At the Frontier of Immunology: Exploring Cellular Crosstalk Across Time and Space

Olympia Melek Hardy

(BSc, MSc)

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

INSTITUTE OF INFECTION AND IMMUNITY

COLLEGE OF MEDICAL AND VETERINARY SCIENCES

SEP 2025

This ones for you mama.

Abstract

Cellular interactions underpin all biological processes and offer unprecedented insight into mechanisms of action in steady state and disease. The advent of single-cell and spatial technologies has allowed us to resolve these interactions across time and space unveiling novel pathways in infectious and inflammatory disease, yet interpretation and visualisation remains challenging in these multi-faceted high-dimensional datasets. This thesis develops and applies computational and visual approaches to infer, prioritise, and validate cell–cell communication (CCI) in such contexts, demonstrating leveraging spatial information allows us to hone in on biological hypotheses and reduces false positives in ligand–receptor analyses. In Chapter 1, I analyse lethal COVID-19 in a Malawian cohort using histology, high-dimensional imaging, and single-cell transcriptomics from lung, blood, and nasal tissues, integrated with datasets from Northern Hemisphere cohorts. This cellular interaction analysis reveals distinct immune drivers in our cohort: an interferon-gamma programme in lung-resident alveolar macrophages in Malawi contrasted with type I/III interferon responses in blood-derived monocytes reported in USA/European cohorts. These results provide mechanistic insight into fatal disease in an under-represented population, and highlight the value of context-aware cellular inference and validation. In Chapter 2, I introduce cellXplore, a Flask–React interactive visualisation web tool that unifies widely used CCI packages and leverages single cell RNA sequencing with spatial transcriptomics to investigate computed cellular interactions. Through interactive, point-and-click workflows, cellXplore streamlines analysis, allowing customisable interactive plots, and prioritises spatially plausible interactions by overlaying ligand–receptor expression with co-localisation of spatial gene expression. I present three end-to-end user workflows using single cell and spatial transcriptomics data from a 10X Visium parasitic infection and a 10X Xenium breast cancer dataset to show indirect spatial validation of

cellular interactions can be utilised in a user-friendly manner. Lastly in Chapter 3, I extend cellular communication inference to complex datasets, validating cellular interactions and key drivers of inflammation leveraging immunohistochemistry and spatial transcriptomics. A multifactor macrophage–fibroblast atlas spanning four tissues and inflammatory states reveals conserved tissue-resident myeloid–stromal circuits through a APOE+/SPARC+ - SPP1+ axis that underpins inflammation alongside tissue-specific crosstalk reflecting organ microenvironments. A second study applies 10X Visium to intestine 'gut-rolls' across four time points of *Heligmosomoides polygyrus* infection. The analysis uncovers epithelial and immune programs associated with granuloma formation, stem-cell reprogramming, and parasite-driven immunomodulation within a distorted tissue landscape, with cellular interactions validated in the spatial context. Together, these studies shine a spotlight on the power of spatially-aware cellular interaction inference providing insight into COVID-19, tissue-resident myeloid–stromal communication during inflammation, and helminth infection in addition to a novel visualisation tool to unlock new insights from cellular interaction data.

Contents

Abstract v

Acknowledgements xii

Declaration xvi

Abbreviations xvii

1 Introduction 1

1.1 Cellular interaction inference in pathogen mediated disease 1

1.1.1 Cellular interaction inference in COVID-19 2

1.1.2 Emergence of single cell atlases 3

1.1.3 Cellular interaction inference in helminth infections 5

1.2 Introduction to single cell RNA sequencing 6

1.3 Introduction to scRNA-seq analysis 9

1.3.1 Pre-processing and Quality Control 10

1.3.2 Advanced Quality Control 12

1.3.3 Additional preprocessing steps 13

1.3.4 Normalisation 15

1.3.5 Feature Selection and Dimensionality Reduction 17

1.3.6 Dimensionality Reduction for sc-RNA visualisation 18

1.3.7 Clustering 19

1.3.8 Integration and Batch Correction 20

1.3.9 Cell type classification and composition 21

1.3.10 Pseudotime and velocity 22

1.3.11 Differential expression and functional analysis 23

1.4 Introduction to cellular interactions and inference methods 25
1.4.1 L-R Databases 26
1.4.2 scRNA-seq cellular inference methods 27
1.5 Introduction to spatial technologies 31
1.5.1 Spatial Transcriptomics 31
1.5.2 Spatial Proteomics 33
1.5.3 Additional analysis steps in spatial technologies 34
1.5.4 Cell Segmentation 35
1.5.5 Quality Control 36
1.5.6 Spatially Variable Genes and Neighbourhood Analysis 37
1.5.7 Cellular Deconvolution 37
1.5.8 Spatial cellular inference methods 39
1.6 Introduction to single cell visualisation tools 40
1.7 Aims and Objectives 42
1.7.1 Aim 1: Spatially resolved single-cell atlas unveils a distinct cellular signature of fatal lung COVID-19 in a Malawian population 43
1.7.2 Aim 2: cellXplore: a web tool to interactively explore cellular interactions at the single cell resolution 43
1.7.3 Aim 3: Dissecting cellular interactions in big data: Contextualising cellular interactions using atlas-level single cell and sequencing based spatial transcriptomics 43
1.7.4 Publications 43

2 Spatially resolved single-cell atlas unveils a distinct cellular signature of fatal lung COVID-19 in a Malawian population 45

2.1 Abstract 45
2.2 Introduction 46
2.3 Methods 48
2.3.1 Patient Recruitment 49
2.3.2 Minimally invasive autopsy 50
2.3.3 Luminex Multiparameter Cytokine Assay 51

2.3.4 Dissociation of lung cells from frozen samples and single nuclei preparation . . . 51
2.3.5 Single cell and single nuclei partitioning and library preparation . . 52
2.3.6 Single cell processing . . . 53
2.3.7 Hashtag Demultiplexing . . . 57
2.3.8 SNP splitting of multiplexed runs . . . 57
2.3.9 Lung Integration . . . 58
2.3.10 Pseudo-bulk . . . 59
2.3.11 Exploring viral reads in samples . . . 60
2.3.12 Gene panels defining the IFN-γ response . . . 60
2.4 Results . . . 61
2.4.1 Cohort overview . . . 62
2.4.2 scRNA mapping results . . . 63
2.4.3 scRNA quality control results . . . 75
2.4.4 Pulmonary cell scRNA-seq reveals low levels of viral RNA and an IFN-γ dominated response in the Malawi cohort . . . 77
2.4.5 Integration with Human Lung Cell Atlas (HLCA): IFN-γ driven responses in Malawi cohort and type I/III interferon responses in other cohorts . . . 84
2.4.6 Single-cell analysis of nasal cells may be a useful proxy for lung parenchymal responses . . . 90
2.4.7 Stromal cellular interactions are driven by macrophages and vascular interactions by neutrophils . . . 96
2.5 Discussion . . . 103
2.6 Supplemental tables . . . 110

3 cellXplore: a web tool to interactively explore cellular interactions at the single cell resolution 112

3.1 Abstract . . . 112
3.2 Introduction . . . 113
3.3 Methods . . . 116

3.3.1 Legacy cellXplore architecture overview . . . 116
3.3.2 cellXplore Software Architecture . . . 118
3.3.3 Development protocol for cellXplore . . . 120
3.3.4 Input data requirements . . . 121
3.3.5 Case study dataset preprocessing . . . 123
3.4 Results . . . 125
3.4.1 Reanalysis of *T.brucei* dataset . . . 126
3.4.2 Analysis of cellular interactions in active *Trypanosoma brucei* infection using the legacy cellXplore . . . 129
3.4.3 Moving away from legacy cellXplore to the current cellXplore . . . 136
3.4.4 Case study: Workflow 1 using *T.brucei* infection to examine microglia-plasma cell cross talk . . . 147
3.4.5 Case study: Workflow 2 confirming interactions between astrocytes and microglia in *T.brucei* infection . . . 151
3.4.6 Case study: Workflow 3 using patient matched single cell and Xenium of a Breast Cancer tumour . . . 158
3.5 Discussion . . . 167
3.6 Appendix . . . 173

4 Dissecting cellular interactions in big data: Contextualising cellular interactions using atlas-level single cell and sequencing based spatial transcriptomics 176
4.1 Abstract . . . 176
4.2 Introduction . . . 177
4.3 Methods . . . 180
4.3.1 Annotation of the full atlas dataset . . . 181
4.3.2 Cellular interaction inference of the macrophage-fibroblast atlas . . . 181
4.3.3 *H. polygyrus* Visium dataset processing . . . 182
4.3.4 Preparing the intestine single-cell RNA sequencing reference datasets 183
4.3.5 Cell2location analysis of *H. polygyrus* infected and naïve mice intestine . . . 184

4.3.6 Cellular communication inference of *H. polygyrus* infection in the murine intestine . . . 185
4.4 Results . . . 185
4.4.1 Section 1: Identifying cellular interactions in complex atlas level data 186
4.4.2 Expanding the macrophage-fibroblast atlas to the full dataset using SingleR . . . 186
4.4.3 Comparing overlap of cellular interactions across the synovium, lung, skin and heart in homeostasis and disease . . . 195
4.4.4 Focusing cellular inference on lung, skin and synovium in homeostasis and disease . . . 200
4.4.5 Validating interactions in skin and synovium . . . 210
4.4.6 Section 2: Identifying cellular interactions leveraging spatial transcriptomics during *H. polygyrus* infection . . . 213
4.4.7 Integration of *H. polygyrus* Visium datasets across time . . . 213
4.4.8 Differential expression analysis reveals temporal gene expression patterns over infection . . . 215
4.4.9 Cellular interaction analysis shows decreased Wnt signalling during *H. polygyrus* infection . . . 217
4.4.10 Crypts and villi show changes during *H. polygyrus* infection . . . 219
4.4.11 Molecular characterisation of the *H. polygyrus* granulomas and the surrounding tissue niche . . . 221
4.4.12 Using cellular deconvolution to identify spatial niches . . . 227
4.4.13 Cell-to-cell interactions and signalling pathways . . . 229
4.5 Discussion . . . 232

5 Discussion **239**

Acknowledgements

TLDR: Thank you to everyone I have ever come across during my PhD, you have really made it remarkable!

I would like to thank my supervisors first and foremost, in particular Professor Thomas Otto who has guided me through these crazy four years harnessing the emotional chaos with pragmatism every step of the way. I'm not sure if I achieved the optimum level of pragmatism but I think I reached an acceptable pseudo-pragmatism that keeps things interesting. Our journey starts from grilling me in my masters viva to coaxing me back to rainy Glasgow to work as a bioinformatician before starting the PhD, then flying me to sunny Montpellier, then back to rainy Glasgow then writing up in sunny Turkey. A good mixture of sunshine and rainy days if you ask me! I've been provided with so many amazing opportunities from teaching on countless courses, working in amazing collaborations, and visiting lovely places so thank you for not only your guidance academically but also personally. I would also like to thank Professor Mariola Kurowska-Stolarska who has given me tremendous support and guidance throughout the PhD, always making time for me whenever I needed for much needed chitchats. Not to mention the opportunities she's provided over the years to collaborate and develop my immunology knowledge. I would also like to thank Professor Mark Coles who supervised me during my short stay at the Kennedy Institute of Rheumatology at Oxford and supported me allowing to collaborate within his group. I think we will always be somewhat bonded by the thought of sticky shoes on the floor of Jesters nightclub. I'd also like to thank my assessors Professor Simon Milling and Dr. Kevin Bryson who never failed to make time for me when needed, and

provided many entertaining APR meetings which will always stick with me. Next, I owe most of the success of finishing this PhD to my amazing lab group all past, present and adjacent members. I thank Scott Arkison first and foremost as without the invaluable reminder of how not to break servers, run up memory on home drives, installing packages and the like this PhD would not have been possible. I thank you for our Friday morning (or whenever I'd stroll in) chats and bringing a zest to the office that served as such a pick-me-up in gloomier times. I thank Dr. Ross Laidlaw for being my partner in crime since day one, running the scClinic almost every Wednesday afternoon, creating the Advanced Single Cell course, board games, festivals, public engagement and all the rest of it. Thank you also for your unwavering support and guidance throughout the PhD both academic and emotional, I really learnt a lot over my time. I thank Andrew McCluskey for joining me come rain or shine outside the building for de-stress briefs and other associated activities that kept me going throughout my PhD at the best of times. I thank my sweet ladies Yiyi, Joy and Brenda who just forever bring me joy and happiness even in the gloom of Glasgow, never failing to put a smile on my face. I have loved our emotional support system over the course of my PhD and how we always have shared such precious times together. Next, I want to thank my lab group away from home starting with Dr. Lucy MacDonald who has been nothing short of a mentor to me over the course of my PhD. You have supported me in every single way imaginable and taught me so much over my time, giving me so many opportunities along the way. I want to thank Dr. Lavinia Colletto who has been an emotional anchor, always resilient and stable through the constant chaos. Your unconditional support has meant so much to me even and I know how difficult I can be to reach at times so thank you for always being there. I want to thank Jack Frew for being my RACE partner in crime and never failing to put a smile on my face or take a blood sample whichever comes first. I want to thank Dr. Domenico Somma for being everything in both Otto and Kurowska-Stolarska labs. 5 stars, 10/10. You have not only offered me support but continue to spread your pearls of wisdom and joy across so many people, you are and always will be so impressive! I want to thank Dr Ted Simakou for all of our chats over the years bonding over our shared yearning being in a sunnier Mediterranean place. I want to just thank both lab groups completely for being the best group of people I could wish to see most days for four years. Next, I want to thank members that have come and

gone that played such a role in my PhD. Hannah Bialic, I don't even know where to begin. Thank you for being the big ball of energy that you are, not only did we bond through public engagement and our love of science but also literally everything else. I want to thank Dr. Vicky Bolton who opened my world up and was always there for me. I want to thank Dr. Kyle Cunningham for always being there for me during the best of times. Next, I want to thank all the people on level 6 for being amazing and putting up with my cowboy boots clacking up and down the hallway, Barbara, Ruby, Gabriel, Grace to name a few. I want to thank Dr. Marianne Donald and Carla Johnson for our after-work pints and the fun that came with it. I want to thank the Computational Biology community for all the amazing monthly talks and the socials that went with it Alex, Kieran, Jake, Vinny, Fran, Holly, Lucas. I want to thank John Cole for all the amazing opportunities and support he has offered me over my PhD, from teaching to pub quizzes which have soothed the soul. I'd also like to thank Dr. Kathryn Crouch for your support and guidance over the PhD, it has been absolutely invaluable and I'll miss our little pass-by chats in the hallway. I want to thank the teaching staff of the Bioinformatics Masters in particular Graham and Mark who have allowed me to get involved in teaching over the years and I'll miss that dry old humour from you Graham. All in all I want to thank everybody for making my time over the past four years. I have to sum it up otherwise I'll be typing forever. I want to reserve a special thanks to Connor Tennant who has been my lifeline during the PhD where without his support and hours spent in the pub, I wouldn't have been able to be this mentally sound by the end of it. All the chats we connected on made some elements of doing a PhD less lonely and for that I can't be grateful enough. I can't wait to celebrate with you when your time comes. Thank you for everything, you're the best.

Finally, on to my nearest and dearest, I want to thank my mama for always supporting me and trying to ask about what it is exactly that I do. No matter what I've done in life you've always been immensely proud and that really drives me to be the best that I can be. I love you lots and congrats to you too, it was all worth it. I want to also thank my dad who always wanted me to be a doctor, maybe a different kind though… I want to thank

my brother and sister who have ventured up to Glasgow to come and visit from time to time, those times will always be dear to me. I want to also thank Lara and Frankie who have put up with me being so far away for the past 4 years, having to make do with Bristol weekends when I can. You guys are always with me! I want to give the biggest thanks of all to Kıvanç, my heart and soul. You have supported and inspired me constantly and push me to be the best that I can be and have the confidence to recognise it. I love you so much and thank you for all that you do. I want to thank the Eren family especially Selma teyze who has made my writing up experience like no other, making sure I ate well and had everything I needed. I think I have covered everyone, *I hope I have covered everyone.* Thank you!!

Declaration

I declare that, except where explicit reference is made to the contribution of others, that this dissertation is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Olympia Melek Hardy

Abbreviations

- mRNA: Messenger RNA
- RNA-seq: Bulk RNA sequencing
- scRNA: Single cell RNA sequencing
- GEM: Gel bead-in-emulsion
- UMI: Unique molecular identifier
- cDNA: Complementary DNA
- snRNA: Single nuclei RNA sequencing
- QECH: Queen Elizabeth Central Hospital
- SARS-CoV2: Severe acute respiratory syndrome coronavirus 2
- Covid-19: Coronavirus disease 2019
- ARDS: Acute respiratory distress syndrome
- LRTD: Lower respiratory tract disease
- PCR: Polymerase chain reaction
- MITS: Minimally invasive autopsy sampling
- IMC: Imaging mass cytometry
- CHAMPS: Child Health and Mortality Prevention Surveillance
- HLCA: Human Lung Cell Atlas
- SSA: Sub-Saharan Africa
- HIV: Human Immunodeficiency Virus
- CCI: Cell-cell interactions
- L-R: Ligand-receptor

- ST: Spatial transcriptomics
- HAT: Human African Trypanosomiasis
- UMAP: Uniform Manifold Approximation and Projection
- PCA: Principal Component Analysis
- sNN: Shared nearest neighbour
- t-SNE: t-distributed stochastic neighbour embedding
- HVG: Highly variable gene
- DE: Differential expression
- FDR: False discovery rate
- TPR: True positive rate
- FISH: Fluorescence in situ hybridization
- SVG: Spatially variable gene
- GNN: Graph Neural Networks
- RCTD: Robust Cell Type Decomposition
- DWLS: Dampened Weighted Least Squares

Chapter 1

Introduction

1.1 Cellular interaction inference in pathogen mediated disease

The biological motivation of this thesis is emphasising the impact of single cell RNA sequencing (scRNA) and spatial technologies in the inference of cellular communication and how it transforms the research landscape of pathogen-mediated disease. Through leveraging information from scRNA-seq datasets and cells in their spatial context to infer cellular interactions we have uncovered new mechanisms of action, delineated key drivers of disease on both the molecular and cellular level and described new potential targets for immunotherapies. We demonstrate the power of cellular communication inference, modelling ligand and receptor gene expression at the single-cell resolution in COVID-19, inflammatory disease, and parasite infection. Furthermore, this thesis touches on the breadth of tools and visualisation techniques available to interpret cellular communication, including a novel visualisation tool empower bench biologists to explore their data in both modalities.

1.1.1 Cellular interaction inference in COVID-19

Since the COVID-19 pandemic outbreak SARS-CoV-2 infections are still a challenge with approximately 700 million cases recorded and 7 million deaths reported worldwide since 2019[1]. Although the illness is classified as respiratory, the severity in which it affects patients is highly variable and phenotypically demonstrates itself as a systemic disease involving multiple organs showing symptoms such as hyperinflammation, cytokine storms and lung alveolar damage[2–6]. Recent applications of single-cell and spatial transcriptomics have significantly expanded our understanding of COVID-19 pathology, particularly in organ-specific and tissue-contextual settings such as the blood and lung. Early studies during the pandemic utilised single cell transcriptomics to describe the immune landscape of bronchoalvelolar cells taken from BALF fluid and found that the cytokine storm is the main driver of creating a proinflammatory macrophage microenvironment in severe COVID-19 patients, along with notable CD8+ T cell expansion in moderate cases[2]. Atlas-level datasets of the lung in fatal COVID-19 have also been developed such as Melms et al. who found that COVID-19 lungs were characterised by a dense infiltration of monocyte-derived and alveolar macrophages paired with weakened T cell responses[4]. Through cellular interaction inference, they investigated lung remodelling in the stromal-immune interface, describing a TGF-β signalling mechanism that induces lung fibrosis increasing the production of IL1-β. Spatial technologies such as IMC have been utilised to investigate the spatial landscape of the lung unveiling severe immune cell infiltration, infected alveolar epithelial cells and expansion of stromal cells such as fibroblasts[7]. Another study utilised spatial transcriptomics to reveal a conserved immune signalling circuit in severely damaged area of the lung, marked by interactions between IFN-γ-expressing cytotoxic lymphocytes and pro-inflammatory macrophages. This IFN-γ-driven response drove upregulation of chemokines such as CXCL9, CXCL10, and CXCL11, facilitating recruitment of CXCR3+ infiltrating immune cells[8]. More recently, multimodal analysis of the COVID-19 immune landscape have been reported leveraging both single cell and spatial technologies to uncover key cellular interaction pathways. Lee et al.[9] constructed a spatial and single-cell atlas of alveolar damage progression in COVID-19 by integrating

histopathology-defined regions with sc/snRNA-seq and spatial transcriptomics datasets across 33 lung samples[9]. The revealed distinct macrophage subsets associated with acute COVID-19 damage driven IFN-α signatures and collagen-driven proliferative fibroblast activity in late stage lung fibrosis. Through utilising cellular interaction inference, they found a SPP1/osteopontin signalling axis in macrophages as a key intercellular driver in the initial stages of alveolar injury. Together, these studies demonstrate the importance of single cell and spatial technologies that can provide a detailed, spatial map of cellular transitions and signalling pathways that underlie pro-inflammatory and pro-fibrotic processes in COVID-19 lung injury.

1.1.2 Emergence of single cell atlases

Large scale single cell atlases have emerged in recent years not only targeting particular diseases such as COVID-19 but also whole organisms, tissues and pan-tissue cell types. Consortia such as the Human Cell Atlas (HCA) were pioneers in the field of creating single cell atlases that combined multiple studies to standardise atlases that typically contain millions of cells[10]. Since then the emergence of species specific single cell atlases have been established such as the Tabula Sapiens[11], Tabula Muris[12] and the Fly Cell Atlas[13] for humans, mice and drosophila respectively, that span millions of cells across multiple organs. Tissue-focused atlases demonstrate how integrated references enable consensus annotations and case–control comparisons. For example, the Human Lung Cell Atlas[14] spanning over 2.4 million cells in health and disease, while disease-oriented consortia like the Human Tumour Atlas Network[15] provide insights into cellular mechanisms of disease. Single cell atlases have also been curated to provide comprehensive insight into biological processes such as developmental lineages from mouse gastrulation and early organogenesis[16] to comprehensive maps of human foetal gene expression[17]. In addition to providing cell state discovery and disease endotyping, single cell atlases serve as useful resources for reproducible cell type annotation allowing for reference mapping and deep-learning approaches of unlabelled single cell datasets. In addition to this, large

scale integration efforts across multiple studies, donors and modalities allow the development of new tools[18] and for meta-analyses that may reveal additional insights that were previously unreported in the original studies[19]. Furthermore, single cell atlases have been utilised to serve as a ground truth, in the absence of single cell RNA data, for spatial deconvolution using cell type signatures to resolve cells captured in space[20]. Other consortia aim to answer more targeted questions about the mechanisms of inflammatory disease, such as the Immune-Mediated Inflammatory Disease Biobanks in the UK (IMID-Bio-UK) consortium (http://www.imidbio.co.uk/) that aims to identify common pathways in immune-mediated inflammatory diseases such as rheumatoid arthritis, inflammatory bowel disease and psoriasis. In these contexts, cellular communication inference offers unprecedented insight into potential drivers of disease and ligand-receptor interactions that may be inferred from atlas-level data. Cellular interaction inference has been implemented in single-cell tissue atlases to investigate shared mechanisms of action such as exploring interacting cells in the human intestine[21] and their role in intestinal organisation, and across tissues such as unveiling shared and distinct cellular cross-talk in cancer[22]. Thus, by applying cellular interaction inference with atlas-level data, shared or distict pathways and interactions can be explored providing new insights or potential immunomodulatory targets. In addition to this, tailored cell-cell interaction databases have also been developed leveraging single cell atlases, such as scAgeCom[23] which curates its database on cellular interactions inferred from the Tabula Muris Senis[24] and Calico murine ageing cell atlas[25] to provide cellular interactions that are directly involved in ageing and cellular senescence. More recently, cellular communication atlases have been proposed such as CellCommuNet[26] which aggregates multiple single cell atlases and infers cellular communication across tissues, cell types and disease states demonstrating the breadth of knowledge that can be inferred from single cell atlases through cellular communication inference. Collectively, these resources serve not only as reference datasets but also as platforms for comparative analysis, and the development of various computational methods for integration, benchmarking and inference across complex biological systems.

1.1.3 Cellular interaction inference in helminth infections

In addition to viral and inflammatory diseases, cellular interaction inference can also shine light on host-parasite interactions and the mechanism of infection. Helminths are multicellular parasitic worms that chronically infect over a billion people worldwide, contributing significantly to global morbidity through immunomodulation, malnutrition, and tissue pathology[27]. Among them, *Heligmosomoides polygyrus* is a widely studied system used to investigate host–parasite interactions and immune regulation mouse models. *H. polygyrus* is a murine intestinal nematode that establishes chronic infections in the small intestine by inducing a type 2 immune response, characterised by interleukin (IL)-4, IL-5, and IL-13 secretion from Th2 cells, goblet cell hyperplasia, eosinophilia, and alternatively activated macrophages[28]. Orally administered larvae invade the small-intestinal wall, develop and mature in submucosal granulomas and emerge back into the lumen as adults and feed on host intestinal tissue. The adult worms then anchor themselves to the villi, reproduce and lay eggs that escape in the host faeces[29]. In the murine host, there is a complex interplay between the gut epithelia and the immune system. The innate immune response is key to releasing type 2 cytokines that influence the polarisation of the adaptive immune system and gut epithelial physiology[29]. Dendritic cells are the main innate immune cell responsible for priming Th2 responses against active helminth infection, with studies showing a compromised Th2 response during infection when this population is depleted[30]. Another key player are alternatively activated macrophages that have high expression of Ym1, RELM-α and arginase-1 in response to helminth-driven Th2 responses[29]. In the gut epithelia, during *H. polygyrus* infection, epithelial-sensing of the parasite invasion releases alarmins such as IL-25 and IL-33, which activate group-2 innate lymphoid cells (ILC2s) and primes type 2 immune responses[31]. In addition to this, IL-25 derived from tuft cells drives ILC2 production of IL-13, which in turn expands tuft cells and goblet cells creating a 'weep-and-sweep' defence to tackle parasitic infection[32]. In addition to epithelial-immune cell effects, the parasite infection is a potent inducer of regulatory T cells and is commonly used to explore mechanisms of immune tolerance, host–microbiota–parasite dynamics, and mucosal immunology[33]. In addition to the immune system, a hall-

mark of *H. polygyrus* infection is strong immunoregulation by the parasite to the host. The parasite produces excretory–secretory (HES) products that directly modulate host pathways to favour survival during chronic infection. An example of this is the molecular mimicry of TGF-β. The secreted TGF-β mimic (Hp-TGM) engages TGF-β receptors and CD44 to induce regulatory T cells and dampen anti-parasite immunity[34]. Research into *H. polygyrus* infection has been instrumental in uncovering how helminths manipulate host immunity to promote long-term survival and tissue homeostasis. However, to date the use of single cell and spatial technology to investigate the host response in different tissue microenviroments is limited. Haber et al. conducted a single-cell transcriptomic survey of the mouse small intestinal epithelium and assessed how its cellular composition changes in response to various infections, including *Heligmosomoides polygyrus*[35]. They found that the infection significantly increased the abundance of goblet cells, and tuft cells, epithelial cells implicated in type 2 immune responses and mucosal defence. However the study, although profiled the stromal cell landscape of the infected gut, did not investigate cell-cell interactions that may be associated with remodelling of the gut epithelium during parasitic infection. Thus, leveraging cellular communication inference at the single cell transcriptomic level remains an exciting prospect in not only helminth research but also extracellular parasite host-response mechanisms.

1.2 Introduction to single cell RNA sequencing

To understand the molecular mechanisms of cellular responses, we can assess cells in a variety of ways, such as interrogating genomic DNA sequences, messenger RNA (mRNA) sequences, and protein expression[36]. Biological processes are underpinned by protein-protein interactions however, simultaneously capturing the thousands of proteins expressed by the genome in a single cell, more commonly known as the proteome is still challenging. Thus,

as a proxy for protein expression, we can analyse the transcriptome, a collection of messenger RNA molecules, whose expression can be indicative of protein expression and can be extrapolated to describe cellular phenotypes and cell states[36]. The field of transcriptomics has provided valuable biological insight into gene expression patterns through the use of hybridisation microarrays to the emergence of ultra-high-throughput sequencing techniques such as bulk RNA sequencing (RNA-seq)[37]. Although all cells in the body share nearly identical genomes, each cell expresses only a subset of genes, resulting in distinct transcriptomes in different cell types[38]. This heterogeneity of transcriptomes is further exemplified in similar cell types that occur within different environmental niches, cellular processes, and perturbation states[39,40]. Thus, to capture the stochastic nature of gene expression between cells, conventional bulk RNA sequencing is limited as it only provides an average expression profile for a set of cells, missing the inherent heterogeneity of cell-cell variability[38,41]. This led to the first single cell RNA sequencing (scRNA-seq) study in 2009, which revolutionised molecular biology by allowing the measurement of transcriptome profiles for individual cells at an unprecedented scale and resolution[42]. Here, tissues are digested during the single-cell dissociation step, followed by single-cell isolation to profile the mRNA in each cell separately[43]. Two main approaches for scRNA-seq were developed, the first being plate-based approaches, such as Smart-seq, which isolate individual cells into microwell plates and enable full-length transcript coverage with high sensitivity. This approach is limited in throughput, requiring lower cell numbers; however, the increased sensitivity facilitates the discovery of rare cell-type populations[43–46]. In contrast, microfluidic droplet-based methods, including Drop-seq and 10x Genomics Chromium, scale up single cell profiling by encapsulating thousands of cells with uniquely barcoded beads in nanolitre droplets[47,48]. These platforms provide a cost-efficient and highly scalable approach to scRNA-seq, but typically capture only the 3' or 5' ends of transcripts with lower sensitivity for rare genes. In this introduction, we will focus on the 10x Genomics Chromium microfluidic system that enables massively parallel scRNA-seq through gel bead-in-emulsion (GEM) technology, as this approach is the most commonly adopted and was used to generate the data described in the thesis. With this technique, individual cells are encapsulated with barcoded gel beads and reverse transcription reagents within the droplets. Inside each droplet, the cell is lysed, releasing mRNA molecules

that hybridise to bead-bound oligonucleotides containing a cell-specific barcode, a unique molecular identifier (UMI), and a poly-dT sequence. This process uniquely tags each transcript with its cell-of-origin and molecule-specific information, enabling the quantification of transcripts. After reverse transcription, the emulsion is broken and barcoded cDNA is recovered, amplified, and prepared into sequencing libraries for Illumina platforms[36,48].

However, these technologies come with limitations, namely that single-cell measurements of transcriptional states inherently carry greater uncertainty than bulk RNA-seq, primarily due to the limited amount of starting material available per cell[49]. Because only a small fraction of transcripts present in a cell are captured during sequencing, technical biases and noise give rise to a high noise-to-signal ratio. High-throughput droplet-based techniques typically recover only 5–20% of a cell's RNA content, while plate-based methods achieve higher capture efficiencies of 30–40%[46,50]. The capture efficiency of poly-adenylated mRNAs and subsequent conversion and amplification of cDNA is still an open-ended problem[51]. This directly impacts the detection of lowly expressed genes whereby despite the gene being expressed, it is missed by current scRNA-seq technologies leading to drop-out events where zero counts for a gene occur[36]. One factor in scRNA-seq that can mitigate this effect is altering the sequencing depth of the run by increasing the number of reads in the run. For tasks such as un-biased cell type classification the standard is to sequence between 30,000 reads per cell. If the sequencing depth is on the lower end of this range, populations can be described however, granular details such as gene co-expression, cellular communication and regulatory networks will be missed[36]. Other strategies exist to reduce techniclal biases and improve true gene detection such as spike-ins where some protocols can adopt the use of a known mixture of poly-adenylated mRNAs. Through leveraging the read-out from the spike-ins we can assess how much variation is derived from technical artefacts, batch effects and true biological signal[52,53]. However, spike-ins suffer from degradation and capture efficiency, thus UMIs were introduced in 3' sequencing technologies such as 10X Chromium to reduce amplification bias and to facilitate more accurate estimation of true molecule counts[54], rendering spike-ins redundant. The concept of this is that the cDNA is amplified before sequencing to increase its probability

of being measured, the UMI tag allows us to distinguish between amplified copies of the same mRNA molecule and reads that are derived from separate mRNA molecules that are transcribed from the same gene[55]. Despite these strategies, the number of detected transcripts can still vary dramatically between cells, resulting in some cells having much less reliable expression profiles than others. For example, most genes are transcribed in a short amount of time followed by periods of inactive transcription, a phenomenon called transcriptional bursting, which can cause variation in the total transcripts expressed for a given cell[56]. Thus, as a result of this temporal transcriptional fluctuation, the sparsity of single cell data are attenuated and lead to a high frequency of drop-out events where transcripts are not detected resulting in a large number of zero counts[36]. In addition to this, cell size variation can also affect the absolute number of transcripts detected in each cell, further complicated by cellular processes such as cell cycle and proliferation[53,57,58]. However, these intrinsic uncertainties at the single cell level can be mitigated by computational approaches that can take advantage of the information shared between cells, from evaluating cells at the read-level post-sequencing and at the gene-level post-analysis, resulting in a more robust and well-resolved view of the underlying transcriptional landscape[49].

1.3 Introduction to scRNA-seq analysis

The next section will describe the key steps in the computational pipeline to analyse scRNA-seq data and how they handle detection of both technical biases and biological variation shown in Figure 1.1. The pipeline can be divided into three main sections, data preprocessing steps, identification of different cellular compartments and compositions, and downstream functional analysis that elucidates molecular mechanisms of action.

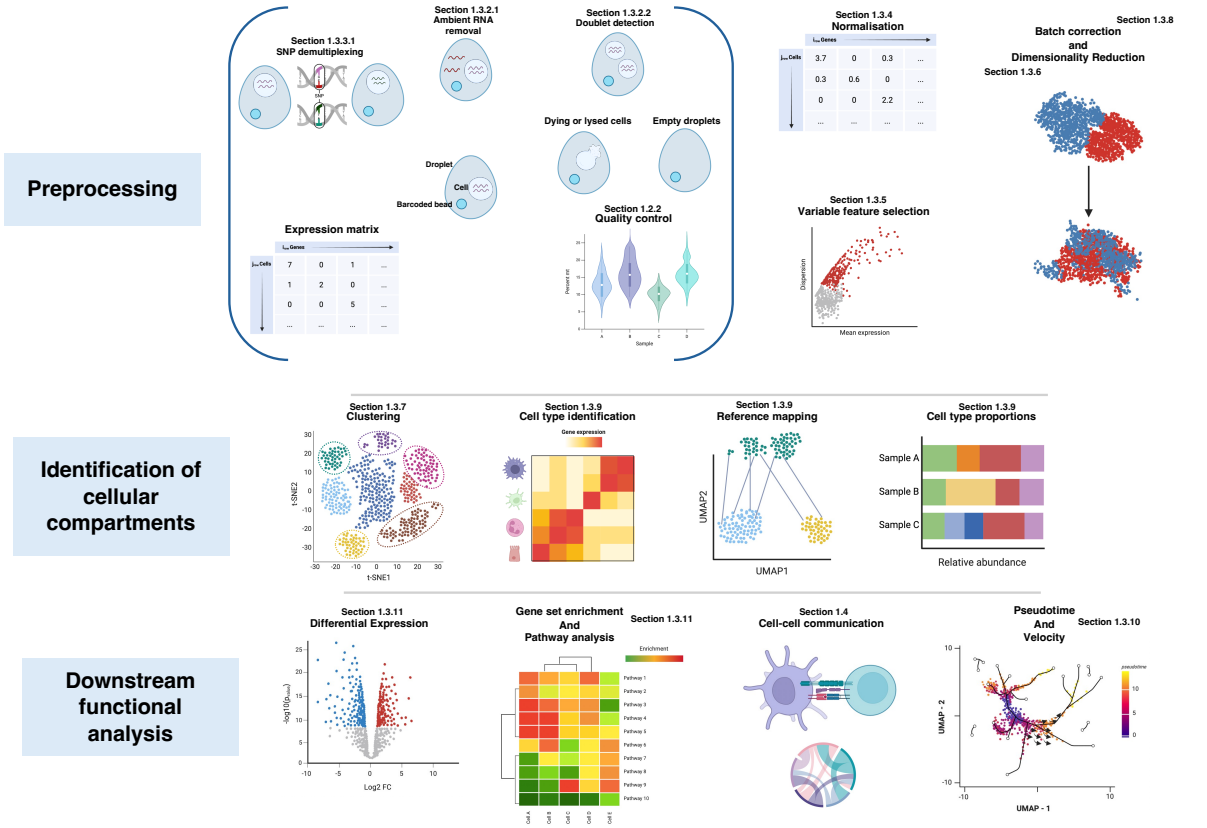


Figure 1.1: Graphical schematic showing the key analysis steps of a single cell RNA sequencing pipeline. Preprocessing of single cell data encapsulates steps such as various quality control steps shown within brackets (Quality Control, SNP Demultiplexing, Doublet Detection, Ambient RNA Removal), normalisation, variable feature selection and batch correction and dimensionality reduction. Identification of cellular compartments comprises clustering, cell type identification, reference mapping and cellular composition. Downstream functional analysis refers to differential gene expresson, functional gene set enrichment and pathway analysis, cellular communications and pseudotime and velocity. Overview figure created in PowerPoint, using icons from the BioRender library.

1.3.1 Pre-processing and Quality Control

Single cell RNA data are subject to inherent and random noise that can obscure the true biological signal; therefore, adequate preprocessing of the data is necessary to remove confounding sources of variation. Once reads have been mapped to a reference genome and the UMIs have been quantified, the output is a digital matrix containing UMI counts that represent gene expression of each gene in every cell. The next stage is to identify low-quality cells in each sample or replicate that can be evaluated using three covariates,

the number of counts per cell, the number of captured genes per cell, and the proportion of mitochondrial reads per cell[59,60]. The distributions of these can be visualised using various plots to determine a threshold cut-off to exclude outliers that may suggest low-quality cells in the data. For example, cells that exhibit low sequencing depth, a limited number of detected genes, and a high proportion of mitochondrial gene expression often indicate damaged or low-quality cells. In such cases, cytoplasmic mRNA may have leaked due to the cell membrane being damaged during the sequencing, leaving predominantly mitochondrial mRNA intact[55]. It is important to consider all three factors in unison when determining outliers, as observing any one aspect in isolation could result in misinterpretation of the underlying biology. For example, cells with low counts or expressed genes could indicate quiescent populations or cell types with inherently lower transcriptional profiles, such as neutrophils[61,62]. Likewise, cells with higher mitochondrial counts could indicate metabolically active cells that may be taking part in respiratory processes[63]. Furthermore, cells with higher transcriptional counts could be explained with the size of the cell, as mentioned previously. Thus, the underlying biology of the system needs to be considered so that permissive thresholds can be selected to avoid missing biologically meaningful cells. Lastly, we can filter on the gene level after we have interrogated cell-level quality metrics. Raw count matrices for humans and mice can contain more than 65,000 genes with 20,000 protein-coding genes, not all of which will be expressed in many cells and thus are deemed uninformative of the underlying biological system or cellular heterogeneity. These uninformative genes can be filtered away, for example, filtering out genes expressed in fewer than 10 cells across the dataset. For datasets with high drop-out rates this parameter should be adjusted accordingly to not exclude too much of the data[55].

1.3.2 Advanced Quality Control

1.3.2.1 Ambient RNA Removal

There are additional quality control measures we can take to ensure we are only selecting quality cells for our downstream analysis. Technical noise can arise from cell-free RNA that is present in the cell solution and assigned to another cells native RNA during library construction, termed ambient RNA[43]. This can lead to complications in cell type identification as transcripts of different cell types can contaminate the true cell type signature of another cell, potentially leading to ambiguous cell type populations downstream[64]. Additionally, noise from cell-free RNA can mask true biological signal and lead to misinterpretation of gene expression profiles so therefore should be removed. Ambient RNA removal methods exist such as SoupX which aims to estimate ambient RNA contamination by leveraging both the raw matrices that have empty droplet and background noise information, the filtered matrices containing true cells and clustering information output by the CellRanger mapping algorithm[64]. Another algorithm, CellBender, uses an unsupervised Bayesian model that requires no a priori knowledge of cell expression profiles[65]. Ambient RNA removal can be performed as a first step in the computational pipeline as both tools output a corrected expression matrix where background noise derived from ambient RNA have been removed, and can therefore be used for all downstream analyses.

1.3.2.2 Doublet Detection

Another quality control consideration is the simple assumption in scRNA-seq analysis is that each droplet contains the RNA of an intact cell. However low-quality cells can violate this assumption whereby droplets can fail to capture a cell, capture multiple cells or capture ambient RNA from a lysed cell[43]. During the encapsulation process of a cell in the nanolitre droplet, cells may fail to be captured leading to an empty droplet that will

demonstrate zero-counts in the expression matrix. These will usually be filtered out using the threshold strategies by selecting a lower cut-off for the number of transcript counts. However in some cases the droplet can capture two cells termed doublets that result in an increased transcript count compared to that of a droplet that has captured a single cell. These can be removed using the upper threshold limits however there are computational methods that have been created to aid in doublet detection. Doublets can be categorised into two groups, heterotypic doublets that are doublets formed by two different cell types and homotypic doublets, formed by the same cell type. Most doublet detection methods iteratively generate artificial doublets by randomly sampling cells and combining them, then compare them against measured cells with a priori knowledge[43]. Popular doublet detection packages used in single cell analysis are DoubletFinder[66], Scrublet[67], scDblFinder[68] and DoubletDecon[69] that work on these principles labelling each cell in the data as either a doublet or a singlet and are performed on each sample before further downstream processing.

1.3.3 Additional preprocessing steps

1.3.3.1 SNP Demultiplexing

Depending on experimental design there may be contexts where additional preprocessing steps must be performed. One example is this is pooling samples together to reduce technical variation in scRNA-seq workflows. When multiple patient samples are pooled into a single scRNA-seq run to reduce costs and minimise technical variation, their donor identities can be recovered computationally using SNP-based demultiplexing methods. Tools such as Souporcell[70] leverage single nucleotide polymorphisms (SNPs) expressed in RNA to cluster cells by genotype without requiring prior donor genotypes. First, Souporcell identifies candidate SNPs from pooled transcriptomic reads and constructs an allele count matrix that records the number of reference and alternative alleles detected in

each cell. It then applies a probabilistic mixture model to group cells with similar allele fractions into distinct genotype clusters, corresponding to individual donors. Cells that display mixed SNP profiles spanning two clusters are flagged as sample-level doublets. Although its accuracy depends on sufficient SNP coverage from sequencing depth and enough genetic diversity between donors to accurately genotype donor samples. Alternative SNP demultiplexing tools also exist such as Vireo[71] that uses a variational Bayesian inference model to infer genotypes and methods like Demuxlet that require known donor genotypes a priori from SNP arrays or whole exome sequencing to match SNPs to the donor reference[72]. Once these steps have been completed we can be confident that the cells we proceed with are of sufficient quality, have been demultiplexed if necessary and de-noised for technical variation. However, it is important to note that quality control in scRNA-seq is an iterative process and cell quality cannot be fully determined a a priori. Downstream steps such as clustering or cell type annotation may indicate that quality control thresholds or methods may need to be revisited[55].

1.3.3.2 CITE-seq

In 2017 Stoeckius et al. proposed Cellular Indexing of Transcriptomes and Epitopes by sequencing (CITE-seq), a multi-modal single-cell profiling technology that enables the simultaneous measurement of RNA expression and surface protein abundance in the same cell[73]. CITE-seq overcomes the inability of scRNA-seq in robustly quantifying cell surface protein levels, which are often functional mediators of cell identity and immune signalling. By implementing oligonucleotide-barcoded antibodies known as Antibody-Derived Tags (ADTs), they are captured during standard droplet-based library preparation in scRNA-seq and sequenced alongside mRNA to produce a unified multi-modal single-cell molecular profile. CITE-seq is particularly impactful in single cell transcriptomics, where transcript levels do not always correlate with protein abundance due to proteomic factors such as post-transcriptional regulation, protein turnover, and even cell-state dependent surface marker expression[74]. Thus, the technology has enabled higher-resolution cell type clas-

sification[75], improved identification of rare immune populations and cell states[76,77] that may have been missed from scRNA-seq alone due to poor protein-level correlation or drop-out rates in droplet-based approaches. In addition to this, CITE-seq can be utilised for cell hash tagging approaches (HTO) where unique oligo-antibodies are used against ubiquitously expressed cell surface proteins to derive cells from multiplexed runs to their original tissue or sample, identify cell multiplets, and 'super-load' droplet-based systems to reduce sequencing costs[78]. First, the ADTs are quantified and demultiplexed through a mapping approach to identify antibody barcodes, cell calling based on the number of mRNA reads associated with each barcode, and UMI quantification. Many tools exist for this such as tools embedded in the Seurat pipeline[18] and demultiplexing pipelines such as cellhashR[79]. ADTs are normalised to correct for non-specific binding, batch effects in protein libraries through normalisation methods such as centred log ratio (CLR)[18,73] or de-noised and scaled by background (dsb)[80]. ADT and mRNA levels can then be projected jointly in the same latent space using dimensionality reduction approaches like weighted-nearest neighbours (WNN) that leverage both transcriptomic and proteomic information[18]. Downstream, both differential expression on the transcriptomic level can be conducted and differential abundance analysis on the protein level to comprehensively characterise single cells in both modalities[18].

1.3.4 Normalisation

The next step in the single cell analysis pipeline is to normalise our data so that the gene expression profiles of cells are directly comparable to one another while removing technical variability and preserving biological differences. Technical effects include differences in library size (total number of unique molecular identifiers, UMIs, per cell), capture efficiency, amplification bias, and batch effects from reagents or sequencing runs as previously described[55]. If normalisation is not performed cells with higher total RNA capture would appear to express higher gene levels, and thus bias clustering, dimensionality reduction, and differential expression analyses downstream. Many different methods

have been proposed for normalisation with the most common being library size normalisation adapted from bulk-RNA sequencing where counts are normalised by a uniform scaling factor that is proportional to the count depth per cell. This method divides gene counts for a given cell, multiplies it by a scale factor usually 10,000 then applies a natural log transformation to account for zero-counts. This stabilises variance in the data and corrects for differences in sequencing depth but assumes that most genes are not differentially expressed between cells, meaning that only low or medium expressed genes are accurately normalised[55]. Furthermore, the log-transformation does not fully stabilise the mean–variance relationship, and biases arise if a small number of genes dominate the expression profile of certain cells, distorting the normalisation for other genes[81]. Other methods have been proposed such as scran normalisation which estimates size factors for pooled cells that have similar transcriptional profiles, and then deconvolves them back to individual cells. This strategy reduces noise introduced by dropouts and provides more robust normalisation for heterogeneous cell populations[82]. This can be an alternative to library size normalisation however scran does not explicitly account for high dropout rates, which can introduce noise into the size factor estimation in extremely sparse datasets. Furthermore, although the method corrects for library size differences, it still does not fully stabilise the mean–variance relationship in the data, meaning that a log-transformation step is still required and residual technical variance can remain[81]. To directly address mean-variance stabilisation scTransform was developed which fits a regularised negative binomial regression model for each gene, and regressing out sequencing depth and other technical covariates. It then returns Pearson residuals, which are variance-stabilised and homoscedastic. This method simultaneously normalises and transforms the data, reducing the need for separate scaling steps and has been shown to improve clustering and integration by mitigating technical artifacts while preserving biological heterogeneity[83]. However, there is still open discourse as to which normalisation method to use and should be chosen carefully depending on the nuances of the biological dataset. A benchmark study found that a simple two-step approach of proportional fitting of counts by library size followed by log-transformation with a pseudocount followed by an additional fitting step performed consistently well when compared to more complex model-based approaches[84].

Conversely, another study concluded that variance stabilising methods perform best but more simplistic library size normalisation is sufficient for scRNA-seq normalisation and more complex approaches should only be implemented when the dataset demands it, such as extreme sparsity of the data[85].

1.3.5 Feature Selection and Dimensionality Reduction

After completing normalisation, we must handle a large dimensional expression matrix consisting of thousands of genes and potentially thousands of cells, however many of these genes are uninformative to the biology of the system. Thus, in this high dimensional space, single cell data suffers from the curse of dimensionality where distortion of distances between data points and technical noise can obscure true biological structure[86,87]. Dimensionality reduction mitigates these issues by capturing the most informative sources of variation while filtering out random fluctuations inherent to sparse scRNA-seq data. The first step of reducing the dimensionality of scRNA-seq datasets is feature selection where the dataset is filtered to keep only genes that explain the most variation in the data called highly variable genes (HVGs)[52]. Most methods bin genes in the data according to their mean expression and HVGs are selected by their highest variance-to-mean ratio[55]. These genes ideally drive the biological variation in the data and can be used to separate main subpopulations downstream without impacting the identification of smaller subpopulations[43]. After HVG selection, the dimensions of the data set can be further reduced by dimensionality reduction algorithms that aim to summarise and visualise the expression matrix in a low-dimensional space such as principal component analysis (PCA)[88]. PCA is a linear approach that aims to summarise a dataset via its top N principal components namely orthogonal planes that are statistically uncorrelated perpendicular axes drawn through the gene expression data that captures distinct non-overlapping sources of variability[89].

1.3.6 Dimensionality Reduction for sc-RNA visualisation

Although PCA fails to capture the structure of the data when compared to non-linear methods it is a pre-processing step to pass the embeddings for non-linear dimensionality reduction visualisation techniques such as t-SNE (t-distributed stochastic neighbour embedding)[90] or UMAP (Uniform Manifold Approximation and Projection)[91] and can adequately summarise key differences in the data[55]. t-SNE is widely used for visualizing single-cell RNA-seq data in two or three dimensions and converts pairwise distances between cells into probabilities that reflect similarity, then optimises the distances in a low-dimensional embedding that preserves local neighbourhood relationships while maintaining local structure rather than global distances[55,90]. UMAP is a newer dimensionality reduction method that builds a k-nearest neighbour (kNN) graph of local relationships in the high-dimensional space and then projects it in a low-dimensional embedding[91]. A k-nearest neighbour (kNN) graph is a mathematical representation of the local similarity structure between cells in a scRNA dataset. Each cell is treated as a node and is connected by edges to its k most similar neighbours, usually determined by using Euclidean or cosine distance in the PCA space[92]. The parameter k controls the neighbourhood size, with smaller values emphasising very local relationships and larger values capturing broader structures at the expense of fine-grained resolution. For example, if we set $k = 3$, the kNN graph would take each cell and find its 3 closest neighbours, iteratively forming a distance graph that preserves the high-dimensional topology of the dataset. In contrast to t-SNE, the objective of UMAP is to preserve both local and global structure, facilitating the representation of continuous biological processes, such as cell differentiation[93]. Non-linear dimensionality reduction methods such as t-SNE and UMAP have become the standard in visualising scRNA-seq data however, there is still scepticism about the reliability and interpretability of high-dimensional data projected in low-dimensional embeddings. When projecting high-dimensional gene expression data into two or three dimensions, inevitably we lose information and introduce distortions in both local and global structure. Thus, neighbourhood relationships between cells in the lower-dimensional space often differ significantly from those in the original high-dimensional space[94]. This can lead to

biological misinterpretation of cell types that appear to be transcriptionally similar but are distinct subpopulations. Therefore, although these techniques are the accepted practice, low-dimensional embeddings serve only as a useful visualisation tool and distances should not be solely interpreted as biological similarity.

1.3.7 Clustering

The next stage of the analysis pipeline is to group cells with similar transcriptional profiles to identify distinct cell types, states, or subpopulations within heterogeneous tissues, a process called clustering[55]. Clustering algorithms exist in many flavours such as partition-based methods such as k-means clustering[95] or density-based approaches such as DBSCAN[96]. However, the most popular and widely used are graph-based clustering approaches such as Louvain or Leiden clustering as they are less computationally intensive and are scalable for large scRNA datasets. The Louvain algorithm works by iteratively optimising a modularity function, that defines how well a graph is split into clusters, grouping nodes (cells) into distinct communities that maximize within-cluster connectivity. However, it suffers from a known limitation where some resulting clusters may be poorly connected or fragmented internally, leading to suboptimal partitions[97]. The Leiden algorithm was introduced as an improvement to handle the partial clustering instability in Louvain clustering by adding an additional step that merges and splits sub-clusters based on their internal connectivity[98]. The optimised modularity function includes a resolution parameter, where the user can choose the granularity of the cluster partitions. By altering this parameter the lower the resolution the more granular the clusters will be, the higher the resolution the more finer the clusters will be but are subject to patterns emrging that are noise-driven[99]. Once cells have been clustered, transcriptionally distinct subpopulations are revealed that can then be interrogated at the gene level to discover cellular heterogeneity, differential gene expression and trajectory inference. However, the

results of the clustering can be sensitive to the upstream steps of the analysis such as normalisation, reduction of dimensionality, and the chosen parameters of the methods[55]. Therefore, clustering should be complemented by marker gene validation and biological knowledge to ensure that identified populations are meaningful.

1.3.8 Integration and Batch Correction

The above steps are sufficient when analysing data that may come from a single sample or donor, however, most single cell studies require the analysis of multiple samples deriving from different patient-donors, different preparation protocols or disease conditions. Each of these factors contribute technical differences, also known as batch effects, which need to be corrected for as they can obscure true biological signal and lead to misleading clustering or downstream differentially expressed genes[55]. There have been many different methods developed for integration of scRNA data, the most common being canonical correlation analysis (CCA)[100] and Harmony[101] which perform well for simple integration tasks with straightforward batch effects[102]. CCA implements an anchor-based strategy where pairs of transcriptionally similar cells from the different datasets are used to compute a non-linear transformation to the data that is projected into a shared integrated space[103]. Harmony models batch effects in the PCA space as additive factors and corrects them through aligning cells with similar biological profiles across batches[101]. Deep learning integration methods have emerged such as scVI (single-cell Variational Inference) which uses a variational autoencoder (VAE) to probabilistically model gene expression counts to integrate cells from different batches into a harmonized latent space using batch information as a latent variable[104]. The above are examples of unsupervised integration methods that require no prior labels for the integration task, however there are tools that can leverage this information and perform semi-supervised integration. One example is scANVI[105] which extends scVI by incorporating partial cell type labels for better alignment and label transfer, another is STACAS that leverages cell type label information to integrate datasets with partial population overlap which preserves dataset-specific biology

while still aligning common populations[106]. Despite extensive tool development to handle batch correction in single cell data, benchmarking single-cell integration methods remains challenging due to the lack of universally accepted benchmark metrics to measure the efficiency of the integration. This is mainly derived from the absence of a ground truth in in biological data, making it difficult to objectively evaluate whether integration preserves or distorts underlying biology[107]. Another limitation is the confounding of batch effects with biological variation in datasets. If cell type or condition is correlated with batch, integration methods may remove true biological signals or under-correct for batch effects. Finally, evaluation metrics to measure the efficiency of an integration method can differ substantially and prioritise different aspects of integrations such as batch removal or preserving biological context. There is still ambiguity of which method to use depending on the biological question at hand, and the inherent properties of the scRNA dataset[107,108].

1.3.9 Cell type classification and composition

Once integration and clustering have been performed, we can now identify what cell populations are present in the scRNA data based on their gene expression profile. This process can be performed in a number of different ways that can be categories into unsupervised, semi-supervised, or supervised approaches. In unsupervised classification, the clusters are annotated post-hoc using top marker gene expression, by identifying genes that are highly expressed in each cluster and matching them to known cell type markers from literature or public databases[97]. This approach is flexible and allows for the discovery of novel or unexpected populations, but it depends heavily on correct interpretation, time intensive and is sensitive to clustering resolution. Conversely, supervised cell type classification methods rely on a reference dataset in which the cell type labels are already known. Popular tools such as SingleR[109], scPred[110] typically involve mapping query cells into a shared feature space, for example the PCA space, with the reference and assigning the most likely label based on a threshold similarity. Supervised classification is generally faster and more reproducible than manual annotation, but it depends on the quality of

the reference dataset, and may mislabel cells that are not represented in the reference. Lastly, semi-supervised approaches combine both approaches by using labelled reference data to guide the annotation of shared populations, while allowing for de-novo discovery of novel or dataset-specific cell types. Tools such as scANVI[105] and Seurat's reference mapping[103] allow users to integrate a labelled atlas with new data and transfer cell type labels via kNN graphs in a shared embedding space. These methods also provide confidence scores for predictions, to distinguish confidently labelled cells from ambiguous or novel populations. Once we have identified what cell type populations are present in our data we can visualise this in terms of relative proportions to understand what cell type populations are changing across the different co-variates of our data such as disease condition or patient-to-patient variability.

1.3.10 Pseudotime and velocity

For datasets that capture a dynamic biological process such as cell type development or parasitic life cycles we can implement methods such as pseudotime and RNA velocity to reconstruct and order cells along the process using their transcriptional profiles. Pseudotime refers to a latent, continuous variable of arbritary time that approximates the temporal ordering of cells along a biological process—such as differentiation—based on their transcriptional similarity[111]. Pseudotime algorithms typically leverage the low dimensional space of scRNA data like the PCA space and then construct a graph or trajectory that connects cells in this space. Two popular methods are Monocle which constructs a minimum spanning tree to represent cell state transitions and Slingshot that fits simultaneous smooth curves (lineages) through low-dimensional embeddings, using cluster centroids as anchors[111–113]. To extend pseudotime analysis RNA velocity methods can infer the transcriptional dynamics and directionality of each cell using the abundance of unspliced and spliced transcripts[114]. Newly transcribed (unspliced) RNA represents the up-regulation phase of gene expression, while spliced RNA represents the steady-state or decay phase. RNA velocity methods such as Velocyto, scVelo and CellRank can model

splicing kinetics, to estimate the directionality of a differentiation process, identify terminal and initial cell state populations and key genes that drive these processes[114–117]. By leveraging the information from RNA velocity and pseudotemporal ordering we can gain insight beyond the static snapshot of scRNA data and estimate RNA kinetics and cell differentiation processes.

