra-tumour heterogeneity and the various subtypes and their mixtures; as well as plasticity, which can be observed through the presence of intermediate cell types.

Transcriptomic profiles of specific cell populations also provide us with further insights into their biology. This type of data provides an opportunity to identify new molecular markers, in particular transcriptomic signatures that are cell type specific.

Many other types of downstream analyses can be performed of such a dataset. Though in some cases, methods designed for bulk data can easily be reused or adapted, single-cell resolution offers many new opportunities (e.g. using copy number estimation to separate cancerous cells from the normal ones, infer their lineage order, ...) which the many custom-developed algorithms make good use of.

There are some complications too. There are fewer reads per sample in single-cell transcriptomics than bulk, and the resulting matrices are sparse - even genes which are expressed will not always be detected. This causes difficulties in downstream analyses. Cell type identification, for example, remains non-trivial even when there are known and reliable markers available. In most cases, decisions and assignments are made in a cluster-based manner to compensate for this.

Though there are many single-cell datasets already published, they are much more limited in scope and quantity (in number of patients) than what is available for bulk transcriptomics. It is also a method that is less accessible in general, and unlikely to be routinely used in the clinic in the foreseeable future.

As we start diving into the exploration of cell communities, we inevitably reach the question of interaction of cell-cell interactions. Several algorithms were proposed to mine such cell-cell interactions from single-cell RNA sequencing (scRNA-seq) data[64–67]. With it, by separating the expression signals to individual cells, we are able to tell if certain genes are expressed within the same cells, or different populations. But physical location is also of concern - are ligands and receptors being expressed in proximity of each other? The proposed algorithms are all are limited by the lack of spatial resolution in single-cell data. This leads us to the latest iteration of transcriptomic technologies.

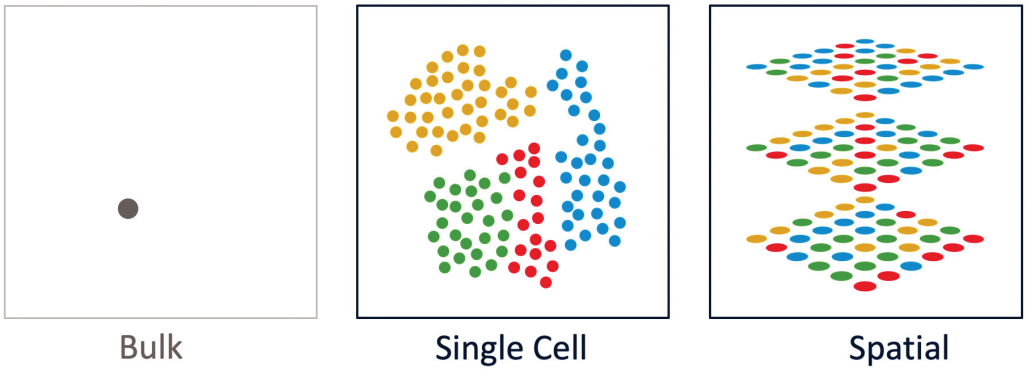

Figure 8: Transcriptomics at different levels of resolution: bulk, single-cell and spatial. Source: 10X Genomics. (2019).[68]

1.2.4.3 Spatial transcriptomics

In recent years, there has been a big growth in development and use of spatial transcriptomics. The ability to investigate the spatial organisation of gene expression patterns and correlate them to other types of imaging has brought a whole new dimension to the field of transcriptomics, cancer research included.

There are several different types of spatial approaches in principle.

The first relies on physical separation/isolation of regions of interest (ROI) (tomo-seq[69], GeoMx[70]), followed by a bulk (or single-cell) transcriptomic protocol performed on each of the isolated regions. This could range between a specific section that is of interest to the study, to a systematic microdissection of the entire tissue sample.

The *in situ* hybridization approach[71] (seqFISH[72], MERFISH[73]) is based on targeting transcripts with labelled probes and capturing their location via fluorescence microscopy. Multiple rounds of hybridization and different colours are used to encode for each gene with a unique barcode. *In situ* sequencing[74] (ISS[72], STARmap[75]) is also image based, but profiles 1–2 nucleotides at a time.

The NGS barcode approach (Visium[76], slide-seq[77]) involves binding barcodes encoding spatial location to the transcripts themselves before performing RNA-seq, and assigning each to its corresponding location during processing. The resolution of this approach is typically less than single-cell.

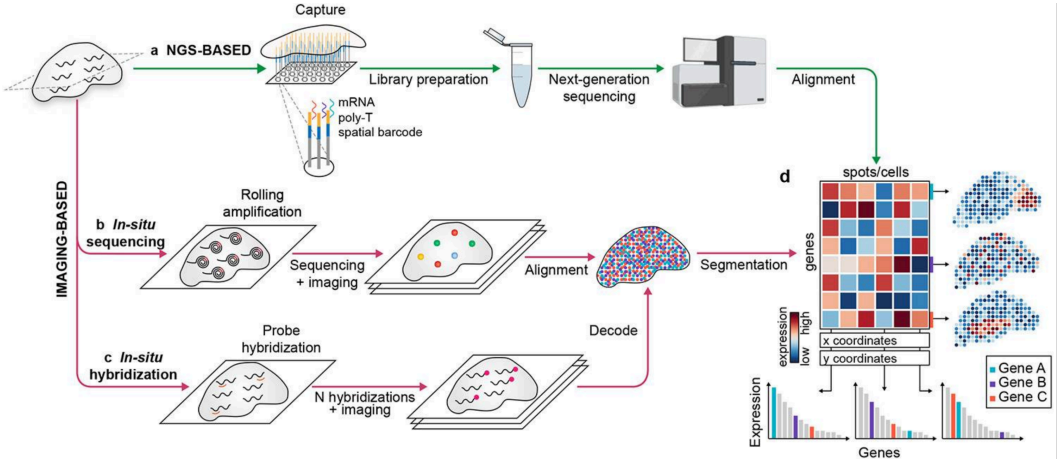

Figure 9: Fully spatial approaches to transcriptomics: NGS barcode, *in situ* sequencing, and *in situ* hybridization. Source: Rao et al. (2021).[78]

Nonetheless, it is this approach that has achieved widespread popularity (thanks to 10X Genomics

Visium platform[76] and its relative accessibility), and the majority of spatial data produced currently is of this kind.

The original version of the Visium method, which is the most relevant to this work, produces two levels of information — first imaging of the given tissue section is performed, typically with either hematoxylin-eosin (H&E) staining or immunofluorescence (IF).

The second level is the spatial gene expression data, which is obtained using spatial barcoding. Predefined areas on the slide bear location tags, which are then bound to mRNA molecules. Bulk RNA-seq is performed and location information corresponding to the image layer is extracted. The slides contain square shaped capture areas, 6.5 mm in each dimension, so only small sections of tissue can be analysed. This is more appropriate for tissue structure specific analyses than general analyses attempting to produce results representing the entire tissue.

The capture areas contain 4,992 spots arranged in a regular hexagonal grid pattern. These are the predefined locations for which we are to obtain gene expression data, but typically not all of them will be covered by tissue. The spots are 55 µm in diameter, 100 µm apart and typically overlap with 1 to 10 cells.

Later iterations of Visium offer larger capture areas or simultaneous analysis of RNA and protein expression[79]. Recently released, Visium HD offers single cell-scale resolution and continuous tissue coverage[80].

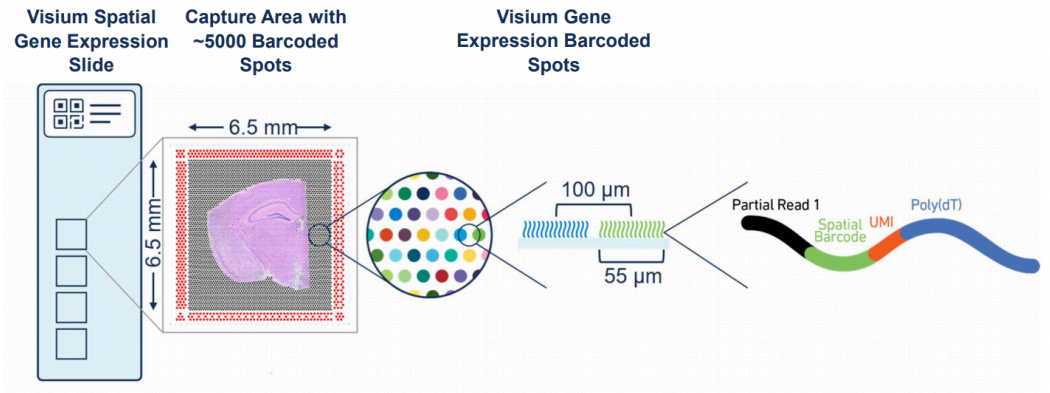

Figure 10: The Visium Spatial Gene Expression Slide v2. Source: 10x Genomics (2023).[81]

The added spatial dimension of this technology adds completely different opportunities for investigation of tissue organisation and cell communities. We are able to connect the gene expression pattens to the pathological images, and associate them with the structures found there; and thus make a direct link between transcriptional profiles and physical phenotypes. We are also able to

use this spatial axis to compare genes with each other, and look for transcripts with similar spatial distributions.

Some of the difficulties of handling this kind of data are similar to single-cell transcriptomics (e.g. sparsity, noisiness, high dimensionality). Additionally, the field of spatial transcriptomics is still very young, so there is little consensus when it comes to many methodological aspects of downstream analysis.

1.2.4.4 Clinical use

Though genetic testing is the more common and traditional approach, transcriptomics is gradually making its way into the clinic.

Transcriptomic data is already an important component of basic research, as well as pre-clinical and clinical studies, where the data collection is largely limited to the bulk version. Though the technology is advancing, it does not seem realistic to expect single-cell or spatial methods to be part of routine clinical decision-making in the foreseeable future. This is a question of accessibility, but also other practical limitations. The samples necessary for spatial transcriptomic, for example, must be larger compared to bulk measurements, and require resection rather than biopsy alone.

Still, there are ways to utilise the resources and research results from these more advanced iterations of technology and transfer them to the clinic indirectly.

The increased resolution offers precision, and allows us to isolate biological signals in a less noisy manner. Returning to bulk data and picking out these signals with the help of the resulting more precise signatures is one way to close the circle.

A lot of work has also been done on deconvolution of bulk RNA-seq based on a relevant (if not specifically paired) single-cell dataset.

Transcriptomics has the potential to improve early detection capabilities, as well as influence clinical management of individual patients and deliver on the promises of personalised medicine. Though the overwhelming majority of screening and response predicting biomarkers are still more traditional, novel transcriptomic-based solutions like the recently FDA-approved ColoSense[82] are arriving.

1.3 Lung cancer

Lung cancer is the second most common and the most deadly cancer type worldwide (2.2 million estimated new cases and 1.8 million deaths in 2020). In Europe, it is the 3rd most common and the most deadly cancer type (480,000 new cases and 380,000 deaths in 2020). It is the 6th most common cause of death worldwide[1].

1.3.1 The human lungs

In their main respiratory role, they are the site of gas exchange, but they also act as a barrier filtering out foreign particles. Additionally, they are involved in specific instances of precursor activation, degradation of chemical mediators, and neuroendocrine factor secretion.

The human lungs are located in the thoracic cavity of the chest. The right lung has three lobes and is larger than the left one with only two, as the latter shares space with the heart and has a cardiac notch to accommodate it. The lung is functionally divided into the conducting zone, which leads outside air to the site of respiration, and the respiratory zone, where gas exchange takes place.

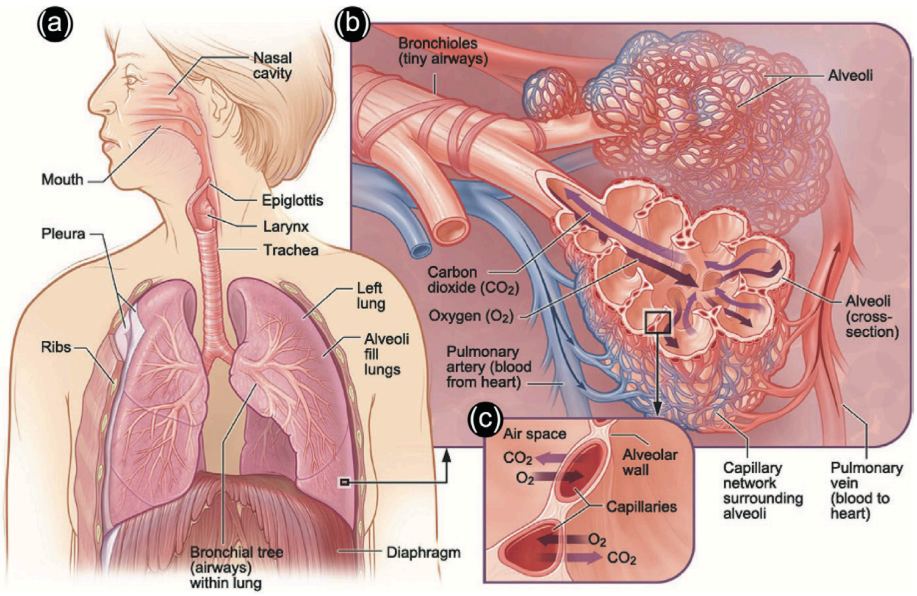

Figure 11: The anatomy of human respiratory system and lungs. Source: Reicherter. (2015)[83]

The conduction system begins outside the lungs with the nasal cavities, nasopharynx, larynx and trachea. In the lung, it consists of the branched hierarchical structure of bronchi and ends with

the terminal bronchioles. In addition to being moved through the airways, the air is also warmed, moisturised and filtered for any particles in this section.

The respiratory system starts with the respiratory bronchioles and includes alveolar ducts, alveolar sacs and alveoli. These are the structural and functional units of the respiratory system and provide a very large surface area for gas exchange between the lungs and the blood.

Alveoli are composed of type I and type II cells (pneumocytes). The former are flat epithelial cells with fragile membranes allowing gas exchange and cover the vast majority of the surface, though the populations are fairly balanced. The latter produce a surfactant protein and lipid mixture that reduces surface tension and enables lung expansion. Alveolar macrophages are also present as a last line of barrier against inhaled particles. They are also involved in surfactant turnover. The alveolar wall contains a network of capillaries (composed of capillary I and II cells), which sometimes comes in direct contact with the epithelium in what is called the blood-gas barrier; elsewhere there is supporting tissue composed of fibroblasts. Small passages are present between certain alveoli to compensate for any obstruction.

The respiratory epithelium is composed of different cell populations attached to a layer of basal cells. Ciliated cells are most represented - their role is to remove clear inhaled particles trapped in mucus, which is secreted by the goblet cells. Other cell types are rarer - tuft cells have a chemosensory role, ionocytes maintain airway fluid balance, club (secretory) cells produce secretoglobins (replacing goblet cells deeper in the respiratory tree). A small population of pulmonary neuroendocrine cells is spread throughout the airway epithelium. These act as an airway sensor and respond to signals by secreting neuropeptides and neurotransmitters. Other glandular cells are present in submucosal glands (SMGs): mucous and serous cells.

Airways are supported by cartilage up until and including bronchi; bronchioles being too small for that are supported by smooth muscle; chondrocytes and airway smooth muscle cells are involved respectively. The lungs have a double blood supply - bronchial arteries carry oxygenated blood to the airways through the bronchial circulation, while the pulmonary circulation brings deoxygenated blood from the heart and returns it back oxygenated for body-wide distribution. Vascular smooth muscle cells control pulmonary vessels, which are lined by venous and arterial endothelial cells. Lymphatic vessels are also present and lined by lymphatic endothelial cells. Lung pleura encases and protects the lung. It is composed of mesothelial cells.

Finally, the lung has an immune cell population consisting of both innate and adaptive immune response cells. Given the relative exposure of the lung as a tissue, these play a key protective role.

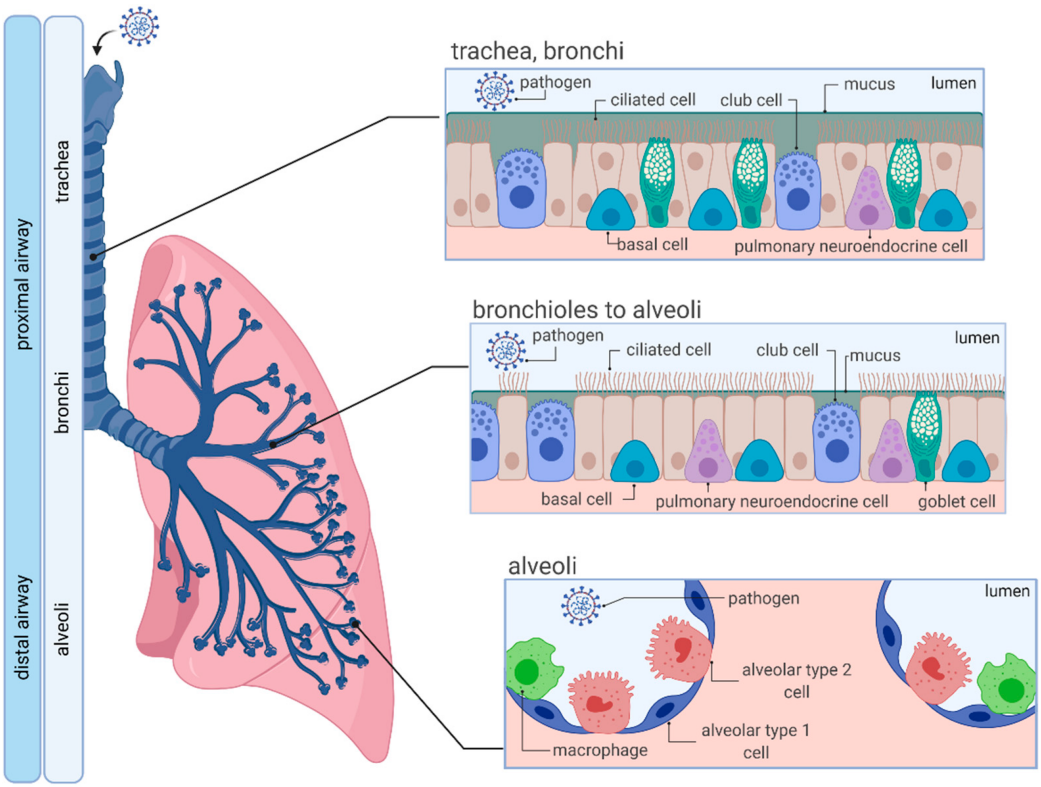

Figure 12: A simplified schematic of the human lung and the cellular composition of its epithelium. Source: Heinen et al. (2021)[84].

Lung-specific genes include those encoding lung surfactant proteins (SFTP family), secretoglobins and serpin peptidase inhibitors. They are enriched for the process of respiratory gas exchange, as well as immune response, regulation of cell proliferation and lung tissue development.

1.3.2 Histological classification of lung cancers

The 2021 WHO classification of lung cancers[85] distinguishes between different histological groups, the main ones being the adenocarcinomas (LUAD, 40-50%), the squamous cell carcinomas (LUSC, 20-30%), and the neuroendocrine (NE) carcinomas, which include the small-cell lung carcinomas (SCLC, 10-15%)[86] and the large cell neuroendocrine lung carcinomas (LCNEC, ∼3%)[87]. The adenocarcinomas are a diverse histology with five distinct growth patterns. Examples of main histologies are collected in Figure 13.

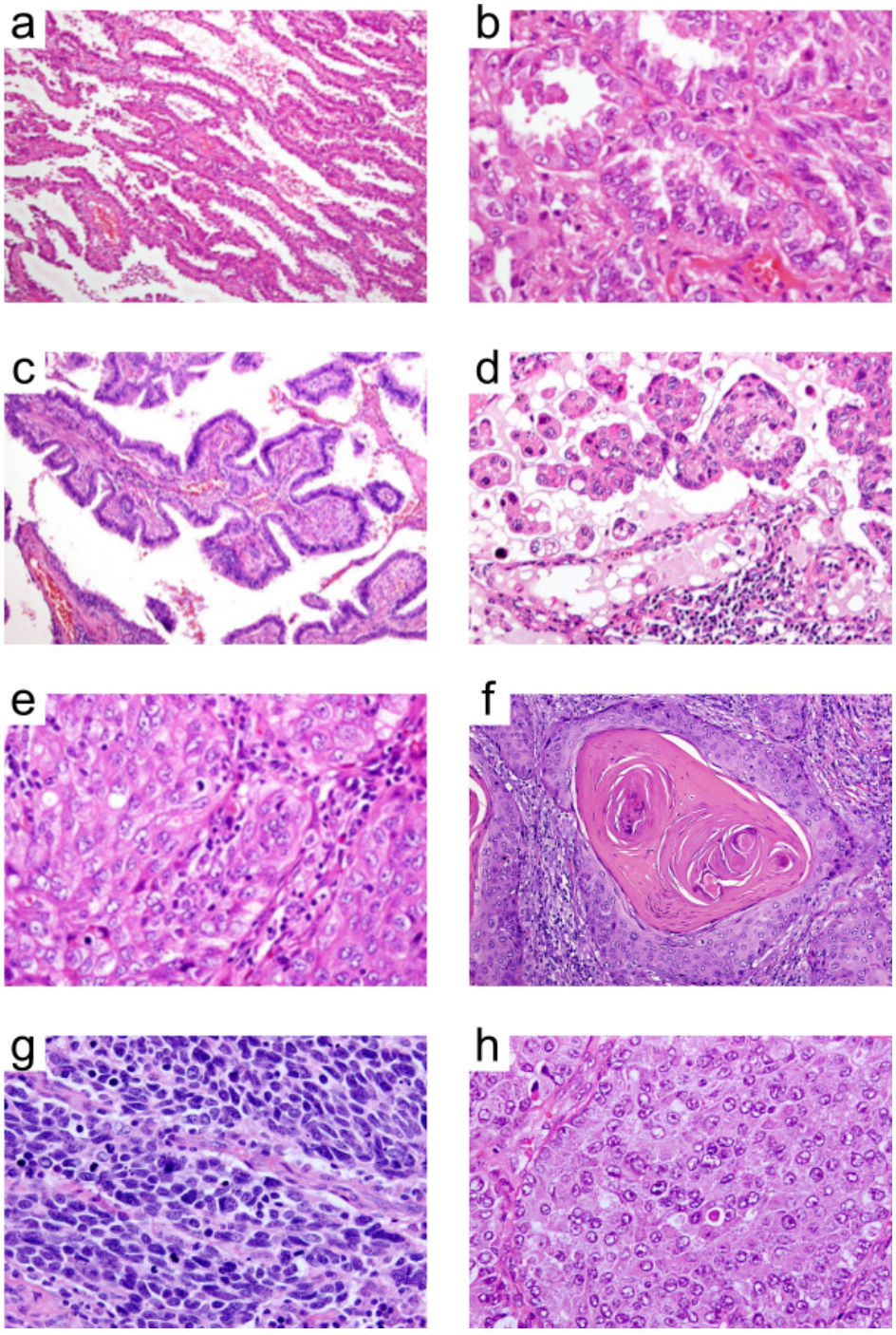

Figure 13: H&E stained tissue of the main histologies of lung cancer: a-e) major adenocarcinoma growth patterns (respectively: lepidic, acinar, papillary, micropapillary and solid), f) squamous cell carcinoma, g) small-cell carcinoma, h) large cell neuroendocrine carcinoma. Source: adapted from Travis (2011)[88].

In the literature, the common practice is to distinguish small-cell lung carcinomas from all other lung cancer types, called non-small cell lung carcinomas (NSCLCs), as SCLCs correspond to the most aggressive and resistant form of lung cancer (5-year survival rate of around 6% in SCLC and 24% in NSCLC).

1.3.2.1 Mixed histology

Around 5% of primary lung cancers present a mixed histology (also known as combined histology), where distinct histological components are present within the same tumour[89]. They typically have higher malignancy and poorer prognosis[90]. All major histological groups can appear in mixed histologies with each other: adenosquamous carcinoma, LUAD/LUSC combined to a neuroendocrine histology (SCLC/LCNEC), or the two neuroendocrine histologies together (SCLC + LCNEC).

A study of transcriptomic and whole-exome sequencing of nine microdissected tumours of 19 different histological samples has indicated that the histology of a tumour is associated with its transcriptomic profile, rather than genetic, though the impact of the latter cannot be completely excluded, particularly not in all cases[46]. The transformation of NSCLC into SCLC, for example, has been closely associated with TP53 and RB1 co-mutation[91]. Other histological transformations, such as between LUAD and LUSC, have also been reported in certain conditions[92,93].

1.3.3 Environmental factors

Smoking is the leading cause and risk factor of lung cancer by far, with advanced age being the other major one[94]. Both duration (years) of smoking and cigarettes smoked per day increase the risk of cancer development[95]. Second-hand exposure to smoke also affects non-smokers[96]. Other important environmental factors include air pollution[97], radon exposure[98] and working with hazardous chemicals[99].

1.3.4 Genetic factors

Lung cancer is heavily driven by environmental factors. Our understanding of lung cancer in never smokers (LCINS), estimated to represent about 10-25% patients, is lacking in general[100]. Inherited lung cancer risk is particularly poorly understood. The prevalence of adenocarcinomas with onco-

genic EGFR mutations among LCINS has been associated with several factors, including germline mutations[101].

1.3.5 Standard of care

1.3.5.1 NSCLC treatment and immunotherapy

With NSCLC, particularly adenocarcinoma, identification of driver mutations is key for treatment selection, as they can be treated with targeted therapy for specific mutations. Patients with an *EGFR* mutation (frequent in females, non-smokers, adenocarcinomas) respond well to receptor tyrosine kinase inhibitors (RTKIs) and poorly to checkpoint inhibitors (perhaps due to low mutational heterogeneity; typically excluded from later clinical studies). Other actionable aberrations include *ALK* fusion (also poor response to checkpoint inhibitors), *BRAF*V600e mutation, *MET* exon 14 skipping mutation, *RET* fusion and *ROS1* fusion. In squamous carcinoma, loss of function (*PTEN*) and gain of function (*PIK3CA*) mutations appear to be more common[102].

For patients with stage I-III resectable disease, surgery with chemotherapy is still the standard. Nivolumab (PD-1 inhibitor) was first tested to treat advanced refractory squamous NSCLC, which lacked effective standard therapy. Clinical trials showed about 20% response and about 3 months improvement of median OS in both squamous and nonsquamous NSCLC compared to docetaxel. 5-year pooled OS was also notably higher (13.4% vs 2.6%), demonstrating a durable response and a significant clinical benefit in patients with advanced NSCLC[103]. This success was followed by other immune checkpoint inhibitor (ICI) clinical trials.

Pembrolizumab has demonstrated a 2-month median OS improvement on PD-L1 positive cases (>1%) and an almost 6-month improvement in the >50% subpopulation, where it also improved median PFS[104]. Atezolizumab (PD-L1 inhibitor) improved median OS by around 4 months regardless of PD-L1 status, though ORR was higher in the positive group[105].

Success in second line treatment has encouraged study of checkpoint inhibitors as first line of treatment for patients in advanced NSCLC. Pembrolizumab has a 44.8% response rate compared to 27.8% of traditional platinum-based chemotherapy in patients with >50% PD-L1, an increase in median PFS of 4 months, and double 5-year survival rate (32% vs 16%). It has also shown a smaller but still significant improvement in >1% PD-L1 patients[106]. Atezolizumab also provides a significant benefit to PD-L1 high cases[107].

Combination of pembrolizumab with chemotherapy is beneficial compared to chemotherapy

alone, regardless of PD-L1 status (increased median OS and PFS in both squamous and nonsquamous NSCLC)[108]. No direct comparison with pembrolizumab alone on the PD-L1 high group has been made yet, so the decision is made based on disease load and tolerance of chemotherapy. Combination of nivolumab and ipilimumab (PD-1 + CTLA-4) significantly improves survival compared to standard platinum chemotherapy, and particularly for PD-L1 positive cases, with a third of patients surviving 3 years regardless of PD-L1 status (ORR 45.3%)[109].

1.3.5.2 SCLC treatment and immunotherapy

Limited stage patients undergo radio- and chemotherapy. In 75-98% of cases, *TP53* mutation is identified. As there is no targeted therapy for SCLC, identification of driver mutations does not guide treatment.

No treatment options for extensive disease produce long-lasting response, with median survival 1-2 years from diagnosis. Prior to immunotherapy, the options considered were hospice supportive care, additional chemotherapy cycles or clinical trials. Chemotherapy with Topotecan demonstrated improved OS with a more commonly maintained quality of life compared to best supportive care. This was a standard drug for recurrent/metastatic SCLC for years[110].

PD-1 blockade is now a treatment option for previously treated patients with no PD-1/PD-L1 treatment history. Pembrolizumab has been FDA approved for heavily pretreated patients with ORR 19.3% and median OS 7.7 months (compared to 4.4 from other studies)[111]. Nivolumab has been FDA approved for recurrent/metastatic SCLC that has progressed after more than 2 lines of chemotherapy, with ORR 11.9% and 1-year OS 28.3%[112]. Combination of nivolumab and ipilimumab shows an improved ORR of 21.9%, but this is not reflected in OS or PFS, possibly due to frequent treatment discontinuation caused by lack of tolerability[113]. PD-L1 hasn't proven to be a biomarker for SCLC response.

PD-1 blockade has also been studied for use in combination with chemotherapy as first line of treatment. Atezolizumab added to carboplatin and etoposide resulted in a 2-month increase of median OS (51.7% vs 38.2% 1-year OS), despite similar levels of ORR (around 60%)[114]. Durvalumab's addition to etosposide and platinum chemotherapy has also prolonged median OS by 3 months[115]. Combination of pebrolizumab and etoposide has shown an increase in PFS (1-year rate of 13.6% vs 3.1% in standard treatment), with ORR 70.3% vs 61.8% and 2-year OS 22.9% vs 11.2%, though median OS did not significantly improve[116].

1.3.5.3 LCNEC treatment

Surgical removal is the standard approach for early stage disease, and perioperative chemotherapy seems to be favourable. Due to its rarity and limited evidence for any specific treatment approach, clinical management of this disease is not standardized or optimal. It can be treated either by SCLC-type (platinum-etoposide) or NSCLC-type (gemcitabine/taxane/pemextred with platinum) chemotherapy. In type I LCNEC (wild-type *RB1*), the latter shows superior overall survival, while the two approaches are equally successful in type II LCNEC (*RB1* loss) according to a retrospective study[117]. 41 interventional clinical trials including pulmonary LCNEC have been registered at clinical trials.gov to date, but only 6 of those focus on LCNEC specifically and no other types of cancer[118].

1.3.6 Molecular classifications

1.3.6.1 Adenocarcinoma

1.3.6.1.1 Transcriptomic classification

Several initial studies proposed gene expression based classification of adenocarcinomas. Despite being limited in scope, they were relatively consistent in proposing three distinct transcriptomic subtypes[119–121].

1.3.6.1.2 The three transcriptomic subtypes

The first study to evaluate the subtypes more systematically was published in 2006[122]. The subtypes were identified independently in three microarray data cohorts of 31, 72, and 128 LUAD patients. The subtypes were named after the histologies of cancer types their respective gene expression profiles correlated with: bronchioid, squamoid, and magnoid after bronchioalveolar carcinoma (outdated histology type), squamous cell carcinoma, and large-cell carcinoma.

The subtypes are linked to clinical outcomes: the bronchioid subtype has better survival in early-stage disease, but worse in late-stage. In advanced disease, it is the squamoid subtype that is linked with better survival.

The bronchioid subtype is more frequent in non-smokers and female patients and more likely to have an *EGFR* mutation. It is less likely to form a distant metastasis. It expresses markers of type

II pneumocyte differentiation and stains positively for mucin. The squamoid subtype is more frequent in smokers and male patients, and is more likely to carry a *KRAS* mutation. It tends to be less well differentiated. The magnoid subtype is characterised by a high level of inflammation.

1.3.6.1.3 The Cancer Genome Atlas Research Network subtype nomenclature

In 2014, the TCGA research network proposed a revised nomenclature for the transcriptional subtypes in an attempt to coordinate it with the histopathological, anatomic and mutational classifications of lung adenocarcinoma. It would become the new standard: the terminal respiratory unit (TRU) subtype instead of bronchioid, the proximal-inflammatory (PI) subtype instead of squamoid, and the proximal-proliferative (PP) subtype instead of magnoid[123].

1.3.6.1.4 Genetic classification

Two studies further established the three transcriptomic subtypes of adenocarcinoma and explored their association with genomic properties on published (504 LUAD tumours) and novel (116 and 230 LUAD tumours) cohorts[123,124].

The TRU subtype has the most cases of *EGFR* mutation and overexpression, and the cancer development seems to depend on it exclusively, while the rare instances in other subtypes typically co-occur with *TP53* mutation. Kinase fusion expressions are also specific to this subtype.

The PP subtype is characterised by *STK11*, *KEAP* and *KRAS* mutations, as well as the greatest chromosomal instability, most copy number alterations, DNA hypermethylation, and tumour mutation burden. DNA repair pathway is overexpressed in response to the extensive damage.

The PI subtype is characterised by *TP53* mutation, frequently in co-occurrence with *NF1* mutation.

1.3.6.1.5 EGFR, KRAS, and TP53 mutational status based classification

In 2020, a simplified genetic classification of adenocarcinomas was published, based on only the mutational status of genes that are routinely tested in the clinic[125]. The development cohort consisted of 283 cytology samples of lung adenocarcinoma. The validation cohort combined 63 cytology samples from the development cohort that included survival data with another 428 surgical biopsy samples.

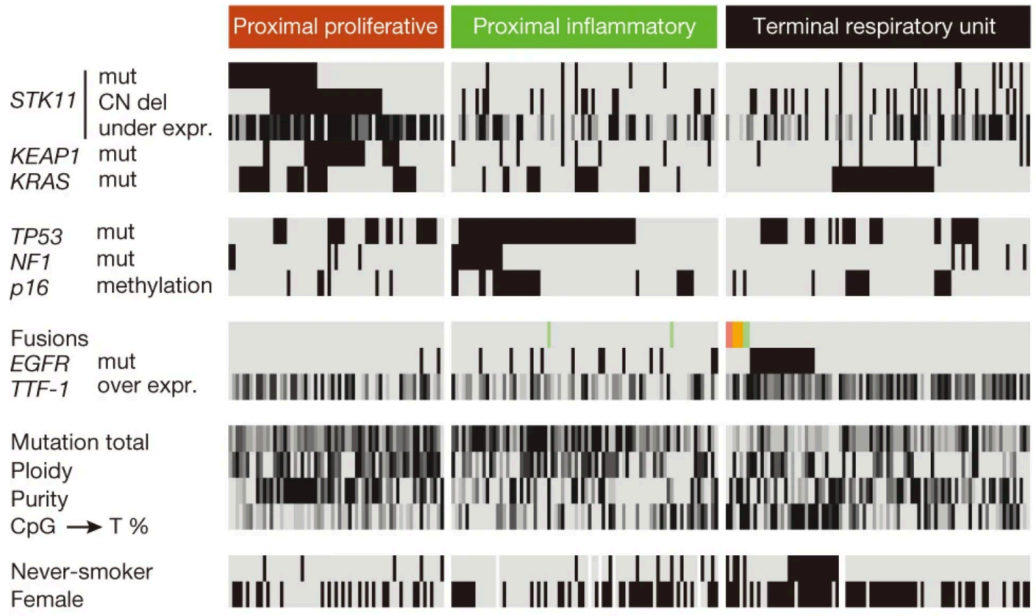

Figure 14: Molecular subtypes of lung adenocarcinoma. Source: The Cancer Genome Atlas Research Network. (2014)[123].

The proposed classes are simplified equivalents of the three molecular classes typically proposed in the literature: simplified terminal respiratory unit (sTRU) subtype - only *EGFR* mutated; simplified proximal-proliferative (sPP) subtype - only *KRAS* mutated; simplified proximal-inflammatory (sPI) subtype - only *TP53* mutated. Additionally, *KRAS/TP53* subtype, *EGFR/TP53* subtype, non-TRUPPPI subtype (none of the three genes mutated), and "no mutation" subtype (none of the genes detected on the panel used mutated) are included.

1.3.6.2 Other molecular classifications

1.3.6.2.1 DNA methylation classification

A CpG island methylator phenotype (CIMP) subtype classification was first proposed in an initial study on a training cohort of 41 and validation cohort of 87 LUAD patients[126]. It was further established in the TCGA network study of lung adenocarcinoma[123].

Aberrant CpG island (CGI) promoter DNA hypermethylation is associated with absence of transcription and occurs in cancers at a number of known tumour suppressor genes[127]. In CIMP, first described in colorectal adenocarcinoma[128], hundreds of CPIs become hypermethylated in a

coordinated manner during tumour progression.

Three classes were identified by both studies: a CIMP-high group, which is significantly altered, a CIMP-low group, which appears normal, and an in-between CIMP-intermediate group. The PI transcriptional subtype has the strongest association with CIMP-high and CIMP-intermediate, while the PP subtype is mostly CIMP-low.

Two additional studies on cohorts of 335 LUAD patients and 150 and 503 LUAD patients were published in 2020 and 2021 respectively[129,130]. Each of them proposes seven subtypes, linked to clinical outcomes.

1.3.6.2.2 Proteomic classification

A study published in 2019 profiled a cohort of 87 lung adenocarcinomas[131]. The previously described transcriptomic subtypes were also recovered based on protomic data (mass spectrometry and protein arrays) and further described via subtype-specific proteogenomic network models.

The PI subtype is characterised by interactions that drive IFN-γ signalling and inflammation, significantly overexpresses PD-L1, and also overexpresses CTLA4. With high immune infiltration and tumour mutational burden (TMB), it has all the properties to be the likeliest suitable subtype for immune checkpoint inhibitor treatment.

The TRU subtype is characterised by overactive EGFR signalling and downstream activity. It is the likeliest candidate for EGFR inhibitor response as well as for response to treatment targeting other members of the signalling cascades in question (e.g. PRKCE40, RPS6KA1).

The PP subtype is characterised by cell-cycle and glycolytic biological processes and an immune-cold environment. The majority of tumours do not possess targetable oncogene mutations, but there are potentially exploitable metabolic vulnerabilities (e.g. glutaminase inhibition) in certain cases.

1.3.6.2.3 Multiomic classification

Many classifications integrating multiple layers of omic data have been proposed over the years, starting with the TCGA research network study, each proposing a different number and division of subtypes[123,132,133].

A recently published study[134] defined a classification based on differentially expressed radiation-associated genes, immune-related genes, methylation data, and somatic mutation data on 446 LUAD samples from the TCGA cohort.

Three subtypes were identified: CS1, CS2, and CS3. CS3 shows a relatively low mutational burden and the best prognosis, as well as higher chromosomal stability. CS2 has the worst prognosis, and the highest frequency of *TP53* aberrations and the highest levels of immune cell infiltration. CS1 has the highest frequency of *KRAS* mutations and the lowest levels of immune cell infiltration. A bioinformatic chemosensitivity prediction indicated CS2 as the most sensitive to docetaxel and cisplatin, and CS1 as most sensitive to paclitaxel. CS2 also emerged as the likeliest subtype to benefit from PD-1 immunotherapy.

1.3.6.3 Squamous cell carcinoma

1.3.6.4 Transcriptomic classification

Initial attempts at LUSC classification described two subtypes each - one with a better and one with a worse prognosis[135–137].

1.3.6.4.1 Primitive, classical, secretory, and basal subtype

The first large and systematic classification study was published in 2010 by Wilkerson et al[138]. It defined what became the standard transcriptomic classification of lung squamous cell carcinoma on five published cohorts, totalling 382 LUSC patients, and evaluated it on a new validation cohort of 56 patients.

The unsupervised consensus clustering proposed four classes: primitive, classical, secretory, and basal. The subtypes show certain similarities to normal lung cell types. The basal subtype is similar to lung basal cells in cell adhesion and epidermal development functional expression, *S100A2* and *KRT5* basal cell markers, and overexpression of genes with basement membrane product localisation. The secretory subtype in turn resembles lung secretory cells, overexpressing surfactant and mucin. The primitive subtype is characterised by poorly differentiated cells, proliferation, and the worst survival outcome. The classical subtype displays all the typical characteristics of LUSC - typical of males, highest smoking prevalence, and overexpression of *TP63*. It is also the most common subtype (37%). A later study published in 2014[88] further showed that the primitive subtype was enriched in basaloïd squamous cell carcinoma, a rare histological type, with high aggressivness.

1.3.6.4.2 Intra-tumour heterogeneity of transcriptomic subtypes

A study published in 2018 evaluated the heterogeneity of transcriptional subtypes within lung carcinomas[139]. It included samples from 35 regions of 10 tumours, including two cases of LUSC with three samples each. The analysis points to the presence of ITH (basal and secretory subtype) in one case, while it is not conclusive in the other.

1.3.6.5 Genetic classification

1.3.6.5.1 The Cancer Genome Atlas Research Network study

The TCGA study on lung squamous cell carcinoma published in 2012 has reiterated the previously described transcriptional subtypes, and associated each of them to specific genetic characteristics based on a cohort of 178 patients[140]. The classical subtype is characterized by alterations in *KEAP1*, *NFE2L2* and *PTEN*, with notable hypermethylation and chromosomal instability. It is anticorrelated with amplification of *FGFR1* and *WHSC1L1*. The classical subtype is the most associated with the 3q26 amplicon, though it was present in all subtypes. The primitive expression subtype is associated with *RB1* and *PTEN* alterations, and the basal expression subtype with *NF1* alterations.

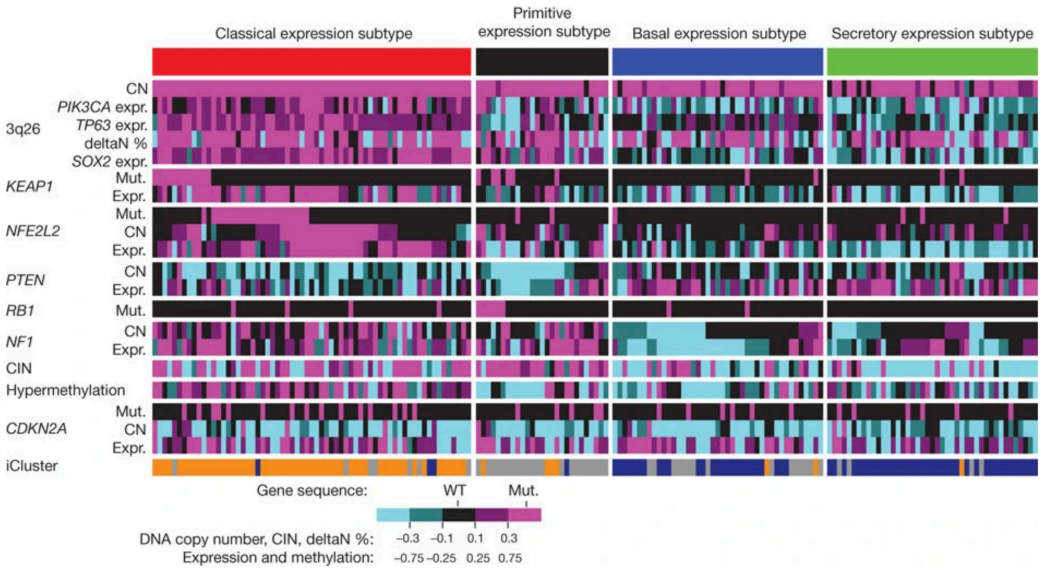

Figure 15: Molecular subtypes of lung squamous cell carcinoma. Source: The Cancer Genome Atlas Research Network. (2012)[140].

1.3.6.6 Other molecular classifications

1.3.6.6.1 DNA methylation classification

Two studies published in 2021 analysed methylation arrays of the TCGA cohort (N=504) in parallel[141,142]. Each of them proposed seven different subtypes. The subtypes show some clinical differences, but they are not described in detail, and do not appear consistent between the two studies.

1.3.6.6.2 Metabolomic classification

A recent study identified four tumour subtypes and four stroma subtypes of lung squamous cell carcinoma based on metabolite patterns of 330 LUSC patients[143]. Though preference showed in some cases, the tumour and stromal subtypes were not consistently paired in the same way overall.

T1 (PD-L1$^-$CD3$^-$CD8$^-$) is characterized by low immune cell infiltration, low DNA damage and poor prognosis, while T2 (PD-L1$^+$CD3$^+$CD8$^+$) has high infiltration and good prognosis. T3 also has good survival, T4 is also PD-L1$^+$.

Stromal subtypes are associated with immunological features and prognosis. S1 is linked with chemotherapy benefit, while S2 (PD-L1$^-$CD3$^-$CD8$^-$) is linked to chemotherapy resistance (evaluated on an independent neoadjuvant chemotherapy-treated cohort of 40 patients). S4 (PD-L1$^+$CD3$^+$CD8$^+$) is associated with good prognosis.

1.3.6.6.3 Proteomic classification

A study published in 2019 identified three proteomic subtypes of lung squamous cell carcinoma in a cohort of 108 LUSC patients[144]. Inflamed subtype (40% of cases) is enriched with neutrophils, B-cells, and monocytes, as well as tertiary lymphoid structures (TLS). It also expresses more PD-1. Redox subtype (47% of cases) has reduced immune infiltration. It is enriched for oxidation-reduction and glutathione pathways *NFE2L2* and *KEAP1* mutations, and amplifications in 3q2. The rest of tumours belongs to the mixed subtype, which has reduced expression of both immunological and redox proteins. It might be associated with Wnt signaling function.

1.3.6.6.4 Multiomic classification

A multiomic classification based on CNA, RNA, protein, phosphoprotein, and acetyl protein data from LUSC 108 patients was published in 2021[145]. The five clusters identified were named based on similarity to the previously established molecular subtypes. The basal-inclusive (B-I) subtype is linked to basaloid histology and upregulated metabolic, immune, and estrogen receptor signalling. The epithelial to mesenchymal transition-enriched (EMT-E) subtype is linked with upregulation of EMT, angiogenesis and myogenesis, as well as fibroblast infiltration. Together with B-I, it seems to roughly correspond to a partition of the basal transcriptomic subtype. The classical transcriptomic subtype was completely recovered here. It was linked with alterations of *KEAP1*, *CUL3*, and *NFE2L2* and amplification of *SOX2* and *TP63*. The inflamed-secretory (I-S) subtype corresponds to the secretory subtype and is characterised by a strong upregulation of immune pathways. The proliferative-primitive (P-P) subtype corresponds to the primitive subtype and is marked by an upregulation of proliferative pathways and downregulation of immune signalling.

Tumours with weaker associations to a subtype cluster, a possible indication of intra-tumour heterogeneity, had significantly worse survival.

1.3.6.7 SCLC

1.3.6.7.1 Genetic classification

For a long time, small-cell lung carcinoma was seen as a homogeneous disease. While they established the near universal double mutation of *TP53* and *RB1* that characterises SCLC, neither the smaller scale inital studies[146,147] nor the more in-depth genomic profiling follow-up on a cohort of 110 cases[148] proposed any kind of subtype classification. While the perception of a monolithic entity has since been broken, no classification at the genetic level has been established to date.

1.3.6.7.2 Transcriptomic classification

A study published in 2019 integrated and synthesised many previous attempts to classify small-cell lung carcinomas based on differential gene expression[148–156]. It was the first to propose the now standard division of SCLCs into four classes: SCLC-A, SCLC-N, SCLC-P, and SCLC-Y[157]. Each of the subtypes is characterised by elevated expression of its respective defining transcription factor.

SCLC-A, defined by *ASCL1* expression, and SCLC-N, defined by *NEUROD1* expression, represent the so-called neuroendocrine subtypes. SCLC-A is the most common class, accounting

for about 40-50% of cases, while SCLC-N is the second most common, accounting for about 30%[158,159]. SCLC-A is associated with the classic morphology and high NE marker expression. Further division has previously been proposed based on the level of *HES1* expression[156]. SCLC-A is associated with the variant morphology and a lower level of NE marker expression. Combination cases with a high level of both *ASCL1* and *NEUROD1* expression have also been reported. A study that included 63 samples of the two subtypes in question found diverse patterns of with IHC, including reciprocal expression, pointing to intra-tumour heterogeneity of subtypes[159]. Subtype switching has also been observed, with indication to a hierarchy where the higher NE subtype transitions to a lower NE one[152].

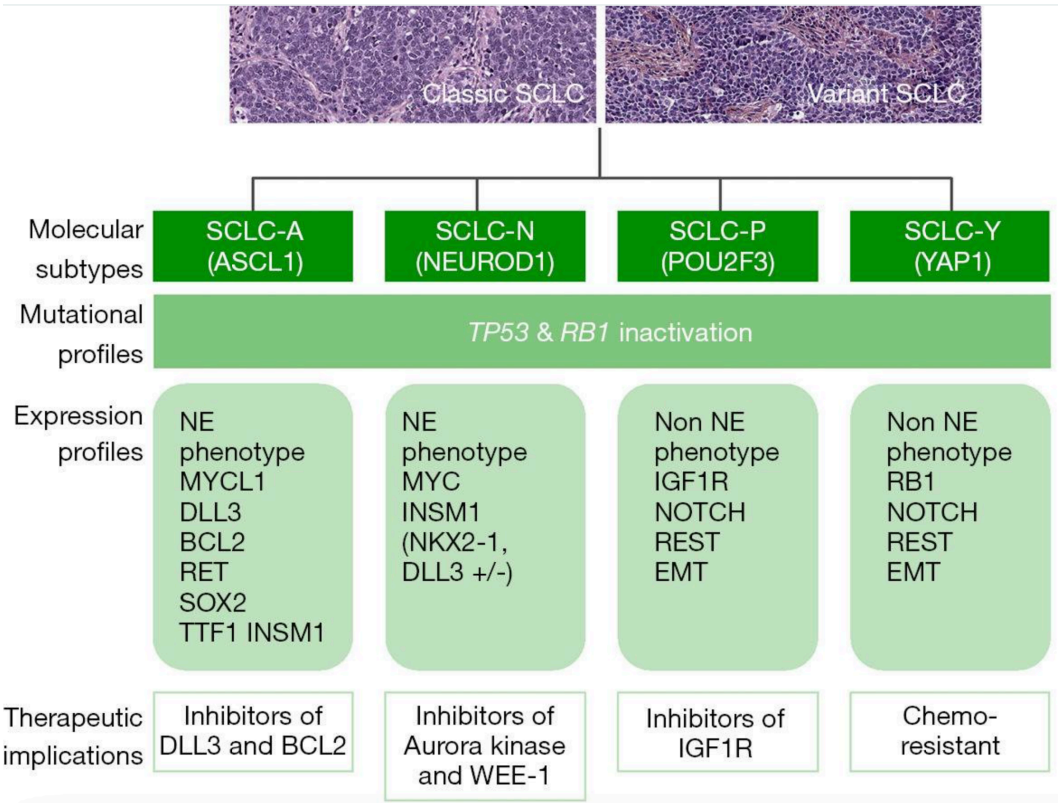

Figure 16: Molecular subtypes of small-cell lung carcinoma. Source: Lantuejoul et al. (2021)[160].

SCLC-P is defined by expression of *POU2F3*. Since it is typically expressed in tuft cells, it is hypothesised that this subtype might originate from such a healthy cell[155]. SCLC-P does not express traditional NE markers.

The proposed SCLC-Y is defined by *YAP1* expression, but has been elusive - its validation efforts have had mixed success[161,162]. The difficulties may or may not be due to its rarity, particularly as a standalone subtype. An alternative triple-negative fourth subtyped SCLC-I characterised by an

inflamed signature has been proposed[163].

1.3.6.8 LCNEC

1.3.6.8.1 Genetic classification

Several attempts to classify LCNEC based on mutational profiles have reached similar conclusions. Two studies, on a cohort of 45 paired tumour and normal surgically resected samples and a cohort of 78 samples (65 surgically resected and 13 biopies) respectively, both proposed two major subclasses[164,165].

A SCLC-like subtype is defined by alteration of both *TP53* and *RB1* (joint inactivation of these is a signature SCLC event). A NSCLC-like subtype is characterized by the absence of such co-alteration (mostly on account of *RB1*, with *TP53* still largely affected) and presence of typical NSCLC-like aberrations in genes such as *STK11*, *KRAS* and *KEAP1*. There is also some indication towards a minor third outlier carcinoid-like subclass with a low mutational burden and hallmark carcinoid *MEN1* alterations. It has been proposed that high-grade neuroendocrine carcinomas could develop from pre-existing carcinoids.

A study on a cohort of 54 of advanced stage disease has reported the PFS (and potentially OS, depending on classification details) of SCLC-like patients to be shorter despite a generally higher ORR to chemotherapy[166].

1.3.6.8.2 Transcriptomic classification

Similarly to the purely genetic classifications, a comprehensive molecular study based on a cohort of 75 LCNEC patients[167] found 2 distinct genomic groups of LCNEC, associated with specific transcriptional patterns.

NSCLC-like genetic LCNEC subtype (type I) has demonstrated a neuroendocrine expression profile (high in *ASCL1* and *DLL3*; low in *NOTCH* pathway), a profile typically associated with SCLC.

SCLC-like genetic LCNEC subtype (type II) has, on the other hand, shown a low neuroendocrine expression profile (low *ASCL1* and *DLL3*, high *NOTCH*) more typically associated with NSCLC. This study was performed before the four transcriptomic subtypes were proposed, but this seems to indicate that the SCLC-like LCNEC subtype bears resemblance to the non-NE subtypes, rather than to SCLC-A or SCLC-N. An upregulation in immune pathways has also been detected in this subtype, which might have implications for response to immune checkpoint inhibitor treatment.

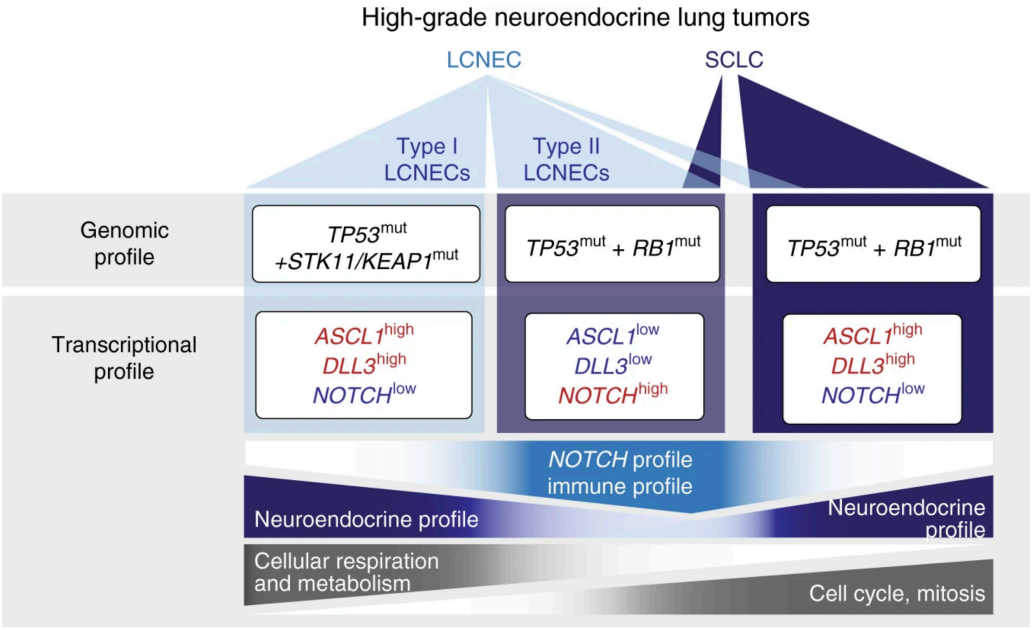

Figure 17: Molecular classification of high grade neuroendocrine lung carcinomas. Source: George et al. (2018).[167]

Both LUAD[168] and SCLC[167] transcriptomic profiling and classification attempts have previously yielded classes of samples that bear resemblance to LCNEC expression patterns.

1.3.7 Cellular atlases

The efforts to study and describe the lung cancer tissue at single-cell level have yielded multiple cell atlases to date. The initial ventures to map the cellular landscape based on self produced data have grown into projects operating on an even larger scale, integrating existing data from many different sources. Other than size alone, this has the added advantage of capturing the information from a larger number and variation of individuals and experimental protocols, thus making the result more general and representative.

The largest focus of such studies and data remains on non-small-cell lung carcinoma, and adenocarcinoma in particular. When small-cell lung carcinoma has been included in the study, it has been its primary focus, as is usual practice in lung cancer research.

Understanding how the cancerous cellular populations relate to the healthy state is also of key importance, and has not been overlooked in the research community.

Below, I briefly overview some of the latest integrative efforts of healthy lung, NSCLC and SCLC cell atlases in this context. I also include visual representations of the atlases in the form of UMAP images, an application of a dimension reduction technique that is typically used to visualise these high dimensional atlases in the 2D space.

1.3.7.1 Healthy lung

The integrated Human Lung Cell Atlas (HLCA) is the latest iteration of a human lung cell atlas developed within the scope of the Human Cell Atlas consortium[169]. Its core component represents healthy lung tissue reference, onto which further healthy or diseased data can be mapped for annotation or comparison. The premise of the atlas is, that any cell types and states which are disease specific, will end up outside the scope of the original core when projecting into the healthy space. (As an elaboration of this premise, a number of datasets, including samples from many different diseases, have been included and projected onto the core in the extended version of the atlas.)

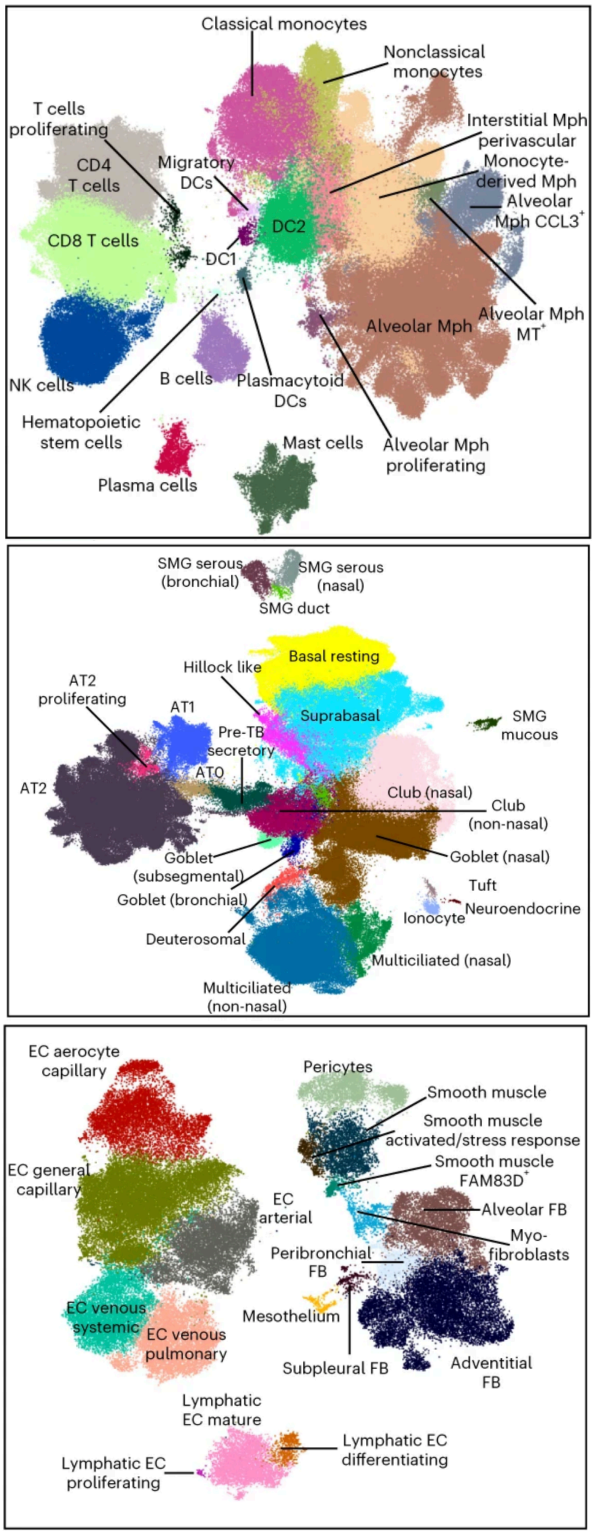

Figure 18: UMAPs of the integrated Human Lung Cell Atlas. Source: Sikkema et al. (2023)[169]

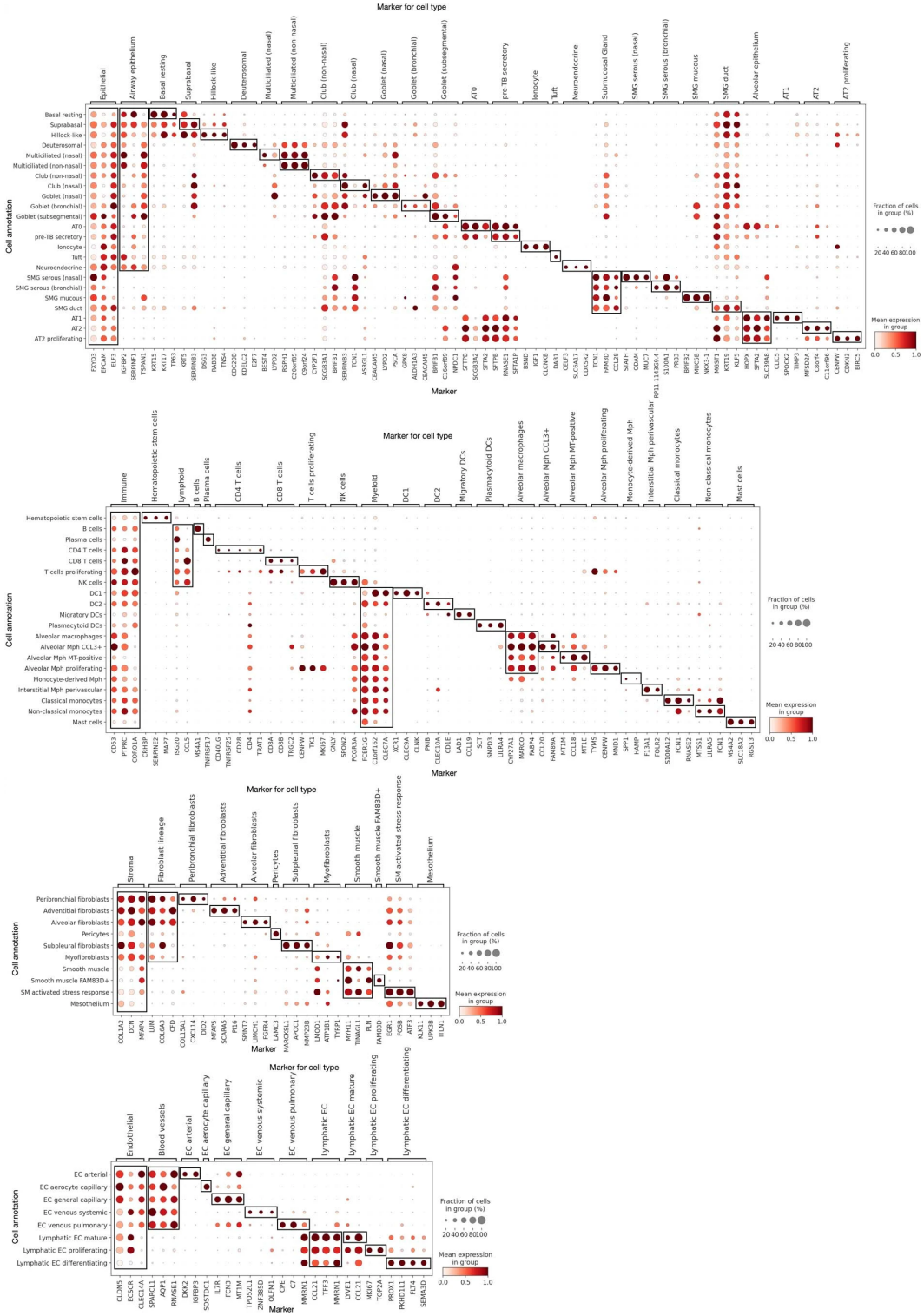

Figure 19: Markers of the 61 cell types of the integrated Human Lung Cell Atlas. Source: Sikkema et al. (2023)[169]

Several previous studies that canvassed the cell types of the healthy human lung are brought together here, along with additional sources, bringing the final core atlas composition up to 14 datasets, containing samples from 107 individuals, and 584,444 cells.

There are no consensus transcriptomic profile-based definitions of cell types, and the question of cell type annotation of scRNA-seq datasets is not a trivial one. Indeed, it is a question very tightly woven into the existence of such atlases in the first place.

The annotations were first harmonised and mapped into a five-level annotation reference, and then revised under the guidance of experts post initial annotation supported by integration of datasets. Only 41% of cell type annotations correspond to the initial ones, while 28% are further refined, and as many as 31% are substantially changed from the original studies, further highlighting the complexity of this issue.

Identification of cell types still does, in large degree, rely on clustering, and consequently the presence of a number of similar cells in the sample and/or reference, which is often not the case in rare cell types[169].

1.3.7.2 Non-small-cell lung carcinoma

There are many non-small-cell lung carcinoma scRNA-seq datasets published, with lung adenocarcinoma being particularly well covered.

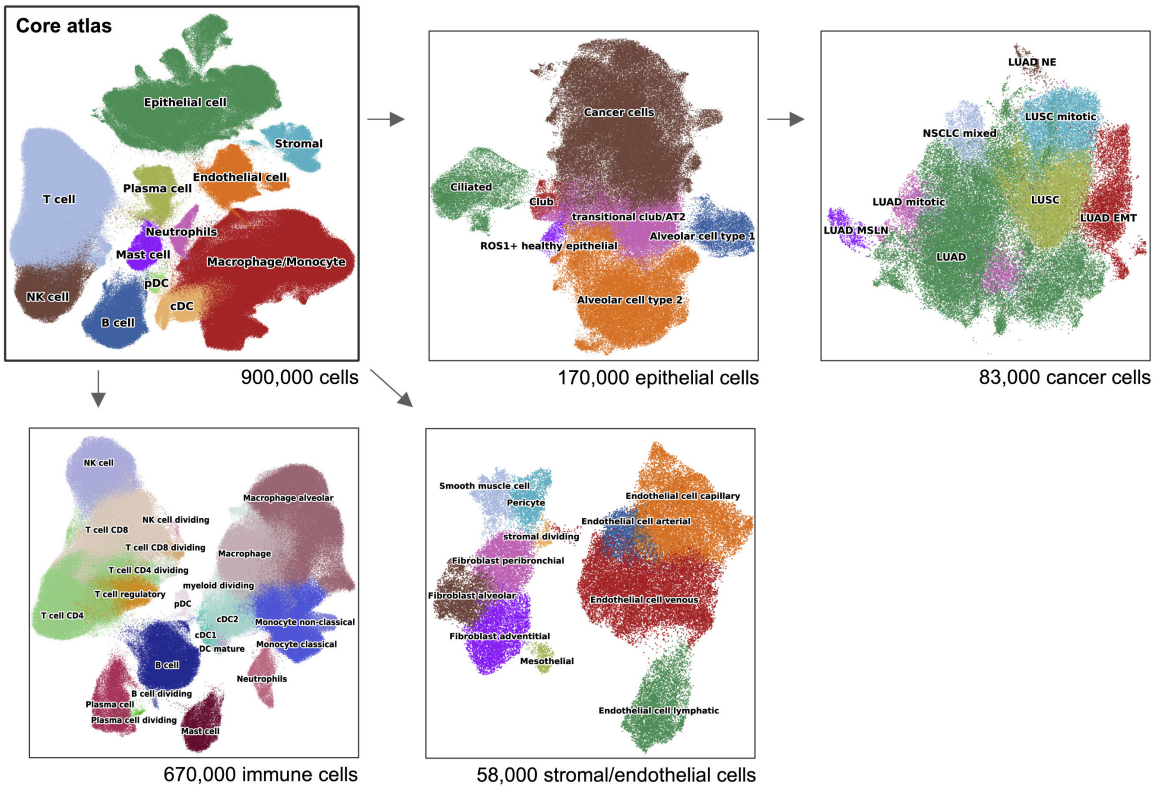

Figure 20: UMAPs of the non-small-cell lung carcinoma atlas. Source: Salcher et al. (2022)[170]

A recent large-scale integrated effort has combined 21 datasets containing 505 samples from 212 patients and 86 control individuals (156 adenocarcinomas, 41 squamous cell carcinomas, 15 not otherwise specified; 176 primary tumours, 45 metastases) into a core atlas of 898,422 cells (83,439 tumoural)[170].

The cell typing of the cancer cells is limited to assignment of post annotation-guided integration clusters based on known markers. The distances of gene expression between annotated clusters remain small and borders between annotated clusters are fuzzy.

1.3.7.3 Small-cell lung carcinoma

Though there is far less data available for this subset of lung cancer, an atlas has been generated and published within the scope of the Human Tumor Atlas Network consortium[171].

The atlas contains 155,098 cells (54,523 tumoural) from 21 fresh SCLC clinical samples, sourced from 19 patients. Additional adjacent normal tissue and adenocarcinoma samples are included in the atlas as a reference.

Though canonical subtypes (SCLC-A, SCLC-N, SCLC-P, SCLC-Y) can be identified based on their markers, and a clinically interesting PLCG2-high subpopulation has been identified within the atlas, further exploration of types and states within cancer cell populations is difficult. The tendency of cells to cluster with the rest of cells from the same tumour is very strong, and integrating samples from different patients without removing all signals of interest completely remains an elusive task.

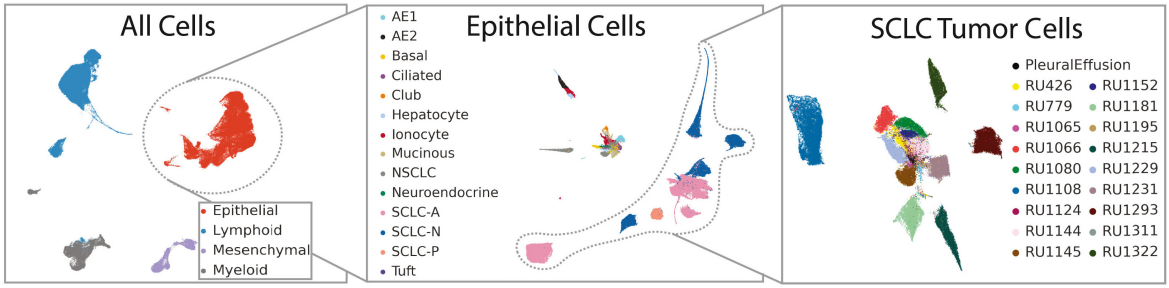

Figure 21: UMAPs of the small-cell lung carcinoma atlas. Source: Chan et al. (2021)[171]

As it is, a transcriptomic reference of such tumours and their corresponding microenvironments is still an invaluable resource.

1.3.8 Cell states and plasticity in lung tumours

A single-cell analysis of tumour and matched normal tissue samples from 16 early-stage lung adenocarcinoma patients revealed that diversity among cancer cells was strongly linked to specific drivers. The malignant epithelial cells displayed loss of lineage identity and differed strongly between patients. *KRAS*-mutant patients in particular clustered closely together and showed a unique transcriptional reprogramming, with strikingly reduced differentiation, few copy number events, and

high variability amongst the cells. A specific subpopulation of alveolar intermediate cells (originating from type II pneumocytes) characterized by high *KRT8* expression showed a low differentiation and increased plasticity as well as a specific transcriptional profile, which was found to be enriched in adenocarcinomas. It has been proposed these cells transition into malignant ones, a process specific for *KRAS*-mutant adenocarcinoma patients[172].

Another single-cell study profiled 16 patients of different histological subtypes. It has demonstrated that single-lineage and mixed-lineage tumours in the same patients originated from the same cell, highlighting the plasticity involved in acquisition of multiple-lineage features[173].

High cell plasticity has also been observed in SCLC, where a study based on 12 patient single-cell profiles identified several types of hybrid cell states: epithelial/mesenchymal, tumour/immune, hybrid of two SCLC subtypes and hybrid of SCLC and adenocarcinoma cells, pointing to potential lineage transitions[174].

2 Objectives

THE PURPOSE of this thesis is to establish a framework for characterisation of intra-tumour heterogeneity of lung cancers. The work set out here centres around transcriptomic data based characterisation, and is focused on single-cell and spatial technologies in particular. The key objectives are both methodological and applied.

2.1 Characterisation of intra-tumour heterogeneity

Single-cell transcriptomics based ITH characterisation includes several different aspects:

1. Study of cell types and cell-cell communication
2. Study of factors that modulate these communities
3. Study of relationships between communities and aggressiveness of cancer

Spatial transcriptomic based ITH characterisation is similar, but involves cell communities organised in tissue, i.e. niches, and potential deconvolution of the single-cell level information.

2.2 Methodological objectives

The first objective of this thesis was to support these points of ITH characterisation methodologically. Three algorithms to this point are presented in the next chapter of this work.

2.2.1 Marker identification

A marker identification algorithm addresses the first two points mentioned above - it first reveals the factors modulating the cell types; later, these are used to identify and reveal the presence of cell communities in single-cell, spatial or bulk transcriptomic data, and even further communities/niches in the last case. Though there are many existing marker identification algorithms, the objective was to develop a method capable of identifying markers more suitable for knowledge transfer, prioritising higher specificity for a given cell type. The other key objective was to capture promising candidates even at low expression levels by simulating scenarios of higher availability of reads.

2.2.2 Spatial pattern classification

Gene expression spatial patterns might point to an underlying organised cell community in tissue. A classifier of spatial patterns identifying patterns pointing to structures of biological interest (barriers, gradients, ...) and those of no interest (random distribution) contributes to niche identification, the equivalent of first point listed above for spatial transcriptomic data. The objective was to first explore spatial pattern types through a custom spatial data generator and then develop a classifier separating the classes defined.

2.2.3 CNV inference

A copy number variation inference algorithm addresses the final point listed above. Though such algorithms already exist, the objective was to improve on their limited sensitivity on single-cell data, where the number of reads is typically low.

2.3 Lung cancer cell communities

The final objective of this thesis was to canvas the cell types in all major histologies of lung cancer based on available published single-cell transcriptomic datasets through the construction of a unified lung cancer atlas, along with annotation and corresponding markers suitable for knowledge transfer.

3 Results

The following chapter is split into four sections. The first three contain methodological results in the form of a single-cell marker identification algorithm, a spatial pattern classifier and a copy number variation inference algorithm. The final section of the chapter is dedicated to the unified lung cancer cell atlas.

3.1 SCherlock algorithm - robust cell type markers identification

Contribution: I participated in the design and fully implemented the code of the SCherlock algorithm, as well as all extensive benchmarking. An article presenting this method is in the works, which I will co-sign as first author.

3.1.1 Introduction

Numerous single-cell transcriptomic atlases covering different tissues, organs of the human body, stages of development and diseases have been published in recent years since the outset of the Human Cell Atlas Project[63]. They often include extensive cell type annotations, sometimes at a very granular level, with many reporting newly discovered cell types[175–177]. However, the utility of such atlases is limited by the lack of robust, usable and transferable gene markers of cell populations, supported by a statistical framework.

SCherlock is a novel method that allows identification of such markers, as well as cases where this is not possible. Two rounds of scoring are performed. The initial theoretical scoring serves to eliminate the majority of genes from consideration. In a departure from standard practice in the field[178,179], binomial rather than negative binomial distribution is used to model counts in preliminary estimations despite its limitations, as its parametrisation allows for simple consideration of scenarios with increased sensitivity, key to this algorithm. The final scores are obtained via more computationally expensive empirical scoring of a limited number of top scoring candidate genes performed on expression measurements directly.

3.1.2 Methods

3.1.2.1 Algorithm overview

In a general sense, the algorithm consists of several steps that are performed in sequence. Each is described in further detail in the following sections. There are multiple alternatives to choose from at some of them, in accordance with user needs. The steps include:

1. Theoretical distribution parameter estimation, where parameters are fit to the dataset
2. Theoretical distribution scoring, where initial marker scores are computed based on this theoretical distribution, with the following sub-steps:
   (a) Multiplication factor introduction enables consideration of situations of higher sensitivity
   (b) Cross-cell type correction penalises genes that are markers of multiple cell types
   (c) Aggregation, ordering and filtering finalises the theoretical scores and identifies candidate genes for further evaluation
3. Empirical distribution scoring, where marker scores are recomputed on data for top candidate genes
4. Metamarker construction, where individual marker power can be combined for increased sensitivity

3.1.2.2 Theoretical distribution parameter estimation

For each gene individually, a distribution is fitted to model its expression in cell type A, and another one to model its expression in the complement (all the other cell types) $\bar{A}$.

3.1.2.2.1 Notation

Counts of gene i in cell j in sample s: $x_{ij}^{(s)}$

Counts of gene i in all cells of cell type A in sample s:

$y_{iA}^{(s)} = \sum_{j \in A} x_{ij}^{(s)}$

Total number of counts across the entire set of genes G for all cells of cell type A in sample s:

$N^{(s)} = \sum_{i \in G} y_{iA}^{(s)} = \sum_{i \in G} \sum_{j \in A} x_{ij}^{(s)}$

Similarly, for the complement of all other cell types:

$y_{i\bar{A}}^{(s)} = \sum_{j \notin A} x_{ij}^{(s)}$

$N'^{(s)} = \sum_{i \in G} y_{i\bar{A}}^{(s)} = \sum_{i \in G} \sum_{j \notin A} x_{ij}^{(s)}$

3.1.2.2.2 Binomial distribution fitting

The binomial distribution $B(n, p)$ is parametrised by the number of trials n and the probability of success p. In this context, n represents the number of all transcripts (counts) in a cell of a given cell type or its complement, while p represents the proportion of transcripts that belong to the gene in question.

$\widehat{p_{iA}} = median_s(\frac{y_{iA}^{(s)}}{N^{(s)}}), \widehat{nA} = median_s(\frac{N^{(s)}}{|A|})$

$\widehat{p_{i\bar{A}}} = median_s(\frac{y_{i\bar{A}}^{(s)}}{N'^{(s)}}), \widehat{n\bar{A}} = median_s(\frac{N'^{(s)}}{|\bar{A}|})$

A median value across the samples is used when a column containing sample annotation is specified. If not, everything is considered a single sample and the estimates are computed on the entire dataset directly, which amounts to a mean calculation approach.

3.1.2.3 Theoretical distribution scoring

After fitting the theoretical distributions of each gene for cell type A and its complement, its initial score as a marker for A is computed. For a given threshold l, any value of the marker equal or greater to it is assigned to belong to cell type A. There are three different scoring modes available:

- 'diff'

  We find the threshold l, where the difference between sensitivity and the false positive rate (FPR) is the largest, with the difference itself being the score:

  $max_l(sensitivity_l - FPR_l)$

- 'sensFPRzero'

  The threshold l is set to the lowest value, where FPR reaches zero. The score is the sensitivity calculated at this threshold:

  $sensitivity_{FPR=0}$

- 'sensPPV99'

  The threshold l is set to the lowest value, where the positive predictive value (PPV) reaches 0.99 (which is not guaranteed to happen, the score is zero otherwise). The score is the sensitivity calculated at this threshold:

  $sensitivity_{PPV=0.99}$

For any given number of counts c and the distributions of gene i in A and $\bar{A}$ estimated in the previous steps, we consider the values of their cumulative distribution functions:

$\alpha_c^{(i)} := cdf_A^{(i)}(c)$

$\beta_c^{(i)} := cdf_{\bar{A}}^{(i)}(c)$

The confusion matrix at threshold l can then be computed as follows:

	$\#counts < l$	$\#counts \geq l$
A	$\alpha_l^{(i)}$	$1 - \alpha_l^{(i)} = sensitivity_l$
$\bar{A}$	$\beta_l^{(i)}$	$1 - \beta_l^{(i)} = FPR_l$

The scores can then be computed as:

- 'diff'

  $max_l(sensitivity_l - FPR_l) = max_l(\beta_l^{(i)} - \alpha_l^{(i)})$

- 'sensFPRzero'

  $sensitivity_{FPR=0} = 1 - \alpha^{(i)}_{argmax_l(\beta_l^{(i)})}$

- 'sensPPV99'

  $PPV_l = \frac{1-\alpha_l^{(i)}}{2-\alpha_l^{(i)}-\beta_l^{(i)}}$

  $sensitivity_{PPV=0.99}$ = not guaranteed to exist

3.1.2.3.1 Multiplication factor introduction

The power of markers to distinguish between cell types is dependent on the number of total reads available (see Figure 22). As the single-cell context typically represents the one with the fewest reads, we introduce a multiplication factor k, which enables us to consider scenarios of increased sensitivity (higher number of reads available), in order to better grasp the true marker potential of genes.

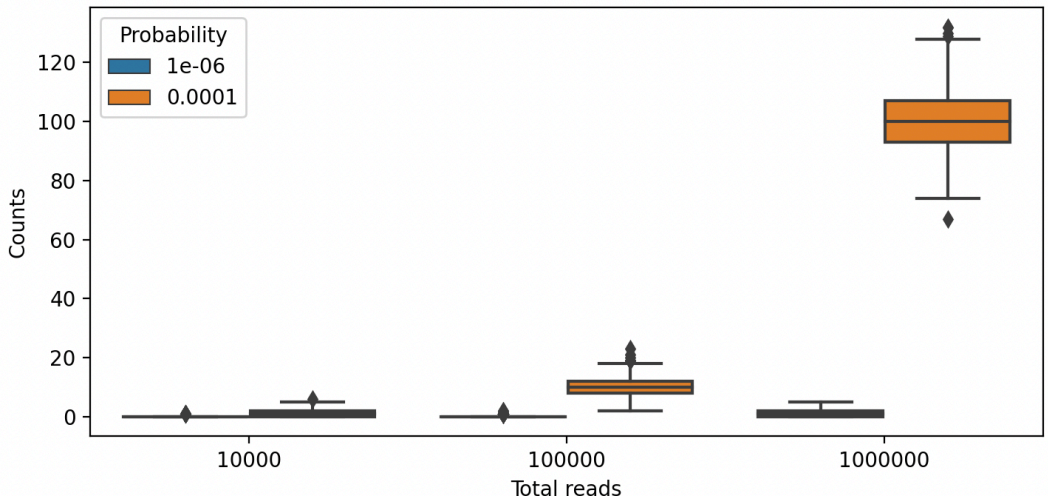

Figure 22: Simulated counts for a theoretical example marker gene in its cell type population ($p = 10^{-4}$) and the rest ($p = 10^{-6}$) as a function of total reads. $N = 1000$ simulations were ran for each scenario.

We simply multiply the first parameter of binomial distribution, representing the number of total reads, with k. We test a specified range of k values (1, 10, 25, 50 and 100 by default). Using pairs of distributions $B(k * \widehat{nA}, \widehat{p_{iA}})$ and $B(k * \widehat{n\bar{A}}, \widehat{p_{i\bar{A}}})$, and the chosen scoring mode, we compute the initial score for each value k:

$$score_k(i, A)$$

3.1.2.3.2 Cross-cell type correction

For each gene, we divide the initial marker scores for each cell type by their sum.

$$score'_k(i, A) = \frac{score_k(i, A)}{\sum_{A \in \{celltypes\}} score_k(i, A)}$$

3.1.2.3.3 Aggregation, ordering and filtering

The final scores are computed either as a mean or as a max across the values of k post correction:

$$score_{mean}(i, A) = \frac{1}{|k|} \sum_{k} score'_k(i, A)$$

$$score_{max}(i, A) = max_k(score'_k(i, A))$$

Genes with identical scores are ordered by their expression levels within the cell type in question (the more expressed ones having priority).

In the filtering step, only genes with final scores > 0.5 are kept for each cell type. This ensures that the metamarker of a given cell type cannot contain a gene which is a better marker of another cell type, as the corrected score would fall below 0.5 in such case.

Additionally, while genes with very low expressions are being considered, a lower limit on the absolute number of reads in the target cell type for a potential marker can be set. In order to eliminate patient biased discoveries, a minimum number of patient with non-zero number of reads in the target cell type can also be set.

In case of many remaining candidates, further cuts can be made for the sake of practicality (max 100 genes per cell type considered by default).

3.1.2.4 Empirical distribution scoring

Marker scores are recalculated on the empirical distributions. For the multiplication factor part, an equal number of metacells are simulated by pooling k random cells and summing their count profiles. As this simulation step is computationally expensive, the empirical scores are calculated only on the best candidates for each cell type, preselected based on theoretical scores.

3.1.2.5 Metamarker construction

Individual marker genes can be combined into more powerful metamarkers with increased sensitivity and robustness.

There are different potential approaches to metamarker construction. Various criteria might be imposed on genes for inclusion and cutoff decisions. One straightforward way used in the examples here is to begin at the top of empirical scoring the table and include all genes with scores greater than 0.5. The behaviour of a metamarker is evaluated by computing the collective counts and treating it as a single gene marker.

3.1.3 Results

3.1.3.1 Specificity of markers

SCherlock deviates from existing marker methods and published marker lists accompanying cell atlases, which typically focus on p-value and log-fold change as the marker quality criteria. When pure, high-quality markers exist, such criteria will retrieve them. In absence of such markers, less specific markers scoring highly on both fronts rise to the top, which have less power to distinguish between cell types. In fact, the same genes can be proposed as markers of multiple cell types, as seen in the example of Hao et al. (2021)[180] in Figure 23.

The markers obtained on the same dataset by SCherlock ($k = [1, 10, 25]$, 'diff' scoring, mean aggregation, minimum of 10 reads and 3 patients at target cell type) are shown in Figure 24.

With SCherlock, specificity rather than sensitivity is prioritised. The cross-cell type correction in combination with filtering eliminates any genes which are stronger or equally strong markers of other cell types from consideration. It also penalises weaker associations. The example in Figure 25 illustrates this difference of priority. We can see that JCHAIN, proposed by Hao et al. as a marker of pDCs, is much more expressed in plasma cells and plasmablasts than in pDCs, meaning that its expression is not specific to pDCs. SCherlock identifies KRT5 as a marker of pDCs. In the context of only PBMCs, where it has been proposed, it appears to indeed be expressed in pDCs specifically (NB: KRT5 being known as a basal epithelium marker, it would likely be discarded by SCherlock in a wider context).

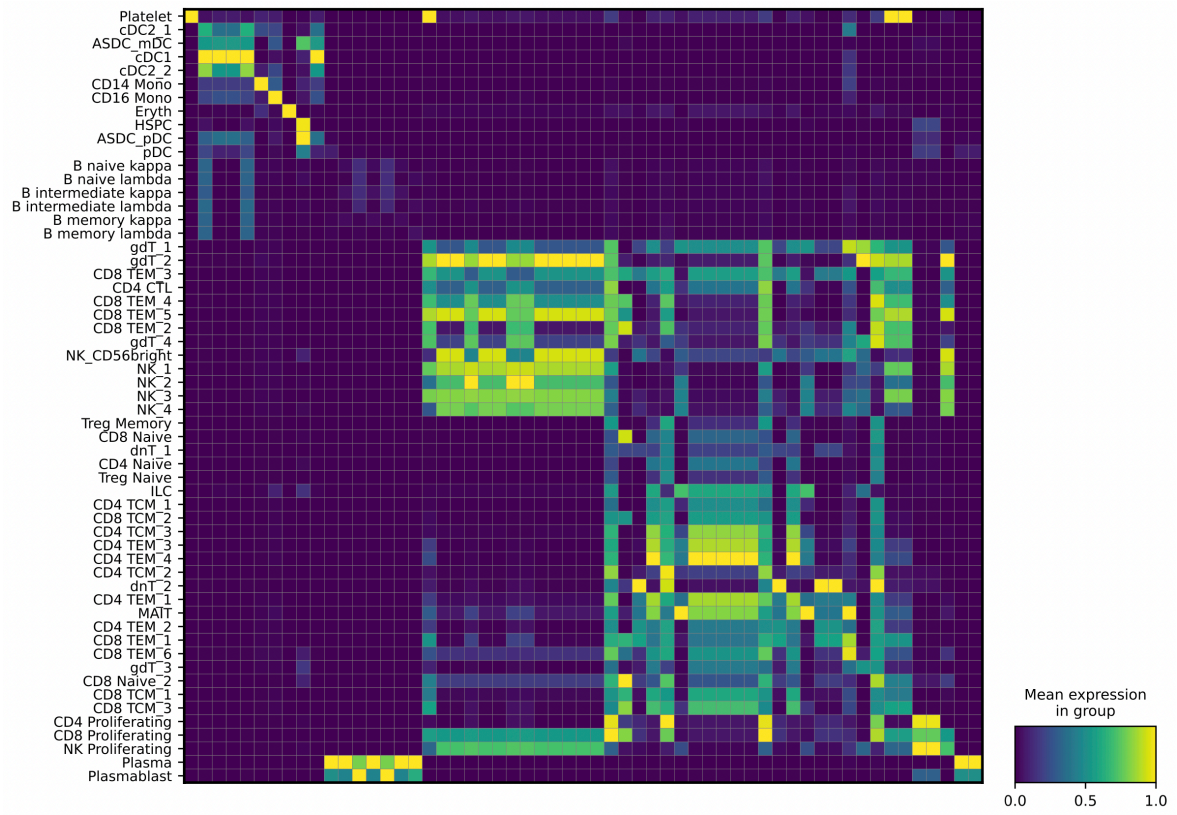

Figure 23: Matrixplot of proposed marker genes with highest log-fold change for each cell type in Hao et al. (2021)[180]. Cell types are shown in rows, their respective markers are plotted in columns. Mean expression is scaled by gene.

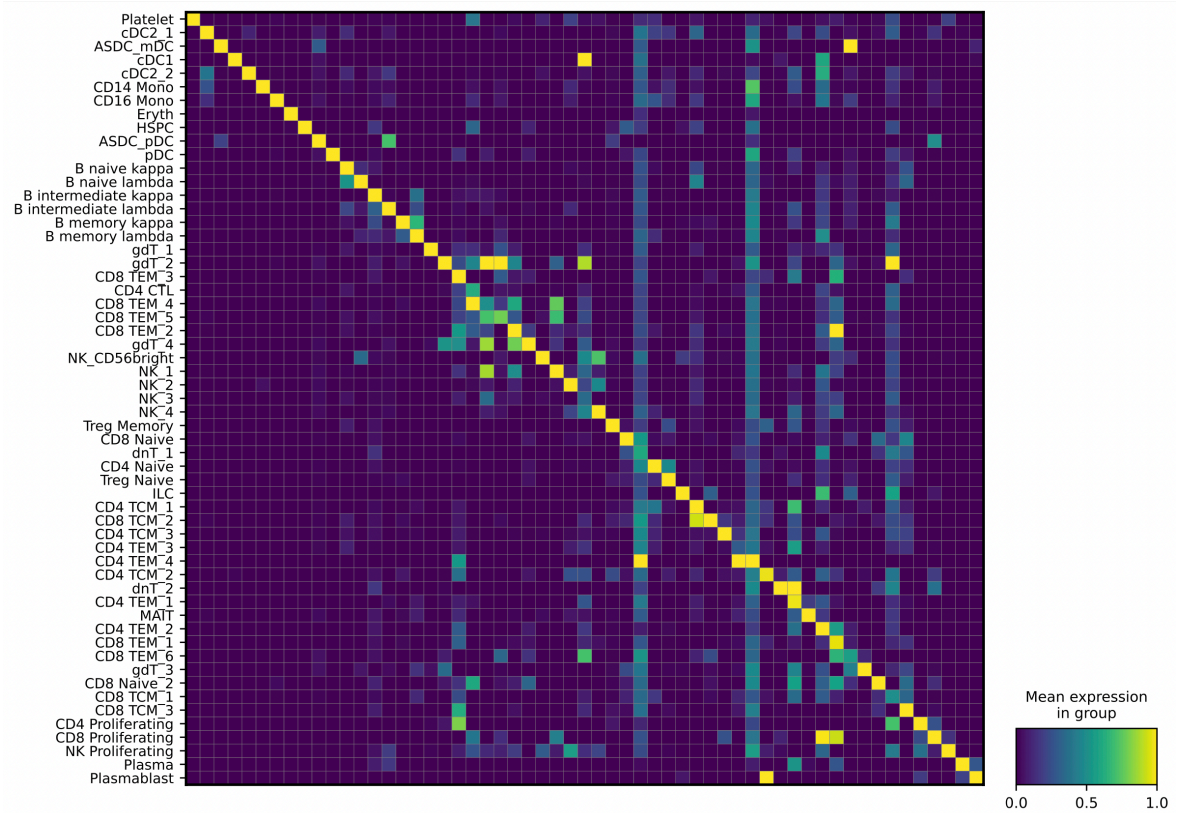

Figure 24: Matrixplot of the highest scoring marker genes according to SCherlock evaluation on Hao et al. (2021)[180]. Cell types are shown in rows, their respective markers are plotted in columns. Mean expression is scaled by gene. Though top scoring candidates are shown for all cell types, no genes meet the conditions to be identified as markers for ASDC pDC, B intermediate lambda, B memory lambda, CD4 CTL, CD4 Proliferating, CD4 TCM 1-2, CD4 TEM 1-4, CD8 Proliferating, CD8 TCM 1&3, CD8 TEM 1-6, NK 1&3-4, dnT 1, and gdT 2 cells.

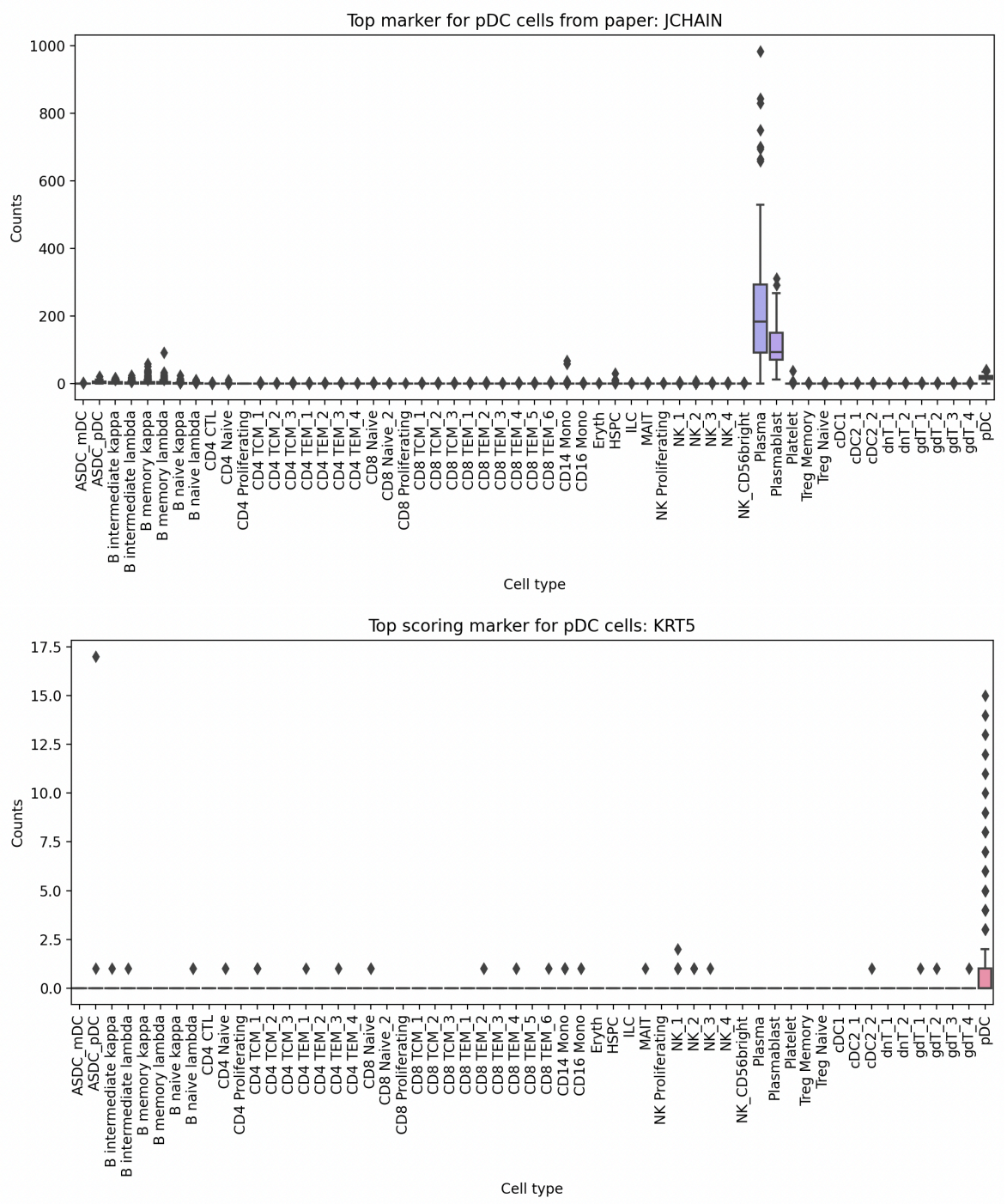

Figure 25: Expression of *JCHAIN* (top), marker for pDC cells from Hao et al. (2021)[180] with the highest log-fold change, and *KRT5* (bottom), highest scoring marker gene according to theoretical score, not featured on the list of markers supplied with the original paper.