1.3.11 Differential expression and functional analysis

When we have identified cell type populations we can perform differential expression (DE) analysis on the raw normalised count data to identify genes whose expression levels vary significantly between groups, such as clusters, conditions, or pseudotime lineages. Unlike bulk RNA-seq, scRNA-seq data is sparse, overdispersed, and has a high dropout rate, requiring specialised statistical models to account for the gene expression distribution. Common cell-level DE approaches include non-parametric tests such as the Wilcoxon rank-sum test or MAST which is a generalised mixed effect model that observes expression using a hurdle model to account for the bimodal distribution of zero and non-zero values[97,118]. Pseudobulk DE can also be performed where by aggregating counts across cells from the same group or sample can mitigate technical noise, the impact of dropouts and overdispersion[119]. By creating pseudo-bulk profiles from scRNA data we can implement bulk-RNA DE methods such as DESeq2[120], edgeR[121] and limma[122] which yield more reliable p-values and better control of false discovery rates compared to single cell-level methods that often underestimate variance[123]. However, differential expression is an open-ended problem in scRNA analysis and the consensus between different differential expression methods is remarkably low[124,125]. Current methods for DE analysis in single-cell RNA-seq continue have statistical trade-offs between the true positive rate (TPR) and precision. Methods that are optimised for high TPR tend to identify a larger number of genes as differentially expressed, but this comes at the cost of an increased rate of false positives, for example a pitfall of the MAST algorithm. On the other hand, approaches that prioritise high precision may fail to detect subtle differences, resulting in

lower TPR and potentially missing biologically relevant genes[123] as seen in methods that implement pseudobulk counts. Pseudoreplication bias is the key driver of this, where individual cells, rather than biological replicates, are treated as independent observations. This violates a key assumption of standard statistical tests and leads to a substantial inflation of the false discovery rate (FDR), as cells derived from the same individual or sample can share biological and technical variance[126]. Thus, selection of DE methods is critical to ensure robust statistical inference of genes that are true biological signal and not false positives. Once differentially expressed genes are identified, functional enrichment analysis is used to interpret their biological pathways and mechanisms. Tools such as Gene Ontology (GO) enrichment, KEGG pathway analysis, and Reactome pathway mapping can identify over-represented biological processes, molecular functions, and pathways. In single-cell workflows, this is often facilitated by packages such as clusterProfiler[127]. Alternatively, methods such as Gene Set Enrichment Analysis (GSEA) can be applied directly to ranked gene lists, either by log-fold change or p-values, allowing for detection of coordinated gene set expression without relying on strict DE thresholds[128]. For more context-specific analyses, tools such as AUCell[129] and UCell[130] GSVA[131] or Seurat's gene module scoring[132] can score the activity of gene sets at the single-cell level such as interferon responses. For example, the Seurat's gene module scoring function computes a cell-wise average expression of a target gene set, subtracting the average expression of control gene sets with matched expression bins to account for background variation. Newer tools exist such as PROGENy[133], DoRothEA[134,135] and decoupleR[136] that can infer pathway and transcription factor activity based on downstream targets. Together, these methods enable the interpretation of cellular functions and gene regulatory programs that drive transcriptional heterogeneity in single-cell data.

1.4 Introduction to cellular interactions and inference methods

Lastly, another functional analysis that can be completed in scRNA data analysis is the inference of cellular communications, the main focus of this thesis. Cell-cell interactions (CCI) leverage many diverse molecules to coordinate a response across tissues in both homeostasis and disease such as ligands, receptors, structural proteins and metabolites. Signalling pathways driving cell-cell communications are mediated by various protein interactions for example between ligand-receptor, receptor-receptor and extracellular matrix-receptors[137]. Downstream signalling occurs in 'receiver' cells that are expressing a cognate surface receptor to a given 'sender' cell triggering a change in transcription factor activity and gene expression[137,138]. Understanding how the modified gene response ultimately leads to altered interactions between the cell and its native microenvironment, can provide invaluable insight into functional pathways that contribute to disease and developmental processes[137]. CCI inference can be achieved by using the gene expression levels of a given ligand and its corresponding receptor as an indirect measure of protein expression. Many computational tools have been developed utilising manually curated L-R databases and statistical methods to quantitively evaluate the probability of two cell types interacting based on this assumption (Figure 1.2).

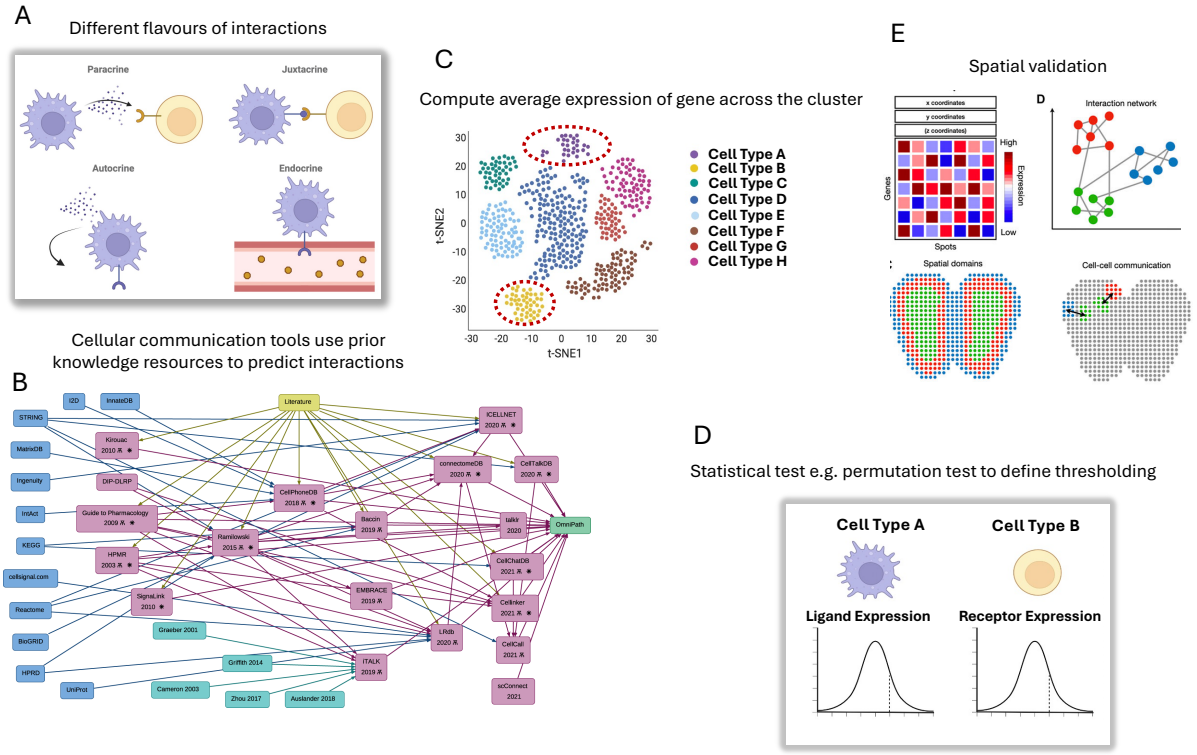


Figure 1.2: Graphical schematic showing an overview of cellular interaction inference. A) Showing the different types of cellular interactions, paracrine where the cell releases ligand received on the receptor of another cell. Juxtacrine where the interaction is cell-to-cell contact. Autocrine where the cell is releasing ligand that is received on its own receptor. Endocrine where the ligand is released into the vasculature and is received by the receptor of a cell. B) Figure demonstrating the interconnectivity of cellular interaction databases that contain a priori knowledge, figure taken from Dimitrov et al[139] C) Schematic showing the expression is taken over the average expression of a cluster. D) Statistical test usually based on a thresholding of expression of ligand and receptor in cell types and assigned a significance from techniques such as permutation-testing. E) Spatial validation where applicable by projecting expression into space. Overview figure created in PowerPoint, using icons from the BioRender library.

1.4.1 L-R Databases

Ligand–receptor (L-R) interaction databases provide the foundation of cellular inference by providing a priori knowledge from curated experimentally validated or computationally predicted ligand–receptor pairs. These databases often include additional metadata about ligand receptor pairs such as complex formation, subunit composition, and func-

tional pathway information. There are general biological databases such as Reactome[140], KEGG[141,142] and STRING[143] that feed into the general framework of ligand-receptor databases but also manually curated databases such as CellPhoneDB[144], Ramilowski (FANTOM5)[145] and CellChatDB[146]. Moreover, OmniPath[147] provides a cellular interaction database that encompasses all the mentioned database sources, allowing the user to select particular databases of interest, or filter for relevant interactions. Many interaction inference methods incorporate one or multiple L-R databases, usually tailored for mouse or human interactions.

1.4.2 scRNA-seq cellular inference methods

The most popular cell-cell communication inference tools use a permutation-based approach to identify putative ligand-receptor interactions. Firstly, a communication score for each L-R pair is computed based on the average expression of a given ligand and receptor gene across a cell cluster and then a significance value is obtained through cluster label permutation. Popular permutation-based packages like CellPhoneDB[148], CellChat[146] and ICELLNET[149] also consider multimeric protein complexes, a biological aspect of CCI that had not addressed. Network based methods use properties of connections between genes involved in ligand-receptor signalling to build networks of interaction relationships. NicheNet[150] and SoptSC[151] not only infer CCI through gene expression networks but also take into account co-expression of downstream signalling target genes when generating an interaction score for a given L-R. To handle more complex experimental designs such as multi-patient multi-condition cohorts NicheNet has been extended to MultiNicheNet[152] which leverages differentially expressed ligand-receptor interactions while taking into consideration experimental design such as multiple samples, conditions and batches. Recently, tensor-based tools have emerged, for example scTensor[153], which uses a complex mathematical model to predict L-R interactions from matrix operations. The tool considers all interacting cell pairs simultaneously to model a many-to-many relationship of L-R interactions that can span across multiple cell-type pairs, rather than alternative packages

that assume L-R co-expression is one-to-one[137,153]. Furthermore, emergence of CCI tools that aim to putative L-R interaction from trajectory inference methods may allow insight into how cell communication is changing over time or through dynamic developmental processes. TraSig[154] aims to utilise scRNA information to characterise CCI over dynamic cell processes such as immune response modelling whereby using average expression of L-R genes is inadequate to reflect changes in cellular communication along a trajectory. Most popular cellular inference tools are written in R, however LIANA+[155] a Python based cellular inference tool aims to bring together tools and databases in a comprehensive Python framework where the user has the flexibility to choose a singlular or multiple cellular inference methods and compare their agreement similarity and overlap. Although CCI tools are becoming more popular in facilitating mechanisms of disease one of the biggest caveats is that protein abundance does not correlate directly with transcript expression. CCI packages assume that gene expression reflects protein abundance and that from this inferred protein abundance we can relate it to protein-protein interaction (PPI) strength ignoring aspects such as post-translational modifications. This aspect has been mitigated by packages such as NicheNetR and CellChat that use protein expression datasets to optimise their framework. Another drawback to inference of CCI from gene expression is that cell signalling is spatially constrained, a pivotal dimension that is not preserved in scRNA-seq data. Interacting cells are usually in close proximity to each other due to limited spatial diffusivity of the expressed ligand, or to achieve activation through physical contact with adjacent cells[138]. An overview of the tools mentioned here is shown in Figure 1.3.

Tool	Method	Interaction inference	Features	Language
CellphoneDB	Permutation-based	Computes L-R score and assigns significance using cluster-label permutation	Includes multimeric protein complexes and has a comprehensive database	Python
CellChat	Permutation-based	Computes L-R score and assigns significance using cluster-label permutation	Includes multimeric protein complexes and contains extensive visualisation functions	R
ICELLNET	Permutation-based	Computes L-R score and assigns significance using cluster-label permutation	Can be used on bulk and single cell RNA datasets	R
NicheNet	Network-based	Constructs gene networks and downstream targets to prioritise L-R	Leverages downstream targets to infer the importance of a cellular interaction	R
SoptSC	Network-based	Constructs gene networks and downstream targets to prioritise L-R	Leverages downstream targets and builds cell-level and pathway level interaction networks	MATLAB/R
MultiNicheNet	Network-based	Uses differential expression accounting for co-variates then constructs gene networks and downstream targets to prioritise L-R	Performs differential expression to account for multi-sample/multi-condition experimental designs	R
scTensor	Tensor-based	Implements tensor/matrix operations to infer cellular interactions	Considers many-to-many relationships	R
TraSig	Trajectory-based	Uses trajectories and pseudotime to order interactions	Models dynamic CCI processes	Python
LIANA+	Framework	Wrapper for multiple CCI packages and databases	Allows freedom to choose/compare different CCI methods and databases	Python

Figure 1.3: Table showing an overview of mentioned cellular interaction inference methods. Including the statistical basis for the tool, how it infers interactions, features of the tool and the implemented language.

The output and visualisation strategies of cellular interaction inference tools are widely conserved and are shown in Figure 1.4. Cellular interaction results are presented in a tabular data frame and contain fields such as sender/source, i.e. the cell type that is expressing the ligand, receiver/target i.e the cell type that is expressing the receptor, the ligand/receptor, an interaction metric such as mean expression of the interaction or communication probability, a p-value for significance and functional information. Popular visualisations of cellular interactions consist of a dotplot or circos plot showing interacting cell types and ligand-receptor pairs, a heatmap showing the number or strength of interactions. Other visualisations can use low-dimensional or spatial embeddings such as projecting the ligand-receptor expression back to the UMAP or spatial coordinates.

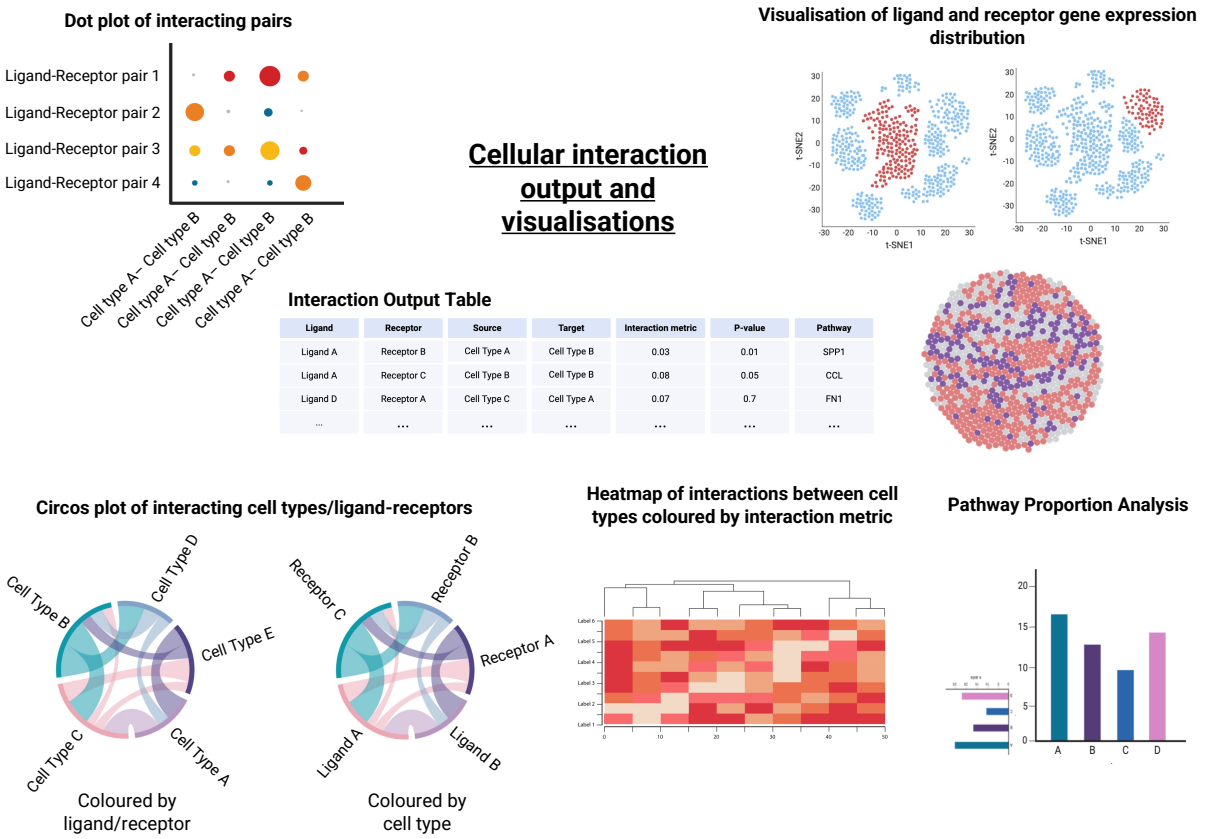


Figure 1.4: Graphical schematic showing the key output and visualisations of cellular interactions. Output of interactions is a tabular data frame containing fields pertaining to quantitive metrics of each interaction per row. Visualisations include a dotplot, circos plot, heatmap, Upset plot, and contextual visualisation such as ligand/receptor expression in space or in the UMAP space. Overview figure created in PowerPoint, using icons from the BioRender library.

1.5 Introduction to spatial technologies

1.5.1 Spatial Transcriptomics

Spatial transcriptomics (ST) and proteomics enable the mapping of gene and protein expression within the spatial context of intact tissues, allowing us to molecularly profile cells in their native spatial context. Early low-plex methods such as RNAscope use fluorescence in situ hybridization (FISH) to detect a small number of transcripts (typically less than 20) at a single-molecule resolution, facilitating validation of specific targets such as ligand and receptor expression with high precision[156]. Similarly, low-plex proteomic techniques such as imaging mass cytometry (IMC) extends multiplexing to approximately 40–50 proteins by combining metal-conjugated antibodies with laser ablation and mass spectrometry, allowing for spatial proteomic profiling at subcellular resolution[157]. These techniques offer a mechanism to cross-validate findings from cellular inference analysis derived from single cell transcriptomics at a high spatial resolution but are limited in throughput and available targets. High-plex spatial methods have emerged that can profile thousands of genes simultaneously. Technologies such as 10x Genomics Visium combine spatially barcoded oligonucleotide arrays with RNA-seq and histological imaging to capture the transcriptome across tissue sections at near-single-cell resolution, enabling whole-transcriptome analysis while preserving spatial information[158]. Each Visium slide contains four capture areas, each comprising of 5,000 spots (each $\sim$55 µm in diameter and spaced $\sim$100 µm apart), where each spot is densely packed with oligonucleotides that contain a spatial barcode, a unique molecular identifier (UMI), and a poly(dT) tail. The tissue is then permeabilised to release the mRNA while preserving tissue structure. Released transcripts hybridise to the poly(dT) capture probes on the spots that are located directly underneath the tissue. The spatial barcode on each probe thus tags the mRNA according to its

position in the tissue so that when the output gene expression matrices are reconstructed, each row represents a gene and each column a spot. Then by overlaying the histological image and spot coordinates, we can visualise whole transcriptome spatial gene expression patterns across the tissue[159]. Other pseudo-bulk high-plex spatial platforms exist such as NanoString's GeoMx Digital Spatial Profiler (DSP) that uses UV-cleaving oligonucleotide barcodes from selected regions of interest (ROIs) to detect RNA or protein probes[160]. NanoString's CosMx Spatial Molecular Imager (SMI) and 10x Genomics Xenium are the latest generation of in-situ spatial transcriptomics platforms that enable high-plex RNA profiling with true single-cell and subcellular resolution. Both platforms interrogate hundreds to thousands of RNA targets directly in intact tissue sections without requiring tissue dissociation or capture arrays, thus preserving spatial context while achieving cellular granularity. CosMx operates using a multiplexed fluorescence imaging strategy, in which RNA targets are detected by hybridisation of oligonucleotide probes that contain sequence-specific barcodes. These barcodes are revealed over multiple iterative cycles of fluorescent imaging and probe stripping, with each cycle detecting a subset of targets using combinatorial barcoding. CosMx currently supports up to 6,000 genes per assay in both fresh-frozen and FFPE tissues, with additional support for multiplexed protein detection[161]. In contrast, Xenium employs a hybridisation-based cyclic readout that detects RNA molecules using padlock probes followed by rolling-circle amplification (RCA) to generate spatially localised fluorescent amplicons that are then decoded over multiple imaging rounds[162]. Tissue sections are stained with DAPI and optional membrane markers, and image-based segmentation is combined with spatial transcript quantification to output cell-by-gene matrices with spatial coordinates.

1.5.2 Spatial Proteomics

Spatial proteomics encompasses a diverse set of in situ protein profiling technologies, many of which evolved from immunohistochemistry and iterative imaging techniques[163]. Cyclic Immunofluorescence (CycIF) implement repeated rounds of antibody staining and imaging to measure up to 40–100 proteins while preserving tissue morphology[164]. CODEX (Co-Detection by Indexing) uses DNA-barcoded antibodies and iterative fluorophore read-outs to achieve 50–60 protein targets without destruction of the tissue, allowing for deep mapping of cell types in a range of FFPE or fresh-frozen tissue samples[165]. Other techniques combine spectrometry with antibody binding such as Multiplexed Ion Beam Imaging (MIBI-TOF)[166], that quantifies proteins in their spatial context by using isotope-labelled antibodies detected by time-of-flight mass spectrometry, providing subcellular resolution for 40-plex panels. Another method, Imaging Mass Cytometry (IMC) similarly uses metal-conjugated antibodies but detects them using laser ablation coupled to mass cytometry, allowing the profiling of 30–40+ proteins at cellular resolution with minimal spectral noise deriving from autofluorescence[157]. These techniques have allowed the proteomic profiling of cell type compositions of tumours[167], psoriatic arthritis[168] and neurodegenerative disease[169]. More recently, high-resolution imaging mass cytometry (HR-IMC) has emerged to spatially profile proteins within their subcellular spatial context such as nuclei and mitochondria[170]. This platform detected chemotherapy-induced perturbations of patient-derived ovarian cancer cells that was previously undetected with conventional IMC[170]. Spatial proteomic technologies however, share a major caveat in the limited number of antigens they are able to profile, which is unrepresentative of the complete complexity of the proteome in a cell or tissue. When considering protein processes, the antigens profile may still not consider biological processes such as alternative splicing and post-translational modifications[171,172].

1.5.3 Additional analysis steps in spatial technologies

Spatial technologies, both transcriptomic and proteomic, introduce some analysis steps that differ from standard single-cell RNA-seq (scRNA-seq) due to the addition of the spatial tissue axis. While both modalities begin with normalisation, dimensionality reduction, and clustering, ST data require additional image integration and spatial quality control steps (Figure 1.5). In this section 10X technologies will be focused on as these have been used in the thesis, however fundamental concepts can also be applied to ST methods from other platforms.

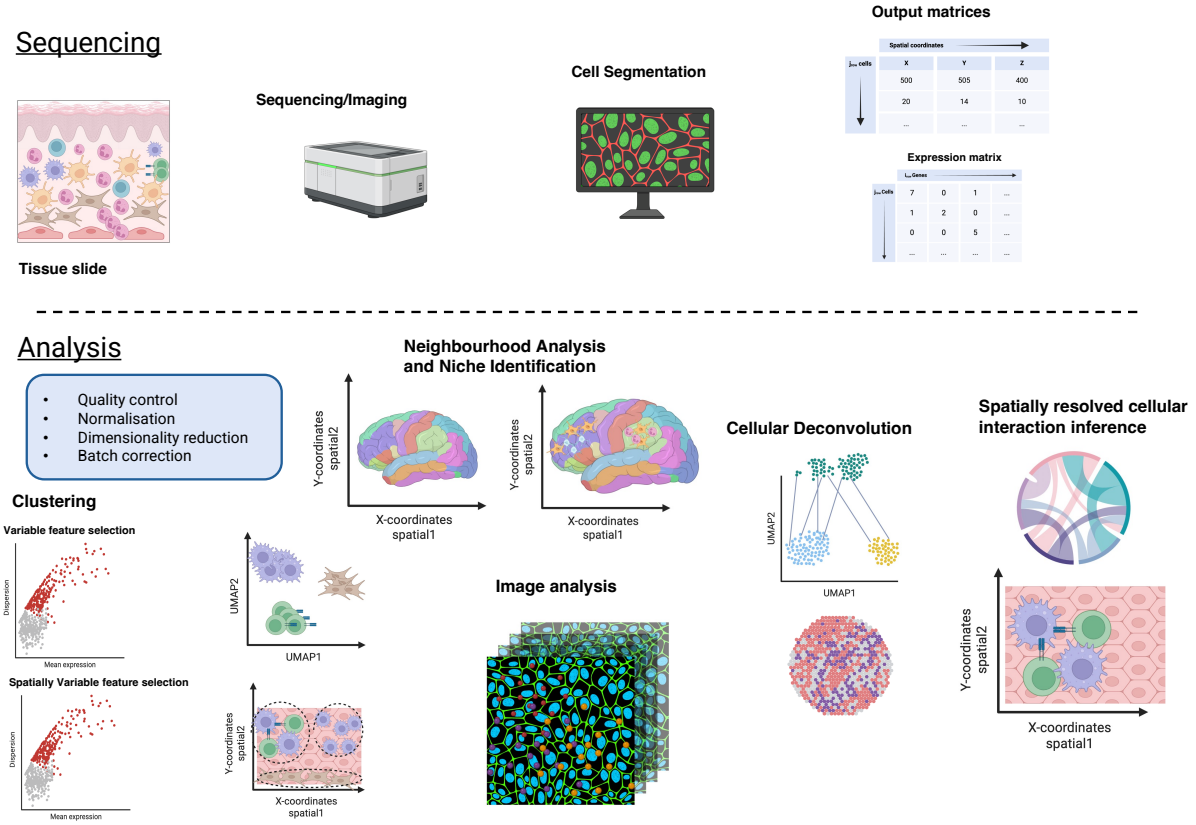


Figure 1.5: Graphical schematic showing the key analysis steps of a spatial analysis pipeline. Preprocessing of spatial data encapsulates steps such as cell segmentation and steps similar to single cell RNA analysis such as normalisation, batch correction and dimensionality reduction. Additional analysis steps are outlined using variable genes to cluster in the expression space and spatially variable genes to cluster in the spatial context, neighbourhood analysis and niche identification, cellular deconvolution for pseudobuk spatial data, spatially aware cellular communications inference, and image analysis. Overview figure created in PowerPoint, using icons from the BioRender library.

1.5.4 Cell Segmentation

The main aim of cell segmentation is to identify single-cell boundaries between cells and is essential in spatial transcriptomics as transcripts that have spatial coordinates must be assigned to individual cells to construct the cell-by-gene matrix. There are two main approaches for cell segmentation algorithms, the first being image-based approaches and the second being transcript-based. Cellpose is a generalist deep-learning method that predicts object probabilities from nuclear DAPI and cell membrane staining to reconstruct cell masks from fluorescence images[173–175], but does not assign transcripts to cells. Another algorithm, Mesmer, is a large-scale, supervised whole-cell segmentation model that integrates nuclear and cytoplasmic staining in tissue sections, and leverages existing annotated histology datasets[176]. In contrast, Baysor is based on transcript information and fits a Bayesian mixture model over transcript distribution across the tissue along with nuclei DAPI staining to infer cell assignments and boundaries making it particularly useful when the tissue organisation is dense[177]. In practice, these methods are complementary and can be used together such as image-based models like Cellpose or Mesmer to gain information on image staining paired with transcript-based approaches such as Baysor that can leverage the spatial patterning of transcripts themselves. More recently combinatory approaches have emerged such as segger which is a graph neural network (GNN)-based method that utilises both image and transcript information to segment subcellular spatial transcriptomic data[178]. The question of cell segmentation is an active area of tool development and impacts all downstream analysis in particular cellular communication as misinformed segmentation will impact cell distances and boundaries. Other strategies are being developed such as multi-modal segmentation provided by 10X Genomics that uses protein markers to delineate the nuclear, interior and boundary regions of a cell and uses a deep learning model paired with nuclear expansion to segment cells with cell segmentation masks being output as part of the 10X Xenium workflow[179].

1.5.5 Quality Control

ST data require additional quality control metrics such as aligning the gene expression matrix to histological images, filtering low-quality spots, and evaluating tissue coverage, often through image-based QC metrics such as spot-tissue overlap. For probe-based technologies such as Xenium, aspects of the data such as low-confidence transcript assignment must be examined that indicate unreliable transcripts resulting from non-specific binding during hybridisation cycles, misreads in the decoding process, and poorly designed probes that have a low signal-to-noise ratio. At the cellular level, QC metrics derived from the segmentation output such as the number of transcripts per cell, number of genes, cell area, and nucleus area are used to filter out low-quality or cells that may be artefacts. For instance, cells with extremely low transcript counts, low gene complexity, or outlier cell areas could indicate empty segments that do not contain a cell, tissue/cell debris, or merged cells that the segmentation could not determine and should be removed from downstream analysis. An additional technical artifact in Xenium data is the 'border effect', which refers to reduced transcript detection accuracy around the edges of the tissue section. Cells located near the tissue boundary are more likely to be partially captured due to the physical limitations of tissue sectioning. Moreover, the efficiency of probe hybridisation and transcript amplification is compromised at the tissue margins as a result of uneven distribution of reagents or optical mistakes during imaging. These factors contribute to a higher rate of dropouts and an inflated number of zero-count genes in edge cells. Additionally, segmentation algorithms often struggle to accurately segment incomplete or irregularly shaped border cells, leading to over- or under-segmentation of the cells. Thus, visual inspection of transcript density maps and spatial distributions of quality metrics such as total transcripts and number of genes across the spatial axis can also help identify and filter out affected regions[162].

1.5.6 Spatially Variable Genes and Neighbourhood Analysis

In order to understand the spatial organisation of cell types within tissues we can identify spatial niches using neighbourhood analysis. By leveraging the spatial information of cells or spots, we can compute spatial graphs in which nodes represent cells and edges define physical proximity, typically based on either a fixed radius or k-nearest neighbour graph[180,181]. The graphs provide information for identifying which cell-type pairs co-occur more or less frequently than expected by chance, so we can detect biologically meaningful spatial associations or functional tissue microenvironments. Many tools have been developed to identify spatial niches in ST data, such as network based methods like CellCharter[182] and BANKSY[183] to deep learning methods that leverage graph neural networks (GNNs) such as GraphST[184] and more recently scNiche[185], which applies multiple GNNs to generate multi-view representations for characterising complex cellular neighbourhoods. We can also observe which genes in the data are driving these cellular neighbourhoods by identifying spatially variable genes (SVGs), which are genes whose expression exhibits spatial correlation across the tissue. Many tools have been developed to identify SVGs such as SpatialDE[186] that uses Gaussian processes, SPARK[187] that implements generalised linear spatial models, and correlation distance-based methods such as Moran's I used by Seurat and Squidpy[180]. Together, these tools provide a comprehensive framework for dissecting how gene expression and spatial proximity shape the architecture of tissue microenvironments.

1.5.7 Cellular Deconvolution

In instances where you have low-resolution ST data the identification of cell types and tissue niches may be challenging. For example, in technologies such as imaging mass cytometry[157] we can use spatial information of potential interacting cells but at the cost of cellular resolution due to limited marker panels. Conversely, technologies such as Visium,

where you have full-transcriptome information but pseudobulk spots that contain multiple different cell types it can be difficult to assign cell type information for niche and cellular interaction discovery. Therefore, cellular deconvolution techniques that implement integration of scRNA-seq data and ST data can attempt to overcome these limitations by using commonly measured genes to estimate a degree of similarity between the two datasets, assign cell types to the spatial data and infer spatial origins of the scRNA-seq[138]. Machine learning techniques are pivotal to the integration of scRNA-seq and ST datasets where computational models can identify common structures in the high dimensional data and project them into the same latent space for analysis. Many computational tools are under development to achieve this using underlying machine learning networks to build models describing gene expression in spatial context. For example, stPlus[188] is a reference based method that uses an autoencoder which enforces loss of information from the input spatial and scRNA-seq datasets whereby the model is trained on the cell embeddings of the scRNA-seq to predict gene expression in the ST dataset via a weighted k-nearest neighbour. Another reference-based method is cell2location[189], a Bayesian model, which aims to integrate ST and scRNA datasets through estimation of cell type signatures from scRNA data to then be used to decompose mRNA counts at a given spatial location into the reference cell types. Other popular methods include, RCTD (Robust Cell Type Decomposition) that models spot-level gene expression as a weighted combination of reference cell-type profiles[190] and Dampened Weighted Least Squares (DWLS) which uses a constrained least-squares problem with dampening weights that reduce the impact of highly expressed genes, allowing for more accurate estimation of both abundant and rare cell types[191]. Additionally, deep learning methods such as Tangram[192] have emerged that integrate spatial context and graph-based embeddings to enhance resolution and spatial accuracy of cell-type mapping. Thus, these deconvolution methods enable high-resolution reconstruction of tissue architecture from low-resolution spatial data, facilitating downstream analysis of cell-type localisation, tissue niches and most importantly cellular interactions.

1.5.8 Spatial cellular inference methods

Various tools have been developed or extended to incorporate spatial information in cellular inference for example, CellChat[146] now includes a spatial mode in which users can incorporate cell or spot coordinates with a distance threshold to construct a spatial adjacency matrix. The matrix is then used to filter or weight ligand–receptors so that only interactions between spatially neighbouring cells are considered. Furthermore, version 2.0 of the CellChat database annotates ligands and receptors as either short-range or long-range signalling, allowing more power to filter out false positives when spatial distances are available. In contrast, CellPhoneDB[144] does not currently implement native spatial functionality, however, it allows for the definition of spatial niches prior to computing interactions and infers interactions only on these neighbouring pairs. Tools such as Giotto[181], stLearn[193], and Squidpy[180] rely on spatial graphs or cell adjacency matrices to filter ligand–receptor interactions that occur between physically adjacent cell types. More complex methods have been developed to model spatial gene regulation and cellular interactions beyond classical ligand–receptor co-expression utilising spatial transcriptomic data only. For example, SpatialDM[194] models ligand and receptor gene expression as spatially variable molecular interactions to test for significant spatial co-occurrence between gene pairs that co-occur in space more than expected by chance. The tool also tests for co-occuring genes across conditions allowing for differential gene pair expression. Additionally there are correlation-based tools such as SVCA, which breaks down variable gene expression into spatial and cell-intrinsic components, to allow the quantification of how the microenvironment has an effect on gene expression[195]. Another tool is SpaCeNet[196] which constructs spatial gene regulatory networks by estimating partial correlations between genes across neighbouring cells, capturing both intra and intercellular interactions. Finally, spaCI[197] applies causal inference to spatial transcriptomics data, modelling how the gene expression of one cell influences nearby cells using various graphical models and Granger causality. Most recently however, NicheCompass[198] has been developed which is a graph-based deep-learning framework that learns interpretable cell embeddings based on intercellular communications from spatial transcriptomics, both spot-based and single

cell based, and multi-omics data. By constructing neighbourhood graphs based on physical proximity and leveraging these interaction embeddings, NicheCompass enables the quantitative identification and characterisation of spatial niches that are driven by cellular communications. The development of cellular inference methods based on spatial transcriptomics is still an active area in the field, and over time many methods will emerge that will not only refine cellular inference of ST data but extend to other technologies such as spatial proteomics and multiomics.

1.6 Introduction to single cell visualisation tools

As demonstrated above, leveraging both modalities of data to investigate cellular interactions not only has unveiled disease mechanisms but also provides an exciting venture into other areas of immunobiology. Currently, with the development of various methodologies and tools to analyse single cell and spatial transcriptomics data, the need for accessible unified platforms to analyse data following best practices is critical. Many studies have focused on benchmarking tools that aim to focus on a common task such as integration, differential expression and cellular interactions. However, there has been limited progress in the development of interactive visualisation tools that require no coding experience for bench biologists. Attempts have been made to document and track the number of tools developed for single cell analysis such as scrna-tools.org[199] and The Awesome Single Cell repository[200] that are manually curated databases derived from preprints, publications and software repositories[201]. As of 2022 there were almost 1,400 tools in the scRNA-tools repository with speculation this would exceed 3,000 tools by the end of 2025 with an increasing amount moving away from the R programming language and being implemented in Python. Zappia et al[201] speculate that could be down to a few different reasons, the first being the scalability of R compared to Python in terms of memory and computational ef-

ficiency. As the size and complexity of single cell and spatial datasets increases, sometimes reaching millions of cells, machine learning approaches are becoming favourable usually built in Python that can handle these types of analyses. Interestingly, another point they discuss is the increased popularity in Python based scRNA analysis tools could be the shift in researchers coming from a purely computational background that chose to develop tools for scRNA and spatial analysis in the language that they prefer. This is echoed by the fact that two thirds of analysis tools are not available from centralised software repositories such as CRAN, Bioconductor or PyPI and are only available on GitHub. This trend of tool development in the field of single cell further perpetuates the barriers bench biologist face when trying to implement these tools, if there is a lack of programming knowledge or how to utilise repository websites such as GitHub. There are in some cases, tools that are built around an existing framework including workflow managers such as Snakemake[202,203] and Nextflow[204] which aid in distribution and installation. However, to properly address this issue, the community could shift their focus on developing tools that bridge the gap between programmer and biologist by implementing interactive web interfaces that allow biologists to explore their own single cell and spatial data. Out of the documented tools in the scrna-tool database 78 of them were produced for visualising single cell data however, only 13 of them contained some element of interactivity, and only one tool InterCellar[205] was tailored for cellular communication visualisation in single cell data. InterCellar is an Shinyapp that focuses on the visualisation of cellular interactions where a fully preprocessed object and precomputed cellular interactions are uploaded, and it provides multiple visualisation functionalities. The plots are interactive and aim to facilitate bench biologists to explore their cellular interaction results however does not extend to spatial technologies incorporating spatial information. Another tool that has since been developed is ezSingleCell[206] that extends their functionality beyond cellular communications and incorporate spatial technologies in their analysis pipeline. ezSingleCell provides multiple modules that focus on different analysis tasks such as differential expression analysis, cell-cell communication and cellular deconvolution of spatial data. It is distributed as a web-application that does not require installation and as a shinyApp that can be installed by a user for offline-analysis. However, maintenance and distribution

of these tools remains a challenge, such as the URL for ezSingleCell being inaccessible and the GitHub repository containing little to no documentation about how the tool can be implemented. In addition to this, the above tools are written and distributed in R, thus bringing into question their scalability for large scale spatial and atlas-level datasets.

1.7 Aims and Objectives

This thesis aims to demonstrate how we can utilise cellular interactions to further elucidate molecular mechanisms of action in infectious diseases such as COVID-19, which then sparked the motivation to develop a visualisation tool to allow the ease of discovery of these interactions, and application of cellular interaction inference using single cell and spatial transcriptomic data to investigate shared pathways of inflammatory disease and host responses during parasitic infection.

1.7.1 Aim 1: Spatially resolved single-cell atlas unveils a distinct cellular signature of fatal lung COVID-19 in a Malawian population

1.7.2 Aim 2: cellXplore: a web tool to interactively explore cellular interactions at the single cell resolution

1.7.3 Aim 3: Dissecting cellular interactions in big data: Contextualising cellular interactions using atlas-level single cell and sequencing based spatial transcriptomics

1.7.4 Publications

The majority of the work presented in this thesis has been published or is under review as follows:

Chapter 1:

Spatially resolved single-cell atlas unveils a distinct cellular signature of fatal lung COVID-19 in a Malawian population - *Nature Medicine* (2024).[207]

Chapter 3:

Human Fibroblast–Myeloid cell tissue atlas across lung, synovium, skin and heart - under review in *Arthritis & Rheumatology* (2025).[208]

Chapter 3:

Spatial transcriptomics reveals recasting of signalling networks in the small intestine following tissue invasion by the helminth parasite Heligmosomoides polygyrus - under review in *Nature Communications* (2025).[209]

Other published works where I have provided expertise of cellular interaction inference include the study below, but was omitted from the thesis:

- *Synovial tissue myeloid dendritic cell subsets exhibit distinct tissue-niche localization and function in health and rheumatoid arthritis* - *Cell Immunity* (2024).[210]

Chapter 2

Spatially resolved single-cell atlas unveils a distinct cellular signature of fatal lung COVID-19 in a Malawian population

2.1 Abstract

Postmortem single-cell studies have transformed understanding of lower respiratory tract diseases (LRTD) including COVID-19 but there is almost no data from African settings where HIV, malaria and other environmental exposures may affect disease pathobiology and treatment targets. This chapter presents the single cell analysis of lethal COVID-19 in a Malawi cohort in a multi-centre collaborative effort published in Nature Medicine[207]. We used histology and high-dimensional imaging to characterise fatal lung disease in Malawian adults with (n=9) and without (n=7) COVID-19, and generated single-cell transcriptomics data from lung, blood and nasal cells. Data integration with other cohorts showed a conserved COVID-19 histopathological signature, driven by contrasting immune and inflammatory mechanisms: in the Malawi cohort, by response to interferon-gamma (IFN-γ) in lung-resident alveolar macrophages, in USA and European cohorts

by type I/III interferon responses, particularly in blood-derived monocytes. In addition to this, our study provides open-source data resources of atlas level single cell data and highlights the importance of studying the cellular mechanisms, in particular multi-modal cellular inference, of disease in under-represented populations, indicating shared and distinct targets for treatment.

2.2 Introduction

There has been significant progress towards the creation of a human cell atlas utilising scRNA-sequencing (scRNA-seq) and high-dimensional cellular imaging data[10,211]. The human cell atlas is transforming our understanding of cells and their states in health and disease and is rapidly becoming a major resource for the development of novel treatments and vaccines[212]. There are currently single cell atlases such as the Tabula Muris[11] and the Tabula Sapiens[213] that aim to profile millions of single cells across multiple organs. Furthermore, there are many reported COVID-19 atlases that aim to profile various tissues at the single cell level in acute, chronic and fatal settings[3,4,6,7,9,214]. Yet, data within these human atlases is heavily biased towards populations in the Northern hemisphere. Populations in sub–Saharan Africa (SSA) are particularly under-represented[215]. For the discovery of therapeutic targets, genetic and environmental factors in different demographic populations may lead to important differences in cell development and cell-compositions in different organs, thus effecting cellular responses to diseases, vaccines and therapies[216,217]. Capturing data from SSA populations is critical to assure that everyone can benefit from the treatment advances derived from the human cell atlas.

Previous single cell studies investigating the mechanisms into COVID-19 pathology have indicated that immunomodulation plays a critical role in COVID-19 outcomes. Single-cell data from lung tissue facilitated identification of specific immunomodulatory targets[2–4,14,212,218,219]. Apart from our high-dimensional imaging study from a Brazilian cohort[220] these data are, thus far, exclusively from populations in Northern hemisphere, similar to most clinical trial data validating their efficacy. For future waves or epidemics of SARS-CoV2 or related viruses, this knowledge gap needs to be addressed. Indeed, given minimal intensive care, the benefit of preventing progression to or deterioration from severe disease by immunomodulation is even more important in SSA due to under-representation in the literature for immunomodulatory targets and socioeconomic factors that impact healthcare. While immunomodulatory therapies can be lifesaving, they can also be harmful[221]. Immunomodulation has focused on two opposing strategies: augmenting the inflammatory response to aid viral clearance or attenuating inflammatory response to reduce pathogenic hyperinflammation. Extensive studies in northern hemisphere cohorts have established that, by the time patients present with life-threatening illness, viral loads are declining, hyperinflammation generally predominates and thus anti-inflammatory interventions are more effective[221,222]. Given evidence that repeated exposure to malaria and other parasitic infections can induce immune tolerance[223–225], we hypothesised that the balance may be different in patients in SSA where these infections are more prevalent providing either a protective or more susceptible immune response. While sometimes this clinical context may be immunoprotective, in those who progress to severe disease, a tolerance-skewed response might blunt immune-mediated viral clearance, leading to a more viral-driven pathology. However, the reverse is also possible. High pathogen exposure can induce an accelerated inflammatory response on re-exposure to pathogens[216]. Either scenario might impact cellular responses driving pathogenesis in the lung and have important implications for informing which therapy may be effective in SSA populations.

To address some of these knowledge gaps we conducted an autopsy study in well-characterised patients at a large public hospital in Malawi, a low-income country in SSA with high rates of malaria, TB and HIV. We undertook detailed histopathological analysis and scRNA-seq on lung, blood and nasal cells and imaging mass cytometry (IMC) to spatially resolve the immune landscape of the lung. We conducted all tissue processing, cell dissociation and scRNA-seq library preparation on site in Malawi, with much of the data prepared on fresh samples. There are so far no studies from any settings that included characterisation across all these modalities. Thus, to fully understand the context of our data in contrast to other populations, we needed to use data from patient cohorts from different regions of the world to enable comparisons (Figure 2.2). Taken together, our data highlight how COVID-19 has a similar histopathological pattern in our SSA cohort to other Northern and Southern Hemisphere cohorts. However, we found a contrasting immune response signature in the SSA cohort, driven by proliferation of lung-resident alveolar macrophages and interferon gamma (IFN-γ).

2.3 Methods

In this section, I will detail all of the methods used in this study both experimental and analytical that relate to the single cell analysis and cellular interaction inference. As a disclaimer, the sections pertaining to patient recruitment, autopsy procedure and preparation of single cell/nuclei libraries were written by Dr. Christopher Moxon and can be read in full in our Nature paper[207], in addition to all of these steps being completed by Dr. Moxon and his team lead by James Nyirenda in Malawi. All analysis involving the imaging mass cytometry proteomic data was conducted by Dr. João Da Silva Filho at the University of Glasgow and can also be found in more detail in the paper[207]. Orthogonal validation staining using RNAscope was completed by Dr. Vanessa Herder at the Center of Viral Research (CVR) at the University of Glasgow. All remaining analysis focusing on the single cell including preprocessing, demultiplexing of samples, core analysis and

cellular inference that are detailed in this method section is work carried out by myself and will be the focus of this chapter with reference to work done by others where necessary. Comprehensive methods of the work completed outside the scope of my own work can be read in our paper[207].

2.3.1 Patient Recruitment

We recruited patients aged 45-75 admitted to Queen Elizabeth Central Hospital (QECH), Blantyre between October 2020 and July 2021 during which there were two epidemiological waves driven by different variants of SARS-CoV2: Beta (December 2020-February 2021) and Delta (May-July 2021)[226]. Patients admitted with respiratory signs were routinely tested for SARS-CoV2 at QECH. We recruited cases into three groups based on clinical criteria: **1)** a Covid-19 group (n=9) with clinical features suggesting acute respiratory distress (ARDS, oxygen requirement and either respiratory signs on clinical examination or chest x-ray changes or both) and who had at least one nasal swab positive for SARS-CoV2 on admission; **2)** A non-Covid-19 LRTD (lower respiratory tract disease) group (n=5) who had clinical signs of ARDS but were negative for SARS-CoV2 on admission and during hospitalisation; **3)** a no LRTD, COVID-19 negative group (n=2) who had no oxygen requirement and no clinical signs of LRTD and for whom the admission and any subsequent nasal swabs were negative for SARS-CoV2 on PCR (Figure 2.2). The study only recruited cases who died between 12 midnight and 12 noon to minimise the post-mortem interval and to avoid doing any autopsies at night.

2.3.2 Minimally invasive autopsy

We used minimally invasive sampling (MITS) to conduct autopsies with large-bore needle biopsies of organ samples rather than full autopsy[227]. Being more culturally acceptable, MITS is widely used to determine cause of death in paediatric studies[227–229], showing good concordance with full autopsy[228]. From our ongoing paediatric MITs studies in Malawi, we adapted protocols for adult patients with Covid-19 to obtain tissue suitable for single cell RNA-sequencing (scRNA-seq) and imaging mass cytometry (IMC) based on the protocol from the Child Health and Mortality Prevention Surveillance (CHAMPS) network but with adaptations. In particular, a larger calibre needle (11 gauge) was used for biopsies to obtain larger tissue samples. Samples were taken from the brain from supraorbital sampling from both left and right sides. From each lung, samples were taken from lower middle and upper zones from a single entry-point, angling the needle to sample different areas. Nasal cells were collected from the nasal inferior turbinate using curettes (ASL Rhino-Pro, Arlington Scientific). Two curettes were collected from each nostril and the cells placed immediately into ice cold Hypothermosol (StemCell). Cells were transported on ice in a cold box immediately to the lab and were spun at 300g for 5 minutes either for immediate processing for scRNA-seq or were stored in Cryostor 10. Nasal fluid was collected using matrix strips (Nasosorption, Hunt Developments) where one strip was used per nostril.

2.3.3 Luminex Multiparameter Cytokine Assay

Cytokine levels were measured in plasma and nasal fluid samples using Luminex with the Inflammation 20-Plex Human ProcartaPlexTM Panel (Thermofisher, EPX200-12185-901) according to the manufacturers protocol and levels measured with a Luminex MagPix device. Data were $\log_2$ transformed and visualised with ComplexHeatmap[230] in R[231] using a Z-score for each cytokine. For the statistical tests of genes associated with the IFN-γ pathway we used a Welch Two Sample t-test. No significant differences for those genes were found between the Covid-19 and LRTD samples.

2.3.4 Dissociation of lung cells from frozen samples and single nuclei preparation

Lung samples were dissociated both from fresh samples and from slow frozen samples that had been stored in liquid nitrogen. Slow frozen cells were defrosted using a defrosting protocol described previously. Fresh or defrosted frozen cells were then dissociated adapting methods developed previously[232]. Briefly, cells were dissociated in a buffer containing 400$\mathrm{mg\,ml^{-1}}$ of Liberase DL (Sigma), 32 U/ml^{-1} of DNAse I (Roche) and 1.5% BSA in PBS (without calcium and magnesium). The tissue was put in buffer (4 times weight:volume) in a GentleMACS C-tube (Miltenyi 130-096-334) minced with scissors and then ran on a GentleMACS dissociator (130-093-235) on programme "C-lung 01_02". Dissociation was achieved by warming tissue on an orbital shaker in a chamber at 37°C for 30 minutes and running "C-lung 01_02" twice more; once at 15 minutes and once at 30 minutes. Enzyme was neutralised by diluting with 10ml of ice cold 20% FBS with 32U/ml of DNase and the sample was filtered through a 100-μm strainer (352360) and samples were subsequently kept on ice with all centrifuge and antibody incubation steps at 4°C. Cells were pelleted by spinning at 300g for 5 minutes and red cells removed by incubation with ACK buffer for 5 minutes. For frozen cells debris were removed using a debris removal

solution (Miltenyi, 130-109-398) according to the manufacturers protocol. Single nuclei were prepared from snap frozen lung samples as described previously[3]. Briefly, frozen lung tissue was kept on dry ice/liquid nitrogen until processing was started. Tissue was added to a gentle MACS C-tube containing 2ml of freshly prepared nuclei extraction buffer which contained RNAse inhibitors; 0.2 U/μL RNaseIN Plus RNAse inhibitor (Promega) and 0.1 U/μL SUPERasin RNAse inhibitor (Thermofisher scientific). Dissociation was achieved by running the C-tube on GentleMACs dissociator on program "m_spleen_01" for 1 minute. The sample was then filtered using a 40 μM strainer. The C-tube and strainer were rinsed using a buffer containing 0.1% enzymatics RNAse inhibitor (Enzymatics). Sample was then pelleted by spinning at 500g for 10 minutes at $40°C$. Pellet was then resuspended in 500μl of 1xST without RNAse Inhibitor. The sample was then filtered using 35μM strainer, a 10 μL volume was loaded on haemocytometer for counting.

2.3.5 Single cell and single nuclei partitioning and library preparation

10X 3' 3v chemistry was used for all samples. For fresh lung samples we loaded 10,000 cells into one channel of a 10X chip (1000120). For fresh nasal and blood samples we labelled the nasal and blood samples with different hashtags and pooled them at a 1:1 ratio and loaded 10,000 – 20,000 cells. For frozen nuclei and single cell samples we pooled samples from 3-6 different cases aiming for equal ratios and loaded 20,000 – 40,000 cells per nuclei. Libraries were prepared according to the manufacturers protocol and sequenced with an Illumina NextSeq2000. To make these data available for analysis by others, reads were submitted to ArrayExpress (E-MTAB-13544).

2.3.6 Single cell processing

2.3.6.1 Processing of the raw reads

5' scRNA-seq data along with the 3' snRNA-seq runs were demultiplexed using Cell Ranger[48] 'mkfastq'. The reads were then aligned with Cell Ranger (v7.0) 'count' to generate transcript count matrices including those that mapped to intronic regions on the genome. Transcript reads were mapped to the human GRCh38 reference genome which was concatenated with the SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome, GenBank MN908947.3) and HIV (human immunodeficiency virius 1, GenBank AF033819.3) genome as additional chromosomes aiming to capture viral reads in our cohort.

2.3.6.2 Ambient RNA removal

To reduce potential noise driven from empty droplets or ambient RNA captured in our samples we used the tool SoupX (v1.6.2)[64]. Data were read into R (v4.2)[231] including the raw unfiltered expression matrices and clustering information required as input. Expression matrices for each sample were then corrected for and used for downstream analyses.

2.3.6.3 Quality control and filtering

Data were analysed using the Seurat package (v4.3)[18] with quality control carried out on individual samples. Thresholding for mitochondrial genes that were expressed in our cells allowed us to exclude dying cells and any doublets that may be present. In addition to this, we chose to keep cells that were expressing more than 150 genes, to maximise discovery of cell types that may express lower levels of genes.

2.3.6.4 Normalisation and variance stabilisation

After filtering away cells, the samples were merged and normalised using the *SCTransform()* function, selecting the top 3,000 variable genes to drive the downstream clustering and regressing out the effects of mitochondrial gene expression ribosomal gene expression. The effect of cell cycle were also regressed out using the *CellCycleScoring()* function to determine expression of cell cycle related genes across cells.

2.3.6.5 Integration

Integration of samples was performed by first running a principle component analysis (PCA) on all merged data objects, the embeddings for which were then fed into the standard Harmony (v0.1.1)[101] integration pipeline.

2.3.6.6 Clustering and dimensionality reduction

Principle components for each integrated object were then visualised using the *ElbowPlot()* function and the appropriate number of principle components were selected to be used to generate the Uniform Manifold and Approximation projection (UMAP). The same PC's were used to determine the k-nearest neighbours for each cell for the shared nearest neighbour (SNN) graph construction followed by clustering at varying resolutions depending on the dataset.

2.3.6.7 Cluster marker identification and cell type annotation

Identification of cluster markers for cell type annotation for the lung and nasal datasets were calculated by running *FindAllMarkers()* on the non-batch normalised expression values using the MAST differential expression algorithm[118]. We specified that genes must be expressed in at least 25% of cells (*min.pct = 0.25*) with a log fold change of 0.25. Cell types were manually annotated leveraging returned cluster marker genes and canonical cell type markers reported from existing literature and curated datasets[3,4,14]. Cell type clusters in the peripheral blood dataset were annotated using the consensus label transfer algorithm SingleR (v2.0.0)[109]. Cells in our cohort were mapped against the Azimuth Reference PBMC atlas[18,233] to obtain cell type label predictions. Cells with mapping scores below 75% confidence were reanalysed and manually annotated using the cluster marker identification steps described above. A majority of poorly mapped cells were reannotated as neutrophils which were an absent cell type in the PBMC reference dataset.

2.3.6.8 Gene ontology and pathway analysis

Differentially expressed (DE) genes across conditions were calculated using the *FindMarkers()* function using MAST. Genes were defined as DE with a significance threshold of <0.05 and a log fold change threshold of 0.25. Gene set enrichment analysis (GSEA) was done using the fgsea package[234] to determine what pathways were enriched in COVID-19 out of the 50 canonical hallmark gene sets as described in Msigdb (v.7.5.1)[128,235,236].

2.3.6.9 Module scoring

Gene module scoring was calculated using the *AddModuleScore()* function of gene sets taken from MsigDB and AmiGO 2[237–239]. Gene sets that were associated with the lambda (GO:0034342), alpha (GO:0035455), interferon beta (GO:0035456), interferon gamma (GO:0034341) and TNF (HALLMARK_TNFA_SIGNALING_VIA_NFKB).

2.3.6.10 Cell-cell communication analysis

Inference of cellular communications was computed using the multinichenetR (v1.0.3) package[152] that allows the prediction of interacting cells based on differentially expressed genes across conditions. We set the minimum number of cells per sample/condition to 5 and searched for the top 250 targets with a log fold change cutoff of >0.5 being expressed in at least 10% of cells.

2.3.7 Hashtag Demultiplexing

Hashtag reads were quantified using CITE-seq-Count (v1.4.4)[240] yielding a matrix containing the hashtag counts per cell and subsequently demultiplexed using the consensus calling algorithm cellHashR (v1.0.1)[79]. The following methods were tested BFFcluster[79], BFFraw[79], GMM-Demux[241], Seurat HTODemux[100] and DropletUtils hashedDrops[242,243] to maximise accuracy of the identification of nasal and peripheral blood cells. Of the methods tested, HTODemux resulted in the highest number of singlets which were then selected for downstream analysis, filtering away doublets and cells that failed to be classified by the algorithm.