3.1.3.2 Contextuality of markers

Markers are typically provided at the most granular level of a given published cell atlas, which often goes beyond established cell types. It is hardly surprising that not all of them would have specific transcriptomic markers to differentiate them from all the rest (see Figures 23 & 25).

On the flip side, an annotation level that is too coarse might include a mixture of cell types that does not possess a common marker.

PBMC

B	CD4 T	CD8 T	other T	DC	Mono	NK	other
B naive B intermed B memory Plasmablast	CD4 CTL CD4 naive CD4 prolif CD4 TCM CD4 TEM Treg	CD8 naive CD8 prolif CD8 TCM CD8 TEM	MAIT dnT gdT	ASDC cDC1 cDC2 pDC	CD14 mono CD16 mono	NK NK prolif NK CD56^{br}	Eryth HSPC ILC Platelet
λ, κ *λ, κ* *λ, κ* *Plasma, Plasmablast*	- - - *#1-3* *#1-4* *Treg naive, Treg memory*	*#1-2* - *#1-3* *#1-6*	- *#1-2* *#1-4*	*ASDC_mDC, ASDC_pDC* - *#1-2* -	- -	*#1-4* - -	- - - -

Figure 26: The hierarchy of cell type annotation levels in Hao et al. (2021)[180].

Identifying an appropriate level of annotation granularity and scope is an iterative task. Levels of the annotation hierarchy provided with the cell atlas provide a good starting point, which is then further refined by mixing them together, merging and eliminating categories as necessary. An example of such a refining is included here for the Hao et al. (2021)[180] dataset. The original hierarchy is shown in Figure 26, while a set of SCherlock marker genes for the dataset at a revisited level of granularity is shown in Figure 27.

Limitations of what is possible are not the only factor at play, the construction must correspond to the requirements of the end goal. Any known or expected absences in the target data or application might mean there are types we can eliminate from consideration, as they will not be part of the operating context. Their removal has the potential to open up possibilities that would not be viable otherwise. In some cases, it might furthermore be useful to eliminate any intermediate/transitional types from consideration when defining markers.

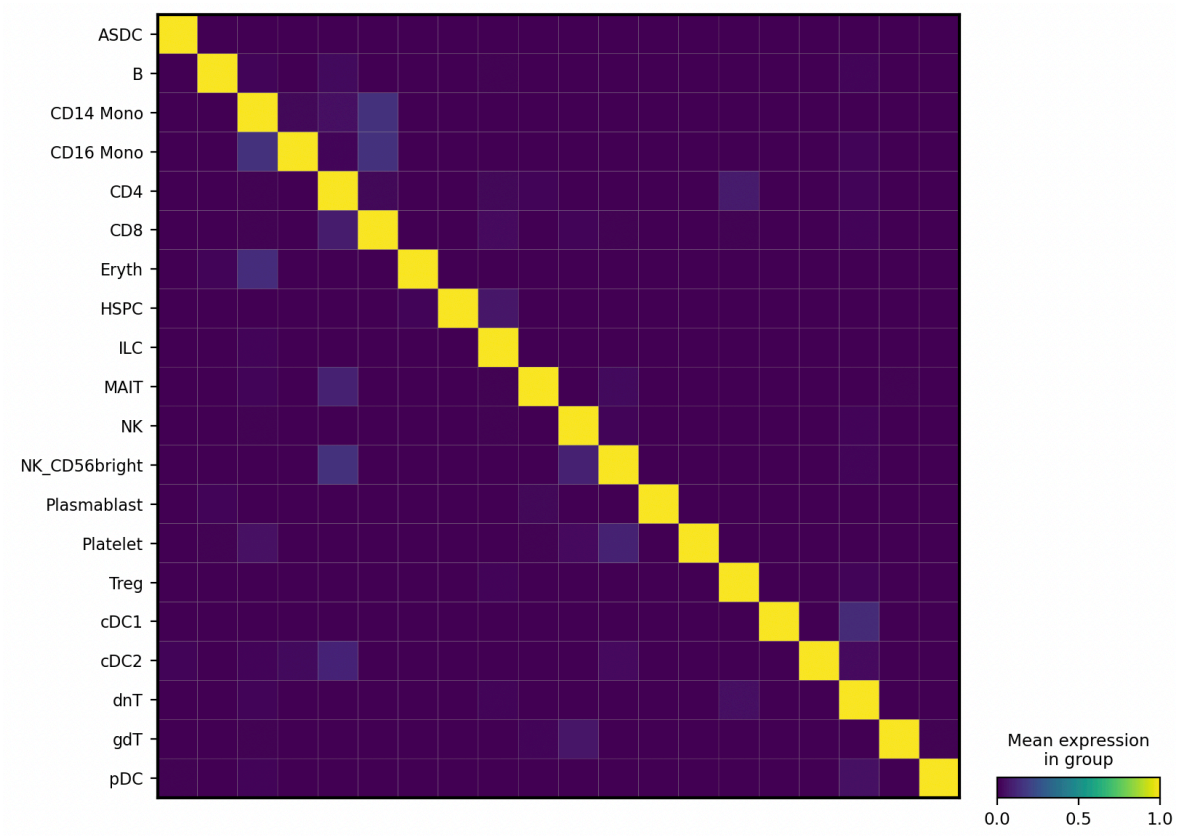

Figure 27: Matrixplot of the highest scoring marker genes according to SCherlock evaluation on the revised annotation level for Hao et al. (2021)[180]. Cell types are shown in rows, their respective markers are plotted in columns. Mean expression is scaled by gene.

3.1.3.3 Detection of markers with low levels of expression

Though SCherlock gives priority to genes with higher expression levels when scores are otherwise tied, its emphasis on specificity rather than sensitivity, and especially its particular consideration of scenarios of increased sensitivity, mean that genes with very low levels of expression are in no way deprioritised.

Top scoring markers genes for the revisited scope of cell types of the Hao et al. (2021)[180] dataset include biologically coherent examples of both highly expressed (Figure 28) and lowly expressed (Figure 29) nature.

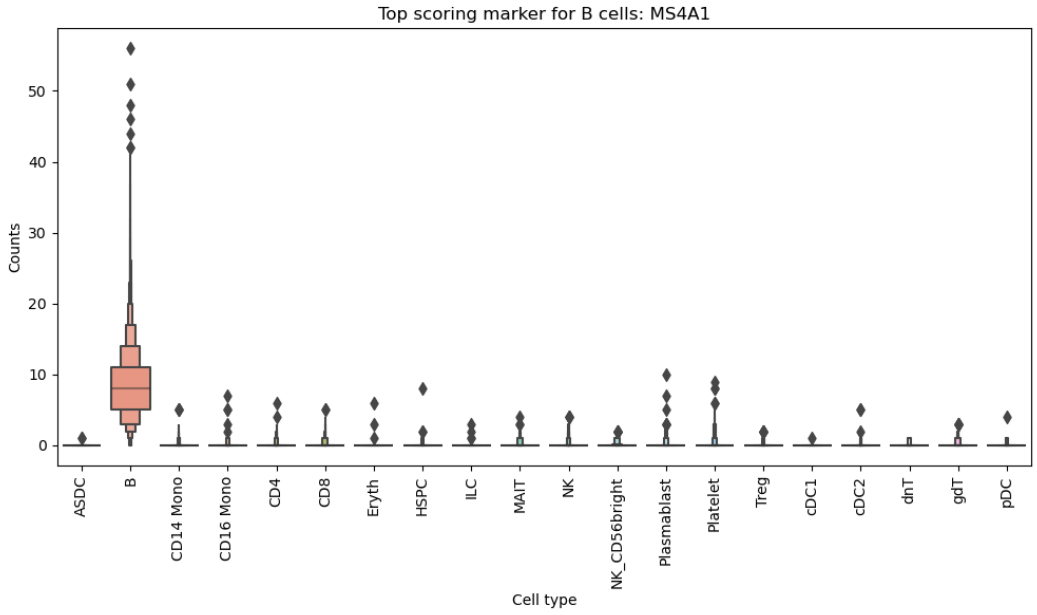

Figure 28: Expression of *MS4A1* (*CD20*), the highest scoring marker gene for B cells according to SCherlock evaluation on Hao et al. (2021)[180].

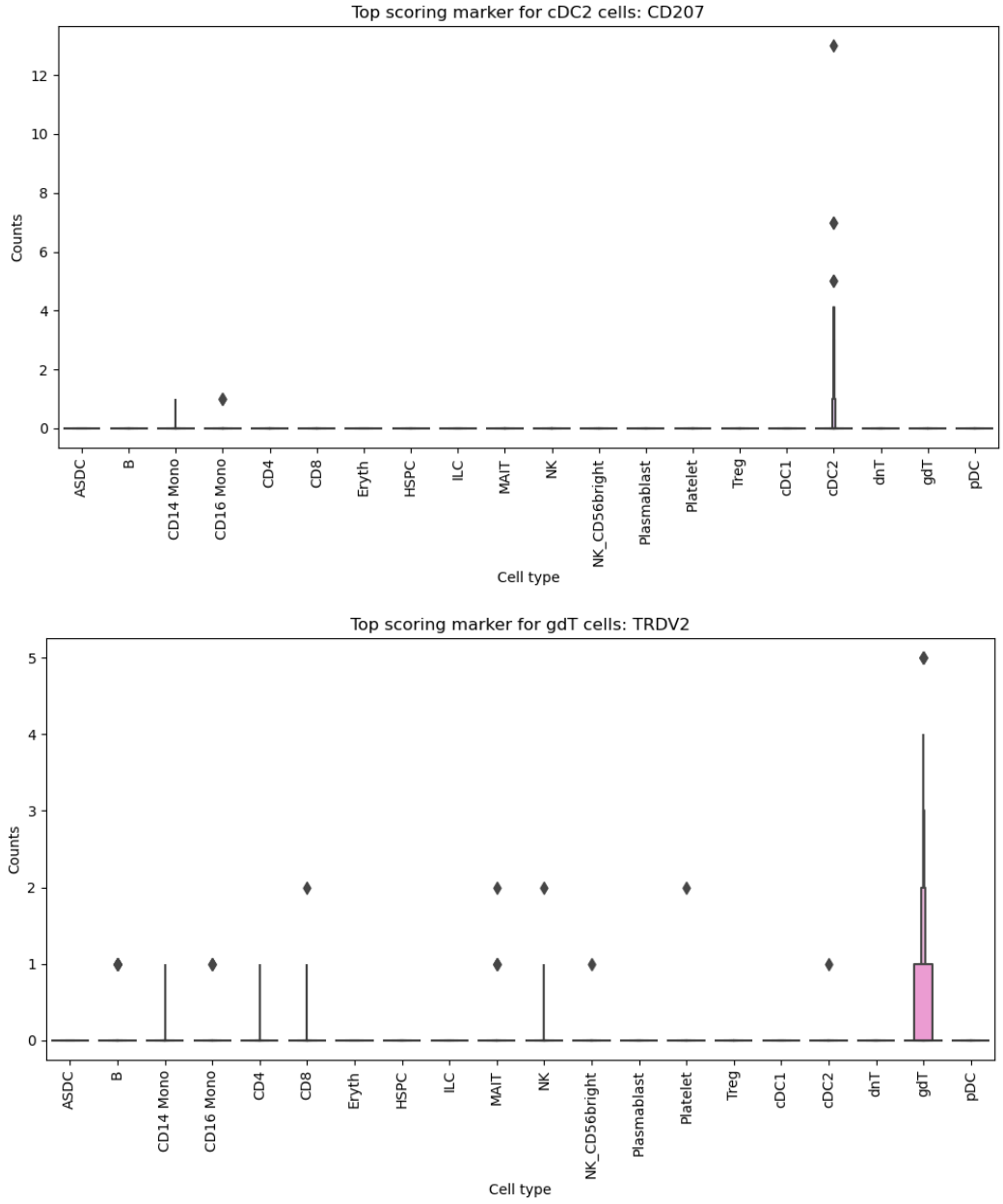

Figure 29: Expression of *CD207* (top) and *TRDV2* (bottom), the highest scoring marker genes for cDC2 and gdT cells respectively according to SCherlock evaluation on Hao et al. (2021)[180].

3.1.4 Discussion

Direct transfer of knowledge via robust markers is established practice when it comes to immune cells (MCPcounter[181], CIBERSORTx[182]), but not otherwise.

SCherlock is a single-cell data marker identification method, which opens up new avenues to harness the vast number of existing single-cell atlas resources. The markers it identifies are very specific, and can be well suited for use in identification and deconvolution contexts.

There are several limits to the method. The specificity often comes with a trade-off of lower sensitivity. This can be overcome in several ways - multiple markers combined into metamarkers; contexts with more reads (e.g. application in bulk transcriptomics); treatment of clusters rather than individual cells.

Furthermore, for the desired specificity, any cross-expression of markers is penalised relatively harshly. The assumption here is that the cell types are labelled correctly. Misannotation of cell types in the source dataset could cause potential issues down the line.

Another potential concern is that we aren't necessarily able to tell if a low level expression marker detected only in a few cells is truly a marker of the entire population of that type, or specific of a hidden subpopulation. It is important to validate annotation levels that are available, but that only covers any subtypes included in the annotation. Issues might arise in applications in case of subpopulations in a different balance.

In a deconvolutional context, biases might arise from the number of reads being cell type dependant, and favour cells with high expression levels (e.g. tumour) in bulk contexts.

3.1.5 Supplementary results

To complete the results presented above, I present below the top markers, identified using SCherlock, of the other immune cell populations from Hao et al. Of note, several of these markers are reported in the literature (mostly in proteinatlas.org) as markers of the corresponding cell types (e.g. CYP1B1 and AC106712 for monocytes, HBM for erythrocytes, SLIT3 for NK cells, SDC1 for B lineage, GP9 for platelets, ERICH5 and KCTD19 for DCs).

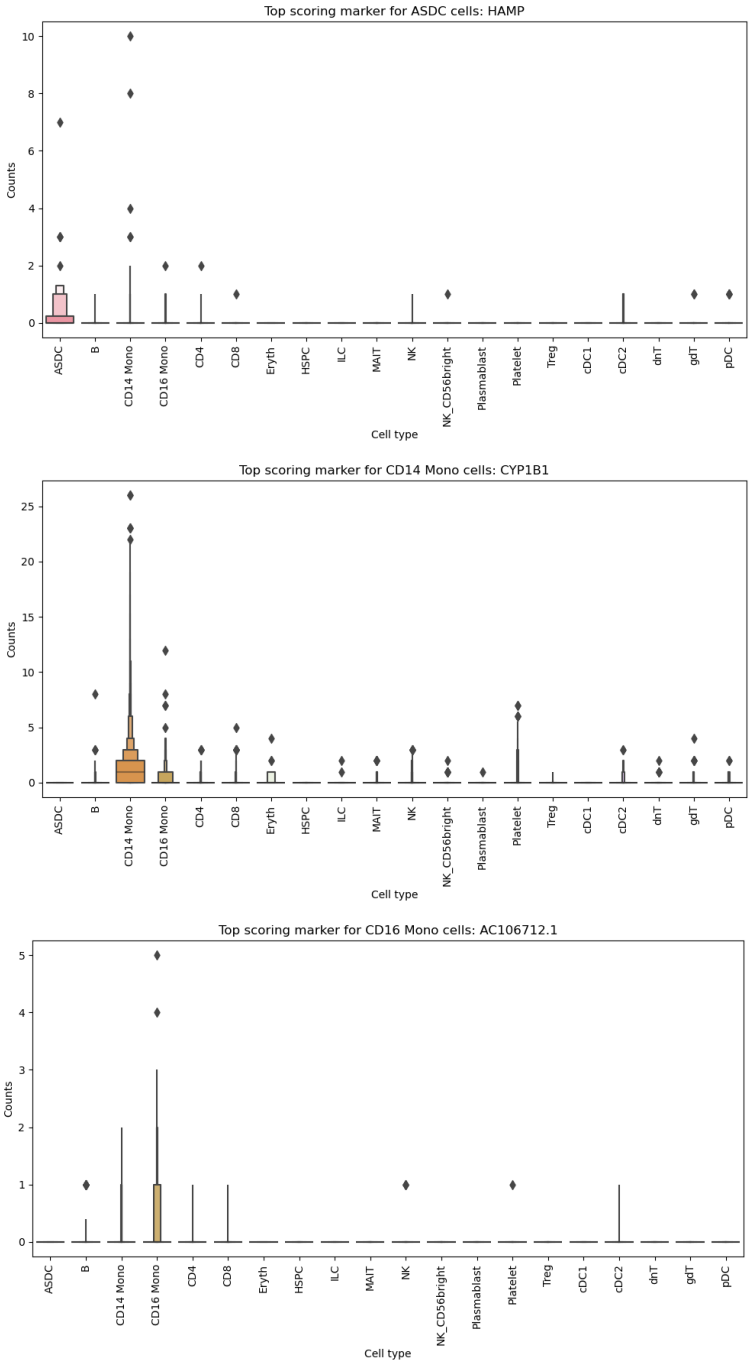

Figure 30: Expression of *HAMP* (top), *CYP1B1* (middle), and *AC106712.1* (bottom), the highest scoring marker genes for ASDC, CD14 Mono and CD16 Mono cells respectively according to SCherlock evaluation on Hao et al. (2021)[180].

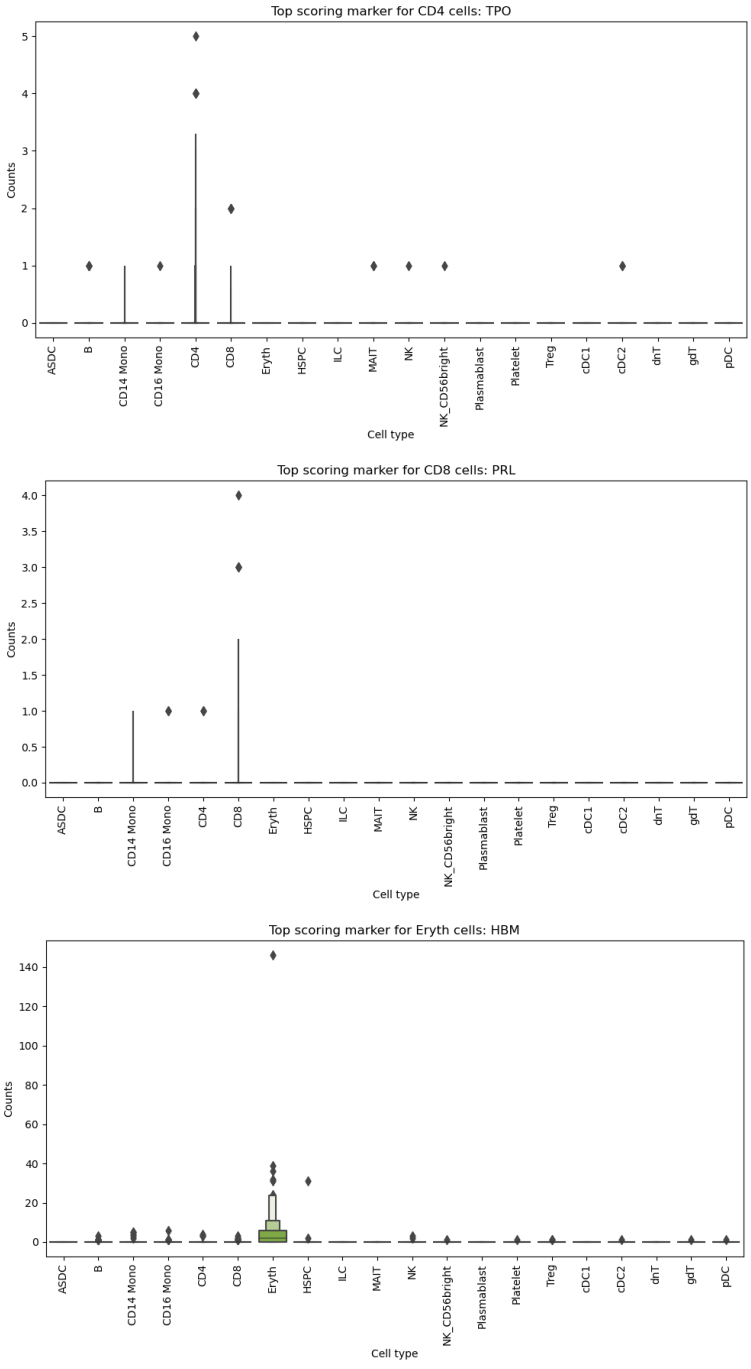

Figure 31: Expression of *TPO* (top), *PRL* (middle), and *HBM* (bottom), the highest scoring marker genes for CD4, CD8 and Eryth cells respectively according to SCherlock evaluation on Hao et al. (2021)[180].

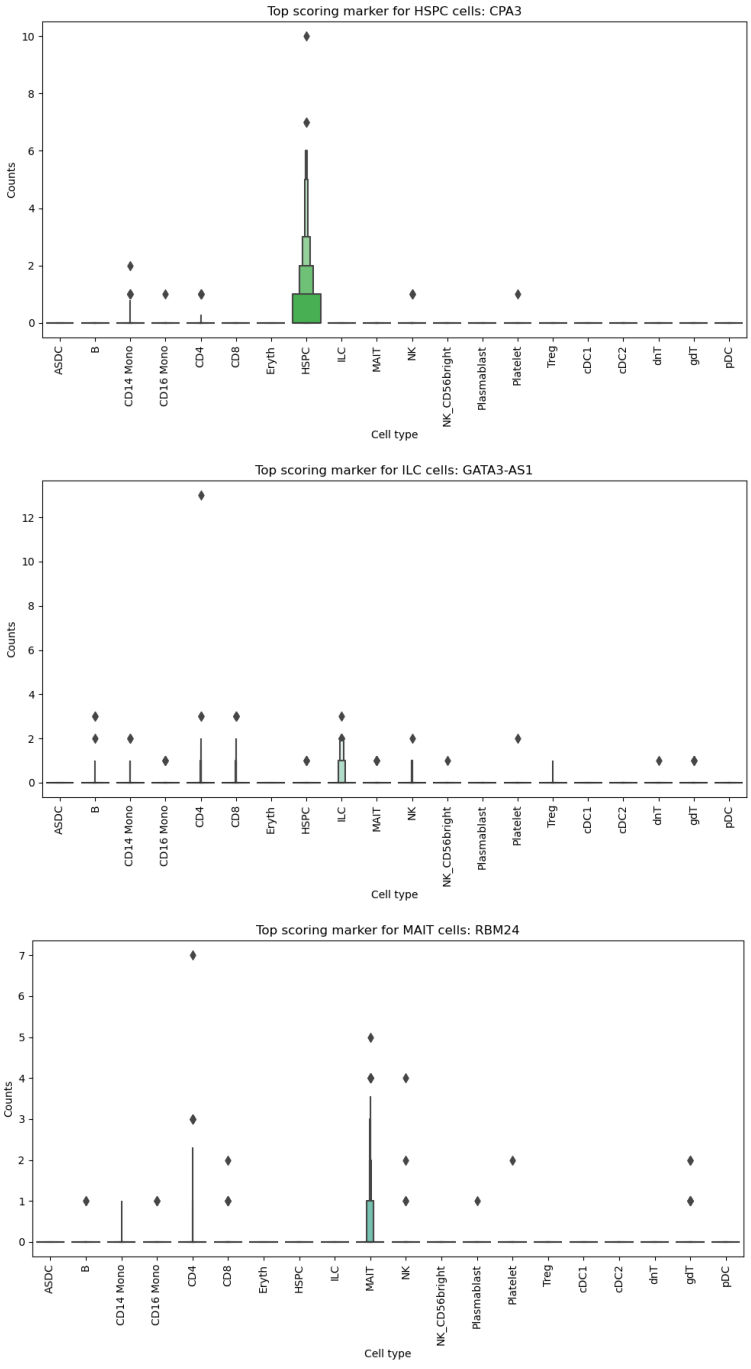

Figure 32: Expression of *CPA3* (top), *GATA3-AS1* (middle), and *RBM24* (bottom), the highest scoring marker genes for HSPC, ILC and MAIT cells respectively according to SCherlock evaluation on Hao et al. (2021)[180].

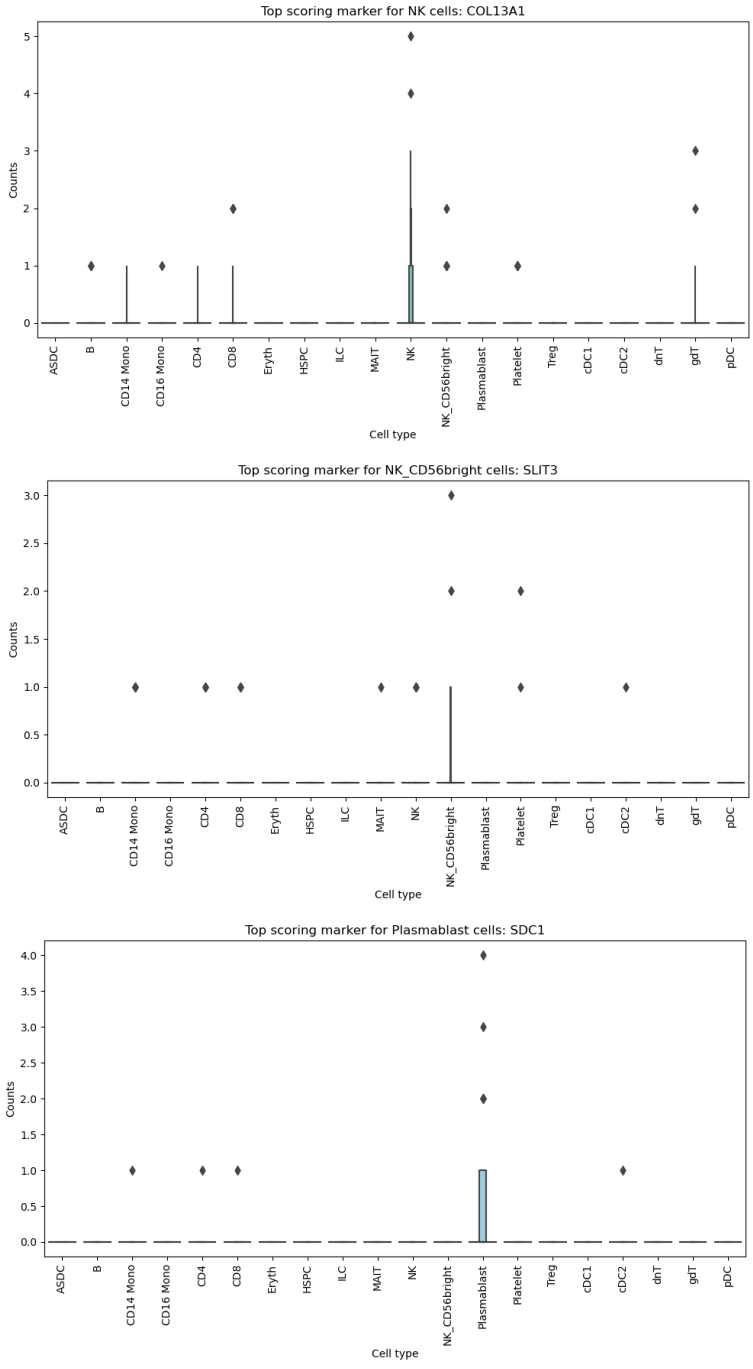

Figure 33: Expression of *COL13A1* (top), *SLIT3* (middle), and *SDC1* (bottom), the highest scoring marker genes for NK, NK CD56bright and Plasmablast cells respectively according to SCherlock evaluation on Hao et al. (2021)[180].

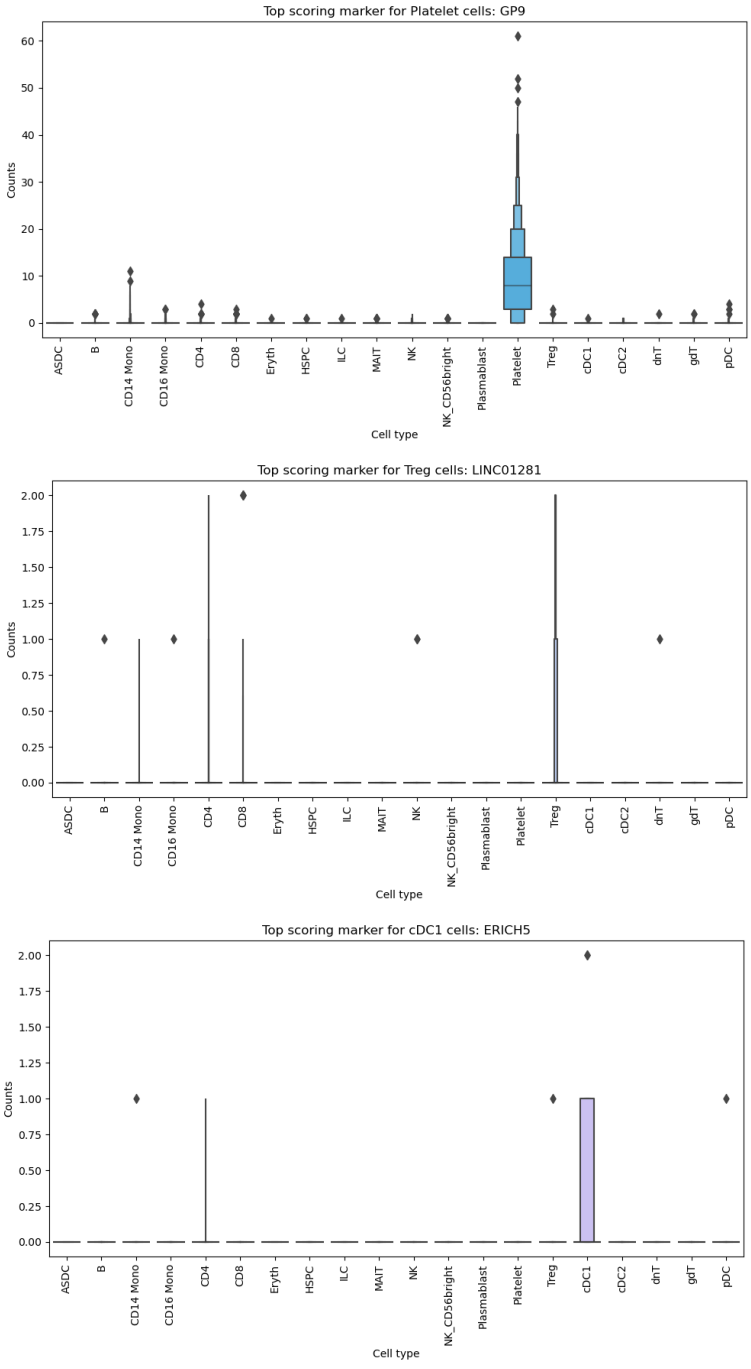

Figure 34: Expression of *GP9* (top), *LINC01281* (middle), and *ERICH5* (bottom), the highest scoring marker genes for Platelet, Treg and cDC1 cells respectively according to SCherlock evaluation on Hao et al. (2021)[180].

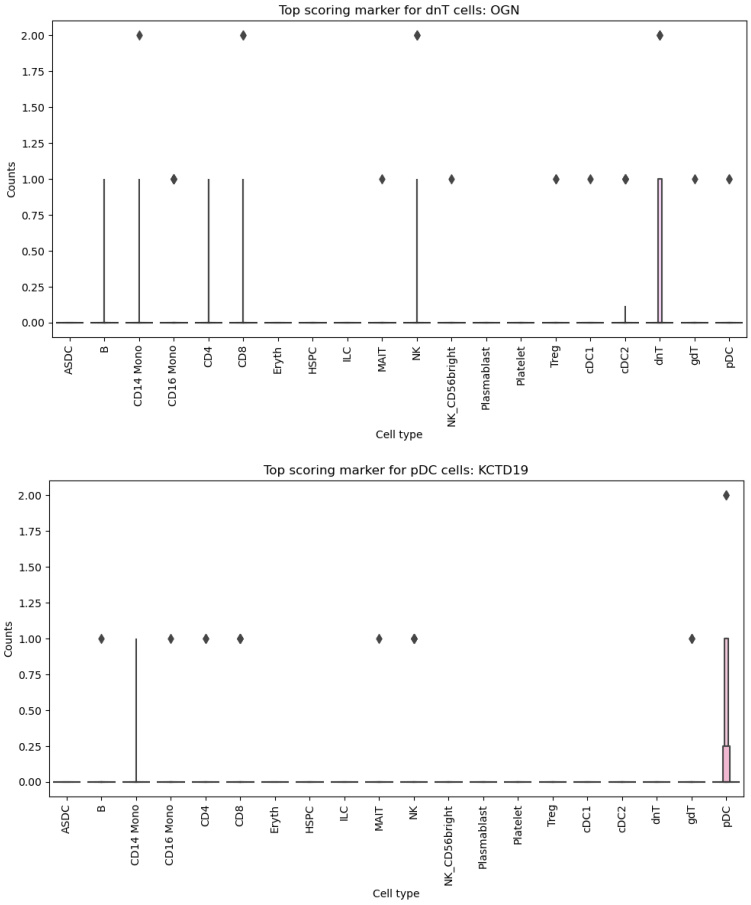

Figure 35: Expression of *OGN* (top), and *KCTD19* (bottom), the highest scoring marker genes for dnT and pDC cells respectively according to SCherlock evaluation on Hao et al. (2021)[180].

3.2 Spattern algorithm - spatial pattern classification

Contribution: I conceived and implemented the code of Spattern and all benchmarks.

3.2.1 Introduction

The study of spatial patterns of gene expression is a natural extension of the spread of spatial transcriptomic technologies. Recent methodological efforts have focused on the identification of spatially variable genes (SVG), i.e genes with a spatial distribution that isn't uniformly random[183–187]. Organisation of spatial patterns has mostly been limited to clustering of the SVGs identified[184,186], or factor analysis[188].

This work centres around recognition of several archetypal spatial pattern classes, predefined based on statistical properties of the gene expression data and resembling specific arrangements of biological interest, uncovering underlying molecular structures and cell niches, in the service of ITH characterisation. To the best of my knowledge, no published work to date has attempted this kind of pattern classification. For now, the classifier is limited to simulated data from a custom spatial data generator also presented here.

Though the focus of this work is on gene expression, the same approach can be extended to any kind of spatial signal.

3.2.2 Methods

3.2.2.1 Spatial data generator

The spatially data generator uses a hexagonal grid in the style of Visium. Each of the classes representing a distinct biological situation corresponds to a distinct underlying model distribution, attempting to capture its essential properties in a simplified way. Spatial expression profiles are simulated individually by gene. For a given pattern class (uniform, gradient, structure, or barrier), the underlying probability density of the eventual expression profile is first randomly generated. Then, each transcript is randomly assigned to a pixel in the grid in accordance with this underlying probability.

Inputs:

- Number of reads N
- Underlying pattern class: uniform | point-source gradient | line-source gradient | | structure | barrier
- Grid dimensions $p \times m$ (default: 30 x 30; 900 pixels)

Output:

- A count matrix for the $p \times m$ pixels; the counts sum up to N

3.2.2.1.1 Uniform

The uniform case represents the situation where there is no underlying structural pattern to gene expression, and the reads are distributed in space in a uniformly random fashion. The underlying density function takes a constant value across all pixels.

3.2.2.1.2 Gradient

The gradient represents a gradually polarised gene distribution in space. The generator includes both point- and line-source gradient input options.

The line-source gradient shape is defined by parallel contour lines. At the first step, the line angle is chosen uniformly at random. The density at a given point is set to the value of intercept of its respective contour line. Then, a constant is added so that the least dense point takes the value of one before scaling to obtain a true density function.

The point-source gradient shape is defined by concentric contour lines. At the first step, the point is chosen uniformly at random in the image space, along with the sign. Density at a given point is set to the distance to the source point (multiplied with the sign). A constant is again added so that the least dense point takes the value of one before scaling to obtain a true density function.

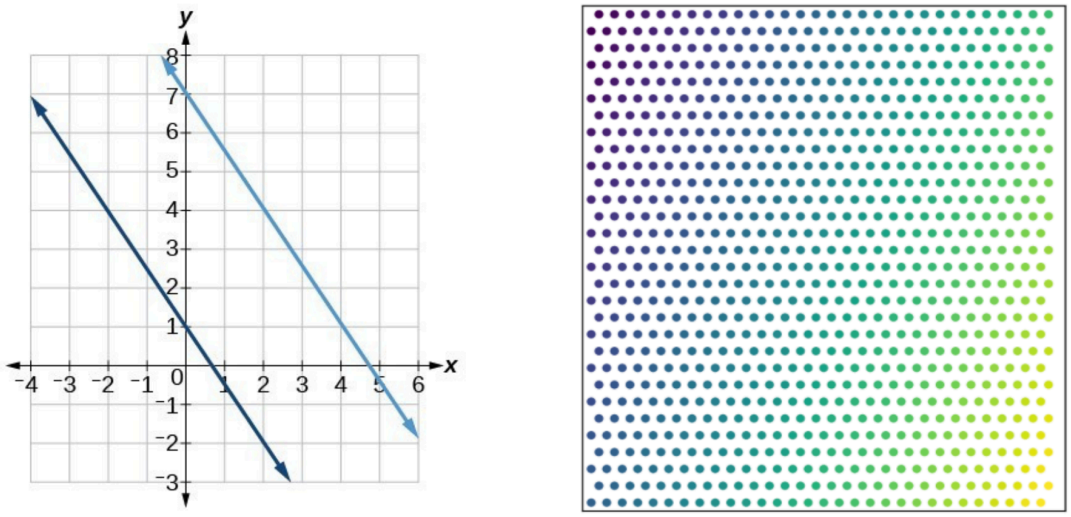

Figure 36: Line-source gradient underlying distribution contour lines (left) and density (right) example. Source: OpenStax[189] (left).

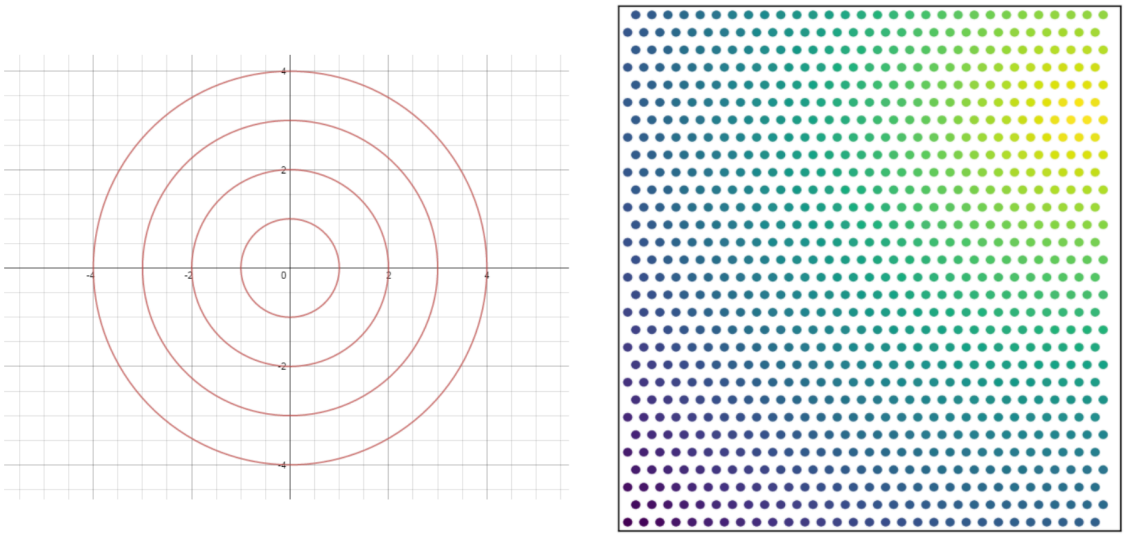

Figure 37: Point-source gradient underlying distribution contour lines (left) and density (right) example. Source: Mathematics[190] (left).

3.2.2.1.3 Structure

This case is in line with a molecular structure specific gene expression pattern. The structure in question is represented by an elliptical shape in the underlying probability distribution.

The first focal point chosen uniformly at random in image space. The second focal point is chosen in relation to the first one, uniformly at random at $(-5, -2] \cup [2, 5)$ for each coordinate. The sum of distances to the focal point is uniformly at random $[1, 1.5)$ times the distance between them. Probability density within the structure is 20 time higher than outside the structure.

3.2.2.1.4 Barrier

This is an equivalent of the structure scenario, with the gene expression being associated with the barrier of a structure rather than its full space. it is represented by the space in the proximity of the underlying distribution defining ellipse.

The first focal point chosen uniformly at random in image space. The second focal point is chosen in relation to the first one, uniformly at random at $(-7, -2] \cup [2, 7)$ for each coordinate. The sum of distances to focal point is uniformly at random $[1, 1.5)$ times the distance between them increased by 1. Probability density of any point with a distance to the ellipse lesser than 0.75 is 20 time higher than any point falling outside this scope.

3.2.2.1.5 Examples of generated data

Figure 39 shows examples of the final randomly generated expression samples for different categories. Further generated examples for varying values of N are collected at the end of the section.

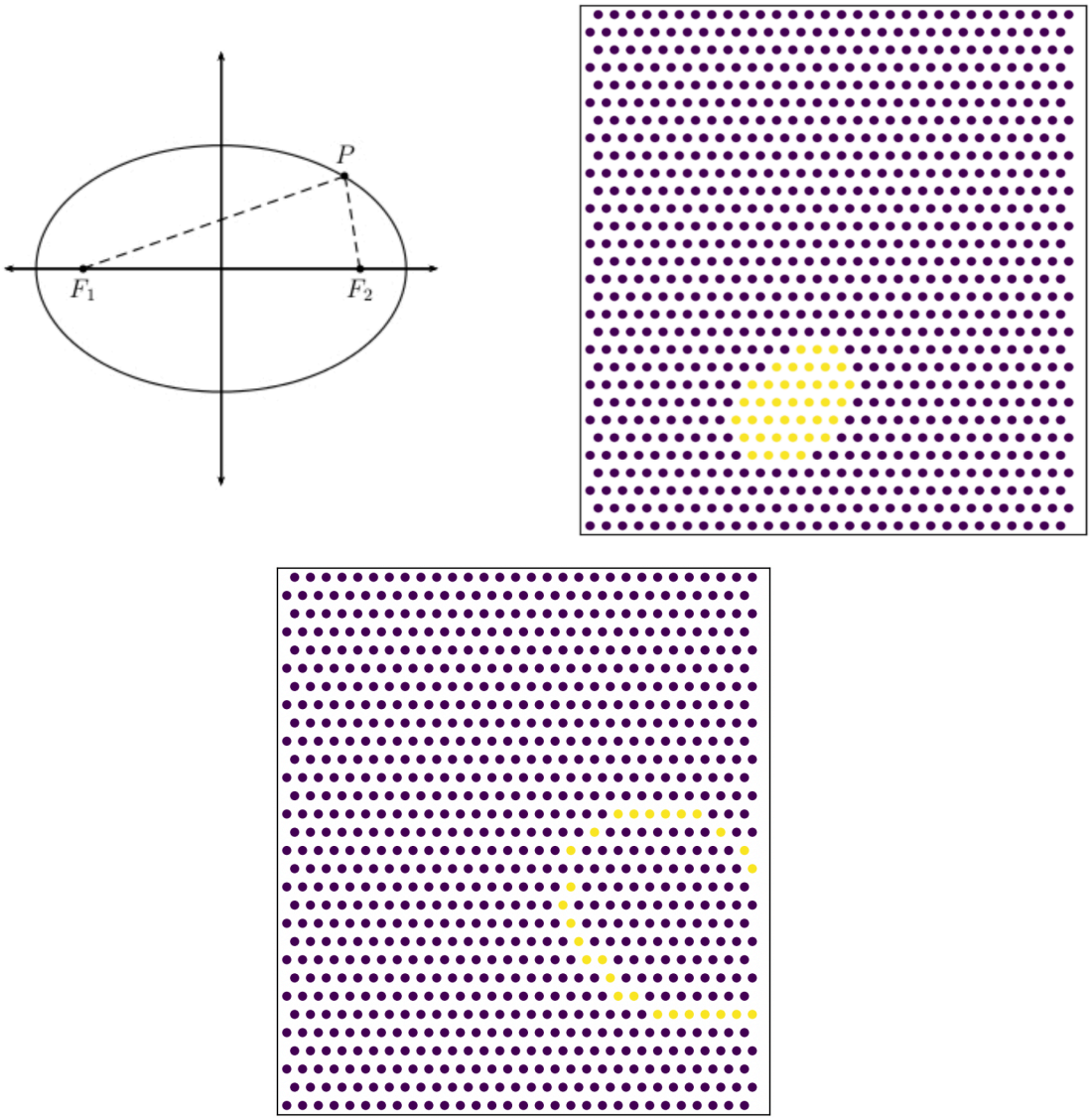

Figure 38: Ellipse defining elements (top left) and corresponding distribution density example for a structure (top right) and a barrier (bottom). Source: calculator.net[191] (top left).

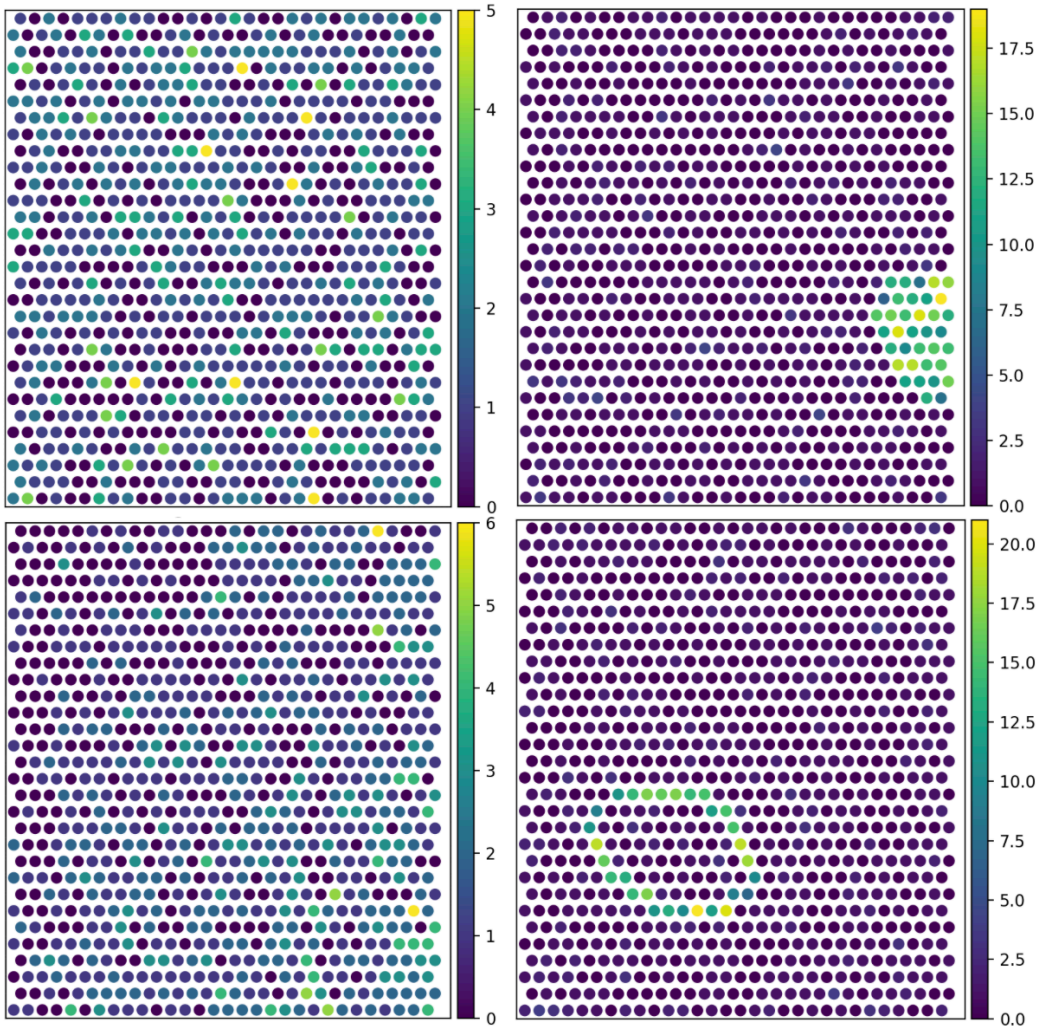

Figure 39: Simulated expression examples with $N = 1000$ for uniform (top left), line-source gradient (bottom left), structure (top right), and barrier (bottom right).

3.2.2.2 Spatial pattern classifier

A classifier was trained on generated data in order to classify different types of spatial patterns. It was based on two types of graph-derived features.

3.2.2.2.1 Feature type 1: Number of large neighbours

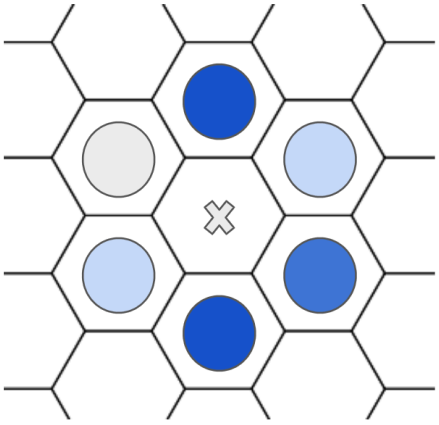

Figure 40: Pixel neighbourhood.

We consider expression of gene G in the neighbourhood of a given pixel. How many neighbours are in the top quantile q according to its expression, compared to the entire space? The answer is an integer between 0 and 6 for any given pixel, and looking at the overall answers across all pixels gives a discrete distribution.

$f_i^q = \#$ of pixels with exactly i neighbours with expression of gene G greater than q

If the distribution is random, pixels with extreme values of expression won't be often found together; in the opposite case we expect areas with concentration of such pixels, with the shape affecting proportions between values.

The classifier uses $f_2^{0.5}$, $f_3^{0.5}$, $f_4^{0.5}$, $f_5^{0.5}$, $f_6^{0.5}$; $f_2^{0.95}$, $f_3^{0.95}$, $f_4^{0.95}$, $f_5^{0.95}$, and $f_6^{0.95}$ as input.

3.2.2.2.2 Feature type 2: Expression-based random walks

Given a starting pixel, we move to one of its neighbours at random at each step, with probabilities of each pixel proportional to its expression of gene G.

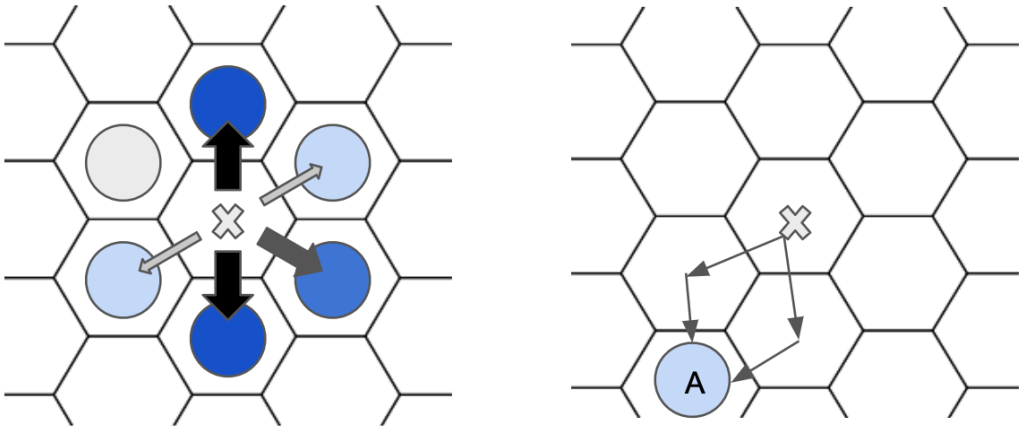

Figure 41: Schematic of a step of an expression-based random walk (left); all paths length two leading to pixel A from pixel X (right).

We look at paths of fixed length l. $P(A|X)$ is the probability that we end up in pixel in A, if we started in X. For each pixel, we compute the sum of probabilities over all possible starting pixels: $\sum_X P(A|X)$.

g_l^q = quantile q of the distribution of $\sum_X P(A|X)$ across all pixels A and paths of length l

The idea behind such features is that in more structured data, we will end up in the same positions more frequently, whereas if data is more diffuse, we will often take many different paths.

The classifier uses $g_2^{0.9}$, $g_2^{0.95}$, $g_2^{0.99}$; $g_3^{0.9}$, $g_3^{0.95}$, $g_3^{0.99}$; $g_4^{0.9}$, $g_4^{0.95}$, and $g_4^{0.99}$ as input.

3.2.2.2.3 Decision tree classifier

A simple decision tree classifier (using the Gini index for the quality of the split) was trained on a total of 19 graph-derived features of the two kinds, with z-score transformation applied before use.

The training and the test set contained 4000 generated samples each: 1000 uniform, 1000 gradient (500 line- and 500 point-source), 1000 structure, and 1000 barrier types.

Classifiers were trained for two levels of transcript abundance: $N = 1000$ and $N = 100$.

3.2.2.3 Results

3.2.2.3.1 Classifier 1 ($N = 1000$ transcripts)

The chosen feature set achieves separation between the four classes (we do not distinguish between the two types of gradients for the purposes of classification.), which form distinct clusters in the UMAP projection (Figure 42).

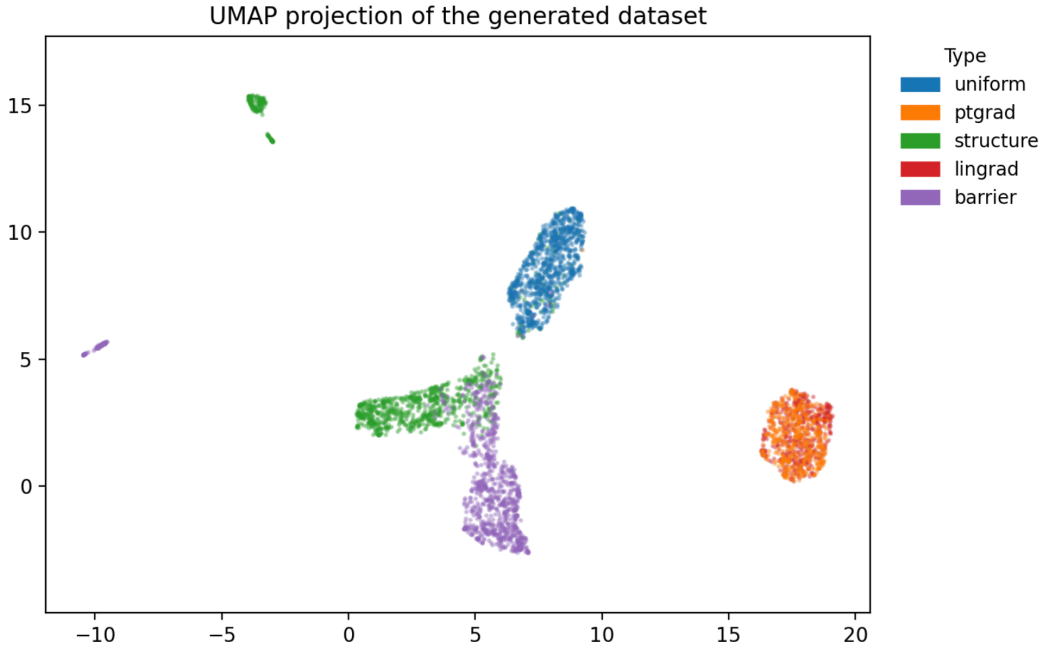

Figure 42: UMAP projection of the training dataset with $N = 1000$ transcripts per sample.

The classifier performance on training data has perfect accuracy of 1 (Table 1), while the accuracy on the test dataset is 0.93 (Table 2).

	uniform	gradient	structure	barrier
uniform	1000	0	0	0
gradient	0	1000	0	0
structure	0	0	1000	0
barrier	0	0	0	1000

Table 1: Training dataset confusion matrix for classifier 1

The gradient class is most consistently well separated from the rest, while the uniform and the barrier class are sometimes confounded with the structure class, but not with each other.

	uniform	gradient	structure	barrier
uniform	935	14	56	2
gradient	6	979	2	0
structure	50	0	867	56
barrier	9	7	75	942

Table 2: Test dataset confusion matrix for classifier 1

3.2.2.3.2 Classifier 2 ($N = 100$ transcripts)

When fewer transcripts are available for a given gene, the separation and clustering of classes based on the chosen feature set are not as clear (Figure 43).

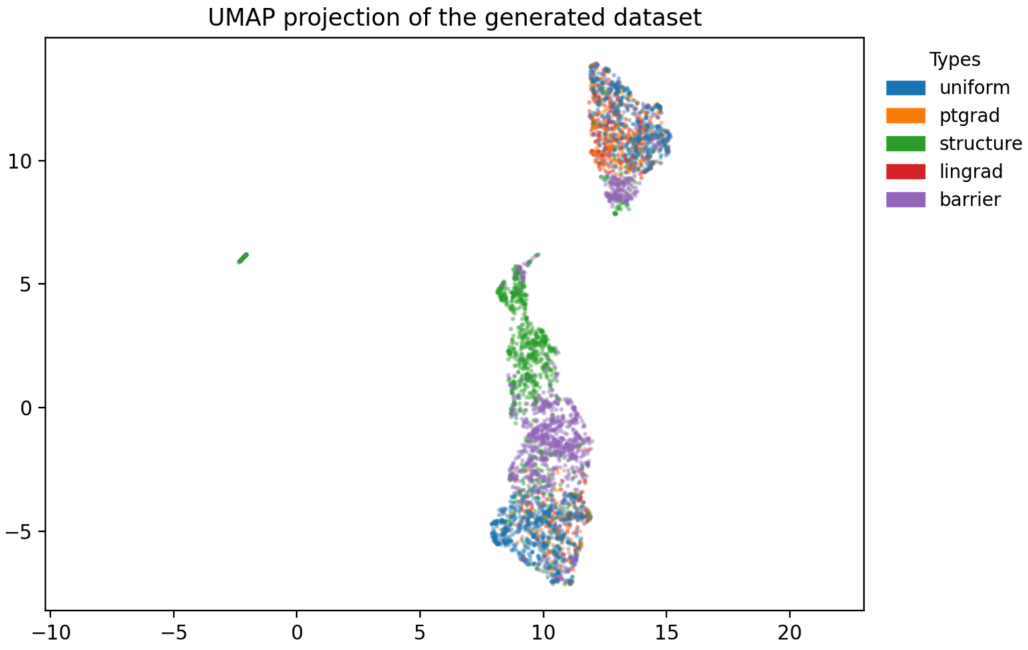

Figure 43: UMAP projection of the training dataset with $N = 100$ transcripts per sample.

The classifier performance on training data has perfect accuracy of 1 (Table 3), while the accuracy on the test dataset is only 0.65 (Table 4).

With fewer transcripts, the gradient, which was previously most consistently correctly predicted, becomes the worst performing class, frequently confounded with the uniform one. Looking at the generated examples with this number of transcripts in Figures 46 and 47, it is difficult to distinguish between the two even by visual inspection. The other two classes perform much better, though

	uniform	gradient	structure	barrier
uniform	1000	0	0	0
gradient	0	1000	0	0
structure	0	0	1000	0
barrier	0	0	0	1000

Table 3: Training dataset confusion matrix for classifier 2

	uniform	gradient	structure	barrier
uniform	595	361	41	29
gradient	328	500	59	72
structure	51	73	755	144
barrier	26	66	145	755

Table 4: Test dataset confusion matrix for classifier 2

they are still confounded significantly more frequently than before (most often with each other).

3.2.3 Discussion

The preliminary classification results on the synthetic data indicate the possibility of identifying specific spatial patterns of biological interest based on simple neighbourhood graph-derived features. Simulated data also provides us with an insight on what is or is not realistically separable at a given number of transcripts, and how many reads might be necessary for a given pattern to be detectable.

Further work is necessary to firstly more closely approximate and ultimately transfer the model to real spatial transcriptomic data, and validate its performance and utility in that context. Some real life examples of relevant expression patterns are gathered in Figures 44 and 45. Figure 44 shows examples of barrier patterns with very sharp spatial definition. While they are not elliptical in shape, they are clear enough that generalisation of Spattern to such cases should not require extensive effort. In fact, considering the feature design and the ability of Spattern to seamlessly mix the two types of gradient in a single category, such a difference in barrier shape is unlikely to play a key role.

The examples shown in Figure 45, on the other hand, are more complex. The pattern of expression of chemokines immediately after infection resembles a point-source gradient, which appears to gradually turn into a kind of closed-end line-source gradient. But separation of such a particular case from a barrier has not yet been considered in the simulated data version of Spattern.

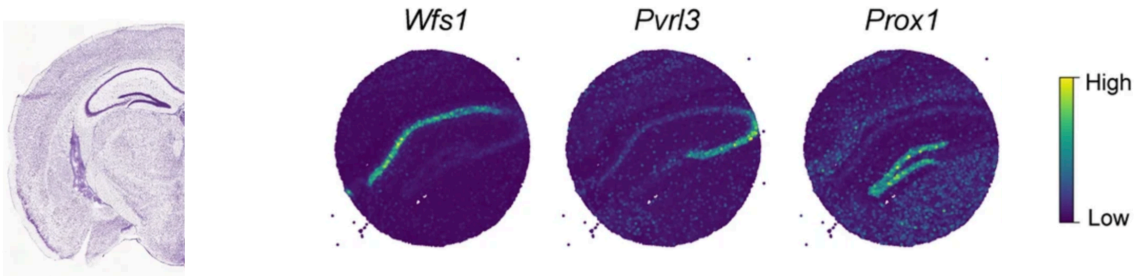

Figure 44: Spatially resolved expression of *WFS1*, *PVRL3*, and *PROX1* in the mouse brain hippocampus. Source: adapted from Zhou et al. (2024)[192].

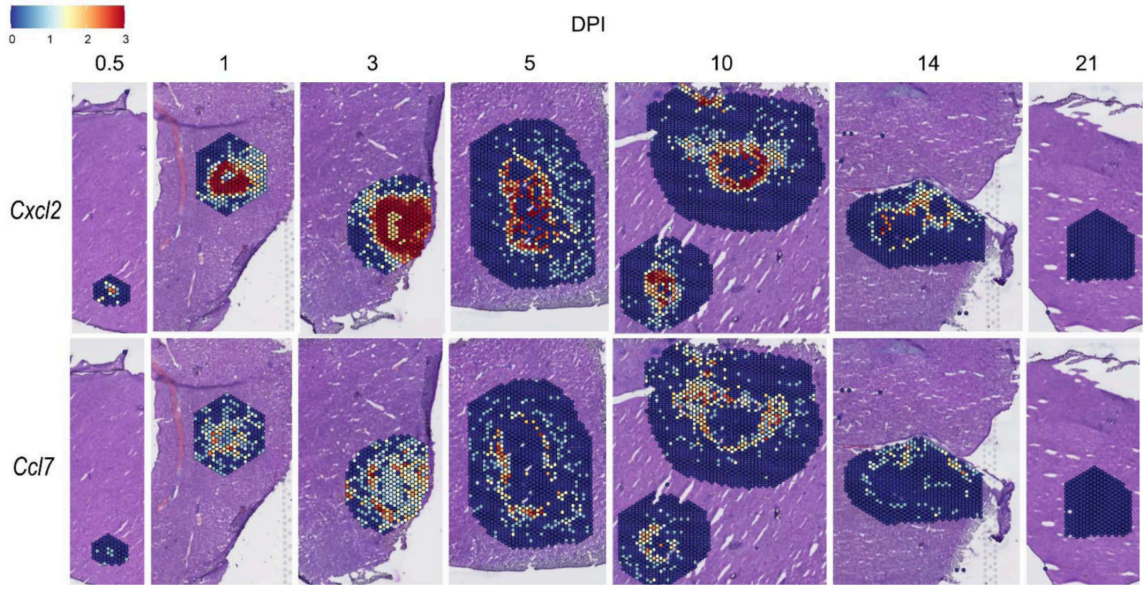

Figure 45: Spatially resolved expression of *CXCL2* and *CCL7* at various days post-infection (DPI) in a liver granuloma response in a mouse infection model. Source: adapted from Amason et al. (2024)[193].

Finally, though a transcriptomic context has been assumed here, the same approach can be considered for any type of spatial signal.

3.2.4 Supplementary results

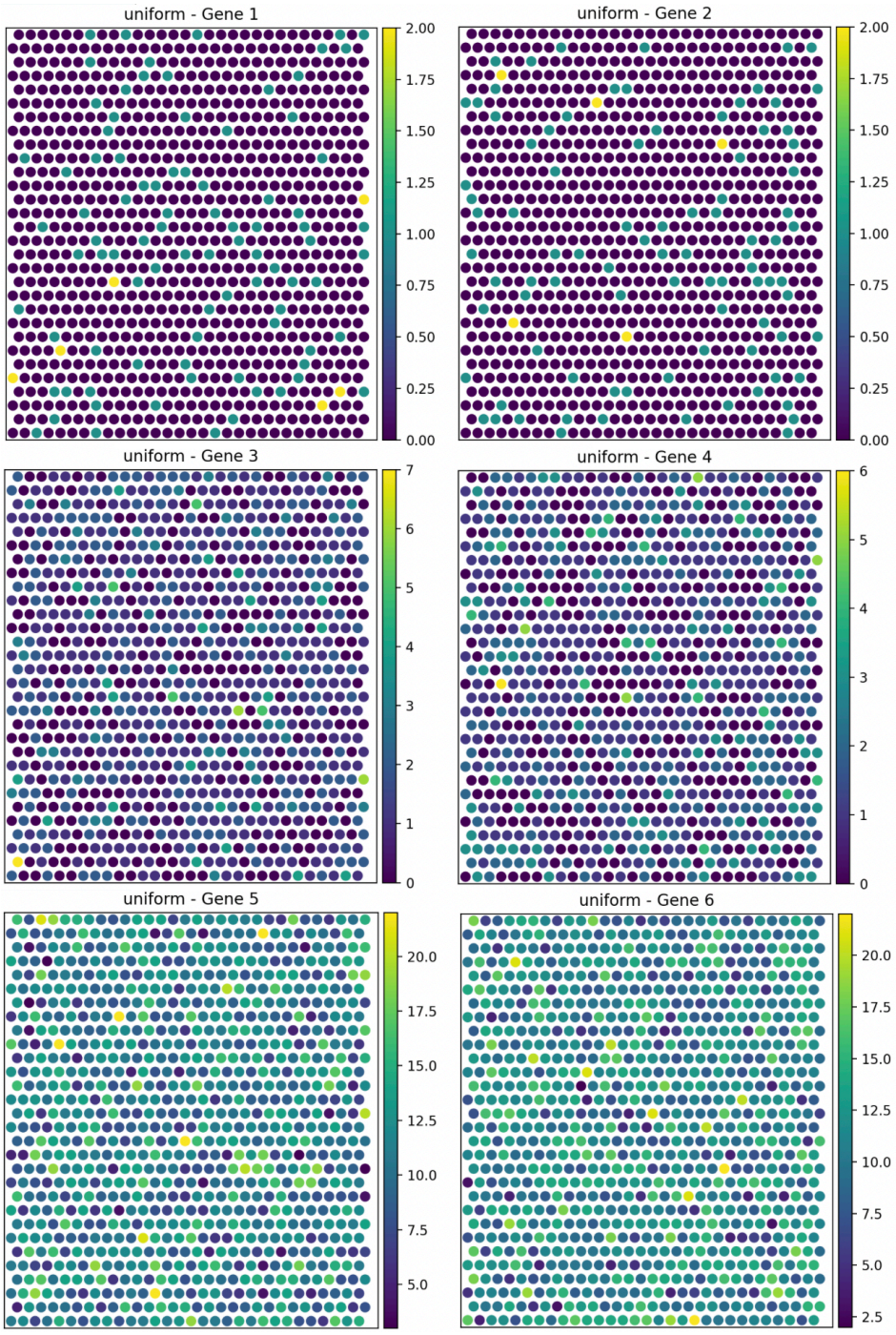

Figure 46: Simulated uniform expression distribution examples for $N = 100$ (top), $N = 1000$ (middle), and $N = 10000$ (bottom).

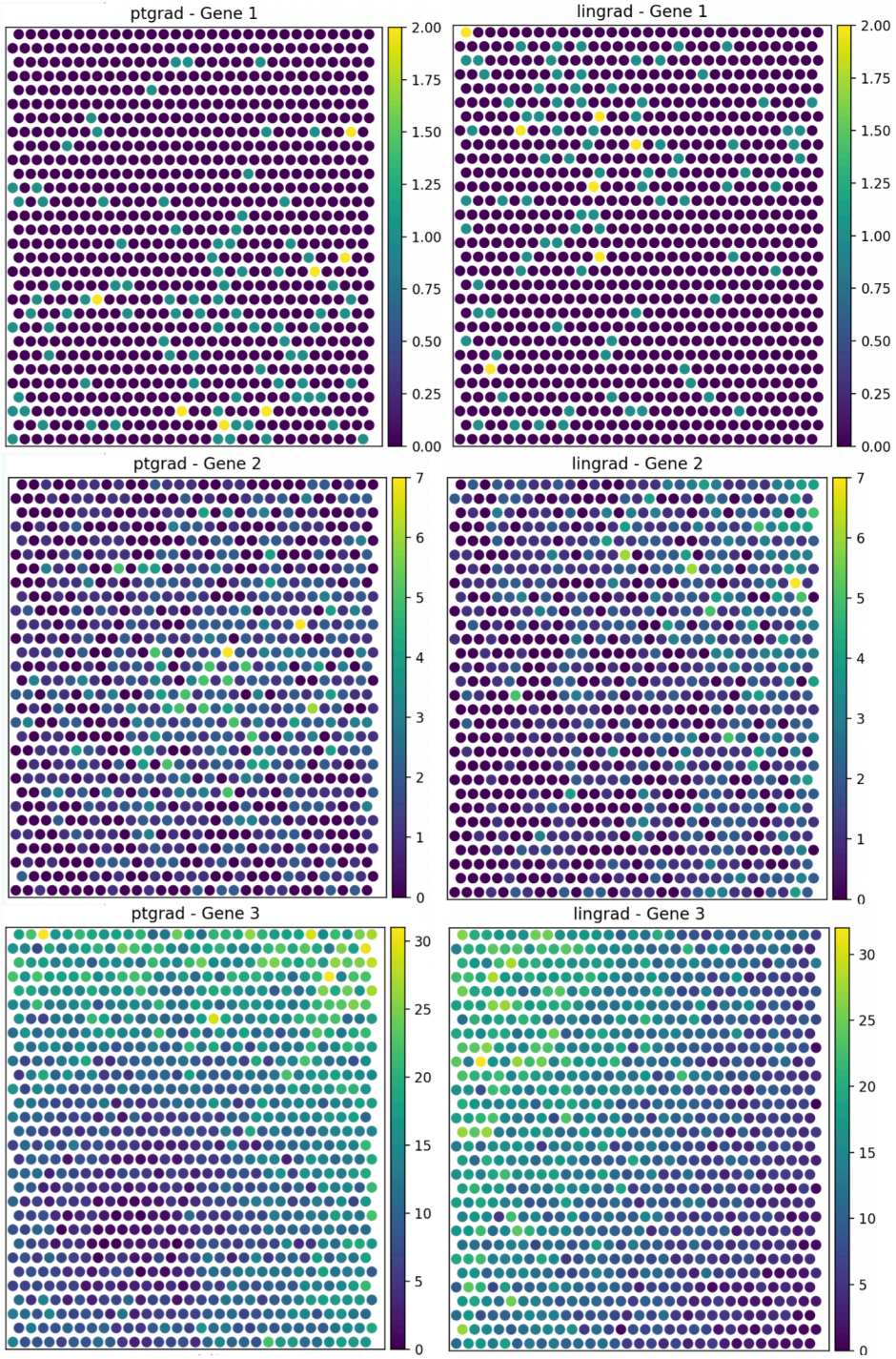

Figure 47: Simulated expression distribution examples of the point-source (left) and point-source (right) gradient patterns for $N = 100$ (top), $N = 1000$ (middle), and $N = 10000$ (bottom).

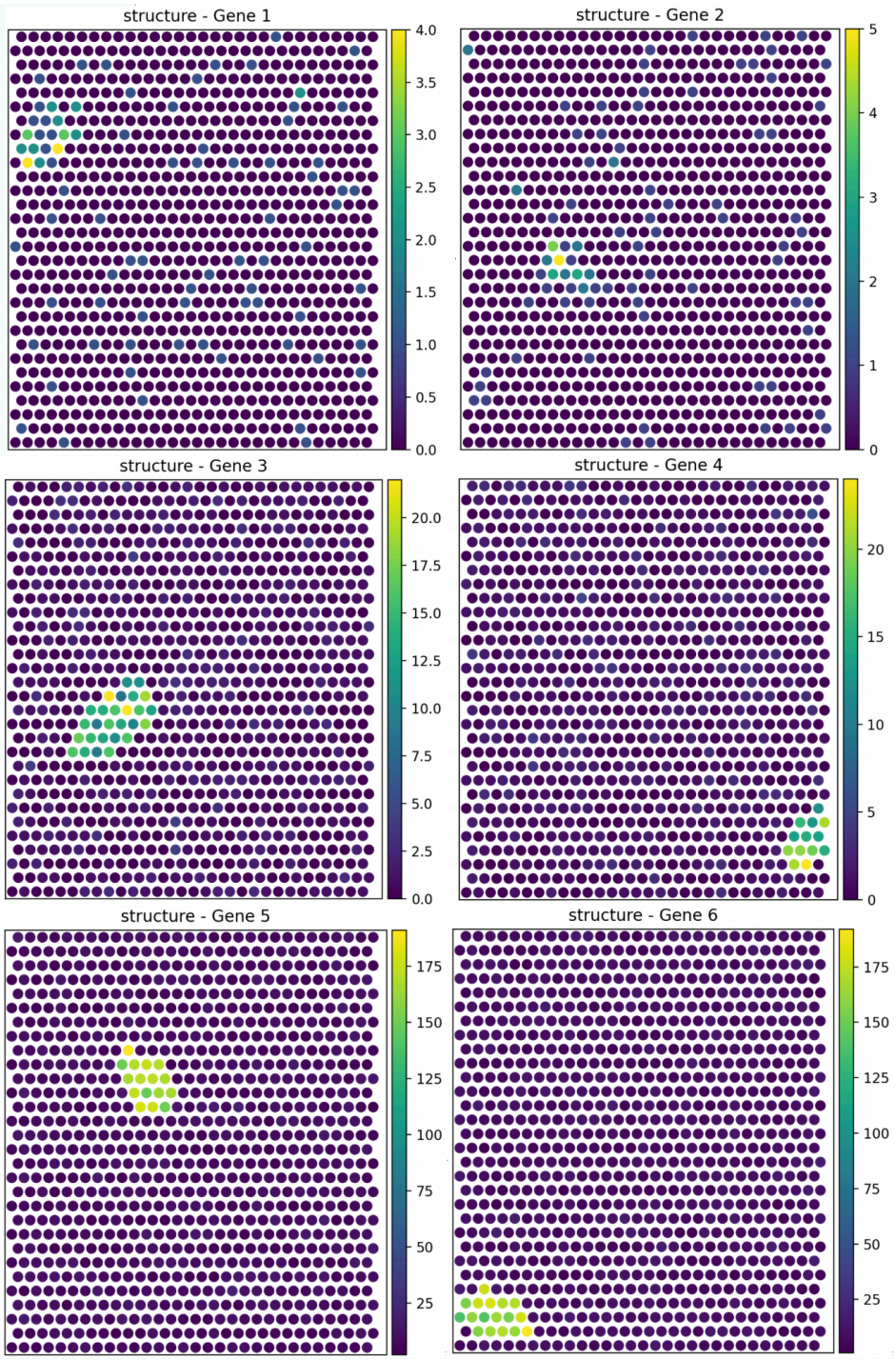

Figure 48: Simulated expression distribution examples of the structure pattern for $N = 100$ (top), $N = 1000$ (middle), and $N = 10000$ (bottom).

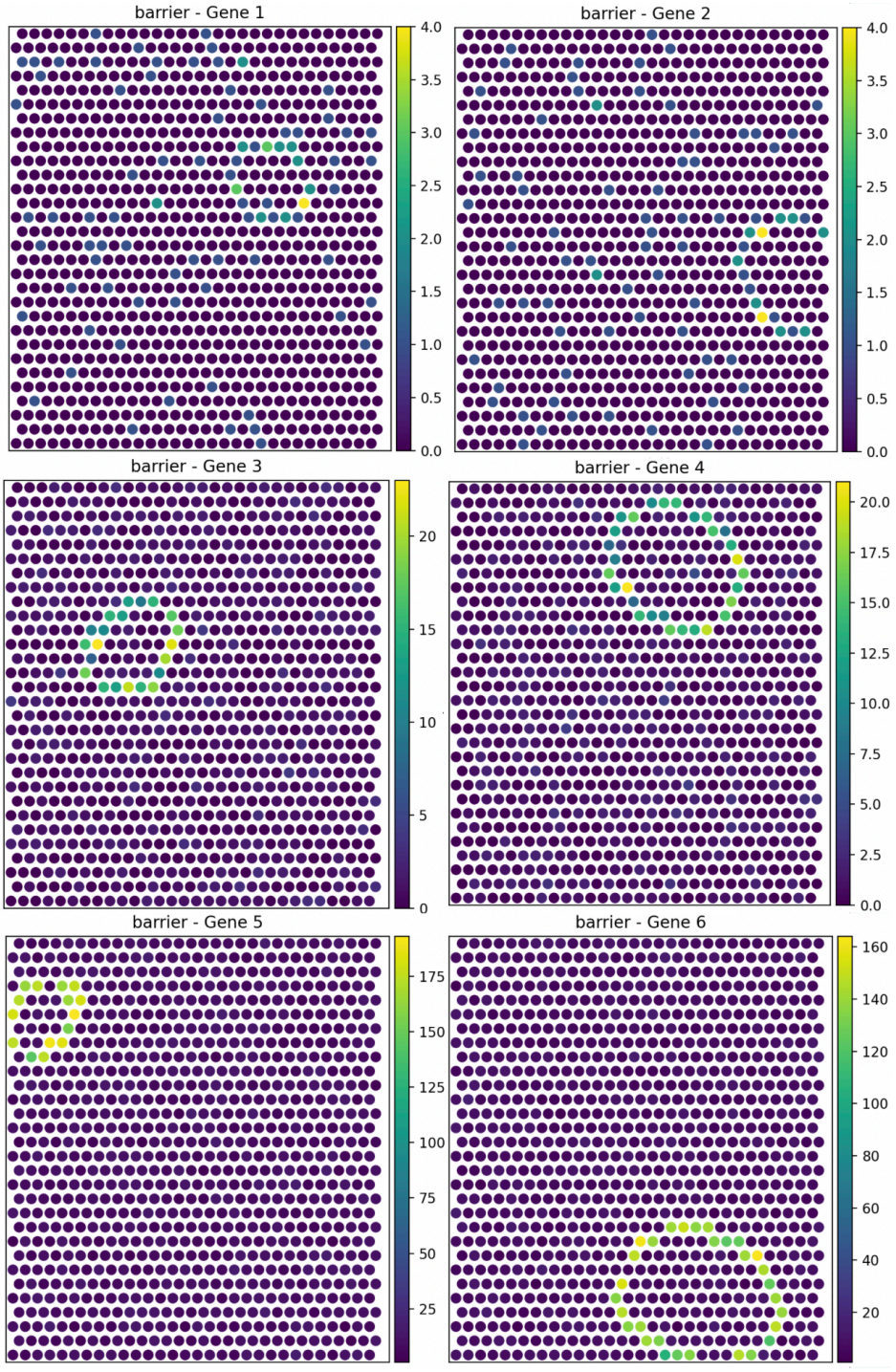

Figure 49: Simulated expression distribution examples of the barrier pattern for $N = 100$ (top), $N = 1000$ (middle), and $N = 10000$ (bottom).

3.3 FastCNV algorithm - transcriptome-based CNV inference

Contribution: I participated in the code review and benchmark comparison to other methods. An article presenting this method is in preparation, which I will co-sign as an author.

3.3.1 Introduction

DNA copy number variations (CNVs) are structural variations of the genome, where copies of genes are either gained via duplication events, or lost via deletion. Like other types of genetic aberrations, CNVs are common in cancer, which is characterised by hallmark genomic instability[7].

Inference of CNVs from transcriptomic data is based on the study of average gene expression within the window frames along the length of the genome, a notable increase in expression suggesting a copy gain, and decrease a copy loss. Comparison against a diploid reference is necessary for an accurate evaluation. The principle has been adapted to single-cell and spatial transcriptomic data[194–197].

Because single-cell and spatial transcriptomic data typically has few reads per cell/spot, there is little signal, and the sensitivity of existing methods is limited. The proposed FastCNV increases both sensitivity and speed of calculation by pooling reads from multiple sources.

3.3.2 Methods

The algorithm can be run on either single-cell or spatial transcriptomic data.

3.3.2.1 Reference assignment

Though the method can operate without it, a reference population will typically be assigned at input. This could be either spots of normal tissue, paired healthy tissue samples, or simply a subset of cell types as a proxy for normal.

3.3.2.2 Amplification step

In order to increase the sensitivity of the method, observations from similar cells or spots are aggregated into meta-cells or meta-spots and treated as one. Cells/spots of the same annotation are first clustered using the Louvain algorithm[198] (default Seurat[199] implementation). Cells are grouped

within each individual cluster only, added together until the threshold (set to 30,000 reads by default) is reached. In the tumour portion, the data of each patent is treated separately, while in the reference portion potential mixing is not an issue.

3.3.2.3 CNV calling

The CNV calling step itself is not original to our method. The same approach that inspired inferCNV[197] is used on the aggregated dataset of meta-cells/spots.

First, mean expression across all cells/spots is computed, and only most highly expressed genes (7000 by default) are considered. The expression of each gene is standardised (truncated z-score between -3 and 3). For each genomic window, all cells/spots falling within the set quantile (0.01 and 0.99 by default) values of the reference set are assigned a zero score. The quantiles of the rest are scaled to fall between -1 and 1, with values of > 0 considered chromosomal gains, and values < 0 considered chromosomal losses. If a reference is provided, the final scores are corrected on the values obtained from that population.

3.3.3 Results

Heatmaps of CNV calls on example colorectal cancer single-cell[200] and spatial transcriptomic[201] data are shown in Figures 50 & 51. As expected, annotation groups that were assigned as reference data at input show very little copy number variation, while tumour cells show the greatest number of copy number gains and losses. In both cases, the variation of calls within the tumour cell portion points towards subclonality.

Visualisation of non-zero CNV call fraction for each cell of the single-cell example on a UMAP (Figure 52) reveals four clusters of tumour cells with different degrees of CNV aberrations, in line the hypothesis of subclonality. Visualisation of the CNV fraction for the spatial example on the Visium slide (Figure 53) shows two distinct areas of the tumour with notably different levels of CNV calls, indicating subclonality with a very clear spatial organisation.

Clustering profiles based on CNV calls (shown in Figures 54 & 55) results in tumour cells/areas forming multiple distinct clusters, which is in line with expectations as well as previous observations regarding subclonality. The tumour clusters in the single-cell example are extremely pure, while there is some mixing in the spatial example, which can be explained by the somewhat lower resolution of this method (not quite single-cell).

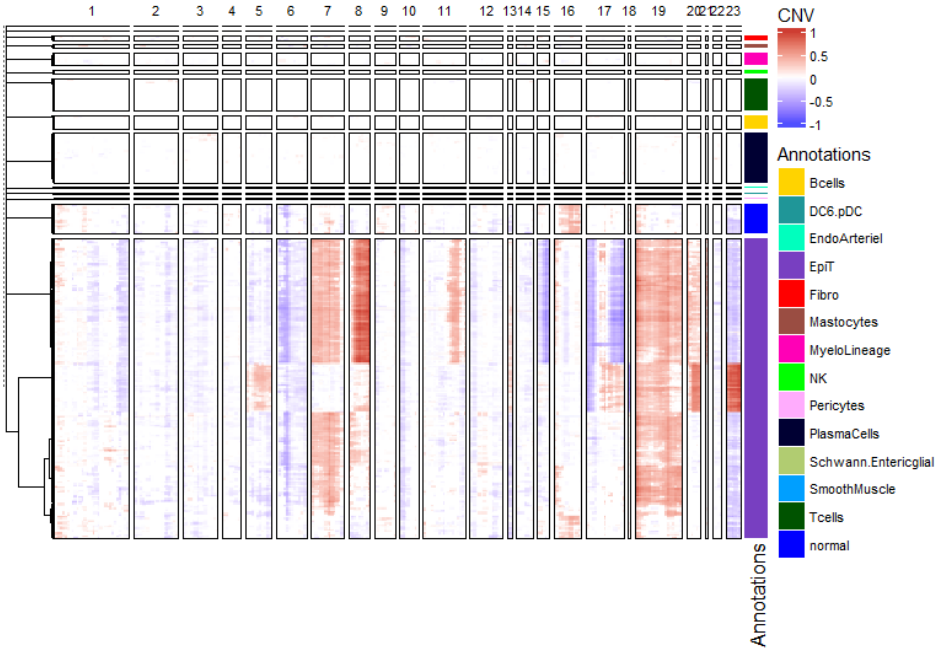

Figure 50: CNV heatmap for a single-cell colorectal cancer example dataset.

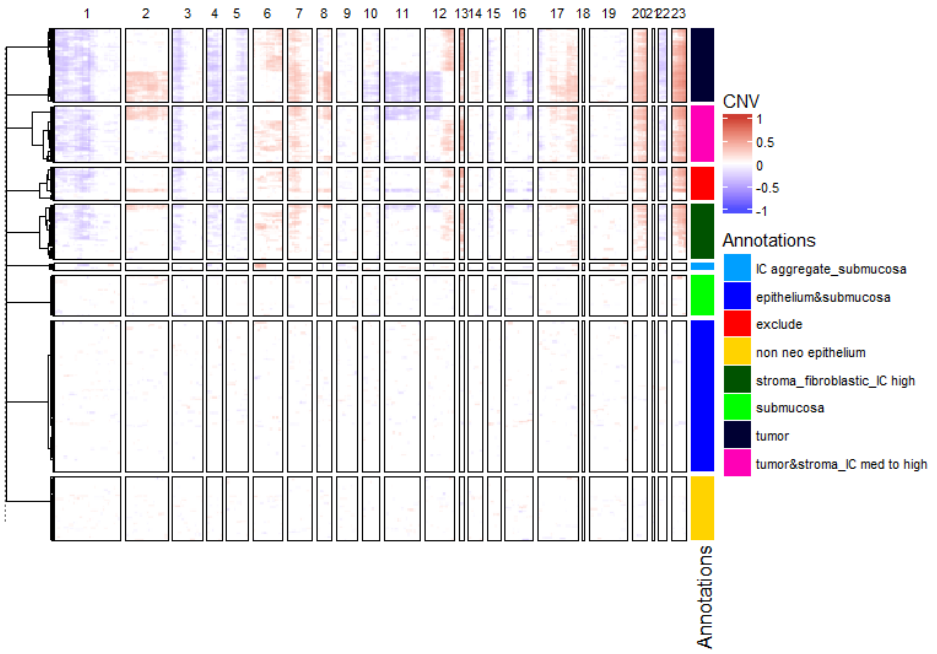

Figure 51: CNV heatmap for a spatial colorectal cancer example slide.

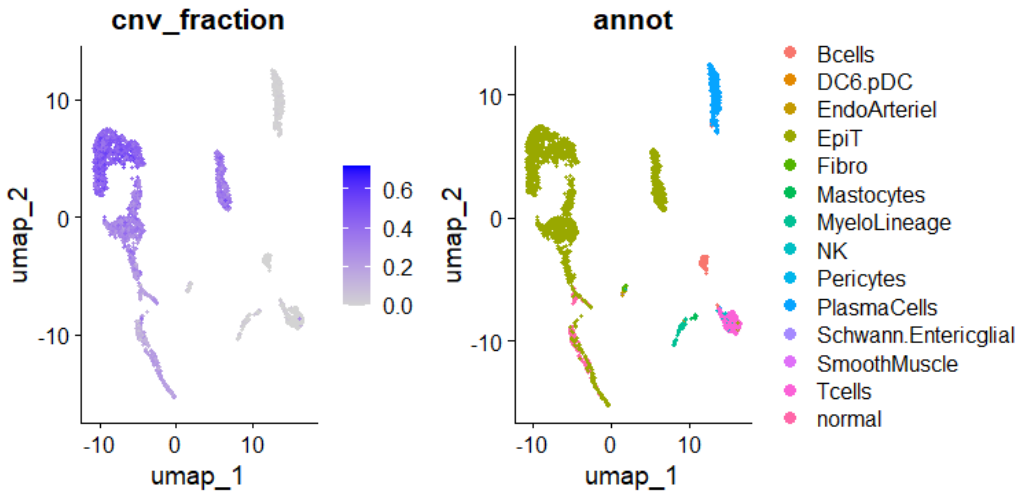

Figure 52: UMAP of the single-cell example dataset with fraction of non-zero CNV calls (left) and cell type annotation (right).

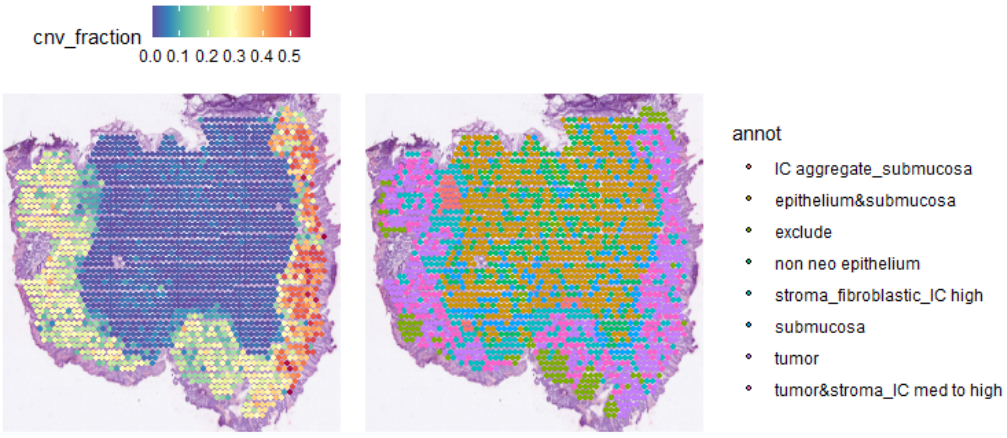

Figure 53: Visium slide of the spatial example with fraction of non-zero CNV calls (left) and spot annotation (right).

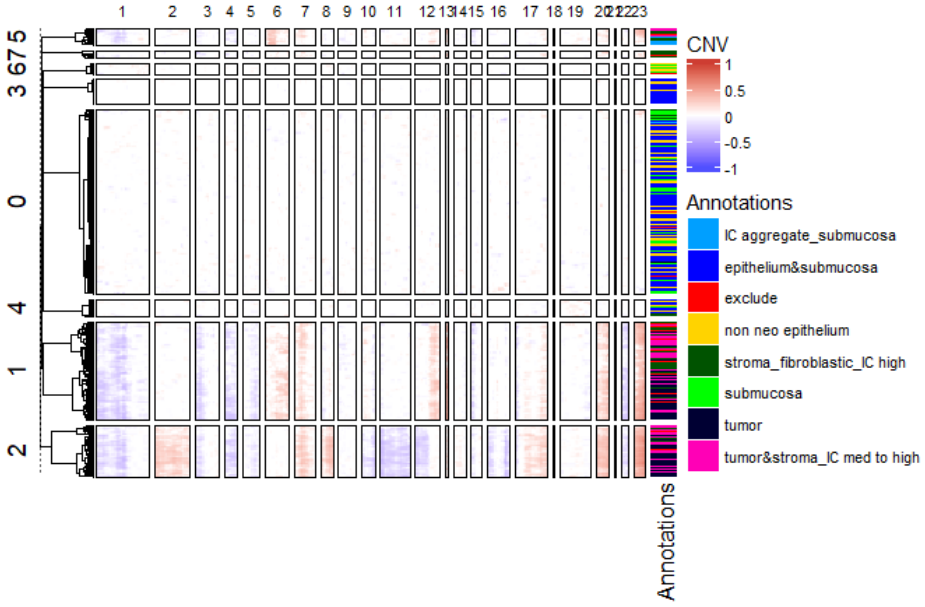

Figure 54: CNV heatmap with CNV calls based clusters for a spatial colorectal cancer example slide.

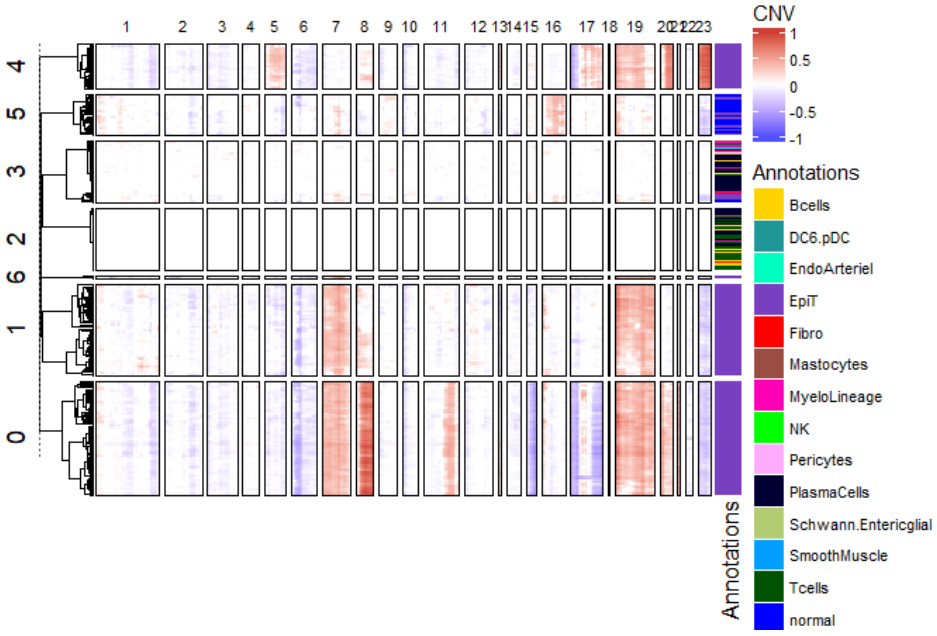

Figure 55: CNV heatmap with CNV calls based clusters for a single-cell colorectal cancer example dataset.

3.3.4 Discussion

While CNV inference from single-cell/spatial transcriptomic data is not new, and the approach of genomic windows utilised here is already established in the literature, the additional amplification step and the introduction of meta-cells or meta-spots proposed in FastCNV increase both sensitivity and speed of CNV analysis of such data.

CNV analysis is most relevant in cancer contexts, where a diverse repertoire of aberrations can be found. CNV based clustering can reveal subclonality of these aberrations, which can be involved in various aspects of tumour development and progression.

3.4 Unified lung cancer cell atlas

Contribution: I conceived and realised all of the analyses involved in building the atlas in their entirety. Though the atlas is currently a draft and the work is ongoing, an article presenting this work is in the early stages of preparation, which I will co-sign as the first author.

3.4.1 Introduction

Several attempts to describe the cells and states of lung cancer tissue and healthy lung tissue comprehensively in cellular atlases exist in the literature. Some of the recent integrative efforts include the integrated Human Lung Cell Atlas (HLCA)[169], a non-small cell lung cancer atlas[170], and a small-cell lung cancer atlas[171], all described in more detail in Chapter 1.

So far the all efforts have been focused on either NSCLC (or one of its subhistologies) or SCLC exclusively, with the other category serving as a small contextual background at most.

Here, a unified lung cancer atlas is proposed, based on published single-cell data sources. All major lung cancer histologies are included in a balanced fashion. Healthy tissue samples are also included. Cell type annotations have been revisited in the context of marker analysis performed with the SCherlock algorithm.

3.4.2 Methods

3.4.2.1 Atlas composition

The lung cancer data was obtained from existing published (meta) sources: the integrated NSCLC and SCLC atlas[170,171]. The initial goal was to roughly balance the three major histologies (LUAD, LUSC and SCLC) in the composition of the unified atlas.