2.3.8 SNP splitting of multiplexed runs

Demultiplexing of combined sample runs were carried out using the Souporcell[70] algorithm to identify distinct genotypes and assign cells to different individuals. For each run, we set the number of clusters (k) to the expected number of genotypes in the run (k=2–6), and cell barcodes were assigned to each cluster. Cluster barcodes were then used to subset the input BAM file across human leukocyte antigen (HLA) loci of the multiplexed runs, under the assumption that these would be distinct regions of the genome for each individual. Using Integrative Genomics Viewer (IGV)[244,245], we visualized single nucleotide polymorphism (SNP) distributions at a set allele frequency of 0.2 and compared the subset BAM files to BAM files from individual runs. Iteratively, Souporcell clusters were assigned to samples through the following rationale: **1)** matching SNP distributions to independent sequencing runs, **2)** through mapping to sex chromosomes or **3)** through the process of elimination where an independent sequencing run genotype was not available. For some samples we found that multiple cells that had been allocated to a given individual shared the same genotype, therefore confirming that the Souporcell algorithm had failed to identify distinct genotypes for the number of patients we had multiplexed in our

runs. To further investigate this, we prepared a test BAM file generated from barcodes belonging to individual runs (Cos11-L, Cos12-L, Cos14-L and Cos16-L) and partitioned the ratio of reads coming from each sample (0.5:2:5:10 respectively). This provided a ground truth of expected number of cells for the Souporcell algorithm to classify into the appropriate number of clusters (k=4). Repeating the above strategy, we observed that the algorithm failed to classify cells belonging to the genotype with the lowest ratio. Therefore, we concluded that where we had a low number of cells deriving from a case within a multiplexed run due to technical limitations, there may not have been sufficient information available for the SNP clustering algorithm to correctly identify the genotype. In scenarios where Souporcell failed to identify the expected number of genomes, we assigned cluster barcodes to matching genotypes from independent sample runs regardless of expected k. After successful demultiplexing, we identified which cells derived from which patient and were able to proceed with downstream single-cell analyses as outlined above.

2.3.9 Lung Integration

The Human Lung Cell Atlas (HLCA)[246] was downloaded from the cellxgene[247] data portal, containing over 2.4 million lung cells in health and disease, including those within our cohort. The atlas was filtered down to exclude irrelevant datasets that were not directly comparable to our cohort, retaining cells that were taken from the lung and lung parenchyma. This included studies predominantly originating from the northern hemisphere, observing the lung cellular landscape in COVID-19, pneumonia and absence of lower respiratory tract disease. To mitigate the discordance of unique cell types included within the HLCA we selected cell types that were broadly annotated at *ann_level_3* that harmonised with our analyses (AT1, AT2, EC arterial, EC capillary, EC venous, Fibroblasts, Innate lymphoid cell, NK, Macrophages, Monocytes, T cell lineage). These cell type populations included the immune compartment and stromal populations that are pivotal for viral response in the lung. After setting these thresholds we yielded an atlas consisting of over 1 million cells. To avoid lack of power for downstream analyses

with our cohort, we randomly subsampled each cell type within each disease condition to create a normalised atlas of 100,000 cells to integrate with our lung atlas. Both atlases were normalised using *SCTransform()* regressing out potential confounding features such as cell cycle effects, mitochondrial and ribosomal gene expression. Next, common variable features were found between the two datasets using *SelectIntegrationFeatures()* set at 3,000 genes. The datasets were then combined into a single object and were analysed using the canonical analysis pipeline including generating PCA embeddings to form the bases of integration by harmony[101]. UMAP coordinates were obtained using the first 38 principle components and the data were broadly clustered at a resolution of 0.2. Manual cluster annotation was performed by running *FindAllMarkers()*, leveraging canonical cell type markers. To increase the granularity of the T-cell population, these were subset out and reclustered following the recommended pipeline including re-running SCTransform[83] and reintegration with harmony[101].

2.3.10 Pseudo-bulk

To be able to compare the nasal and blood scRNA-Seq with the Luminex, we pseudo-bulked the all single-cell data for each tissue, using the Seurat function *AverageExpression()*. All cells were assigned to a unified identifier ('pseudo_cluster') to pool cells belonging from different cell type clusters together. After executing the above function, this generated a pseudo-bulk expression matrix with the average gene expression of the Luminex panel genes across all cell types stratified by case. The data were plotted similar as the Luminex data, using ComplexHeatmap[230] and a Z-score of the counts, see (Figure 2.24). For the statistical tests of genes associated with the IFN-γ pathway we used a Welch Two Sample t-test. No significant differences for those genes were found between the Covid-19 and LRTD samples.

2.3.11 Exploring viral reads in samples

As mentioned above, we mapped our 10X scRNA-Seq reads against a combined reference of the human genome with the genomes of human genome and a HIV and Covid-19 reference. We found no reads mapping against the HIV genome. (Figure 2.10) summarizes the cells with evidence of at least two Covid-19 UMIs within a cell. Interestingly, one sample (cov12) had over 300,000 reads mapped against the Covid-19 genome, for the nasal and blood sample, resulting in many cells being Covid positive. After closer inspection, we realised that this must come from dying cells with very high Covid load, that burst in the process and contaminated other cells. This explains also why SoupX filtered those counts out as environmental contamination.

2.3.12 Gene panels defining the IFN-γ response

To investigate the various interferon responses we used genes that are associated with each gene ontology term. In particular we examined the interferon gamma response which were defined with the following 125 genes: *CD74, TLR2, CCL16, TLR3, CCL25, SHFL, CAMK2A, CALCOCO2, HPX, SYNCRIP, CDC42, ADAMTS13, IFITM2, IFITM3, ACTR2, ACTR3, STXBP4, SIRPA, SLC26A6, MEFV, RAF1, GBP7, CCL26, IL23R, WAS, IL12RB1, GBP6, CASP1, IL12B, KYNU, CCL14, CALM1, GBP2, GBP1, MRC1, TYK2, CD58, ASS1, DAPK3, CD47, GCH1, RAB7B, SLC11A1, SNCA, NUB1, RAB20, STAT1, CCL3, CD40, IRF1, CXCL16, CLDN1, FLNB, XCL2, EDN1, CDC42EP4, CCL15, CCL3, L1GSN, CCL22, GAPDH, CX3CL1, STXBP1, STXBP3, LGALS9, CCL24, RAB43, CCL19, KIF5B, WNT5A, MYO1C, TP53, GBP3, IFITM1, CCL11, ACTG1, TNFSLC30A8, FASLG, CCL20, VAMP3, CCL17, CCL7, IFNGR2, SLC22A5, CCL8,*

BST2, CCL13, PDE12, DAPK1, XCL1, CITED1, ZYX, CIITA, IFNG, AQP4, CCL21, AIF1, CDC42EP2, CCL5, CCL2, STX4, IRF8, JAK2, HLA-DPA1, STX8, RPL13A, IFNGR1, TRIM21, CYP27B1, GBP5, GBP4, VIM, HCK, VPS26B, CCL4, UBD, ACOD1, CCL18, CCL2, 3NOS2, TLR4, SP100, JAK1, RPS6KB1.

2.4 Results

Here the results of the study will be presented with a focus on the single cell RNA sequencing analysis of lung, nasal and blood samples from our Malawi cohort. Outlined in Figure 2.1 is the full scope of the study where we additionally carried out histopathology of lung samples comparing lesions from patients with fatal Covid-19 and LRTD. In addition to this we also completed an imaging mass cytometry analysis of lung samples to profile the cellular composition of the tissue within its native context using a 40-antibody panel. The analysis of the imaging mass cytometry will not be covered in this chapter but will be referenced particularly in the cell-cell interaction part of the results. The additional analyses carried out in this project can be found in our paper and read in full[207].

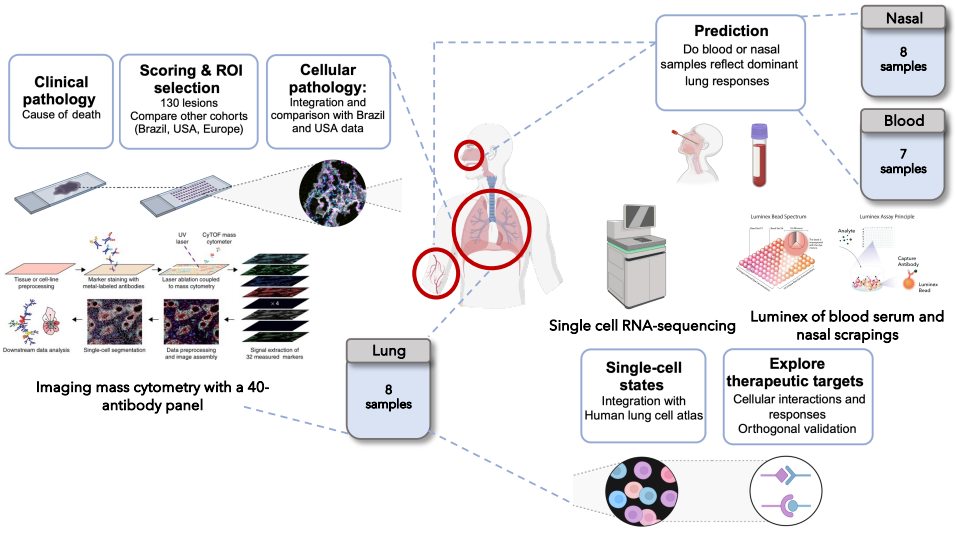


Figure 2.1: Graphical overview showing the complete scope of the study.

2.4.1 Cohort overview

Below is the clinical characteristics of our Malawi cohort Figure 2.2. We note that compared to existing Covid-19 studies that our cohort is deemed young with a median age of 56. Furthermore, a proportion of our patients have underlying co-morbidities, in particular HIV, however we found no reads mapping to the viral genome and no evidence of impact on T-cell levels due to the small sample size. Detailed in the table is the number of patients where various modalities of data were collected either for single cell RNA sequencing, imaging mass cytometry or Luminex analysis.

	Case	Diagnosis	HIV	Sex	Age (yr)	PMI (hr)	Obese/ Under	Pre-morbidity	S.S to death	Lung sc/sn	Nasal sc	Blood sc	Lung IMC	Nasal Lx	Blood Lx
Covid-19	3	C19	1	M	55-60	9	↑	DM2, HT	7		●	●	●	●	●
	5	C19	1	M	55-61	5	↑↑	DM2, HT	6				●		●
	6	C19	1	M	50-55	4	↑↑	DM2, HT	7	●		●	●	●	●
	7	C19	1	F	50-55	5	↑↑↑	Cancer	4		●		●	●	●
	8	C19	1	F	45-50	6	↑↑↑	HT, A	8				●	●	●
	9	C19	0	F	50-55	5.5	↑↑	None	8				●	●	●
	12	Sepsis+C19	0	F	60-65	9	↓	None	29 (2)*	●	●	●	●	●	●
	13	C19	0	F	70-75	5.5	↑↑	DM2, HT	20	●	●		●	●	●
	15	C19	0	M	55-60	10.5	→	None	5	●	●	●	●	●	●
No LRTD	2	TB	1	M	45-50	9	↓↓	None	13	●		●			●
	16	B. Pneum.	0	F	60-65	2.5	→	HT	9				●	●	
LRTD	1	TB	1	F	50-55	3	→	None	17					●	●
	4	L. Cancer	1	F	60-65	3	↓↓↓	None	10	●	●		●	●	●
	10	B. Pneum.	0	M	60-65	9.5	↓	HT	5	●	●	●	●	●	●
	11	Sepsis	1	F	50-55	10.5	↓↓	None	4				●	●	
	14	Stroke	0	F	50-55	9	→	None	2	●	●	●	●	●	●

Figure 2.2: Detailed summary of clinical characteristics of our Malawi cohort. PMI stands for post-mortem interval measured in hours. Arrows indicate patient weight and are denoted as follows: ↑, overweight; ↑↑, obese; ↑↑↑, morbidly obese; ↓, underweight; ↓↓, severely underweight. Pre-morbidity conditions are abbreviated as follows: DM2, type 2 diabetes mellitus; HT, hypertension. S.S. to death, indicates symptom start to death, showing the number of days between the first COVID-19 symptoms and death. Lung IMC, imaging mass cytometry; Lung sc, lung cell single-cell RNA-seq; Nasal sc, nasal cell single-cell RNA-seq; Blood sc, blood cell single-cell RNA-seq; Nasal Lx, nasal Luminex, and Blood Lx, blood Luminex. A dot for each of these parameters indicates that data are available for that case.

To investigate the differences between our Malawi cohort and cohorts recruited in the northern hemisphere we integrated our single cell data with existing studies. The metadata comparing our cohort to northern hemisphere cohorts is detailed in Figure 2.3.

Cohort	Group	N=	Age	SS to death	% HIV	Lung sc	Nasal sc	Blood sc
Malawi	Covid19	9	56 (48-72)	7 (1-20)	56	5	5	4
	LRTD	5	60 (49-60)	10 (5-17)	60	3	2	2
	No LRTD	2	51.5 (51-52)	3 (2-4)	50	1	1	1
Human Lung Cell Atlas (Sikkema et al.)	Covid19	60	50 (21-77)	NR	NR	60	0	0
	LRTD	13	NR	NR	NR	13	0	0
	No LRTD	178	49 (20-76)	NR	NR	178	0	0
USA (TM Delorey et al.)	Covid19	16	NR (30 - >89)	17.5 (1-41)	NR	16	0	0
USA (JC Melms et al.)	Covid19	19	72.8 (58-84)	(<4-63)	NR	19	0	0

Figure 2.3: Summary of characteristics of our Malawi cohort versus published cohorts that we have used as comparators. Abbreviations: SS to death = symptom start to death in days. Lung sc = lung cell single-cell RNA-seq, denotes the number of cases with scRNA-seq data from lung tissue. nasal sc = Nasal cell single-cell RNA-seq, denotes the number of cases with scRNA-seq data from nasal tissue. blood sc = blood cell single-cell RNA-seq, number of cases with this data.

2.4.2 scRNA mapping results

The below tables outline the mapping results of the single cell/nuclei sequencing runs where we aimed to sequence at a depth of at least 30,000 mean reads per cell. This was to ensure that we achieved sufficient depth for the demultiplexing algorithms to identify individual genotypes of multiplexed patient runs.

Runs	Number of cells	Mean reads per cell	Median genes per cell	Sequencing Type
001R-003-004L-014R	12788	38659	1254	SN
003-007-Nasal	2180	69519	669	SC
006-001-012-003-007-008-PBMCs	12351	40291	438	SC
008R-006L-015L	39468	39352	2059	SN
C7plus11plus4	47870	38324	1916	SN
Co6plus1-ns	6028	34531	901	SN
Cos11-L	6044	35652	1634	SC
Cos11-N	4867	49362	728	SC
Cos12-L	6933	43463	779	SC
Cos12NplusP	8208	43043	1888	SC
cos13-L	779	65210	650	SC
Cos13NplusP	4265	36721	1610	SC
Cos14-L	4294	30924	612	SC
Cos14PbplusN	6294	41647	1043	SC
Cos15PbplusN	9210	34766	1050	SC
Cos16-L	11101	40199	1676	SC
Cos16PbplusN	12493	26802	1155	SC
Co6-Lu	498	49145	1600	SC

Table 2.1: Mapping statistics of the CellRanger output for each individual and multiplexed runs.

For cases 12-16, we had hashtagged nasal and blood samples that were collected from the same individual. The table below outlines the results from the quantification of the hashtag reads and subsequent demultiplexing highlighting how many cells from each tissue type were recovered.

Sample	% mapped	% unmapped	Total cells	Nasal	Blood	Negative	Doublet	Total HTO Reads	HTO Reads per cell
Cos16PbplusN	98	2	10311	1796	2245	4309	1961	9526733	924.12
Cos15PbplusN	98	2	8127	492	1708	5747	180	12120440	1491.38
Cos14PbplusN	80	20	6164	742	1676	3640	106	8162964	1324.51
Cos13NplusP	96	4	4276	134	237	1630	2275	8032828	1878.58
Cos12NplusP	97	3	7421	648	1543	4979	254	4488264	604.81
Total			**36299**	**3812**	**7409**	**20305**	**4776**		

Table 2.2: Hashtag demultiplexing summary statistics from CITE-seq-COUNT and Seurat HTODemux.

In addition to the multiplexed hashtag runs, we also had sequencing runs that were multiplexed across patients. The table shows the results from clustering of single nucleotide polymorphisms (SNPs) and shows how many cells were assigned to each genotype. We note that not all expected genotypes were recovered and we proceeded with assigned cells only, filtering away unassigned cells and doublets.

Sample	Singlet	Doublet	Unassigned	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5	Cluster_6	Percentage Doublets	Percentage Unassigned	Expected Genotypes	Actual Genotypes	Number of cells	Reads per Cell
Cos6plus1-ns	2556	22	2664	1176	1380	-	-	-	-	0.42%	50.82%	2	2	5242	14390
008R-006L-015L	32742	2511	3658	10125	9194	13423	-	-	-	6.45%	9.40%	3	2	38911	19780
C7plus11plus4	24097	3	2778	8074	7962	8081	-	-	-	0.01%	10.34%	3	1	47742	8932
003-007-Nasal	1604	44	589	355	1249	-	-	-	-	1.96%	26.33%	2	2	2237	67748
001R-003-004L-014R	5037	46	7373	1258	1473	1100	1206	-	-	0.37%	59.19%	4	2	12456	28940
006-001-012-003-007-008-PBMCs	8940	805	2621	1058	867	1455	1268	2803	1489	6.50%	21.20%	6	4	12366	40242

Table 2.3: Summary statistics for the output of the Souporcell algorithm: Summary of multiplexed sample splitting.

After obtaining the results of the demultiplexing algorithm we assigned each cluster to a patient genotype. The clusters here refer to the clusters in Table 2.3 which were iteratively identified through SNP clustering patterns across variable HLA regions (See Methods 2.3.8).

Cos6plus1-ns	Number of cells	Case ID
Cluster 1	1176	6
Cluster 2	1380	1
008R-006L-015L		
Cluster 1	10125	6
Cluster 2	9194	15
Cluster 3	13423	15
C7plus11plus4		
Cluster 1	8054	11
Cluster 2	7962	11
Cluster 3	8081	11
003-007-Nasal		
Cluster 1	355	7
Cluster 2	1249	3
001R-003-004L-014R		
Cluster 1	1258	1
Cluster 2	1473	14
Cluster 3	1100	14
Cluster 4	1206	14
006-001-012-003-007-008-PBMCs		
Cluster 1	1058	12
Cluster 2	867	6
Cluster 3	1455	12
Cluster 4	1268	1
Cluster 5	2803	6
Cluster 6	1489	3

Table 2.4: Summary statistics for the output of the Souporcell algorithm: Summary of genotype assignments.

The assignment process was as follows, first samtools was used to filter the CellRanger output BAM file to the HLA regions on chromosome 6 and then subset the relevant cells using a barcode list. The BAM file was then indexed to be read into IGV.

```
samtools view -h /export/III-data/otto/oh21b/COSMIC_fastqs/
   COSMIC_Mapping_Results/COSMIC_Cos15PbplusN/outs/
   possorted_genome_bam.bam chr6:29941260-33089696 | perl ~to16r/Bin/
   cellranger.filterBAMvalidbarcodes.pl /export/III-data/otto/oh21b/
   COSMIC_fastqs/COSMIC_Mapping_Results/COSMIC_Cos15PbplusN/Soup/
   barcodes.tsv | samtools view -Sb - > /export/III-data/otto/oh21b/
   COSMIC_fastqs/COSMIC_Mapping_Results/COSMIC_Cos15PbplusN/Soup/
   bam_cos15.MHC.bam

samtools index /export/III-data/otto/oh21b/COSMIC_fastqs/
   COSMIC_Mapping_Results/COSMIC_Cos15PbplusN/Soup/bam_cos15.MHC.bam
```

Listing 2.1: Code snippet used to filter BAM to HLA regions and valid cell barcodes

After the HLA region BAM files are generated, we navigated to the HLA regions and viewed the SNP genotype distribution at an allele frequency set to 0.2. As shown in Figure 2.4 two patient SNP samples are visualised that demonstrate a clear distinct genotype between the different samples. We can also demonstrate with the Souporcell algorithm that similar cell numbers are recovered compared with the original cell counts from CellRanger, 10737 and 6045 respectively, losing less than 50 cells that failed to be assigned.

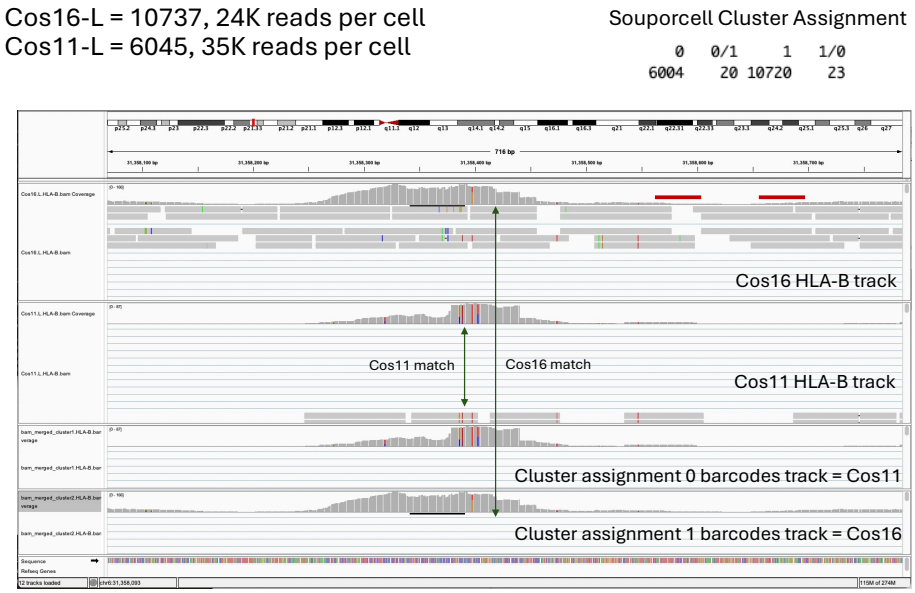


Figure 2.4: Screenshot of the IGV viewer visualising a merged BAM file from Cos16-L and Cos11-L to sanity check the performance of Souporcell to recover both genotypes. Upper tracks show SNP distribution at 0.2 allele frequency of the HLA-B region of Cos16 and Cos11 respectively. Lower tracks show the SNP distribution of the cells assigned to each Souporcell cluster with k=2.

However, we found that for some of the multiplexed samples we were observing duplicate genotypes shared across multiple samples despite setting the k parameter to the number of expected genotypes as shown in Figure 2.5.

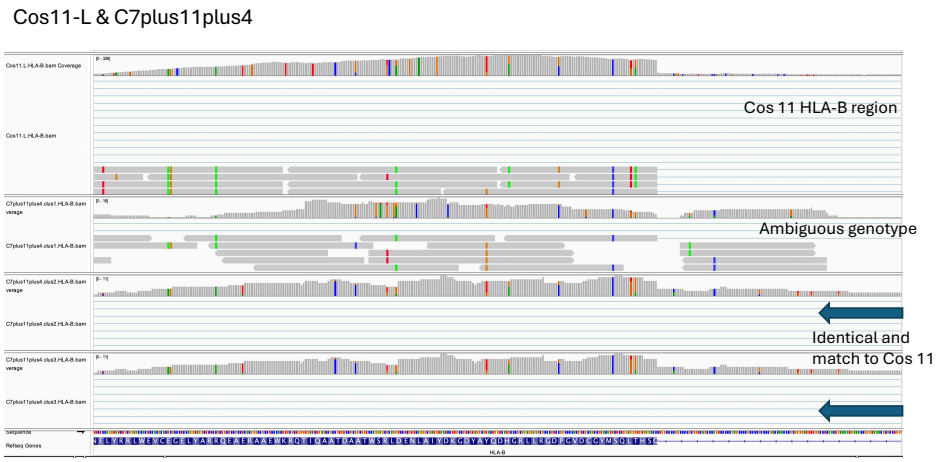


Figure 2.5: Screenshot of the IGV viewer visualising a BAM file from a multiplexed run C7plus11plus4 that should contain 3 distinct genotypes and independent sample Cos11-L to attempt to identify the genotypes. Upper tracks show SNP distribution at 0.2 allele frequency of the HLA-B region of Cos11. Lower tracks show the SNP distribution of the cells assigned to each Souporcell cluster with k=3. Arrows indicate positive identification of the Cos11 genotype, with an ambiguous mixture genotype also annotated.

We suspected the sequencing depth was an issue however despite deeper sequencing the same effect was still observed. Since we are handling post-mortem tissue that is of lower starting quality than other tissues we thought the algorithm may not be able to handle a high level of low-quality cells or that the cells coming from particular patients are lost by the time sequencing occurs. So investigate this further we created a dummy BAM file from 4 patients of which were run separately where we knew the genotypes a priori, Cos16-L, Cos14-L, Cos12-L and Cos11-L, and randomly subsampled the BAM files in a 10:5:2:0.5 ratio (Cos16-L = 57% total cells, Cos14-L = 29% total cells, Cos12-L = 11% total cells, Cos11-L = 3% total cells). We also wanted to observe the effect of varying the k parameter in the Souporcell algorithm and see if it affected genotype recovery. We first set the k parameter to k=*3* to observe which genotypes were recovered shown in Figure 2.6 and recovered three out of the four expected genotypes with Cos11-L failing to be identified. We also note that although the percentage of cells are representative of

the true sample cell ratios for Cos16, the algorithm assigned over double the expected percentage of cells to Cos12 when compared to Cos14. This could be down to the varying sequencing depth between the sample with Cos12 being sequenced deeper than Cos14 therefore introducing a bias in the algorithm assignment (Table 2.1).

k = 3

	0	1	2
Cell Barcodes	11005	3594	5263
% of cells	55	18	26
Genotype match	Cos16	Cos14	Cos12

True Sample read ratios
Cos16-L = 57%
Cos14-L = 29%
Cos12-L = 11%
Cos11-L = 3%

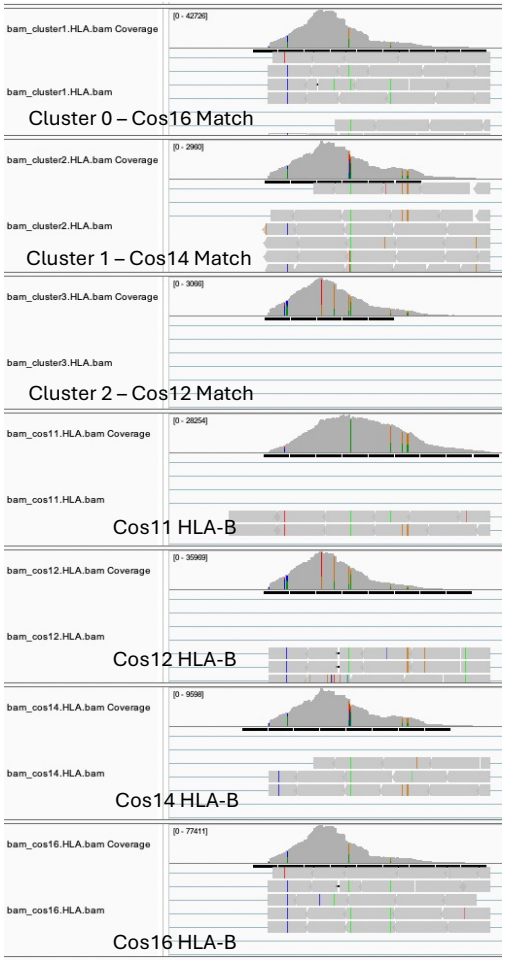


Figure 2.6: Screenshot of the IGV viewer visualising a merged BAM file from Cos16-L, Cos14-L, Cos12-L and Cos11-L to test varying patient cell ratios against the performance of Souporcell to recover all genotypes. Table shows the cluster assignments, proportion of total cells the matched genotype. Lower tracks show SNP distribution at 0.2 allele frequency of the HLA-B region of Cos16-L, Cos14-L, Cos12-L and Cos11-L respectively. Upper tracks show the SNP distribution of the cells assigned to each Souporcell cluster with k=3. Tracks have been annotated with the Souporcell cluster and their respective genotype match.

As setting the k parameter to less than the expected genotypes failed to identify all the patients, we moved to set k=4 to see if we observe the same behaviour (Figure 2.7).

k = 4

	0	1	2	3
Cell Barcodes	3362	5838	4287	5596
% of cells	18	31	22	29
Genotype match	Cos14	Cos12	Cos16	Cos16

True Sample read ratios
Cos16-L = 57%
Cos14-L = 29%
Cos12-L = 11%
Cos11-L = 3%

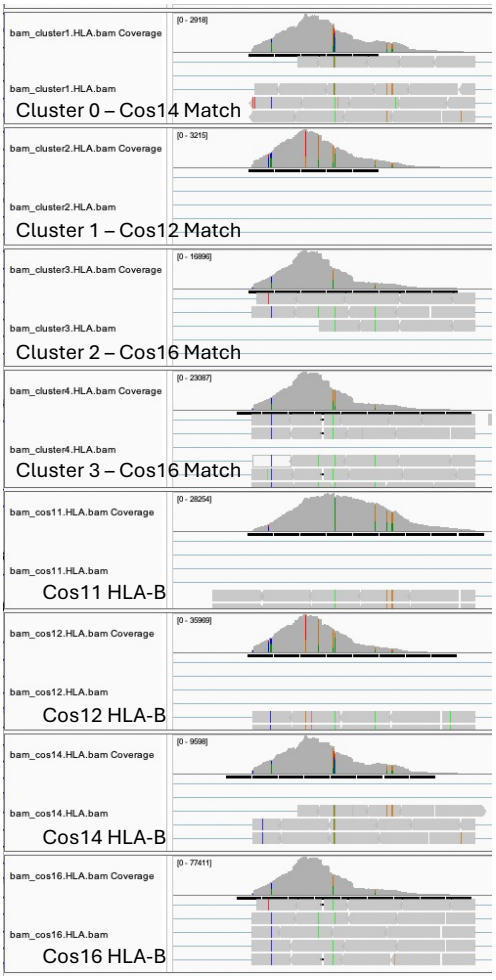


Figure 2.7: Screenshot of the IGV viewer visualising a merged BAM file from Cos16-L, Cos14-L, Cos12-L and Cos11-L to test varying patient cell ratios against the performance of Souporcell to recover all genotypes. Table shows the cluster assignments, proportion of total cells the matched genotype. Lower tracks show SNP distribution at 0.2 allele frequency of the HLA-B region of Cos16-L, Cos14-L, Cos12-L and Cos11-L respectively. Upper tracks show the SNP distribution of the cells assigned to each Souporcell cluster with k=4. Tracks have been annotated with the Souporcell cluster and their respective genotype match.

Despite the k parameter being set to the true number of expected genotypes the same result was observed where Cos16 was represented and Cos12 was overrepresented in the genotype assignment. Similarly to the previous iteration, the Cos11 genotype was failed to be identified with only three genotypes being recovered so we decided to overcluster the data by setting k=5 to see if we could force the algorithm to resolve rarer populations with the Cos11 genotype (Figure 2.8). Unfortunately, overestimating the number of genotypes still failed to accurately represent the number of true samples in the data. This then led to the hypothesis that if the cells deriving from a particular patient are of low quality or have died resulting in low cell numbers then the Souporcell algorithm is unable to recover the genotype by clustering SNPs alone and further patient genotype information is required. Thus, the genotype assignment of cells to patient samples was an iterative processes leveraging information from independent samples and visually comparing the SNP distribution over the HLA regions. Where genotypes were identical, as shown in Figure 2.5, these Souporcell clusters were assigned to the nearest matching genotype and patients whose genotype was unable to be recovered were excluded from the study moving forward.

k = 5

	0	1	2	3	4
Cell Barcodes	2743	5930	3991	3283	3288
% of cells	14	31	21	17	17
Genotype match	Cos16	Cos12	Cos16	Cos14	Cos16

True Sample read ratios
Cos16-L = 57%
Cos14-L = 29%
Cos12-L = 11%
Cos11-L = 3%

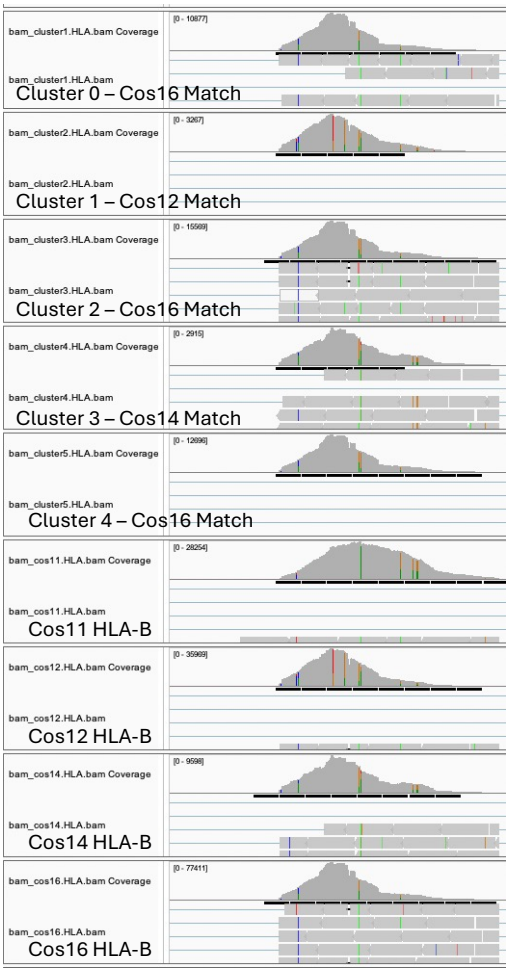


Figure 2.8: Screenshot of the IGV viewer visualising a merged BAM file from Cos16-L, Cos14-L, Cos12-L and Cos11-L to test varying patient cell ratios against the performance of Souporcell to recover all genotypes. Table shows the cluster assignments, proportion of total cells the matched genotype. Lower tracks show SNP distribution at 0.2 allele frequency of the HLA-B region of Cos16-L, Cos14-L, Cos12-L and Cos11-L respectively. Upper tracks show the SNP distribution of the cells assigned to each Souporcell cluster with k=5. Tracks have been annotated with the Souporcell cluster and their respective genotype match.

2.4.3 scRNA quality control results

Once all the samples were assigned, the following quality control metrics were performed on each of the samples for each tissue atlas outlined in the tables below.

Sample	nFeature_RNA cut-off (lower)	nFeature_RNA cut-off (upper)	Percent MT cut-off
Cos-1.1	150	5000	15
Cos-1.2	150	5000	10
Cos-6.1	150	5000	15
Cos-6.2	150	5000	10
Cos-6.3	150	4000	10
Cos-11.1	150	3000	5
Cos-11.2	150	8000	15
Cos-12.1	150	6000	15
Cos-13.1	150	4000	10
Cos-14.1	150	5000	10
Cos-14.2	150	6000	15
Cos-15.1	150	5000	5
Cos-15.2	150	8000	10
Cos-16.1	150	8000	10

Table 2.5: QC thresholds for each lung sample in the Malawi lung atlas. Samples with the same number with decimal points (e.g. Cos1.1) are samples that have been retrieved from SNP clustering genotype assignment with Souporcell and are cells deriving from the same patient but a different run.

Sample	nFeature_RNA cut-off (lower)	nFeature_RNA cut-off (upper)	Percent MT cut-off
Cos-3	150	7000	15
Cos-7	150	7000	50
Cos-12	150	8000	50
Cos-13	150	3000	15
Cos-14	150	9000	40
Cos-15	150	9000	50
Cos-16	150	7000	25

Table 2.6: QC thresholds for each nasal sample in the Malawi nasal atlas.

Sample	nFeature_RNA cut-off (lower)	nFeature_RNA cut-off (upper)	Percent MT cut-off
Cos-1	150	7000	25
Cos-3	150	3500	25
Cos-6	150	4000	25
Cos-12.1	150	7000	25
Cos-12.2	150	6000	25
Cos-13	150	2500	15
Cos-14	150	6000	10
Cos-15	150	8000	25
Cos-16	150	5000	20

Table 2.7: QC thresholds for each blood sample in the Malawi blood atlas. Samples with the same number with decimal points (e.g. Cos-12.1) are samples that have been retrieved from SNP clustering genotype assignment with Souporcell and are cells deriving from the same patient but a different run.

Once quality control filtering was completed, the resulting tissue atlases used the following parameters shown in Table 2.8 to obtain the appropriate UMAPs shown in this chapter.

Atlas	N dims	Resolution
Broad Lung Atlas	32	0.2
Immune Lung Atlas	28	0.5
Nasal Atlas	30	0.3
Blood Atlas	30	0.5
Integrated Lung Atlas	38	0.5

Table 2.8: Number of principal components (N dims) and clustering resolution used for each tissue atlas.

2.4.4 Pulmonary cell scRNA-seq reveals low levels of viral RNA and an IFN-γ dominated response in the Malawi cohort

To explore cellular responses in the lung at greater depth in our Malawi cohort, including in alveolar macrophages, we utilized scRNA-seq and single nuclei-sequencing (snRNA-seq) from 4 Covid-19 cases, 3 LRTD cases, and 1 non-LRTD case. Integrating over 66,000 cells resulted in 16 cell clusters composed of a mixture of immune and stromal cells (Figure 2.9).

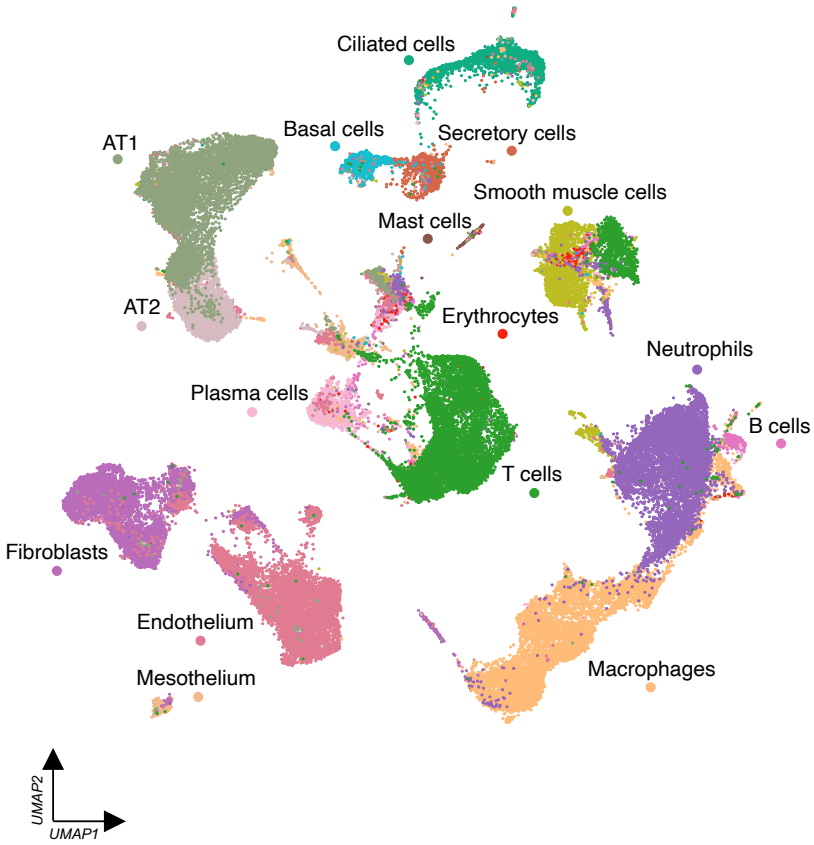
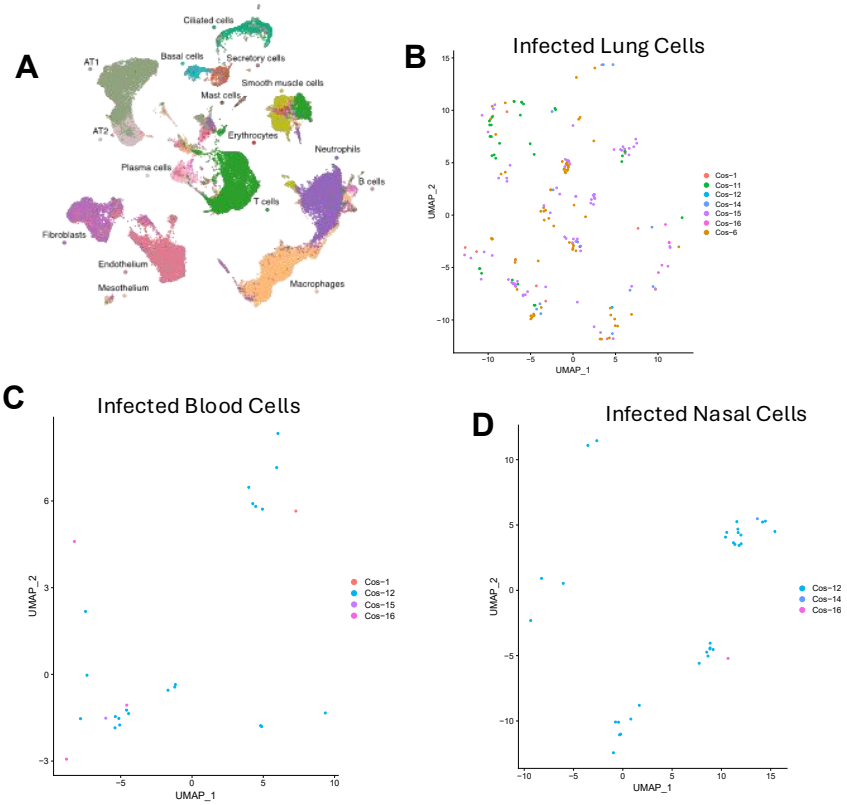


Figure 2.9: UMAP visualisation of 66,882 lung cells across our cohort, coloured by broad cell types.

SARS-CoV2 transcripts have been detected in scRNA-seq data in other postmortem cohorts. We detected few SARS-CoV2 reads suggesting that at the time of death, there was minimal replicating virus (Figure 2.10). This is contrary to our initial prediction of tolerance and viral escape predominating in SSA populations but is consistent with our other data supporting inflammatory rather than direct viral-driven pathogenetic mechanisms.



E

Tissue	Lung	Blood	Nasal	Total
Cos1	10	1	X	11
Cos6	69	X	X	69
Cos11	30	X	X	30
Cos12	1	21	32	54
Cos14	12	X	1	13
Cos15	87	1	X	88
Cos16	5	3	1	9
Total	214	26	34	274

Figure 2.10: Minimal SARS-CoV2 reads in single-cell data of lung, nasal and blood cells. (A) Lung reference as in Fig 3a. (B-D) UMAPs indicate the cells in which we found reads that mapped to the SARS-CoV2 genome, coloured by case. E) Table showing absolute cell numbers per case that contain expression of UMIs that passed quality control steps that map to the SARS-CoV2 genome in the lung, peripheral blood and nasal compartment.

To identify cell types cluster marker analysis was performed which revealed canonical cell type gene expression in line with existing literature for cell types in the lung immune and stromal cell compartments (Figure 2.11). A detailed breakdown of cell numbers across cell types can be found in the Supplemental Tables section (Tables 2.9, 2.10, 2.11).

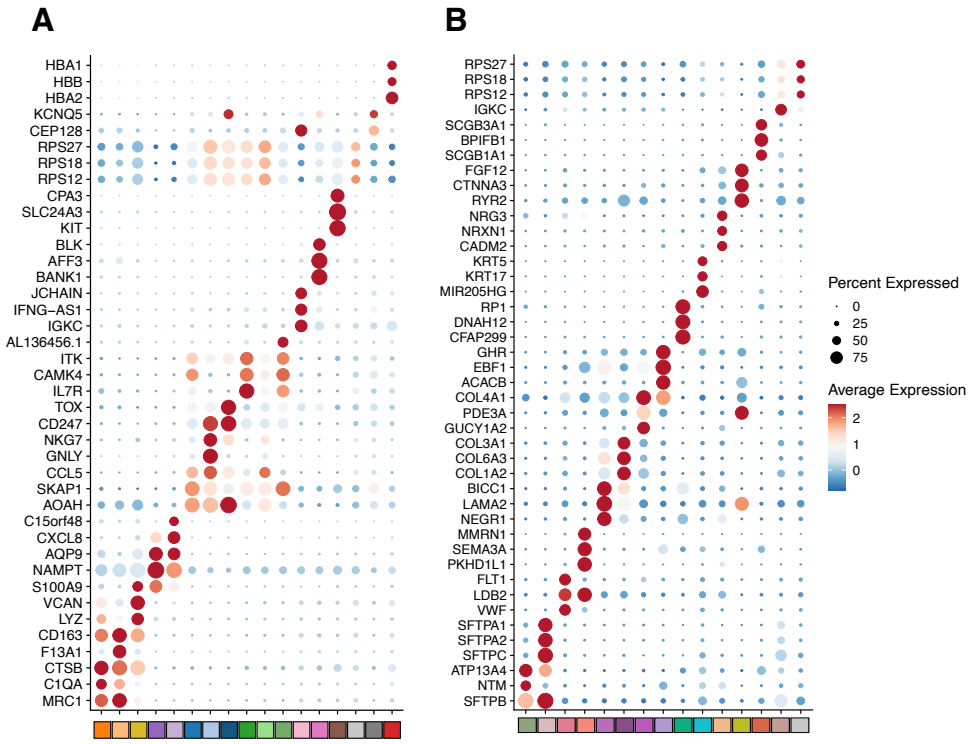


Figure 2.11: Top cluster markers characterising immune and stromal cell populations in the lung in the Malawi cohort. A) Dotplot showing the average expression of top 3 cluster markers for each cell type in the lung immune compartment (Figure 2.12). B) Dotplot showing the average expression of top 3 cluster markers for each cell type in the lung stromal compartment (Figure 2.13).

We then undertook finer annotation of immune (Figure 2.12) and stromal/vascular cell pools (Figure 2.13). We identified alveolar and interstitial macrophages and monocyte-derived macrophages, consistent with monocyte/macrophage populations identified by IMC. Both mature and immature neutrophils were present.

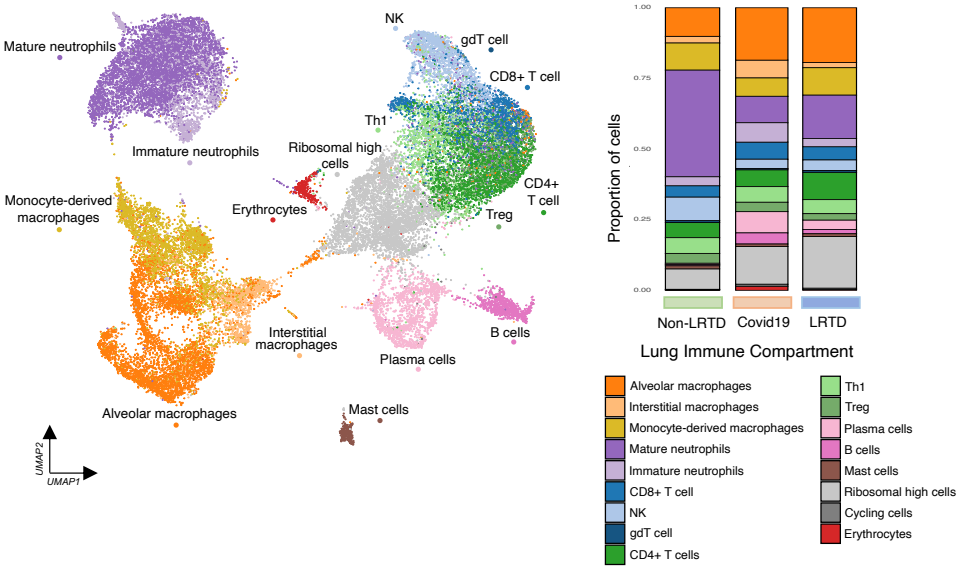


Figure 2.12: UMAP visualisation of 33,504 lung cells reclustered at a higher resolution to characterise the immune landscape, coloured by cell type.

Stromal cells included adventitial and alveolar fibroblasts as well as type I and II pneumocytes (AT1, AT2) and basal, secretory, and ciliated epithelial cells (Figure 2.13). Cell proportions should be interpreted with caution given few cases per group, but they showed cell diversity expansion in the Covid-19 and LRTD groups not observed or absent in the non-LRTD group.

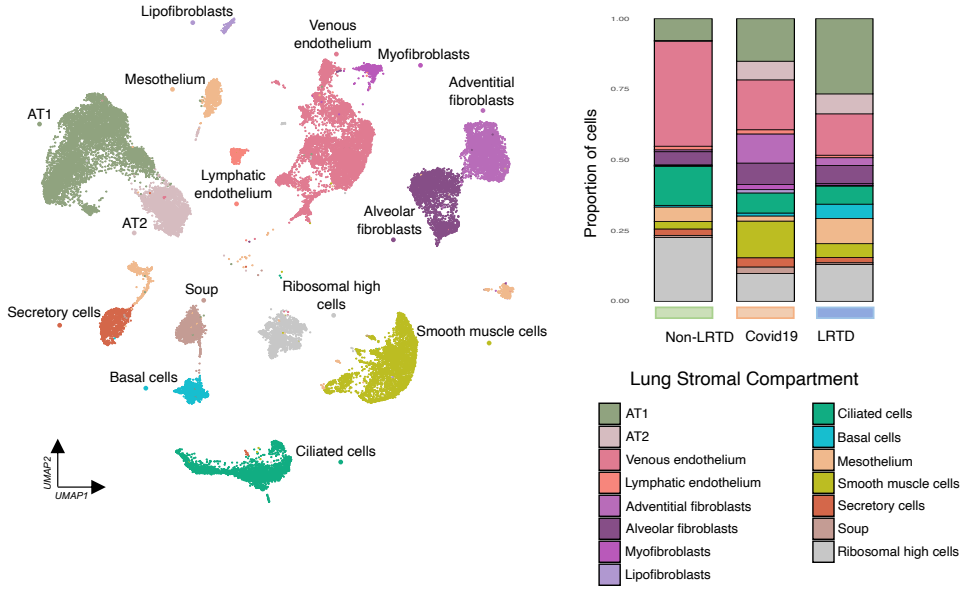


Figure 2.13: UMAP visualisation of 33,378 lung cells reclustered at a higher resolution to characterise the stromal landscape, coloured by cell type. To note, cells assigned 'soup' were not able to be clearly defined by canonical cell type markers and were indicative of multiplets.

Principal differences in Covid-19 compared to LRTD were in myeloid cells, particularly alveolar macrophages (Figure 2.14), while few genes were expressed at higher levels in lymphocytes, mast cells, or stromal cells. This could be down to few cell numbers in the non-LRTD group for which we only had one sample, thus differential gene expression lacked sufficient power. As our patient cohort had low values of days between symptom onset to death we decided to focus on the key players in acute COVID-19 infection including myeloid cells and T-cells that may contribute to the cytokine storm and immune cell infiltration to the lungs.

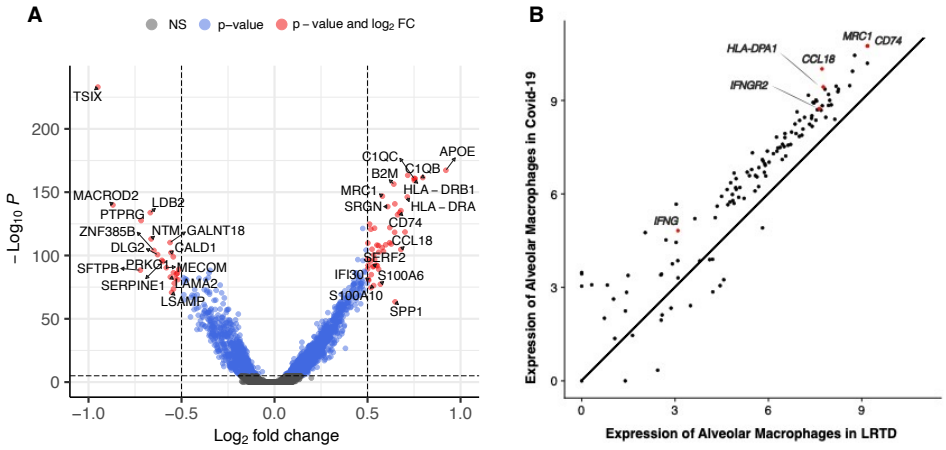


Figure 2.14: A) Volcano plot showing top differentially expressed genes in alveolar macrophages in COVID-19 compared to LRTD with a significant adjusted p-value (<0.05) and a log-fold change of more than 0.5 using MAST followed by Bonferroni multiple test correction. B) Plot shows expression levels of different IFNG module genes in alveolar macrophages between Covid19 and LRTD cases. Line is at 1:1 ratio hence dots to the left of the line indicate genes with higher expression in Covid19 cases and to the right of the line indicates genes with higher expression in LRTD. The IFNG receptors IFNGR1 and IFNGR2

In alveolar macrophages, top differentially regulated genes included markers of tissue residency (*C1QC*, *C1QB*) and factors shown to mediate lung fibrosis (*CCL18*) and apoptosis (*S1006*), as well as activation and recruitment of other myeloid cells (*SPP1*). IFN-γ response protein (*IFI30*) and MHC proteins (*HLA-DRA*, *HLA-DRB1*) were all up-regulated, indicating response to IFN-γ.

This IFN-γ dominant response contrasts with Type I and III dominant interferon responses shown to be critical in pathogenesis in Northern hemisphere Covid-19 cohorts. Given our IMC data[207] indicating a prominence of alveolar macrophages in the immune response and in alveolar damage, we analysed alveolar macrophage interferon response modules: IFN-γ response pathways were strongly up-regulated in Covid-19 compared to LRTD. IFN-β, IFN-λ, and TNF responses were also up-regulated but to a lesser degree. Across other myeloid cell IFN responses were heterogeneous, and TNF response was up-regulated in the LRTD group in CD4+ T-cells (Figure 2.15).

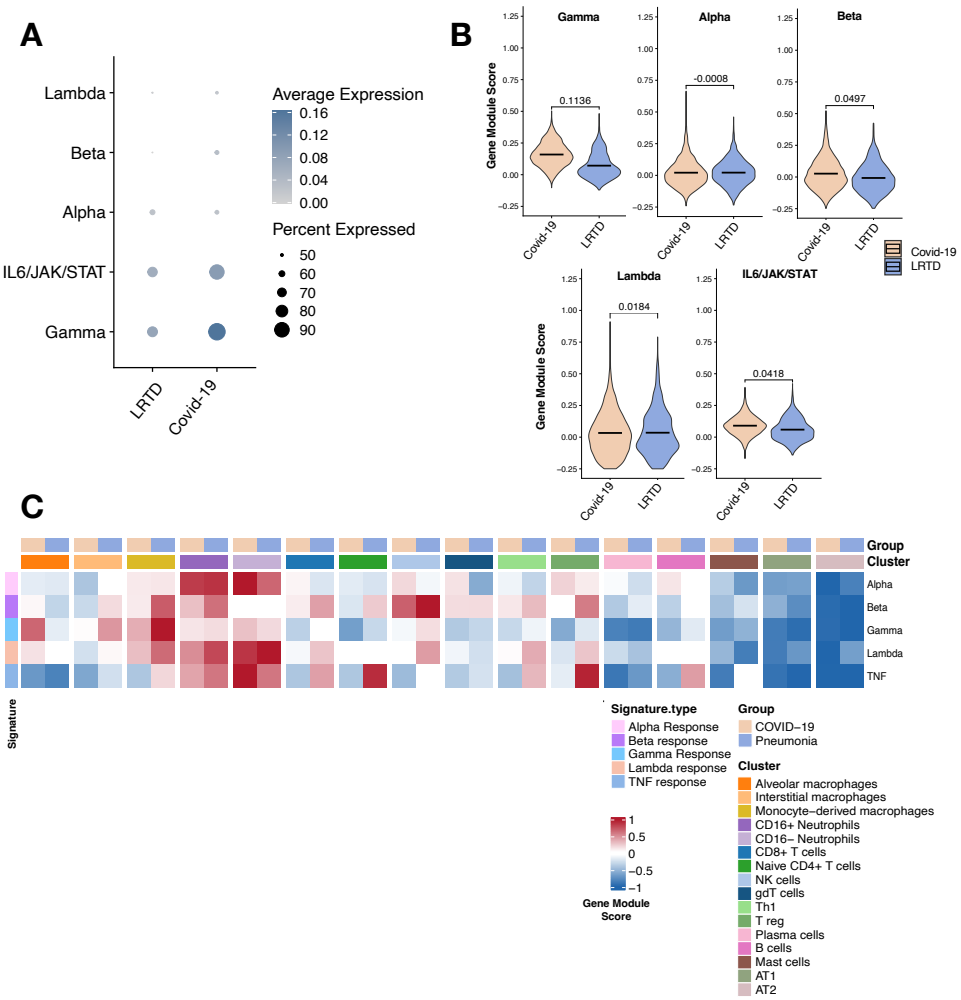


Figure 2.15: A) Dotplot showing the average gene module score of interferon response pathways across alveolar macrophages in Covid-19 and LRTD. B) Violin plots showing the gene module score across alveolar macrophages in gene sets associated with the gamma, alpha, beta, lambda and IL6 response in Covid-19 compared to LRTD. Black lines indicate the mean value across all cells, with the log fold change between means across conditions annotated above the plots. C) Heatmap showing the mean gene module score across cells in gene sets associated with the alpha, beta, gamma, lambda and TNF response. Cell types have been grouped by Covid-19 and LRTD to show the difference in response and module score values have been scaled between –1 and 1.

2.4.5 Integration with Human Lung Cell Atlas (HLCA): IFN-γ driven responses in Malawi cohort and type I/III interferon responses in other cohorts

To validate the IFN-γ response in the Malawi cohort compared to type I and III interferon and IL1-dominant responses described in Northern hemisphere cohorts, and to understand the implications for distinct, therapeutic approaches, we integrated our single-cell data with multi-cohort Covid-19 (5 cohorts, 60 cases), LRTD (1 cohort, 13 cases), and non-LRTD (23 cohorts, 178 cases) data from the Human Lung Cell Atlas (HLCA) (Figure 2.16).