3.4.2.1.1 SCLC data

As the availability of single-cell SCLC data is much more limited compared to NSCLC one, all SCLC samples except pleural effusion and RU1108 (quality control issues) from the SCLC atlas[170] were included. The data used consists of 19x SCLC samples (60091 cells) from 17 patients.

3.4.2.1.2 LUAD and LUSC data

The availability of LUAD data is much greater than that of LUSC (see Figure 56). In order to balance the SCLC data as well as possible, as well as make good use of datasets with the most comprehensive annotations, data from three source studies was selected as an initial starting point.

All samples from Goveia et al. (2020)[202] were included: 3x LUAD [8578 cells], 4x LUSC [12608 cells], as well as 6 x paired tumour adjacent tissue [15182 cells].

Two unclassified NSCLC samples (P11, P42) from Wu et al. (2021)[203] were discarded. The data used consists of 18x LUAD [16739 cells] and 22x LUSC [41120 cells].

The Laughney et al. (2020)[204] datasets consists of 13x LUAD [16977 cells], along with 4x tumour adjacent tissue [9426 cells], and was used in entirety. It includes mutational profile annotations and was partially also used in the SCLC dataset as the NSCLC background reference source.

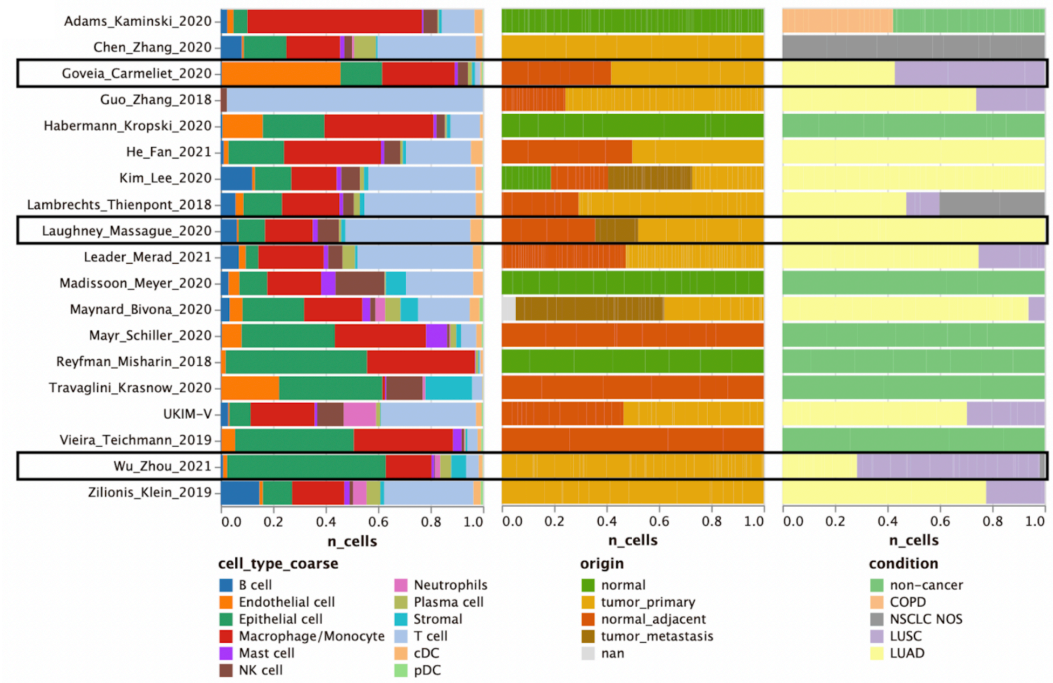

Figure 56: Breakdown of the NSCLC atlas[170] datasets. The unified lung cell atlas includes data sourced from three studies: Goveia et al. (2020)[202], Laughney et al. (2020)[204], and Wu et al. (2021)[203]; as well as additional indivdual samples from other sources. Source: adapted from Salcher et al. (2022)[170].

3.4.2.1.3 Neuroendocrine NSCLC data

Due to an interest in LCNEC and a lack of publicly available single-cell datasets at the time of work, an analysis of neuroendocrine marker expression (CHGA, SYP, CD56) in NSCLC samples was performed. A number of samples with at least partial expression of at least one of the markers were identified. A handful of most significant expressors were chosen for inclusion in the atlas.

Samples LX680 from Laughney et al. (2020)[204] and P10 from Wu et al. (2021)[203] were already included in the dataset.

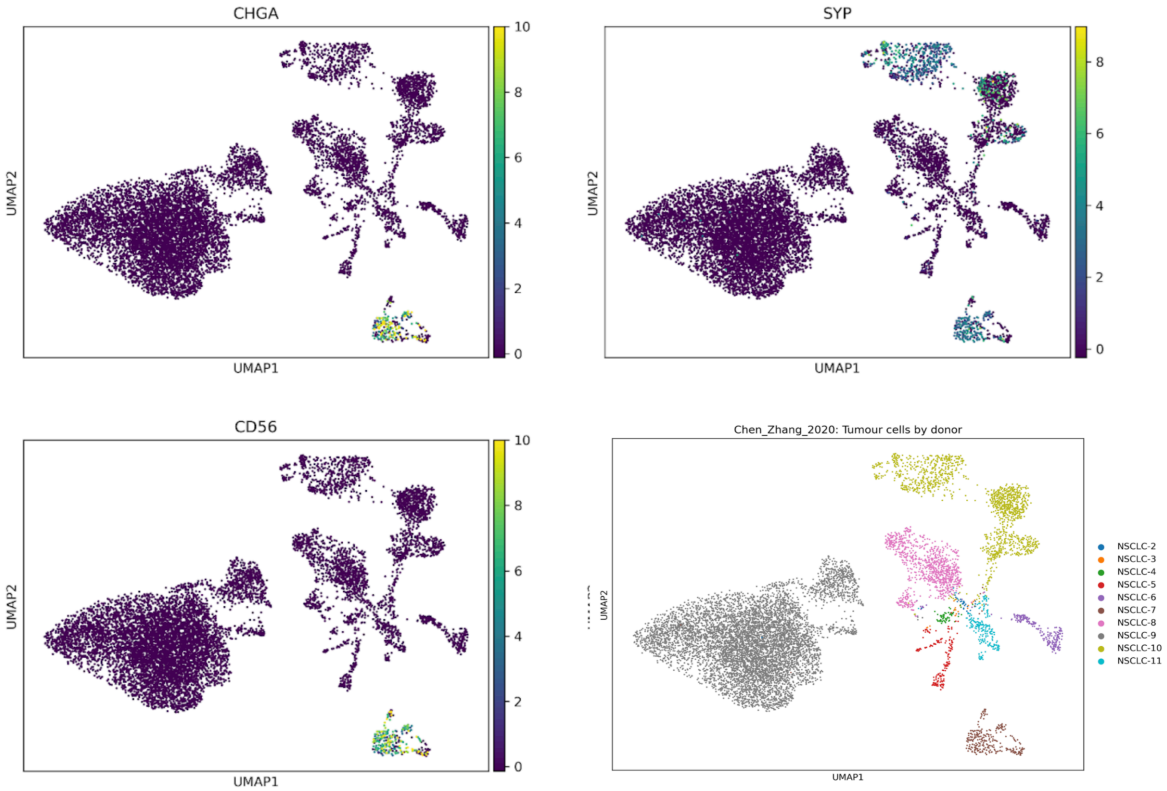

Figure 57: UMAP of the Chen et al. (2020)[205] dataset with *CHGA* (top left), *SYP* (top right), *CD56* (bottom left) expression and sample annotation (bottom right). NSCLC-7 expresses all three neuroendocrine markers, while NSCLC-10 expresses *SYP* in all three subclusters, but to different degrees.

Six further samples were added into the atlas: NSCLC-7 and NSCLC-10 from Chen et al. (2020)[205] (see Figure 57), 184CO and 192C from Adams et al. (2020)[206], TH155 from Maynard et al. (2020)[207], and P12 from Salecher et al. (2022)[170] dataset.

3.4.2.1.4 Dataset split

The complete atlas consists of 223,619 cells. In order to validate the robustness of results obtained, it is split into two halves, randomly across each sample: the first half containing 111,809 cells, and the second containing 111,810 cells.

3.4.2.2 Atlas analysis

The atlas was first integrated and clustered as a whole. After a coarse level split was identified, the process was repeated at the level of each of the subsets individually in an isolated manner. In some cases, a sample subset of HLCA[169] was integrated into the subset as an additional projection onto healthy data with detailed annotation.

3.4.2.2.1 Data integration

Data integration at every step was performed with an autoencoder-based method scVI[208]. A set of 2000 top highly variable genes was identified each time, with sample donor serving as the batch. In the case of subsets, a limited number of cases with three or fewer cells per donor were first filtered out. The parameters from the scVI tutorial were replicated to set up the model each time: two hidden layers and thirty dimensions in the latent space. Leiden clustering was performed on the integrated data space.

3.4.2.2.2 CNV analysis

Copy number variation analysis was performed with the FastCNV algorithm described in the previous chapter. The myeloid and lymphoid cells were used as reference data for the CNV analysis. Genomic windows of size 250 were used.

3.4.2.2.3 Marker analysis

Marker analysis was performed with the SCherlock algorithm, described in the first section of this chapter. The markers were computed with the 'diff' scoring approach, multiplication factor values of $k = [1, 10, 25]$ and mean aggregation approach. Gene filtering parameters of minimum 3 patients and 10 reads per candidate marker for the cell type in question, and maximum 100 genes retained were applied before empirical scoring.

3.4.2.2.4 Reannotation

The clusters obtained by clustering integrated data have been analysed with regard to annotations from original sources. Further exploration with canonical markers has been performed for verification and to address inconsistencies. The final annotations were assigned by cluster. The tumour cell subtypes come from source pathological annotations. The level of granularity was determined iteratively and concurrently with the marker analysis.

3.4.3 Results

The results shown in here are largely based on the first half of the dataset. The equivalent figures for the second half of the dataset are contained in the supplementary result at the end of the section.

The dataset has been integrated with the scVI[208] algorithm, with batch normalisation taking place at the level of donors. Leiden clustering has been performed on the integrated data (see Figure 58). The majority of the resulting clusters span data of several donors from several different datasets, though a couple are limited to cells of a single donor (unsurprisingly, further inspection reveals these to be individual tumour cases).

We first look at the split into the four lineage groups according to the annotations of each of the two sources (Figure 59). The two sources are consistent to a high degree, and each of the leiden clusters can be assigned to one of the four lineages according to the majority type (and further confirmation via canonical markers).

The resulting epithelial subset contains 45914 cells, the lymphoid subset 31022, the myeloid subset 23091 and the mesenchymal subset 11782. We proceed to reintegrate and study each of them in isolation.

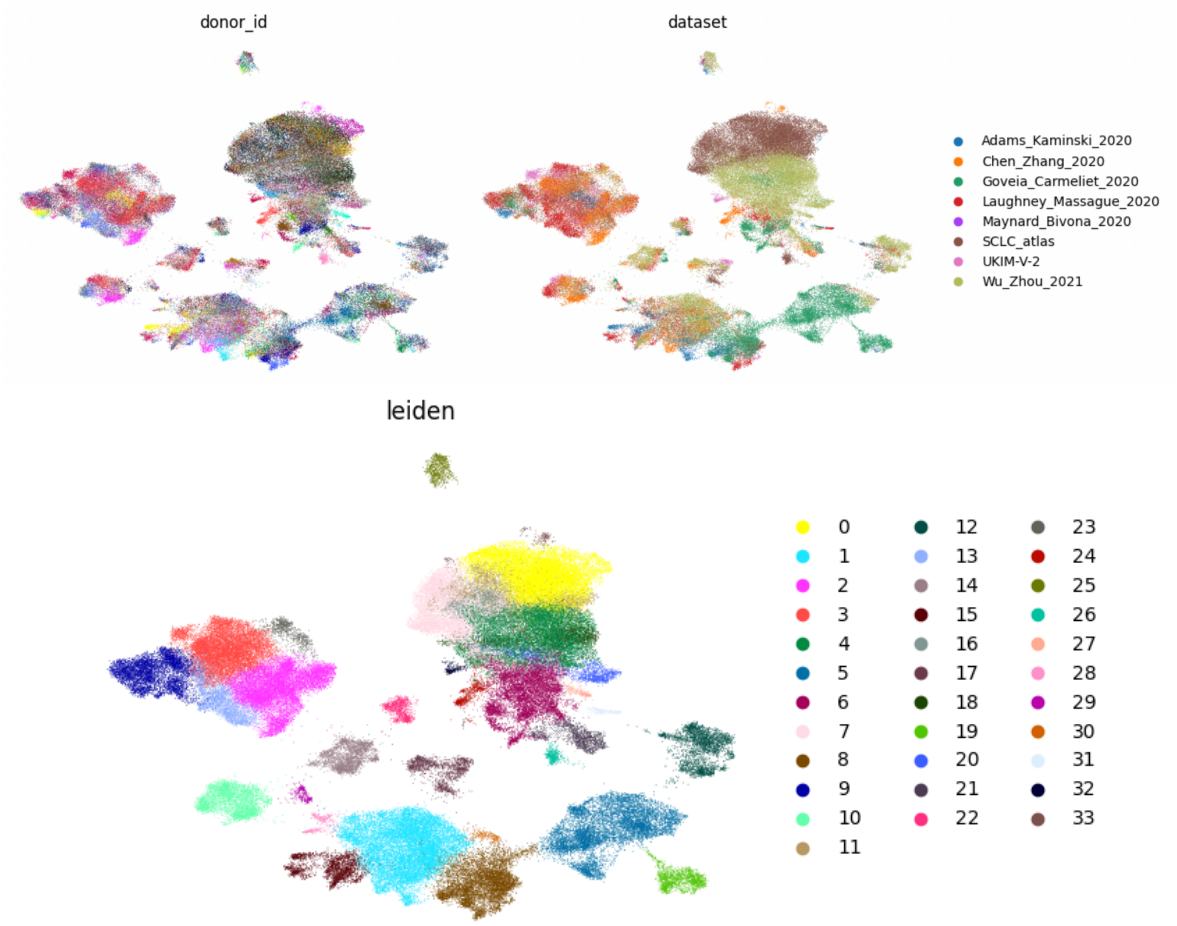

Figure 58: UMAPs of the first half of the dataset post scVI integration annotated by donor (top left), source dataset (top right) and leiden cluster (bottom).

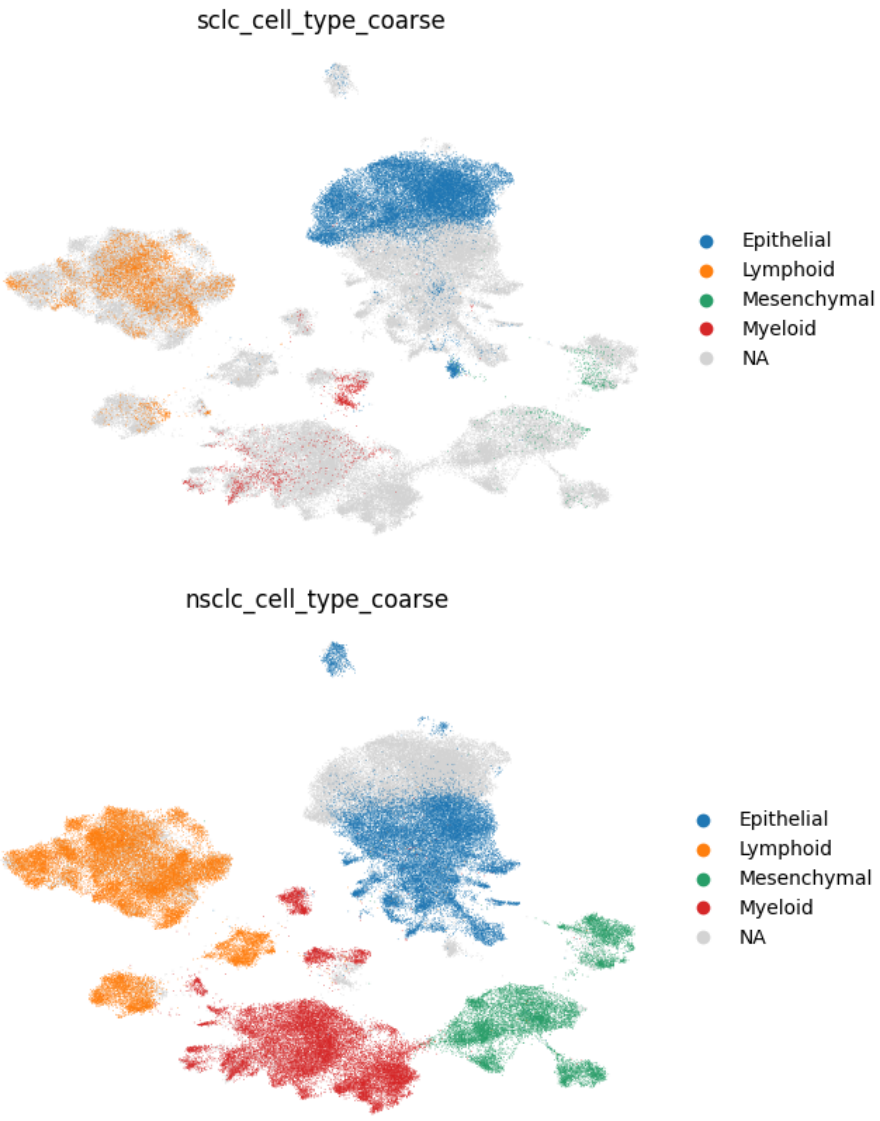

Figure 59: UMAPs of the first half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom).

3.4.3.1 The mesenchymal subset

The mesenchymal subset is the smallest one. Figure 60 shows the integrated dataset annotated according to both sources. The annotations from the NSCLC atlas are much more granular than the ones from the SCLC one, yet they are largely related to the anatomical location rather than function.

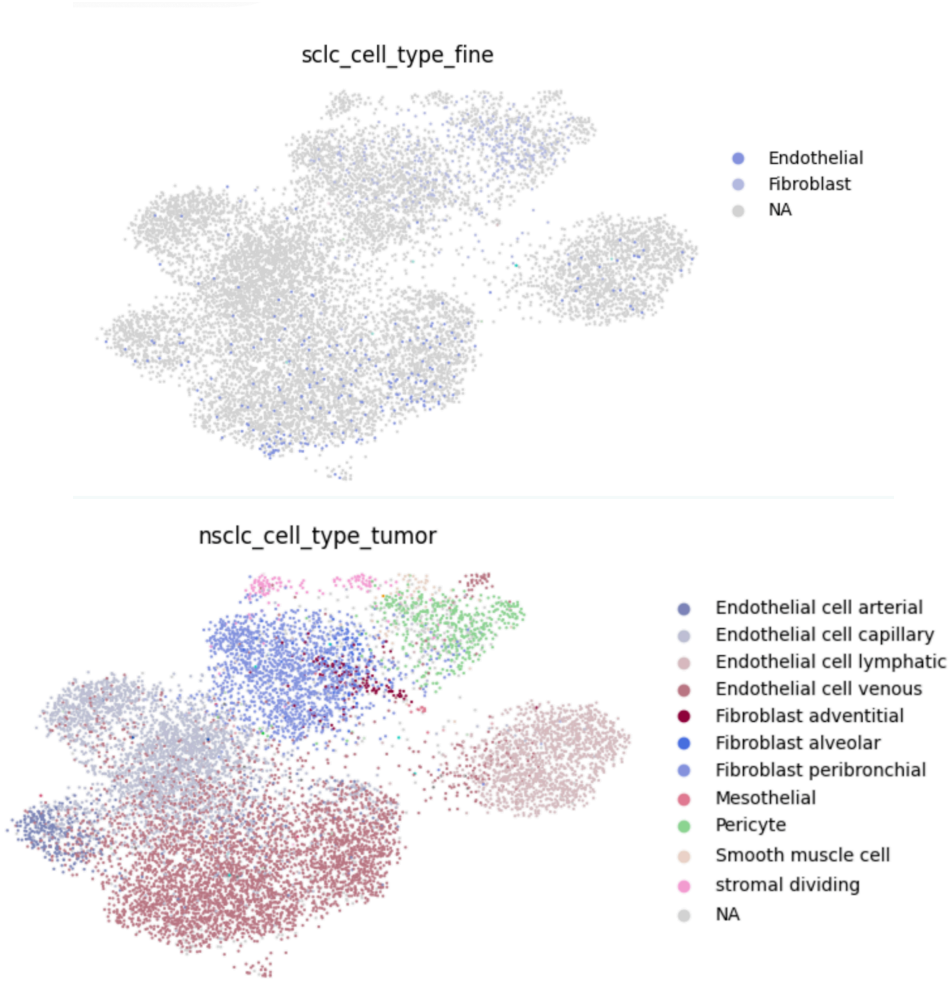

Figure 60: UMAPs of the mesenchymal subset of the first half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom).

There is not much room for inconsistencies between the two annotation sources, but the fibroblast half of the SCLC subset includes a significant portion of cells that maps to pericytes on the NSCLC side.

This is further reiterated upon the addition of a healthy mesenchymal cell sample from HLCA[169], consisting of 2000 (or all available if less) cells for each cell type at the finest level of annotation. The resulting projection is shown in Figure 61. Again, a significant portion of SCLC-annotated fibroblasts clusters away from its assigned group and maps to close to pericytes from both other sources. The pericytes of cancer patient donors of both types do appear to differ somewhat from the healthy ones, the former clustering more closely with smooth muscle cells.

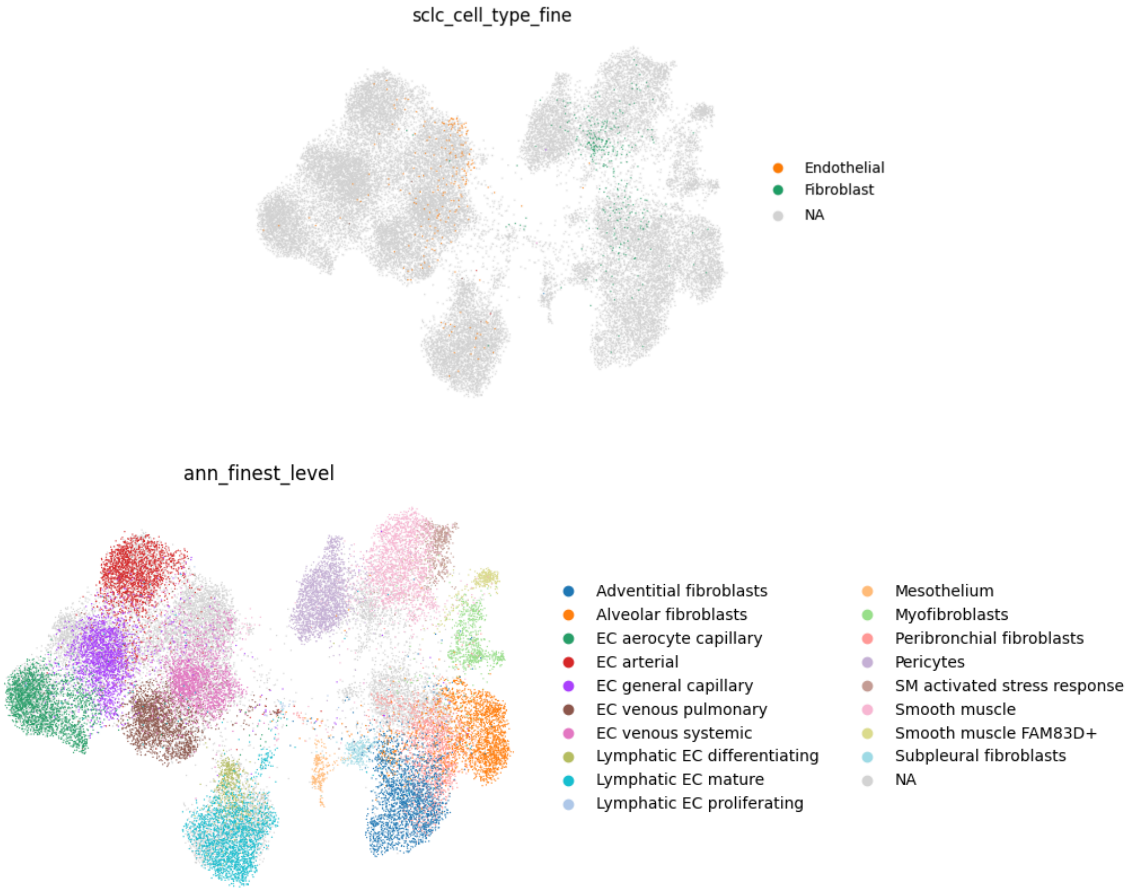

Figure 61: UMAPs of the mesenchymal subset of the first half of the dataset post scVI integration with additional healthy data annotated by SCLC atlas coarse annotation (top) and healthy lung cell atlas annotation (bottom).

3.4.3.2 The myeloid subset

The annotation of the myeloid subset is once again more detailed on side of the originally larger NSCLC atlas source. The integrated dataset annotated according to both sources is shown in Figure 62. There is a notable inconsistency when it comes to the annotation of dendritic cells from the SCLC atlas subset - a large portion of these map anywhere but any of the dendritic cell subtypes. The clusters of the NSCLC annotation are largely well recovered.

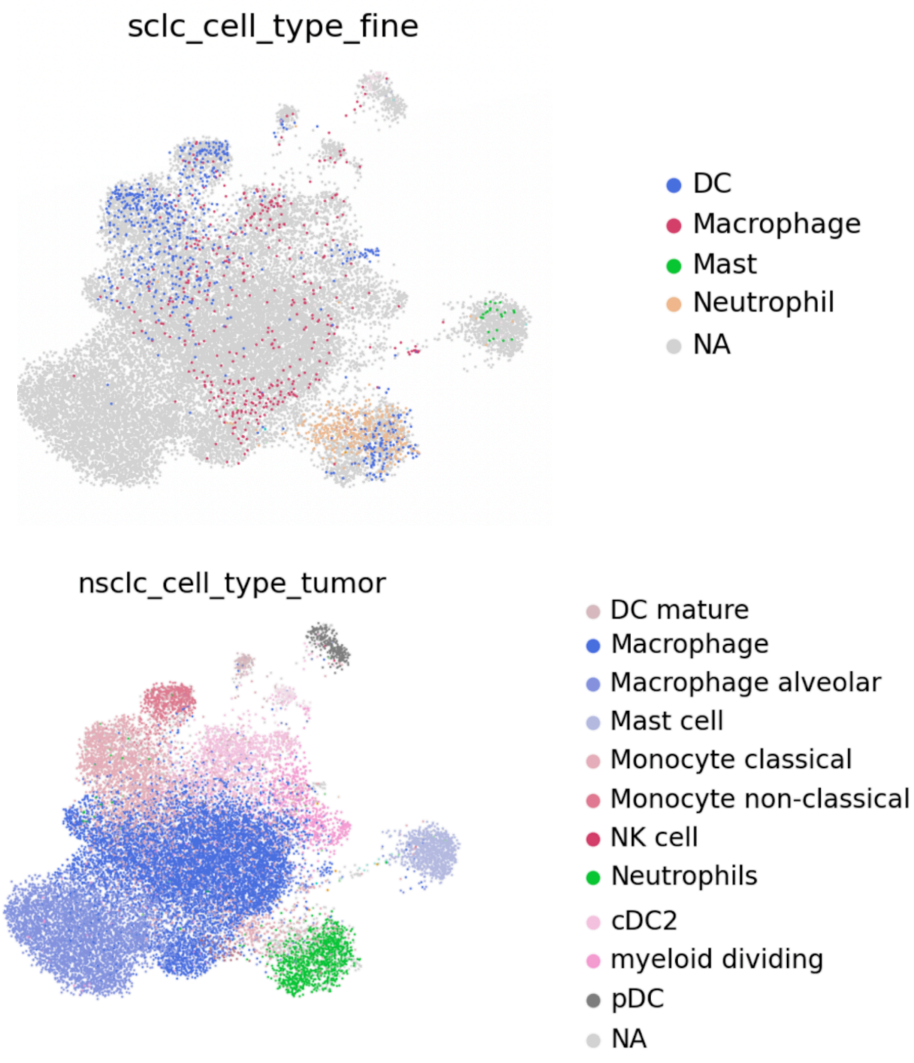

Figure 62: UMAPs of the myeloid subset of the first half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom).

Inspection of canonical markers of the four subtypes of dendritic cells identified here (cDC1, cDC2, DC mature, and pDC) in Figure 63 further disputes the possibility of these being actual dendritic cells, as their expression is clearly contained to their respective clusters as following the NSCLC atlas source annotation.

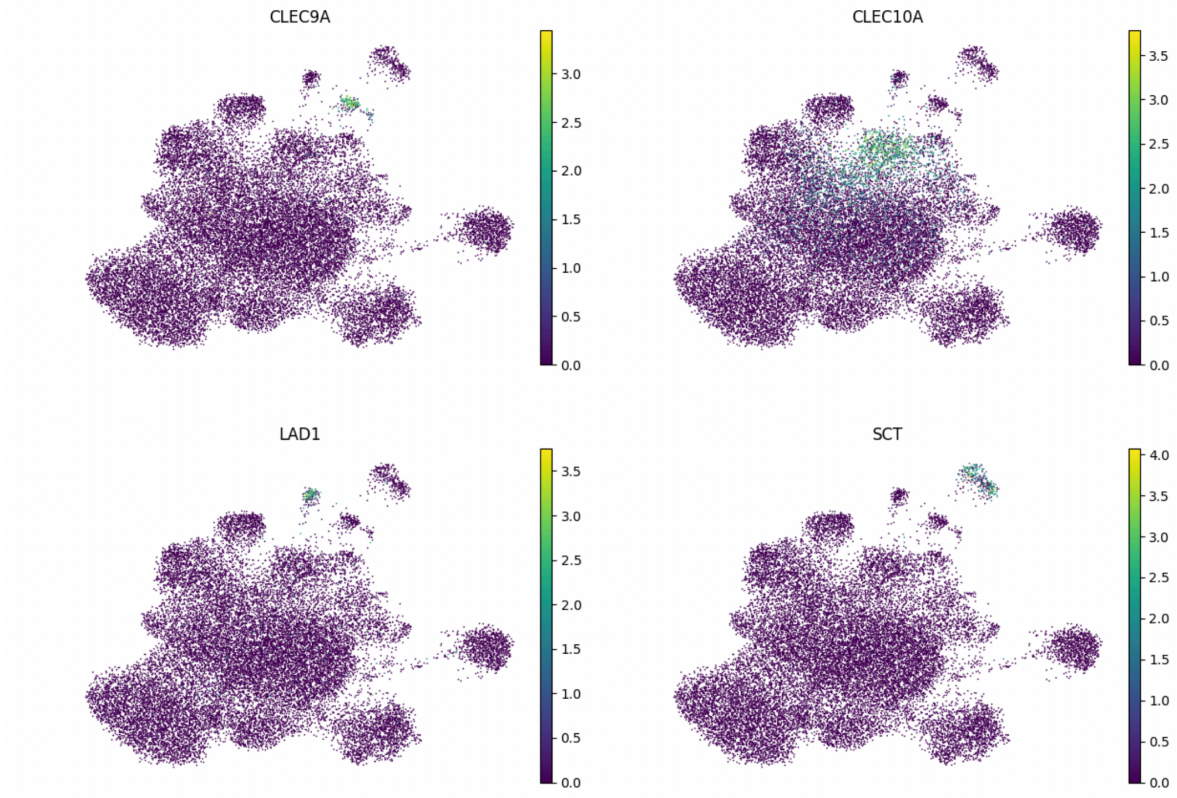

Figure 63: UMAPs of the myeloid subset of the first half of the dataset post scVI integration annotated by the expression of markers of four DC cell types: cDC1 (top left), cDC2 (top right), DC mature (bottom left), and pDC (bottom right).

3.4.3.3 The lymphoid subset

The integrated lymphoid subset annotated according to both sources is shown in Figure 64. There is clear separation of B cells and plasma cells, while the T and NK cells show less distinction, despite extremely granular annotation from the NSCLC source. The annotations at this level are not well recovered by clusters.

No NK cells are included in the SCLC annotation, but a significant portion of what is annotated as

T cells clearly maps to the NK cell section of the NSCLC part of the subset. NK marker expression shown in Figure 65 further supports the validity of this mapping, as the previously annotated T cells show levels of expression consistent with the rest of the NK cell cluster.

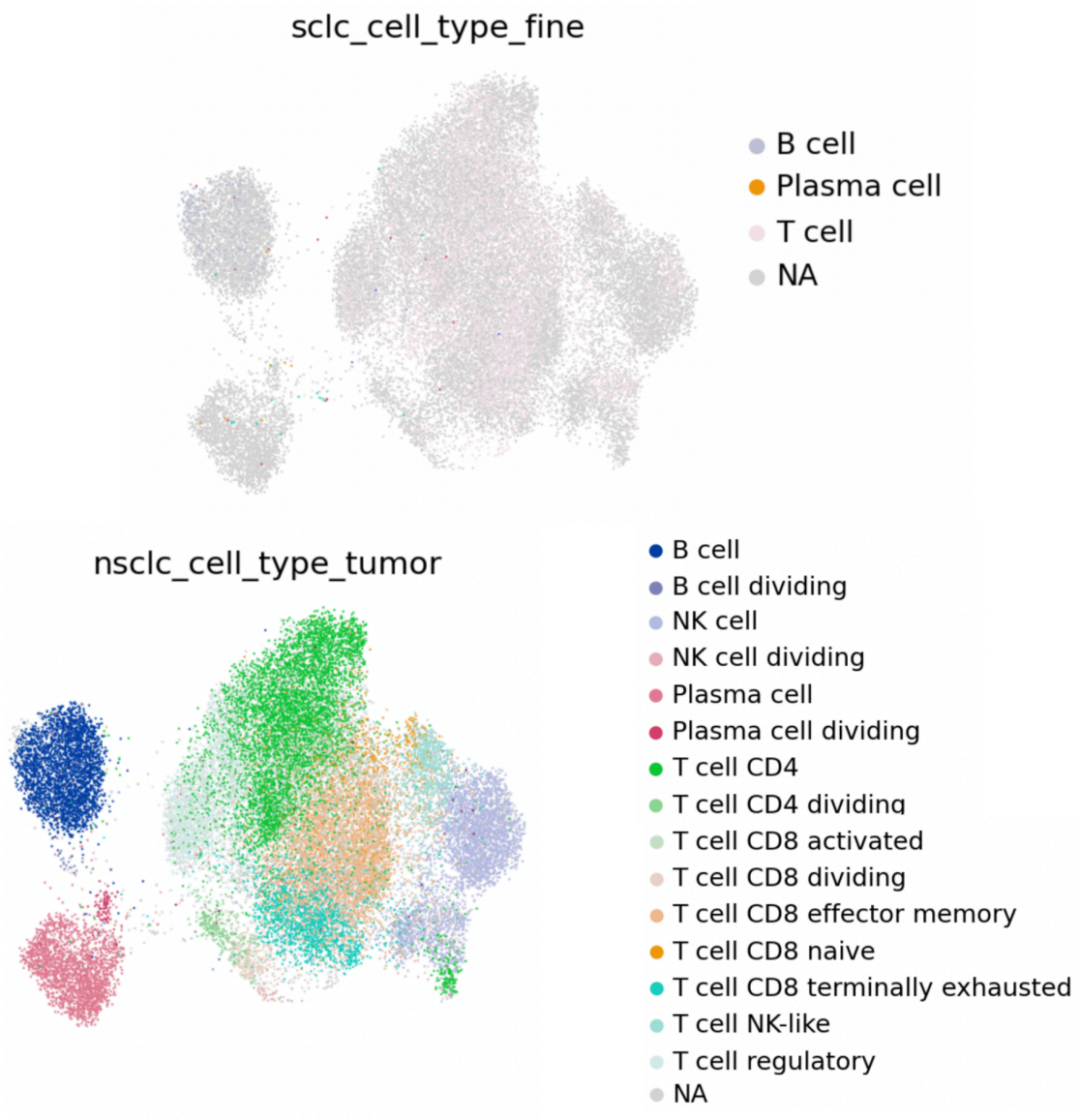

Figure 64: UMAPs of the lymphoid subset of the first half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom).

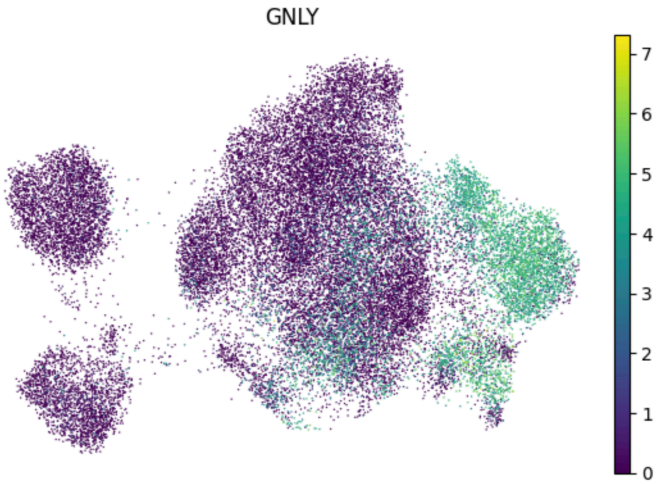

Figure 65: UMAP of the lymphoid subset of the first half of the dataset post scVI integration annotated by the expression of a marker of the NK cell type.

3.4.3.4 The epithelial subset

Figure 66 shows the integrated epithelial dataset annotated according to both sources. The annotations are detailed on both sides, which is in contrast with the few clear clusters arising.

Though there are few healthy cells included in either part of the subset, the SCLC source includes detailed annotations of very rare populations. Though only the SCLC patients of the atlas were selected (and not the small NSCLC background reference), a few stray cells were annotated as NSCLC, which can be presumed a misannotation.

The NSCLC source annotations, on the contrary, are mostly focused on the tumour cells, and consist of clinical histology, as well as very specific functional features.

While SCLC and NSCLC tumour cells show separation, it appears that gene expression profile based integration of tumour cells (even upon repetition of experiment in isolation from the rest of epithelial cells) does not leave enough of variation in the signal to allow tumour cell subtype characterisation via clustering.

CNV calling and CNV-based clustering was performed as an alternative avenue, but did not yield useful results. Much like with non-integrated gene expression, the patient specific signal was strong enough to drown out anything else potentially present in the mix, which is reflected in the completely patient based CNV clusters seen in Figure 67.

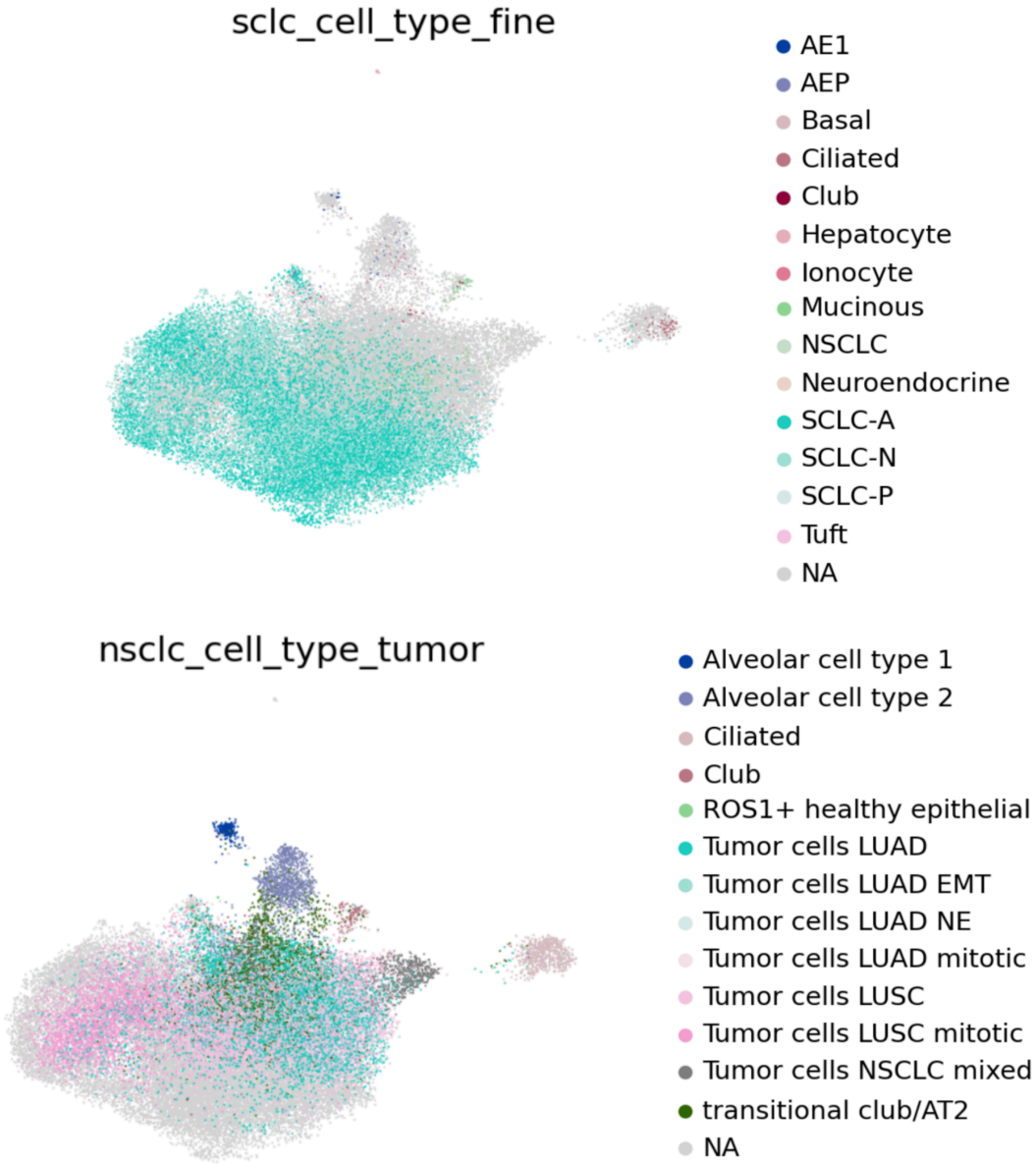

Figure 66: UMAPs of the epithelial subset of the first half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom).

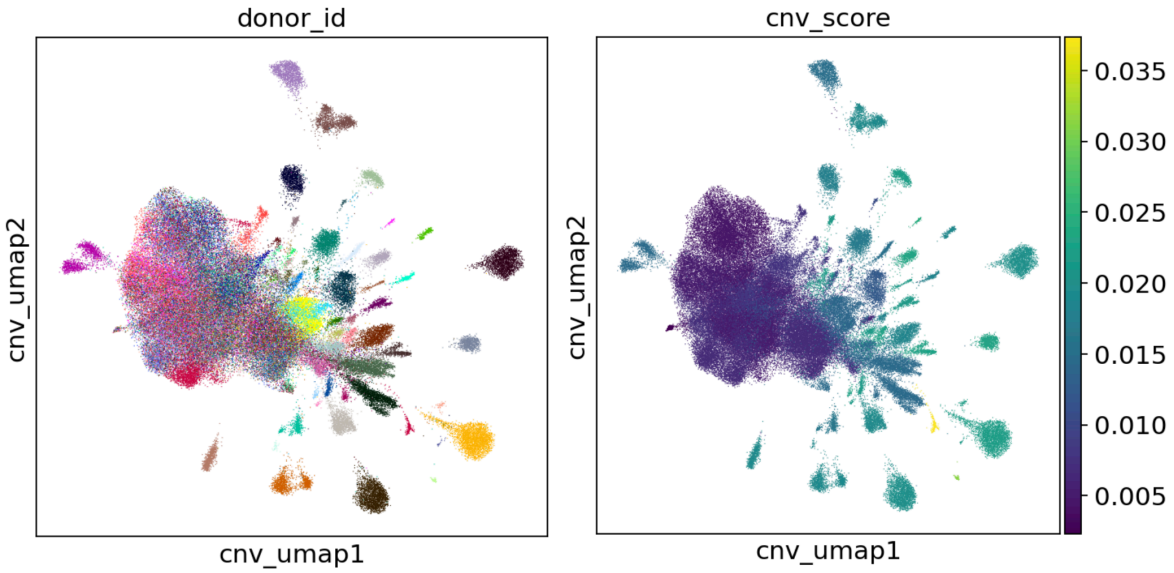

Figure 67: UMAPs of the CNV clusters of the first half of the dataset post FastCNV analysis annotated by the expression of donor (left) and CNV score (right).

3.4.3.5 Marker analysis

The markers proposed by the source atlases are shown in Figures 68 and 69. The NSCLC atlas proposes a number of markers without any kind of scores or ranking information. Though cell type annotation reaches an incredibly granular degree (all levels collected in Table 5), the markers are only proposed for the second coarsest one.

The specificity of markers observed varies largely between cell types. While some (e.g. ciliated cells) have a number of strong markers available, others are highly confounded with other types (e.g. plasma cells).

The SCLC atlas only proposes markers of SCLC subtypes (Figure 69). Though they undoubtedly carry signal, there is still a level of noise in the expression across other cell types.

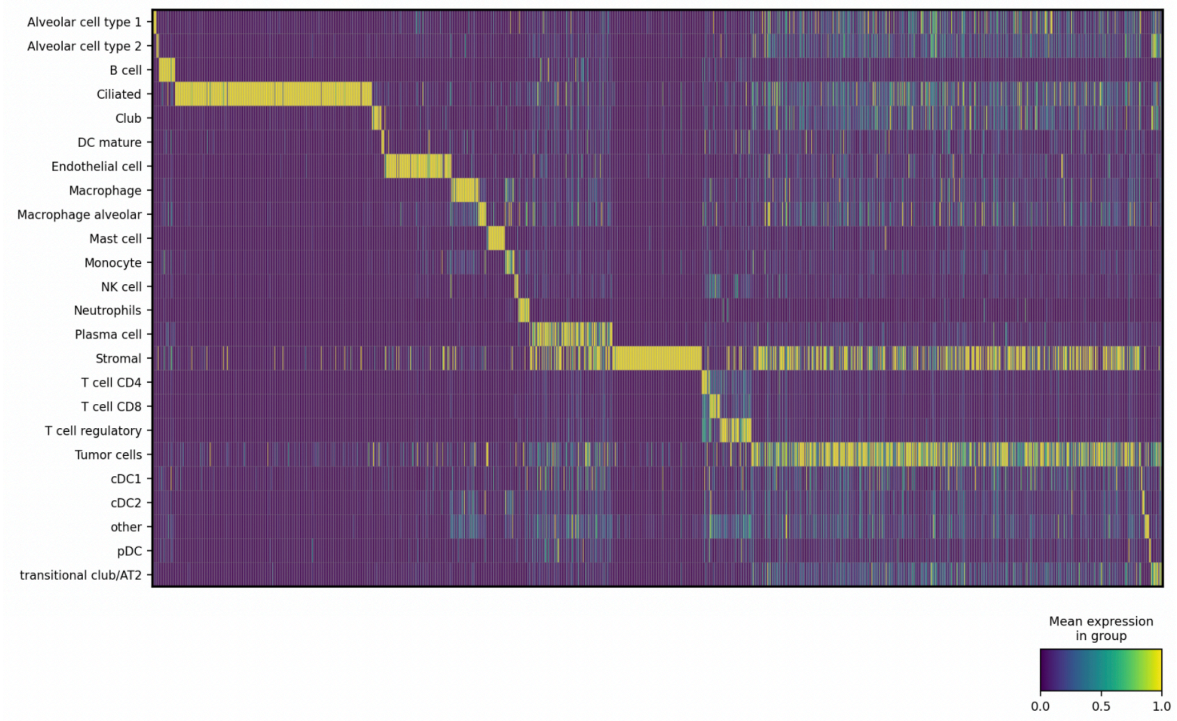

Figure 68: Matrixplot of proposed marker genes for each cell type in the NSCLC atlas[170]. Cell types are shown in rows, their markers are plotted in columns. Mean expression is scaled by gene.

Level 1	Level 2	Level 3	Level 4	Tumour level
Epithelial	Alveolar cell type 1	Alveolar cell type 1	Alveolar cell type 1	Alveolar cell type 1
	Alveolar cell type 2	Alveolar cell type 2	Alveolar cell type 2	Alveolar cell type 2
	Ciliated	Ciliated	Ciliated	Ciliated
	Club	Club	Club	Club
	transitional club/AT2	transitional club/AT2	transitional club/AT2	transitional club/AT2
			ROS1+ healthy epithelial	ROS1+ healthy epithelial
	Tumor cells	Tumor cells	Tumor cells	Tumor cells LUAD
				Tumor cells LUAD EMT
				Tumor cells LUAD NE
				Tumor cells LUAD mitotic
				Tumor cells LUSC
				Tumor cells LUSC mitotic
				Tumor cells NSCLC mixed
Lymphoid	B cell	B cell	B cell	B cell
			B cell dividing	B cell dividing
	T cell regulatory	T cell regulatory	T cell regulatory	T cell regulatory
	T cell CD4	T cell CD4	T cell CD4	T cell CD4
			T cell CD4 dividing	T cell CD4 dividing
	T cell CD8	T cell CD8	T cell CD8 activated	T cell CD8 activated
			T cell CD8 dividing	T cell CD8 dividing
			T cell CD8 effector memory	T cell CD8 effector memory
			T cell CD8 naive	T cell CD8 naive
			T cell CD8 terminally exhausted	T cell CD8 terminally exhausted
			T cell NK-like	T cell NK-like
	Plasma cell	Plasma cell	Plasma cell	Plasma cell
			Plasma cell dividing	Plasma cell dividing
	NK cell	NK cell	NK cell	NK cell
			NK cell dividing	NK cell dividing
Myeloid	DC mature	DC mature	DC mature	DC mature
	Macrophage	Macrophage	Macrophage	Macrophage
	Macrophage alveolar	Macrophage alveolar	Macrophage alveolar	Macrophage alveolar
	pDC	pDC	pDC	pDC
	cDC1	cDC1	cDC1	cDC1
	cDC2	cDC2	cDC2	cDC2
	Neutrophils	Neutrophils	Neutrophils	Neutrophils
	Monocyte	Monocyte classical	Monocyte classical	Monocyte classical
		Monocyte non-classical	Monocyte non-classical	Monocyte non-classical
	Mast cell	Mast cell	Mast cell	Mast cell
		myeloid dividing	myeloid dividing	myeloid dividing
Mesenchymal	Endothelial cell	Endothelial cell arterial	Endothelial cell arterial	Endothelial cell arterial
		Endothelial cell capillary	Endothelial cell capillary	Endothelial cell capillary
		Endothelial cell venous	Endothelial cell venous	Endothelial cell venous
		Endothelial cell lymphatic	Endothelial cell lymphatic	Endothelial cell lymphatic
	Stromal	Fibroblast lung	Fibroblast adventitial	Fibroblast adventitial
			Fibroblast alveolar	Fibroblast alveolar
		Fibroblast peribronchial	Fibroblast peribronchial	Fibroblast peribronchial
		Mesothelial	Mesothelial	Mesothelial
		Pericyte	Pericyte	Pericyte
		Smooth muscle cell	Smooth muscle cell	Smooth muscle cell
		stromal dividing	stromal dividing	stromal dividing
	*other**			

Table 5: Annotation levels of the NSCLC atlas[170] (names adjusted for consistency, additional neutrophile level omitted). *Includes a handful of subtypes excluded from the rest of level 2.

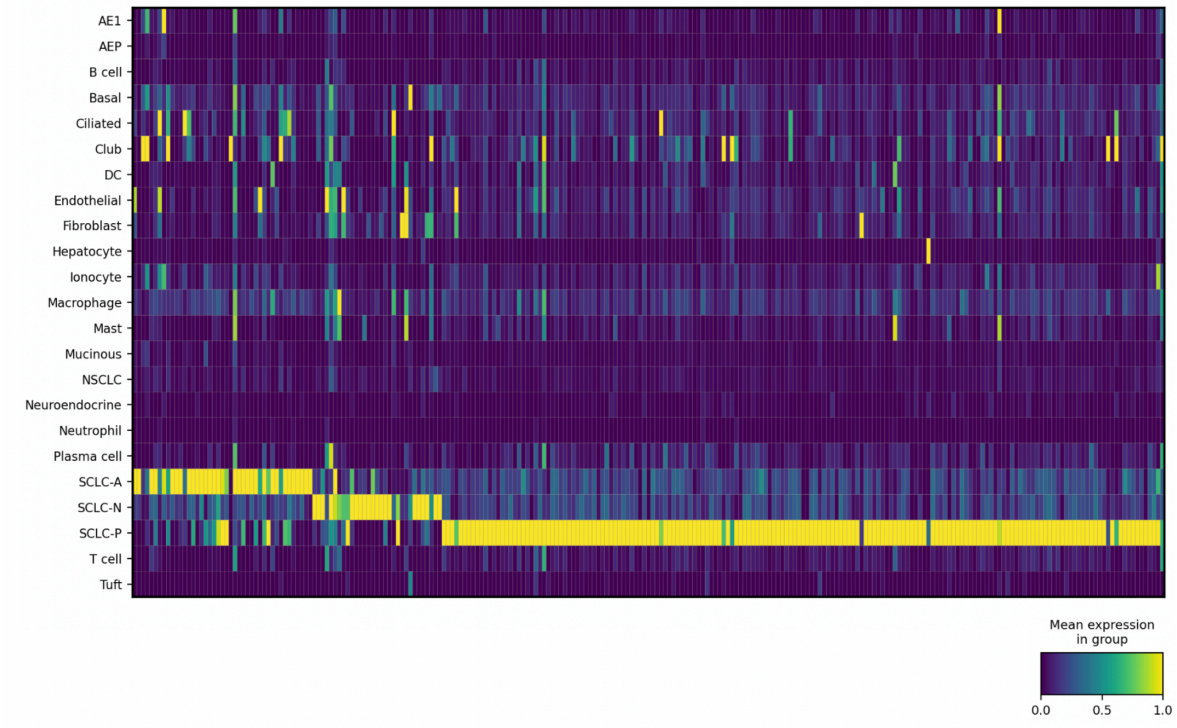

Figure 69: Matrixplot of proposed marker genes for cell types in the SCLC atlas (only markers for the three subtypes of SCLC were proposed)[171]. Cell types are shown in rows, their markers are plotted in columns. Mean expression is scaled by gene.

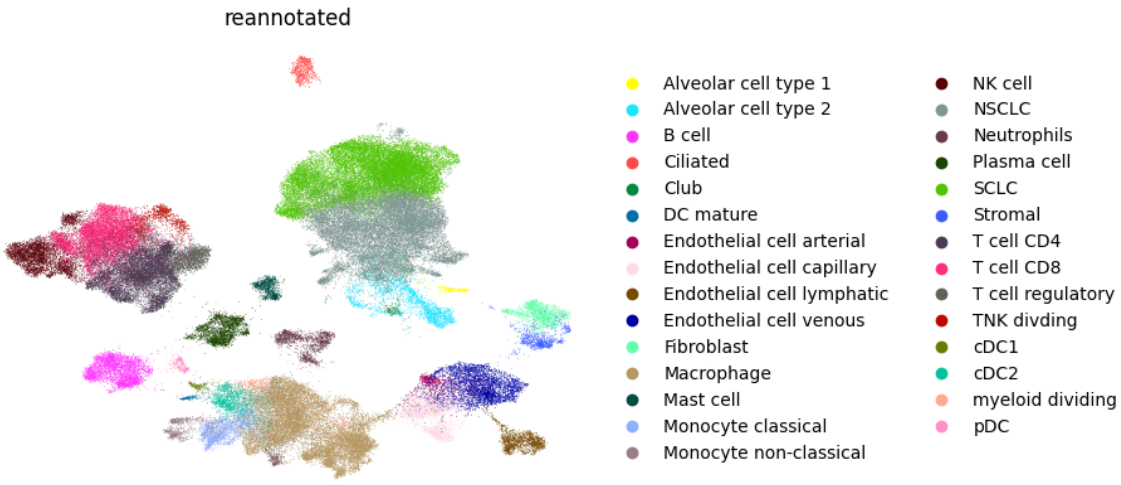

Figure 70: UMAPs of the first half of the dataset post scVI integration reannotated.

Our reannotation of the unified lung cell atlas is shown in Figure 70. The level of granularity has been iteratively fixed at the degree where markers of satisfactory quality can be identified for each cell type by the SCherlock algorithm. The matrixplot of top scoring markers is shown in Figure 71, while the full list of these markers and their scores is available in the supplementary results.

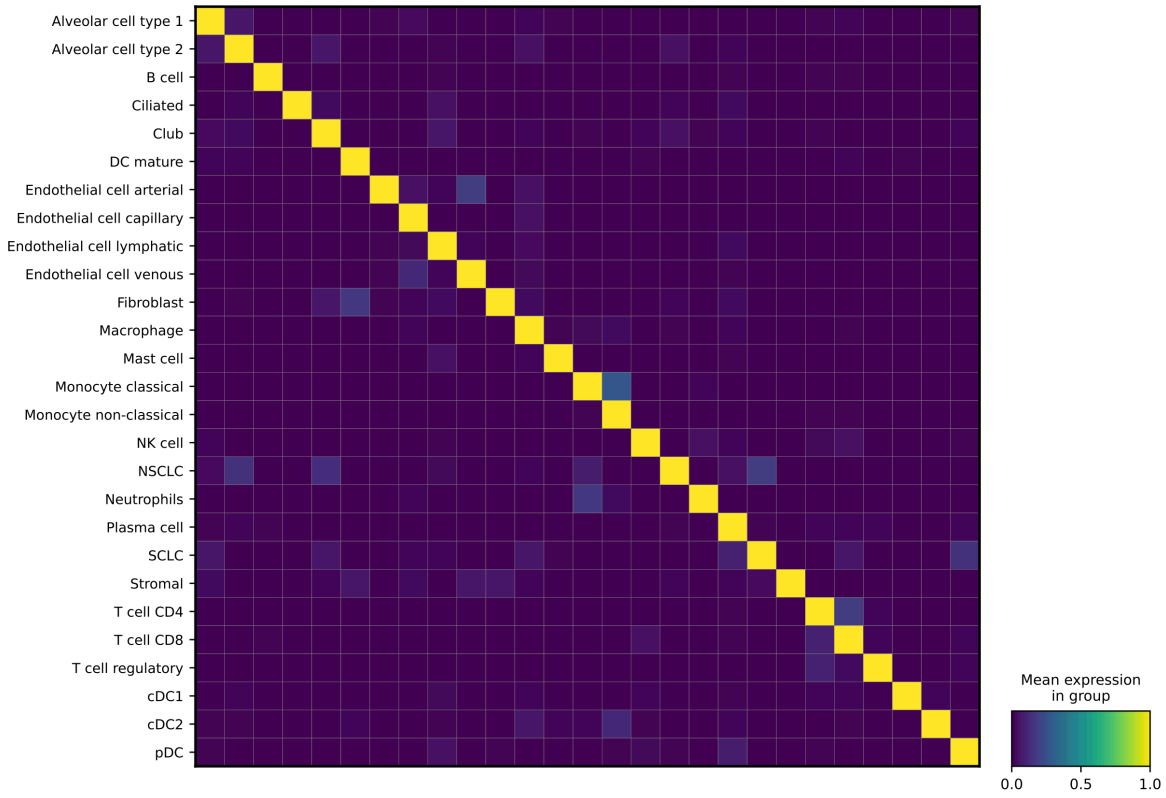

Figure 71: Matrixplot of proposed top scoring marker genes for reannotated cell types in the unified lung cancer cell atlas. Cell types are shown in rows, their markers are plotted in columns. Mean expression is scaled by gene.

3.4.4 Discussion

The unified lung cancer cell atlas described here unites gene expression data of lung cancers of all major histologies, including likely unidentified cases of LCNEC.

Several issues were identified in the source annotations and addressed in the reannotation provided with the atlas. Marker analysis was performed applying the SCherlock algorithm and a set of ranked

markers and their scores provided for each cell type. The markers improve on the specificity and performance as a set compared to what was proposed with the original sources.

One future direction where success has been limited so far is further characterisation of tumour cells. Balancing the need to remove the overwhelming tumour specific noise with the need to retain enough information about the interesting commonalities between them has proven difficult. Neither expression-based data integration nor a CNV-based approach have yielded desired results. In the next steps, we plan to explore their characteristics through predefined signatures of function and healthy cell identity.

3.4.5 Supplementary results

The supplementary results contain the list of top scoring markers of the reannotated cell types and their corresponding SCherlock scores, as well as all UMAP figures for the second half of the dataset.

Cell type	Gene symbol	SCherlock score
Alveolar cell type 1	MYRF	0.795
Alveolar cell type 2	ROS1	0.785
B cell	MS4A1	1.000
cDC1	CLEC9A	0.871
cDC2	CD1A	0.917
Ciliated	TEKT1	1.000
Club	TSPAN8	0.571
DC mature	CCL19	0.690
Endothelial cell arterial	DKK2	0.870
Endothelial cell capillary	CA4	0.710
Endothelial cell lymphatic	ACKR2	0.843
Endothelial cell venous	SELE	0.839
Fibroblast	DPT	0.922
Macrophage	FABP3	1.000
Mast cell	GCSAML	0.991
Monocyte classical	S100A12	0.604
Monocyte non-classical	LILRA1	0.518
Neutrophils	CXCR1	0.920
NK cell	KLRF1	0.946
NSCLC	TMPRSS11E	1.000
pDC	PACSIN1	0.761
Plasma cell	SPAG4	0.757
SCLC	CGHA	1.000
Stromal	FHL5	0.942
T cell CD4	CD40LG	0.946
T cell CD8	CD8B	0.862
T cell regulatory	FOXP3	1.000

Table 6: Top scoring marker genes and their respective SCherlock scores for reannotated cell types in the unified lung cancer cell atlas.

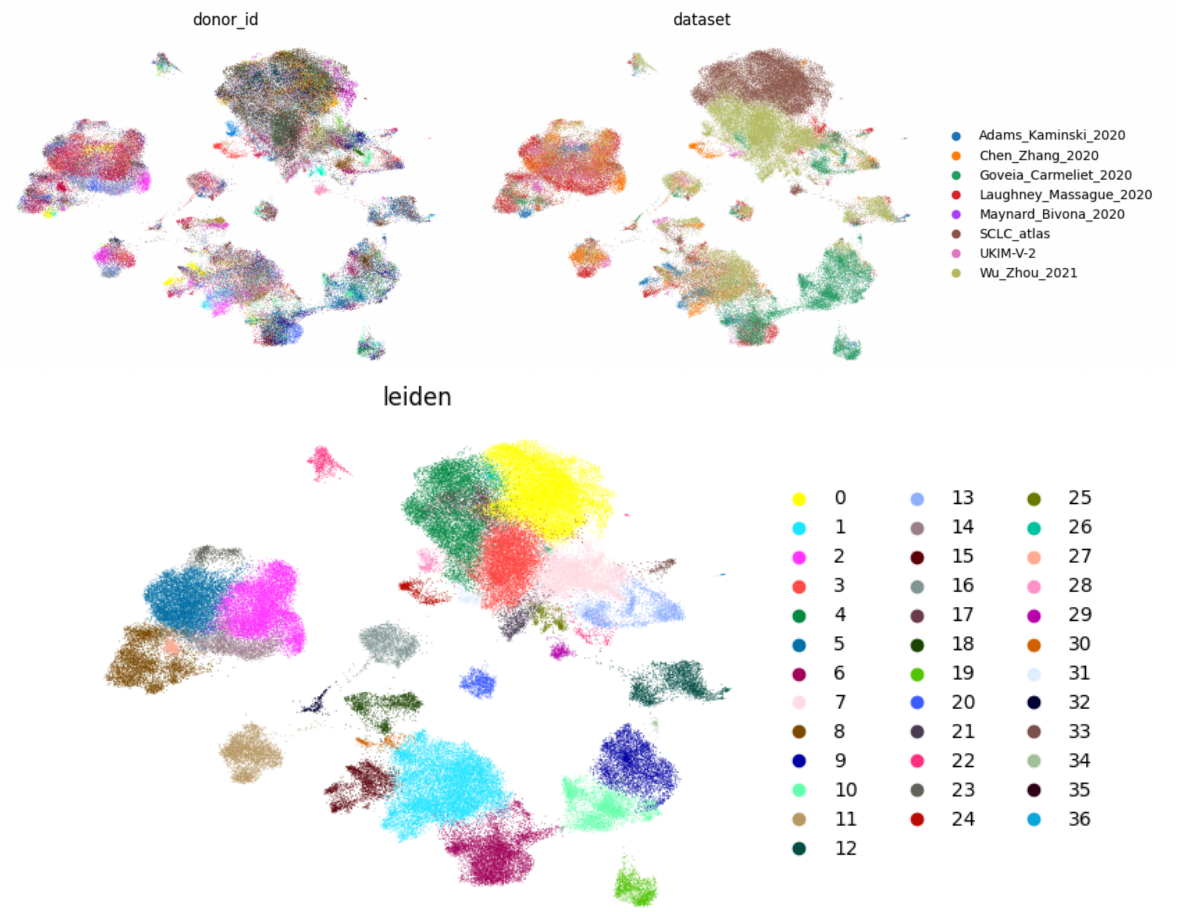

Figure 72: UMAPs of the second half of the dataset post scVI integration annotated by donor (top left), source dataset (top right) and leiden cluster (bottom).

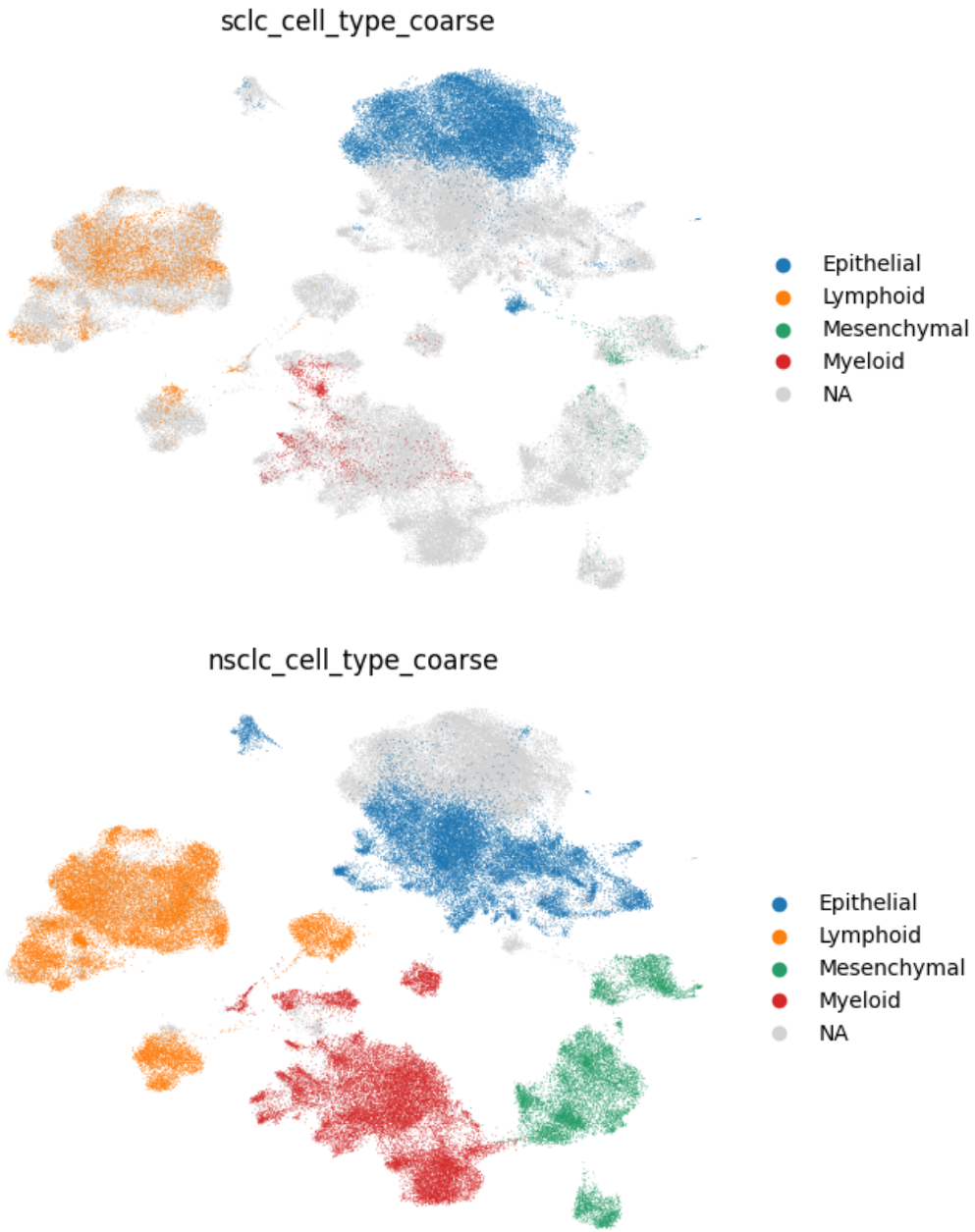

Figure 73: UMAPs of the second half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom).

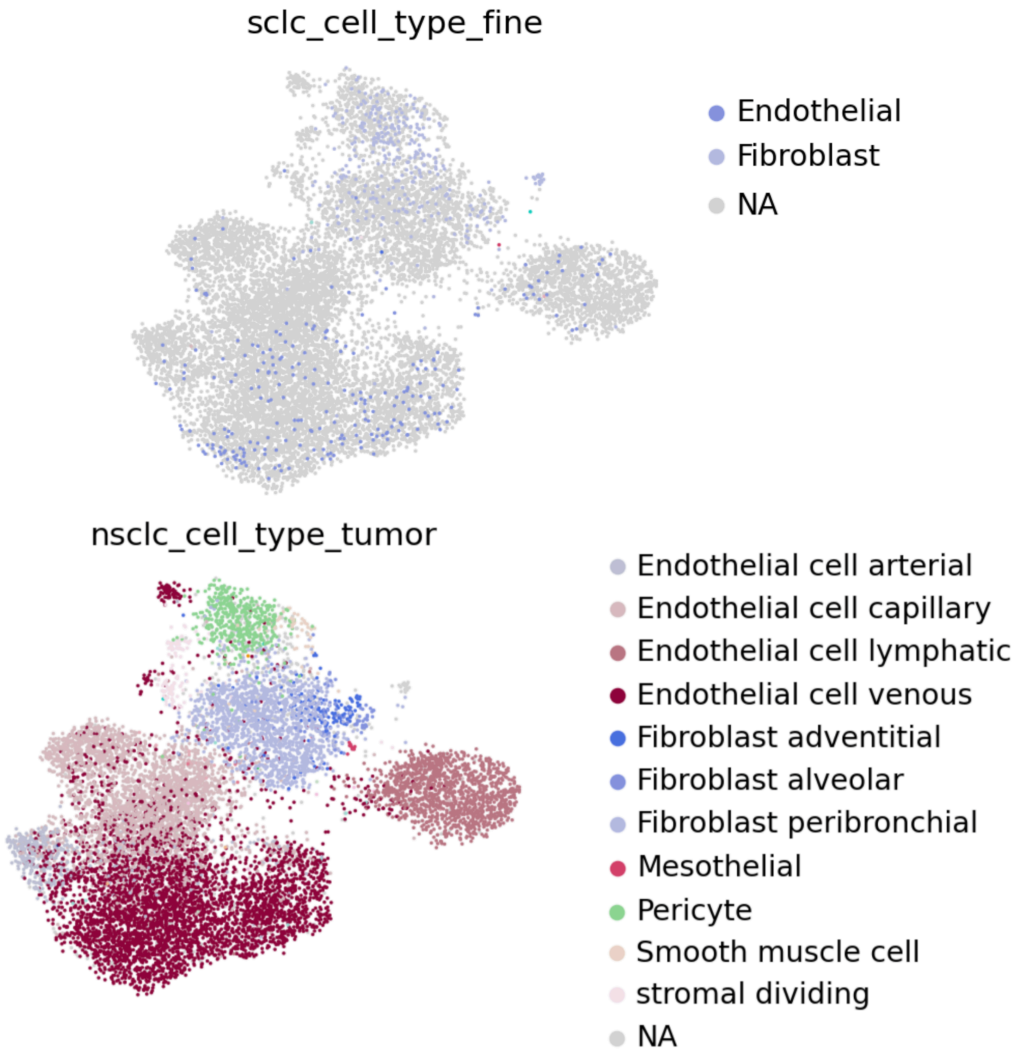

Figure 74: UMAPs of the mesenchymal subset of the second half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom).

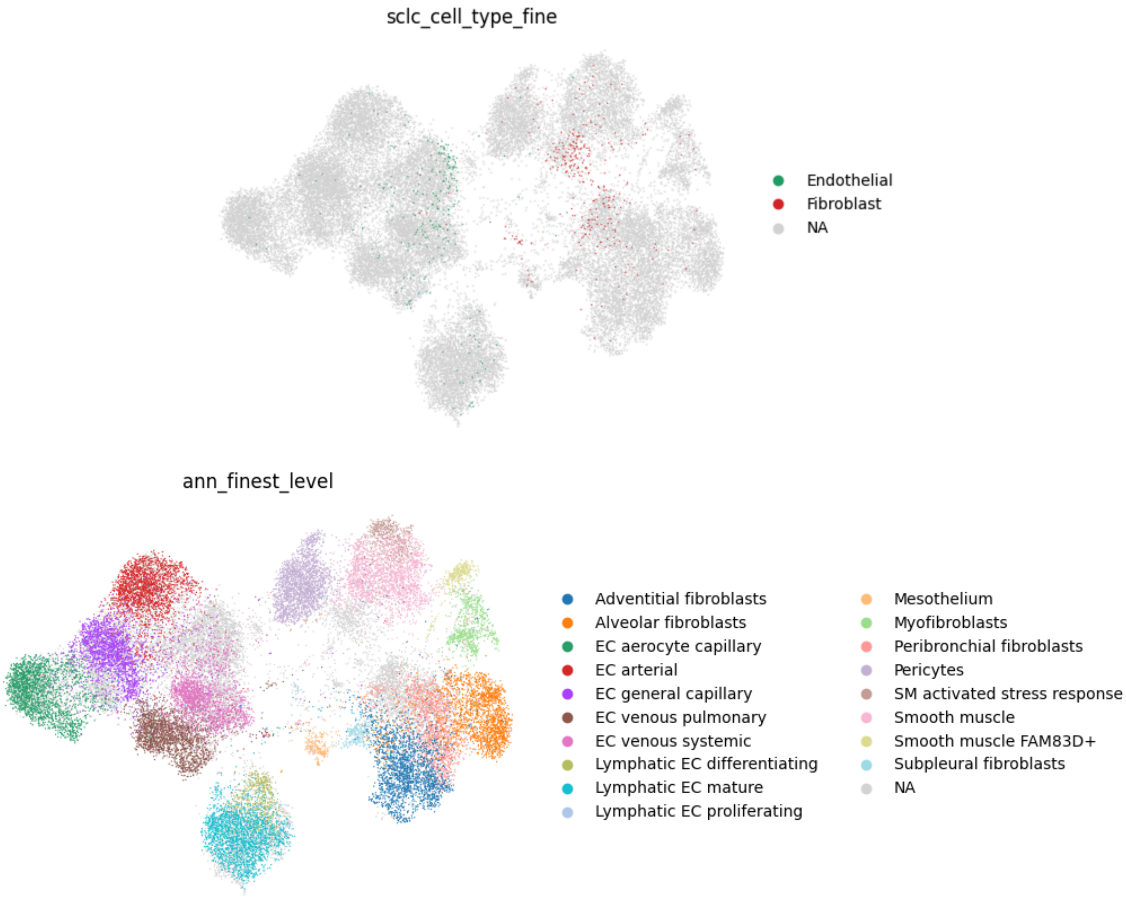

Figure 75: UMAPs of the mesenchymal subset of the second half of the dataset post scVI integration with additional healthy data annotated by SCLC atlas coarse annotation (top) and healthy lung cell atlas annotation (bottom).

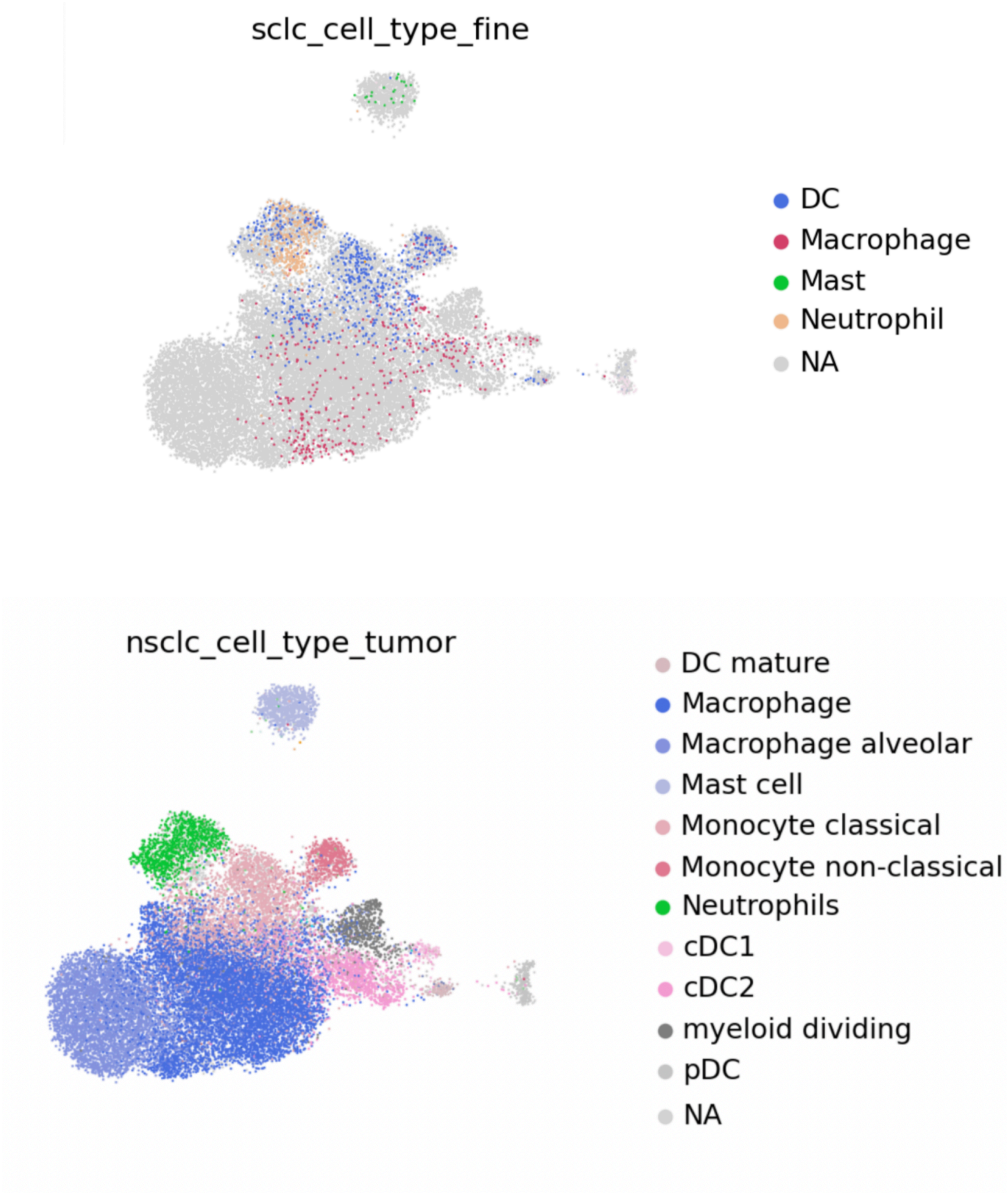

Figure 76: UMAPs of the myeloid subset of the second half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom).

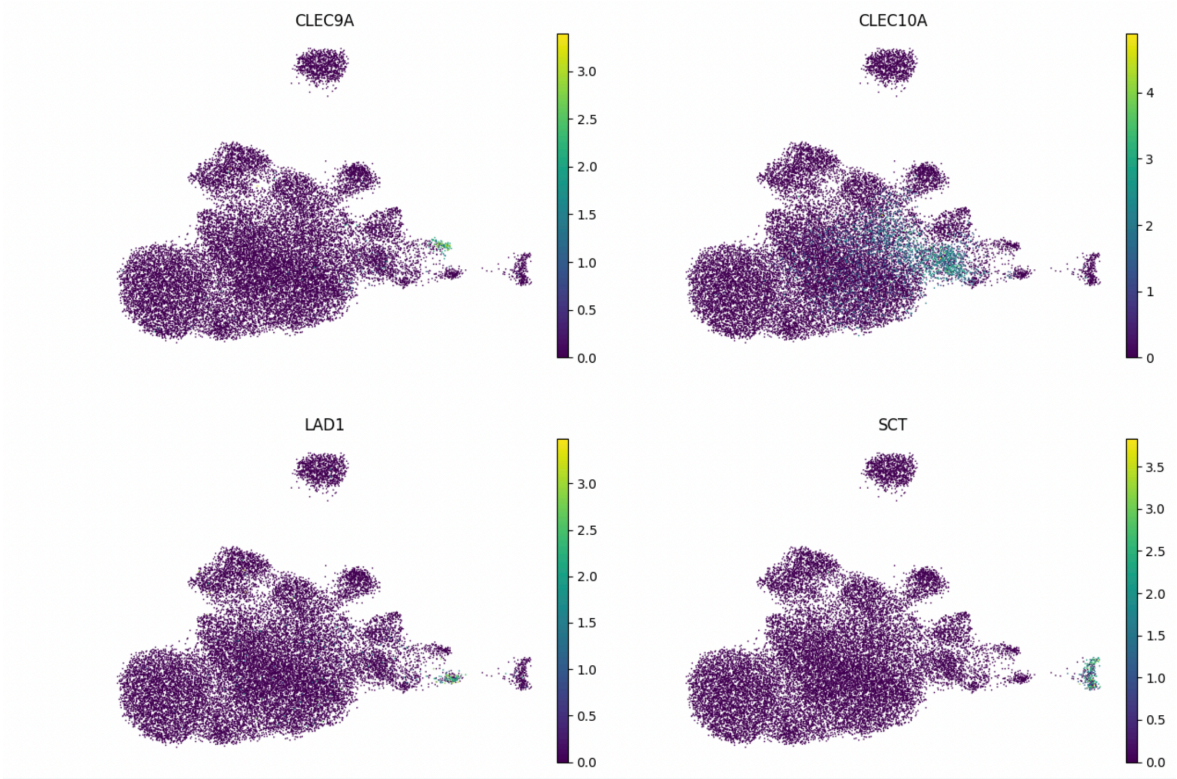

Figure 77: UMAPs of the myeloid subset of the second half of the dataset post scVI integration annotated by the expression of markers of four DC cell types: cDC1 (top left), cDC2 (top right), DC mature (bottom left), and pDC (bottom right).

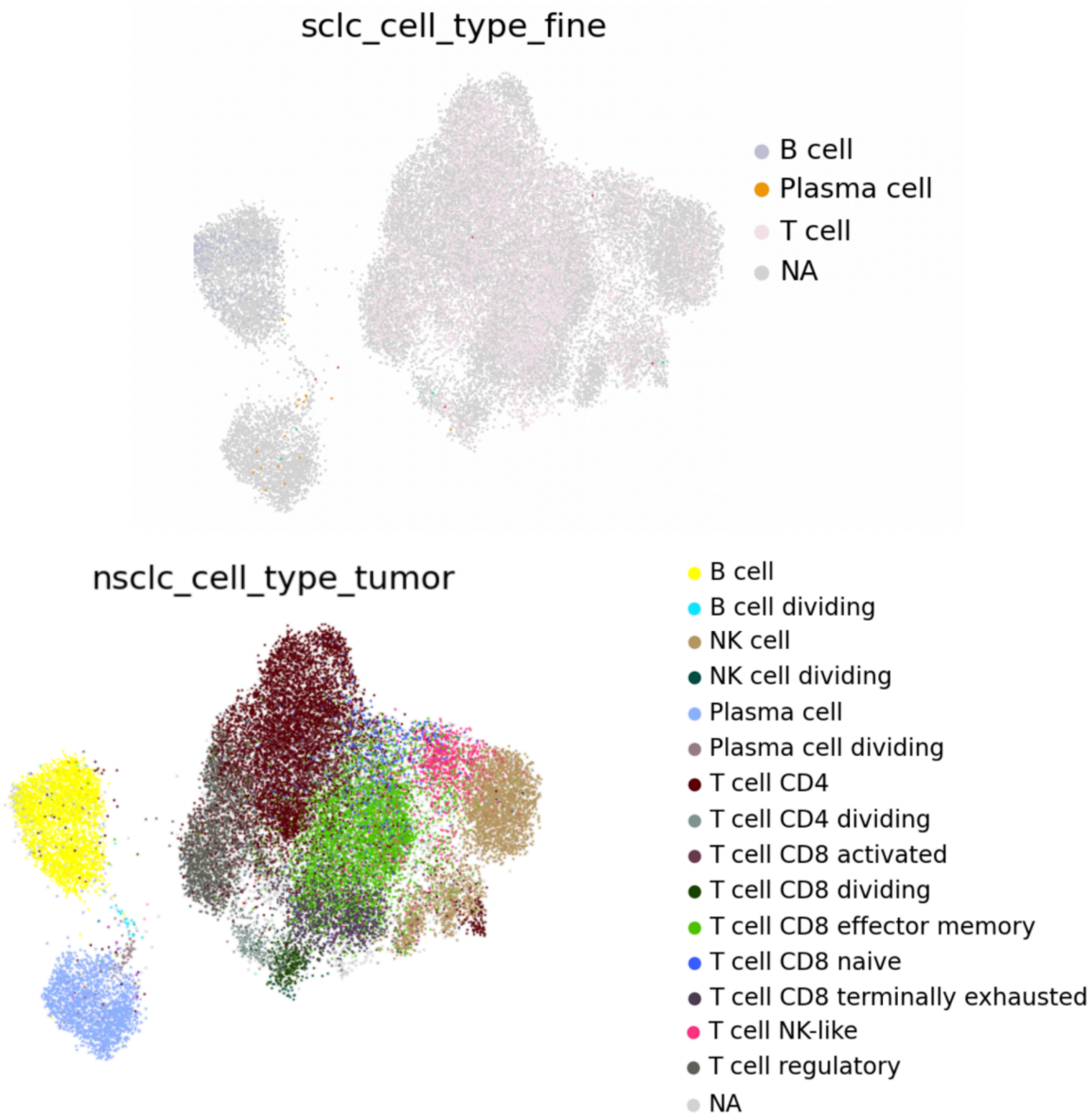

Figure 78: UMAPs of the lymphoid subset of the second half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom).

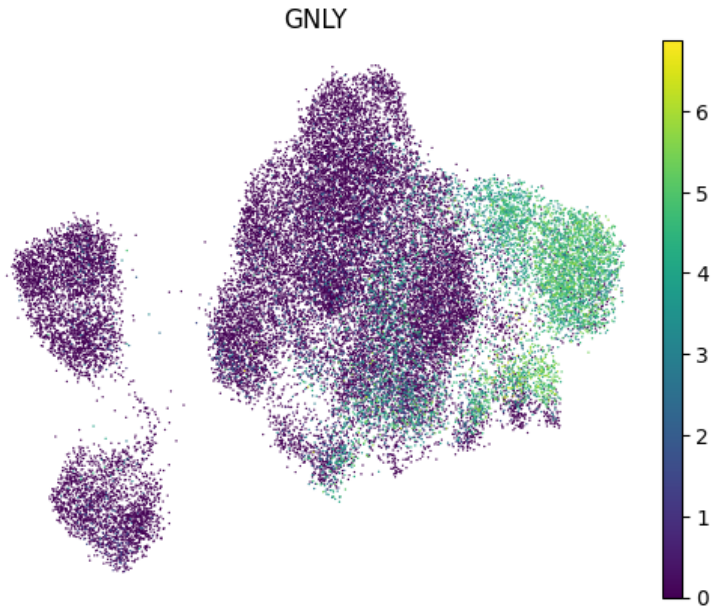

Figure 79: UMAP of the lymphoid subset of the second half of the dataset post scVI integration annotated by the expression of a marker of the NK cell type.

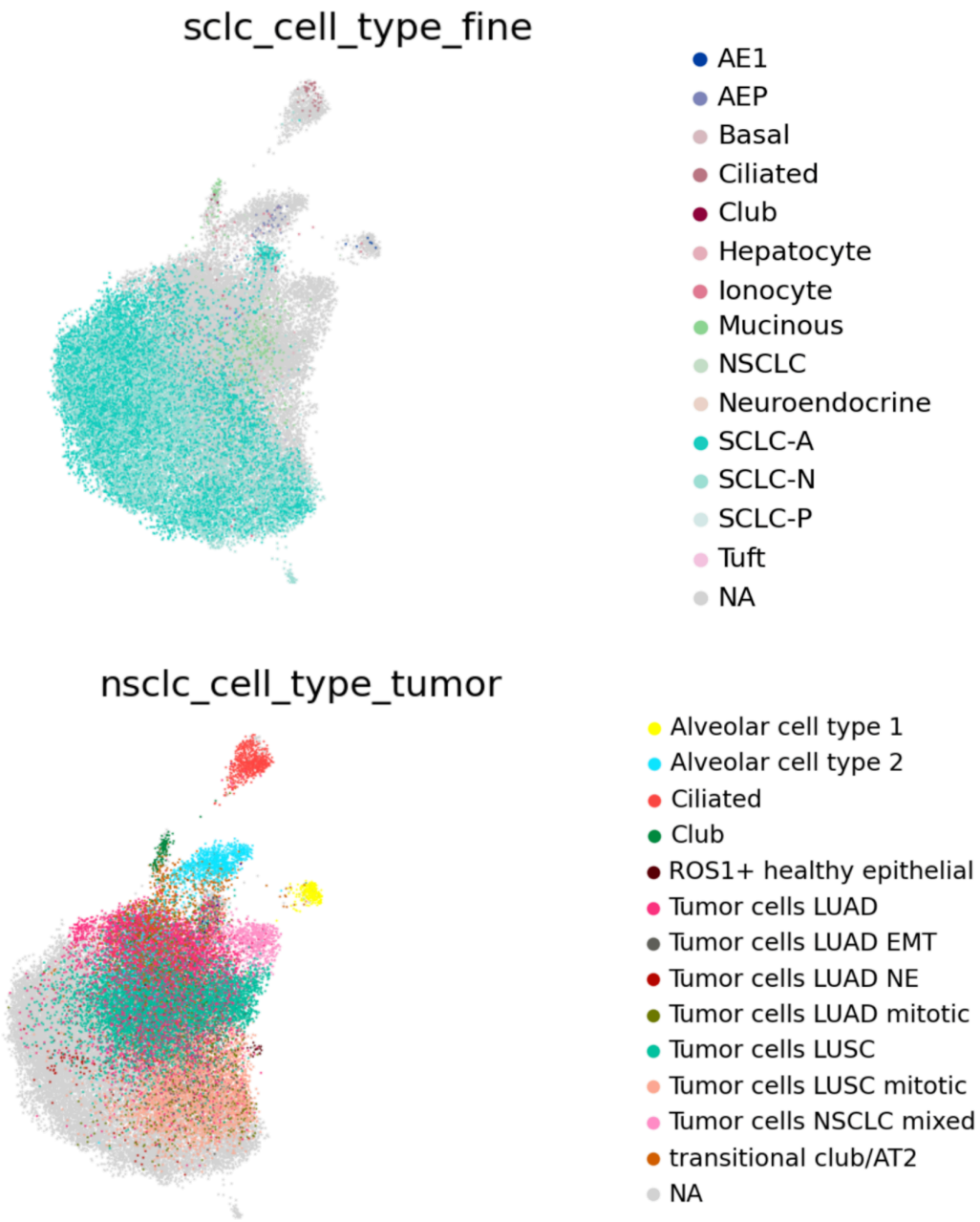

Figure 80: UMAPs of the epithelial subset of the second half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom).

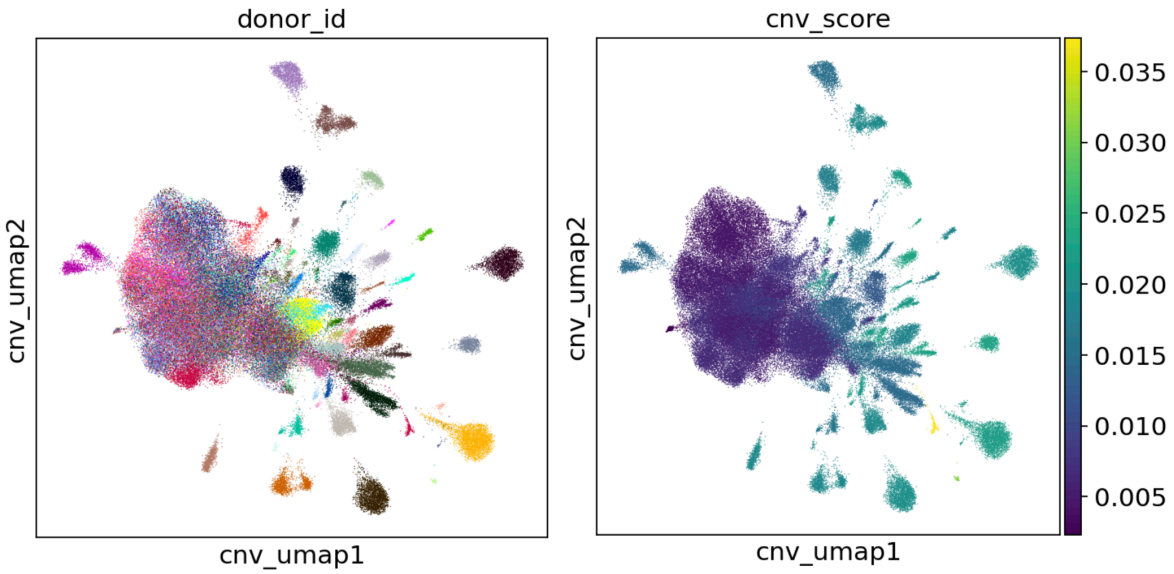

Figure 81: UMAPs of the CNV clusters of the second half of the dataset post FastCNV analysis annotated by the expression of donor (left) and CNV score (right).

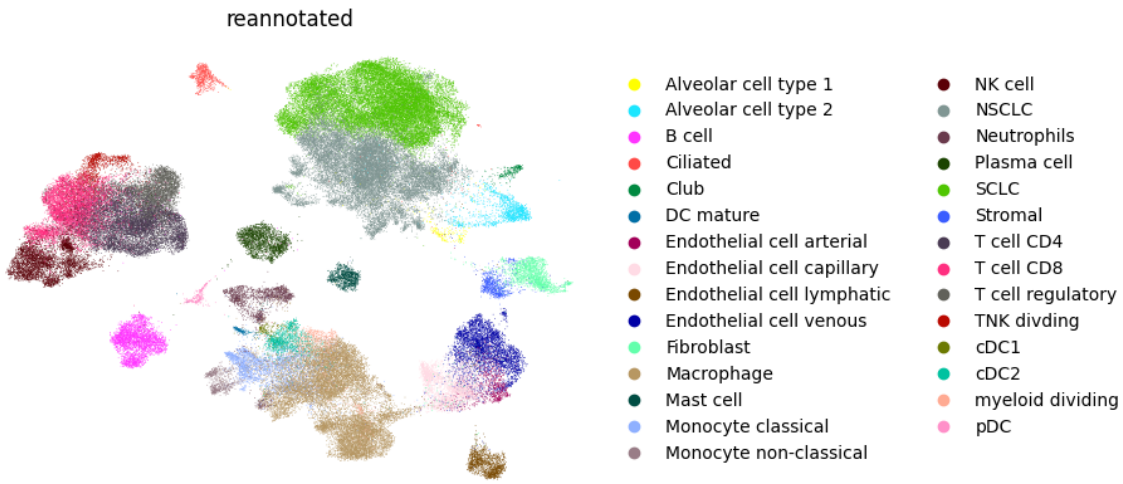

Figure 82: UMAPs of the second half of the dataset post scVI integration reannotated.

4 Discussion

In this work, I presented several advancements in the study of intra-tumour heterogeneity, comprising of three methodological parts and a unified lung cancer atlas.

In the methodological portion of the thesis, I presented three algorithms, which address three separate problems, but are all directly related to the study of ITH at the level of single-cell and spatial transcriptomic data.