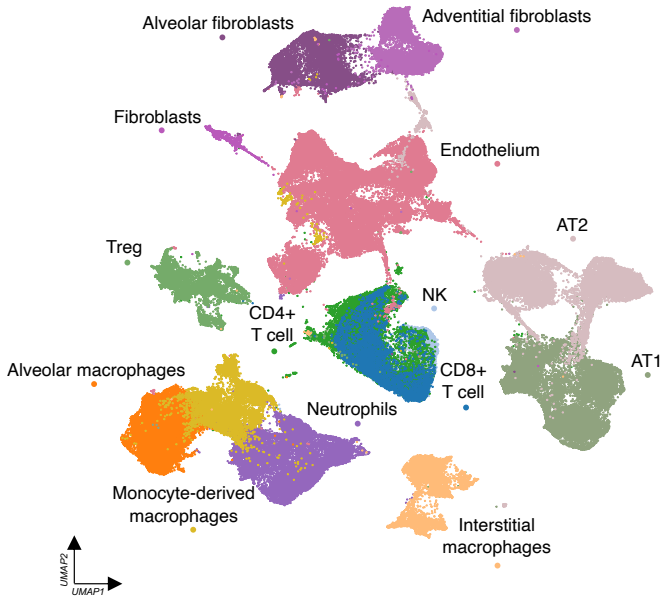


Figure 2.16: UMAP visualisation of 147,935 lung cells deriving from integrating cells from Covid-19, LRTD and non-LRTD cases from our cohort with cells from the human lung cell atlas (HLCA) from non-LRTD, LRTD and Covid-19 cases. Clusters are coloured by cell type.

Differential expression analysis was completed comparing our Malawi COVID-19 cohort to the non-LRTD cases in the HLCA atlas with up-regulated antigen presentation genes in myeloid cells like alveolar macrophages, such as *HLA-DRA*, *HLA-DRB1*, *HLA-B*, *HLA-A* expected in lung inflammation and also captured in our previous analysis (Figure 2.14). Additionally, genes relating to a type II interferon repsonse were also observed to be

up-regulated, such as *IFNGR2* and *JAK2*. We also wanted to investigate the differences between COVID-19 cases in our Malawian cohort to COVID-19 cases in the HLCA. Similarly, amongst the top differentially expressed genes we had upregulation of MHC class II related genes and interferon type II related genes upregulated in immune cell populations. To gather a more global view of the key differences in Malawian COVID-19 cases and COVID-19 cases derived from the HLCA, we used pathway analysis to profile cellular response differences between our cohort and cohorts in the HLCA (Figure 2.17). Pathways indicative of IFN-γ response were increased across all cell types in the Malawi cohort. Furthermore, *IFNG* (IFN-γ gene) was specifically increased in the Malawi cohort in CD4+ and CD8+ T-cells versus HLCA Covid-19 and non-LRTD groups, suggesting that macrophages are responding to IFN-γ produced by T-cells. Other inflammatory pathways showed a mixture of up and down-regulation in the Malawi cohort compared to HLCA cohorts, including *IL6/JAK/STAT* and *TNF-NFKB*, key targets for therapies being used in Covid-19. Many of the other interferon-response genes were more up-regulated in the HLCA cohorts or had a heterogeneous distribution across cells, although notably monocyte-derived macrophages generally had a higher interferon response in HLCA Covid-19 cohorts.

Figure 2.17: A) Heatmap showing pathway analysis for differentially expressed genes in our COVID-19 cohort compared to the HLCA COVID-19 cohort. Shown are the 50 canonical hallmark gene sets (for list see Supplemental information) coloured by the normalised enrichment score for each cell type. Gene ontology pathways of interest are indicated by arrows (IL6 JAK STAT3 SIGNALING, green, TNFA SIGNALING VIA NFKB, blue, INTERFERON GAMMA RESPONSE, orange). B) Dot plot showing the average expression of top differentially expressed genes in the lung alveolar macrophages that contribute the highest in the hallmark gene set "INTERFERON GAMMA RESPONSE" pathway in our COVID-19 cohort compared to the HLCA COVID-19 cohort.

We also examined the gene expression of known inflammatory mediators to look for cell-specific expression of cytokines and chemokines in the lung across the Malawi and HLCA cohorts (Figure 2.18). This revealed increased expression of *IFNG* by CD4+/CD8+ cells in the Malawi cohort when compared to, not only the Malawi LRTD cases, but also

the HLCA Covid-19 cases. This increase of expression is notably absent when comparing Covid-19 and LRTD cases within the HLCA cohort, indicating the potential production of *IFNG* by T-cells and the response by alveolar macrophages is specific to our Malawi cohort.

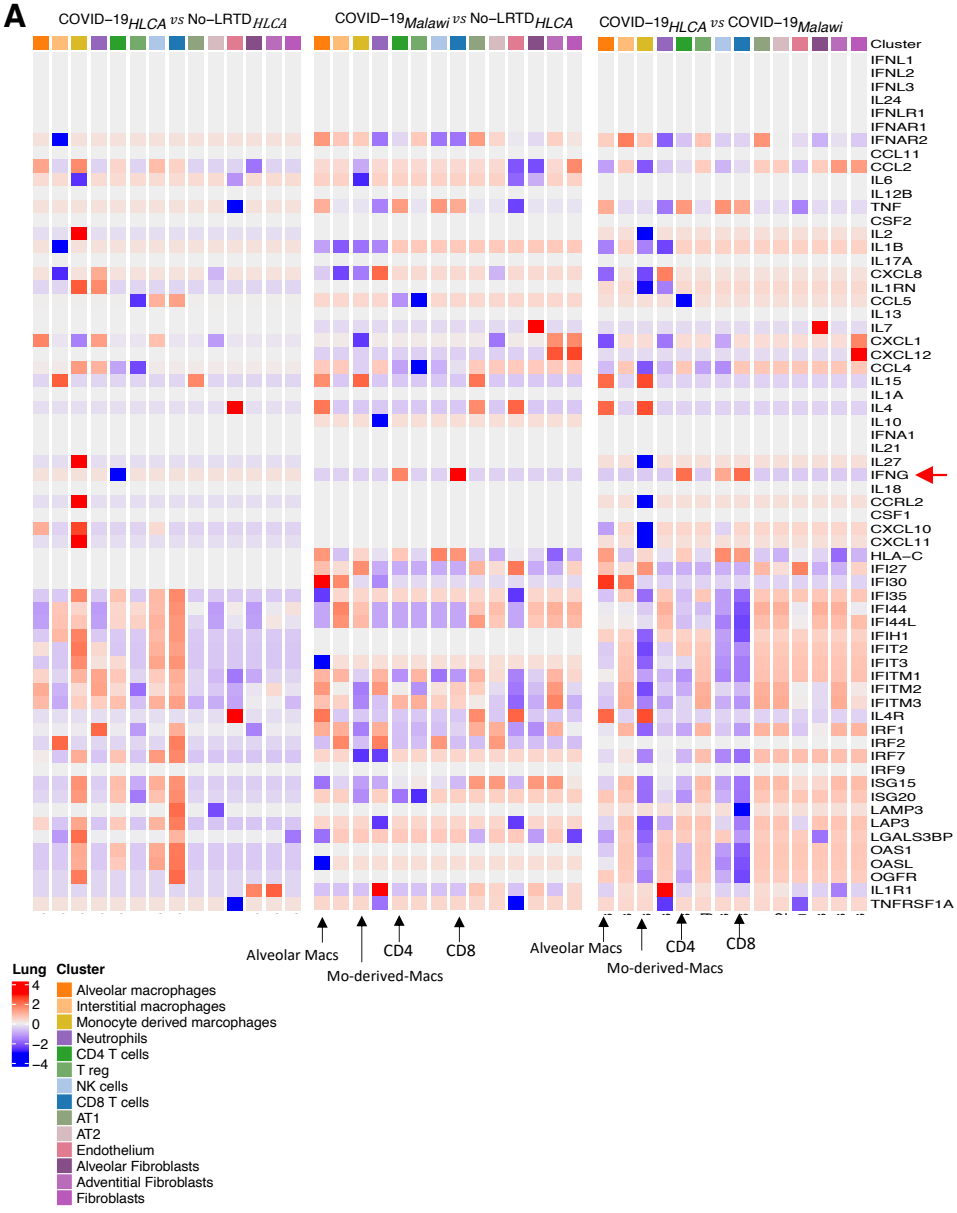


Figure 2.18: A) Heatmaps showing the log fold change of up/down-regulated interferon response genes taken from immunologic gene sets involved in the immune response. Comparisons include the change in interferon response in cells from the HLCA COVID-19 cohort compared to HLCA control cases (left), the Malawi COVID-19 cohort compared to control cases from the HLCA (middle) and interferon responses from our COVID-19 cohort compared to the HLCA COVID-19 cohort (right).

These data show many shared inflammatory pathways between Malawi and HLCA cohorts but with an amplified IFN-γ response in the Malawi cohort, highlighting IFN-γ production from CD4+/CD8+ T-cells and response in alveolar macrophages.

To further validate these findings, we wanted to ensure that this IFNG response persisted in cohorts that were more directly comparable with our study. We found two existing studies conducted by Melms et al.[4] and Delorey et al.[3] where lung autopsy single cell/single nuclei had been conduced on lethal COVID-19 on patients in the US. This patient cohort consisted predominantly of Caucasian and Hispanic patients and had comparable clinical metadata to our Malawi study (Figure 2.19). We integrated our 9 Covid-19 cases with 16 cases from Delorey and 19 cases from Melms, yielding 200,000 lung cells across 21 clusters of immune and stromal compartments, using 38 PCs at a clustering resolution of 0.3 (Table 2.8). Gene module score analysis of the alveolar macrophages showed a marked increase in response to interferon gamma as seen when compared to the HLCA cohort. When observing the results from gene ontology analysis we can see an increase in the interferon gamma response in alveolar macrophages as well as response to type II interferon when compared to northern hemisphere post-mortem cohorts.

Together with these two integrations we can demonstrate that the response to interferon gamma in the macrophage compartment is increased in the Malawi cohort when compared to the northern hemisphere cohorts in lethal and non-lethal Covid-19 comparisons.

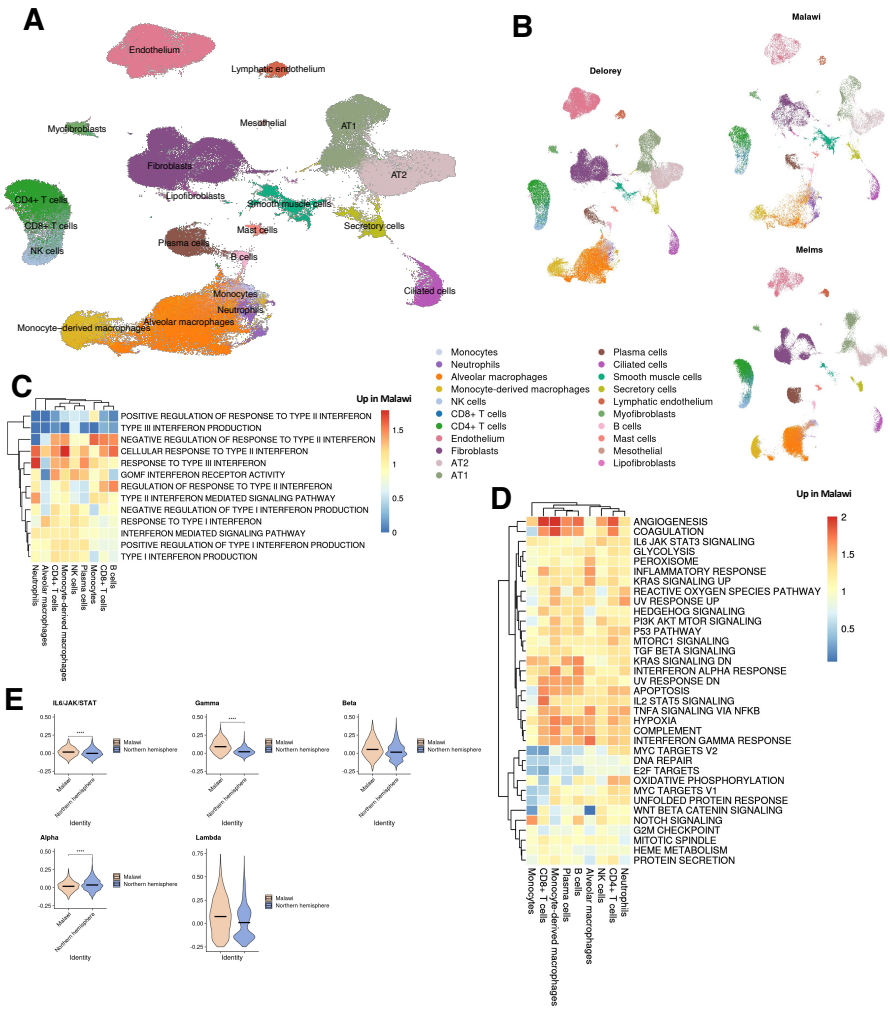


Figure 2.19: A) UMAP visualisation of 200,000 lung cells deriving from integrating cells from Covid19 cases from our cohort with cells from two existing Northern Hemisphere (USA) lung autopsy study lethal Covid19 cases. Clusters are coloured by cell type. B) UMAP visualisation to show integration across the three datasets; Malawi being our cohort, Delorey and Melms being the two Northern hemisphere cohorts that were combined and termed 'Northern hemisphere' for the additional figure panels. C) Heatmap showing pathway analysis for differentially expressed genes in our COVID-19 cohort compared to the Northern hemisphere COVID-19 cohorts. Shown are all Biological Process gene sets that are affiliated with interferon response taken from MsigDB coloured by the normalised enrichment score for each immune cell type. D) Heatmap showing pathway analysis for differentially expressed genes in our COVID-19 cohort compared to the Northern hemisphere COVID-19 cohorts. Shown are relevant filtered canonical hallmark gene sets coloured by the normalised enrichment score for each cell type. E) Violin plots showing the gene module score across CD8+ T cells in gene sets associated with the gamma, alpha, beta, lambda and TNF response in Malawi COVID-19 compared to Northern hemisphere COVID-19. Black lines indicate the median value across all cells, with asterisks to denote the significance level (ns = non-significant, **** = p <= 0.0001).

2.4.6 Single-cell analysis of nasal cells may be a useful proxy for lung parenchymal responses

While lung is the principal organ involved in severe and fatal Covid-19 disease, lung samples are not easily accessible during life. For future Covid-19 waves or other emerging diseases it would be invaluable to predict lung responses using nasal or blood samples that can readily be obtained. We performed scRNA-seq on nasal cells in 8 cases (5 Covid-19; 2 LRTD and 1 non-LRTD) and peripheral blood mononuclear cells in 7 individuals (4 Covid-19, 2 LRTD and 1 non-LRTD). We recovered 8,098 nasal cells which mapped to ten clusters composing immune and stromal cells (Figure 2.20) and 13,350 blood cells (Figure 2.21). Mapping statistics and additional parameters are described in the scRNA mapping and quality control section in Table 2.7 and Table 2.6.

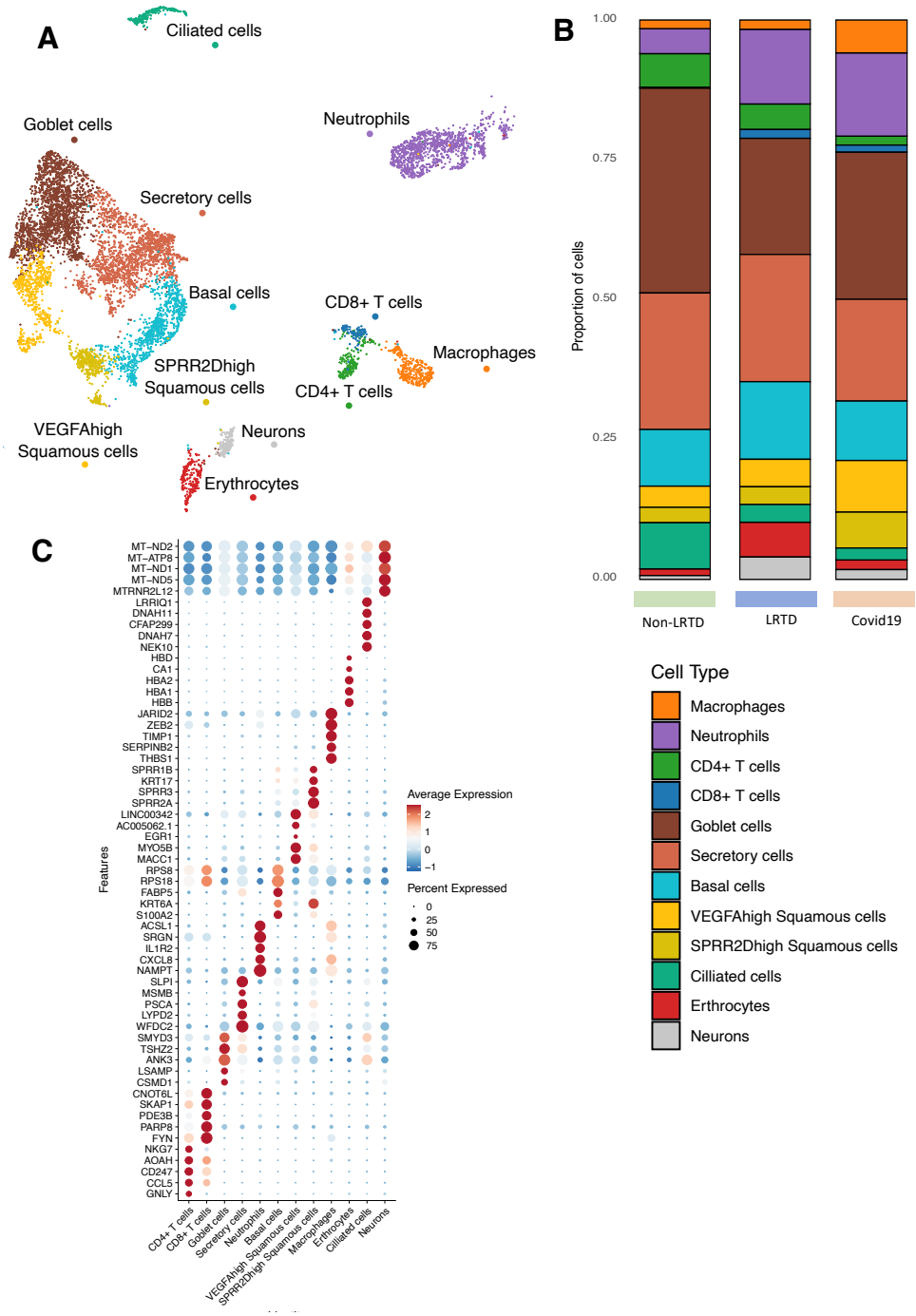


Figure 2.20: A) UMAP visualisation of 8,098 nasal cells across our cohort, coloured by broad cell types. B) Cell type proportion bar plots of cell types from nasal scrapings, grouped by disease group. C) Dotplot showing the average expression of top 5 cluster markers for each cell type in the nasal compartment.

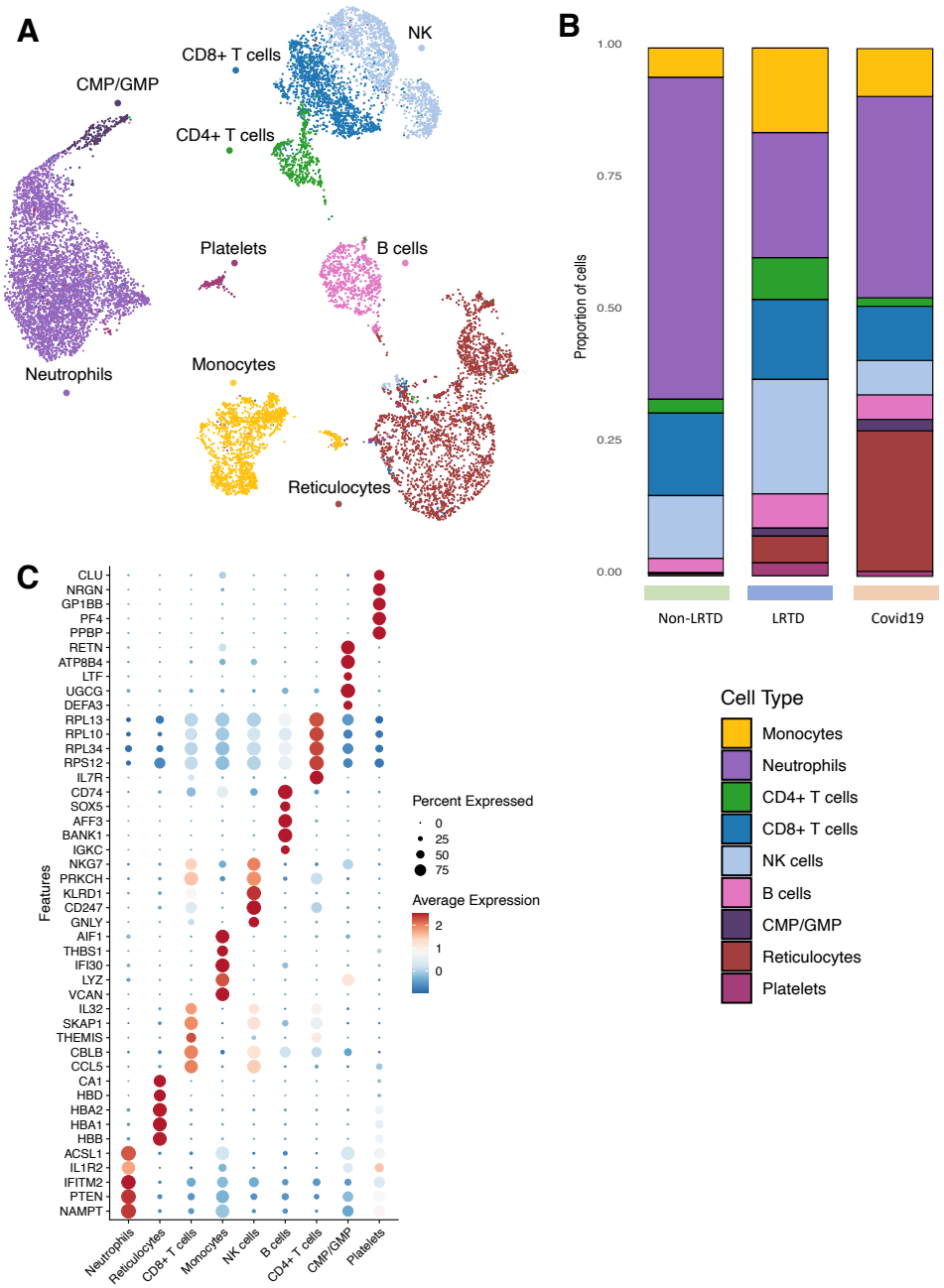


Figure 2.21: A) UMAP visualisation of 13,350 blood cells across our cohort, coloured by broad cell types. B) Cell type proportion bar plots of cell types from whole blood, grouped by disease group. C) Dotplot showing the average expression of top 5 cluster markers for each cell type in the blood compartment.

Nasal macrophages had several similar differentially expressed genes in the Covid-19 versus LRTD cases that mirrored lung alveolar macrophage responses including *SPP1* and *C1QB*, genes indicative of proliferation (*LGALS1*, *TMSB10*) and *MHCII* genes (*HLA-DPB1*, *HLA-DQA1*) (Figure 2.22).

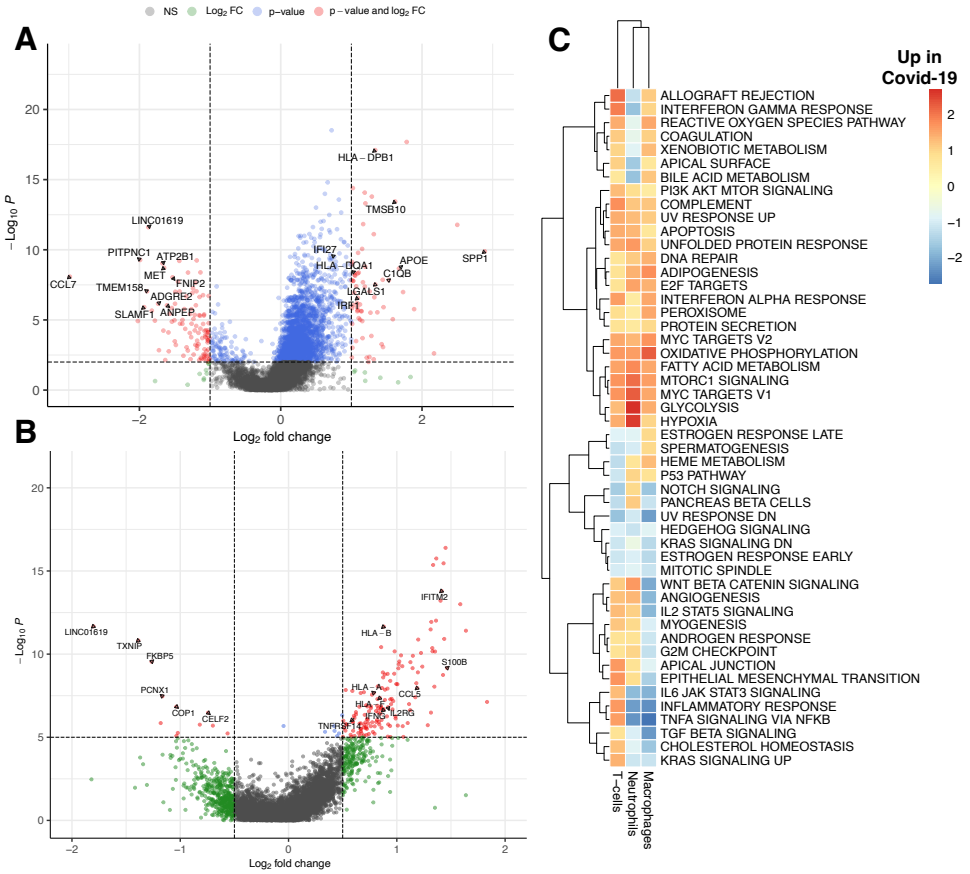


Figure 2.22: A-B) Volcano plots showing top differentially expressed genes in nasal macrophages and T-cells in COVID-19 compared to LRTD with a significant adjusted p-value (<0.05) and a log-fold change of more than 0.5 using MAST followed by Bonferroni multiple test correction. C) Heatmap showing pathway analysis for differentially expressed genes in our COVID-19 cohort compared to the LRTD cohort. Shown are the 50 canonical hallmark gene sets (for list see Supplemental information) coloured by the normalised enrichment score for each cell type.

MHC class II gene up-regulation is a canonical response to IFN-γ and consistent with this there was *IFNG* (IFN-γ gene) up-regulation in T-cells in the Covid-19 cases in comparison to LRTD cases (Figure 2.22). Pathway analysis showed higher levels of IFN-γ response in macrophages and T-cells, further validating an IFN-γ response in these cells (Figure 2.22). Blood monocytes in Covid19 versus LRTD cases had up-regulation of the alarmin *S100A12* and of genes involved in inflammation (*AREG*) and vascular damage (*NDRG1*) but not in genes indicative of IFN-γ response and *IFNG* was not up-regulated in T-cells (Figure 2.23).

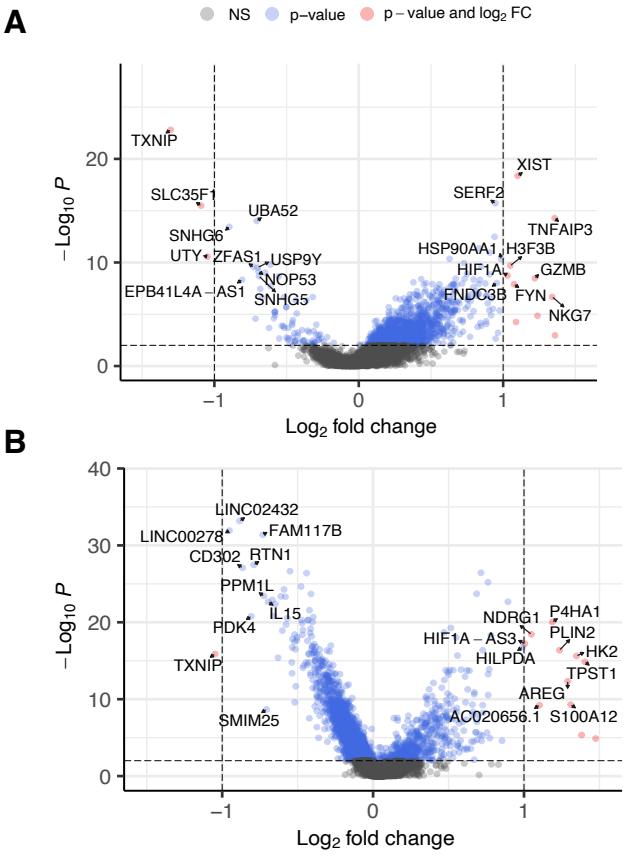


Figure 2.23: A-B) Volcano plots showing top differentially expressed genes in blood T-cells and monocytes in COVID-19 compared to LRTD with a significant adjusted p-value (<0.05) and a log-fold change of more than 0.5 using MAST followed by Bonferroni multiple test correction.

Hence, in our small cohort, nasal cells better paralleled lung response than blood cells, supporting previous Covid-19 and non-Covid-19 studies that highlighted the utility of nasal cells for understanding respiratory immune responses.

Since scRNA-seq is not available in most settings, we assessed the extent to which cytokine responses (Luminex) in plasma or nasal fluid could distinguish the inflammatory or IFN-γ response in Covid-19 versus LRTD cases. In nasal fluid there was a trend towards several cytokines being higher in Covid-19 cases than in LRTD cases, but none significant and no clear difference for IFN-γ (Figure 2.24).

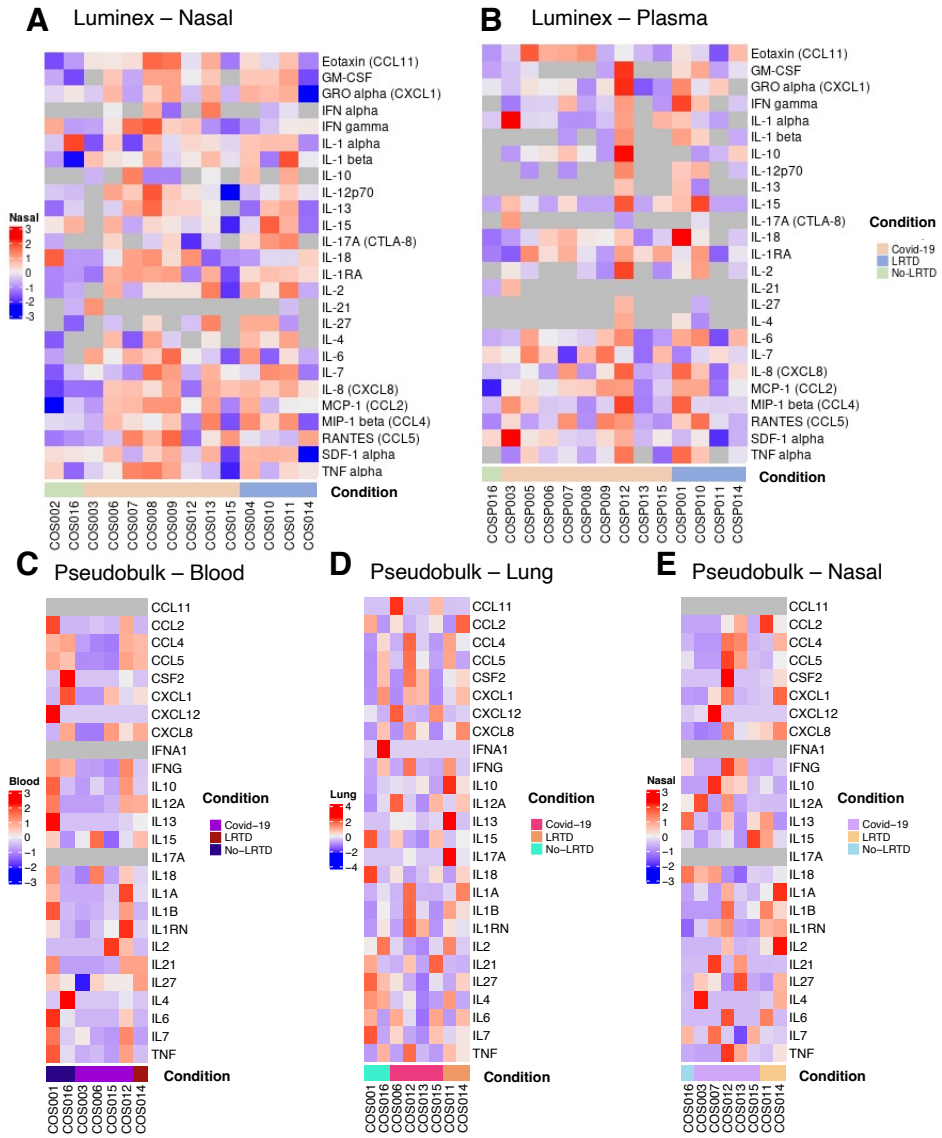


Figure 2.24: A-B) Heatmaps showing cytokine signatures in different tissues. Values are plotted as z-score (grey mean not measured). Samples are grouped by their disease type. Luminex data of Nasal (A) and Plasma (B). Data were transformed with a log2 and for the visualistion with ComplexHeatmap in R with a Z-score by gene. For the statistical tests we compared levels of IFN-y, IL6, IL8, TNF and IL1b in nasal fluid and plasma between the Covid-19 and LRTD samples using a Welch Two Sample t-test which was non-significant for all comparisons, we did not correct for multiple comparisons. (C-E) Pseudobulk heatmaps showing cytokines included in the Luminex panel on the transcriptomic level in the peripheral blood, lung and nasal compartment per patient. As for Luminex we compared levels of IFN-y, IL6, IL8, TNF and IL1b in nasal, blood and lung cells between the Covid-19 and LRTD samples using a Welch Two Sample t-test which was non-significant for all comparisons, we did not correct for multiple comparisons.

There was no clear blood circulating cytokine response pattern, and no circulating cytokine levels were significantly higher in Covid-19 compared to other groups (Figure 2.24). A pseudo-bulk sequencing approach in blood, nasal and lung cells also did not distinguish a clear IFN-γ or any other specific inflammatory cytokine signature between Covid-19

and LRTD cases (Figure 2.24). Single-cell methods identified an interferon signature and T-cell-macrophage axis, bulk cytokine and gene expression approaches did not. Given very small numbers per group this is perhaps unsurprising. It may stem from greater discriminatory power of single-cell methods and is supportive of the value of single-cell approaches, particularly in small cohorts.

2.4.7 Stromal cellular interactions are driven by macrophages and vascular interactions by neutrophils

To validate our findings of the role of IFN-γ responding alveolar macrophages in lung parenchymal pathology and neutrophil interactions in vascular pathology, and to predict novel molecular interactions to target therapeutically we used cell interaction methods. First, unbiased receptor-ligand analysis of our scRNA-seq data highlighted that a large proportion of the imputed interactions in the lung involved alveolar macrophages, interacting with fibroblasts, epithelial cells and other immune cells including CD4+ T-cells (Figure 2.25), in keeping with our findings from cell proportions, cellular histology and scRNA-seq. Many interactions involved with antigen presentation were predicted expressed by alvelolar macrophages such as *HLA-DRA*, *HLA-DRB1* and particularly *HLA-DMA-CD9* interaction with vascular endothelium involved in macrophage activation[248]. Other interactions related to immune cell infiltration such as *ALOX5AP-ALOX5* between alveolar macrophages and alveolar type II pneumoncytes. Additional interactions such as *ITGAM* and *ITGB2* integrins expressed by alveolar macrophages indicating immune cell recruitment up-regulated in COVID-19 when compared to LRTD.

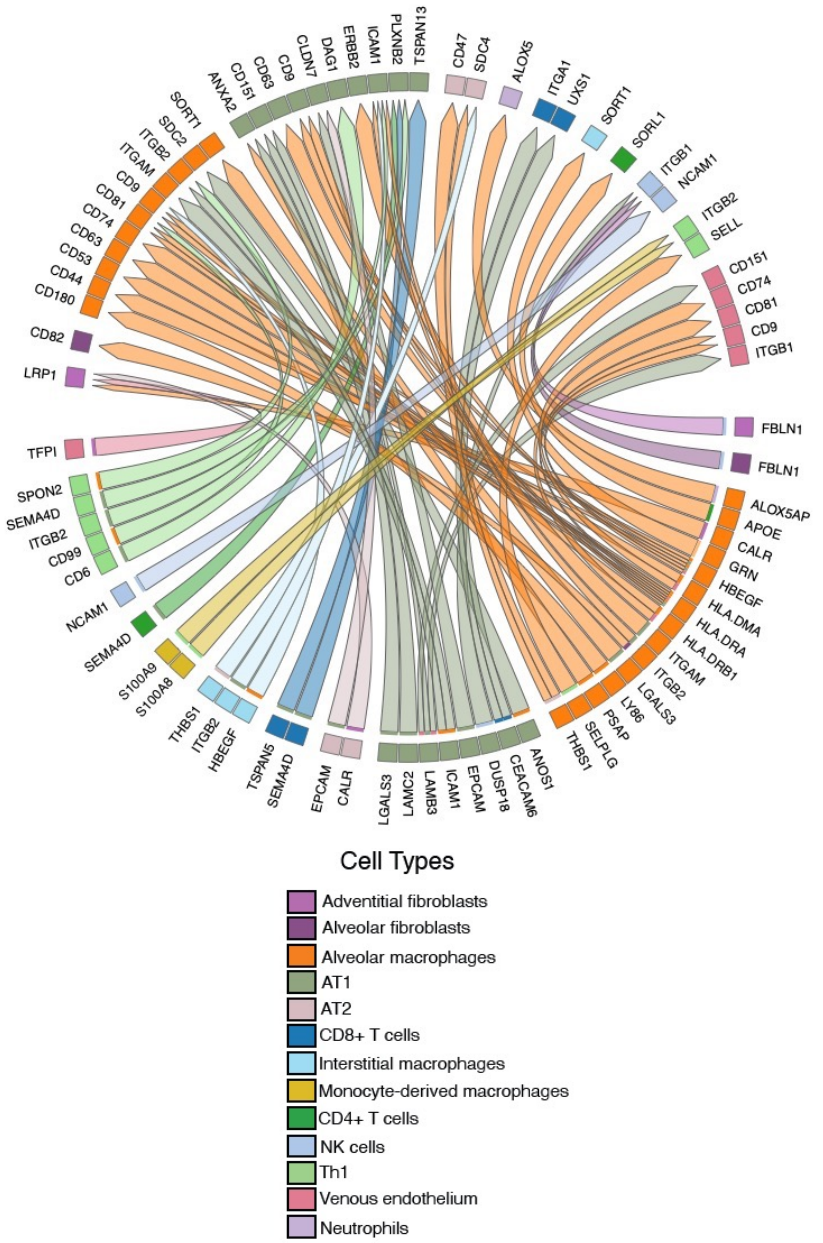


Figure 2.25: Circos plot showing the top 50 differentially expressed interactions upregulated in our COVID-19 cohort compared to LRTD. Segments are coloured by cell type with ligands and receptors labelled on the outside. Direction of the arrows show the senders of communications i.e. expression of ligand, and receiver of communications. Inner tracks on sender segments are coloured by the receiving cell type for ease of interpretation.

To validate these interactions in a spatial context we leveraged imaging mass cytometry analysis completed by Dr. João Da Silva Filho in Covid-19 and LRTD cases in our Malawi cohort. Here we had 130 representative regions of interest containing specific pathological lesions or normal lung areas (9 Covid, 3 LRTD, 2 Non-LRTD cases), with cell types delineated with a 39 metal-conjugated antibody panel. Neighbourhood enrichment analysis

was completed using the IMC data to identify cells located close to each other with greater than expected frequency as an indicator of their likelihood to be interacting (Figure 2.26). An additional integrative analysis using IMC data from cohorts of different demographics, Brazil and USA was also completed and can be read in our paper[207].

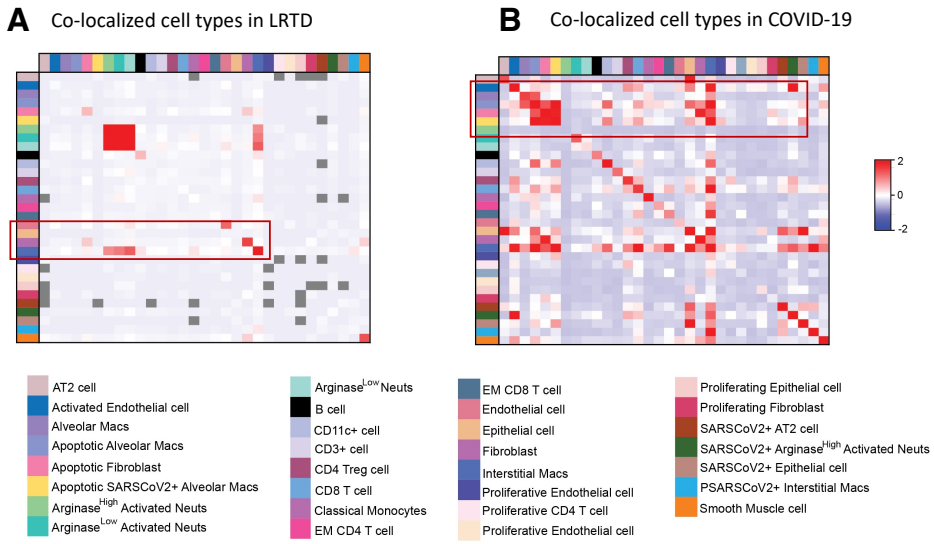


Figure 2.26: A-B) Heatmaps showing co-localised cell types as shown by the IMC providing insight into potentially interacting cell types in the lung.

The neighbourhood analysis was critical to hone in on particular cell types of interest to complete a more targeted cellular interaction approach involving cell types that are co-localising with each other. In the non-LRTD there were no significant interactions. The LRTD group was completely dominated by neutrophil interactions (Figure 2.26). In the Covid-19 group several neighbourhood enrichments were prominent – principally alveolar macrophages (with and without SARS-CoV2-S and apoptosis) with apoptotic fibroblasts and to a lesser extent type II pneumocytes (Figure 2.26).

This supports the role in pathogenesis of alveolar macrophages including the apoptotic population present in Malawi but not USA or Brazil cohorts inferred from previous IMC integration and analysis. In contrast, the most prominent neighbourhood enrichment for neutrophils was between SARS-CoV2-S+, Arghigh neutrophils and activated endothelial cells implicating neutrophils in endothelial/vascular pathology.

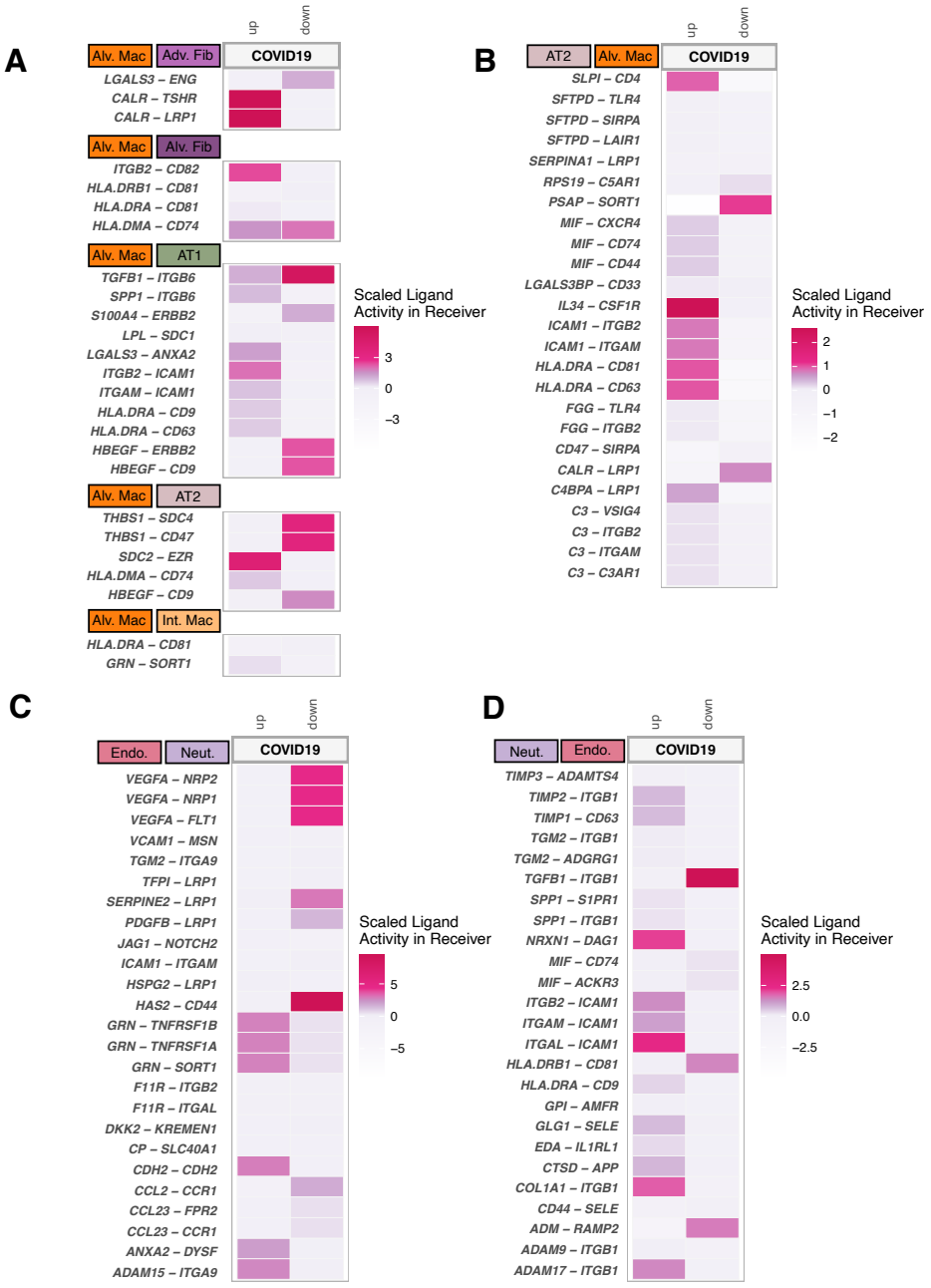


Figure 2.27: A) Heatmap showing up/down-regulated interactions in COVID-19 compared to LRTD driven by AT2 pneumonocytes to alveolar macrophages. Coloured boxes indicate cell type with the ligand-expressing cell type followed by the receptor-expressing cell type. B) Heatmap showing up/down-regulated interactions in COVID-19 compared to LRTD driven by lung alveolar macrophages to lung epithelial cells and interstitial macrophages. Coloured boxes indicate cell type with the ligand-expressing cell type followed by the receptor-expressing cell type. C) Heatmap showing up/down-regulated interactions in COVID-19 compared to LRTD driven by lung endothelium to neutrophils. Coloured boxes indicate cell type with the ligand-expressing cell type followed by the receptor-expressing cell type. D) Heatmap showing up/down-regulated interactions in COVID-19 compared to LRTD driven by neutrophils to lung endothelium. Coloured boxes indicate cell type with the ligand-expressing cell type followed by the receptor-expressing cell type.

We then looked at validated interactions in Covid-19 in closer detail in scRNA-seq data focusing on the co-localising cell types enriched in the COVID-19 IMC neighbourhood analysis. Macrophage interactions were frequently from ligands on type II pneumocytes to receptors on alveolar macrophages (Figure 2.27), in keeping with type II pneumocytes cells generally being a principle infected cell. Several of these interactions involved macrophage inhibitory factor (*MIF*) from type II pneumocytes with *CD74*, *CD44* and *CXCR4* on macrophages, a classical response chain in macrophages and a key initiator of proliferation, chemotaxis and activation. *ICAM-1* on type II pneumocytes was predicted to signal to integrins (*ITGB2-ITGAM*) on alveolar macrophages, an interaction involved in cellular attachment during recruitment. Another strong predicted interaction was *IL-34-CSF1R*, involved in triggering macrophage activation and chemotaxis. Reciprocally, there were several interactions between alveolar macrophages and epithelial cells consistent with our IMC data that indicate their role in alveolar pathology. These included *SPP1* and *TGFβ* with type II pneumonocyte integrins (*ITGB6*) (Figure 2.27), interactions implicated in lung pathology and fibrosis. We identified multiple neutrophil interactions with endothelial cells indicating processes involved in neutrophils attachment to the vascular wall (e.g., *ITGAL-ICAM-1*) and of activation by neutrophil granule proteins (*GRN-TNFRSF1A*) (Figure 2.27), providing molecular validation supporting their role in coagulation, endothelial activation and vascular pathology indicated by IMC.

To further validate the IFN-γ response in Malawian patients, we integrated both the scRNA-seq and IMC data and mapped gene expression profiles onto IMC cells using a recently developed pipeline[249] to observe projected gene expression onto the IMC protein data.(Figure 2.28). The integrated output showed upregulation of IFN-γ response genes, including*HLA-DR*, *IFI30* and *APOE*, and the inducible component of the IFN-γ receptor (*IFNGR2*) in tissue-resident CD206high alveolar and interstitial macrophages. Notably, the IFN-γ response was most prominent in the SARS-CoV-2+ and apoptotic CD206high

macrophage populations, predicted to interact with apoptotic fibroblasts and type II pneumocytes in the neighborhood analysis. Thus, mapping scRNA-seq data onto our IMC data not only validates the IFN-γ response but also implicates these IFN-γ-responding cells in lung stromal cell damage.

Additionally, with work completed by Dr. Vanessa Herder at the Centre for Virus Research (CVR), we validated the IFN-γ response using in situ hybridization staining across patients and 138 ROIs which highlighted significantly higher numbers of *IFNGR2+* cells in patients with COVID-19 than in non-LRTD controls but not between non-LRTD patients and patients with LRTD. *IFNGR2* was predominantly in CD206high cells, which could be observed in diffuse alveolar damage lesions. In contrast, the number of *IFNG+* cells was not significantly increased in patients with COVID-19, validating findings from scRNA-seq. Thus, multiple orthogonal methods demonstrate an IFN-γ response in CD206high lung-resident macrophages, and this is best explained by the responsiveness of these cells rather than increased inflammation and *IFNG* production.

These data highlight the value of a combined scRNA-seq and IMC approach. They provide spatial and receptor-ligand validation for roles of alveolar macrophages in molecular processes that are plausibly involved in alveolar damage and lung fibrosis, and for neutrophils in endothelial activation. The data predict specific molecular interactions involved in these processes. If validated by further work, some of these interactions may be plausible targets for intervention, e.g., *MIF* for which several small molecules are in clinical development for therapy in inflammatory disorders.

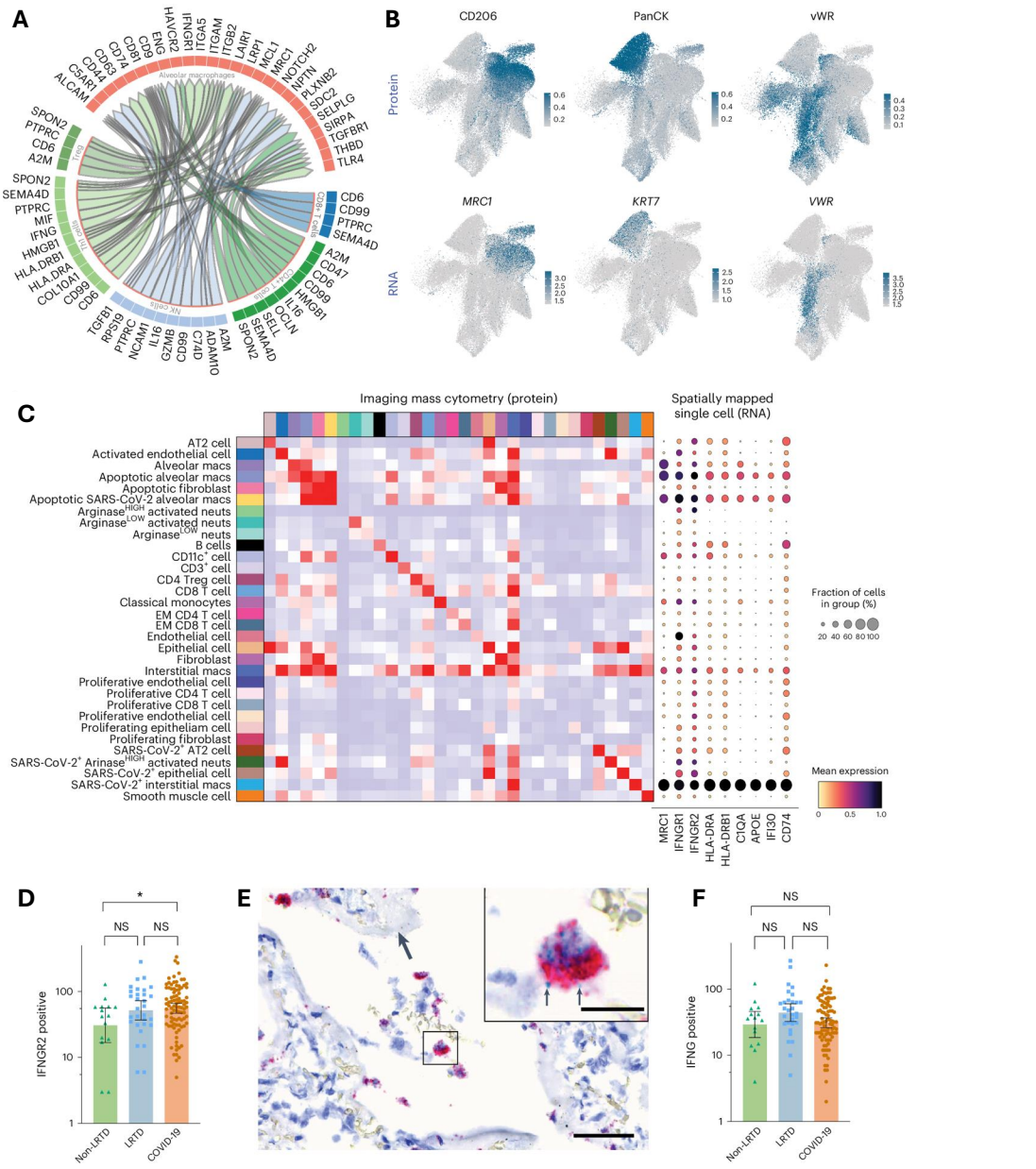


Figure 2.28: A) Circos plot showing the top cell–cell interactions from immune cells to alveolar macrophages in Malawian patients with COVID-19 versus Malawian patients with LRTD. Segments are colored by cell type with ligands and receptors labeled on the outside. Direction of the arrows shows the senders of communications that are expressing a given ligand to the receiver cell type expressing its cognate receptor. Inner tracks on sender segments are colored by the receiving cell type for ease of interpretation. B) UMAP plots to show expression levels of different hallmark proteins in different clusters by IMC and then below RNA levels from scRNA-seq data imputed by MaxFuse. C) Heatmaps showing co-localized cell types from IMC data, providing insight into potentially interacting cell types in the lung in patients with COVID-19. D-F) Quantification of mRNA in situ staining for IFNGR2 and IFNG in tissue. In total, 138 ROIs were taken based on multiple sampled areas from the left and right lung in patients with COVID-19 (n=9), LRTD (n=3) and non-LRTD (n=2). Separate TMA sections were dual stained for either IFNGR2 and MRC1 (CD206) or IFNG and CD3E mRNA by in situ hybridization, and then the number of cells positive for each stain within respective cells of interest IFNGR2 in CD206+ cells and IFNG in CD3E+ cells was analyzed by automatic quantification. Each dot represents the quantities of positive cells in an independent tissue core that were used as replicates for analysis in D and F. These data were log transformed and analyzed using one-way ANOVA and Tukey's multiple comparison test to adjust for multiple comparisons and a pre-defined alpha level of 0.05. Coloured bars show the geometric mean, and error bars show the 95% confidence interval. D) Compared to non-LRTD patients (green), there were significantly higher numbers of IFNGR2+ cells in patients with COVID-19 (orange) but not in patients with LRTD (blue)(P=0.0441). E) Co-staining of IFNGR2 (red) and CD206 gene (green) using mRNA probes in lungs of patients infected with SARS-CoV-2. Lung of patients with COVID-19 shows, in the periphery of the damaged alveolar space fibrin (empty arrows) and in the lumen of the alveoli, cells with macrophage morphology expressing IFNGR2 (red signal, rectangle). The insert shows a higher magnification of the rectangle with a macrophage expressing CD206 in green (black arrows) and abundant IFNGR2 in red. Scale bars, 60µm and 15µm, respectively. F) No significant difference was observed in quantities of IFNG+ cells among the different groups (Non-COVID-19;green LRTD;blue, COVID-19;orange), (NS, not significant; P=0.111).

2.5 Discussion

In a multicentre collaborative approach spanning cross-continents with groups in Malawi and Glasgow, we conducted minimally invasive autopsies on fatal COVID-19 and other LRTD and non-LRTD cases in a Malawian population and characterised pulmonary, blood and nasal immune responses using scRNA-seq and IMC. While other studies have used these techniques for COVID-19 investigations in other settings, this is the first such study in a SSA population and to our knowledge some of the first scRNA-seq data from lung samples in any SSA population. Furthermore, conducting this study during the pandemic, facing logistical and technical challenges, using novel techniques for tissue extraction and implementing a multi-modal approach to orthogonally validate findings demonstrate the technical achievement of this unique collaborative effort. Our de-identified data are provided open access, including tools for visualising single-cell and histology data, making an important resource for furthering the global understanding of COVID-19 pathogenesis, immune responses in SSA populations and more widely for the human cell atlas. Access to all fully annotated objects are interactively hosted for exploration hosted on Glasgow-based servers using the cellxgene-VIP[250] platform here: Lung Atlas: https://cellatlas-cxg.mvls.gla.ac.uk/COSMIC/view/COSMIC_Lung_Atlas.h5ad/, Lung Immune Atlas: https://cellatlas-cxg.mvls.gla.ac.uk/COSMIC/view/COSMIC_Lung_Immune_Atlas.h5ad/, Lung Stromal Atlas: https://cellatlas-cxg.mvls.gla.ac.uk/COSMIC/view/COSMIC_Lung_Stromal_Atlas.h5ad/, Nasal Atlas: https://cellatlas-cxg.mvls.gla.ac.uk/COSMIC/view/COSMIC_Nasal_Atlas.h5ad/, Blood Atlas: https://cellatlas-cxg.mvls.gla.ac.uk/COSMIC/view/COSMIC_Blood_Atlas.h5ad/, Histopathology slides on virtual microscope: https://covid-atlas.cvr.gla.ac.uk, and the IMC: https://cellatlas-cxg.mvls.gla.ac.uk/COSMIC/view/COSMIC_IMC_Lung.h5ad/.

Given that many parasitic infections induce immune tolerance we hypothesised that there might be an attenuated immune response in SSA populations, blunting immune-mediated viral clearance and leading to high viral-loads in individuals who present with life-threatening disease. If so, pathology might be driven by direct-viral effects rather than hyperinflammation, indicating a need for different treatment approaches from Northern hemisphere cohorts. In fact, we found a robust immune response and comparatively low levels of virus, surprisingly even in highly immunosuppressed cases with HIV. Our data indicate that pathology is driven by inflammation, with many similarities to other non-African cohorts. These similarities are reassuring, indicating that many principles for diagnosis and treatment can likely be extrapolated from more extensively investigated populations. However, there were also differences that may have implications for therapy, in particular IFN-γ responses were upregulated in comparison to a large multi-country integrated HLCA dataset and additional existing fatal Covid-19 lung atlases. IFN-γ was produced by T-cells, with alveolar macrophages the principle responding cells. Spatially-resolved IMC neighbourhood analysis and scRNA-seq receptor-ligand analysis orthogonally validated these processes alongside further validation using in situ staining of lung tissue. In contrast *IL6* and *TNF* responses were not as prominent. scRNA-seq of nasal cells also identified *IFNG* upregulation in T-cells and evidence of IFN-γ response in macrophages in a sample type that is readily accessible, supporting prior data on the utility of nasal cells as an accessible proxy for lung responses.

There is cross-over between the responses of different interferons and IFN-γ signal has been detected previously in COVID-19 lung[3], yet it is interesting to consider why there was such a marked upregulation in our cohort compared to the large integrated HCLA dataset and other post-mortem cohorts. IFN-γ response has been shown to be a key component of effective immunity to malaria and is augmented in malaria exposed individuals, in part through epigenetic changes termed trained immunity[224]. Increased IFN-γ response was a key difference in SSA (Gabon) versus European individuals exposed to controlled human malaria infection and a correlate of protection[223]. While type I/III interferons are more typically involved in clearance of SARS-CoV2 and other respiratory viruses[251],

IFN-γ also plays a role, particularly in macrophages[252]. Considering our data with these prior studies we propose that trained immune responses to prior infections may favour an accelerated macrophage IFN-γ response. We hypothesise that this may be a double-edged sword in COVID-19 in SSA: such an accelerated trained response may generally be protective (through more rapid viral clearance), but in a subset of patients it may lead to accelerated hyperinflammation and collateral tissue damage. This hypothesis is supported by the short time between symptom onset and death in our cohort already with a clear hyperinflammatory response. Further exploration of macrophage responses in both SSA and non-SSA populations is therefore warranted. Considering the potential for translation the existing therapies for COVID-19 target JAK/STAT (Baricitinib), IL6 (Tocilizumab/sarilumab) or TNF (infliximab)[221,222]. JAK/STAT signaling is a conserved pathway for interferon responses including IFN-γ Thus our data, if corroborated, support potential efficacy of Baricitinib over other treatments. Baricitinib is a small molecule (tablet) and thus highly suited to wide distribution[222].

Our data have several limitations. Our cohort was small and in a single centre. Although single-cell methods have a higher capacity to resolve complex data in small sample sizes, many analyses in our study were underpowered. It is thus unclear how representative our data are of the wider Malawi or other SSA populations. Studies in other settings and ideally large multi-centre studies, are needed. While this would be a complex undertaking, we have demonstrated that single-cell methods are feasible in a SSA setting, and our study provides useful templates. While lung samples cannot readily be obtained in live patients, post-mortem studies have limitations: cells may change or degrade; pathological processes present early in disease are likely missed. Yet, post-mortem studies in northern hemisphere settings with longer post-mortem intervals identified validated targets[3]. Furthermore, the overall quality of our single cell data was impacted due to its post-mortem nature, leading to a change in sample preparation from fresh/frozen single cell sequencing to frozen single nuclei sequencing. Quality of the cells were further attenuated due to logistical limitations of conducting sample collection and processing of post-mortem tissue across two countries during a global pandemic. The technical impact of this created challenges at various stages

of the analysis such as introducing batch effects during integrations between not only patients but also cohorts. Also studies that we used for comparisons also had significant variation in methods and demographics from ours which may induce noise and bias. We used data-integration methods which reduce but do not eliminate this. We first adopted fresh single cell dissociated tissue samples that were collected at the Queen Elizabeth Central Hospital in Blantyre Malawi during the pandemic. Shallow sequencing of the initial fresh lung samples was performed in Glasgow and after preliminary mapping and analysis of the quality of reads it became apparent that the quality of the tissue was insufficient for downstream analysis. Mapping results indicated that we had a low fraction of reads within the cells indicating a large amount of ambient RNA and a low fraction of reads confidently mapping to the reference transcriptome. This prompted investigation of the quality of the FASTQ reads themselves, where we observed that the read quality was poor. This suggested that the tissue by the time of sequencing was of very poor quality and many of the cells had lysed or were degraded making downstream inference challenging. We then decided to change our tissue protocol to use single nuclei snap frozen tissue samples after seeking advice from the Delorey et al[3] authors who found that for post-mortem tissue this method preserved the integrity of the tissue sufficiently for downstream analysis. After this, we performed deeper sequencing of the tissue samples and found a vast improvement in mapping quality of the runs and proceeded with the analysis. Due to constraints of the study, we opted to multiplex multiple patient samples together and adopt a SNP clustering approach to try and computationally extract distinct genotypes to recover our patient samples. We realised that due to the quality of the post-mortem tissue that not all patient samples could be recovered by the Souporcell algorithm[70] and we lacked genotyping data from individual patients in the study. To overcome this, we matched SNP distribution patterns over highly variable HLA regions, an approach that is detailed in the results of this chapter. This method enabled us to recover cells from patients that were well represented in each run, while cells present only at very low proportions could not be recovered, likely because the starting material was insufficient to distinguish them as unique genotypes or because most of those cells coming from those patients had already died. The SNP matching over the HLA region was performed on a qualitative basis by the human eye and lacked in quantitative robustness however, we consolidated

this using a dummy example of merged BAM files from individual patient runs and found the same affect. For most of the nasal and blood samples we used a hashtag multiplexing approach where nasal and blood cells from the same patient were pooled and sequenced in a single run. This required mapping and demultiplexing of the hashtags to separate out the tissue populations for each patient run. Initially, the demultiplexing of the hashtags resulted in a large proportion of cells that were assigned as 'Negative' i.e. unable to be distinguished by the demultiplexing algorithm. To troubleshoot this I implemented a pipeline that runs the demultiplexing step using multiple different demultiplexing tools however the result remained the same irrespective of method. We then increased the depth of the sequencing of the hashtags to see if this improved the demultiplexing and reduce the number of negative cells, which marginally improved the proportions of recovered cells in the two tissues. To investigate this further we attempted an iterative clustering approach were cells were clustered together to see where they are distributed across the UMAP using information from the singlet cells that were assigned 'Nasal' or 'Blood'. In addition to this, there are distinct cell types in each tissue that would aid the separation in the clustering such as epithelial cell types in the nasal tissue that would be absent in the blood. This approach still resulted in ambiguous cell type assignment so proceeding we chose to only include the singlet cells that had successfully been assigned by the demultiplexing algorithm. This failure to accurately demultiplex the multi-tissue runs meant that we suffered a reduction in cell numbers for both tissue atlases which made the downstream analysis of some cell types challenging due to low cell numbers. Furthermore, as a result of the exclusion of samples either from low-quality fresh runs or failure to identify distinct genotypes in the multiplexed runs, it meant that overall we had very small sample numbers for this study. This meant that, we could not make comparisons between our COVID-19 patients and patients that had no reported LRTD as there was only one sample acting as our 'control' group across all tissue atlases. Thus, our comparisons were restricted to comparing our COVID-19 patients and our LRTD patients to find distinct mechanisms of action within our Malawi cohort. The low sample size effect was less pronounced in the lung tissue analysis however, in the nasal and blood the sample size was incredibly low, affecting the power and robustness of the results presented in this chapter. Despite the limitations, we recovered sufficient cell numbers to complete a robust analysis of the key

drivers of lethal COVID-19 pathology in our Malawian cohort particularly in the lung compartment. That being said, this limited us in examining other interesting aspects of the data such as the impact of HIV infection on COVID-19 pathology. In our patient cohort we uniquely recruited an even number of patients that tested positive for HIV and those that were uninfected with the virus. We hypothesised that the influence of the immune system when faced with COVID-19 may differ when already actively engaged with an additional viral infection, an infection that interestingly impacts T-cell numbers[253]. Of the patients we recovered we examined T-cell numbers to investigate this however T-cell numbers were variable irrespective of HIV status. To attempt to observe viral reads from either COVID-19 or HIV in our infected patients we mapped our data to a concatenated genome containing the human, HIV and SARS-CoV-2 genome however we found no evidence of HIV reads in our tissue atlases and very few SARS-CoV-2 reads. This could be a result of the preprocessing of the data and running SoupX[64] on each sample to remove ambient RNA. As the mapping of the data indicated that we had large amounts of ambient RNA in each of our runs, the correction of this effect was deemed mandatory before proceeding with downstream analysis to avoid the negative impact of ambient RNA. However, SARS-CoV-2 has been widely reported to promote cell death in late stage chronic infection though mechanisms such as apoptosis and pyroptosis causing infected cells to burst and contribute to tissue damage[254]. Thus, by correcting for ambient RNA, the algorithm SoupX may have removed ambient viral reads from lysed cell contents indirectly influencing the extent of viral reads in our cohort. In addition to this, we also wanted to investigate the effect of different SARS-CoV-2 variants in our cohort as we recruited patients over the beta and delta variant waves. However, we only had one patient that was infected with the beta variant and thus insufficient sample size impacted power and robustness. Nonetheless, we demonstrate how we can still gain invaluable insights from post-mortem tissue and how we can leverage spatial data to perform a targeted cellular interaction analysis. By focusing our cellular inference on co-localised cell types in the IMC data, we found further evidence that there was a type II interferon response in alveolar macrophages that was absent from northern hemisphere cohorts. Furthermore, we orthogonally validated this interaction using RNAscope that showed an increase of IFNGR2 in our COVID-19 lung samples highlighting the power

of using multiple modalities of data in cellular interaction inference. The evidence for this interferon response is clear in our cohort and absent in the northern hemisphere cohort when we compared our lung data. This could be down to the time frame of the disease where our cohort had a significantly shorter symptom start to death period due to lack of accessible ventilation equipment unlike the northern hemisphere cohorts mainly recruited from hospitals across the US. This shortened window of disease could suggest that the first responders of the viral infection in the lung was tissue-resident cells that were already present in the tissue niche at the site of infection. When the viral infection persists for a longer period of time such as in the northern hemisphere cohort it may allow sufficient time for monocyte recruitment from the blood and monocyte-derived macrophage differentiation to occur. However this would need to be further investigated in time course infection studies and cannot be answered by the results in this chapter alone. Together, this chapter presents a valuable unique dataset elucidating the mechanisms of lethal COVID-19 in a demographic cohort that was not represented during the pandemic. Lastly, our data serves as a single cell resource to be used by the community to investigate inflammatory mechanisms across three tissues, aided by interactive atlases that are readily available in the publication[207].

2.6 Supplemental tables

Cell Type	Covid-19	LRTD	Non-LRTD
Alveolar macrophages	3079	1748	814
Interstitial macrophages	1049	164	191
Monocyte-derived macrophages	1079	875	773
CD16+ Neutrophils	1555	1381	3039
CD16- Neutrophils	1138	262	254
Naive CD4+ T cells	992	873	434
Th1	920	449	447
T reg	551	202	278
CD8+ T cells	1009	423	328
NK cells	616	390	719
B cells	653	137	37
Plasma cells	1251	301	39
Mast cells	145	86	86
Cycling cells	115	40	14
Erythrocytes	241	26	14
Adventitial fibroblasts	2327	346	18
Alveolar fibroblasts	1695	805	121
AT1	3376	3356	199
AT2	1483	891	6
Basal	248	636	17
Ciliated cells	1597	808	358
Lipofibroblast	259	37	3
Lymphatic endothelium	323	107	32
Mesothelial	392	1129	130
Myofibroblast	409	79	7
Ribosomal high cells	2221	1651	584
Secretory cells	734	243	57
Smooth muscle cells	2927	613	70
Venous endothelium	3985	1852	959

Table 2.9: This table contains all cell counts for annotated lung cells included in this study split by disease group.

Cell Type	Covid-19	LRTD	Non-LRTD
Macrophages	44	286	9
Neutrophils	354	724	26
CD4+ T cells	121	78	35
CD8+ T cells	42	62	1
Basal cells	368	517	59
Ciliated cells	84	104	48
Goblet cells	550	1279	213
Secretory cells	603	885	142
SPRR2Dhigh Squamous cells	84	312	16
VEGFAhigh Squamous cells	130	449	22
Neurons	106	85	4

Table 2.10: This table contains all cell counts for annotated nasal cells included in this study split by disease group.

Cell Type	Covid-19	LRTD	Non-LRTD
Monocytes	782	619	47
Neutrophils	3297	917	518
CMP/GMP	184	57	2
CD4+ T cells	137	305	22
CD8+ T cells	888	586	133
NK cells	564	838	101
B cells	401	251	23
Reticulocytes	2303	197	0
Platelets	80	95	3

Table 2.11: This table contains all cell counts for annotated blood cells included in this study split by disease group.

Chapter 3

cellXplore: a web tool to interactively explore cellular interactions at the single cell resolution

3.1 Abstract

Cells communicate through many diverse molecules such as ligands, receptors (L-R), structural proteins and metabolites to coordinate a response across tissues in both homeostasis and disease. With the advent of single cell RNA-sequencing (scRNA-seq), cell-cell interaction (CCI) inference and spatial transcriptomics, many computational tools have been developed to predict interacting cell types. Examination of CCI's has been invaluable within the scope of immunology and cancer in elucidating mechanisms of action within disease however, a key limitation of existing CCI tools is that the output is often large and complex which poses a challenge to correct interpretation requiring bioinformatic analyses. In addition to this, cellular interaction inference leads to many false positives thus leveraging spatial transcriptomics to observe ligand-receptor expression in space can allow us to hone in on biologically meaningful signal. To mitigate these problems, we have

developed a Flask-React web tool cellXplore to facilitate CCI analysis in a user-friendly manner consisting of a web interface with click and point functionality. This provides a shared platform bringing together widely-used existing CCI packages, allowing users to develop customisable analysis pipelines and interpret results with interactive data visualisations. We demonstrate the functionalities of cellXplore using three distinct workflows applied to a Visium and Xenium dataset with matched scRNA-seq data to build a comprehensive view of the interactome in its native context in parasitic infection and in the breast cancer tumour microenvironment.

3.2 Introduction

We demonstrate in the previous chapter that leveraging multi-modal data that provides ligand-receptor expression and a spatial axis can allow us to orthogonally validate inferred interactions. Through inference of cellular interactions from the lung single cell data we could integrate the imaging mass cytometry proteomic data to determine spatially co-localised cell types and project ligand-receptor expression in a shared dimensional space unveiling the critical role cellular interaction inference plays in the elucidation of underlying immunomodulatory mechanisms in disease. The advent of single-cell RNA sequencing (scRNA-seq) has allowed the dissection of cellular heterogeneity in tissues, reflecting the inherent stochasticity and variability of gene expression in each cell. This technology has given rise to the development of many computational tools that utilise manually curated databases and statistical methods to quantitively evaluate the probability of two cell types interacting based on ligand receptor gene expression[144,146,148,150,152]. However, cell signalling is spatially constrained, a pivotal dimension that is not preserved in scRNA-seq data. Interacting cells are usually in close proximity to each other due to limited spatial diffusivity of the expressed ligand, or to achieve activation through physical contact with adjacent cells, thus giving rise to spatial patterns of interaction within tissues. Spatial transcriptomic methods have revolutionised insight into cellular interactions and function

by maintaining the spatial localisation of cells within their native context[185,255,256]. By preserving the organisation of cells, previously lost through tissue dissociation, elucidation of the effects of the tissue microenvironment allows deeper biological insight that cannot be answered through single cell-RNA sequencing alone. Identification of CCIs in spatial data utilise spatial location by filtering out returned co-expressing L-R pairs from existing databases if it is physically impossible for communication to occur[144,146]. Other methods compute probabilistic models of inter-cell variation driving spatially variable genes which is indicative of a CCI occurring therefore reducing the number of false positives in scRNA-seq results[194,196]. Despite the insights developed from these tools there still remains uncertainty surrounding the validity of the results, agreement amongst methods and robustness[139]. In addition to this, there is yet to exist a platform that comprehensively brings together both modalities of data, cellular interaction results to support hypotheses in a complementary manner with no coding background needed. Although cell-cell interaction visualisation platforms exist, namely Intercellar[205], there are none to date that address the fundamental limitation of cellular communications inference from scRNA-seq data; the absence of spatial context. Another tool that has since been developed is ezSingleCell[206] that extends their functionality beyond cellular communications and incorporate spatial technologies in their analysis pipeline. It is distributed as a web-application that does not require installation and as an R shinyApp that can be installed by a user for offline-analysis. However, the tool is inaccessible such as the URL being broken and the GitHub repository containing little to no documentation about how the tool can be implemented. In addition to this, the above tools are written and distributed in R, thus bringing into question their scalability for large scale spatial and atlas-level datasets. Here we present cellXplore an interactive visualisation platform that allows exploratory analysis of CCI localised within its native context with no prior computational skills required. The primary aim of cellXplore is to provide a shared platform streamlining the downstream visualisation of pre-computed cell-cell interactions inferred from scRNA-seq and spatial transcriptomic data. Through allowing the development of customisable interactive data visualisations, we allow the user to interrogate their data and draw biological conclusions thus bridging the gap between biologist and programmer. cellXplore requires a fully pre-processed object containing ligand-receptor information from their single-cell data

that can be generated from supported packages such as CellPhoneDB[144,148], CellChat[146] and NicheNetR[150]. cellXplore allows for filtering and in-depth examination of L-R pairs of interest, visualisation of selected L-R pairs at the transcriptomic level then followed by validation within their spatial context. Lastly, users can export results through multiple visualisation options ensuring reducibility and application for publications. First, we detail two iterations of the tool, the legacy cellXplore built in the cellxgeneVIP[247,250] framework and the current cellXplore built using Vitessce[257] visualisation components. Then we demonstrate three potential analytical workflows in which a user can interrogate their interaction results leveraging information from single cell and spatial data. The first workflow the user can leverage their spatial data by interactively selecting a region of interest that contains harmonised annotations with the single cell. Then, cell types present in the selection will be highlighted in the single cell data and, if they are present, the cellular interaction results will be filtered for interactions between the cell types of interest. We then apply this workflow using a single cell dataset of the murine brain during *Trypanosoma brucei* infection that also contains patient matched Visium data at va rious time points post-infection[258] to investigate microglia cross-talk. The second workflow aims to allow the user to be able to search for the presence of a particular ligand-receptor interaction, validate this in a spatial context, and confirm the expression in the single cell data. This workflow is implemented to validate cellular interactions of interest in the *Trypanosoma brucei* dataset that were previously identified with the first workflow. Lastly, we demonstrate a third workflow in which we wanted the user to select cell types of interest in the single cell, investigate the interacting ligand-receptor pairs, search multiple ligand-receptor pairs in the spatial data and validate again with the single cell gene expression. To give a working example we applied cellXplore to a publicly available Breast Cancer dataset where Xenium was performed with patient matched single cell sequencing data[162] to investigate cellular interactions between epithelial cells and the tumour microenvironment.

3.3 Methods

Here the methods first detail the software architecture of the legacy cellXplore, the first iteration of the tool that was built within the cellxgene framework. The results of this tool can be viewed and compared to the current version of cellXplore in the Results section. Then, the methods proceed to outline the software architecture and functionality of the current working version of cellXplore with a step-by-step tutorial available in more detail here (https://cellxplore-app.readthedocs.io/en/latest/). Finally, any additional preprocessing steps of the exemplar datasets used to demonstrate functionality are outlined such as cellular interaction inference.

3.3.1 Legacy cellXplore architecture overview

The front-end of the legacy cellXplore is built with a WebGL library, a Javascript API that allows interactive 2D and 3D graphics alongside various additional packages such as D3 that allow complex interactivity functionalities. The app is a client-server model that involves a Python-based backend design built for Scanpy[92] to allow single cell computational tasks. cellxgeneVIP[250] is a wrapper over the original cellxgene[247] framework that uses a client-side JavaScript panel plugin that allows user input to be communicated through to the server. Developers can add various custom functions to the plugin without changing the underlying cellxgene source code implementing both Python and R modules. Here we have installed a custom cellxgeneVIP client on a private development server with alterations to the source code to allow for compatibility with cell-cell interaction packages including CellPhoneDB[148] and CellChat[146]. Interactive plotting functionality is implemented using plotly in the Python back-end and its Javascript implementation in the front-end. Circos plotting functionality is implemented using rpy2 to allow inter-

operability across programming languages to allow plotting modules from R packages including ggplot2 and circlize. The GitHub repository to the codebase can be accessed here https://github.com/olympiahardy/cellXplore_v1 with the backend code found in *VIPInterface.py* and the frontend code found in *interface.html*.

Legacy backend packages

Category	Package	Version
Core environment	Python	3.8.5
Sc packages	Scanpy; AnnData; diffxpy	1.6.1; 0.7.4; 0.7.4
Data handling	pandas; NumPy; PyArrow	1.2.1; 1.19.5; 1.0.1
Visualisation	Matplotlib; Seaborn; Plotly	3.3.4; 0.11.1; 4.8.1
R interface	rpy2	3.3.5
R environment	R; Seurat; fgsea; ComplexHeatmap	3.6.3; 3.2.3; Bioconductor;
cellxgene	cellxgene	0.15

Table 3.1: Main backend packages used in the legacy cellXplore.

Legacy frontend packages

Category	Library	Version / Role
Core framework	jQuery; jQuery UI	3.4.1; 1.10.3
Styling	Bootstrap; Font Awesome	3.3.7; 4.7.0
Interactive tables	DataTables core; DataTables Buttons	1.10.20; 1.6.1
Visualisation	d3	3.x
Visualisation	Plotly	5.11.0

Table 3.2: Main frontend packages used in the legacy cellXplore.

3.3.2 cellXplore Software Architecture

The current cellXplore is built as a Flask[259]-React web app where the user can create multiple interactive cellular interaction visualisations plotted using the D3 Javascript library[260]. Currently, cellXplore is hosted on a development server *oh-cxg-dev-mvls.gla.ac.uk* and can be accessed for each dataset at http://oh-cxg-dev-mvls.gla.ac.uk/breastcancer and http://oh-cxg-dev-mvls.gla.ac.uk/braintbrucei or can be launched locally from the GitHub repository (https://github.com/olympiahardy/cellXplore_App), with all dependencies to be installed with a convenient environment.yml file and package.json for the backend and frontend installation respectively.

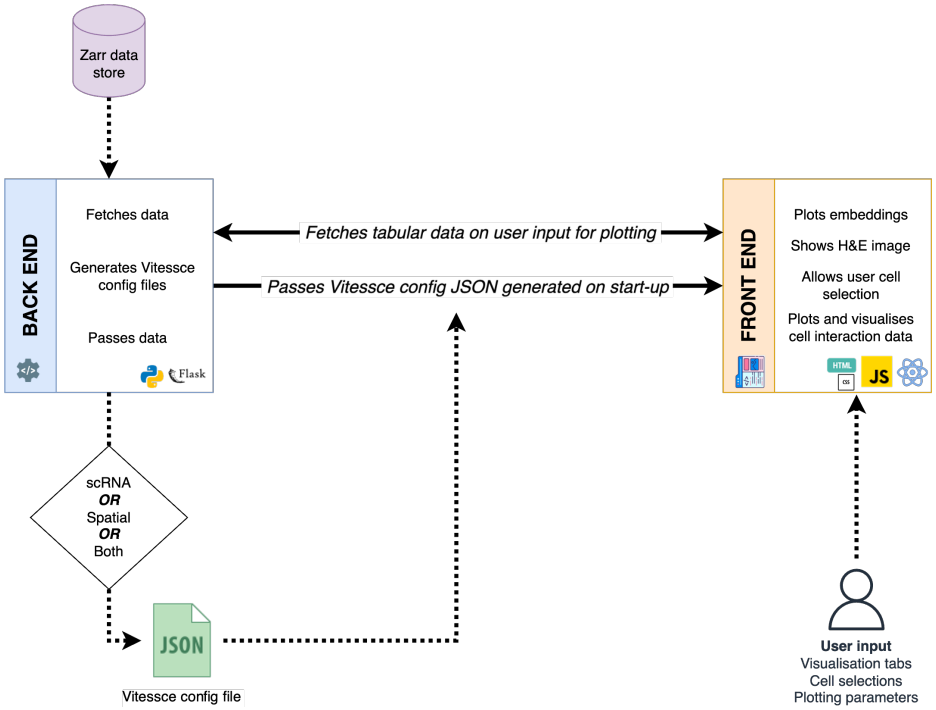


Figure 3.1: Schematic demonstrating the software architecture of cellXplore

As shown in Figure 3.1, the backend architecture follows the Flask framework and consists of a single python file that contains multiple REST endpoints for fetching interaction data. Each endpoint serves JSON payloads for the frontend that function as distinct plotting and visualisation functions that are described in more detail in the Visualisation Tab section.

In addition to this, the backend also contains functions to generate a JSON configuration file that is parsed by the Vitessce visualisation components[257]. The JSON config file contains descriptors to inform the frontend what Javascript components are required and how to access the necessary paths in the Zarr data structure. Additionally, it also instructs how to link different views together and how they should be laid out on the user interface. A detailed example of a dummy JSON config file for Vitessce is explained in more detail in the appendix of this chapter (Appendix 3.1). The frontend architecture adopts a modular, component-based design using React, enabling reusability of data across different visualisation tabs. Each plotting function is encapsulated within its own JavaScript file, facilitating maintainability and further ease of development of the tool. The front end also handles all plotting functionality and filtering logic, thus implementing this on the client side, eliminates the need for asynchronous backend-frontend communication. This way, the tool reduces computational load, resulting in faster response times and a smoother user experience.

Backend packages

Category	Package	Version
Core environment	Python	3.10.15
Web framework	Flask; Flask-CORS; Flask-Caching	3.0.3; 5.0.0; 2.3.0
ASGI server	Uvicorn	0.32.0
Data structures	AnnData; SpatialData	0.10.9; 0.3.0
Data formats	Zarr; OME-Zarr; h5py	2.18.3; 0.9.0; 3.12.1
Data handling	pandas; NumPy; PyArrow; xarray	2.2.3; 1.26.4; 17.0.0; 2024.11.0
sc/Spatial packages	Scanpy; Squidpy	1.10.3; 1.6.1
Cell-cell communication	LIANA+	1.4.0
Visualization	Plotly; Matplotlib; Seaborn; Vitessce	5.24.1; 3.9.2; 0.13.2; 3.4.1
Distributed computing	Dask; Distributed	2024.10.0; 2024.10.0
Networking	Requests; aiohttp	2.32.3; 3.10.10

Table 3.3: Main backend packages used in the current cellXplore.

Frontend package versions

Category	Package	Version
UI framework	React; React DOM	18.3.1
Bundling/server	Vite; @vitejs/plugin-react	5.4.14; 4.3.2
UI components	@mui/material; @mui/system; @mui/styled-engine	5.15.14; 5.15.14; 6.4.0
CSS Styling	@emotion/styled	11.14.0
sc/Spatial visualisation	@vitessce/dev	3.5.11
Interactive plotting	d3;	7.9.0;
UI tab utilities	react-select; react-tabs	5.10.1; 6.1.0 — selects/tabs
Exporting	html2canvas; jspdf	1.4.1; 3.0.1
Static server	http-server	14.1.1

Table 3.4: Main frontend package versions used in the current cellXplore

3.3.3 Development protocol for cellXplore

Development and testing of cellXplore were carried out on a dedicated development server *(oh-cxg-dev.mvls.gla.ac.uk)*, configured with 24 Intel Core CPU cores, 30 GB of RAM, and running Ubuntu 20.04.5 LTS. cellXplore was initiated from scratch using a Vite-based build format to establish the frontend and backend components. During development, the frontend was launched with *npm run dev*, while the backend was executed as a Python file, allowing the interface to be tested via a web browser on a local machine. Code changes were version-controlled with GitHub, organised into two branches: *main* for stable releases and *local_dev* for ongoing development. Once changes were finalised, the repository was pushed to GitHub and subsequently deployed to the development server (oh-cxg-dev.mvls.gla.ac.uk) using *git pull*. Deployment was performed via SSH tunnelling, where the repository was cloned, dependencies were installed with *npm install*, and the frontend was compiled using *npm run build*. The backend was then launched within a screen session, enabling the tool to remain accessible after logout from the server. Finally, to ensure dataset-specific hosting and configuration a separate Git branch was created for each dataset. To achieve this, the GitHub repository was cloned twice on the development server so that each dataset could run from an independent codebase. In addition

to this, each instance of the application was configured to serve on a distinct local port, enabling multiple datasets to be hosted simultaneously. Access to these dataset instances was managed using Nginx as a reverse proxy. For each dataset, a location block was added to the Nginx configuration file, mapping a unique URL path to the respective local port. For example, requests to /braintbrucei/ were forwarded to port 5001, while /breastcancer/ was mapped to port 5000. This setup allowed multiple datasets to be accessed in parallel under the same domain (oh-cxg-dev.mvls.gla.ac.uk) while isolating their backend processes so that multiple screen sessions can be run in parallel without interference.

3.3.4 Input data requirements

The user must have a fully preprocessed single cell object where cell types have been annotated, dimensionality reductions computed and cell-cell interaction analysis inferred saved to a Zarr store.

The table below (Table 3.5) provides an overview of mandatory and optional requirements for cellXplore. The user must have a precomputed single cell object saved to a Zarr store that contains dimensionality reductions, categorical annotations such as cell type labels and cellular interaction results. Spatial data is not mandatory however, where available, the spatial data must contain spatial locations and harmonised cell type labels with the single cell data. In addition to this, the user may also include any image data such as histology slides but are not mandatory for the tool to function.

Input Data	**Requirement**	**Details**
(Mandatory) General Requirements		
Accepted Zarr stores	Mandatory	`AnnData` or `SpatialData`
Dimension reductions	Mandatory	Embeddings stored in `obsm` (e.g., UMAP, PCA)

Input Data	Requirement	Details
Categorical annotations	Mandatory	Cell types, clusters, or metadata stored in `obs`
Cell-cell interactions	Mandatory	Interaction matrix must be stored in `.uns`
(Optional) Spatial Data Requirements		
Spatial coordinates	Mandatory	Location embeddings (e.g., spot or cell centroid) in `obs` or spatial metadata
Spot/cell-level annotations	Mandatory	Region or spatial labels stored in `obs`
Images	Optional	H&E or histology images as OME-TIFF or OME-Zarr
Joint Analysis Requirements (Single Cell + Spatial Data)		
Harmonised annotations	Mandatory	Consistent cell type or cluster field shared across both datasets

Table 3.5: This table contains the mandatory and optional requirements for input data for cellXplore.

3.3.5 Case study dataset preprocessing

3.3.5.1 10X Visium *T.brucei* murine brain infection

All data were publicly available and were taken from this study[258] investigating changes in the murine brain during *Trypanosoma brucei* infection. The raw Visium spatial transcriptomics data and images were downloaded and re-analysed using Giotto[181], a spatial analysis toolbox. First, for each infection time point, a GiottoVisiumObject was created using filtered Visium counts from the SpaceRanger output and the low-resolution histology image. Objects were then filtered for only barcodes present within the tissue region, and spots with gene counts more than 50 and normalised using a scale factor of 6,000. Highly variable features were then calculated and PCA was computed and UMAP embeddings were obtained using the 10 principle components. To obtain clusters, a sNN network was constructed using the *createNearestNetwork* function and was fed into *doLeidenCluster*. To obtain spatial cluster markers *findMarkers_one_vs_all* was performed and the top 10 cluster markers were plotted using *plotMetaDataHeatmap*. Spatial regions were then annotated using publicly available resources such as the Tabula Muris[11] and the Human Protein Cell Atlas[261](`proteinatlas.org`) to check for orthologous gene expression. Once spatial regions had been annotated and cellular deconvolution analysis was performed using the SpatialDWLS[262] algorithm that combines cell type enrichment with the dampened weight least squares algorithm. For this, the fully preprocessed and annotated single cell RDS object was downloaded from Zenodo and contained the following cell types detailed in the paper: *Microglia 1, Microglia 2, Microglia 3, Microglia 4, Astrocyte 1, Astrocyte 2, Pericytes/Tanycytes, Endothelial, Ependymocytes, B cells/Oligo, T cells*. The data was read into R and subset for each time point where *FindAllMarkers* from the Seurat package was ran to obtain signature marker genes for each cell type. Then, using the top 20 marker genes for each cell type a signature matrix was constructed using the *makeSignMatrixDWLSfromMatrix* for the SpatialDWLS algorithm. Deconvolution results were obtained by running *runDWLSDeconv* and were then exported as a CSV file. Spatial coordinates,

reduction embeddings, metadata and normalised counts were then extracted from the Giotto object and exported as CSV files. For cellular inference analysis the preprocessed single cell object was analysed using CellChat (v1) following the comparative analysis protocol. Firstly, the dataset was split based on infection status and converted into a CellChat object was completed separately for each condition. Overexpressed genes were first identified using *identifyOverExpressedGenes* followed by overexpressed ligands using the *identifyOverExpressedInteractions* function. Following this the comunication probability was inferred using *computeCommunProb* and interactions between cell type groups containing less than 10 cells were filtered out. The remaining interactions were then inferred on a pathway level using *computeCommunProbPathway* and the total aggregated cellular interaction network was calculated using *aggregateNet*. Finally, the interaction dataframe results for each object were extracted from the CellChat objects using *subsetCommunication*, with manual addition of a 'condition' column termed 'Infected' or 'Uninfected' to distinguish conditions and exported as a CSV. The preprocessed single cell RDS object was then converted to a H5AD file compatible with Python using the SeuratDisk package functions *SaveH5Seurat* and *Convert*. Once the spatial, single cell, deconvolution and interaction results have been exported, all data were read into Python. An Anndata object was constructed with the single cell data, the spatial data was saved to the 'uns' slot along with the deconvolution and interaction results each assigned to their own layer, 'spatial', 'deconvolution' and 'CellChat_Interactions' respectively. For the assignment of cell types in the Visium data, cellXplore sets a spot threshold cut-off at 60% to determine what the dominant cell-type is at that spot location based on the deconvolution results.

3.3.5.2 10X Xenium and single cell sequencing of Breast Cancer

The fully preprocessed single cell and Xenium datasets were downloaded from the 10X Genomics website (https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast) and contained harmonised annotations of cell types across both datasets. Cellular inference was completed using the LIANA+[155] package on the single cell data using the *liana.method.cellchat* function, returning a results table of interactions that were expressed in at least 10% of cells in the cluster. Cellular interaction pathways were annotated using omnipathDB[147,263] package and added to the results table where available. In the case where ligand-receptor pairs had no functional annotation information they were assigned the label 'Unknown' in the pathway column. Subsequently, the table was saved to the 'uns' layer of the data object. Both datasets were then saved to a Zarr store ready to be used in cellXplore.

3.4 Results

In this section, the functionalities of cellXplore will be demonstrated, using two spatial transcriptomics datasets with paired scRNA-seq data where cellular interactions have been precomputed. First, the legacy cellXplore will be presented showing its existing applications and then the current version of cellXplore will be presented, outlining three distinct workflows a user may wish to take when utilising the tool.

3.4.1 Reanalysis of *T.brucei* dataset

Human African trypanosomiasis (HAT) is a disease caused by the parasite *Trypanosoma brucei* that leads to neurological dysfunction, more commonly known as sleeping sickness[264]. Quintana et al elaborate on the molecular mechanisms of immune cells during neuroinflammation at the parasite-host interface in the murine brain during various stages of infection[258]. This dataset serves as an ideal paradigm for the development of cellXplore as it had sample-matched single cell and Visium data and therefore was selected for analysis. The single cell data were taken from the murine brain at naive, 25, and 45 days post-infection with the cellular annotations reported in the paper (Figure 3.2). The data contains representative homeostatic cell types in the brain such as microglia and astrocytes, including an immune cell compartment of T cells and B cells.

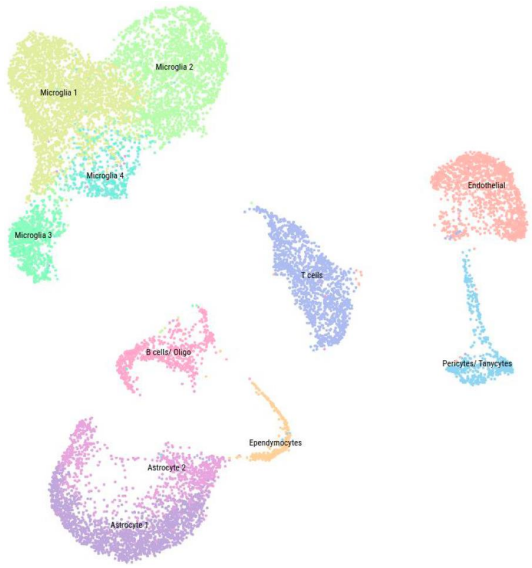


Figure 3.2: UMAP visualisation of annotated cell types presented in the Quintana et al paper

The Visium data in the paper were used to validate interesting findings in the single cell data by projecting gene expression in space. However, to further develop the functionality of cellXplore I wanted to elucidate cell types likely to be present in the various spatial regions of the brain using cell-type deconvolution techniques. When doing this we can

get a more comprehensive understanding of our downstream interaction analysis by not only defining the different anatomical regions of the brain but also what particular cell types may reside there. The spatial data was re-analysed and clustered revealing 7 distinct brain compartments (Figure 3.3) that were annotated using reference datasets detailed in the Methods section of this chapter. Identified regions included the basal ganglia (*Pde1b, Gpr88*), the hippocampus (*Cnih2, Wipf3*), the cerebral (*Nov, Atp2b4*) and upper cerebral cortex (*Clstn1, Egr1*), the thalamus (*Prkcd, Ramp3*), hypothalamus (*Resp18, Nap1l5*), and white matter (*Fth1, Tpt1*). Defined brain regions were further cross-validated with anatomical diagrams of the murine brain, that provided additional confidence that the spatial organisation of the annotated clusters were correct.

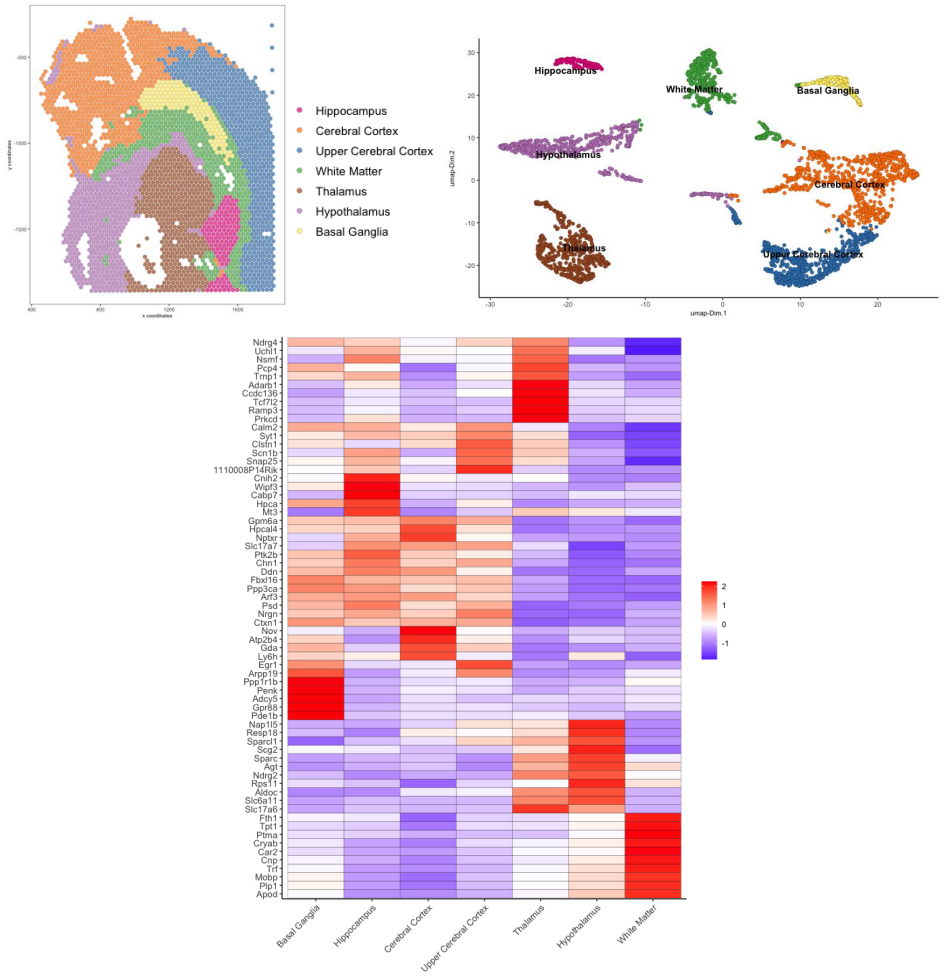


Figure 3.3: Top left) Spatial projection of the naive slide showing annotated regions of the murine brain. Top right) UMAP of the spot clusters annotated with brain regions. Bottom) Heatmap showing the top 10 marker genes for each annotated cluster

Once spatial regions were defined, cell type deconvolution was performed to elucidate what potential cell types were localised to the brain compartments. This step is critical when considering cellular inference with pseudo-bulk spatial data, and a prerequisite for cellXplore when using Visium data. The results of the deconvolution are shown in Figure 3.4 in naive state to demonstrate the estimated cell type composition at homeostasis.

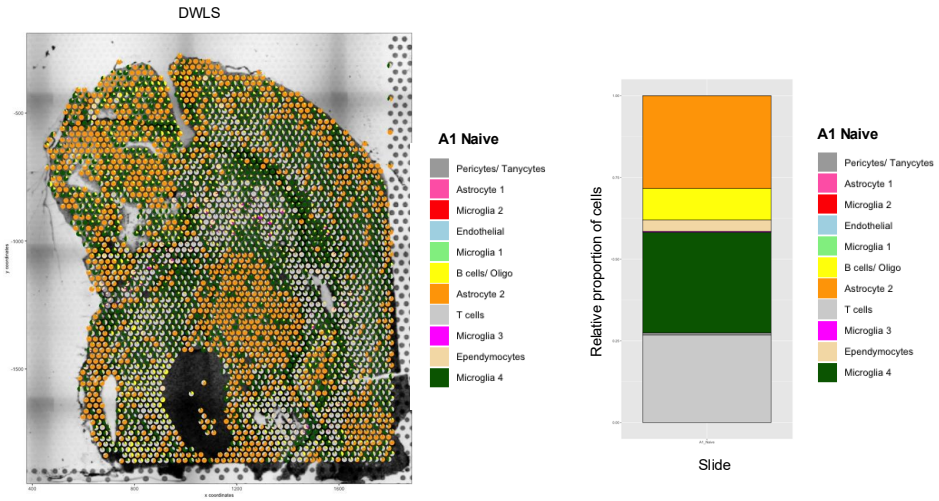


Figure 3.4: Left) Spatial slide showing the cell type deconvolution of the uninfected murine brain. Estimated cell type proportions per spot are visualised as a pie chart. Right) Stacked proportion bar plot showing the overall relative proportion of cell types across the whole slide using a 60% cut-off to determine spot annotation.

We can now see the structural composition of the murine brain on a cellular level with astrocytes concentrated in regions such as as the thalamus whereas we observe a more heterogenous composition of cell types in areas like the cerebral cortex that has co-localisation with various adaptive immune cells and microglia. Thus, cell type deconvolution can give us deeper insights into potential interactors, which cellXplore can utilise to filter out false positives from precomputed cellular interaction results.

3.4.2 Analysis of cellular interactions in active *Trypanosoma brucei* infection using the legacy cellXplore

Implementing the legacy cellXplore, built within the cellxgene framework, we can first see how this workflow can be demonstrated using the *T.brucei* dataset. Once the tool is launched from the command-line with the dataset we can see the homepage with a view of the single cell data in a UMAP representation as shown in Figure 3.5. The data can be coloured by various categorical metadata shown in the sidebar and labels can be toggled for ease of reading using the top toolbar.

Figure 3.5: Screenshot of the T.brucei single cell dataset implemented in the legacy cellXplore

The cellXplore visualisation tab can be accessed by clicking 'Cell Cell Interaction Analysis' located on the top right of the page which opens a window to visualise interaction data. The legacy cellXplore supported cellular inference analysed with two popular packages CellPhoneDB and CellChat, where the user selects either the 'CellPhoneDB Interaction Analysis' or 'CellChat Interaction Analysis' tab (Figure 3.6).

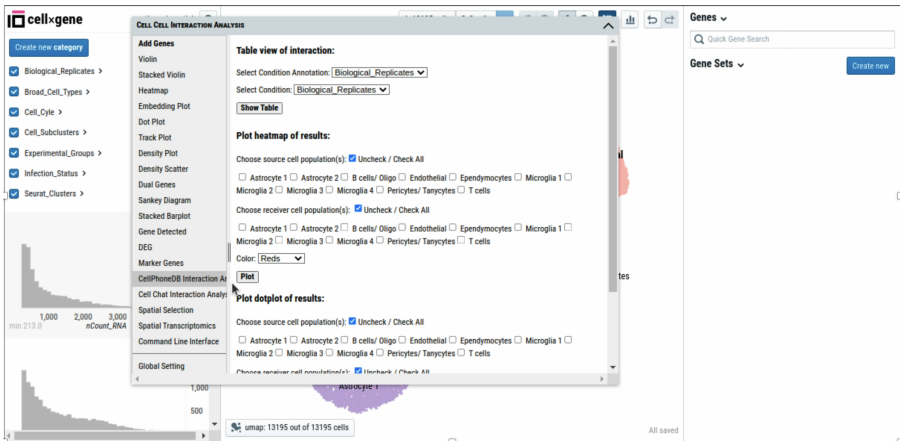


Figure 3.6: Screenshot of the legacy cellXplore visualisation plugin tabs that offer visualisation functionality

When selected, the analysis tab offers different visualisation functionalities such as a table view of interactions for each condition, a heatmap of interacting cell types and a dot plot of ligand-receptor pairs (Figure 3.7). Users can select or deselect cell type populations of interest using tick boxes and change the colour maps of the plots for customisation. Once the user is happy with the selection they can click on the plot button allowing dynamic rendering of results.

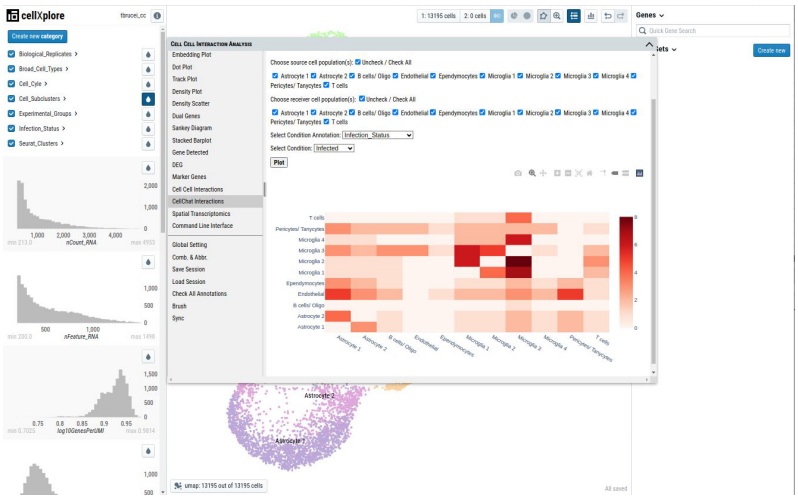


Figure 3.7: Screenshot of the legacy cellXplore visualisation plugin CellChat Interactions tab showing a frequency heatmap of cellular interactions during T.brucei infection

Using cellXplore to visualise cellular interaction results inferred with the CellChat package revealed increase micgroglia cross-talk in *T.brucei* infection compared to naive state (Figure 3.8). Analysis revealed an increase in chemokine signaling during infection and a notable increase in interaction strength between microglia and astrocytes, expressing *Psap* and *Gpr37l1* that plays a part in astrocyte migration and neuroprotective function. Other interactions of interest include *Lgals9-Ighm* between microglia and B-cells suggesting a role that microglia regulate B cell signaling during infection.

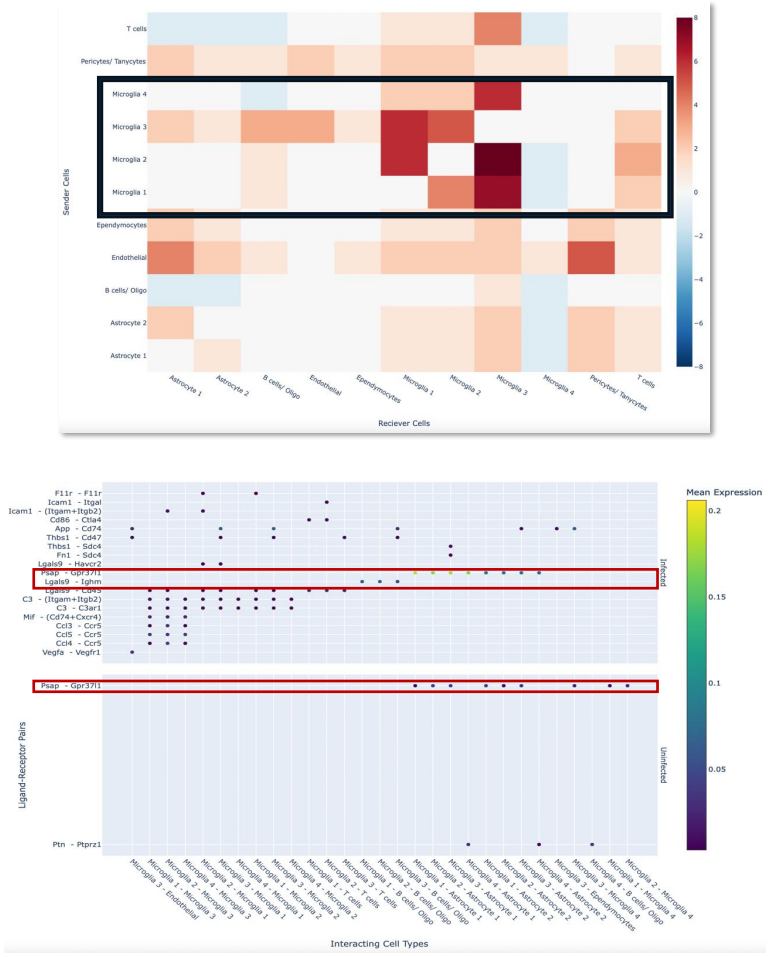


Figure 3.8: Top) Heatmap showing the differential number of interacting cell types in naive state versus infected. The black box highlights the increased cross talk across microglia subsets. Bottom) Dot plot showing interactions between microglia subsets and other cell types in the data in naive versus uninfected. Red boxes highlight increased interaction between microglia and astrocytes and also an infection specific interaction between microglia and B cells

The original paper describes microglia-plasma cell crosstalk during infection, so utilising cellXplore to investigate this, we found that there are different patterns of interactions between various microglia subsets, a result not fully elucidated in the paper (Figure 3.9). In particular, we can see that Microglia 3 is more pro-inflammatory interacting with B-cells through *Spp1* and *Pecam1*, a phenotype absent from the other subsets. We can also recapitulate interactions reported in the paper using cellXplore, such as *Il10* hypothesised to regulate pro-inflammatory responses in microglia.

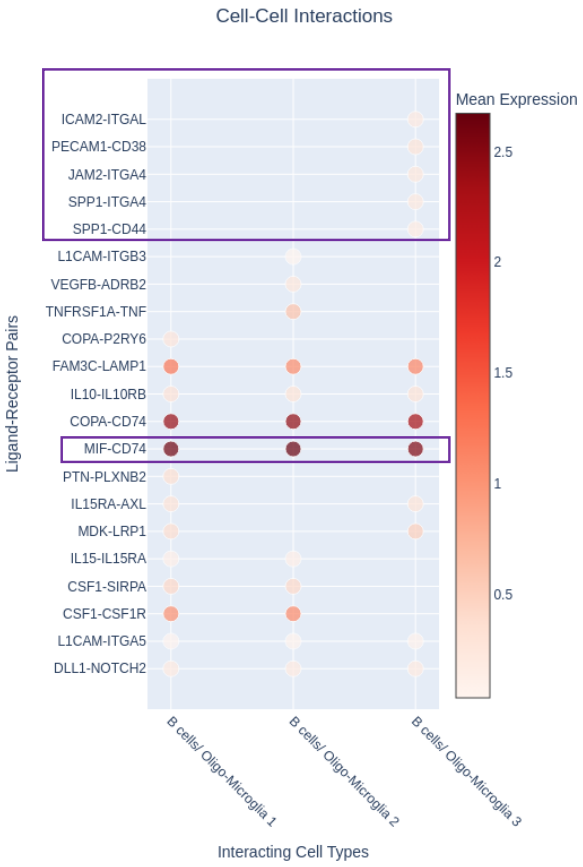


Figure 3.9: Dotplot showing interactions between B-cells and various microglia subsets during infection. Purple boxes highlight microglia subset specific interactions and shared interactions across all subsets.

Another interaction of interest was a high expression of *Mif* and *Cd74* between B-cells and all microglia subsets. We then used cellXplore to visualise the co-expression of this ligand-receptor in the single cell data. We found that in infection both the expression of *Mif* and *Cd74* are increased and localised to both microglia and B-cells (Figure 3.10).

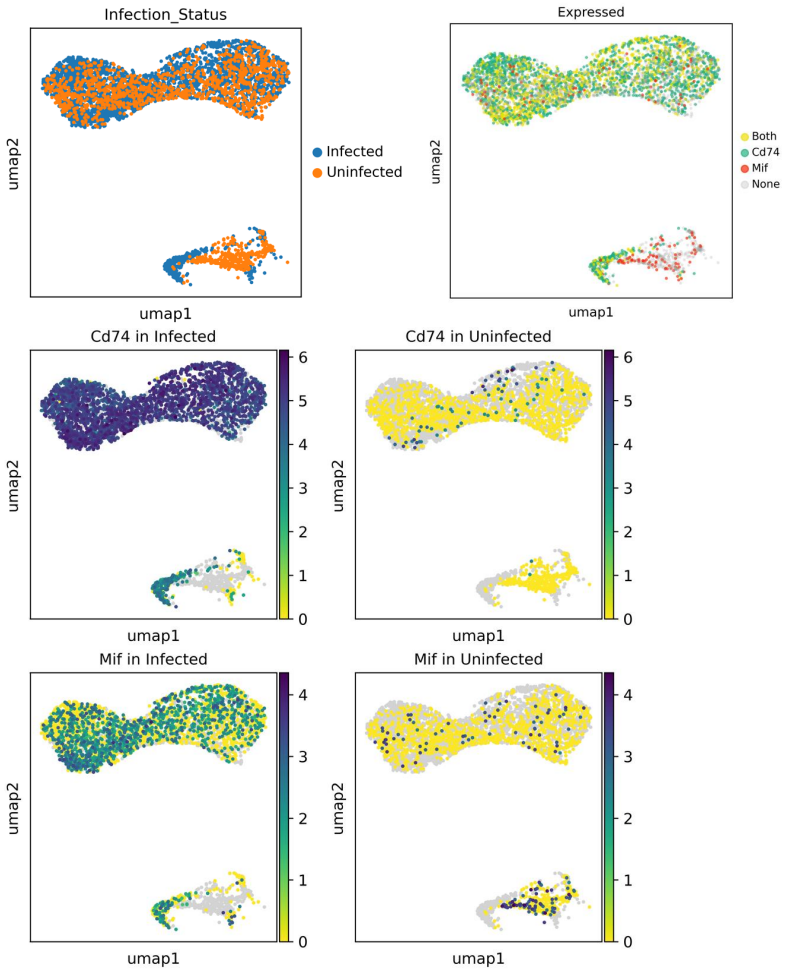


Figure 3.10: Top left) UMAP showing cells coloured by their infection status. Top right) UMAP showing cells coloured by expression of Mif, Cd74, or both. Left panels) UMAPs showing Cd74 and 'Mif in infected cells. Right panels) UMAPs showing Cd74 and Mif in uninfected cells.

By applying cellXplore on this dataset we can visualise reported interactions and discover potential interactions that may not have been previously reported. To further validate our findings from the single cell analysis we can also leverage spatial data and identify co-localised cell types. By navigating to the 'Spatial Selection' tab and selecting the relevant slide of interest, we can generate an interactive spatial plot of the spot coordinates of the Visium data coloured by the region annotations (Figure 3.11).