The first method proposed was SCherlock, a marker identification algorithm for single-cell data. It is designed in particular with consideration for knowledge transfer and differences of contexts in mind. The markers identified are highly specific, sometimes at the expense of lower sensitivity. This is a trade-off made with an awareness of single-cell transcriptomics being an inherently low read context. The issue is potentially irrelevant in the case of transfer into bulk and pseudo-bulk transcriptomics, where more reads are available (we are somewhat in control of this with the choice of multiplication factor parameters, which simulate higher sensitivity contexts and directly influence maker prioritisation). In contexts where sensitivity remains an issue, this can be compensated by the use of metamarkers combining multiple genes, or treatment of clusters, rather than individual instances. One of the advantages of SCherlock is that it can be used to check the granularity of a proposed cell atlas: each time that a proposed cell type comes with no specific marker, it raises questions about the cell type in question. Is it a transitional state between 2 other types (which might explain why it has no specific marker), does it mix different subtypes, or is it actually identical to another subtype in the atlas? Ultimately, some subtypes are by essence transitional (e.g. transit amplifying cells in the colon mucosa), so an absence of specific markers doesn't mean they don't exist biologically, but it does alert us to the (im)possibility of deconvoluting them. The view of markers in this work maximises their suitability for identification and deconvolution purposes. It also represents a deviation from how markers are considered and treated in general in the existing literature.

The second methodological part consisted of a spatial pattern classifier, trained to distinguish between several predefined spatial distribution categories of biological interest when encountered in tissue gene expression. For now, the classification is limited to simulated data from the custom built generator designed to mimic real life data in key characteristics. Though significant further effort is required for extension to real life datasets, this work establishes the concept of distinguishing between different types of spatial patterns based on graph-derived properties of local neighbourhoods. Furthermore, the use of custom generated data for classifier training enables us to explore the impact of factors such as the number of reads per sample on successful distinction between

different pattern types. This allows us to adjust expectations accordingly, but can also encourage us to consider new alternative strategies and seek other solutions to the limitation at hand.

The final method proposed was FastCNV, a copy number variation inference algorithm for single-cell and spatial transcriptomic data. While the method is largely based on practices used in other already established algorithms (such as the genomic window approach), it introduces a unique amplification step. This feature addresses two particular concerns encountered in the context of CNV inference on transcriptomic data at this resolution. The pooling of similar cells or spots allows for increased sensitivity, mitigating the issues arising from a limited number of reads typically available in data coming from these technologies. On the other hand, reduced computational time is a direct consequence of treatment of meta-cells or spots, rather than individual instances. Since the dimensionality of such datasets is typically high, this reduction in computational demands is very welcome. Identifying CNVs in tumour cells is important when building a tumour cell atlas based on scRNA-seq data, because CNVs are most often patient/sample specific, and drive patient/sample specific tumour cell clusters in class discovery analysis. Being informed of the CNVs allows to clearly identify the signal at the source of these patient specific clusters. It may also provide a basis for regressing out the effect of CNVs on the cellular transcriptome.

In the final part of the results chapter, I presented a unified lung cancer cell atlas. To the best of my knowledge, it is the first lung cancer cell atlas whose scope covers all major lung cancer histologies. The data was sourced from several published studies via two existing integrative efforts. Several issues with source annotations were identified and addressed. The resulting atlas is reannotated at a level of granularity that is recoverable by clustering and at which satisfactory cell type specific markers have been identified by the SCherlock algorithm. A big part of interest in a unified lung cancer cell atlas is the study and characterisation of tumour cells. Isolating the signal of interest - of similarities beyond individual cases - has proven challenging. On the one hand, the strong signals of each individual tumour have the tendency to drown out everything else, while on the other, correcting for this with data integration appears to remove signals of interest as well. To counter this issue, tumour cells will be studied through expression signatures of function and healthy cell identity in the future stages.

4.1 Limitations

A limitation of the markers identified using SCherlock is that they are controlled only in the context of the training data, and typically might be irrelevant in another context. For example, *KRT5*, identified as a marker of pDCs in the context of PBMCs, would of course be irrelevant in the

context of an epithelial tissue, being a basal epithelium marker. Additionally, as any method aiming to identify markers, it can lead to identifying false positive markers when using a limited number of samples or datasets as input.

Spattern has yet to be validated on real spatial transcriptomics data. Furthermore, it will probably have to rely on metagene signal, both to reduce signal sparsity and to capture the co-presence of multiple populations at the same time. Consequently, its results will depend heavily on the appropriate selection of relevant metagenes.

Pooling the reference cells/spots from different samples strengthens the accuracy of FastCNV results; when this is not feasible, FastCNV might show a reduced sensitivity.

The draft of a unified atlas of lung cancer cells proposed here is designed to improve the deconvolution of bulk and spatial transcriptome of lung tumours. It needs a benchmark for this application, notably because the abundance of RNA greatly varies from one cell type to another, which translates into systematic bias in the cell type proportions estimated by deconvolution.

4.2 Conclusion

I contribute here a framework for characterizing intra-tumour heterogeneity of lung cancers, using tumour transcriptomics data. This framework includes three methods and a unified, revisited, lung cancer cell atlas.

My main methodological contribution concerns the SCherlock algorithm for the identification of cell type specific markers using scRNA-seq data. It is motivated by the fact that existing methods tend to favour sensitivity at the single-cell level over specificity. This strategy, which makes sense at the single-cell level, is very problematic for bulk/spatial transcriptome deconvolution. It generates highly correlated markers across different cell types, resulting in unreliable deconvolution results. A side effect of this strategy is also that it doesn't penalize spurious cell types, related to hierarchies of cell types that are too granular or not enough granular. SCherlock has no real competitors, to my knowledge, as existing methods are not pursuing the same goal (they look for markers to be used in single-cell data context).

The contribution regarding the FastCNV algorithm, to infer Copy Number Variations from single-cell/spatial transcriptome (tumour) data, is less original and impactful than SCherlock, as good equivalents exist. Nonetheless, the algorithm is more sensitive and much faster than its competitors. This gain in sensitivity is useful to better identify cells/spots with tumour content, which

is not always an easy task.

My last methodological contribution concerns the Spattern method, for the identification of spatial patterns in spatial transcriptome data. It still needs further benchmarking. While the underlying idea is very common in image analysis (pattern recognition is an active field of research since a long time), it is not yet widely used for spatial transcriptomics data analysis. Helping to classify the shape of various signals, notably those related to cell communities, might prove useful to interpret the underlying biology.

The draft of a unified lung cancer cell atlas, proposed here, takes advantage of the SCherlock and FastCNV methods to properly identify cell types and their markers. Thanks to SCherlock, this new atlas comes with a granularity of cell types that is amenable to deconvolution.

While this work delivers a framework for the characterisation of intra-tumour heterogeneity of lung cancers, the proposed methods are intended to be applied in a much broader framework for data analysis of single-cell and spatial transcriptomics data.

4.3 Perspectives

From the lung cancer cell atlas presented here, a direct prospect is its use as a reference to deconvolve the bulk and spatial transcriptome of lung tumours in order to establish their cellular composition. This first step should enable us to directly measure the degree of intra-tumour heterogeneity of tumours, and possible co-occurrences between certain tumour cell populations and certain immune and stromal microenvironments. These cell communities, made up of tumour/immune/stromal cells, are of direct interest in understanding the biology of lung tumours. They may also provide important information for predicting the prognosis and/or the response to treatments.

Bibliography

(*1*) Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: a cancer journal for clinicians 71*, 209–249.

(*2*) Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell 144*, 646–674.

(*3*) Amin, A. R., Karpowicz, P. A., Carey, T. E., Arbiser, J., Nahta, R., Chen, Z. G., Dong, J.-T., Kucuk, O., Khan, G. N., Huang, G. S., Mi, S., Lee, H.-Y., Reichrath, J., Honoki, K., Georgakilas, A. G., Amedei, A., Amin, A., Helferich, B., Boosani, C. S., Ciriolo, M. R., Chen, S., Mohammed, S. I., Azmi, A. S., Keith, W. N., Bhakta, D., Halicka, D., Niccolai, E., Fujii, H., Aquilano, K., Ashraf, S. S., Nowsheen, S., Yang, X., Bilsland, A., and Shin, D. M. (2015). Evasion of anti-growth signaling: A key step in tumorigenesis and potential target for treatment and prophylaxis by natural compounds. *Seminars in Cancer Biology 35*, S55–S77.

(*4*) Chen, X., Zhang, T., Su, W., Dou, Z., Zhao, D., Jin, X., Lei, H., Wang, J., Xie, X., Cheng, B., Li, Q., Zhang, H., and Di, C. (2022). Mutant p53 in cancer: from molecular mechanism to therapeutic modulation. *Cell Death & Disease 13*, 974.

(*5*) Fares, J., Fares, M. Y., Khachfe, H. H., Salhab, H. A., and Fares, Y. (2020). Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduction and Targeted Therapy 5*, 28.

(*6*) Liberti, M. V., and Locasale, J. W. (2016). The Warburg Effect: How Does it Benefit Cancer Cells? *Trends in Biochemical Sciences 41*, 211–218.

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

(*8*) Prieto García, C. USP28 reguliert Plattenepithelzell-Onkogenese und DNA-Reparatur über ΔNp63-DeubiquitinierungUSP28 regulates Squamous cell oncogenesis and DNA repair via ΔNp63 deubiquitination, Ph.D. Thesis, Universität Würzburg, 2022.

(*9*) Nia, H. T., Munn, L. L., and Jain, R. K. (2020). Physical traits of cancer. *Science 370*, eaaz0868.

(*10*) Rouvière, O., Melodelima, C., Hoang Dinh, A., Bratan, F., Pagnoux, G., Sanzalone, T., Crouzet, S., Colombel, M., Mège-Lechevallier, F., and Souchon, R. (2017). Stiffness of benign and malignant prostate tissue measured by shear-wave elastography: a preliminary study. *European Radiology 27*, 1858–1866.

(*11*) Carrara, S., Di Leo, M., Grizzi, F., Correale, L., Rahal, D., Anderloni, A., Auriemma, F., Fugazza, A., Preatoni, P., Maselli, R., Hassan, C., Finati, E., Mangiavillano, B., and Repici, A. (2018). EUS elastography (strain ratio) and fractal-based quantitative analysis for the diagnosis of solid pancreatic lesions. *Gastrointestinal Endoscopy 87*, 1464–1473.

(*12*) Shahryari, M., Tzschätzsch, H., Guo, J., Marticorena Garcia, S. R., Böning, G., Fehrenbach, U., Stencel, L., Asbach, P., Hamm, B., Käs, J. A., Braun, J., Denecke, T., and Sack, I. (2019). Tomoelastography Distinguishes Noninvasively between Benign and Malignant Liver Lesions. *Cancer Research 79*, 5704–5710.

(*13*) Piñeiro Fernández, J., Luddy, K. A., Harmon, C., and O'Farrelly, C. (2019). Hepatic Tumor Microenvironments and Effects on NK Cell Phenotype and Function. *International Journal of Molecular Sciences 20*, 4131.

(*14*) Chmiel, P., Rychcik-Pazyrska, P., and Stec, R. (2023). Defining Tumor Microenvironment as a Possible Target for Effective GEP-NENs Immunotherapy—A Systematic Review. *Cancers 15*, 5232.

(*15*) Yang, D., Liu, J., Qian, H., and Zhuang, Q. (2023). Cancer-associated fibroblasts: from basic science to anticancer therapy. *Experimental & Molecular Medicine 55*, 1322–1332.

(*16*) De Visser, K. E., and Joyce, J. A. (2023). The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. *Cancer Cell 41*, 374–403.

(*17*) Philip, M., and Schietinger, A. (2022). CD8+ T cell differentiation and dysfunction in cancer. *Nature Reviews. Immunology 22*, 209–223.

(*18*) Polanczyk, M. J., Walker, E., Haley, D., Guerrouahen, B. S., and Akporiaye, E. T. (2019). Blockade of TGF-β signaling to enhance the antitumor response is accompanied by dysregulation of the functional activity of CD4+CD25+Foxp3+ and CD4+CD25−Foxp3+ T cells. *Journal of Translational Medicine 17*, 219.

(*19*) Togashi, Y., Shitara, K., and Nishikawa, H. (2019). Regulatory T cells in cancer immunosuppression - implications for anticancer therapy. *Nature Reviews. Clinical Oncology 16*, 356–371.

(*20*) Dagogo-Jack, I., Valiev, I., Kotlov, N., Belozerova, A., Lopareva, A., Butusova, A., Samarina, N., Boyko, A., Xiang, Z., Johnson, M., Degryse, S., Keane, F. K., Sequist, L. V., Lanuti, M., Fowler, N., Mino-Kenudson, M., and Bagaev, A. (2023). B-Cell Infiltrate in the Tumor Microenvironment Is Associated With Improved Survival in Resected Lung Adenocarcinoma. *JTO Clinical and Research Reports 4*, 100527.

(*21*) Weng, Y., Yuan, J., Cui, X., Wang, J., Chen, H., Xu, L., Chen, X., Peng, M., and Song, Q. (2024). The impact of tertiary lymphoid structures on tumor prognosis and the immune microenvironment in non-small cell lung cancer. *Scientific Reports 14*, 16246.

(*22*) Zhang, S., Liu, W., Hu, B., Wang, P., Lv, X., Chen, S., and Shao, Z. (2020). Prognostic Significance of Tumor-Infiltrating Natural Killer Cells in Solid Tumors: A Systematic Review and Meta-Analysis. *Frontiers in Immunology 11*, 1242.

(*23*) DeNardo, D. G., and Ruffell, B. (2019). Macrophages as regulators of tumour immunity and immunotherapy. *Nature Reviews Immunology 19*, 369–382.

(*24*) Veglia, F., Sanseviero, E., and Gabrilovich, D. I. (2021). Myeloid-derived suppressor cells in the era of increasing myeloid cell diversity. *Nature Reviews Immunology 21*, 485–498.

(*25*) Hato, L., Vizcay, A., Eguren, I., Pérez-Gracia, J. L., Rodríguez, J., Gállego Pérez-Larraya, J., Sarobe, P., Inogés, S., Díaz De Cerio, A. L., and Santisteban, M. (2024). Dendritic Cells in Cancer Immunology and Immunotherapy. *Cancers 16*, 981.

(*26*) Syn, N. L.-X., Yong, W.-P., Goh, B.-C., and Lee, S.-C. (2016). Evolving landscape of tumor molecular profiling for personalized cancer therapy: a comprehensive review. *Expert Opinion on Drug Metabolism & Toxicology 12*, 911–922.

(*27*) Marusyk, A., Janiszewska, M., and Polyak, K. (2020). Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. *Cancer Cell 37*, 471–484.

(*28*) Cairns, J. (1975). Mutation selection and the natural history of cancer. *Nature 255*, 197–200.

(*29*) Nowell, P. C. (1976). The Clonal Evolution of Tumor Cell Populations: Acquired genetic lability permits stepwise selection of variant sublines and underlies tumor progression. *Science 194*, 23–28.

(*30*) Gerstung, M. et al. (2020). The evolutionary history of 2,658 cancers. *Nature 578*, 122–128.

(*31*) Maley, C. C., Galipeau, P. C., Finley, J. C., Wongsurawat, V. J., Li, X., Sanchez, C. A., Paulson, T. G., Blount, P. L., Risques, R.-A., Rabinovitch, P. S., and Reid, B. J. (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genetics 38*, 468–473.

(*32*) Mroz, E. A., Tward, A. D., Pickering, C. R., Myers, J. N., Ferris, R. L., and Rocco, J. W. (2013). High intratumor genetic heterogeneity is related to worse outcome in patients with head and neck squamous cell carcinoma. *Cancer 119*, 3034–3042.

(*33*) Dessert, D. Clonal Evolution Model, Pancalogy.com, 2015.

(*34*) Davies, A., Zoubeidi, A., Beltran, H., and Selth, L. A. (2023). The Transcriptional and Epigenetic Landscape of Cancer Cell Lineage Plasticity. *Cancer Discovery 13*, 1771–1788.

(*35*) Rubin, M. A., Bristow, R. G., Thienger, P. D., Dive, C., and Imielinski, M. (2020). Impact of Lineage Plasticity to and from a Neuroendocrine Phenotype on Progression and Response in Prostate and Lung Cancers. *Molecular Cell 80*, 562–577.

(*36*) Zhao, Y., Fang, L. T., Shen, T.-w., Choudhari, S., Talsania, K., Chen, X., Shetty, J., Kriga, Y., Tran, B., Zhu, B., Chen, Z., Chen, W., Wang, C., Jaeger, E., Meerzaman, D., Lu, C., Idler, K., Ren, L., Zheng, Y., Shi, L., Petitjean, V., Sultan, M., Hung, T., Peters, E., Drabek, J., Vojta, P., Maestro, R., Gasparotto, D., Kõks, S., Reimann, E., Scherer, A., Nordlund, J., Liljedahl, U., Foox, J., Mason, C. E., Xiao, C., Hong, H., and Xiao, W. (2021). Whole genome and exome sequencing reference datasets from a multi-center and cross-platform benchmark study. *Scientific Data 8*, 296.

(*37*) Kwon, Y. W., Jo, H.-S., Bae, S., Seo, Y., Song, P., Song, M., and Yoon, J. H. (2021). Application of Proteomics in Cancer: Recent Trends and Approaches for Biomarkers Discovery. *Frontiers in Medicine 8*, 747333.

(*38*) Timp, W., and Timp, G. (2020). Beyond mass spectrometry, the next step in proteomics. *Science Advances 6*, eaax8978.

(*39*) Wik, L., Nordberg, N., Broberg, J., Björkesten, J., Assarsson, E., Henriksson, S., Grundberg, I., Pettersson, E., Westerberg, C., Liljeroth, E., Falck, A., and Lundberg, M. (2021). Proximity Extension Assay in Combination with Next-Generation Sequencing for High-throughput Proteome-wide Analysis. *Molecular & Cellular Proteomics 20*, 100168.

(*40*) Hirahara, K., Vahedi, G., Ghoreschi, K., Yang, X.-P., Nakayamada, S., Kanno, Y., O'Shea, J. J., and Laurence, A. (2011). Helper T-cell differentiation and plasticity: insights from epigenetics. *Immunology 134*, 235–245.

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

(*42*) Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. *Science 316*, 1497–1502.

(*43*) Buccitelli, C., and Selbach, M. (2020). mRNAs, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics 21*, 630–644.

(*44*) Supplitt, S., Karpinski, P., Sasiadek, M., and Laczmanska, I. (2021). Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. *International Journal of Molecular Sciences 22*, 1422.

(*45*) Tong, M., Deng, Z., Yang, M., Xu, C., Zhang, X., Zhang, Q., Liao, Y., Deng, X., Lv, D., Zhang, X., Zhang, Y., Li, P., Song, L., Wang, B., Al-Dherasi, A., Li, Z., and Liu, Q. (2018). Transcriptomic but not genomic variability confers phenotype of breast cancer stem cells. *Cancer Communications 38*, 1–16.

(*46*) Tang, M., Abbas, H. A., Negrao, M. V., Ramineni, M., Hu, X., Hubert, S. M., Fujimoto, J., Reuben, A., Varghese, S., Zhang, J., Li, J., Chow, C.-W., Mao, X., Song, X., Lee, W.-C., Wu, J., Little, L., Gumbs, C., Behrens, C., Moran, C., Weissferdt, A., Lee, J. J., Sepesi, B., Swisher, S., Cheng, C., Kurie, J., Gibbons, D., Heymach, J. V., Wistuba, I. I., Futreal, P. A., Kalhor, N., and Zhang, J. (2021). The histologic phenotype of lung cancers is associated with transcriptomic features rather than genomic characteristics. *Nature Communications 12*, 7081.

(*47*) Ghatak, S., Mehrabi, S. F., Mehdawi, L. M., Satapathy, S. R., and Sjölander, A. (2022). Identification of a Novel Five-Gene Signature as a Prognostic and Diagnostic Biomarker in Colorectal Cancers. *International Journal of Molecular Sciences 23*, 793.

(*48*) Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology 17*, 13.

(*49*) Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.) 270*, 467–470.

(*50*) Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., and Ecker, J. R. (2008). Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell 133*, 523–536.

(*51*) Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA sequencing: the teenage years. *Nature Reviews Genetics 20*, 631–656.

(*52*) Wang, X., Park, J., Susztak, K., Zhang, N. R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications 10*, 380.

(*53*) Fan, J., Lyu, Y., Zhang, Q., Wang, X., Li, M., and Xiao, R. (2022). MuSiC2: cell-type deconvolution for multi-condition bulk RNA-seq data. *Briefings in Bioinformatics 23*, bbac430.

(*54*) Chu, T., Wang, Z., Pe'er, D., and Danko, C. G. (2022). Cell type and gene expression deconvolution with BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in oncology. *Nature Cancer 3*, 505–517.

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

(*56*) The TRACERx consortium, Rosenthal, R., Cadieux, E. L., Salgado, R., Bakir, M. A., Moore, D. A., Hiley, C. T., Lund, T., Tanić, M., Reading, J. L., Joshi, K., Henry, J. Y., Ghorani, E., Wilson, G. A., Birkbak, N. J., Jamal-Hanjani, M., Veeriah, S., Szallasi, Z., Loi, S., Hellmann, M. D., Feber, A., Chain, B., Herrero, J., Quezada, S. A., Demeulemeester, J., Van Loo, P., Beck, S., McGranahan, N., and Swanton, C. (2019). Neoantigen-directed immune escape in lung cancer evolution. *Nature 567*, 479–485.

(*57*) Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., and Soboleva, A. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research 41*, D991–D995.

(*58*) Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., ur-Rehman, S., Saunders, G., Kandasamy, J., Caccamo, M., Leinonen, R., Vaughan, B., Laurent, T., Rowland, F., Marin-Garcia, P., Barker, J., Jokinen, P., Torres, A. C., De Argila, J. R., Llobet, O. M., Medina, I., Puy, M. S., Alberich, M., De La Torre, S., Navarro, A., Paschall, J., and Flicek, P. (2015). The European Genome-phenome Archive of human data consented for biomedical research. *Nature Genetics 47*, 692–695.

(*59*) Karlsson, A., Brunnström, H., Micke, P., Veerla, S., Mattsson, J., La Fleur, L., Botling, J., Jönsson, M., Reuterswärd, C., Planck, M., and Staaf, J. (2017). Gene Expression Profiling of Large Cell Lung Cancer Links Transcriptional Phenotypes to the New Histological WHO 2015 Classification. *Journal of Thoracic Oncology 12*, 1257–1267.

(*60*) Muraro, M. J., Dharmadhikari, G., Grün, D., Groen, N., Dielen, T., Jansen, E., van Gurp, L., Engelse, M. A., Carlotti, F., de Koning, E. J. P., and van Oudenaarden, A. (2016). A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems 3*, 385–394.e3.

(*61*) Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. (2017). Massively parallel digital transcriptional profiling of single cells. *Nature Communications 8*, 14049.

(*62*) Hwang, B., Lee, J. H., and Bang, D. (2018). Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & Molecular Medicine 50*, 1–14.

(*63*) Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J. C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C. P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T. N., Shalek, A., Shapiro, E., Sharma, P., Shin, J. W., Stegle, O., Stratton, M., Stubbington, M. J. T., Theis, F. J., Uhlen, M., Van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N., and Human Cell Atlas Meeting Participants (2017). The Human Cell Atlas. *eLife 6*, e27041.

(*64*) Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., Myung, P., Plikus, M. V., and Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. *Nature Communications 12*, 1088.

(*65*) Efremova, M., Vento-Tormo, M., Teichmann, S. A., and Vento-Tormo, R. (2020). CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nature Protocols 15*, 1484–1506.

(*66*) Browaeys, R., Saelens, W., and Saeys, Y. (2020). NicheNet: modeling intercellular communication by linking ligands to target genes. *Nature Methods 17*, 159–162.

(*67*) Noël, F., Massenet-Regad, L., Carmi-Levy, I., Cappuccio, A., Grandclaudon, M., Trichot, C., Kieffer, Y., Mechta-Grigoriou, F., and Soumelis, V. (2021). Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nature Communications 12*, 1089.

(*68*) Genomics, 1., and Hindson, B. Single-Cell Explorations Are Underway, and Spatial Is the Next Frontier, 2019.

(*69*) Junker, J. P., Noël, E., Guryev, V., Peterson, K. A., Shah, G., Huisken, J., McMahon, A. P., Berezikov, E., Bakkers, J., and van Oudenaarden, A. (2014). Genome-wide RNA Tomography in the Zebrafish Embryo. *Cell 159*, 662–675.

(*70*) Merritt, C. R., Ong, G. T., Church, S. E., Barker, K., Danaher, P., Geiss, G., Hoang, M., Jung, J., Liang, Y., McKay-Fleisch, J., Nguyen, K., Norgaard, Z., Sorg, K., Sprague, I., Warren, C., Warren, S., Webster, P. J., Zhou, Z., Zollinger, D. R., Dunaway, D. L., Mills, G. B., and Beechem, J. M. (2020). Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nature Biotechnology 38*, 586–599.

(*71*) Femino, A. M., Fay, F. S., Fogarty, K., and Singer, R. H. (1998). Visualization of Single RNA Transcripts in Situ. *Science 280*, 585–590.

(*72*) Shah, S., Lubeck, E., Zhou, W., and Cai, L. (2016). In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron 92*, 342–357.

(*73*) Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., and Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science 348*, aaa6090.

(*74*) Ke, R., Mignardi, M., Pacureanu, A., Svedlund, J., Botling, J., Wählby, C., and Nilsson, M. (2013). In situ sequencing for RNA analysis in preserved tissue and cells. *Nature Methods 10*, 857–860.

(*75*) Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G. P., Bava, F.-A., and Deisseroth, K. (2018). Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science 361*, eaat5691.

(*76*) Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsson, S., Codeluppi, S., Borg, Å., Pontén, F., Costea, P. I., Sahlén, P., Mulder, J., Bergmann, O., Lundeberg, J., and Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science 353*, 78–82.

(*77*) Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., Welch, J., Chen, L. M., Chen, F., and Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science 363*, 1463–1467.

(*78*) Rao, A., Barkley, D., França, G. S., and Yanai, I. (2021). Exploring tissue architecture using spatial transcriptomics. *Nature 596*, 211–220.

(*79*) Ben-Chetrit, N., Niu, X., Swett, A. D., Sotelo, J., Jiao, M. S., Stewart, C. M., Potenski, C., Mielinis, P., Roelli, P., Stoeckius, M., and Landau, D. A. (2023). Integration of whole transcriptome spatial profiling with protein markers. *Nature Biotechnology 41*, 788–793.

(*80*) Oliveira, M. F., Romero, J. P., Chung, M., Williams, S., Gottscho, A. D., Gupta, A., Pilipauskas, S. E., Mohabbat, S., Raman, N., Sukovich, D., Patterson, D., Visium HD Development Team, and Taylor, S. E. B. Characterization of immune cell populations in the tumor microenvironment of colorectal cancer using high definition spatial profiling, en, 2024.

(*81*) 10x Genomics Visium for Fresh Frozen samples, 2023.

(*82*) Barnell, E. K., Wurtzler, E. M., La Rocca, J., Fitzgerald, T., Petrone, J., Hao, Y., Kang, Y., Holmes, F. L., and Lieberman, D. A. (2023). Multitarget Stool RNA Test for Colorectal Cancer Screening. *JAMA 330*, 1760.

(*83*) Reicherter, J. M. In *Fundamentals of Polygraph Practice*; Elsevier: 2015, pp 29–60.

(*84*) Heinen, N., Klöhn, M., Steinmann, E., and Pfaender, S. (2021). In Vitro Lung Models and Their Application to Study SARS-CoV-2 Pathogenesis and Disease. *Viruses 13*, 792.

(*85*) Nicholson, A. G., Tsao, M. S., Beasley, M. B., Borczuk, A. C., Brambilla, E., Cooper, W. A., Dacic, S., Jain, D., Kerr, K. M., Lantuejoul, S., Noguchi, M., Papotti, M., Rekhtman, N., Scagliotti, G., Van Schil, P., Sholl, L., Yatabe, Y., Yoshida, A., and Travis, W. D. (2022). The 2021 WHO Classification of Lung Tumors: Impact of Advances Since 2015. *Journal of Thoracic Oncology 17*, 362–387.

(*86*) Zhang, Y., Vaccarella, S., Morgan, E., Li, M., Etxeberria, J., Chokunonga, E., Manraj, S. S., Kamate, B., Omonisi, A., and Bray, F. (2023). Global variations in lung cancer incidence by histological subtype in 2020: a population-based study. *The Lancet Oncology 24*, 1206–1218.

(*87*) Fernandez, F. G., and Battafarano, R. J. (2006). Large-Cell Neuroendocrine Carcinoma of the Lung: An Aggressive Neuroendocrine Lung Cancer. *Seminars in Thoracic and Cardiovascular Surgery 18*, 206–210.

(*88*) Brambilla, C., Laffaire, J., Lantuejoul, S., Moro-Sibilot, D., Mignotte, H., Arbib, F., Toffart, A.-C., Petel, F., Hainaut, P., Rousseaux, S., Khochbin, S., De Reyniès, A., and Brambilla, E. (2014). Lung Squamous Cell Carcinomas with Basaloid Histology Represent a Specific Molecular Entity. *Clinical Cancer Research 20*, 5777–5786.

(*89*) Ruffini, E., Rena, O., Oliaro, A., Filosso, P., Bongiovanni, M., Arslanian, A., Papalia, E., and Maggi, G. (2002). Lung tumors with mixed histologic pattern. Clinico-pathologic characteristics and prognostic significance. *European Journal of Cardio-Thoracic Surgery 22*, 701–707.

(*90*) Deng, P., Hu, C., Zhou, L., Li, Y., and Huang, L. (2013). Clinical characteristics and prognostic significance of 92 cases of patients with primary mixed-histology lung cancer. *Molecular and Clinical Oncology 1*, 863–868.

(*91*) Zeng, J., Ding, X., Ding, J., and Wang, X. (2023). Histological transformation into SCLC: An important resistance mechanism of NSCLC upon immunotherapy. *Frontiers in Immunology 14*, 1275957.

(*92*) Nagasaka, M., Pansare, R. S., Abdulfatah, E., Guan, H., Tranchida, P., and Gadgeel, S. M. (2017). Histologic Transformation in NSCLC with PD-1 therapy. *Journal of Thoracic Oncology 12*, e133–e134.

(*93*) Mariniello, A., Righi, L., Morrone, A., Carnio, S., and Bironzo, P. (2022). Squamous cell histological transformation in a lung adenocarcinoma patient (hyper) progressing upon immunotherapy. *Tumori Journal 108*, NP15–NP19.

(*94*) Walser, T., Cui, X., Yanagawa, J., Lee, J. M., Heinrich, E., Lee, G., Sharma, S., and Dubinett, S. M. (2008). Smoking and lung cancer: the role of inflammation. *Proceedings of the American Thoracic Society 5*, 811–815.

(*95*) Tammemägi, M. C., Katki, H. A., Hocking, W. G., Church, T. R., Caporaso, N., Kvale, P. A., Chaturvedi, A. K., Silvestri, G. A., Riley, T. L., Commins, J., and Berg, C. D. (2013). Selection criteria for lung-cancer screening. *The New England Journal of Medicine 368*, 728–736.

(*96*) Kim, A.-S., Ko, H.-J., Kwon, J.-H., and Lee, J.-M. (2018). Exposure to Secondhand Smoke and Risk of Cancer in Never Smokers: A Meta-Analysis of Epidemiologic Studies. *International Journal of Environmental Research and Public Health 15*, 1981.

(*97*) Loomis, D., Grosse, Y., Lauby-Secretan, B., Ghissassi, F. E., Bouvard, V., Benbrahim-Tallaa, L., Guha, N., Baan, R., Mattock, H., and Straif, K. (2013). The carcinogenicity of outdoor air pollution. *The Lancet Oncology 14*, 1262–1263.

(*98*) *WHO Handbook on Indoor Radon: A Public Health Perspective*; WHO Guidelines Approved by the Guidelines Review Committee; World Health Organization: Geneva, 2009.

(*99*) Field, R. W., and Withers, B. L. (2012). Occupational and environmental causes of lung cancer. *Clinics in Chest Medicine 33*, 681–703.

(*100*) Couraud, S., Zalcman, G., Milleron, B., Morin, F., and Souquet, P.-J. (2012). Lung cancer in never smokers – A review. *European Journal of Cancer 48*, 1299–1311.

(*101*) Carrot-Zhang, J., Soca-Chafre, G., Patterson, N., Thorner, A. R., Nag, A., Watson, J., Genovese, G., Rodriguez, J., Gelbard, M. K., Corrales-Rodriguez, L., Mitsuishi, Y., Ha, G., Campbell, J. D., Oxnard, G. R., Arrieta, O., Cardona, A. F., Gusev, A., and Meyerson, M. (2021). Genetic Ancestry Contributes to Somatic Mutations in Lung Cancers from Admixed Latin American Populations. *Cancer Discovery 11*, 591–598.

(*102*) McLaughlin, J., Berkman, J., and Nana-Sinkam, P. (2023). Targeted therapies in non-small cell lung cancer: present and future. *Faculty Reviews 12*, 22.

(*103*) Horn, L., Spigel, D. R., Vokes, E. E., Holgado, E., Ready, N., Steins, M., Poddubskaya, E., Borghaei, H., Felip, E., Paz-Ares, L., Pluzanski, A., Reckamp, K. L., Burgio, M. A., Kohlhäeufl, M., Waterhouse, D., Barlesi, F., Antonia, S., Arrieta, O., Fayette, J., Crinò, L., Rizvi, N., Reck, M., Hellmann, M. D., Geese, W. J., Li, A., Blackwood-Chirchir, A., Healey, D., Brahmer, J., and Eberhardt, W. E. (2017). Nivolumab Versus Docetaxel in Previously Treated Patients With Advanced Non–Small-Cell Lung Cancer: Two-Year Outcomes From Two Randomized, Open-Label, Phase III Trials (CheckMate 017 and CheckMate 057). *Journal of Clinical Oncology 35*, 3924–3933.

(*104*) Garon, E. B., Rizvi, N. A., Hui, R., Leighl, N., Balmanoukian, A. S., Eder, J. P., Patnaik, A., Aggarwal, C., Gubens, M., Horn, L., Carcereny, E., Ahn, M.-J., Felip, E., Lee, J.-S., Hellmann, M. D., Hamid, O., Goldman, J. W., Soria, J.-C., Dolled-Filhart, M., Rutledge, R. Z., Zhang, J., Lunceford, J. K., Rangwala, R., Lubiniecki, G. M., Roach, C., Emancipator, K., and Gandhi, L. (2015). Pembrolizumab for the Treatment of Non–Small-Cell Lung Cancer. *New England Journal of Medicine 372*, 2018–2028.

(*105*) Rittmeyer, A., Barlesi, F., Waterkamp, D., Park, K., Ciardiello, F., von Pawel, J., Gadgeel, S. M., Hida, T., Kowalski, D. M., Dols, M. C., Cortinovis, D. L., Leach, J., Polikoff, J., Barrios, C., Kabbinavar, F., Frontera, O. A., De Marinis, F., Turna, H., Lee, J.-S., Ballinger, M., Kowanetz, M., He, P., Chen, D. S., Sandler, A., Gandara, D. R., and OAK Study Group (2017). Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. *Lancet (London, England) 389*, 255–265.

(*106*) Reck, M., Rodríguez-Abreu, D., Robinson, A. G., Hui, R., Csőszi, T., Fülöp, A., Gottfried, M., Peled, N., Tafreshi, A., Cuffe, S., O'Brien, M., Rao, S., Hotta, K., Leiby, M. A., Lubiniecki, G. M., Shentu, Y., Rangwala, R., and Brahmer, J. R. (2016). Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. *New England Journal of Medicine 375*, 1823–1833.

(*107*) Herbst, R. S., Giaccone, G., De Marinis, F., Reinmuth, N., Vergnenegre, A., Barrios, C. H., Morise, M., Felip, E., Andric, Z., Geater, S., Özgüroğlu, M., Zou, W., Sandler, A., Enquist, I., Komatsubara, K., Deng, Y., Kuriki, H., Wen, X., McCleland, M., Mocci, S., Jassem, J., and Spigel, D. R. (2020). Atezolizumab for First-Line Treatment of PD-L1–Selected Patients with NSCLC. *New England Journal of Medicine 383*, 1328–1339.

(*108*) Borghaei, H., Langer, C. J., Paz-Ares, L., Rodríguez-Abreu, D., Halmos, B., Garassino, M. C., Houghton, B., Kurata, T., Cheng, Y., Lin, J., Pietanza, M. C., Piperdi, B., and Gadgeel, S. M. (2020). Pembrolizumab plus chemotherapy versus chemotherapy alone in patients with advanced non–small cell lung cancer without tumor PD-L1 expression: A pooled analysis of 3 randomized controlled trials. *Cancer 126*, 4867–4877.

(*109*) Hellmann, M. D., Ciuleanu, T.-E., Pluzanski, A., Lee, J. S., Otterson, G. A., Audigier-Valette, C., Minenza, E., Linardou, H., Burgers, S., Salman, P., Borghaei, H., Ramalingam, S. S., Brahmer, J., Reck, M., O'Byrne, K. J., Geese, W. J., Green, G., Chang, H., Szustakowski, J., Bhagavatheeswaran, P., Healey, D., Fu, Y., Nathan, F., and Paz-Ares, L. (2018). Nivolumab plus Ipilimumab in Lung Cancer with a High Tumor Mutational Burden. *The New England Journal of Medicine 378*, 2093–2104.

(*110*) McNamee, N., da Silva, I. P., Nagrial, A., and Gao, B. (2023). Small-Cell Lung Cancer-An Update on Targeted and Immunotherapies. *International Journal of Molecular Sciences 24*, 8129.

(*111*) Riano, I., Patel, S. R., Liu, S. V., and Duma, N. (2021). Evidence to Date: Evaluating Pembrolizumab in the Treatment of Extensive-Stage Small-Cell Lung Cancer. *Clinics and Practice 11*, 441–454.

(*112*) Ready, N., Farago, A. F., de Braud, F., Atmaca, A., Hellmann, M. D., Schneider, J. G., Spigel, D. R., Moreno, V., Chau, I., Hann, C. L., Eder, J. P., Steele, N. L., Pieters, A., Fairchild, J., and Antonia, S. J. (2019). Third-Line Nivolumab Monotherapy in Recurrent SCLC: CheckMate 032. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer 14*, 237–244.

(*113*) Ready, N. E., Ott, P. A., Hellmann, M. D., Zugazagoitia, J., Hann, C. L., de Braud, F., Antonia, S. J., Ascierto, P. A., Moreno, V., Atmaca, A., Salvagni, S., Taylor, M., Amin, A., Camidge, D. R., Horn, L., Calvo, E., Li, A., Lin, W. H., Callahan, M. K., and Spigel, D. R. (2020). Nivolumab Monotherapy and Nivolumab Plus Ipilimumab in Recurrent Small Cell Lung Cancer: Results From the CheckMate 032 Randomized Cohort. *Journal of Thoracic Oncology: Official Publication of the International Association for the Study of Lung Cancer 15*, 426–435.

(*114*) Horn, L., Mansfield, A. S., Szczęsna, A., Havel, L., Krzakowski, M., Hochmair, M. J., Huemer, F., Losonczy, G., Johnson, M. L., Nishio, M., Reck, M., Mok, T., Lam, S., Shames, D. S., Liu, J., Ding, B., Lopez-Chavez, A., Kabbinavar, F., Lin, W., Sandler, A., and Liu, S. V. (2018). First-Line Atezolizumab plus Chemotherapy in Extensive-Stage Small-Cell Lung Cancer. *New England Journal of Medicine 379*, 2220–2229.

(*115*) Paz-Ares, L. et al. (2019). Durvalumab plus platinum–etoposide versus platinum–etoposide in first-line treatment of extensive-stage small-cell lung cancer (CASPIAN): a randomised, controlled, open-label, phase 3 trial. *The Lancet 394*, 1929–1939.

(*116*) Rudin, C. M., Awad, M. M., Navarro, A., Gottfried, M., Peters, S., Csőszi, T., Cheema, P. K., Rodriguez-Abreu, D., Wollner, M., Yang, J. C.-H., Mazieres, J., Orlandi, F. J., Luft, A., Gümüş, M., Kato, T., Kalemkerian, G. P., Luo, Y., Ebiana, V., Pietanza, M. C., Kim, H. R., and KEYNOTE-604 Investigators (2020). Pembrolizumab or Placebo Plus Etoposide and Platinum as First-Line Therapy for Extensive-Stage Small-Cell Lung Cancer: Randomized, Double-Blind, Phase III KEYNOTE-604 Study. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology 38*, 2369–2379.

(*117*) Derks, J. L., Leblay, N., Thunnissen, E., van Suylen, R. J., den Bakker, M., Groen, H. J. M., Smit, E. F., Damhuis, R., van den Broek, E. C., Charbrier, A., Foll, M., McKay, J. D., Fernandez-Cuesta, L., Speel, E.-J. M., Dingemans, A.-M. C., and PALGA-Group (2018). Molecular Subtypes of Pulmonary Large-cell Neuroendocrine Carcinoma Predict Chemotherapy Treatment Outcome. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 24*, 33–42.

(*118*) The National Library of Medicine (NLM) ClinicalTrials.gov.

(*119*) Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., Loda, M., Weber, G., Mark, E. J., Lander, E. S., Wong, W., Johnson, B. E., Golub, T. R., Sugarbaker, D. J., and Meyerson, M. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences 98*, 13790–13795.

(*120*) Beer, D. G., Kardia, S. L., Huang, C.-C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G., Lizyness, M. L., Kuick, R., Hayasaka, S., Taylor, J. M., Iannettoni, M. D., Orringer, M. B., and Hanash, S. (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. *Nature Medicine 8*, 816–824.

(*121*) Endoh, H., Tomida, S., Yatabe, Y., Konishi, H., Osada, H., Tajima, K., Kuwano, H., Takahashi, T., and Mitsudomi, T. (2004). Prognostic Model of Pulmonary Adenocarcinoma by Expression Profiling of Eight Genes As Determined by Quantitative Real-Time Reverse Transcriptase Polymerase Chain Reaction. *Journal of Clinical Oncology 22*, 811–819.

(*122*) Hayes, D. N., Monti, S., Parmigiani, G., Gilks, C. B., Naoki, K., Bhattacharjee, A., Socinski, M. A., Perou, C., and Meyerson, M. (2006). Gene Expression Profiling Reveals Reproducible Human Lung Adenocarcinoma Subtypes in Multiple Independent Patient Cohorts. *Journal of Clinical Oncology 24*, 5079–5090.

(*123*) The Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature 511*, 543–550.

(*124*) Wilkerson, M. D., Yin, X., Walter, V., Zhao, N., Cabanski, C. R., Hayward, M. C., Miller, C. R., Socinski, M. A., Parsons, A. M., Thorne, L. B., Haithcock, B. E., Veeramachaneni, N. K., Funkhouser, W. K., Randell, S. H., Bernard, P. S., Perou, C. M., and Hayes, D. N. (2012). Differential Pathogenesis of Lung Adenocarcinoma Subtypes Involving Sequence Mutations, Copy Number, Chromosomal Instability, and Methylation. *PLoS ONE 7*, ed. by Deutsch, E., e36530.

(*125*) Ruiz-Cordero, R., Ma, J., Khanna, A., Lyons, G., Rinsurongkawong, W., Bassett, R., Guo, M., Routbort, M. J., Zhang, J., Skoulidis, F., Heymach, J., Roarty, E. B., Tang, Z., Medeiros, L. J., Patel, K. P., Luthra, R., and Roy-Chowdhuri, S. (2020). Simplified molecular classification of lung adenocarcinomas based on EGFR, KRAS, and TP53 mutations. *BMC Cancer 20*, 83.

(*126*) Shinjo, K., Okamoto, Y., An, B., Yokoyama, T., Takeuchi, I., Fujii, M., Osada, H., Usami, N., Hasegawa, Y., Ito, H., Hida, T., Fujimoto, N., Kishimoto, T., Sekido, Y., and Kondo, Y. (2012). Integrated analysis of genetic and epigenetic alterations reveals CpG island methylator phenotype associated with distinct clinical characters of lung adenocarcinoma. *Carcinogenesis 33*, 1277–1285.

(*127*) Sproul, D., Kitchen, R. R., Nestor, C. E., Dixon, J. M., Sims, A. H., Harrison, D. J., Ramsahoye, B. H., and Meehan, R. R. (2012). Tissue of origin determines cancer-associated CpG island promoter hypermethylation patterns. *Genome Biology 13*, R84.

(*128*) Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J. G., Baylin, S. B., and Issa, J.-P. J. (1999). CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences 96*, 8681–8686.

(*129*) Shi, S., Xu, M., and Xi, Y. (2020). Molecular subtypes based on DNA promoter methylation predict prognosis in lung adenocarcinoma patients. *Aging 12*, 23917–23930.

(*130*) Xu, D., Li, C., Zhang, Y., and Zhang, J. (2022). DNA methylation molecular subtypes for prognosis prediction in lung adenocarcinoma. *BMC Pulmonary Medicine 22*, 133.

(*131*) Soltis, A. R., Bateman, N. W., Liu, J., Nguyen, T., Franks, T. J., Zhang, X., Dalgard, C. L., Viollet, C., Somiari, S., Yan, C., Zeman, K., Skinner, W. J., Lee, J. S., Pollard, H. B., Turner, C., Petricoin, E. F., Meerzaman, D., Conrads, T. P., Hu, H., Blackwell, R., Sukumar, G., Bacikova, D., Alba, C., McGrath, E., Polisetti, S., Tuck, M., Chiu, A., Peterson, G., Larson, C., Kvecher, L., Deyarmin, B., Kane, J., Miller, K., Conrads, K. A., Hood, B. L., Makohon-Moore, S. C., Abulez, T. S., Baldelli, E., Pierobon, M., Chen, Q.-r., Rodriguez, H., Hanlon, S. E., Soltis, A. R., Bateman, N. W., Liu, J., Nguyen, T., Franks, T. J., Zhang, X., Dalgard, C. L., Viollet, C., Somiari, S., Yan, C., Zeman, K., Skinner, W. J., Lee, J. S., Pollard, H. B., Turner, C., Petricoin, E. F., Meerzaman, D., Conrads, T. P., Hu, H., Shriver, C. D., Moskaluk, C. A., Browning, R. F., Wilkerson, M. D., Shriver, C. D., Moskaluk, C. A., Browning, R. F., and Wilkerson, M. D. (2022). Proteogenomic analysis of lung adenocarcinoma reveals tumor heterogeneity, survival determinants, and therapeutically relevant pathways. *Cell Reports Medicine 3*, 100819.

(*132*) Chen, F., Zhang, Y., Parra, E., Rodriguez, J., Behrens, C., Akbani, R., Lu, Y., Kurie, J. M., Gibbons, D. L., Mills, G. B., Wistuba, I. I., and Creighton, C. J. (2017). Multiplatform-based molecular subtypes of non-small-cell lung cancer. *Oncogene 36*, 1384–1393.

(*133*) Wang, S., Liang, X., Guo, R., Gong, J., Zhong, X., Liu, Y., Wang, D., Hao, Y., and Hu, B. (2022). Identification of molecular subtypes in lung adenocarcinoma based on DNA methylation and gene expression profiling—a bioinformatic analysis. *Annals of Translational Medicine 10*, 882–882.

(*134*) Zhang, J., Li, Y., Dai, W., Tang, F., Wang, L., Wang, Z., Li, S., Ji, Q., Zhang, J., Liao, Z., Yu, J., Xu, Y., Gong, J., Hu, J., Li, J., Guo, X., He, F., Han, L., Gong, Y., Ouyang, W., Wang, Z., and Xie, C. (2024). Molecular classification reveals the sensitivity of lung adenocarcinoma to radiotherapy and immunotherapy: multi-omics clustering based on similarity network fusion. *Cancer Immunology, Immunotherapy 73*, 71.

(*135*) Inamura, K., Fujiwara, T., Hoshida, Y., Isagawa, T., Jones, M. H., Virtanen, C., Shimane, M., Satoh, Y., Okumura, S., Nakagawa, K., Tsuchiya, E., Ishikawa, S., Aburatani, H., Nomura, H., and Ishikawa, Y. (2005). Two subclasses of lung squamous cell carcinoma with different gene expression profiles and prognosis identified by hierarchical clustering and non-negative matrix factorization. *Oncogene 24*, 7105–7113.

(*136*) Larsen, J. E., Pavey, S. J., Passmore, L. H., Bowman, R., Clarke, B. E., Hayward, N. K., and Fong, K. M. (2006). Expression profiling defines a recurrence signature in lung squamous cell carcinoma. *Carcinogenesis 28*, 760–766.