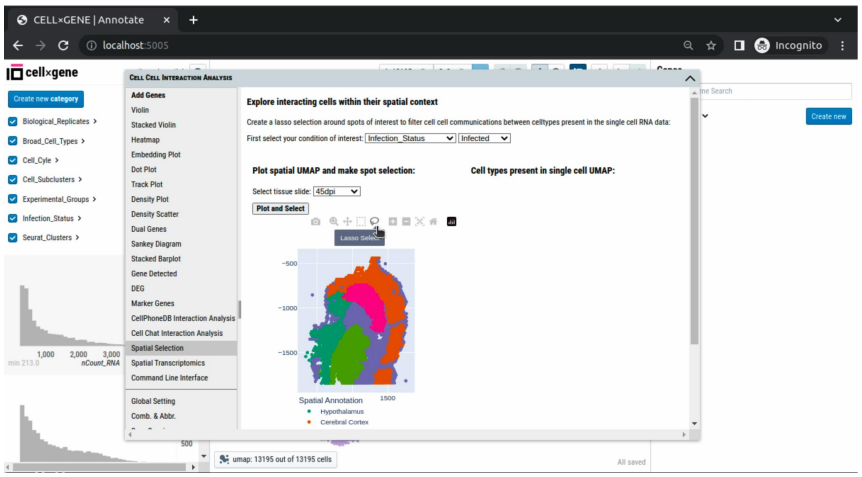


Figure 3.11: Screenshot of the legacy cellXplore visualisation plugin Spatial Selection tab that offers lasso selection interactivity

The plot also contains a lasso tool function that allows the user to select a region of interest in the spatial data. When selected, a UMAP of the single cell data containing the cell types of interest is dynamically plotted and shown next to the Visium data. If the user wishes to create another selection both plots will be updated allowing flexibility of exploration (Figure 3.12).

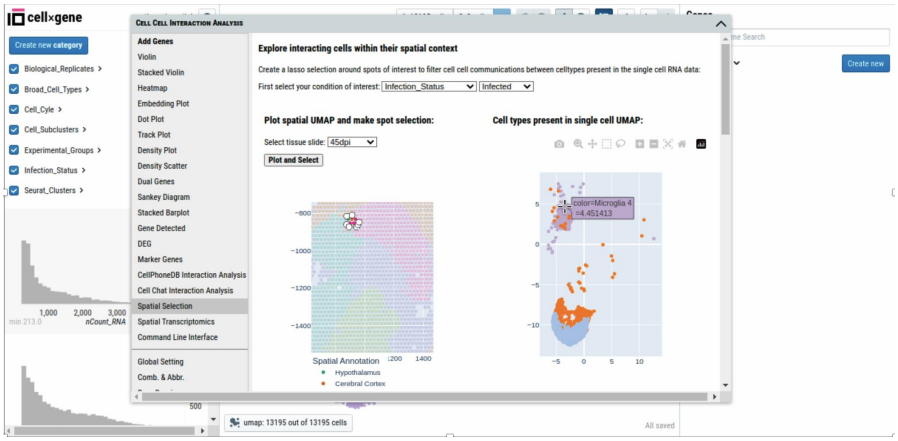


Figure 3.12: Screenshot of the legacy cellXplore visualisation plugin Spatial Selection tab that shows cell types present in the spatial region of interest projected onto their UMAP coordinates in the single cell data

We can see that in the area we selected at the boundary of the hypothalamus and basal ganglia the cell type labels predicted by the cellular deconvolution indicate that microglia and astrocytes are co-localised during infection. The circos plot shows the *Psap-Gpr37l1* interaction we identified earlier in the analysis, giving us more confidence this interaction is likely to occur at both the scRNA-seq transcriptomic level, but also between cell types that are indeed co-localising in space. (Figure 3.13).

Figure 3.13: Screenshot of the legacy cellXplore visualisation plugin Spatial Selection tab that shows a circos plot of interactions that are present between the selected cell types of interest

Below the circos plot we can visualise the interactions between the cell types of interest in a searchable table (Figure 3.14). This gives us quantitative information of potential cellular interactions such as the p-value and communication probability. The table can also be exported to a CSV to be used for further use by the user.

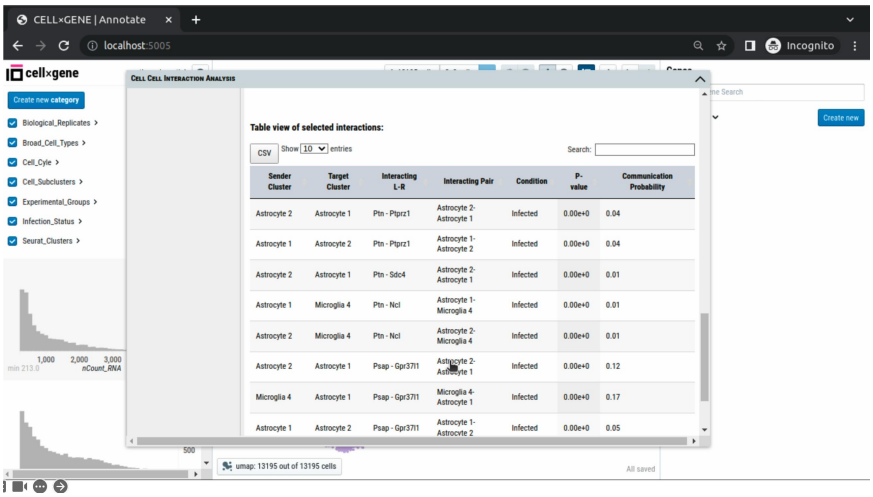


Figure 3.14: Screenshot of the legacy cellXplore visualisation plugin Spatial Selection tab that shows a table view of interactions that are present between the selected cell types of interest

3.4.3 Moving away from legacy cellXplore to the current cellXplore

The legacy cellXplore nicely demonstrates how a user can utilise their spatial data by selecting a region of interest and visualising cellular interactions inferred from paired single cell data. However, this implementation comes with some downfalls that have been addressed with the new implementation of cellXplore.

3.4.3.1 cellXplore improves computational efficiency and data loading by implementing Zarr storage formats

Firstly, the legacy cellXplore faced major issues with large memory datasets being read in as a H5AD file. When the tool was tested with a 10X Xenium dataset that contains large memory images the computational cost negatively impacts the lasso selection resulting in major lag. Thus, we decided to move away from working with H5AD input data formats

that the legacy cellXplore required to a Zarr storage format. This format allows for efficient, scalable storage of large multidimensional arrays by reading data in chunks handling datasets that are larger than memory thus optimising it for out-of-core computation. By allowing this, cellXplore only loads relevant portions of the dataset avoiding complete loading of the data in memory increasing computational efficiency and speed. In addition to this, tabular metadata is stored in a uniform structure allowing for fast structured querying, an example of the directory structure is shown for Anndata (Figure 3.15) and SpatialData respectively (Figure 3.16). Finally, the Zarr storage conveniently integrates with the Python single cell and spatial ecosystems such as Scanpy, Squidpy, Anndata and SpatialData. Commonly used single cell data structures such as AnnData, optimised for tabular data, can be saved to a Zarr store. Additionally, cellXplore is compatible with SpatialData formats that are more tailored to the storage of image based spatial technologies such as Xenium and Cosmx, these can be saved to a Zarr store using wrapper functions provided in Scanpy or SpatialData-io.

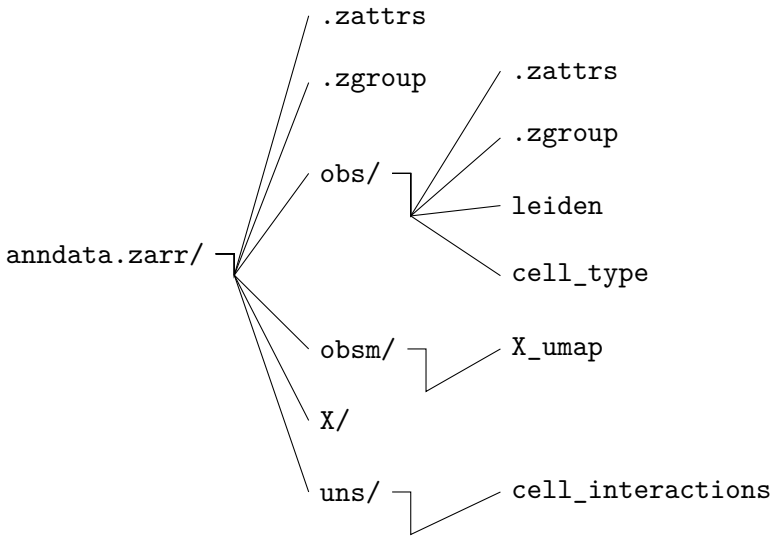


Figure 3.15: Example Zarr directory structure of a `AnnData` object used for single cell transcriptomics visualisation.

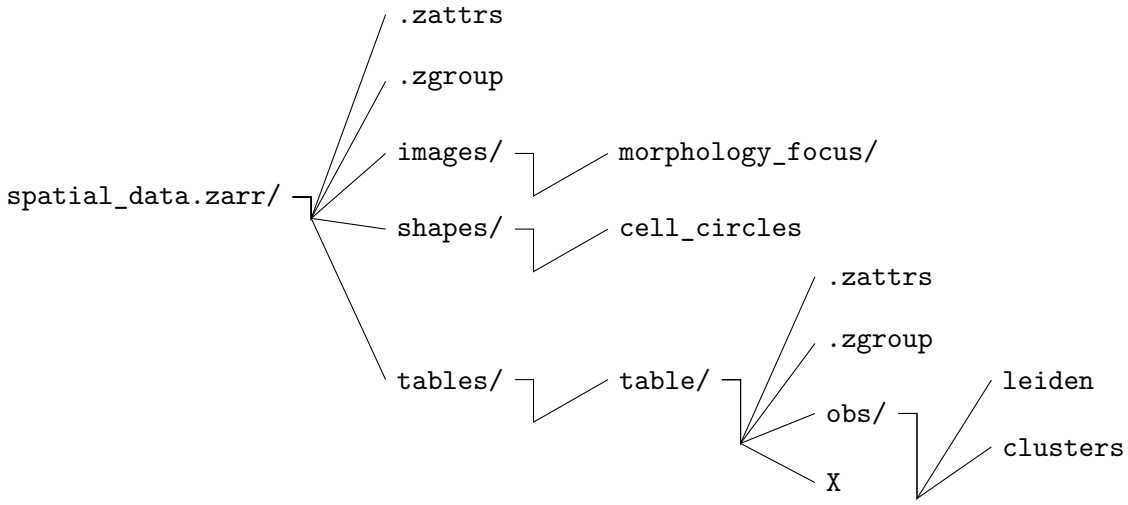


Figure 3.16: Example Zarr directory structure of a `SpatialData` object used for spatial transcriptomics visualisation.

3.4.3.2 Improvements to design UI of cellXplore allows ease of exploration across both data modalities

Another limitation of the legacy cellXplore were the visualisations being limited to the size of the cellxgeneVIP plugin, which was unsuitable for comfortably viewing high dimensional spatial data. To improve this, the new version of cellXplore has a 'Single Cell View' tab that allows the user to explore and visualise their single cell or spatial data using the whole webpage. Using visualisation components from the Vitessce library you can visualise both modalities of data side by side (Figure 3.17), without having to navigate back and forth between the views. The single cell data is projected in the UMAP embedding space using the Scatter Plot component and the spatial data is projected as a spot polygon mask overlaid on the histological image where available using the Spatial component. The Cell Sets component allows the user to visualise their data with any available categorical metadata such as Leiden clusters or annotated cell types in both their single cell and spatial data. Individual or multiple categories of cells can be highlighted by either clicking the label text or by clicking toggle buttons next to each label. Similarly the same may also be done for the spatial view using the Spot Sets component that is located adjacent to the spatial data viewer.

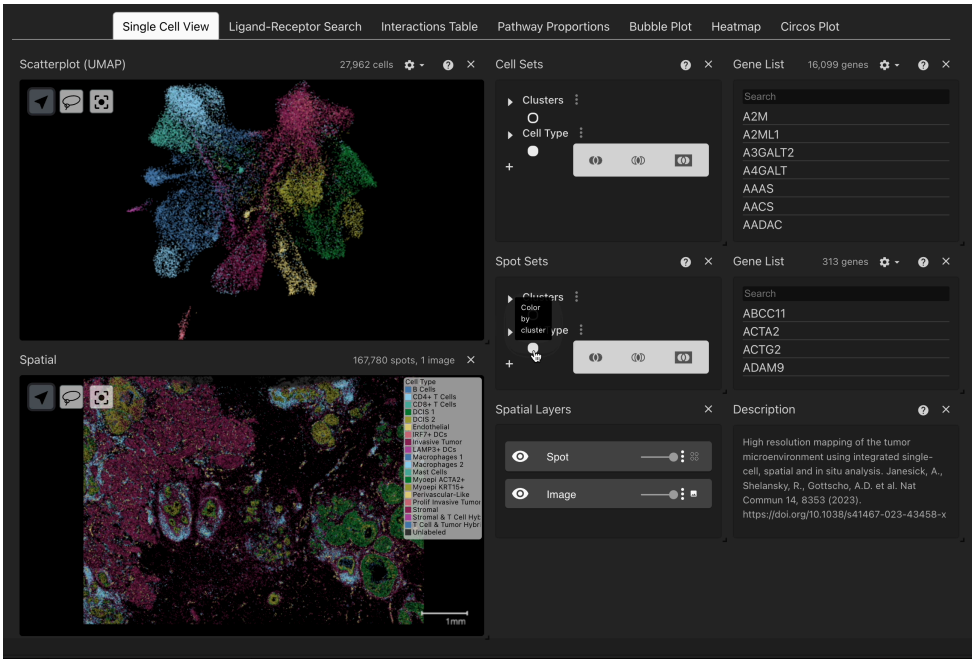


Figure 3.17: Screenshot showing single cell data coloured by metadata in the Single Cell View tab of cellXplore

In addition to visualising categorical metadata we can also visualise gene expression in both the single cell and spatial views using the Gene List component which contains a searchable list of all genes available in the expression matrix of the datasets, a functionality not provided by the legacy cellXplore. By clicking a gene the UMAP component will highlight the genes expression across all cells in the dataset (Figure 3.18). Genes can be ordered by their appearance in the matrix or alphabetically, and where alternative IDs are available, the user can toggle whether to sort by or display these instead of the original gene IDs. Similarly, the same can also be done in the spatial data using the Gene List component adjacent to the spatial data.

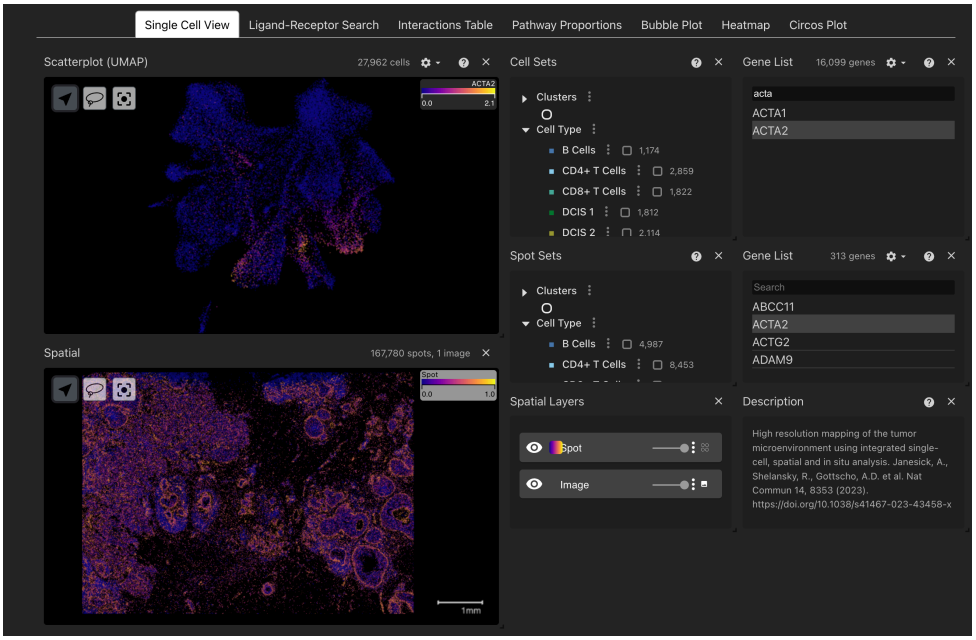


Figure 3.18: Screenshot showing single cell data coloured by ACTA2 gene expression in the Single Cell View tab of cellXplore

The Spatial View component allows users to explore their spatial data on the spot or image level where available. As mentioned above you can colour your spatial data with metadata variables or by gene expression. Additionally, the Spatial Layers component provides user control of showing/hiding the spot level mask or histological image as well as sliders to adjust the opacity of each layer (Figure 3.19).

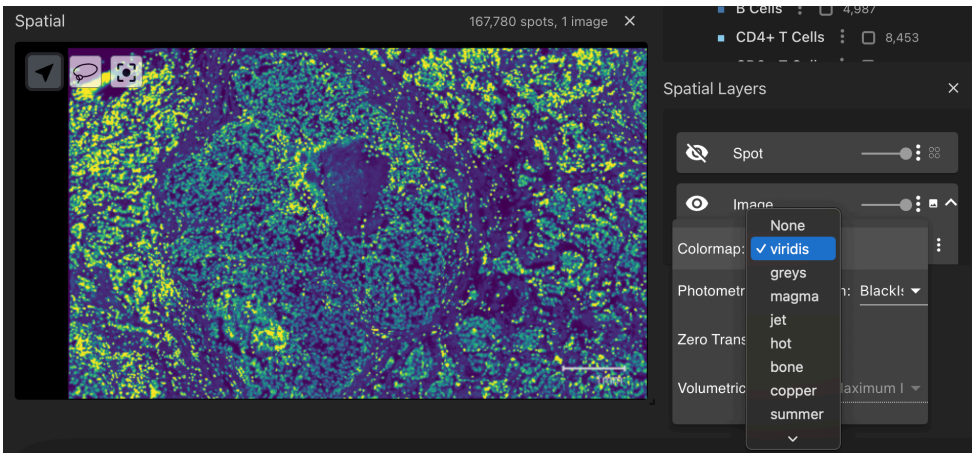


Figure 3.19: Screenshot showing spatial cell data coloured by intensity in the Single Cell View tab of cellXplore

The current cellXplore also provides functionality at the image level where available by allowing the user to utilise the image rather than the spatial embeddings only as implemented in the legacy cellXplore, giving more control over the spatial image settings. By default, the viewer initially shows both the spots and image with the image contrast set to 0 so that the spots are visualised more clearly. The user can control the intensity of the histology image to show staining in various colours or colour maps, simultaneously retaining control over the spot-level polygons so you can see staining and cells overlaid together (Figure 3.20).

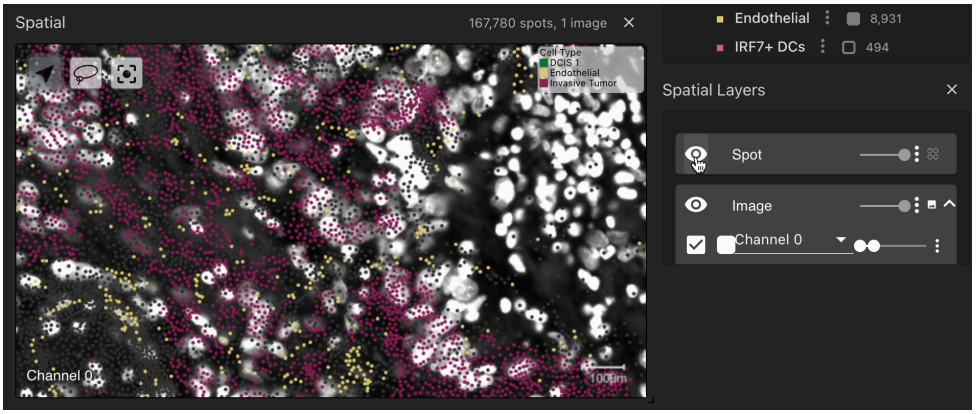


Figure 3.20: Screenshot showing spatial cell data with image and spot mask overlaid coloured by cell type in the Single Cell View tab of cellXplore

3.4.3.3 cellXplore provides advanced data selection and tabular filtering strategies

Filtering strategies in the legacy cellXplore were limited to either metadata categories using tickboxes or spatial selections that refreshes upon each iteration. To improve the flexibility and reproducability of analysis, the current cellXplore gives the user more control over complex cell selections and utilises single cell, spatial data or tabular interaction data. By using the Cell Sets component in the 'Single Cell View' tab we can create selections which can then be passed to the plotting tabs for visualisation. This can be done in two different ways, the first using the toggle buttons of the categorical metadata to select

cell populations of interest. The second is using a convenient lasso tool where the user can click and drag a selection of cells which can be saved and passed to other plotting tabs (Figure 3.21). Once cells are highlighted selections will be stored under 'My Selections' that will appear in the Cell Sets component.

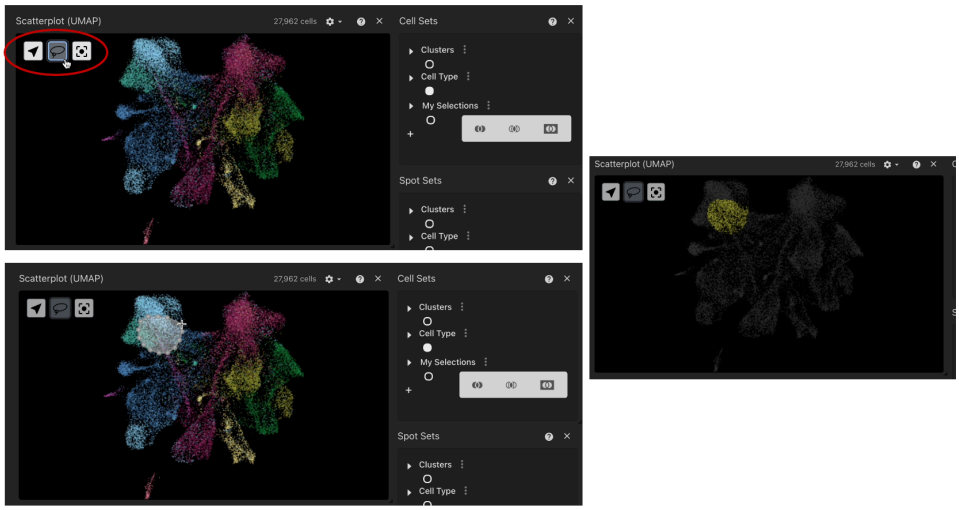


Figure 3.21: Screenshot showing lasso selection in the Single Cell View tab of cellXplore

Both selection strategies can be further manipulated using join functions in three distinct ways (Figure 3.22). The first is a union that takes the union of all selected variables, the second is the difference where cell populations outside of the selection are stored to a new selection, and the last being the intersection which can take the intersection of cell populations in multiple selections.

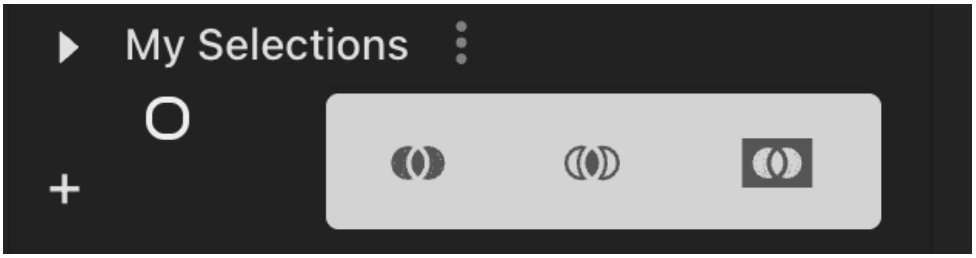


Figure 3.22: Screenshot showing the three different join operations in the Single Cell View tab of cellXplore

Selections of interest can also be renamed by the user and exported as a CSV or JSON which contains the single cell barcodes and the metadata associated with them for further use and reproducibility. Tabular cellular interaction selections can be made in the 'Interactive Table' view where the user can explore their cellular interaction data in an easy table format that has many filtering options to find cellular interactions of interest. Both string and numerical columns can be used to sort the table, with search functionalities and an interactive paginator to control the number of rows displayed. The table view offers a wide range of powerful filtering strategies located above the table to delve into your interaction data (Figure 3.23). These can be performed in two flavours. The first filtering strategy can be utilised by selecting multi-value manual filtering where unique values in the string columns are parsed into drop-down menus. This allows the user to filter on string columns agnostic to the cell-cell interaction package used during the data pre-processing.

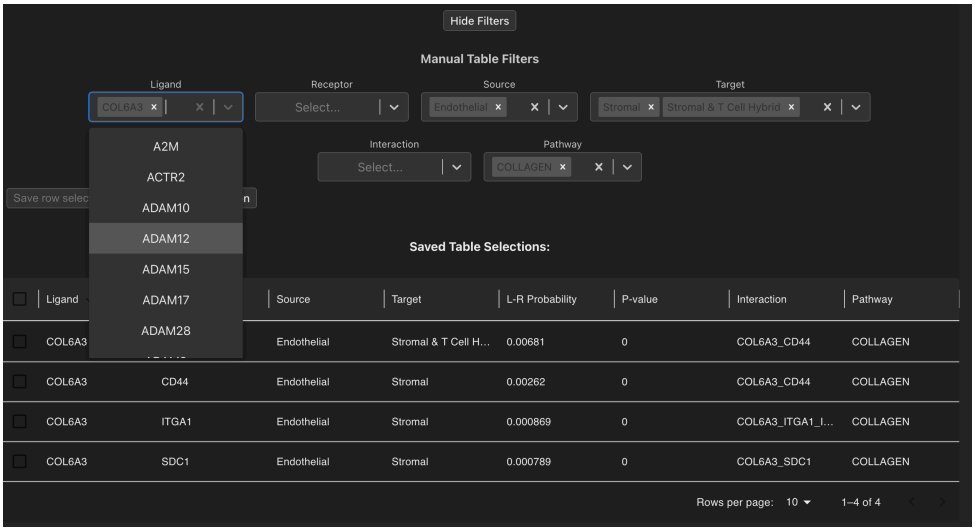


Figure 3.23: Screenshot showing the filtering functionality in the Interaction Table View tab of cellXplore

The second is implementing selections created from the Single Cell View tab that are stored and passed to the Interaction Table View and can be accessed via a convenient drop-down menu (Figure 3.24). When selected, the table will filter and only interactions that arise between the selected cell types of interest.

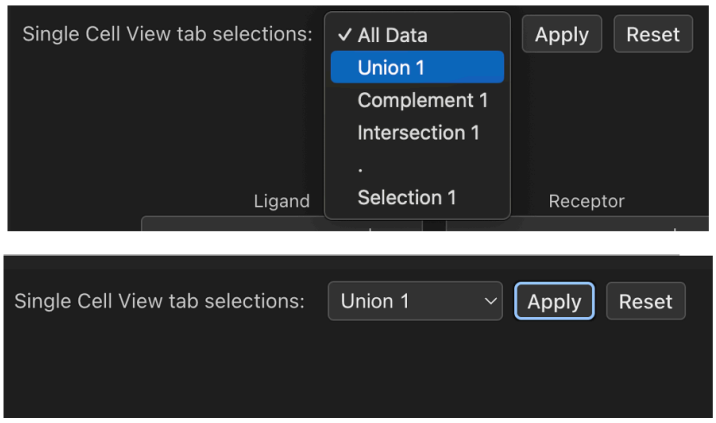


Figure 3.24: Screenshot showing the stored selections made in the Single Cell View tab that can be implemented in the Interaction Table View tab of cellXplore

Individual rows can also be selected in addition to filtering strategies and can be saved and used across additional plotting tabs or exported as a CSV file for further use (Figure 3.25).

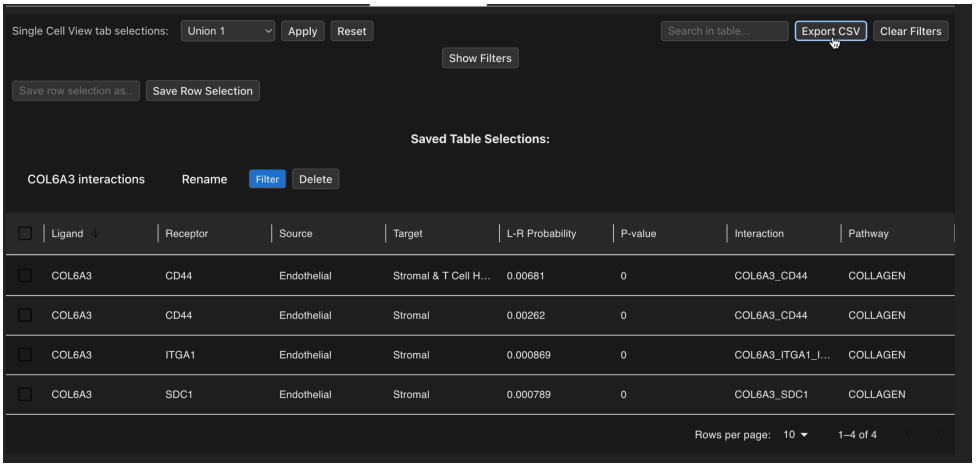


Figure 3.25: Screenshot showing saved selections made by selecting rows in the Interaction Table View tab of cellXplore

Thus, the functionality of saving and setting selections is vastly improved in the current cellXplore, and is not limited to a single selection made in the 'Spatial Selection' tab or static table filtering strategies used in the legacy cellXplore.

3.4.3.4 cellXplore extends interactive plotting functionalities tailored to cellular interaction visualisation

Finally, the cellxgeneVIP framework contained existing plotting functionality that was created for the purpose of differential gene analysis and was inapplicable to cellular interaction visualisation. Therefore, the current cellXplore removed these functionalities and extended plotting visualisations beyond the legacy cellXplore. Each plotting function is interactive and accessible in its own contained tab where selections can be passed in to visualise interactions, which can then be exported in a PDF format for high-quality figure generation. These include a bubble plot, heatmap, circos plot similar to the legacy cellXplore however the current implementation has two additional visualisation views. For example, various cellular inference packages, such as CellChat, provide functional pathway annotations that inform the signalling pathway into which the interaction is fed. The 'Pathway Proportions' tab allows the user to understand what key functional processes occur in our cellular interaction data. The column names of the interaction dataframe are parsed into a drop-down menu where the user can first select the column denoting the pathway information and a grouping variable such as condition or sample to plot a stacked proportion barplot. This shows the relative frequency of the top pathways split by groups, particularly useful for complex datasets that contain multi-condition or multi-sample interaction data (Figure 3.26).

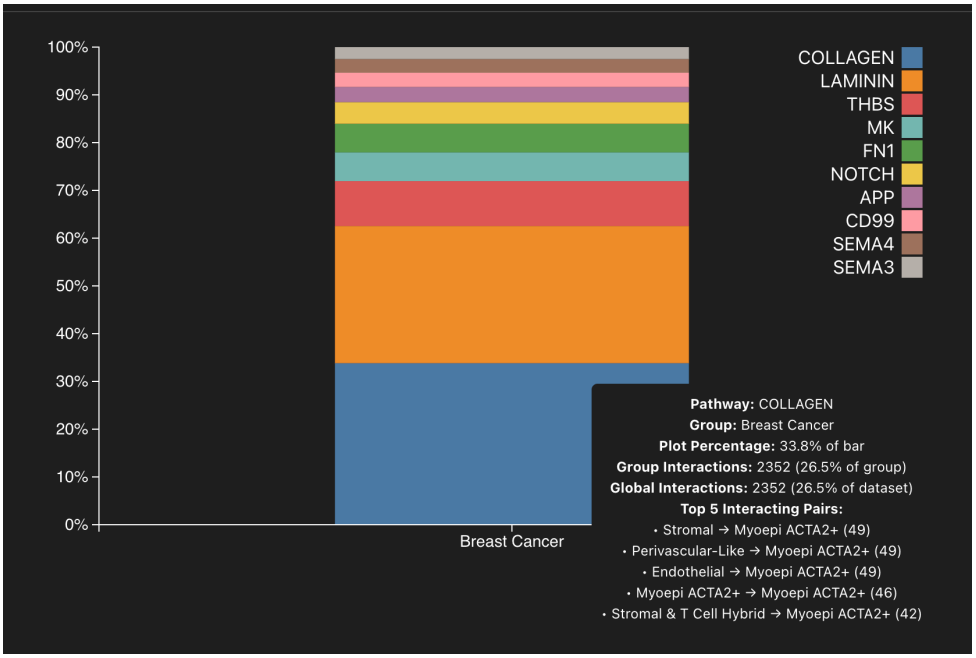


Figure 3.26: Screenshot showing the hover tooltip in the Pathway Proportions View tab of cellXplore

Another new visualisation functionality is the 'Ligand-Receptor Search' tab which contains a dual spatial view that allows users to explore ligands and receptors in their spatial context where available (Figure 3.27). Users can compare two Spatial View components side-by-side, select different genes in each view using the Gene List components and check for spatial co-localisation of ligands and/or receptors. For ease of usability when an area is hovered over in one spatial view then the same location is tracked in the second spatial view highlighted by a white line tooltip.

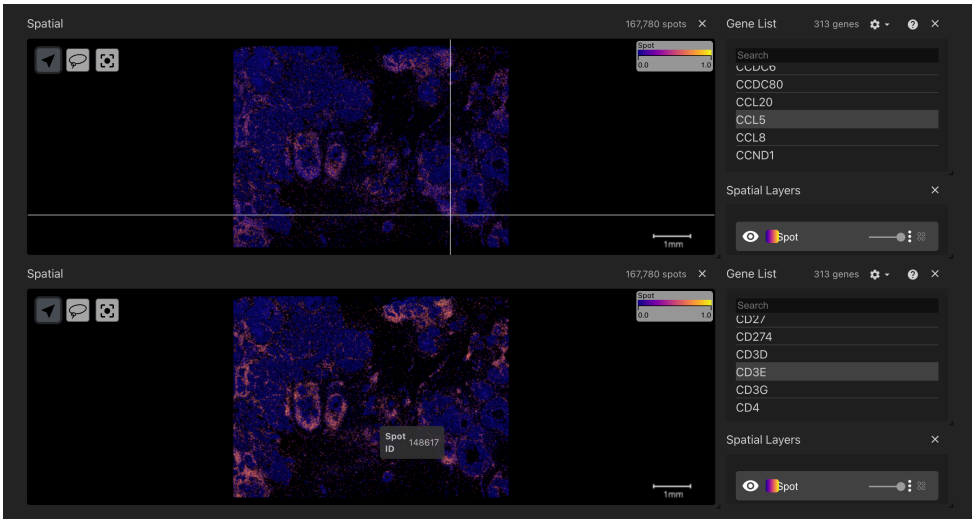
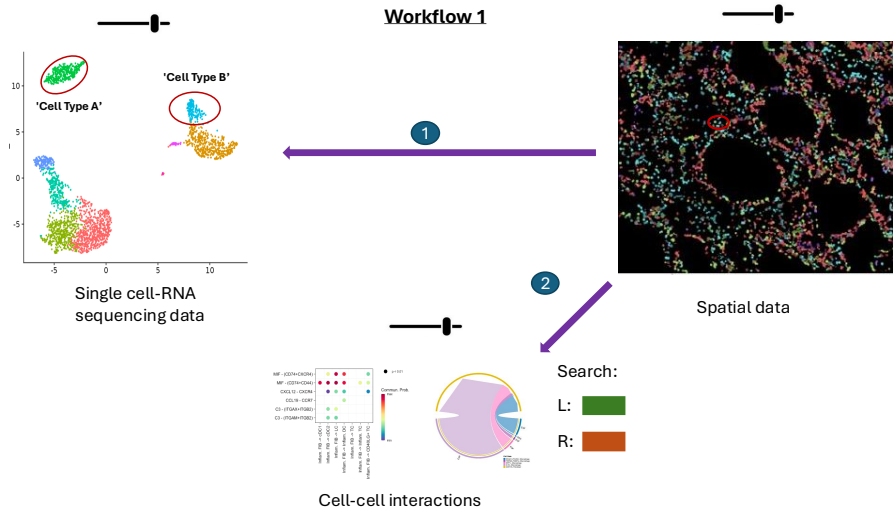


Figure 3.27: Screenshot showing the dual spatial gene expression visualisation in the Ligand-Receptor Search View tab of cellXplore

A more detailed explanation of each plotting visualisation can be found within the cellXplore documentation (https://cellxplore-app.readthedocs.io/en/latest/) and are demonstrated in the workflows outlined below.

3.4.4 Case study: Workflow 1 using *T.brucei* infection to examine microglia-plasma cell cross talk

In workflow one, we wanted the user to be able to input their spatial, single cell, and interaction data and leverage spatial context to filter the interaction table (Figure 3.28). In this scenario, the user can first select a region of interest in the spatial data that contains harmonised annotations with the single cell. Then, cell types present in the selection will be highlighted in the single cell data and, if they are present, the cellular interaction results will be filtered for interactions between the cell types of interest.



Algorithm 1 Spatial-Transcriptomic Interaction Query

Input: Spatial data ST, scRNA-seq data *sc*, interaction table *CCI*
SELECT region $\boldsymbol{r}_1 \in$ ST
if $\boldsymbol{c}_R, \boldsymbol{c}_S \in sc$ **then**
    SELECT $\boldsymbol{c}_R, \boldsymbol{c}_S$ in *sc*
    for each $(\boldsymbol{c}_R, \boldsymbol{c}_S)$ in *sc* **do**
        if $CCI_{\text{IntP}} = (\boldsymbol{c}_R, \boldsymbol{c}_S)$ or $(\boldsymbol{c}_S, \boldsymbol{c}_R)$ **then** and $\boldsymbol{xy}_{\text{dist}} <$ dist
            RETURN matching interactions from *CCI*
        end if
    end for
else
    Throw Error: "No interactions found"
end if

Symbol Definitions

ST Spatial transcriptomics data
sc Single-cell RNA-seq data
CCI Cell-cell interaction table
LR Ligand-receptor pair
$\boldsymbol{r}$ Region of interest
$\boldsymbol{c}_R$ Receiver cell
$\boldsymbol{c}_S$ Sender cell
$\boldsymbol{xy}_{\text{dist}}$ Distance of captured area
IntP Interacting pair ID

Figure 3.28: Overview of user workflow one and analysis query pseudo-code

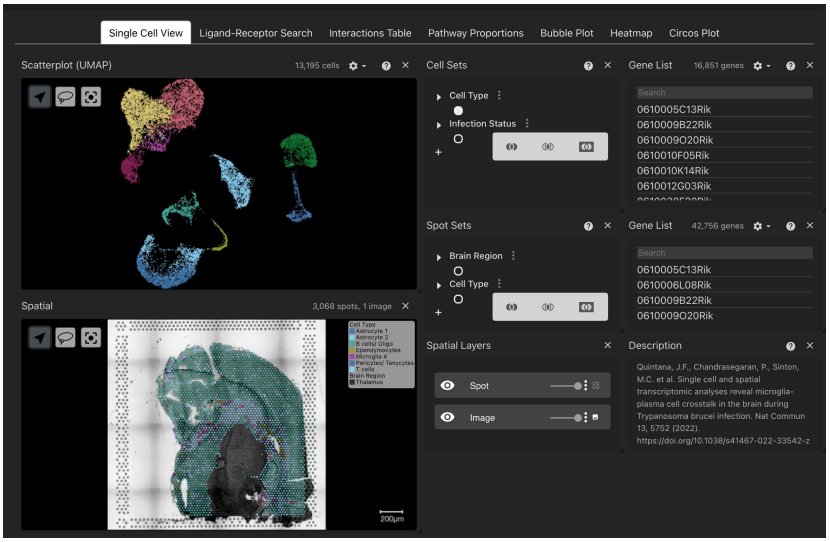


Figure 3.29: Screenshot of the current cellXplore visualising the T.brucei murine brain dataset. Single cell data is coloured by cell type in the top right component. Spatial Visium data is coloured by deconvolution assigned cell types and brain regions.

This workflow was thoroughly demonstrated in the legacy cellXplore and this can be repeated as shown above, the same *T.brucei* dataset is visualised in the current iteration of the tool with a more accessible view to visualise both the single cell and spatial data in one view. Users can colour the data from any existing categorical metadata and also gene expression similar to the legacy cellXplore. Where cellular deconvolution has been computed, the user can leverage this information to assign cell types to their spatial data. In Figure 3.29 we can colour data by multiple metadata categories as shown in the Visium slide. The thalamus region of the brain is highlighted in grey, whereas we can colour other spots outside this region by their assigned cell types. We can see that the white matter is the most heterogenous region abundant in immune cells like microglia and T cells. Using the current implementation of cellXplore we can repeat the analysis of workflow 1 examining microglia and plasma crosstalk, leveraging spatial regions to hone in on cellular interactions of interest demonstrated in the legacy cellXplore. Furthermore, in the current version of cellXplore we can extend our analysis by evaluating the difference in cellular interaction pathways where the data is available (Figure 3.30).

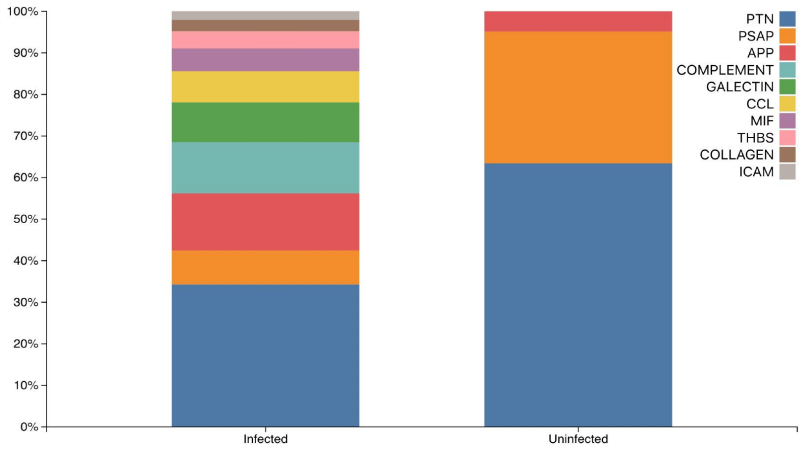
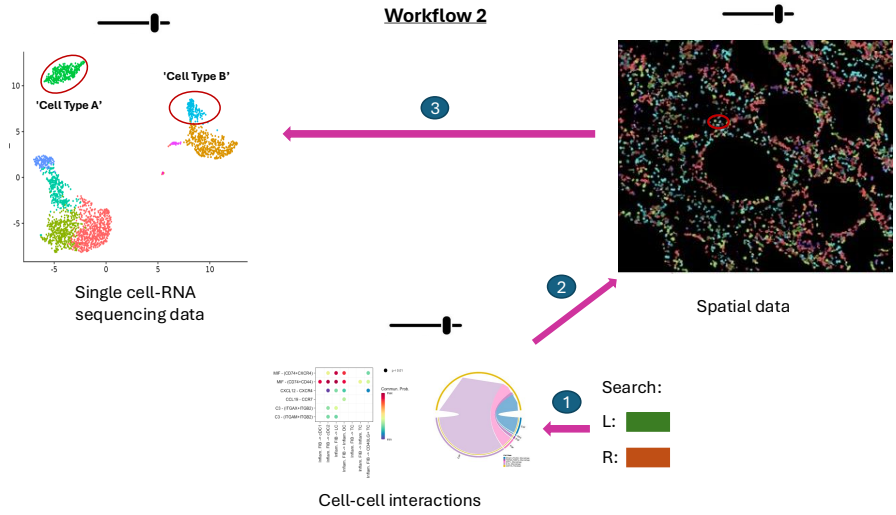


Figure 3.30: Stacked proportion bar plot showing the different signaling pathways involved during naive uninfected state and T.brucei murine brain infection.

This gives us a nice global view of how the interactions are changing across conditions or any grouping variable in our data. Cellular inference tools, such as CellChat, can provide pathway annotation information that can help guide the user to focus on a particular signaling pathway. For example, we observe that during infection the complement and chemokine signaling pathways are activated, as expected during inflammation but are absent in the naive brain. In addition to this, we see the PSAP pathway is substantially decreased in active expression compared to naive. Applying these visualisation strategies, we will now demonstrate how we can complete analysis on the same dataset using workflow 2, to confirm additional interactions between additional cell types we found using workflow 1.

3.4.5 Case study: Workflow 2 confirming interactions between astrocytes and microglia in *T.brucei* infection



Algorithm 2 Ligand-Receptor Interaction Query

SELECT ligand-receptor pair $LR_{(i..j)}$
if $LR_{(i..j)} \in CCI$ **then**
  for each $LR_{(i..j)}$ in CCI **do**
    RETURN CCI where $CCI_L = L_{(i..j)}$ **and** $CCI_R = R_{(i..j)}$
    if $CCI_{\text{sub}} = \text{True}$ **then**
      SELECT $\boldsymbol{c}_R$, $\boldsymbol{c}_S$ in **ST** where $\boldsymbol{xy}_{\text{dist}} < \text{dist}$
      RETURN $LR^2_{(i..j)}$
      for each $LR^2_{(i..j)}$ in CCI **do**
        RETURN CCI where $CCI_L = L_{(i..j)}$ **and** $CCI_R = R_{(i..j)}$
        for each $\boldsymbol{c}_R$, $\boldsymbol{c}_S$ in CCI^2_{sub} **do**
          SELECT $\boldsymbol{c}_R$, $\boldsymbol{c}_S$ in sc
        end for
      end for
    end if
  end for
else
  Throw Error: "No interactions found"
end if

Symbol Definitions

ST Spatial transcriptomics data
sc Single-cell RNA-seq data
CCI Cell-cell interaction table
LR Ligand-receptor pair
$\boldsymbol{r}$ Region of interest
$\boldsymbol{c}_R$ Receiver cell
$\boldsymbol{c}_S$ Sender cell
$\boldsymbol{xy}_{\text{dist}}$ Distance of captured area
IntP Interacting pair ID

Figure 3.31: Overview of user workflow two and analysis query pseudo-code

In workflow two, we wanted the user to be able to search for the presence of a particular ligand-receptor interaction, validate this in a spatial context, and confirm the expression in the single cell data. Previously in our workflow one analysis, we identified a *Psap-Gpr37l1* interaction that occurred between microglia and astrocytes in the brain under both uninfected and infected conditions (Figure 3.13). We can utilise the 'Interactions Table' view to search in our cellular interaction results any interactions that occur between *Psap* and *Gpr37l1* (Figure 3.32).

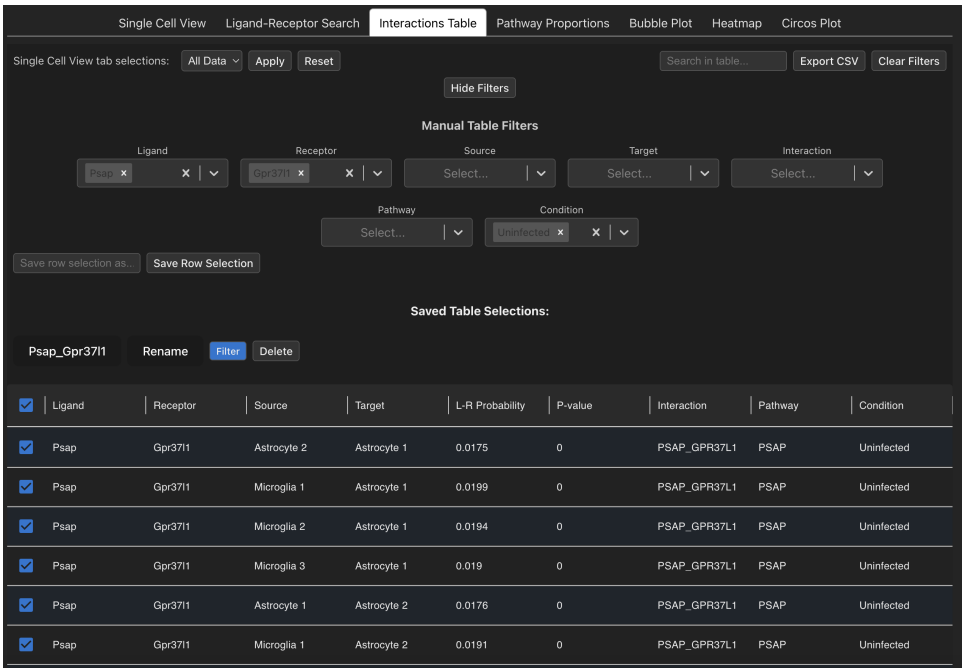


Figure 3.32: Screenshot of the 'Interaction Table' view showing an example selection of Psap-Gpr37l1 interactions in the uninfected murine brain.

As in this example we loaded the naive Visium slide to visualise, we can also exclude interactions occuring between this ligand-receptor pair in the infected conditions. Once we have applied the necessary filtering strategies, we can save the relevant interactions to pass to the plotting tabs to visualise here named 'Psap_Gpr37l1'. We have now established that the interaction of interest exists in the cellular interaction table, but we need to validate the *Psap-Gpr37l1* interaction in its spatial context. We can navigate to the 'Ligand-Receptor Search' tab and search for the spatial expression of each ligand and receptor in space (Figure 3.33).

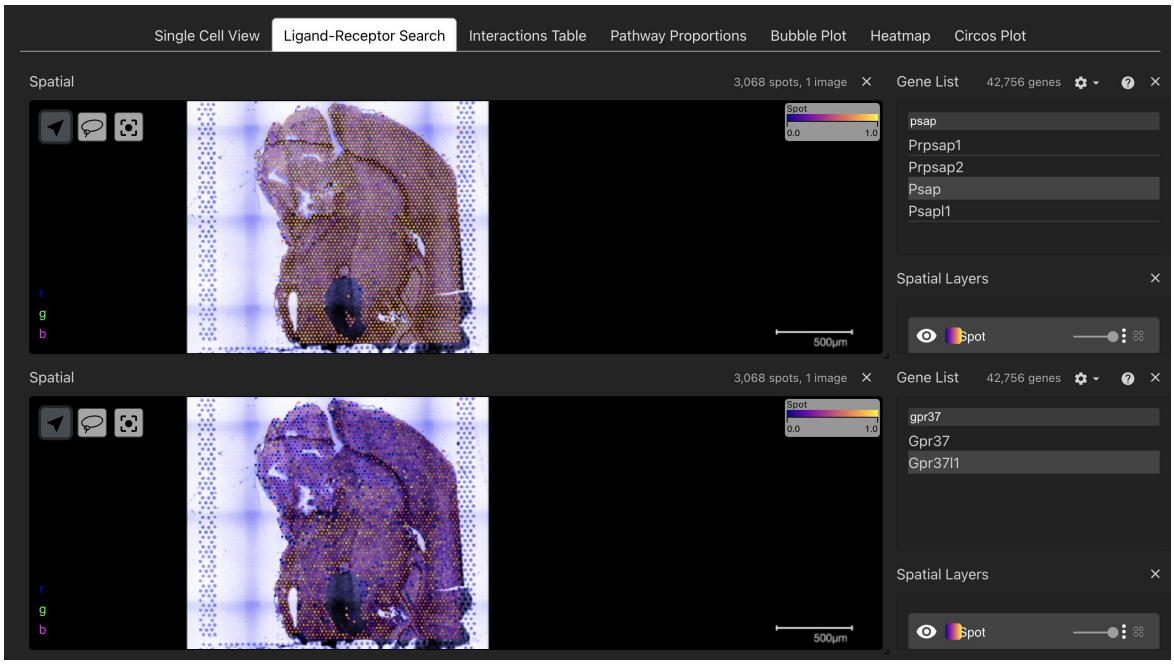


Figure 3.33: Screenshot of the 'Ligand-Receptor Search' view showing spatial expression of Psap (top) and Gpr37l1 (bottom)

From the spatial plots, we can see that *Psap* is ubiquitously expressed across the tissue whereas *Gpr37l1* expression is dimmed in the regions of the basal ganglia and the cerebral cortex. However, if we zoom into the regions of the brain where *Gpr37l1* is highly expressed, we can clearly see co-localisation of *Psap* expression (Figure 3.34). Using the interactive hover tip and zoom functionality we can hone in on spots of interest where a ligand and receptor are both being expressed, increasing our confidence that this ligand-receptor interaction is a true positive.

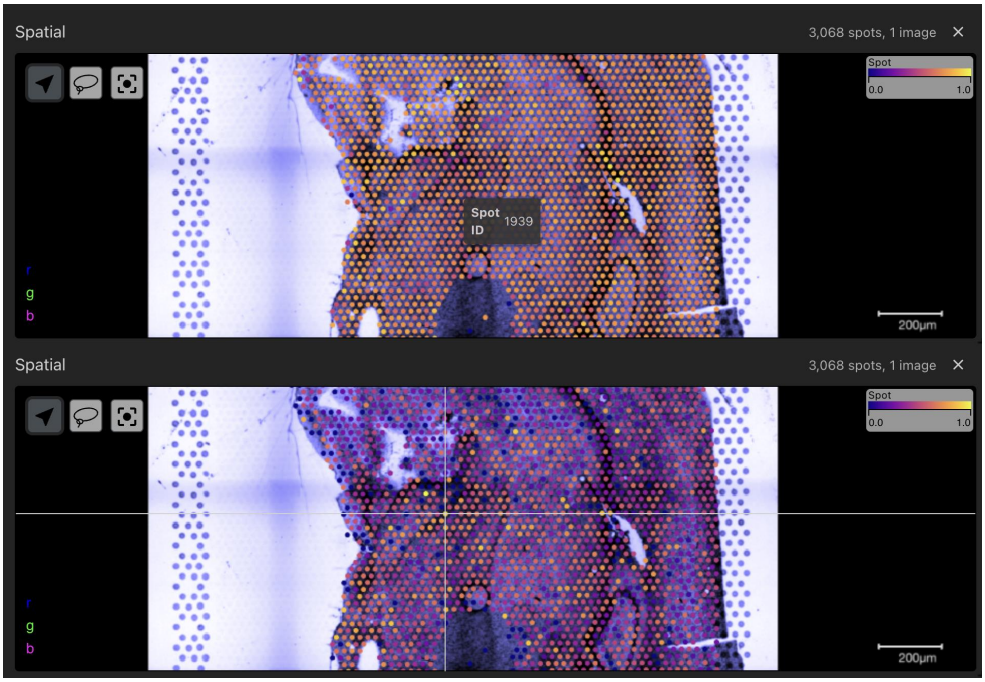


Figure 3.34: Zoomed screenshot of the 'Ligand-Receptor Search' view showing spatial expression of Psap (top) and Gpr37l1 (bottom) aided by the hover tooltip to identify spots/regions of interest

Now that we are confident the expression of the ligand and receptor co-localise we need to ensure that the sender and receiver cell types in the spatial data also co-localise and are present in the single cell data. For this we can select the sender and receiver cell types in both datasets and observe if they harmonise with the gene expression patterns (Figure 3.35). We can see from the localisation of cell types that astrocyte populations co-localise with the microglia 4 subtype in the same regions that correlate to the spatial distribution of ligand receptor gene expression. In areas of the brain like the cerebral cortex and basal ganglia we see less localisation of these cell types and see a more concentrated distribution in the lower regions of the brain slide.

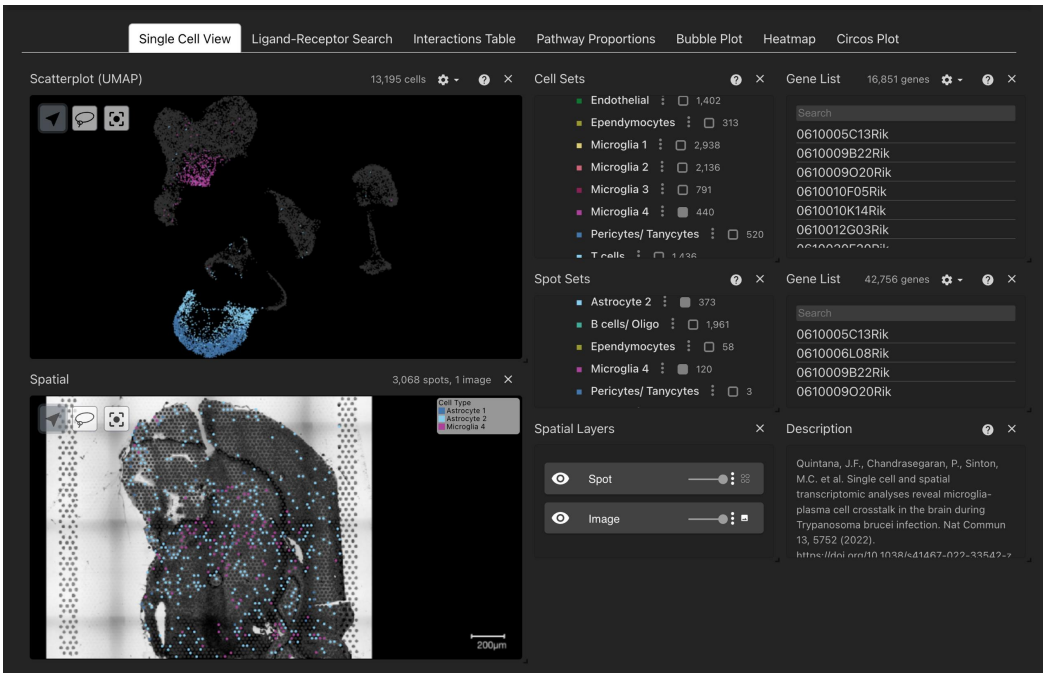


Figure 3.35: Screenshot of the 'Single Cell View' tab showing selected cell types of interest. Microglia and astrocytes are selected in the single cell view (top) and also co-localisation of microglia and astrocytes in the spatial view (bottom)

Finally, we want to conclude the validation of this interaction by checking the expression of *Psap* and *Gpr37l1* in the single cell data. By searching for the expression of these genes in the 'Single Cell View' tab we can nicely see that *Psap* expression is high in our microglia subsets and *Gpr37l1* expression is high in our astrocyte populations (Figure 3.36).