(*137*) Raponi, M., Zhang, Y., Yu, J., Chen, G., Lee, G., Taylor, J. M., MacDonald, J., Thomas, D., Moskaluk, C., Wang, Y., and Beer, D. G. (2006). Gene Expression Signatures for Predicting Prognosis of Squamous Cell and Adenocarcinomas of the Lung. *Cancer Research 66*, 7466–7472.

(*138*) Wilkerson, M. D., Yin, X., Hoadley, K. A., Liu, Y., Hayward, M. C., Cabanski, C. R., Muldrew, K., Miller, C. R., Randell, S. H., Socinski, M. A., Parsons, A. M., Funkhouser, W. K., Lee, C. B., Roberts, P. J., Thorne, L., Bernard, P. S., Perou, C. M., and Hayes, D. N. (2010). Lung Squamous Cell Carcinoma mRNA Expression Subtypes Are Reproducible, Clinically Important, and Correspond to Normal Cell Types. *Clinical Cancer Research 16*, 4864–4875.

(*139*) Lee, W.-C., Diao, L., Wang, J., Zhang, J., Roarty, E. B., Varghese, S., Chow, C.-W., Fujimoto, J., Behrens, C., Cascone, T., Peng, W., Kalhor, N., Moran, C. A., Weissferdt, A., Johnson, F. M., William, W. N., Swisher, S. G., Lee, J. J., Hong, W. K., Heymach, J. V., Wistuba, I. I., Futreal, P. A., and Zhang, J. (2018). Multiregion gene expression profiling

reveals heterogeneity in molecular subtypes and immunotherapy response signatures in lung cancer. *Modern Pathology 31*, 947–955.

(*140*) The Cancer Genome Atlas Research Network (2012). Comprehensive genomic characterization of squamous cell lung cancers. *Nature 489*, 519–525.

(*141*) Huang, G., Zhang, J., Gong, L., Liu, D., Wang, X., Chen, Y., and Guo, S. (2020). Specific Lung Squamous Cell Carcinoma Prognosis-Subtype Distinctions Based on DNA Methylation Patterns. *Medical Science Monitor 27*.

(*142*) Li, X.-S., Nie, K.-C., Zheng, Z.-H., Zhou, R.-S., Huang, Y.-S., Ye, Z.-J., He, F., and Tang, Y. (2021). Molecular subtypes based on DNA methylation predict prognosis in lung squamous cell carcinoma. *BMC Cancer 21*, 96.

(*143*) Wang, J., Sun, N., Kunzke, T., Shen, J., Zens, P., Prade, V. M., Feuchtinger, A., Berezowska, S., and Walch, A. (2023). Spatial metabolomics identifies distinct tumor-specific and stroma-specific subtypes in patients with lung squamous cell carcinoma. *npj Precision Oncology 7*, 114.

(*144*) Stewart, P. A., Welsh, E. A., Slebos, R. J. C., Fang, B., Izumi, V., Chambers, M., Zhang, G., Cen, L., Pettersson, F., Zhang, Y., Chen, Z., Cheng, C.-H., Thapa, R., Thompson, Z., Fellows, K. M., Francis, J. M., Saller, J. J., Mesa, T., Zhang, C., Yoder, S., DeNicola, G. M., Beg, A. A., Boyle, T. A., Teer, J. K., Ann Chen, Y., Koomen, J. M., Eschrich, S. A., and Haura, E. B. (2019). Proteogenomic landscape of squamous cell lung cancer. *Nature Communications 10*, 3578.

(*145*) Satpathy, S., Krug, K., Jean Beltran, P. M., Savage, S. R., Petralia, F., Kumar-Sinha, C., Dou, Y., Reva, B., Kane, M. H., Avanessian, S. C., Vasaikar, S. V., Krek, A., Lei, J. T., Jaehnig, E. J., Omelchenko, T., Geffen, Y., Bergstrom, E. J., Stathias, V., Christianson, K. E., Heiman, D. I., Cieslik, M. P., Cao, S., Song, X., Ji, J., Liu, W., Li, K., Wen, B., Li, Y., Gümüş, Z. H., Selvan, M. E., Soundararajan, R., Visal, T. H., Raso, M. G., Parra, E. R., Babur, Ö., Vats, P., Anand, S., Schraink, T., Cornwell, M., Rodrigues, F. M., Zhu, H., Mo, C.-K., Zhang, Y., Da Veiga Leprevost, F., Huang, C., Chinnaiyan, A. M., Wyczalkowski, M. A., Omenn, G. S., Newton, C. J., Schurer, S., Ruggles, K. V., Fenyö, D., Jewell, S. D., Thiagarajan, M., Mesri, M., Rodriguez, H., Mani, S. A., Udeshi, N. D., Getz, G., Suh, J., Li, Q. K., Hostetter, G., Paik, P. K., Dhanasekaran, S. M., Govindan, R., Ding, L., Robles, A. I., Clauser, K. R., Nesvizhskii, A. I., Wang, P., Carr, S. A., Zhang, B., Mani, D., Gillette, M. A., Green, A., Molinolo, A., Francis, A., Paulovich, A. G., Karnuta, A., Colaprico, A., Hindenach, B., Pruetz, B. L., Kubisa, B., Druker, B. J., Huynh, C., Goldthwaite, C. A., Birger, C., Kinsinger, C. R., Jones, C. D., Rohrer, D., Valley, D. R., Chan,

D. W., Chesla, D., Hansel, D., Ponomareva, E. V., Duffy, E., Burks, E., Schadt, E. E., Fedorov, E. S., An, E., Ding, F., Wilson, G. D., Batra, H., Zhang, H., Maas, J. E., Eschbacher, J., Ketchum, K. A., Rodland, K. D., Hoadley, K. A., Suzuki, K., Um, K. S., Qi, L., Bernard, L., Wiznerowicz, M., Wojtyś, M., Domagalski, M. J., Ellis, M. J., Dyer, M. A., Borucki, M., Anurag, M., Birrer, M. J., Xu, M., Krotevich, M., Roche, N., Edwards, N. J., Vatanian, N., Mucci, N. R., Maunganidze, N., Gabrovski, N., Potapova, O., Fadare, O., Grady, P., McGarvey, P. B., Hariharan, P., Thangudu, R. R., Montgomery, R., Pandurengan, R., Smith, R. D., Welsh, R. J., Mareedu, S., Payne, S. H., Cottingham, S., Singh, S., Tsang, S. X., Cai, S., Gabriel, S., Liu, T., Hiltke, T., Vashist, T., Bauer, T., Sovenko, V., Tourtellotte, W. G., Ma, W., Bocik, W., Hasan, W., Jing, X., Tang, X., Liao, Y., Yvonne, Shutack, Zhang, Z., and Hanhan, Z. (2021). A proteogenomic portrait of lung squamous cell carcinoma. *Cell 184*, 4348–4371.e40.

(*146*) Rudin, C. M., Durinck, S., Stawiski, E. W., Poirier, J. T., Modrusan, Z., Shames, D. S., Bergbower, E. A., Guan, Y., Shin, J., Guillory, J., Rivers, C. S., Foo, C. K., Bhatt, D., Stinson, J., Gnad, F., Haverty, P. M., Gentleman, R., Chaudhuri, S., Janakiraman, V., Jaiswal, B. S., Parikh, C., Yuan, W., Zhang, Z., Koeppen, H., Wu, T. D., Stern, H. M., Yauch, R. L., Huffman, K. E., Paskulin, D. D., Illei, P. B., Varella-Garcia, M., Gazdar, A. F., De Sauvage, F. J., Bourgon, R., Minna, J. D., Brock, M. V., and Seshagiri, S. (2012). Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nature Genetics 44*, 1111–1116.

(*147*) Peifer, M., Fernández-Cuesta, L., Sos, M. L., George, J., Seidel, D., Kasper, L. H., Plenker, D., Leenders, F., Sun, R., Zander, T., Menon, R., Koker, M., Dahmen, I., Müller, C., Di Cerbo, V., Schildhaus, H.-U., Altmüller, J., Baessmann, I., Becker, C., De Wilde, B., Vandesompele, J., Böhm, D., Ansén, S., Gabler, F., Wilkening, I., Heynck, S., Heuckmann, J. M., Lu, X., Carter, S. L., Cibulskis, K., Banerji, S., Getz, G., Park, K.-S., Rauh, D., Grütter, C., Fischer, M., Pasqualucci, L., Wright, G., Wainer, Z., Russell, P., Petersen, I., Chen, Y., Stoelben, E., Ludwig, C., Schnabel, P., Hoffmann, H., Muley, T., Brockmann, M., Engel-Riedel, W., Muscarella, L. A., Fazio, V. M., Groen, H., Timens, W., Sietsma, H., Thunnissen, E., Smit, E., Heideman, D. A. M., Snijders, P. J. F., Cappuzzo, F., Ligorio, C., Damiani, S., Field, J., Solberg, S., Brustugun, O. T., Lund-Iversen, M., Sänger, J., Clement, J. H., Soltermann, A., Moch, H., Weder, W., Solomon, B., Soria, J.-C., Validire, P., Besse, B., Brambilla, E., Brambilla, C., Lantuejoul, S., Lorimier, P., Schneider, P. M., Hallek, M., Pao, W., Meyerson, M., Sage, J., Shendure, J., Schneider, R., Büttner, R., Wolf, J., Nürnberg, P., Perner, S., Heukamp, L. C., Brindle, P. K., Haas, S., and Thomas, R. K.

(2012). Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nature Genetics 44*, 1104–1110.

(*148*) George, J., Lim, J. S., Jang, S. J., Cun, Y., Ozretić, L., Kong, G., Leenders, F., Lu, X., Fernández-Cuesta, L., Bosco, G., Müller, C., Dahmen, I., Jahchan, N. S., Park, K.-S., Yang, D., Karnezis, A. N., Vaka, D., Torres, A., Wang, M. S., Korbel, J. O., Menon, R., Chun, S.-M., Kim, D., Wilkerson, M., Hayes, N., Engelmann, D., Pützer, B., Bos, M., Michels, S., Vlasic, I., Seidel, D., Pinther, B., Schaub, P., Becker, C., Altmüller, J., Yokota, J., Kohno, T., Iwakawa, R., Tsuta, K., Noguchi, M., Muley, T., Hoffmann, H., Schnabel, P. A., Petersen, I., Chen, Y., Soltermann, A., Tischler, V., Choi, C.-m., Kim, Y.-H., Massion, P. P., Zou, Y., Jovanovic, D., Kontic, M., Wright, G. M., Russell, P. A., Solomon, B., Koch, I., Lindner, M., Muscarella, L. A., La Torre, A., Field, J. K., Jakopovic, M., Knezevic, J., Castaños-Vélez, E., Roz, L., Pastorino, U., Brustugun, O.-T., Lund-Iversen, M., Thunnissen, E., Köhler, J., Schuler, M., Botling, J., Sandelin, M., Sanchez-Cespedes, M., Salvesen, H. B., Achter, V., Lang, U., Bogus, M., Schneider, P. M., Zander, T., Ansén, S., Hallek, M., Wolf, J., Vingron, M., Yatabe, Y., Travis, W. D., Nürnberg, P., Reinhardt, C., Perner, S., Heukamp, L., Büttner, R., Haas, S. A., Brambilla, E., Peifer, M., Sage, J., and Thomas, R. K. (2015). Comprehensive genomic profiles of small cell lung cancer. *Nature 524*, 47–53.

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

(*150*) Poirier, J. T., Dobromilskaya, I., Moriarty, W. F., Peacock, C. D., Hann, C. L., and Rudin, C. M. (2013). Selective tropism of Seneca Valley virus for variant subtype small cell lung cancer. *Journal of the National Cancer Institute 105*, 1059–1065.

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

(*152*) Borromeo, M. D., Savage, T. K., Kollipara, R. K., He, M., Augustyn, A., Osborne, J. K., Girard, L., Minna, J. D., Gazdar, A. F., Cobb, M. H., and Johnson, J. E. (2016). ASCL1 and NEUROD1 Reveal Heterogeneity in Pulmonary Neuroendocrine Tumors and Regulate Distinct Genetic Programs. *Cell Reports 16*, 1259–1272.

(*153*) Mollaoglu, G., Guthrie, M. R., Böhm, S., Brägelmann, J., Can, I., Ballieu, P. M., Marx, A., George, J., Heinen, C., Chalishazar, M. D., Cheng, H., Ireland, A. S., Denning, K. E., Mukhopadhyay, A., Vahrenkamp, J. M., Berrett, K. C., Mosbruger, T. L., Wang, J., Kohan,

J. L., Salama, M. E., Witt, B. L., Peifer, M., Thomas, R. K., Gertz, J., Johnson, J. E., Gazdar, A. F., Wechsler-Reya, R. J., Sos, M. L., and Oliver, T. G. (2017). MYC Drives Progression of Small Cell Lung Cancer to a Variant Neuroendocrine Subtype with Vulnerability to Aurora Kinase Inhibition. *Cancer Cell 31*, 270–285.

(*154*) McColl, K., Wildey, G., Sakre, N., Lipka, M. B., Behtaj, M., Kresak, A., Chen, Y., Yang, M., Velcheti, V., Fu, P., and Dowlati, A. (2017). Reciprocal expression of INSM1 and YAP1 defines subgroups in small cell lung cancer. *Oncotarget 8*, 73745–73756.

(*155*) Huang, Y.-H., Klingbeil, O., He, X.-Y., Wu, X. S., Arun, G., Lu, B., Somerville, T. D. D., Milazzo, J. P., Wilkinson, J. E., Demerdash, O. E., Spector, D. L., Egeblad, M., Shi, J., and Vakoc, C. R. (2018). POU2F3 is a master regulator of a tuft cell-like variant of small cell lung cancer. *Genes & Development 32*, 915–928.

(*156*) Wooten, D. J., Groves, S. M., Tyson, D. R., Liu, Q., Lim, J. S., Albert, R., Lopez, C. F., Sage, J., and Quaranta, V. (2019). Systems-level network modeling of Small Cell Lung Cancer subtypes identifies master regulators and destabilizers. *PLOS Computational Biology 15*, ed. by Maini, P. K., e1007343.

(*157*) Rudin, C. M., Poirier, J. T., Byers, L. A., Dive, C., Dowlati, A., George, J., Heymach, J. V., Johnson, J. E., Lehman, J. M., MacPherson, D., Massion, P. P., Minna, J. D., Oliver, T. G., Quaranta, V., Sage, J., Thomas, R. K., Vakoc, C. R., and Gazdar, A. F. (2019). Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nature Reviews. Cancer 19*, 289–297.

(*158*) Ding, X.-L., Su, Y.-G., Yu, L., Bai, Z.-L., Bai, X.-H., Chen, X.-Z., Yang, X., Zhao, R., He, J.-X., and Wang, Y.-Y. (2022). Clinical characteristics and patient outcomes of molecular subtypes of small cell lung cancer (SCLC). *World Journal of Surgical Oncology 20*, 54.

(*159*) Hwang, S., Hong, T. H., Kim, H. K., Choi, Y. S., Zo, J. I., Shim, Y. M., Han, J., Chan Ahn, Y., Pyo, H., Noh, J. M., Lee, H. Y., Kim, H. J., Park, S., Ahn, M.-J., Park, K., Lee, S.-H., Choi, Y.-L., and Kim, J. (2023). Whole-Section Landscape Analysis of Molecular Subtypes in Curatively Resected Small Cell Lung Cancer: Clinicopathologic Features and Prognostic Significance. *Modern Pathology 36*, 100184.

(*160*) Lantuejoul, S., Fernandez-Cuesta, L., Damiola, F., Girard, N., and McLeer, A. (2020). New molecular classification of large cell neuroendocrine carcinoma and small cell lung carcinoma with potential therapeutic impacts. *Translational Lung Cancer Research 9*, 2233–2244.

(*161*) Qu, S., Fetsch, P., Thomas, A., Pommier, Y., Schrump, D. S., Miettinen, M. M., and Chen, H. (2022). Molecular Subtypes of Primary SCLC Tumors and Their Associations With Neuroendocrine and Therapeutic Markers. *Journal of Thoracic Oncology 17*, 141–153.

(*162*) Megyesfalvi, Z., Barany, N., Lantos, A., Valko, Z., Pipek, O., Lang, C., Schwendenwein, A., Oberndorfer, F., Paku, S., Ferencz, B., Dezso, K., Fillinger, J., Lohinai, Z., Moldvay, J., Galffy, G., Szeitz, B., Rezeli, M., Rivard, C., Hirsch, F. R., Brcic, L., Popper, H., Kern, I., Kovacevic, M., Skarda, J., Mittak, M., Marko-Varga, G., Bogos, K., Renyi-Vamos, F., Hoda, M. A., Klikovits, T., Hoetzenecker, K., Schelch, K., Laszlo, V., and Dome, B. (2022). Expression patterns and prognostic relevance of subtype-specific transcription factors in surgically resected small-cell lung cancer: an international multicenter study. *The Journal of Pathology 257*, 674–686.

(*163*) Gay, C. M., Stewart, C. A., Park, E. M., Diao, L., Groves, S. M., Heeke, S., Nabet, B. Y., Fujimoto, J., Solis, L. M., Lu, W., Xi, Y., Cardnell, R. J., Wang, Q., Fabbri, G., Cargill, K. R., Vokes, N. I., Ramkumar, K., Zhang, B., Della Corte, C. M., Robson, P., Swisher, S. G., Roth, J. A., Glisson, B. S., Shames, D. S., Wistuba, I. I., Wang, J., Quaranta, V., Minna, J., Heymach, J. V., and Byers, L. A. (2021). Patterns of transcription factor programs and immune pathway activation define four major subtypes of SCLC with distinct therapeutic vulnerabilities. *Cancer Cell 39*, 346–360.e7.

(*164*) Rekhtman, N., Pietanza, M. C., Hellmann, M. D., Naidoo, J., Arora, A., Won, H., Halpenny, D. F., Wang, H., Tian, S. K., Litvak, A. M., Paik, P. K., Drilon, A. E., Socci, N., Poirier, J. T., Shen, R., Berger, M. F., Moreira, A. L., Travis, W. D., Rudin, C. M., and Ladanyi, M. (2016). Next-Generation Sequencing of Pulmonary Large Cell Neuroendocrine Carcinoma Reveals Small Cell Carcinoma-like and Non-Small Cell Carcinoma-like Subsets. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 22*, 3618–3629.

(*165*) Miyoshi, T., Umemura, S., Matsumura, Y., Mimaki, S., Tada, S., Makinoshima, H., Ishii, G., Udagawa, H., Matsumoto, S., Yoh, K., Niho, S., Ohmatsu, H., Aokage, K., Hishida, T., Yoshida, J., Nagai, K., Goto, K., Tsuboi, M., and Tsuchihara, K. (2017). Genomic Profiling of Large-Cell Neuroendocrine Carcinoma of the Lung. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 23*, 757–765.

(*166*) Zhuo, M., Guan, Y., Yang, X., Hong, L., Wang, Y., Li, Z., Chen, R., Abbas, H. A., Chang, L., Gong, Y., Wu, N., Zhong, J., Chen, W., Chen, H., Dong, Z., Zhu, X., Li, J., Wang, Y., An, T., Wu, M., Wang, Z., Wang, J., Roarty, E. B., Rinsurongkawong, W., Lewis, J., Roth, J. A., Swisher, S. G., Lee, J. J., Heymach, J. V., Wistuba, I. I., Kalhor, N., Yang, L.,

Yi, X., Futreal, P. A., Glisson, B. S., Xia, X., Zhang, J., and Zhao, J. (2020). The Prognostic and Therapeutic Role of Genomic Subtyping by Sequencing Tumor or Cell-Free DNA in Pulmonary Large-Cell Neuroendocrine Carcinoma. *Clinical Cancer Research 26*, 892–901.

(*167*) George, J., Walter, V., Peifer, M., Alexandrov, L. B., Seidel, D., Leenders, F., Maas, L., Müller, C., Dahmen, I., Delhomme, T. M., Ardin, M., Leblay, N., Byrnes, G., Sun, R., De Reynies, A., McLeer-Florin, A., Bosco, G., Malchers, F., Menon, R., Altmüller, J., Becker, C., Nürnberg, P., Achter, V., Lang, U., Schneider, P. M., Bogus, M., Soloway, M. G., Wilkerson, M. D., Cun, Y., McKay, J. D., Moro-Sibilot, D., Brambilla, C. G., Lantuejoul, S., Lemaitre, N., Soltermann, A., Weder, W., Tischler, V., Brustugun, O. T., Lund-Iversen, M., Helland, Å., Solberg, S., Ansén, S., Wright, G., Solomon, B., Roz, L., Pastorino, U., Petersen, I., Clement, J. H., Sänger, J., Wolf, J., Vingron, M., Zander, T., Perner, S., Travis, W. D., Haas, S. A., Olivier, M., Foll, M., Büttner, R., Hayes, D. N., Brambilla, E., Fernandez-Cuesta, L., and Thomas, R. K. (2018). Integrative genomic profiling of large-cell neuroendocrine carcinomas reveals distinct subtypes of high-grade neuroendocrine lung tumors. *Nature Communications 9*, 1048.

(*168*) Garber, M. E., Troyanskaya, O. G., Schluens, K., Petersen, S., Thaesler, Z., Pacyna-Gengelbach, M., van de Rijn, M., Rosen, G. D., Perou, C. M., Whyte, R. I., Altman, R. B., Brown, P. O., Botstein, D., and Petersen, I. (2001). Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of the National Academy of Sciences of the United States of America 98*, 13784–13789.

(*169*) Sikkema, L., Ramírez-Suástegui, C., Strobl, D. C., Gillett, T. E., Zappia, L., Madissoon, E., Markov, N. S., Zaragosi, L.-E., Ji, Y., Ansari, M., Arguel, M.-J., Apperloo, L., Banchero, M., Bécavin, C., Berg, M., Chichelnitskiy, E., Chung, M.-i., Collin, A., Gay, A. C. A., Gote-Schniering, J., Hooshiar Kashani, B., Inecik, K., Jain, M., Kapellos, T. S., Kole, T. M., Leroy, S., Mayr, C. H., Oliver, A. J., Von Papen, M., Peter, L., Taylor, C. J., Walzthoeni, T., Xu, C., Bui, L. T., De Donno, C., Dony, L., Faiz, A., Guo, M., Gutierrez, A. J., Heumos, L., Huang, N., Ibarra, I. L., Jackson, N. D., Kadur Lakshminarasimha Murthy, P., Lotfollahi, M., Tabib, T., Talavera-López, C., Travaglini, K. J., Wilbrey-Clark, A., Worlock, K. B., Yoshida, M., Lung Biological Network Consortium, Chen, Y., Hagood, J. S., Agami, A., Horvath, P., Lundeberg, J., Marquette, C.-H., Pryhuber, G., Samakovlis, C., Sun, X., Ware, L. B., Zhang, K., Van Den Berge, M., Bossé, Y., Desai, T. J., Eickelberg, O., Kaminski, N., Krasnow, M. A., Lafyatis, R., Nikolic, M. Z., Powell, J. E., Rajagopal, J., Rojas, M., Rozenblatt-Rosen, O., Seibold, M. A., Sheppard, D., Shepherd, D. P., Sin, D. D., Timens, W., Tsankov, A. M., Whitsett, J., Xu, Y., Banovich, N. E., Bar-

bry, P., Duong, T. E., Falk, C. S., Meyer, K. B., Kropski, J. A., Pe'er, D., Schiller, H. B., Tata, P. R., Schultze, J. L., Teichmann, S. A., Misharin, A. V., Nawijn, M. C., Luecken, M. D., and Theis, F. J. (2023). An integrated cell atlas of the lung in health and disease. *Nature Medicine 29*, 1563–1577.

(*170*) Salcher, S., Sturm, G., Horvath, L., Untergasser, G., Kuempers, C., Fotakis, G., Panizzolo, E., Martowicz, A., Trebo, M., Pall, G., Gamerith, G., Sykora, M., Augustin, F., Schmitz, K., Finotello, F., Rieder, D., Perner, S., Sopper, S., Wolf, D., Pircher, A., and Trajanoski, Z. (2022). High-resolution single-cell atlas reveals diversity and plasticity of tissue-resident neutrophils in non-small cell lung cancer. *Cancer Cell 40*, 1503–1520.e8.

(*171*) Chan, J. M., Quintanal-Villalonga, Á., Gao, V. R., Xie, Y., Allaj, V., Chaudhary, O., Masilionis, I., Egger, J., Chow, A., Walle, T., Mattar, M., Yarlagadda, D. V., Wang, J. L., Uddin, F., Offin, M., Ciampricotti, M., Qeriqi, B., Bahr, A., De Stanchina, E., Bhanot, U. K., Lai, W. V., Bott, M. J., Jones, D. R., Ruiz, A., Baine, M. K., Li, Y., Rekhtman, N., Poirier, J. T., Nawy, T., Sen, T., Mazutis, L., Hollmann, T. J., Pe'er, D., and Rudin, C. M. (2021). Signatures of plasticity, metastasis, and immunosuppression in an atlas of human small cell lung cancer. *Cancer Cell 39*, 1479–1496.e18.

(*172*) Han, G., Sinjab, A., Rahal, Z., Lynch, A. M., Treekitkarnmongkol, W., Liu, Y., Serrano, A. G., Feng, J., Liang, K., Khan, K., Lu, W., Hernandez, S. D., Liu, Y., Cao, X., Dai, E., Pei, G., Hu, J., Abaya, C., Gomez-Bolanos, L. I., Peng, F., Chen, M., Parra, E. R., Cascone, T., Sepesi, B., Moghaddam, S. J., Scheet, P., Negrao, M. V., Heymach, J. V., Li, M., Dubinett, S. M., Stevenson, C. S., Spira, A. E., Fujimoto, J., Solis, L. M., Wistuba, I. I., Chen, J., Wang, L., and Kadara, H. (2024). An atlas of epithelial cell states and plasticity in lung adenocarcinoma. *Nature 627*, 656–663.

(*173*) Li, Q., Wang, R., Yang, Z., Li, W., Yang, J., Wang, Z., Bai, H., Cui, Y., Tian, Y., Wu, Z., Guo, Y., Xu, J., Wen, L., He, J., Tang, F., and Wang, J. (2022). Molecular profiling of human non-small cell lung cancer by single-cell RNA-seq. *Genome Medicine 14*, 87.

(*174*) Ouyang, H., Chen, J., Xu, C., He, C., Sheng, L., Wang, J., Pan, D., Lin, J., Augert, A., Zhu, Y., and Jia, D. Single-Cell Transcriptomic Analyses of Tumor Ecosystems and Spatial Architectures in Human Small Cell Lung Cancer, en, 2022.

(*175*) Travaglini, K. J., Nabhan, A. N., Penland, L., Sinha, R., Gillich, A., Sit, R. V., Chang, S., Conley, S. D., Mori, Y., Seita, J., Berry, G. J., Shrager, J. B., Metzger, R. J., Kuo, C. S., Neff, N., Weissman, I. L., Quake, S. R., and Krasnow, M. A. (2020). A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature 587*, 619–625.

(*176*) MacParland, S. A., Liu, J. C., Ma, X.-Z., Innes, B. T., Bartczak, A. M., Gage, B. K., Manuel, J., Khuu, N., Echeverri, J., Linares, I., Gupta, R., Cheng, M. L., Liu, L. Y., Camat, D., Chung, S. W., Seliga, R. K., Shao, Z., Lee, E., Ogawa, S., Ogawa, M., Wilson, M. D., Fish, J. E., Selzner, M., Ghanekar, A., Grant, D., Greig, P., Sapisochin, G., Selzner, N., Winegarden, N., Adeyi, O., Keller, G., Bader, G. D., and McGilvray, I. D. (2018). Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nature Communications 9*, 4383.

(*177*) Aizarani, N., Saviano, A., Sagar, Mailly, L., Durand, S., Herman, J. S., Pessaux, P., Baumert, T. F., and Grün, D. (2019). A human liver cell atlas reveals heterogeneity and epithelial progenitors. *Nature 572*, 199–204.

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

(*179*) Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England) 26*, 139–140.

(*180*) Hao, Y., Hao, S., Andersen-Nissen, E., Mauck, W. M., Zheng, S., Butler, A., Lee, M. J., Wilk, A. J., Darby, C., Zager, M., Hoffman, P., Stoeckius, M., Papalexi, E., Mimitou, E. P., Jain, J., Srivastava, A., Stuart, T., Fleming, L. M., Yeung, B., Rogers, A. J., McElrath, J. M., Blish, C. A., Gottardo, R., Smibert, P., and Satija, R. (2021). Integrated analysis of multimodal single-cell data. *Cell 184*, 3573–3587.e29.

(*181*) Becht, E., Giraldo, N. A., Lacroix, L., Buttard, B., Elarouci, N., Petitprez, F., Selves, J., Laurent-Puig, P., Sautès-Fridman, C., Fridman, W. H., and De Reyniès, A. (2016). Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology 17*, 218.

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

(*183*) Svensson, V., Teichmann, S. A., and Stegle, O. (2018). SpatialDE: identification of spatially variable genes. *Nature Methods 15*, 343–346.

(*184*) Sun, S., Zhu, J., and Zhou, X. (2020). Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nature Methods 17*, 193–200.

(*185*) Hao, M., Hua, K., and Zhang, X. (2021). SOMDE: a scalable method for identifying spatially variable genes with self-organizing map. *Bioinformatics 37*, ed. by Robinson, P., 4392–4398.

(*186*) Miller, B. F., Bambah-Mukku, D., Dulac, C., Zhuang, X., and Fan, J. (2021). Characterizing spatial gene expression heterogeneity in spatially resolved single-cell transcriptomic data with nonuniform cellular densities. *Genome Research 31*, 1843–1855.

(*187*) Zhang, K., Feng, W., and Wang, P. (2022). Identification of spatially variable genes with graph cuts. *Nature Communications 13*, 5488.

(*188*) Berglund, E., Maaskola, J., Schultz, N., Friedrich, S., Marklund, M., Bergenstråhle, J., Tarish, F., Tanoglidi, A., Vickovic, S., Larsson, L., Salmén, F., Ogris, C., Wallenborg, K., Lagergren, J., Ståhl, P., Sonnhammer, E., Helleday, T., and Lundeberg, J. (2018). Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nature Communications 9*, 2419.

(*189*) OpenStax Graphs of Linear Functions, 2024.

(*190*) NeilRoy Plot of n concentric circles at once?, Forum post, 2015.

(*191*) Calculator.net Ellipsis, Area Calculator, 2017.

(*192*) Zhou, Y., He, W., Hou, W., and Zhu, Y. (2024). Pianno: a probabilistic framework automating semantic annotation for spatial transcriptomics. *Nature Communications 15*, 2848.

(*193*) Amason, M. E., Beatty, C. J., Harvest, C. K., Saban, D. R., and Miao, E. A. Chemokine expression profile of an innate granuloma, 2024.

(*194*) Patel, A. P., Tirosh, I., Trombetta, J. J., Shalek, A. K., Gillespie, S. M., Wakimoto, H., Cahill, D. P., Nahed, B. V., Curry, W. T., Martuza, R. L., Louis, D. N., Rozenblatt-Rosen, O., Suvà, M. L., Regev, A., and Bernstein, B. E. (2014). Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (New York, N.Y.) 344*, 1396–1401.

(*195*) Tirosh, I., Venteicher, A. S., Hebert, C., Escalante, L. E., Patel, A. P., Yizhak, K., Fisher, J. M., Rodman, C., Mount, C., Filbin, M. G., Neftel, C., Desai, N., Nyman, J., Izar, B., Luo, C. C., Francis, J. M., Patel, A. A., Onozato, M. L., Riggi, N., Livak, K. J., Gennert, D., Satija, R., Nahed, B. V., Curry, W. T., Martuza, R. L., Mylvaganam, R., Iafrate, A. J., Frosch, M. P., Golub, T. R., Rivera, M. N., Getz, G., Rozenblatt-Rosen, O., Cahill, D. P., Monje, M., Bernstein, B. E., Louis, D. N., Regev, A., and Suvà, M. L. (2016). Single-cell

RNA-seq supports a developmental hierarchy in human oligodendroglioma. *Nature 539*, 309–313.

(*196*) Kinker, G. S., Greenwald, A. C., Tal, R., Orlova, Z., Cuoco, M. S., McFarland, J. M., Warren, A., Rodman, C., Roth, J. A., Bender, S. A., Kumar, B., Rocco, J. W., Fernandes, P. A. C. M., Mader, C. C., Keren-Shaul, H., Plotnikov, A., Barr, H., Tsherniak, A., Rozenblatt-Rosen, O., Krizhanovsky, V., Puram, S. V., Regev, A., and Tirosh, I. (2020). Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nature Genetics 52*, 1208–1218.

(*197*) Tickle, T. I., Georgescu, C., Brown, M., and Haas, B. inferCNV of the Trinity CTAT Project. 2019.

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

(*199*) Hao, Y., Stuart, T., Kowalski, M. H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., and Satija, R. (2024). Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology 42*, 293–304.

(*200*) Pelka, K., Hofree, M., Chen, J. H., Sarkizova, S., Pirl, J. D., Jorgji, V., Bejnood, A., Dionne, D., Ge, W. H., Xu, K. H., Chao, S. X., Zollinger, D. R., Lieb, D. J., Reeves, J. W., Fuhrman, C. A., Hoang, M. L., Delorey, T., Nguyen, L. T., Waldman, J., Klapholz, M., Wakiro, I., Cohen, O., Albers, J., Smillie, C. S., Cuoco, M. S., Wu, J., Su, M.-J., Yeung, J., Vijaykumar, B., Magnuson, A. M., Asinovski, N., Moll, T., Goder-Reiser, M. N., Applebaum, A. S., Brais, L. K., DelloStritto, L. K., Denning, S. L., Phillips, S. T., Hill, E. K., Meehan, J. K., Frederick, D. T., Sharova, T., Kanodia, A., Todres, E. Z., Jané-Valbuena, J., Biton, M., Izar, B., Lambden, C. D., Clancy, T. E., Bleday, R., Melnitchouk, N., Irani, J., Kunitake, H., Berger, D. L., Srivastava, A., Hornick, J. L., Ogino, S., Rotem, A., Vigneau, S., Johnson, B. E., Corcoran, R. B., Sharpe, A. H., Kuchroo, V. K., Ng, K., Giannakis, M., Nieman, L. T., Boland, G. M., Aguirre, A. J., Anderson, A. C., Rozenblatt-Rosen, O., Regev, A., and Hacohen, N. (2021). Spatially organized multicellular immune hubs in human colorectal cancer. *Cell 184*, 4734–4752.e20.

(*201*) Valdeolivas, A., Amberg, B., Giroud, N., Richardson, M., Gálvez, E. J. C., Badillo, S., Julien-Laferrière, A., Túrós, D., Voith Von Voithenberg, L., Wells, I., Pesti, B., Lo, A. A., Yángüez, E., Das Thakur, M., Bscheider, M., Sultan, M., Kumpesa, N., Jacobsen, B., Bergauer, T., Saez-Rodriguez, J., Rottenberg, S., Schwalie, P. C., and Hahn, K. (2024).

Profiling the heterogeneity of colorectal cancer consensus molecular subtypes using spatial transcriptomics. *npj Precision Oncology 8*, 10.

(*202*) Goveia, J., Rohlenova, K., Taverna, F., Treps, L., Conradi, L.-C., Pircher, A., Geldhof, V., de Rooij, L. P. M. H., Kalucka, J., Sokol, L., García-Caballero, M., Zheng, Y., Qian, J., Teuwen, L.-A., Khan, S., Boeckx, B., Wauters, E., Decaluwé, H., De Leyn, P., Vansteenkiste, J., Weynand, B., Sagaert, X., Verbeken, E., Wolthuis, A., Topal, B., Everaerts, W., Bohnenberger, H., Emmert, A., Panovska, D., De Smet, F., Staal, F. J. T., Mclaughlin, R. J., Impens, F., Lagani, V., Vinckier, S., Mazzone, M., Schoonjans, L., Dewerchin, M., Eelen, G., Karakach, T. K., Yang, H., Wang, J., Bolund, L., Lin, L., Thienpont, B., Li, X., Lambrechts, D., Luo, Y., and Carmeliet, P. (2020). An Integrated Gene Expression Landscape Profiling Approach to Identify Lung Tumor Endothelial Cell Heterogeneity and Angiogenic Candidates. *Cancer Cell 37*, 21–36.e13.

(*203*) Wu, F., Fan, J., He, Y., Xiong, A., Yu, J., Li, Y., Zhang, Y., Zhao, W., Zhou, F., Li, W., Zhang, J., Zhang, X., Qiao, M., Gao, G., Chen, S., Chen, X., Li, X., Hou, L., Wu, C., Su, C., Ren, S., Odenthal, M., Buettner, R., Fang, N., and Zhou, C. (2021). Single-cell profiling of tumor heterogeneity and the microenvironment in advanced non-small cell lung cancer. *Nature Communications 12*, 2540.

(*204*) Laughney, A. M., Hu, J., Campbell, N. R., Bakhoum, S. F., Setty, M., Lavallée, V.-P., Xie, Y., Masilionis, I., Carr, A. J., Kottapalli, S., Allaj, V., Mattar, M., Rekhtman, N., Xavier, J. B., Mazutis, L., Poirier, J. T., Rudin, C. M., Pe'er, D., and Massagué, J. (2020). Regenerative lineages and immune-mediated pruning in lung cancer metastasis. *Nature Medicine 26*, 259–269.

(*205*) Chen, J., Tan, Y., Sun, F., Hou, L., Zhang, C., Ge, T., Yu, H., Wu, C., Zhu, Y., Duan, L., Wu, L., Song, N., Zhang, L., Zhang, W., Wang, D., Chen, C., Wu, C., Jiang, G., and Zhang, P. (2020). Single-cell transcriptome and antigen-immunoglobin analysis reveals the diversity of B cells in non-small cell lung cancer. *Genome Biology 21*, 152.

(*206*) Adams, T. S., Schupp, J. C., Poli, S., Ayaub, E. A., Neumark, N., Ahangari, F., Chu, S. G., Raby, B. A., DeIuliis, G., Januszyk, M., Duan, Q., Arnett, H. A., Siddiqui, A., Washko, G. R., Homer, R., Yan, X., Rosas, I. O., and Kaminski, N. (2020). Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Science Advances 6*, eaba1983.

(*207*) Maynard, A., McCoach, C. E., Rotow, J. K., Harris, L., Haderk, F., Kerr, D. L., Yu, E. A., Schenk, E. L., Tan, W., Zee, A., Tan, M., Gui, P., Lea, T., Wu, W., Urisman, A., Jones, K., Sit, R., Kolli, P. K., Seeley, E., Gesthalter, Y., Le, D. D., Yamauchi, K. A., Naeger,

D. M., Bandyopadhyay, S., Shah, K., Cech, L., Thomas, N. J., Gupta, A., Gonzalez, M., Do, H., Tan, L., Bacaltos, B., Gomez-Sjoberg, R., Gubens, M., Jahan, T., Kratz, J. R., Jablons, D., Neff, N., Doebele, R. C., Weissman, J., Blakely, C. M., Darmanis, S., and Bivona, T. G. (2020). Therapy-Induced Evolution of Human Lung Cancer Revealed by Single-Cell RNA Sequencing. *Cell 182*, 1232–1251.e22.

(*208*) Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nature Methods 15*, 1053–1058.

List of Figures

1 The established hallmarks of cancer. 21

2 A simplified schematic of the tumour microenvironment and its components. . . 23

3 A simplified schematic of the tumour microenvironment immune cell component tumour suppressing and tumour promoting activity. 24

4 Sources of intra-tumour heterogeneity. 26

5 Intra-tumour heterogeneity of a tumour through progression under the influence of clonal evolution. 27

6 Treatment resistance in the cases of pre-existing subclonal ITH and treatment induced lineage plasticity. 28

7 Cell separation by microfluidics. 32

8 Transcriptomics at different levels of resolution: bulk, single-cell and spatial. . . . 33

9 Fully spatial approaches to transcriptomics: NGS barcode, *in situ* sequencing, and *in situ* hybridization. 34

10 The Visium Spatial Gene Expression Slide v2. 35

11 The anatomy of human respiratory system and lungs. 37

12 A simplified schematic of the human lung and the cellular composition of its epithelium. 39

13 H&E stained tissue of the main histologies of lung cancer. 40

14 Molecular subtypes of lung adenocarcinoma. 46

15 Molecular subtypes of lung squamous cell carcinoma. 49

16 Molecular subtypes of small-cell lung carcinoma. 52

17 Molecular classification of high grade neuroendocrine lung carcinomas. 54

18 UMAPs of the integrated Human Lung Cell Atlas. 56

19 Markers of the 61 cell types of the integrated Human Lung Cell Atlas. 57

20 UMAPs of the non-small-cell lung carcinoma atlas. 59

21 UMAPs of the small-cell lung carcinoma atlas. 60

22 Simulated counts for a theoretical example marker gene in its cell type population ($p = 10^{-4}$) and the rest ($p = 10^{-6}$) as a function of total reads. 69

23 Matrixplot of proposed marker genes with highest log-fold change for each cell type in Hao et al. (2021)[180]. 72

24 Matrixplot of the highest scoring marker genes according to SCherlock evaluation on Hao et al. (2021)[180]. 73

25 Expression of *JCHAIN* (top), marker for pDC cells from Hao et al. (2021)[180] with the highest log-fold change, and *KRT5* (bottom), highest scoring marker gene according to theoretical score, not featured on the list of markers supplied with the original paper. 74

26 The hierarchy of cell type annotation levels in Hao et al. (2021)[180]. 75

27 Matrixplot of the highest scoring marker genes according to SCherlock evaluation on the revised annotation level for Hao et al. (2021)[180]. 76

28 Expression of *MS4A1* (*CD20*), the highest scoring marker gene for B cells according to SCherlock evaluation on Hao et al. (2021)[180]. 77

29 Expression of *CD207* (top) and *TRDV2* (bottom), the highest scoring marker genes for cDC2 and gdT cells respectively according to SCherlock evaluation on Hao et al. (2021)[180]. 78

30 Expression of *HAMP* (top), *CYP1B1* (middle), and *AC106712.1* (bottom), the highest scoring marker genes for ASDC, CD14 Mono and CD16 Mono cells respectively according to SCherlock evaluation on Hao et al. (2021)[180]. 80

31 Expression of *TPO* (top), *PRL* (middle), and *HBM* (bottom), the highest scoring marker genes for CD4, CD8 and Eryth cells respectively according to SCherlock evaluation on Hao et al. (2021)[180]. 81

32 Expression of *CPA3* (top), *GATA3-AS1* (middle), and *RBM24* (bottom), the highest scoring marker genes for HSPC, ILC and MAIT cells respectively according to SCherlock evaluation on Hao et al. (2021)[180]. 82

33 Expression of *COL13A1* (top), *SLIT3* (middle), and *SDC1* (bottom), the highest scoring marker genes for NK, NK CD56bright and Plasmablast cells respectively according to SCherlock evaluation on Hao et al. (2021)[180]. 83

34 Expression of *GP9* (top), *LINC01281* (middle), and *ERICH5* (bottom), the highest scoring marker genes for Platelet, Treg and cDC1 cells respectively according to SCherlock evaluation on Hao et al. (2021)[180]. 84

35 Expression of *OGN* (top), and *KCTD19* (bottom), the highest scoring marker genes for dnT and pDC cells respectively according to SCherlock evaluation on Hao et al. (2021)[180]. 85

36 Line-source gradient underlying distribution contour lines (left) and density (right) example. 88

37 Point-source gradient underlying distribution contour lines (left) and density (right) example. 88

38 Ellipse defining elements (top left) and corresponding distribution density example for a structure (top right) and a barrier (bottom). 90

39 Simulated expression examples with $N = 1000$ for uniform (top left), line-source gradient (bottom left), structure (top right), and barrier (bottom right). 91

40 Pixel neighbourhood. 92

41 Schematic of a step of an expression-based random walk (left); all paths length two leading to pixel A from pixel X (right). 93

42 UMAP projection of the training dataset with $N = 1000$ transcripts per sample. 94

43 UMAP projection of the training dataset with $N = 100$ transcripts per sample. . 95

44 Spatially resolved expression of *WFS1*, *PVRL3*, and *PROX1* in the mouse brain hippocampus. 97

45 Spatially resolved expression of *CXCL2* and *CCL7* at various days post-infection (DPI) in a liver granuloma response in a mouse infection model. 97

46 Simulated uniform expression distribution examples for $N = 100$ (top), $N = 1000$ (middle), and $N = 10000$ (bottom). 98

47 Simulated expression distribution examples of the point-source (left) and point-source (right) gradient patterns for $N = 100$ (top), $N = 1000$ (middle), and $N = 10000$ (bottom). 99

48 Simulated expression distribution examples of the structure pattern for $N = 100$ (top), $N = 1000$ (middle), and $N = 10000$ (bottom). 100

49 Simulated expression distribution examples of the barrier pattern for $N = 100$ (top), $N = 1000$ (middle), and $N = 10000$ (bottom). 101

50 CNV heatmap for a single-cell colorectal cancer example dataset. 104

51 CNV heatmap for a spatial colorectal cancer example slide. 104

52 UMAP of the single-cell example dataset with fraction of non-zero CNV calls (left) and cell type annotation (right). 105

53 Visium slide of the spatial example with fraction of non-zero CNV calls (left) and spot annotation (right). 105

54 CNV heatmap with CNV calls based clusters for a spatial colorectal cancer example slide. 106

55 CNV heatmap with CNV calls based clusters for a single-cell colorectal cancer example dataset. 106

56 Breakdown of the NSCLC atlas[170] datasets. 109

57 UMAP of the Chen et al. (2020)[205] dataset with *CHGA* (top left), *SYP* (top right), *CD56* (bottom left) expression and sample annotation (bottom right). . . 110

58 UMAPs of the first half of the dataset post scVI integration annotated by donor (top left), source dataset (top right) and leiden cluster (bottom). 113

59 UMAPs of the first half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom). . . . 114

60 UMAPs of the mesenchymal subset of the first half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom). 115

61 UMAPs of the mesenchymal subset of the first half of the dataset post scVI integration with additional healthy data annotated by SCLC atlas coarse annotation (top) and healthy lung cell atlas annotation (bottom). 116

62 UMAPs of the myeloid subset of the first half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom). 117

63 UMAPs of the myeloid subset of the first half of the dataset post scVI integration annotated by the expression of markers of four DC cell types: cDC1 (top left), cDC2 (top right), DC mature (bottom left), and pDC (bottom right). 118

64 UMAPs of the lymphoid subset of the first half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom). 119

65 UMAP of the lymphoid subset of the first half of the dataset post scVI integration annotated by the expression of a marker of the NK cell type. 120

66 UMAPs of the epithelial subset of the first half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom). 121

67 UMAPs of the CNV clusters of the first half of the dataset post FastCNV analysis annotated by the expression of donor (left) and CNV score (right). 122

68 Matrixplot of proposed marker genes for each cell type in the NSCLC atlas. . . . 123

69 Matrixplot of proposed marker genes for cell types in the SCLC atlas. 125

70 UMAPs of the first half of the dataset post scVI integration reannotated. 125

71 Matrixplot of proposed top scoring marker genes for reannotated cell types in the unified lung cancer cell atlas. 126

72 UMAPs of the second half of the dataset post scVI integration annotated by donor (top left), source dataset (top right) and leiden cluster (bottom). 128

73 UMAPs of the second half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom). 129

74 UMAPs of the mesenchymal subset of the second half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom). 130

75 UMAPs of the mesenchymal subset of the second half of the dataset post scVI integration with additional healthy data annotated by SCLC atlas coarse annotation (top) and healthy lung cell atlas annotation (bottom). 131

76 UMAPs of the myeloid subset of the second half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom). 132

77 UMAPs of the myeloid subset of the second half of the dataset post scVI integration annotated by the expression of markers of four DC cell types: cDC1 (top left), cDC2 (top right), DC mature (bottom left), and pDC (bottom right). . . . 133

78 UMAPs of the lymphoid subset of the second half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom). 134

79 UMAP of the lymphoid subset of the second half of the dataset post scVI integration annotated by the expression of a marker of the NK cell type. 135

80 UMAPs of the epithelial subset of the second half of the dataset post scVI integration annotated by SCLC atlas coarse annotation (top) and NSCLC atlas coarse annotation (bottom). 136

81 UMAPs of the CNV clusters of the second half of the dataset post FastCNV analysis annotated by the expression of donor (left) and CNV score (right). 137

82 UMAPs of the second half of the dataset post scVI integration reannotated. . . . 137