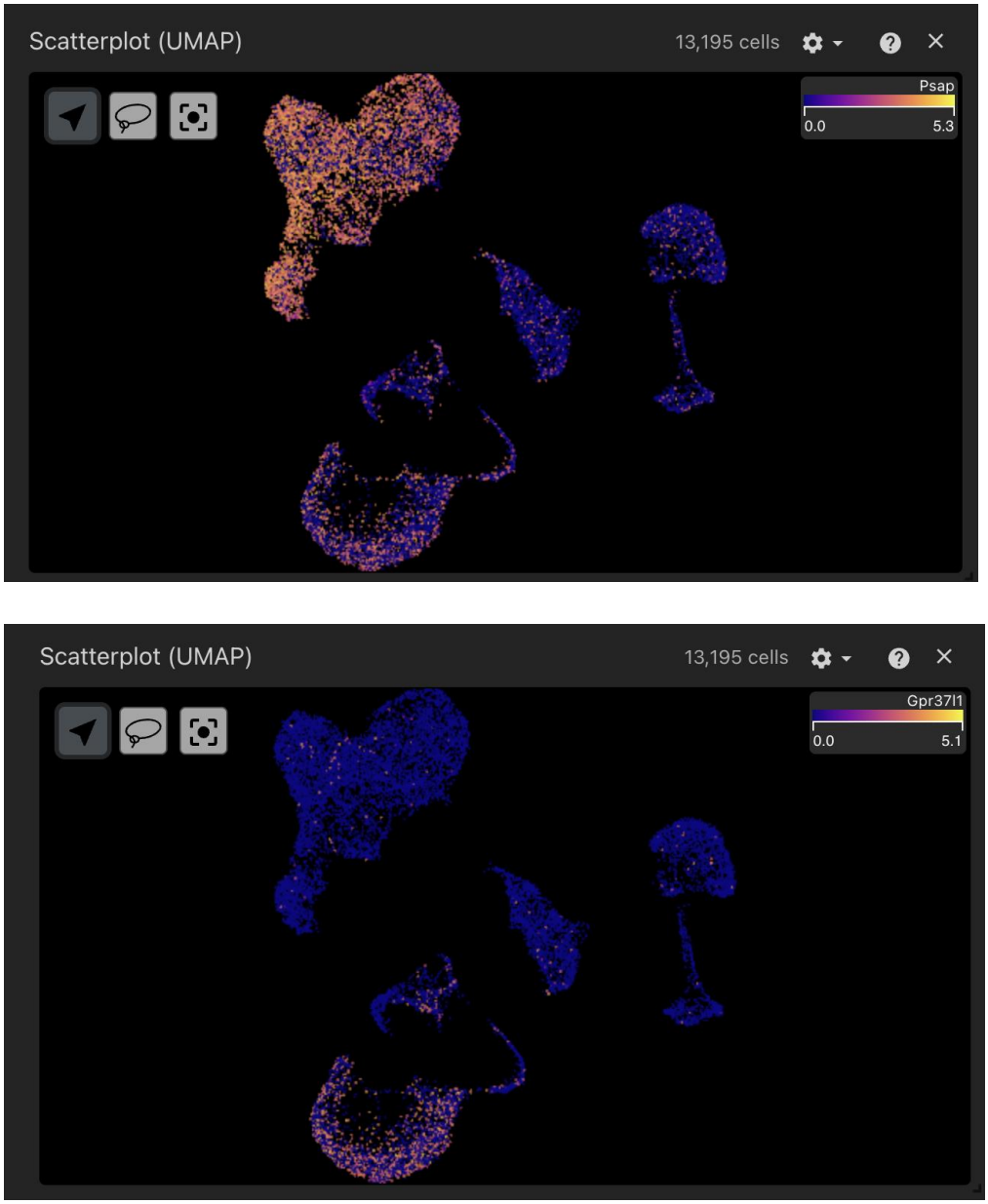


Figure 3.36: Screenshot of the 'Single Cell View' tab scatter plots showing gene expression of Psap (top) and Gpr37l1 (bottom) in the single cell data

Thus, implementing workflow 2 the user can dynamically search for a given ligand-receptor pair, visualise its spatial expression distribution and validate its co-localisation using single cell and deconvolution data. Additionally, we can pass the row selection of *Psap-Gpr37l1* interactions in our uninfected condition and plot the interactions in a frequency heatmap, dot plot and circos plot as shown in Figure 3.37.

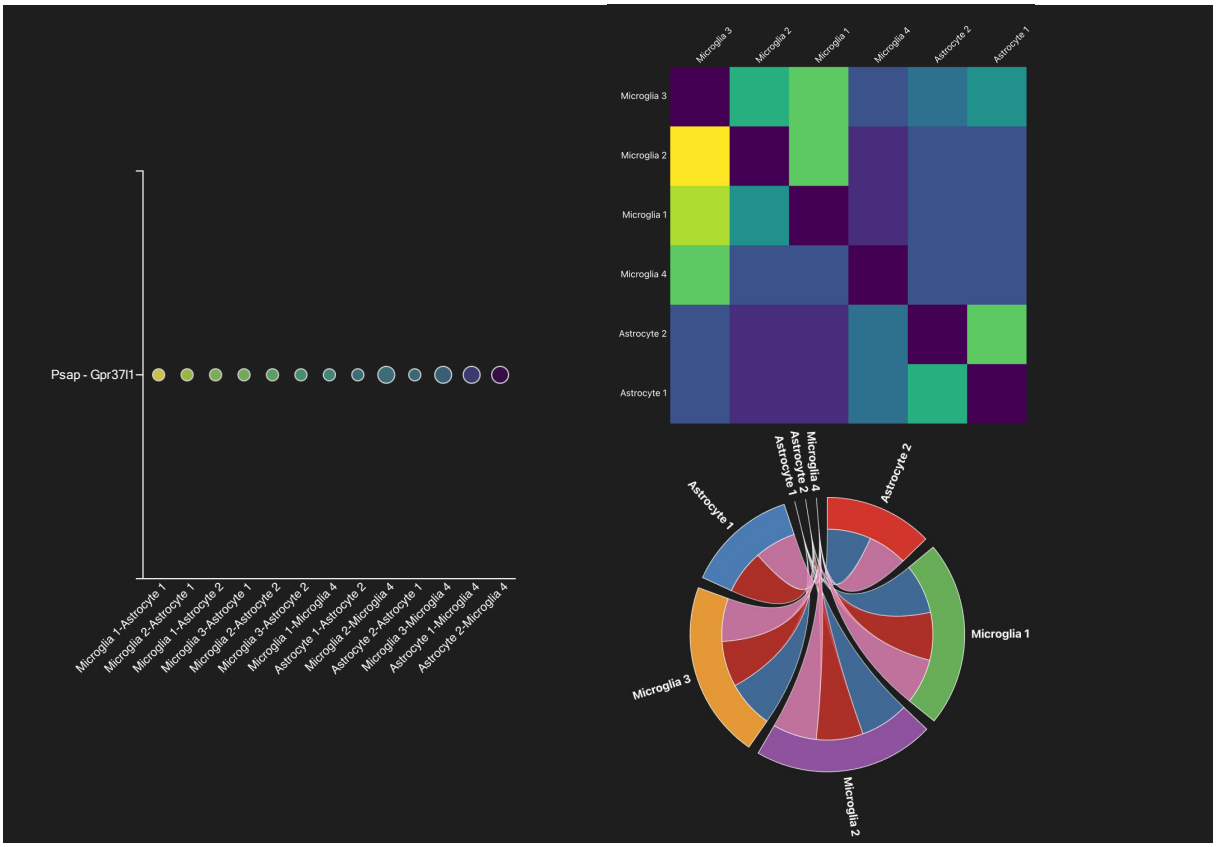
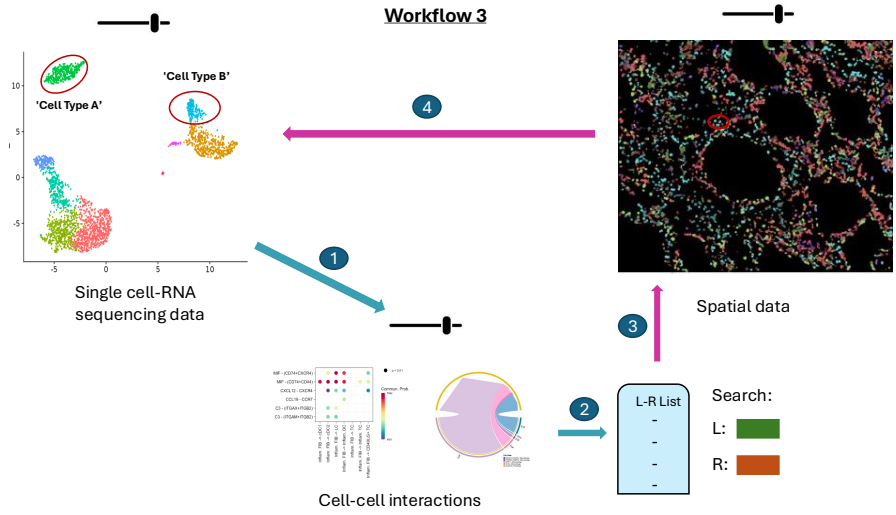


Figure 3.37: Left) Dotplot showing Psap-Gpr37l1 interactions between microglia and astrocyte subsets. Dots are coloured by their interaction probability and size represents their p-value ($p<0.05$). Top right) Frequency heatmap showing the number of interactions between microglia and astrocyte subsets. Bottom right) Circos plot of selected Psap-Gpr37l1 interactions between microglia and astrocytes.

We also wanted to extend the functionality of cellXplore beyond a single spatial technology, so we implemented a publicly available single cell and patient-matched Xenium Breast Cancer dataset to demonstrate interoperability across spatial platforms.

3.4.6 Case study: Workflow 3 using patient matched single cell and Xenium of a Breast Cancer tumour



Algorithm 3 Cell-to-Cell Interaction Lookup by Cell and LR Subset

Select c_R, c_S in sc
for each (c_R, c_S) in sc **do**
  return CCI where $CCI_{c_R} = c_R$ and $CCI_{c_S} = c_S$
  Select $LR_{(i..j)}$ from CCI_{sub}
  if $CCI_{\text{sub}} = \text{True}$ **then**
    Select c_R, c_S in **ST** where $xy_{\text{dist}} < \text{dist}$
    return $LR^2_{(i..j)}$
    for each $LR^2_{(i..j)}$ in CCI **do**
      return CCI where $CCI_L = L_{(i..j)}$ and $CCI_R = R_{(i..j)}$
      for each c_R, c_S in CCI^2_{sub} **do**
        Select c_R, c_S in sc
      end for
    end for
  else
    Throw Error: "No interactions found"
  end if
end for

Symbol Definitions

ST Spatial transcriptomics data
sc Single-cell RNA-seq data
CCI Cell-cell interaction table
LR Ligand-receptor pair
r Region of interest
c_R Receiver cell
c_S Sender cell
xy_{dist} Distance of captured area
IntP Interacting pair ID

Figure 3.38: Overview of user workflow three and analysis query pseudo-code

In workflow three, we wanted the user to select cell types of interest in the single cell, investigate the interacting ligand-receptor pairs, search multiple ligand-receptor pairs in the spatial data and validate again with the single cell gene expression (Figure 3.38). Moving on from pseudo-bulk spatial transcriptomics, we demonstrate this on a publicly available Xenium Breast Cancer dataset with patient matched single cell data shown in cellXplore in Figure 3.39 where individual cells are observed in their spatial context.

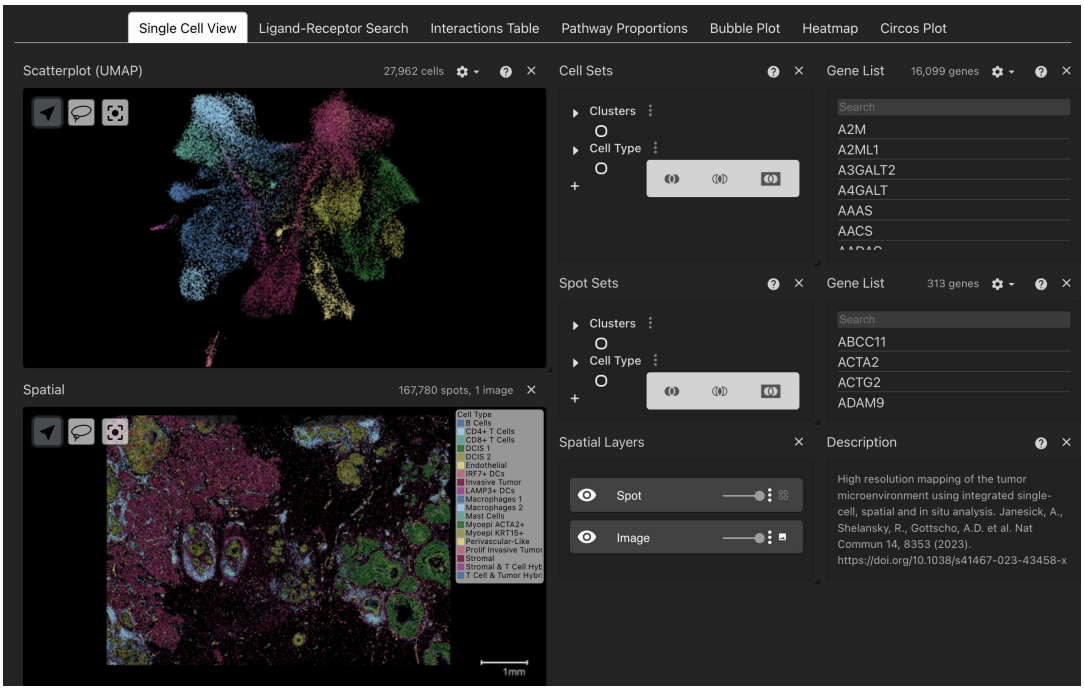


Figure 3.39: Screenshot of cellXplore visualising the single cell and Xenium breast cancer datasets.

The original study took samples from FFPE human breast cancer sections and performed single cell, Visium and Xenium in order to map the tumour microenvironment. They characterised three distinct cancer domains, invasive tumour and two types of ductal carcinoma in situ which they term DCIS1/2 alongside stromal and immune compartments. After computing cellular interaction inference, we can see the key pathways in the tumour microenvironment are collagen and laminin indicating suggesting epithelial changes such as epithelial-mesenchymal transition (EMT) and increased cell invasion (Figure 3.40).

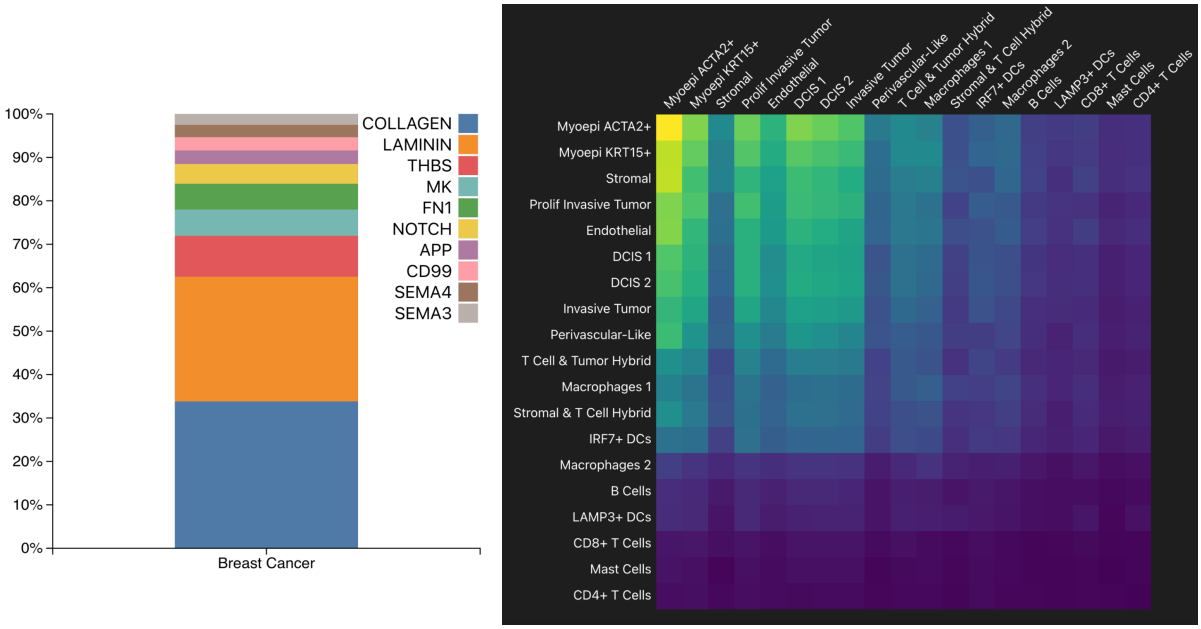


Figure 3.40: Left) Stacked proportion bar plot showing the top 10 interaction signaling pathways in the breast cancer dataset. Right) Frequency heatmap of all interactions between cell types in the breast cancer dataset

When we plot the frequency of interactions between all the cell types we observe the highest number of interactions between the Myoepi ACTA2+ cells and the tumour subtypes. Now we have identified a particular cell type pair of interest we can apply workflow three to see what interactions are occurring in this context. First, we can check to see if the cell types exist in the single cell and gain insight to their spatial distribution across the tissue (Figure 3.41).

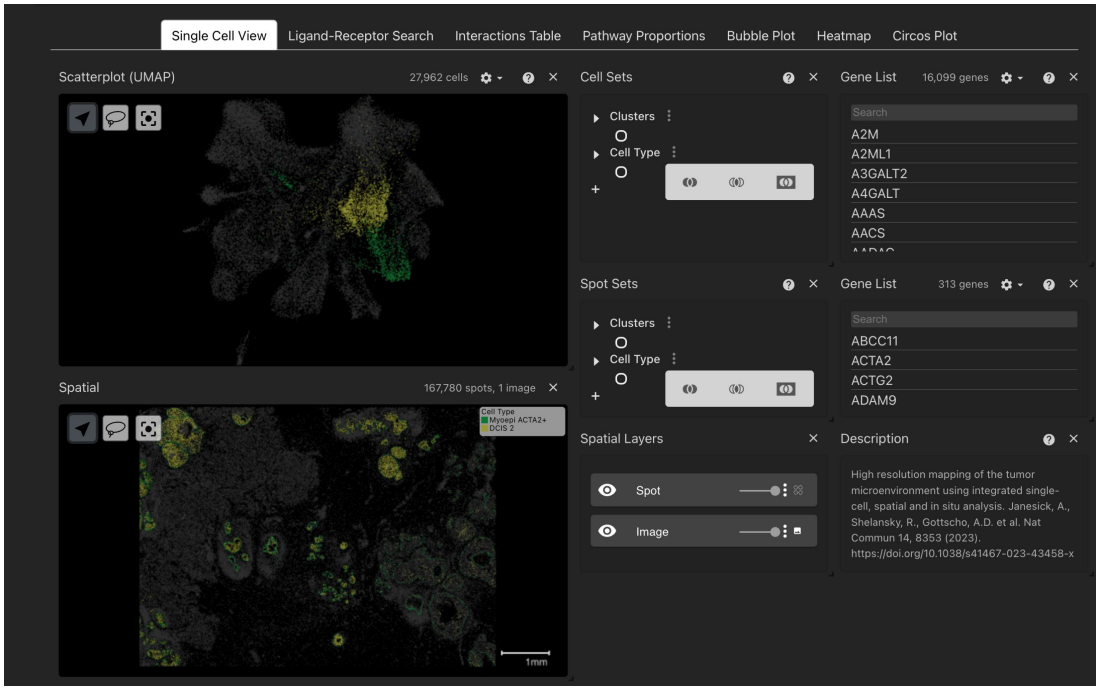


Figure 3.41: Screenshot visualising Myoepi ACTA2+ and DCIS 2 cells in the single cell and Xenium breast cancer datasets.

Here we select the Myoepi ACTA2+ cluster and one of the tumour subtypes DCIS 2 and clearly show close co-localisation suggesting that this myoepithelium cluster may play a pivotal role at the tumour microenvironment interface. As the Xenium dataset is a targeted panel of 313 genes it is useful to be able to filter the cellular interaction results to only contain ligand-receptors that are present in the panel. This is so that we can validate ligand-receptors in the data in their spatial context and consider these high confidence interactions. The Interactions Table tab allows the user to input multiple ligand and receptors to manually filter the table however, we can pull the whole list of genes present in the spatial data by clicking the 'Filter for Spatial Genes' button (Figure 3.42). This pulls the list of genes in the spatial data object and searches the table for ligand-receptor pairs that only occur in the spatial data gene panel.

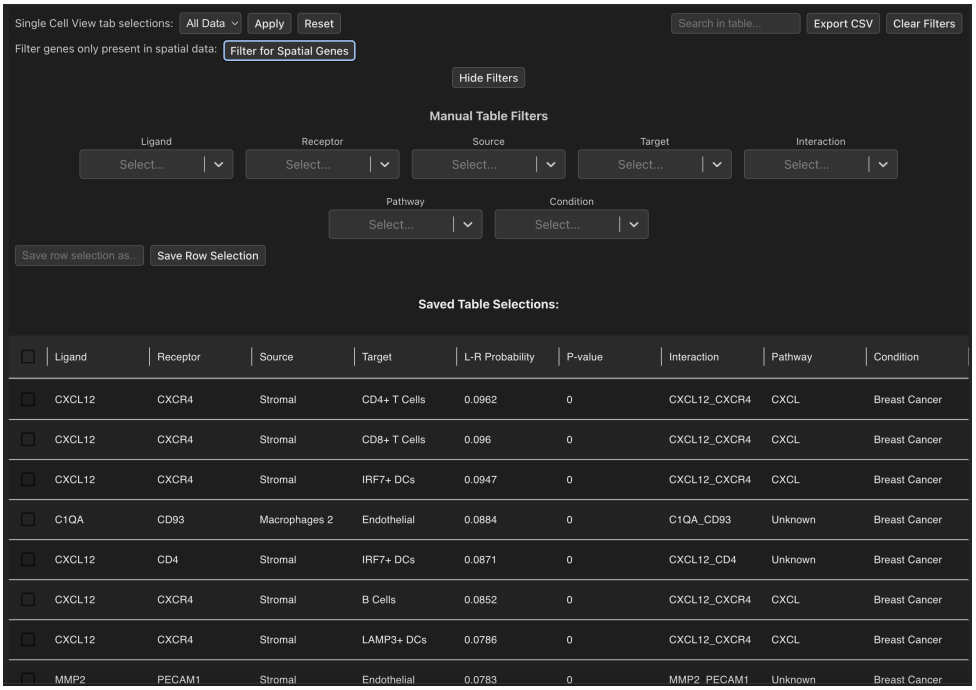


Figure 3.42: Screenshot showing the filtering of the interaction table by using a list of spatial genes

Once we apply this filtering, we go from 40,000 interactions to 4,000 substantially excluding interactions that we cannot orthogonally validate in the Xenium dataset. We can then use the manual filtering functionality to select interactions that occur between the DCIS 2 and Myoepi ACTA2+ clusters, resulting in two interactions shown in Figure 3.43.

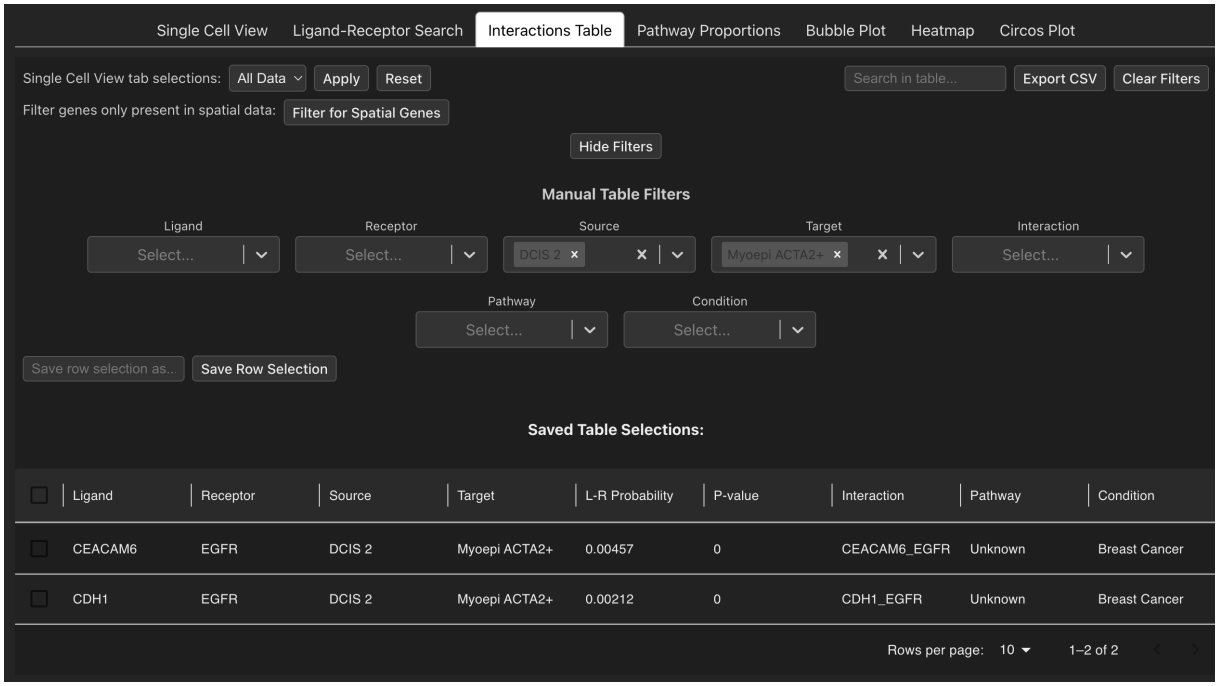


Figure 3.43: Screenshot of spatially relevant interactions between DCIS 2 and Myoepi ACTA2+ cells

Taking the first interaction result of *CEACAM6* being expressed by DCIS 2 cells and *EGFR* being expressed by Myoepi ACTA2+ cells we can check their spatial gene expression to see if they co-localise at the tumour microenvironment interface. As shown in Figure 3.44 we can nicely see that indeed *CEACAM6* is expressed by DCIS 2 cells and other tumour subtypes present in the tissue and *EGFR* is expressed in the surrounding area around the tumour cells.

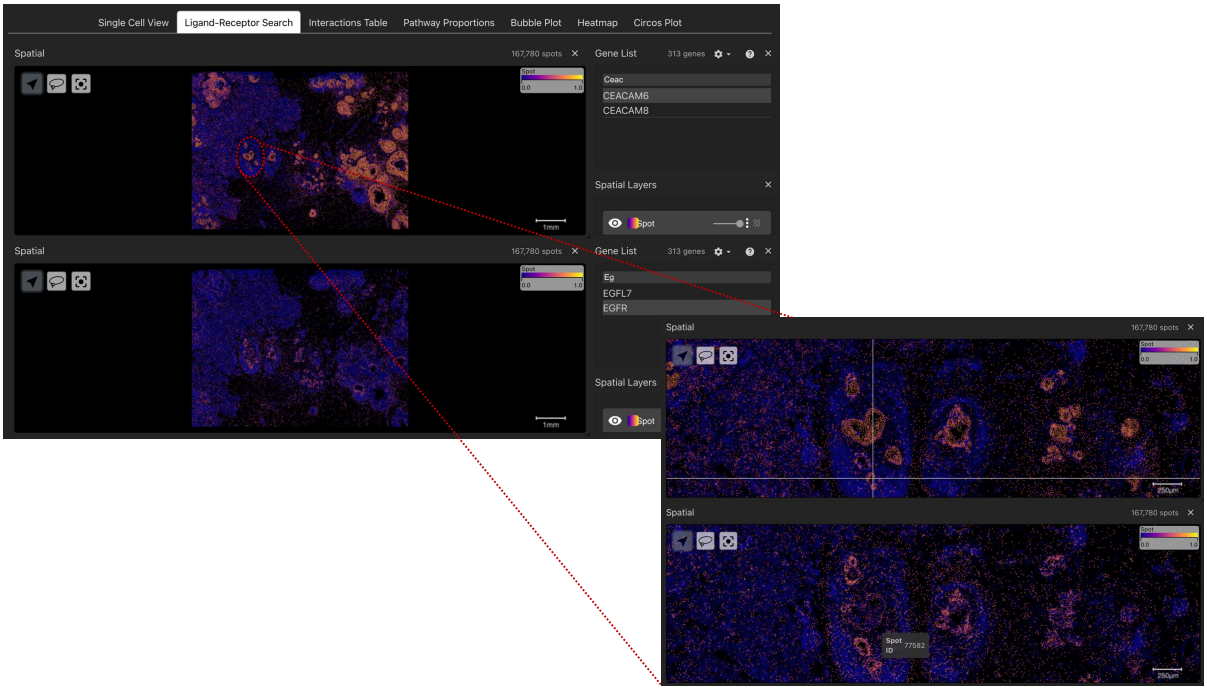


Figure 3.44: Screenshot showing the spatial gene distribution of CEACAM6 and EGFR in the breast cancer Xenium dataset

When we zoom in and take a closer look, it is clear that the spatial gene expression of this ligand-receptor interaction is found at the tumour interface indicating that *CEACAM6* and *EGFR* may play a role in tumour progression. After validating the interaction in the spatial context, we might want to return and check the expression of this ligand-receptor pair in the single cell and ensure that it is expressed by our cell types of interest (Figure 3.45).

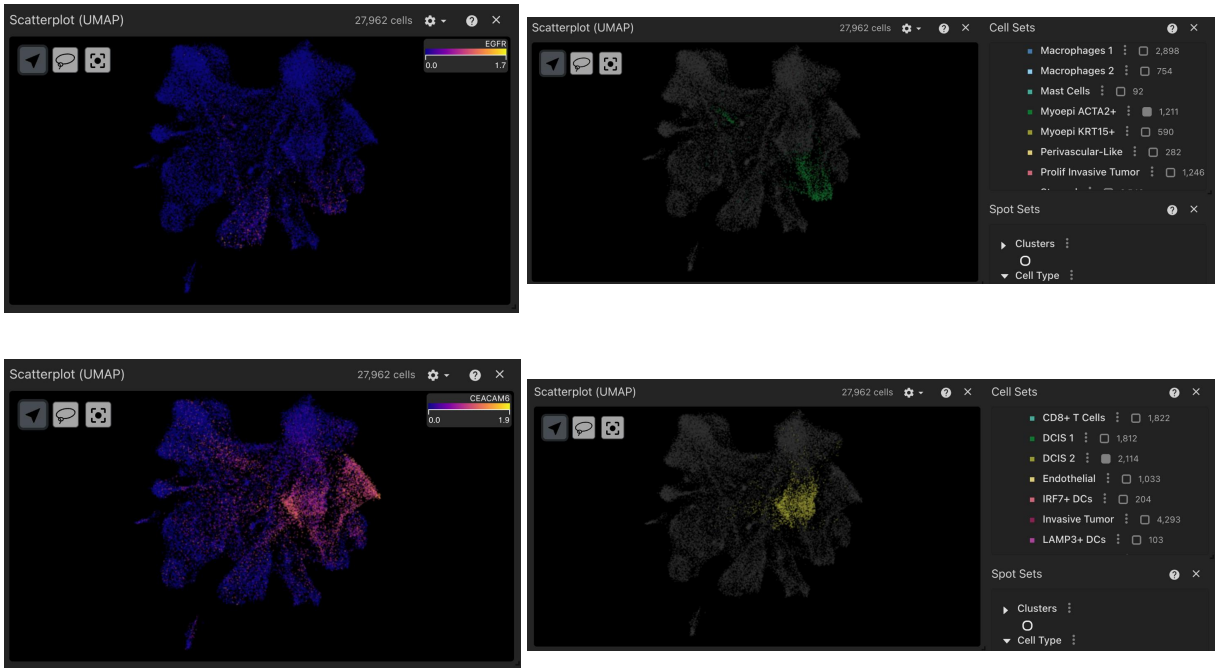


Figure 3.45: Screenshot showing the gene expression of CEACAM6 and EGFR in the single cell dataset alongside the cell type labels projected onto the UMAP

CEACAM6 expression is clearly expressed by the DCIS 2 cells along with other tumour subtypes whereas *EGFR* expression in the Myoepi ACTA2+ cluster is limited to a proportion of cells in the cluster. However, as we validated the spatial expression of these genes we can be confident they exist in space in the tissue section. Finally, we can use other visualisations within cellXplore to quantitively display the interaction such as a bubble plot in Figure 3.46.

Figure 3.46: Bubble plot showing the top 25 cellular interactions between DCIS 2 and Myoepi ACTA2+ cells using a p-value cutoff of 0.05 ($p<0.05$)

By plotting the top 25 interactions inferred from the single cell data we can see the presence of *CEACAM6-EGFR* between DCIS 2 and Myoepi ACTA2+ cells. *CEACAM6* has been shown to regulate cell migration through *EGFR* signaling in various cancers[265–267], thus indicating this interaction is critical in tumour progression. By implementing workflow three, we can show how we can use cellXplore to select cell types of interest, use gene lists to filter our interaction results by prioritising spatial genes and hone in on high confidence interactions that play an important role in disease.

3.5 Discussion

In this chapter, we present cellXplore, a novel web tool to facilitate visualisation and interpretation of cellular interactions. We demonstrate using three distinct workflows how a user may wish to interact with the tool to interrogate cellular interaction results leveraging both single cell and spatial data. Furthermore, we show that cellXplore can be utilised across different spatial platforms providing examples from whole transcriptome Visium pseudo-bulk spatial to imaging-based single cell technologies such as Xenium. Through applying cellXplore using these two datasets, we show different visualisations of cellular interactions such as tabular data, heatmaps, dotplots, circos plots and proportion bar plots. In addition to this, gene expression and spatial distribution of ligand-receptors can be visualised to validate cellular interactions of interest. The interactive interface cellXplore provides flexibility to the user through manual filtering of interactions, selection of regions or cell types of interest and storage of selections to visualise results with no prior coding experience necessary. In the first workflow we re-analysed a Visium dataset with paired single cell data during *T.brucei* infection and recapitulated findings from the original study and revealed varying patterns of microglia-plasma cell cross-talk that were not highlighted in the paper. We observed a decrease in interactions involved in the PSAP signalling pathway in active infection compared to naive controls. This pathway is vital for maintaining lipid homeostasis by degrading sphingolipids in the brain, with PSAP deficiency being shown to contribute to neurodegradation in neurodegenerative diseases such as Parkinsons disease[268,269]. Other pathways were increased that are inline with neuroinflammation and disrupted homeostasis such as the MIF and CCL signalling pathways. The macrophage migration inhibitory factor pathway signals upstream of cytokines and regulated the innate immune response[270]. Studies have reported that the *MIF - CD74/CD44* interaction promotes maintenance, proliferation and survival of microglia and B cells[271]. In addition to this, a cleavage product of this interaction has been reported to induce cell–cell signalling and cell survival in B cells, recapitulating the important role of B-cells in the infection[272]. Then through the second workflow we used a different analysis pipeline

to validate interactions found between microglia and astrocytes that were co-localised together in their spatial context. It has been well reported that microglia-astrocyte cross talk is present in prominent neuroinflammation marked by reactive astrogliosis, microgliosis, and elevated levels of proinflammatory cytokines[273]. We also reported interactions such as C3 expressed on astrocytes and C3aR1 on microglia suggesting that astrocytes can modulate microglial reactivity by activation through the complement signalling pathway in neuroinflammatory conditions[274]. Finally, in the third workflow we show interactions between myoepithelial cells and the tumour microenvironment leveraging all modalities of data, whilst also highlighting the limitation of validating cellular interactions in spatial technologies that utilise targeted panels. This analysis showed a key interaction between CEACAM6 being expressed on DCIS 2 cells and EGFR on myoepithelium. CEACAM6 can activate the ERK/MAPK pathway directly or through EGFR, promoting tumour proliferation, invasion, migration and plays a role in resistance to chemotherapy, making it an attractive target for immunomodulatory therapies[267,275]. The second interaction we identified involved E-cadherin which in most settings has a immunosuppresive role in cancers. However, when interacting with EGFR it promotes a hyper-proliferative phenotype in breast cancer cells and is strongly correlated with breast cancer survival rates[276]. In conclusion, both novel and reported interactions can be investigating using the tool in a wide range of contexts and diseases. Here we also presented two iterations of the web tool, the first being the legacy cellXplore, built within the cellxgene framework, that offered limited visual functionality of spatial transcriptomic data. I outlined the first iteration of cellXplore that was implemented within the cellxgene[247] framework and was a module of the cellxgeneVIP plugin[250] however this was unsuitable for the use-purpose of the tool due to several factors. This then led to the development of the current implementation of cellXplore that was build as a Flask-React app using visualisation components from the Vitessce suite[257] that is tailored to visualise single cell and more importantly spatial data. Firstly, the cellxgene framework was designed to facilitate analysis of differential gene expression in single cell data only. It provided useful plotting functions such as violin plots, gene expression heatmaps, and density plots however, these were not extended to be suitable for spatial analysis. Furthermore, these plotting functionalities were limited to a very small viewing area of the visualisation plugin tab. Thus, when trying

to visualise spatial data, the user is limited to a very small proportion of the available space on the webpage. Another major limitation of the legacy cellXplore was the structure and maintainability of the inherited codebase. The platform was originally forked from the cellxgeneVIP GitHub repository, a visualisation plugin developed by the bxgenomics group, that extended the functionality of the original cellxgene visualisation platform. The backend was implemented as a single Python file consisting of thousands of lines, while the frontend interface was also contained within a large, minimally structured JavaScript file. This lack of modularity and failure to adhere to several core software architecture principles made the codebase extremely difficult to navigate, debug, or test. Furthermore, the code lacked any documentation or commenting of the code, making adding new functionality challenging. Additionally, the simultaneous use of Python, R, and JavaScript in the same environment created a complex dependency ecosystem, leading to frequent and difficult-to-resolve package conflicts during installation. Due to these limitations, the further development of the legacy cellXplore become constrained and underscored the need for a more modular, well-documented, and language-consistent system architecture.

The last limitation of the legacy cellXplore was its need for a .h5ad file as input, which, while widely used in single-cell RNA-seq, is inherently optimised for matrix-like, tabular data. This made it poorly suited for spatial transcriptomics datasets, where high-resolution image data are essential components. The images would be stored in the *'uns'* slot of the AnnData object and when attempted to be loaded into the legacy cellXplore it resulted in computational lag that made spatial viewing challenging. Thus, to overcome this, only the spatial coordinates stored in obsm could be used and visualised, while the actual tissue images were unusable. This was achieved by completely overhauling the old codebase and creating a Flask-React app from scratch, this time aiming to follow correct coding practices. The new architecture allows visualisation and user interaction to be handled entirely by the frontend, while the backend serves only to generate the JSON files needed by the Vitessce visualisation components and serving preprocessed tabular interaction data. In particular, the frontend has been refactored to manage the interactive filtering, tab switching, and plotting logic, which is cleanly organised into separate JavaS-

cript modules that correspond to each visualisation tab. By implementing this modular structure in the frontend, we improve maintainability and debugging while also facilitating potential development of new features in the future. Lastly, cellXplore transitioned to using the Zarr storage format in conjunction with AnnData/SpatialData objects for its data backend. Zarr is an efficient, scalable format for storing large multidimensional arrays and is specifically optimised for out-of-core computation. By reading data in chunks, it can handle datasets that are larger than system memory and only loads relevant portions of the dataset into memory. This eliminates the need for loading the entire dataset into memory, a useful feature for high-resolution tissue images and large-scale spatial transcriptomics data, reducing memory overhead and preventing computational lag. The SpatialData object extends AnnData by incorporating additional spatial modalities—such as transcript coordinates, segmentation masks, geometric shapes, and multichannel images—each organised in their own distinct layers (points, labels, shapes, images, etc.), and has become increasingly popular in the spatial community. By implementing these changes cellXplore now provides a comfortable user experience, adopts better coding practice within its codebase and finally handles large memory high-dimensional data in formats used by the wider community.

Despite many improvements to cellXplore there are still some limitations that need to be addressed by the tool. Some are minor aesthetic improvements of the plotting functions, however more critically, the manner of data ingestion must be developed. Currently, there are some manual preprocessing steps that exist to allow the data to be read in the correct format. This includes ensuring that the cellular interactions are stored in the *'uns'* slot of the data object, formatting any metadata to be visualised as a category type, and checking what data is available for the user interface layout. In Figure 3.47 we propose a potential future improvement schema to facilitate data loading into cellXplore.

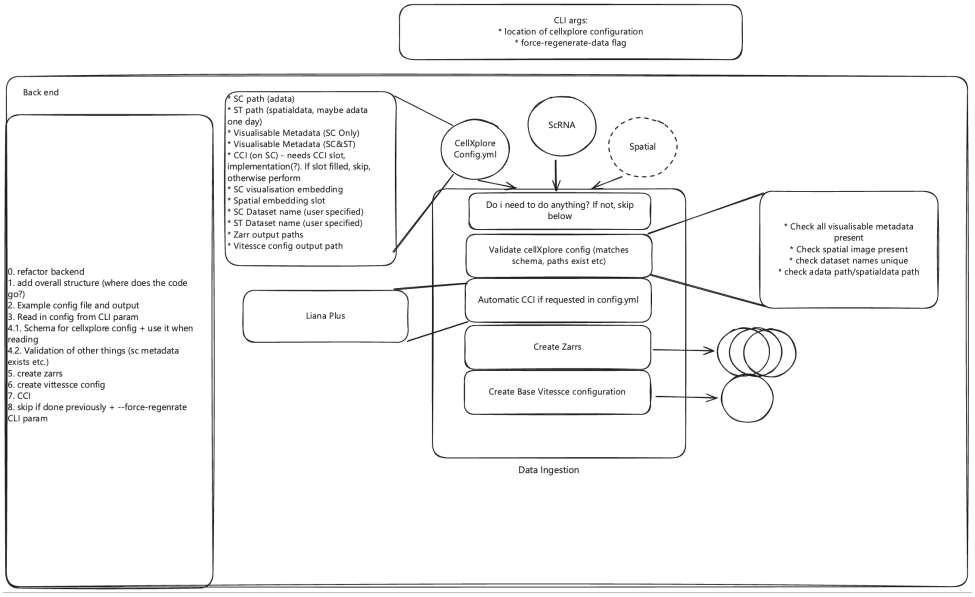


Figure 3.47: A) Schematic demonstrating future work with cellXplore

This would consist of a.yml file that validates all elements of the input data object by first checking the paths to the available data, checking that the metadata is of dtype *category* and taking any additional input from the user such as a description of the dataset. Once this .yml file is loaded into the tool, if necessary single cell or spatial data objects would be converted to a Zarr store and the Vitessce config would be validated and generated. The tool is currently hosted on a development server with static URLs to the datasets presented in this chapter and is not optimised for personal use. We propose in future work for the user to clone the GitHub repository and then launch the tool using CLI parameters to take the .yml file and a cellular interaction flag. The cellular interaction flag would check if cellular interactions had been inferred and in the event they are absent, would compute cellular interactions on the fly using the Liana+[155] package. This package provides a collection of useful Python wrappers of popular cellular inference tools such as CellPhoneDB and CellChat. The CLI flag would also allow the user to select the inference tool they prefer before proceeding to launch cellXplore with their data. Allowing more flexibility in launching and implementing cellXplore, would allow accessibility to users using any dataset, instead of its current state of static URL hosting.

In addition to this, another limitation of cellXplore is the underutilisation of spatial data. The spatial data is leveraged during analysis in a qualitative manner, mainly using visual cues to determine co-localised gene expression. Future development of cellXplore could include quantitative information from spatial analysis such as neighbourhood analysis of co-localised cell types and spatially variable genes. By providing additional visualisations to inform the user of spatial statistics, cellular interaction inference could be more quantitatively constrained spatially, improving validation of true positive interactions. Furthermore, cellXplore could provide more quantitative gene expression visualisations such as violin plots to consolidate gene expression of ligand-receptor pairs in either modality of data. Finally, since the development of cellXplore there have been many cellular inference packages that aim to compute cellular interactions on the spatial data instead of the single cell data only. Although, as demonstrated in this chapter this analysis can be hindered by available gene panels, the tools functionality could be extended to incorporate the results of this type of analysis to allow the user to investigate spatially aware cellular interactions.

Lastly, a nice feature would be the extension of cellXplore to facilitate cellular interaction inference and visualisation of other data modalities such as spatial proteomics. As we observed in the first chapter imaging mass cytometry can be leveraged with single cell data to hone in on interactions of interest. By using multi-modal integration of the Covid-19 lung data we could also project transcriptomic and proteomic read-out into a shared low dimensional space. The potential limitation of utilising cellXplore to visualise the Covid-19 dataset is the requirement of harmonised cell type labels across both datasets and the absence of ligand-receptor expression in the protein panel. However, using multi-modal integration tools such as Maxfuse[249] we can potentially work-around these constraints by using projected expression and cell type labels across both data modalities. Nonetheless, cellXplore can be utilised to explore interactions of statically hosted datasets that mandate a JSON file to be created to host on the Glasgow Atlas server. The tool stands as a useful resource to aid bioinformatic analysis and visualisation of cellular interactions with no coding background required, bridging the gap between computational biologists and

bench scientists. Overall, cellXplore is a versatile visualisation tool that can be used to investigate cellular interaction results. By creating a smooth user interface and experience, cellXplore can be used to facilitate collaborations of cellular inference on big data amongst researchers that lack a coding background. We show the functionality of cellXplore on two distinct datasets and provide different analytical workflows for data interpretation. In future implementations of cellXplore, we hope it can be launched with any dataset and facilitate multi-modal single cell cellular inference exploration.

3.6 Appendix

```
{
  version: "1.0.17", % Version of Vitessce
  name: "Breast Cancer Multi-Modal", % Name of the dataset
  description: "High resolution mapping of the tumor microenvironment
     ...", % Short description of the dataset
  datasets: [{
      uid: "A", % Unique dataset identifier
      name: "Single-Cell RNA",
      files: [{
          fileType: "anndata.zarr", % Type of data store
          url: ".../sc_FPPE_breast_cancer.zarr", % Path to the Zarr
             store
          options: { % Dictionary defining visualisation components
            obsEmbedding: [{ % Defines the dimensionality reductions
                path: "obsm/X_umap",
                dims: [0, 1],
                embeddingType: "UMAP"
              }],
            obsSets: [ % Defines the categorical metadata
              { name: "Clusters", path: "obs/clusters" },
```

```
            { name: "Cell Type", path: "obs/Cell_Type" }
          ],
          obsFeatureMatrix: { % Defines the gene expression matrix
          path: "X" }
        },
        coordinationValues: { % Defines what type of component the
           data is
          obsType: "cell",
          obsSetSelection: "obsSetSelectionScope"
        }}]
  },
  {
    uid: "B",
    name: "Xenium Spatial",
    files: [{
        fileType: "spatialdata.zarr",
        url: ".../Xenium_proper_data.zarr",
        options: {
          obsFeatureMatrix: { "path": "tables/table/X" },
          obsSets: {
            obsSets: [
              { name: "Clusters",
                path: "tables/table/obs/leiden" },
              { name: "Cell Type",
                path: "tables/table/obs/clusters" }
            ],
            tablePath: "tables/table"
          },
          obsSpots: { % Defines the spot-level mask
            path: "shapes/cell_circles",
            tablePath: "tables/table"
          },
          image: { % Defines the spatial image
```

```
51              "path": "images/morphology_focus" }
52            },
53            coordinationValues: {
54              obsType: "spot",
55              obsSetSelection: "obsSetSelectionScope"
56            }}]
57      }],
58    layout: [ % Defines the structural layout of the user interface
59      {
60        component: "scatterplot", % Defines the type of visualisation
              component
61        coordinationScopes: { % Links elements of the component to the
              defined data types
62          dataset: "A",
63          embeddingType: "A",
64          obsType: "A",
65          obsSetSelection: "A"
66        }, % Below defines the component size and layout
67        x: 0.0, y: 0.0, w: 6.0, h: 6.0
68      },
69      ...
70    ],
71    initStrategy: "auto" % Automatic initialisation once all data can
        be accessed
72  }
```

Listing 3.1: Example JSON file to initialise the Vitessce visualisation components in the Single Cell View tab.

Chapter 4

Dissecting cellular interactions in big data: Contextualising cellular interactions using atlas-level single cell and sequencing based spatial transcriptomics

4.1 Abstract

Cellular inference can be extended to large complex datasets where we can implement indirect validation using single cell and spatial data as demonstrated in cellXplore. Here we present two user-cases with validated cellular interactions both experimental and indirect, where cellular interaction inference is applied to an atlas-level dataset and a spatial transcriptomics dataset where the spatial topology of the tissue is warped. In more detail, the first case study focuses on a multifactor single-cell atlas of macrophage and fibroblast populations spanning four tissues in both homeostasis and inflammatory disease. The size and complexity of the dataset required advanced inference and visualisation strategies to resolve context-specific interactions. Our analysis reveals shared and tissue-unique myeloid–stromal phenotypes, identifies conserved pathways of tissue

resident macrophage–fibroblast crosstalk that underpin inflammation, alongside context-specific interactions that reflect the unique microenvironments of different organs. The second case study applies 10X Visium spatial transcriptomics to the intestine, where the non-native orientation of tissue presents challenges in cellular interaction inference. Using the 'swiss-roll' technique to capture the crypt–villus axis, we investigated host–parasite interactions during *Heligmosomoides polygyrus* infection across four time points. This analysis identified immune and epithelial cell programs associated with granuloma formation, stem cell reprogramming, and parasite-driven immunomodulation, providing insight into helminth infection within a distorted tissue landscape. Together, these studies demonstrate the intricacies of cellular interaction inference within atypical contexts, offering new perspectives on myeloid–stromal communication and the spatial dynamics of parasitic infection in the murine intestine.

4.2 Introduction

In previous chapters, we presented an analysis of how cellular interactions elucidate immunomodulatory mechanisms in COVID-19 infection and highlighted the need for an interactive visualisation tool to facilitate cellular interaction interpretation in single cell and spatial transcriptomics data. Building on from this, the focus of this chapter shifts to demonstrate methodological challenges in cellular inference to investigate biological insights in complex datasets. Specifically, we aim to show how cellular interaction analyses can be adapted to datasets where the nature of the experimental design or the tissue architecture complicates standard analysis practice. The first use case is a multifactor single-cell atlas comprising multiple tissues, conditions, and macrophage and fibroblast cell types. Here, the complexity arises from scale and heterogeneity of the data arising from different studies, requiring advanced cellular interaction inference across different biological contexts and coherent visualisation of the interactions of interest. Recently, the complexity of single cell datasets are increasing with many datasets now containing

millions of cells profiled from many samples, different disease states[3,9,214], tissues[14,277,278] and even species[10,11]. Single-cell atlases provide comprehensive reference maps of cellular states across tissues, developmental stages, and disease contexts, generated by integrating large-scale single-cell RNA sequencing datasets, enabling the discovery of rare cell types[19], characterisation of lineage hierarchies[279], and cell type specific states that underpin tissue function and pathology[280,281]. Here we developed a single cell atlas of macrophage and fibroblast populations across four different tissues in homeostasis and inflammatory diseases to identify shared and distinct cellular interactions and indicators of common pathways in inflammation. Tissue-resident macrophages (TRMs) are critical for normal tissue development, physiology, and homeostasis[282]. The majority of TRMs derive from embryonic precursors and are established in tissue before birth, where they are primarily responsible for efferocytosis of apoptotic cells and tissue debris[283,284] and possess unique tissue-specific physiological functions[285]. The ability of TRM to execute such crucial, unique functions is determined by signals in their local niche, which is formed by other tissue resident cells such as fibroblasts and/or epithelial cells as well as soluble mediators[286]. In particular under inflammatory conditions, we observe the infiltration and maturation of proinflammatory monocyte-derived macrophages and expansion of specific fibroblast subtypes[287,288]. Although separate macrophage[289] and fibroblast single cell atlases[281,290,291] have been proposed, tissue-specific interaction between both tissue resident macrophages and stromal cells are not well characterised up to now and have only been reported in a single tissue[292,293]. Therefore, the aim of this study was to identify common and tissue-unique myeloid and stromal phenotypes and to decipher unique and cross-tissue cells and signals involved in pathogenic tissue activation. In this instance, we demonstrate a multi-combinatory approach to inferring cellular interactions in homeostasis and disease across multiple tissues and cell types. The second use case involves a 10X Visium spatial transcriptomics dataset where the tissue is captured in a non-native orientation. The spatial axis allows us to spatially confirm predicted interactions in the native tissue context, filtering out false positives from interactions arising between distant cell types. Spatial transcriptomics has emerged as a powerful tool to unravel the intricacies of gene expression patterns within the complex architecture of tissues[278,294]. However, capturing large or highly structured tissues such as the intestine within the limited 6.5 mm^2 cap-

ture area of the 10x Visium platform presents a significant challenge. To address this, orientation strategies have been developed, most notably the 'swiss-roll' technique, in which the intestinal tissue is longitudinally cut and carefully rolled into a spiral before sectioning. This approach allows a greater length of the intestine to be represented on a single Visium slide while preserving the crypt–villus architecture[295]. Despite the topology of the tissue being warped, spatial studies investigating perturbations in the intestine have been conducted and adopt various techniques of unravelling the intestine such as digital unrolling and calculating an anterior-posterior axis[294,296,297]. This reconstruction of the tissue architecture is critical in cellular interaction inference as interactions that may seem in close proximity to each other in the tissue plane could in fact be distant when the tissue has been unrolled. Here we present a spatial transcriptomics analysis to investigate the localised transcriptional profile within the intestinal epithelium and lamina propria of both naïve mice and mice infected with *Heligmosomoides polygyrus* over infection across 4 different time points. Helminths, or parasitic worms such as *H. polygyrus*, are amongst the most prevalent infectious agents afflicting individuals in developing nations, contributing to a global disease burden as severe as more recognized conditions like malaria and tuberculosis[27]. Upon oral ingestion, larvae swiftly traverse the small intestine's epithelial barrier, establishing in the submucosal tissue by forming granulomas before maturing into adults and returning to the intestinal lumen[298]. Infection initiates a host type 2 immune response, which plays a crucial role in various physiological processes, ranging from safeguarding against parasites to contributing to metabolic adaptation, homeostasis, and tissue regeneration[299,300]. Despite the host's robust immune response, *H. polygyrus* establishes long-term chronic infections attributed to its immunomodulatory effects that allow it to evade the immune system. Most notable is its secretion of proteins that mimic the function of TGF-β[34], a pivotal regulator of the immune system through the induction of T regulatory cells, which inhibit the inflammatory effects of a variety of immune cells[301]. Until recently, investigations into helminth immunomodulation predominantly focused on its downstream impacts on immune cell populations[31,302]. However, in more recent studies, attention has shifted towards unravelling the complex interactions between intestinal helminths and the epithelium, in particular, granuloma formation and epithelial repair[303,304]. During this phase, stem cells in the surrounding areas have been observed

to undergo a ”reversal” to a foetal-like repair phenotype and exhibit a compromised capacity to differentiate into various effector secretory cell subsets, including tuft, goblet, and Paneth cells[304,305]. While the interactions and effects of the nematode on the immune system and epithelium are clearly extensive, it is important to better understand the spatial context in which host-parasite interactions occur and what changes occur across the infected tissue that may favour parasite establishment or clearance. Thus using spatial transcriptomics, we pinpointed immune cell types within the granuloma, and identify potential ligand-receptor pairs mediating communication between tissue sites in granuloma formation and stem cell differentiation providing new insights into the complex interplay between *H. polygyrus* and the intestinal environment. Together, these two case studies demonstrate the application of cellular inference in atypical contexts, providing insight into tissue resident macrophage-fibroblast interactions in complex atlas-level single cell data in homeostasis and disease and to spatially profile host-parasite intestinal interactions over *H. polygyrus* infection.

4.3 Methods

This methods section will detail the methods used to present the results in this chapter for the two analyses. For information about how the normalised macrophage atlas was created and integrated please refer to our bioRxiv preprint here (https://www.biorxiv.org/content/biorxiv/early/2025/03/04/2025.03.04.641204.full.pdf). In the results section Figure 4.1 was created by Dr. Lucy Macdonald providing a comprehensive overview of the datasets included in this study. The normalised macrophage-fibroblast atlas in Figure 4.2 A was created by Dr. Lucy Macdonald after integrating the datasets together and annotating tissue-resident macrophage and fibroblast subsets. Histological staining in Figure 4.16 was completed by Caroline Opselt and her team based at the University of Zurich. Similarly, for methods regarding the sample preparation of the *H. polygyrus* Visium dataset please refer to out bioRxiv preprint here: (https://www.

biorxiv.org/content/10.1101/2024.02.09.579622v1). Mouse experiments and tissue preparation for sequencing was carried out by Dr. Marta Campillo at the University of Glasgow. Figure 4.20 A-F were jointly computed by myself and Dr. Ross Laidlaw at the University of Glasgow. Similarly Figure 4.24 A-C were calculated and plotted by Dr. Ross Laidlaw using a bespoke sectioning algorithm detailed in the paper. Finally, single cell reference datasets detailed in the methods below for the cell type deconvolution were prepared by Dr. Ross Laidlaw.

4.3.1 Annotation of the full atlas dataset

The normalised atlas was read into R and label transfer was performed using SingleR[109] (v2.10.0) using the normalised atlas as a reference dataset and using the wilcox method for marker gene detection: *'full.atlas <- SingleR(test=full.atlas, ref=normalised.atlas, labels=normalised.atlas$celltype, de.method="wilcox")'*. Cell types were then assigned based on the highest phred score for each cell type label. All frequency bar plots were plotted using ggplot2 in R.

4.3.2 Cellular interaction inference of the macrophage-fibroblast atlas

To interrogate ligand-receptor interactions in the synovial tissue microenvironment, we applied CellChat[146] (1.6.1) which is implemented in R. Each tissue was analysed separately for each condition as recommended by the standard package pipeline. We pooled all the inflammatory diseases included in the study into one 'Disease' condition and the healthy controls were assigned 'Healthy' in each tissue. We then performed differential expression between the two conditions for each tissue to obtain statistically significant ligand-receptor interactions expressed in at least 20% of cells through running 'identify-

OverExpressedGenes'. We used a p-value threshold of 0.05 and filtered to obtain ligand and receptors that had a log fold change value of more than 0.25. Next, we aggregated all statistically significant cellular interactions for each tissue into a master table of interactions with a 'Tissue' column to separate out interactions for each tissue. Further processing of the table was performed by filtering to remove interactions that were occurring between the same broad cell type (e.g. fibroblast – fibroblast) to observe the interactions exclusively occurring between the myeloid and stromal compartments. Venn diagrams were plotted with the VennDiagram[306] package in R (v1.7.3) and used the intersection of sender cell type, receiver cell type with their respective interaction appended in a separate column e.g. 'Fibroblast_1'-'Macrophage_1'-'interaction_pair'. Upset bar plots showing overlap were plotted with the UpsetR[307] package (v1.4.0) computed using the same column described above and including functional pathway annotation output by the CellChat package. Plots showing key changes in drivers in homeostasis and disease were obtained through the CellChat package by running 'netAnalysis_signalingRole_scatter'. Circos plots were plotted using the circlize[308] package (v.0.4.12) implemented in CellChat and all heatmaps were created using the ComplexHeatmap[230] package (v2.24.1).

4.3.3 *H. polygyrus* Visium dataset processing

The sample images were analysed and spots were assigned a metadata value according to the type of tissue the spot captured: Crypt, Villus, Peyer's Patch and in infected tissues also Granuloma, including spots that captured *H. polygyrus*. Samples were mapped against the *Mus musculus* mm10 reference using 10X Genomics' Spaceranger version 2.1.1 (10X Genomics) on default parameters except for the loupe alignment JSON file, which was edited so that unlabelled spots were also included in the final mapping output. The Spaceranger mapped Naive, D3, D5 and D7 samples were first read into R using Seurat[97] and underwent quality control to remove spots with high UMI counts. Integration was attempted using harmony[101] with Seurat v5 built-in function 'Integrate' and dimensionality reduction was performed using 15 dimensions and clustering at a resolution of 0.5. After

the failed integration attempt the objects were seperately, for each time point, read into Python and underwent quality control using SCANPY[92], again with spots with high UMI counts detected across the samples removed. The quality controlled naïve, D3, D5 and D7 samples were concatenated together into an AnnData object, with only the genes that were detected in all four of the datasets being present in the concatenated dataset. The expression values were normalised by their total sum, and each cells normalised counts scaled to the median UMI count of the concatenated object. These values were then log1p transformed. A series of differential expressed gene analyses were carried out across the infection time series. The concatenated object was split into two objects for spots labelled 'Crypt' or 'Villi'. For these datasets, each timepoint of interest was compared to its adjacent time points. For example, the D5 sample was compared against the D3 and D7 samples, while the naive sample was compared against the D3 sample. All differentially expressed genes in the study were defined as those with a Benjamini-Hochberg corrected p-value < 0.05, with p-values generated using a Wilcoxon test. Venn diagrams of overlapping genes were performed using the VennDiagram[306] package in R (v1.7.3) and the gene heatmap in Figure 4.18 was created with the ComplexHeatmap[230] package (v2.24.1). To investigate changing Wnt pathway ligands over time custom functions from the sc-toolbox GitHub repository[309] were used. To examine the distance of gene expression from the site of granulomas the semla[296] package was used (v1.2.1).

4.3.4 Preparing the intestine single-cell RNA sequencing reference datasets

The raw expression matrices and metadata of the Xu et al[310] and Haber et al[35] scRNA-seq data were downloaded and loaded into Seurat[97]. For the Xu data, more quality control of the samples was carried out. This consisted of removing cells with remarkably high/low nFeature_counts and also those with a high percentage of mitochondrial reads per cell. The cut-off values can be seen by viewing the relevant code on our GitHub. The metadata of the Xu cells was then simplified, with the cell types being designated as 'low UMI'

being merged with their regular UMI counterparts and specific subsets of cell type e.g. 'DC (Cd103+ Cd11b+)', 'DC (Cd103+ Cd11b-)' and 'DC (Cd103-C2)' being simplified as just 'DC'. Non-immune cells were also removed from the Xu datasets. The Xu and Haber Seurat objects were then merged and converted into H5AD format to be read into Python.

4.3.5 Cell2location analysis of *H. polygyrus* infected and naïve mice intestine

For setting up the model of the Xu and Haber merged scRNA-seq dataset, we used the sequencing run of each of the datasets as the 'batch_key' and included the condition of the datasets (e.g. allergy, parasite infected, naive/control) as a categorical covariate. The training parameters for the training of the Visium slide model in cell2location[189] were chosen as follows. Within both the day 7 and naïve Visium samples there was variation in total UMI counts that could not be explained by the tissue, thus the RNA detection sensitivity parameter was set to 20, as per the recommendation of the cell2location authors. The number of cells per location was chosen to be 50, based on visual observation of the scanned slides. The models were trained on a GPU with 80GB of RAM. The 5% quantile cell abundance was stored in the Visium anndata objects and used for all subsequent analysis and visualisation. Non-negative factorization (NMF) analysis in cell2location was carried out, using the concatenation of the new length and depth coordinates and the original Visium coordinates as the spatial basis for the NMF.

4.3.6 Cellular communication inference of *H. polygyrus* infection in the murine intestine

Spatial niches defined by the NMF factorisation analysis yielded four distinct spatially resolved neighbourhoods termed the villi, upper crypt, lower crypt and granuloma niche. Spots assigned these labels were fed into CellChat[146] (v.2.1.1) alongside the spatial coordinates from the full resolution tissue image to allow resulting interactions to be within spatial constraints. The conversion of spatial coordinates from pixels to micrometres was calculated using the ratio of the theoretical spot size set to 55um over the number of pixels that cover the diameter of the spot. In addition to this, the communication probability of two cells interacting was also restricted with a contact range set to 100 as recommended by the CellChat 10X Visium workflow. The CellChat database used was set to the organism 'mouse' and all functional interaction annotations were used except those classified as 'Non-protein Signalling' to avoid the inclusion of interactions involving synaptic signalling which lies outside the context of the murine intestinal tissue.

4.4 Results

The results section of this chapter will be broadly split into two sections, the first detailing the analysis of the macrophage-fibroblast atlas, including reference mapping to expand the full dataset, initial cellular inference analysis of four tissues and the cellular interaction inference and validation of interactions in tissues that demonstrated high overlap in homeostasis and disease. The second section will detail the spatial analysis of the *H. polygyrus* dataset and insights gained from cellular interaction inference on epithelial repair and immunomodulation in parasitic infection over time.

4.4.1 Section 1: Identifying cellular interactions in complex atlas level data

This work was a collaborative effort across Glasgow and Zurich aiming to identify distinct and shared myeloid and stromal cell populations across four different tissues the lung, skin, synovium and heart. Dataset curation, processing, integration and cellular annotation was completed by Dr. Lucy MacDonald at the University of Glasgow under the supervision of Professor Mariola Kurowska-Stolarska and Professor Thomas Otto during her PhD. Downstream validation of identified interactions were performed at the University Hospital Zurich under the supervision of Professor Caroline Ospelt and colleagues. The complete study and details of the single cell atlas creation can be read in our preprint here (https://www.biorxiv.org/content/biorxiv/early/2025/03/04/2025.03.04.641204.full.pdf), however, for the scope of this chapter the cellular interaction inference portion of the analysis and my contributions will be detailed below.

4.4.2 Expanding the macrophage-fibroblast atlas to the full dataset using SingleR

Dr Lucy MacDonald collected scRNAseq data from 14 public datasets (Figure 4.1), spanning four distinct tissues (heart[311–313], lung[314–316], skin[317–319], and synovium[320–323]), including 162 male and female healthy donors and disease patients, with conditions such as – Heart Failure (HF)[313], Idiopathic Pulmonary Fibrosis (IPF)[314], Systemic Sclerosis-associated Interstitial Lung Disease (SScILD)[314,316], Acne[318], Leprosy[318], Psoriasis[318], Granuloma Annulare (GA)[318], Atopic Dermatitis (AD)[317], Osteoarthritis (OA)[323], Rheumatoid Arthritis (RA)[320–323] and RA that is in sustained clinical remission[320].

Dataset	Tissue	Conditions	Samples	Platform	Cell types	DOI	PMID
Tucker et al. (2020)	Heart	Healthy donor, location-specific (apex, LA, RA, LA, LV, septum, apex) n=6 each	n=36	10x Genomics	Immune, stromal and cardiomyocyte	10.1161/CIRCULATIONAHA.119.045401	32403949
Litvinukova et al. (2020)	Heart	Healthy donor, n=14	n=14	10x Genomics	Immune, stromal and cardiomyocyte	10.1038/s41586-020-2797-4	32971526
Wang et al. (2020)	Heart	Healthy donor, n=14; heart failure, n=6;	n=20	SMARTseq2	Immune, stromal and cardiomyocyte	10.1038/s41556-019-0446-7	31915373
Reyfman et al. (2019)	Lung	Healthy donor, n=8; IPF, n=5; SSc-ILD, n=2	n=14	10x Genomics	Immune, stromal and epithelial	10.1164/rccm.201712-2410OC	30554520
Valenzi et al. (2019)	Lung	Healthy donor, n=5; SSc-ILD, n=10	n=15	10x Genomics	Stromal	10.1136/annrheumdis-2018-214865	31405848
Madisson et al. (2020)	Lung	Healthy donor, n=5	n=5	10x Genomics	Immune, stromal and epithelial	10.1186/s13059-019-1906-x	31892341
Sole-Boldo et al. (2020)	Skin	Young, n=2 and old, n=3 healthy donors	n=5	10x Genomics	Stromal	10.1038/s42003-020-0922-4	32327715
Hughes et al. (2020)	Skin	Healthy donor, n=3; Acne, n=4; GA, n=2; Leprosy, n=4; Psoriasis, n=5	n=18	Seq-Well S3	Immune, stromal and epithelial	10.1016/j.immuni.2020.09.015	33053333
He et al. (2020)	Skin	Healthy donor, n=8; Lesional AD, n=4; Non-lesional AD, n=5	n=16	10x Genomics	Immune, stromal and epithelial	10.1016/j.jaci.2020.01.042	32035984
Stephenson et al. (2018)	Synovium	RA, n=5	n=5	Microfluidic	Immune and stromal	10.1038/s41467-017-02659-x	29476078
Zhang et al. (2019)	Synovium	OA, n=5; RA, n=16	n=21	CEL-Seq2	Immune	10.1038/s41590-019-0378-1	31061532
Alivernini et al. (2020)	Synovium	Healthy donor, n = 4; active RA, n=9; RA in remission, n = 5	n=18	10x Genomics	Immune and stromal	10.1038/s41591-020-0939-8	32601335
Micheroli et al. (2022)	Synovium	RA, n=5	n=5	10x Genomics	Stromal	10.1136/rmdopen-2021-001949	34987094

Figure 4.1: Table showing the datasets Dr. Lucy Macdonald included in the study, heart is coloured in red, lung in green, skin in blue and synovium in purple.

After integration of the tissue atlases with harmony, there were 10 myeloid and 8 stromal cell populations identified across synovium, skin, lung, and heart (Figure 4.2). Due to the size of the dataset cellular annotations were completed using a normalised single cell atlas that contained a subset of the data, 20,000 myeloid cells and 20,000 stromal cells (Figure 4.2 A). The range of myeloid clusters included monocyte-like precursor populations (CD14+S100A12+ and CD16+ISG15+), proinflammatory (IL1B+ and SPP1+) and resolving (TREM2+, NR4A1+ and LYVE1+) macrophages as well as DC phenotypes (CCR7+, CD1c+ and CD207+). Stromal cell clusters included lubricin expressing (PRG4+) fibroblasts, which were the most distinct from all other stromal cell clusters, along with a VCAM1+ fibroblast subtype. Additionally, we characterised a population of CDH19+ fibroblasts, which also expressed transcripts associated with complement activation (C7, CFD) as marker genes and A2M+ fibroblasts that expressed TCF21. Other stromal clusters that we identified included collagen-producing (SPARC+), and pro-inflammatory (APOE+) fibroblasts, producing CXCL12 and C3, as well as a CD34+ fibroblast population (MFAP5+). Finally, we found actin (ACTA2), transgelin (TAGLN) and myosin light chain (MYL9) expressing myofibroblasts-like cells (ACTA2+).

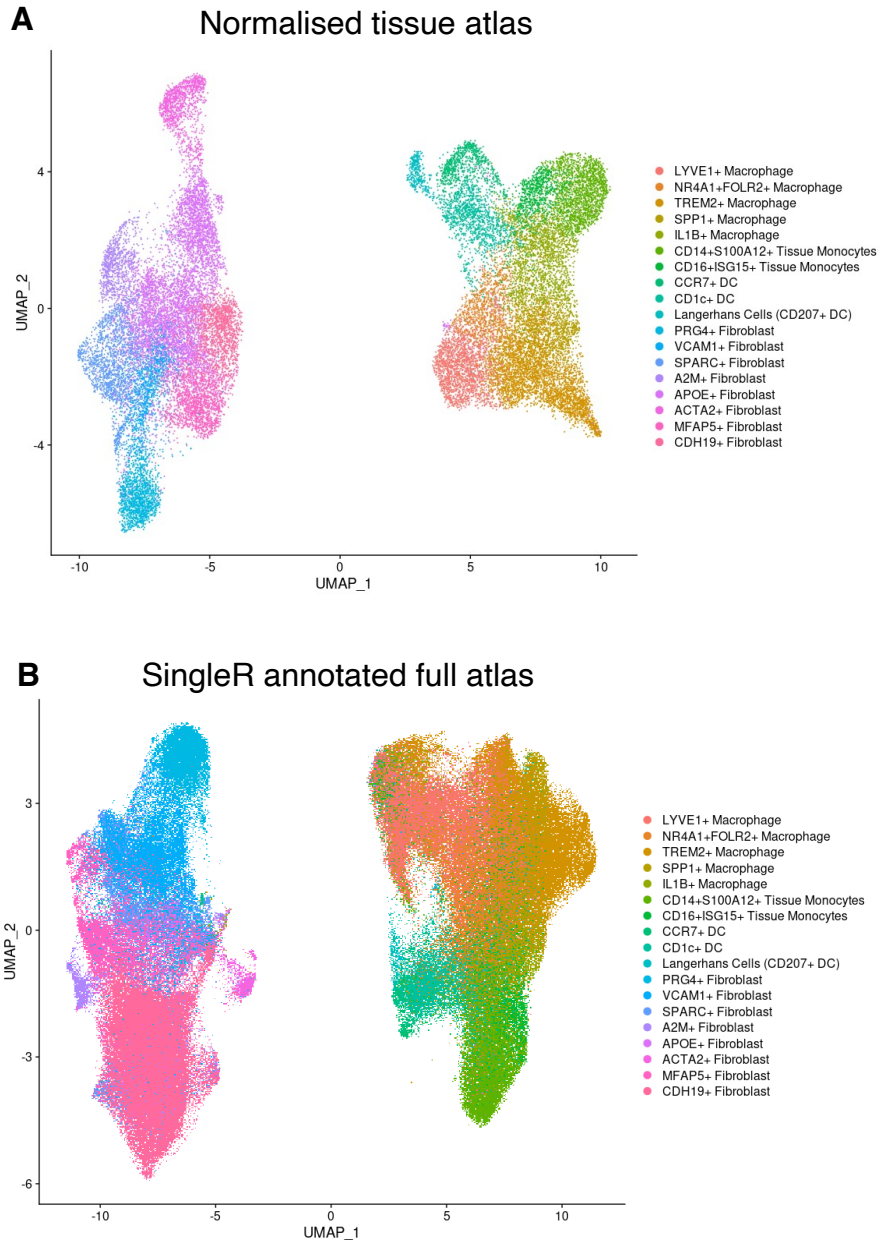


Figure 4.2: A) UMAP showing the normalised macrophage fibroblast (20,000 myeloid, 20,000 stromal cells) atlas split by macrophage and fibroblast subtypes. B) UMAP showing the full cell atlas using SingleR to annotate the rest of the data, split by macrophage and fibroblast subtypes (136,741 myeloid and 117,102 stromal cells).

In order to annotate the dataset consisting of 253,843 cells across all four tissues, I took the normalised annotated atlas Dr MacDonald had preprocessed and used SingleR for label transfer using the normalised atlas as a reference dataset. After obtaining the cellular annotation prediction labels I reclustered the full data at 0.3 resolution shown in Figure 4.2 B. The exact cell numbers for each cell type from the normalised atlas to the full annotated dataset with SingleR can be found in Table 4.1.

Cell type	Normalised cell atlas	SingleR full cell atlas
LYVE1+ Macrophage	1958	24378
NR4A1+FOLR2+ Macrophage	977	12638
TREM2+ Macrophage	3753	45099
SPP1+ Macrophage	1045	11065
IL1B+ Macrophage	2106	11423
CD14+S100A12+ Tissue Monocytes	1924	14201
CD16+ISG15+ Tissue Monocytes	825	3867
CD1c+ DC	1245	9170
Langerhans Cells (CD207+ DC)	332	951
CCR7+ DC	825	3474
SPARC+ Fibroblast	2080	10354
A2M+ Fibroblast	1112	3312
PRG4+ Fibroblast	1370	12389
APOE+ Fibroblast	5949	3367
MFAP5+ Fibroblast	1688	17576
CDH19+ Fibroblast	1411	53043
VCAM1+ Fibroblast	926	14873
ACTA2+ Fibroblast	1181	2663

Table 4.1: Cell counts for each cell type in the normalised atlas and full data atlas after SingleR label transfer.

Next, to ensure the proportion of cell types were still representative of the normalised atlas we plotted cell type proportions of the myeloid cells split across the tissues for both atlases (Figure 4.3).

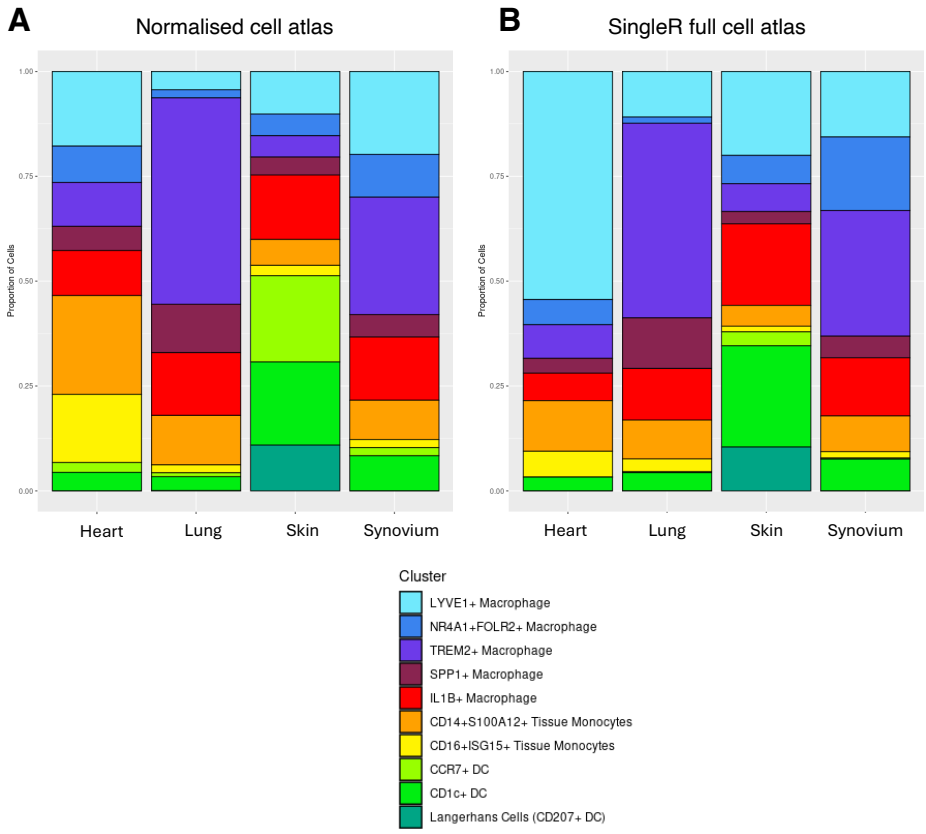


Figure 4.3: A) Stacked proportion barplot showing the relative proportion of macrophage subtypes across each tissue in the normalised macrophage fibroblast atlas split by tissue and coloured by cell type. B) Stacked proportion barplot showing the relative proportion of macrophage subtypes across each tissue in the full cell atlas annotated with SingleR split by tissue and coloured by cell type.

Analysis of myeloid cell clusters indicated initially in the normalised atlas that the heart had a high proportion of CD14+S100A12+ and CD16+ISG15+ monocyte-like precursors, however this decreased in the fully annotated atlas. Instead, we observed an expansion of LYVE1+ macrophages predominantly in the heart. The lung myeloid cell compartment was governed by resolving TREM2+ macrophages that remained unchanged in the fully annotated atlas. Alternatively, the skin myeloid atlas mostly comprised DC phenotypes including CCR7+ and an expanded CD1c+ population in the fully annotated dataset.

Also we saw a proportion of langerin-expressing (CD207+) DCs, also known as Langerhans cells, which were exclusive to the skin. In synovium, we found mixed myeloid fractions without tissue specific dominance of one cell cluster but a reduction in the proportion of CCR7+ DCs in the fully annotated dataset.

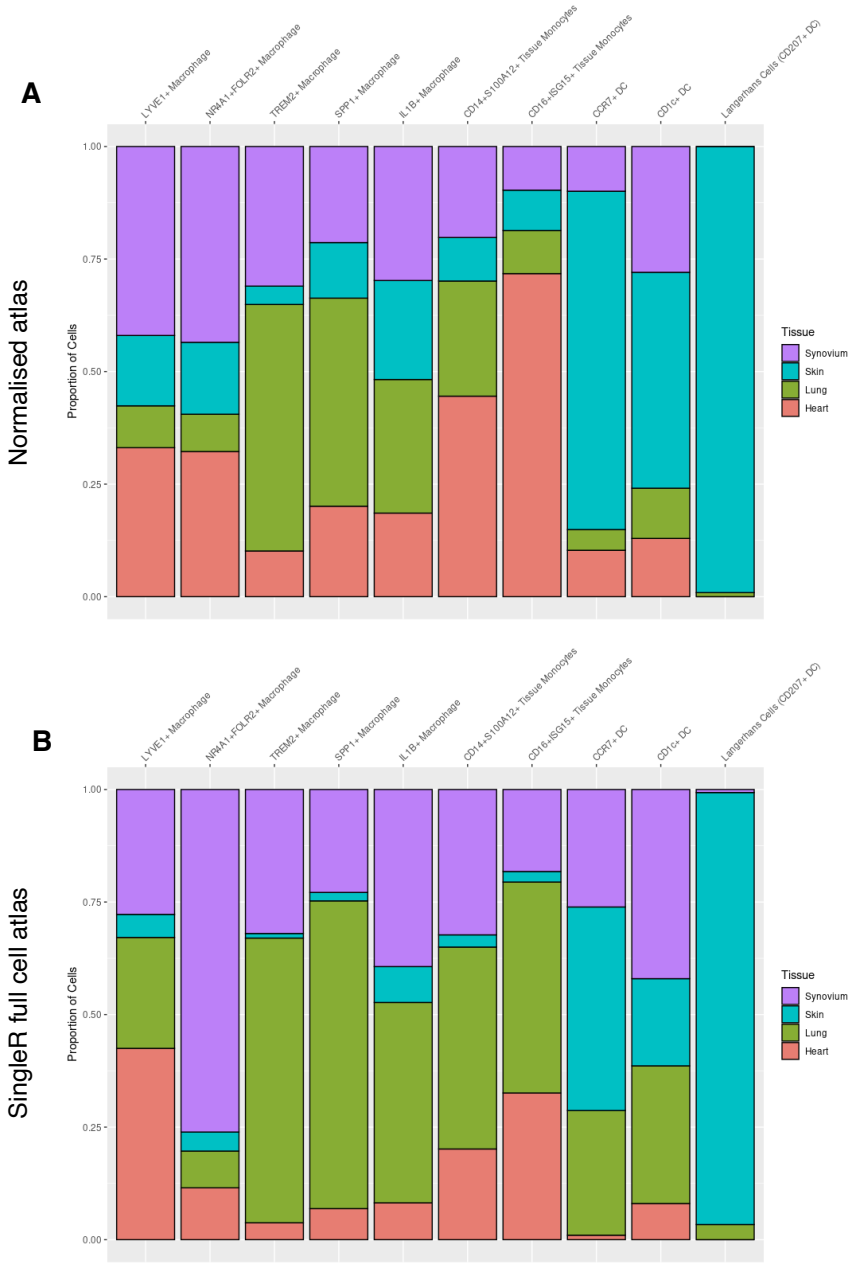


Figure 4.4: A) Stacked proportion barplot showing the relative proportion of macrophage subtypes across each tissue in the normalised macrophage fibroblast atlas split by cell type and coloured by tissue. B) Stacked proportion barplot showing the relative proportion of macrophage subtypes across each tissue in the full cell atlas annotated with SingleR split by cell type and coloured by tissue.

This myeloid cell distribution was also evident when we visualised the abundance of each identified myeloid cell population throughout the analysed tissues (Figure 4.4). In both atlases we see Langerhans (CD207+) DCs were exclusively found in the skin whereas the other myeloid cell types are more heterogenously distributed across the other tissues. LYVE1+ and NR4A1+ macrophages were mostly from the synovium and heart, whilst TREM2+ and SPP1+ phenotypes mainly originated from the lung. We can also see the reassignment of CD14+S100A12+ and CD16+ISG15+ monocyte-like precursors deriving from the heart in the fully annotated cell atlas compared to the normalised cell atlas with them mainly coming from the synovium and lung.

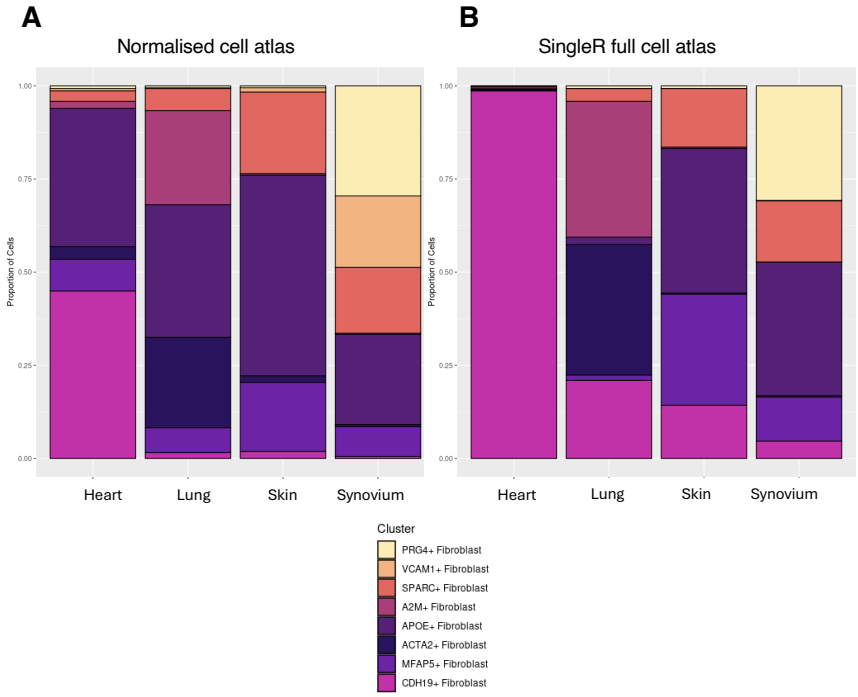


Figure 4.5: A) Stacked proportion barplot showing the relative proportion of fibroblast subtypes across each tissue in the normalised macrophage fibroblast atlas split by tissue and coloured by cell type. B) Stacked proportion barplot showing the relative proportion of fibroblast subtypes across each tissue in the full cell atlas annotated with SingleR split by tissue and coloured by cell type.

Similarly, we performed the same analysis to observe the distribution of stromal cells across the four tissues in both the normal and fully annotated cell atlas. Analysis of the relative proportion of stromal cell clusters across tissues revealed that the heart stromal compartment was predominantly composed of CDH19+ and APOE+ fibroblasts in the normalised cell atlas. However in the fully annotated cell atlas we found a huge expansion of this cell population and observed that CDH19+ fibroblasts were almost exclusively found in the heart (Figure 4.5). In the lung stromal atlas, mainly A2M+, APOE+ and ACTA2+, but also SPARC+ and CDH19+ fibroblasts were found. A2M+ and ACTA2+ fibroblasts were lung specific and were preserved across both atlases. The skin and synovial tissue shared the collagen expressing SPARC+ fibroblasts as well as a proportion of APOE+ and MFAP5+ populations with no skin-specific fibroblast population identified. Finally, the synovium contained the PRG4+ and VCAM1+ populations, which were exclusive to this tissue however this VCAM1+ population was diminished in the fully annotated atlas. In summary, our analysis identified populations of macrophages and fibroblasts that were common to multiple tissues whilst also identifying tissue-unique clusters. PRG4+/VCAM1+ synovial fibroblasts, ACTA2+ and A2M+ lung fibroblasts and CDH19+ heart fibroblasts were tissue-specific fibroblasts and Langerhans (CD207+) DC in the skin tissue specific myeloid cells. When looking at the cell distribution of each identified stromal cell population throughout the analysed tissues (Figure 4.6) the tissue-specificity is more apparent. The proportions between the two atlases remains largely unchanged apart from changes in the proportion of APOE+ and MFAP5+ fibroblasts labelled in the lung in the fully annotated dataset. After having determined common and exclusive stromal and myeloid subpopulations and their representation in the full dataset, we aimed to investigate changes in cell-cell interactions between fibroblasts and macrophages in disease states compared to healthy states. In general, tissue-specificity was more pronounced in the stromal than in the myeloid compartment.

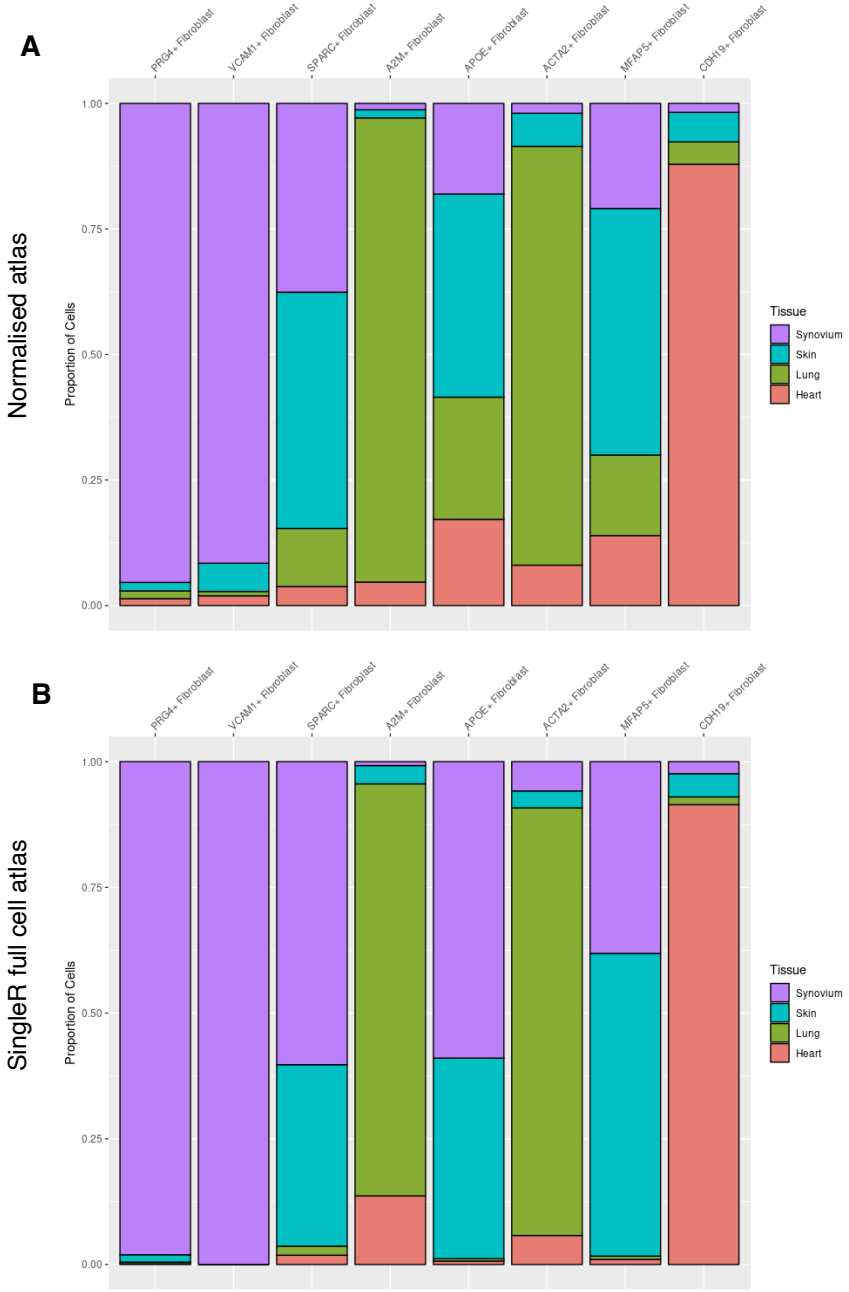


Figure 4.6: A) Stacked proportion barplot showing the relative proportion of fibroblast subtypes across each tissue in the normalised macrophage fibroblast atlas split by cell type and coloured by tissue. B) Stacked proportion barplot showing the relative proportion of fibroblast subtypes across each tissue in the full cell atlas annotated with SingleR split by cell type and coloured by tissue.

4.4.3 Comparing overlap of cellular interactions across the synovium, lung, skin and heart in homeostasis and disease

Once we had labelled the full atlas dataset we could proceed with cellular interaction inference to identify dominant drivers and receivers of communication in the atlas. Cellular inference was performed on the full cell atlas with the various inflammatory disease states pooled into a 'disease' group and was compared with interactions that maintain homeostasis in the healthy control group. We investigated how fibroblast–macrophage interactions vary across tissues in health and disease (Figure 4.7) using the CellChat package[146]. When fibroblasts were considered as senders, i.e. expressing a ligand, (Figure 4.7 A), we observed tissue-specific differences emerging. In the lung, fibroblasts showed weak interactions across myeloid subsets with A2M+ fibroblasts being key drivers in interactions with lung macrophage subsets. The most prominent effects were observed in the skin, where APOE+ and SPARC+ fibroblasts were found to be strongly interacting with IL1B+, NR4A1 FOLR2 and SPP1+ macrophages in inflammatory conditions. Similarly, in the synovium, fibroblasts broadly stimulated macrophages across several subsets, with PRG4+ fibroblasts exclusively showing interactions in the tissue highlighting their tissue specific role in the synovial niche. However in the heart, fibroblast signalling toward macrophages was weak, with no apparent signalling pattern between fibroblasts and macrophages in the tissue.

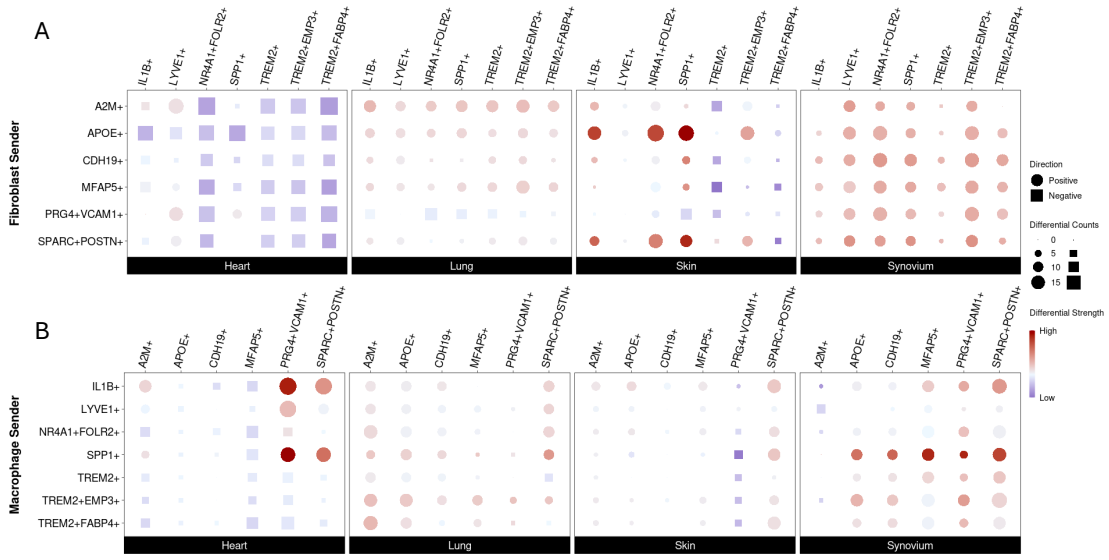


Figure 4.7: A) Bubble plot showing cell interactions where fibroblasts are the source of signal across heart, lung, skin and synovium. B) Bubble plot showing cell interactions where macrophages are the source of signal across heart, lung, skin and synovium. The size of each shape corresponds to the number of differential cell-cell interactions across homeostasis and disease. The cells are coloured by the strength of the interactions i.e. the sum of all the interaction weights. Circles represent a positive change in expression i.e. stronger interactions, squares represent a negative change in expression i.e. weaker interactions.

When macrophages were considered as senders (Figure 4.7 B), the heart showed strong interactions between IL1B+ and SPP1+ macrophages with PRG4+ fibroblasts which should only be exclusive to the synovium. In the lung, communication between macrophages and fibroblast subsets was weak, with TREM2+ macrophages driving communication in disease. In the skin, macrophages interacted with SPARC+ fibroblasts mainly by IL1B+ and SPP1+ macrophages. By contrast, in the synovium, macrophages showed stronger interactions across fibroblast subsets, with SPP1+, TREM2+ and IL1B macrophages interacting with PRG4+, MFAP5+ and SPARC+ fibroblasts.

Figure 4.7 suggested that there were patterns of similarity and difference in the cell types that were communicating across the tissues. We observed SPP1+ macrophages being key drivers across the tissues in particular the synovium and APOE+ and SPARC+ fibroblasts driving communication from the stromal compartment. To get an understanding about

the similarity or uniqueness of cellular interactions occurring in each tissue regardless of whether the sender is myeloid or stromal we wanted to look at the overlap of the interactions in each condition. This would give us a more global view about tissue specific interactions irrespective of cell type (Figure 4.8).

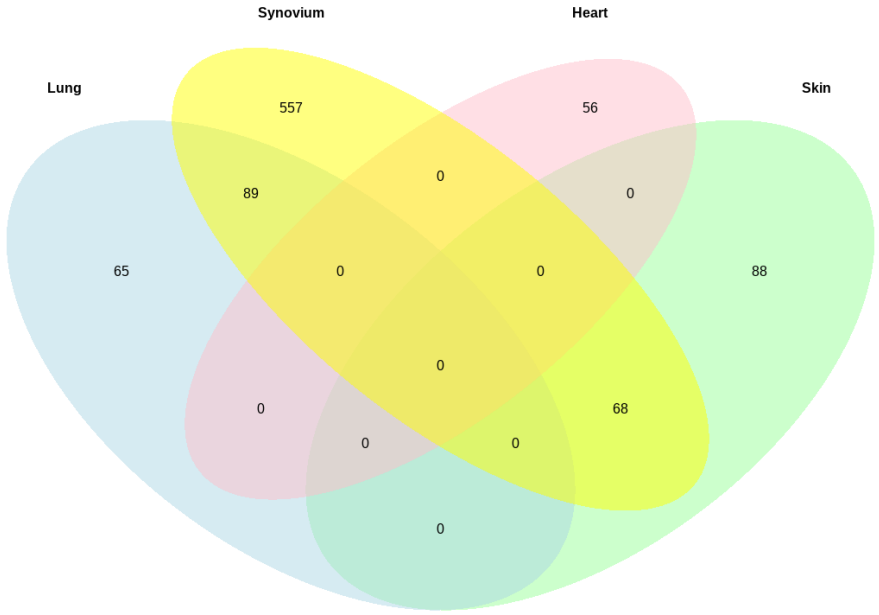


Figure 4.8: Venn diagram showing the total number of cell-cell interactions that are shared and distinct across tissues. Heart is coloured in red, lung in blue, skin in green and synovium in yellow.

We saw that the heart was the most unique tissue and shared no interactions in disease when compared to the other three tissues. The most similar tissues in disease were the lung and synovium which shared 89 interactions. Similarly, the synovium and skin shared 68 interactions in disease, however, shared no interactions with the lung. Overall, the synovium had the highest number of unique interactions when compared to the other tissues suggesting these interactions are specific to the tissue niche in disease. However, to be noted is the synovium contained significantly more cells that the other tissue datasets in disease that may have introduced a bias or skew in the number of interactions returned

with no filtering thresholds (Heart: 240 cells, Lung: 38044, Skin: 14707, Synovium: 63467). To gain a deeper understanding of what pathways these shared and unique interactions play in disease we examined the functional annotation of each overlap in the Upset plot in Figure 4.9.

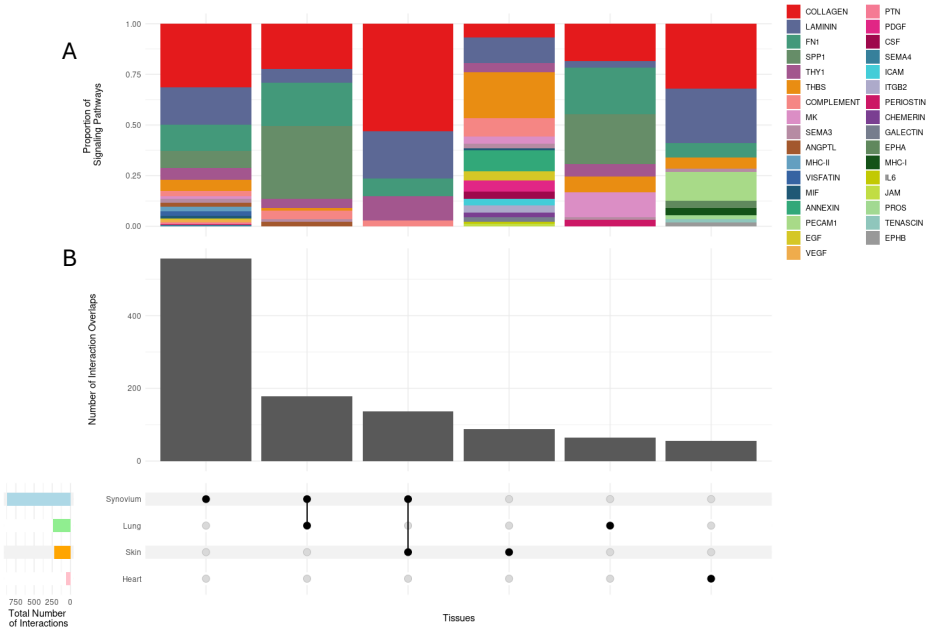


Figure 4.9: A) Stacked proportion barplot showing the relative proportion of cellular interaction pathways in each intersection of either overlapping or distinct cellular interactions between the heart, lung, skin and synovium. Segments are coloured by functional pathway B) Upset plot showing the number of overlapping or distinct interactions in each tissue comparison for the heart, lung, skin and synovium.

Analysis of the proportional contribution of signalling pathways across tissues revealed both conserved and tissue-specific patterns of fibroblast–macrophage communication (Figure 4.9 A). Collagen signalling was evident in every tissue, suggesting extracellular matrix remodelling and fibrosis as a conserved feature of stromal–immune crosstalk during inflammation. Shared interactions between the lung and synovium consisted of interactions in the SPP1 pathway and the FN1 signalling pathway whereas shared interactions between the synovium and skin were predominantly involved in the LAMININ and THY1 pathway. The unique interactions in the synovium shared similarities with pathways shared in the lung and synovium however showed an increase in MHC-II and VISFATIN signalling unique to the tissue. In the skin there was a marked increase in THBS signalling compared to the other tissues along with increase COMPLEMENT signalling and a unique PDGF

signalling absent in the lung, synovium and heart. Similarly, unique interactions in the lung fed into the MK signalling pathway and the PERIOSTIN pathway which was not observed in other tissues. Finally, the heart showed the least number of interactions in disease and demonstrated no overlap with the other tissues (Figure 4.9 B). Alongside collagen and laminin signalling, there was also evidence that suggested PECAM1 signalling played a role in heart inflammation.

To investigate the absence of interpretable signalling in the heart tissue when compared to other tissues, we examined the cell counts of each fine-grained cellular annotation in our cell atlas (Figure 4.10).

Synovium Cell Type Frequency

	Synovium Homeostasis	Synovium Disease
LYVE1+ Macrophage	1107	3634
LYVE1+CD163+ Macrophage	578	2026
LYVE1+MERTK+ Macrophage	7	51
NR4A1+FOLR2+ Macrophage	1826	6476
TREM2+ Macrophage	1818	3406
TREM2+FABP4+ Macrophage	407	648
TREM2+EMP3+ Macrophage	2740	5183
SPP1+ Macrophage	260	2189
IL1B+ Macrophage	95	509
CD14+C5aR1+ Macrophage	1037	4922
CD14+S100A12+ Tissue Monocytes	1078	2961
CD16+ISG15+ Tissue Monocytes	174	542
ISG15+ Macrophage	24	92
CCR7+ DC	7	124
CLEC10A+ Macrophage	1077	2505
Langerhans Cells (CD207+ DC)	0	5
PRG4+VCAM1+ Fibroblast	1178	8167
SPARC+POSTN+ Fibroblast	69	4849
SPARC+KRT1+ Fibroblast	0	0
SPARC+CALD1+ Fibroblast	99	1237
A2M+ Fibroblast	2	13
APOE+ Fibroblast	1297	9553
ACTA2+ Fibroblast	19	85
MFAP5+ Fibroblast	525	3052
CDH19+ Fibroblast	174	1238
TOTAL	15598	63467

Lung Cell Type Frequency

	Lung Homeostasis	Lung Disease
LYVE1+ Macrophage	1409	1952
LYVE1+CD163+ Macrophage	1132	2045
LYVE1+MERTK+ Macrophage	9	30
NR4A1+FOLR2+ Macrophage	377	514
TREM2+ Macrophage	6891	6571
TREM2+FABP4+ Macrophage	6905	7325
TREM2+EMP3+ Macrophage	130	261
SPP1+ Macrophage	1534	5786
IL1B+ Macrophage	450	494
CD14+C5aR1+ Macrophage	2567	3915
CD14+S100A12+ Tissue Monocytes	2742	2868
CD16+ISG15+ Tissue Monocytes	891	951
ISG15+ Macrophage	446	481
CCR7+ DC	47	92
CLEC10A+ Macrophage	892	1714
Langerhans Cells (CD207+ DC)	5	20
PRG4+VCAM1+ Fibroblast	16	16
SPARC+POSTN+ Fibroblast	92	52
SPARC+KRT1+ Fibroblast	0	1
SPARC+CALD1+ Fibroblast	208	184
A2M+ Fibroblast	582	997
APOE+ Fibroblast	49	38
ACTA2+ Fibroblast	165	1353
MFAP5+ Fibroblast	45	18
CDH19+ Fibroblast	550	366
TOTAL	28134	38044

Skin Cell Type Frequency

	Skin Homeostasis	Skin Disease
LYVE1+ Macrophage	159	358
LYVE1+CD163+ Macrophage	272	536
LYVE1+MERTK+ Macrophage	16	27
NR4A1+FOLR2+ Macrophage	181	280
TREM2+ Macrophage	273	88
TREM2+FABP4+ Macrophage	30	7
TREM2+EMP3+ Macrophage	24	32
SPP1+ Macrophage	32	169
IL1B+ Macrophage	510	445
CD14+C5aR1+ Macrophage	259	115
CD14+S100A12+ Tissue Monocytes	151	188
CD16+ISG15+ Tissue Monocytes	18	74
ISG15+ Macrophage	2	8
CCR7+ DC	64	163
CLEC10A+ Macrophage	794	857
Langerhans Cells (CD207+ DC)	195	520
PRG4+VCAM1+ Fibroblast	120	18
SPARC+POSTN+ Fibroblast	884	1446
SPARC+KRT1+ Fibroblast	67	546
SPARC+CALD1+ Fibroblast	2071	2315
A2M+ Fibroblast	38	32
APOE+ Fibroblast	4134	3211
ACTA2+ Fibroblast	35	25
MFAP5+ Fibroblast	3276	2365
CDH19+ Fibroblast	1848	882
TOTAL	15453	14707

?

Heart Cell Type Frequency

	Heart Homeostasis	Heart Disease
LYVE1+ Macrophage	2234	3
LYVE1+CD163+ Macrophage	5348	25
LYVE1+MERTK+ Macrophage	3732	1
NR4A1+FOLR2+ Macrophage	1256	1
TREM2+ Macrophage	1108	0
TREM2+FABP4+ Macrophage	94	0
TREM2+EMP3+ Macrophage	466	0
SPP1+ Macrophage	736	3
IL1B+ Macrophage	194	2
CD14+C5aR1+ Macrophage	1169	0
CD14+S100A12+ Tissue Monocytes	2518	2
CD16+ISG15+ Tissue Monocytes	1280	1
ISG15+ Macrophage	77	0
CCR7+ DC	4	1
CLEC10A+ Macrophage	683	2
Langerhans Cells (CD207+ DC)	0	0
PRG4+VCAM1+ Fibroblast	11	2
SPARC+POSTN+ Fibroblast	29	2
SPARC+KRT1+ Fibroblast	0	0
SPARC+CALD1+ Fibroblast	1315	16
A2M+ Fibroblast	253	10
APOE+ Fibroblast	124	0
ACTA2+ Fibroblast	92	11
MFAP5+ Fibroblast	94	1
CDH19+ Fibroblast	55383	157
TOTAL	78200	240

Cell types outlined in red have been excluded from further cell-cell interaction analysis where there must be at least 50 cells in each condition

Figure 4.10: Table showing cell type counts of each finely annotated cell type annotation split across the four tissues, heart, lung, skin and synovium. Cell type annotations that have less than 50 cells in the cluster are outlines in red and were insufficient for cellular communication inference.

By highlighting cell types that have less than 50 cells in the cluster split by condition, we realised that the heart disease dataset had insufficient cell numbers to complete a robust cellular inference analysis. Although there were representation of cell types in the healthy condition of the heart, the numbers were disproportionally low in comparison. Thus, moving forward the heart was excluded from the rest of the subsequent cellular inference analysis.

4.4.4 Focusing cellular inference on lung, skin and synovium in homeostasis and disease

As the disease heart datasets contained insufficient cell type annotation numbers to run cellular interaction inference as shown in Figure 4.10 it was removed from further analysis. To examine overlap of interactions with the heart excluded, we ran CellChat in two ways, first to find differentially expressed cellular interactions and second to find interactions based on their communication probability, a metric provided by CellChat (Figure 4.11). Differential interaction expression was applied to address the difference in total cell numbers across the tissues and was run with a ligand logfold change = 0.25 and p-value cut-off of 0.05. Interactions returned based on the communication probability alone were only subject to a p-value cut-off of 0.05. We can see that the results of the two thresholding strategies differ with the number of interactions increased when using the more permissive communication probability. When examining the overlap of shared and unique cell-cell interactions between the three tissues we found little overlap in the interactions involved in a state of homeostasis. However, in disease we observed a higher number of shared predicted interactions between the skin and synovium than in the lung. This could be down to the inclusion of inflammatory diseases such as psoriatic arthritis which display a complex plethora of phenotypes that affects both tissues.

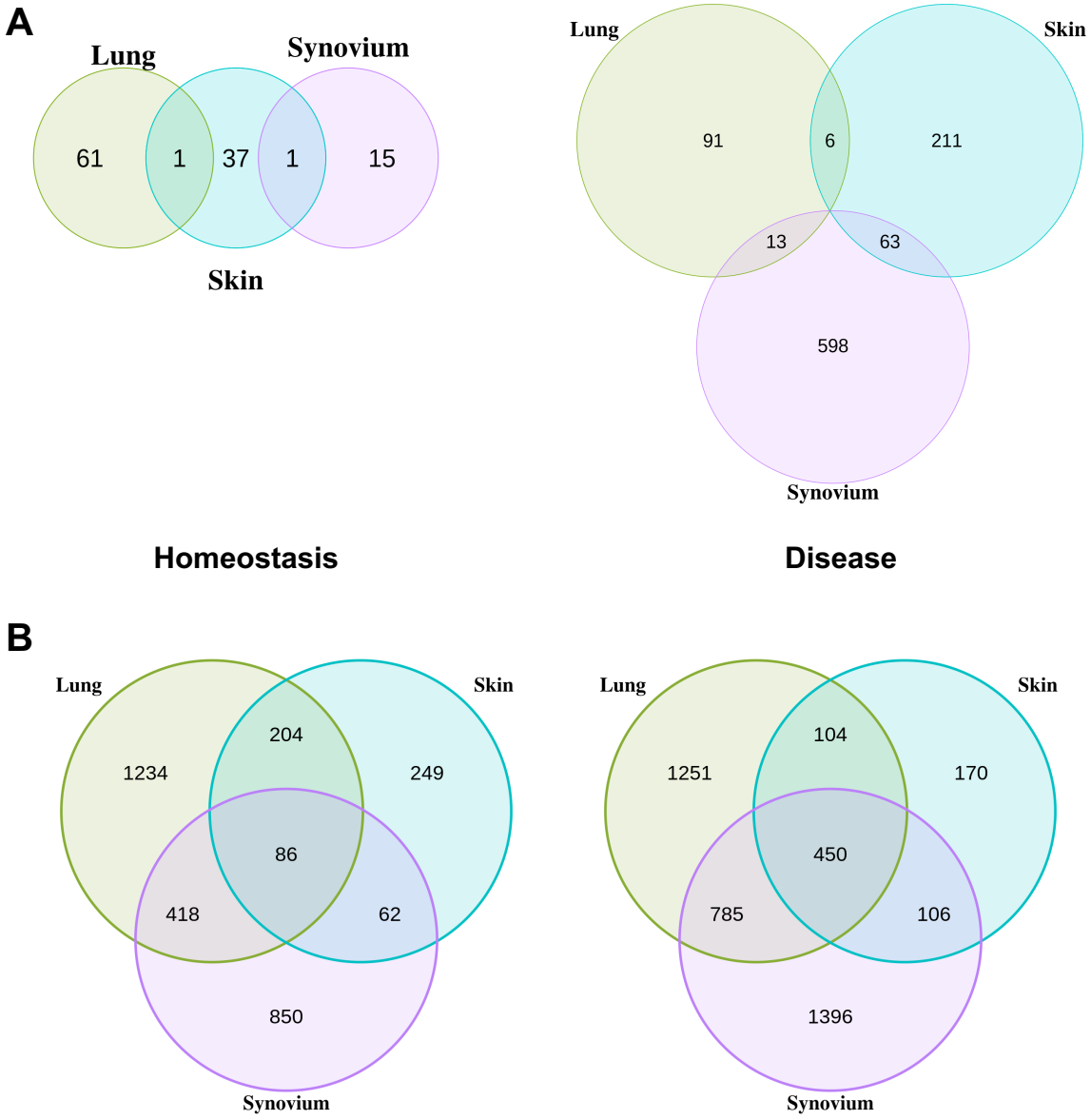


Figure 4.11: A) Venn diagram showing the number of differentially expressed cell-cell interactions that are shared and distinct across the synovium, lung and skin based on ligand and receptor log fold change of above 0.25 and a p-value threshold of 0.05. B) Venn diagram showing the number of cell-cell interactions that are shared and distinct across the synovium, lung and skin based on communication probability with a p-value threshold of 0.05.

After examining the global overlap of interactions between the three tissues we wanted to investigate which myeloid and stromal populations were the key drivers of interactions in homeostasis and disease. To do this we looked at the sum of the number and strength of the incoming interactions, which cell types were more active in receiving communications through receptor expression levels, and outgoing interactions of cell types that were more active at sending communications through ligand expression (Figure 4.12). In homeostasis across the myeloid compartment of the three tissues the activity of CD1c+ DCs was prominent at expressing ligand and receptor. However, in disease the myeloid landscape shifts with similarities and differences between organs. Most notably, the key drivers of inflammation in all tissues were SPP1+ macrophages activated by disease states when compared to homeostasis. The lung had the most drastic change in SPP1+ macrophage activity increasing in both number and ligand activity when compared to the skin and synovium. SPP1 producing macrophages have been reported in a wide range of diseases to increase macrophage polarisation and fibrosis in the lung[324,325] and is correlated with poor prognosis in lung cancer[326]. Furthermore, the lung had increased incoming activity by TREM2+ macrophages not observed in the other tissues. It has been reported that TREM2+ macrophages are pro-fibrotic and regulate alveolar macrophage survival[327]. In the lung and skin we observed increased signalling activity of NR4A1+ resolving macrophages in disease compared to inflammation with activity levels remaining unchanged in the synovium. In the skin and synovium, alongside SPP1+ macrophages we also saw activation of IL1B+ pro-inflammatory macrophages.

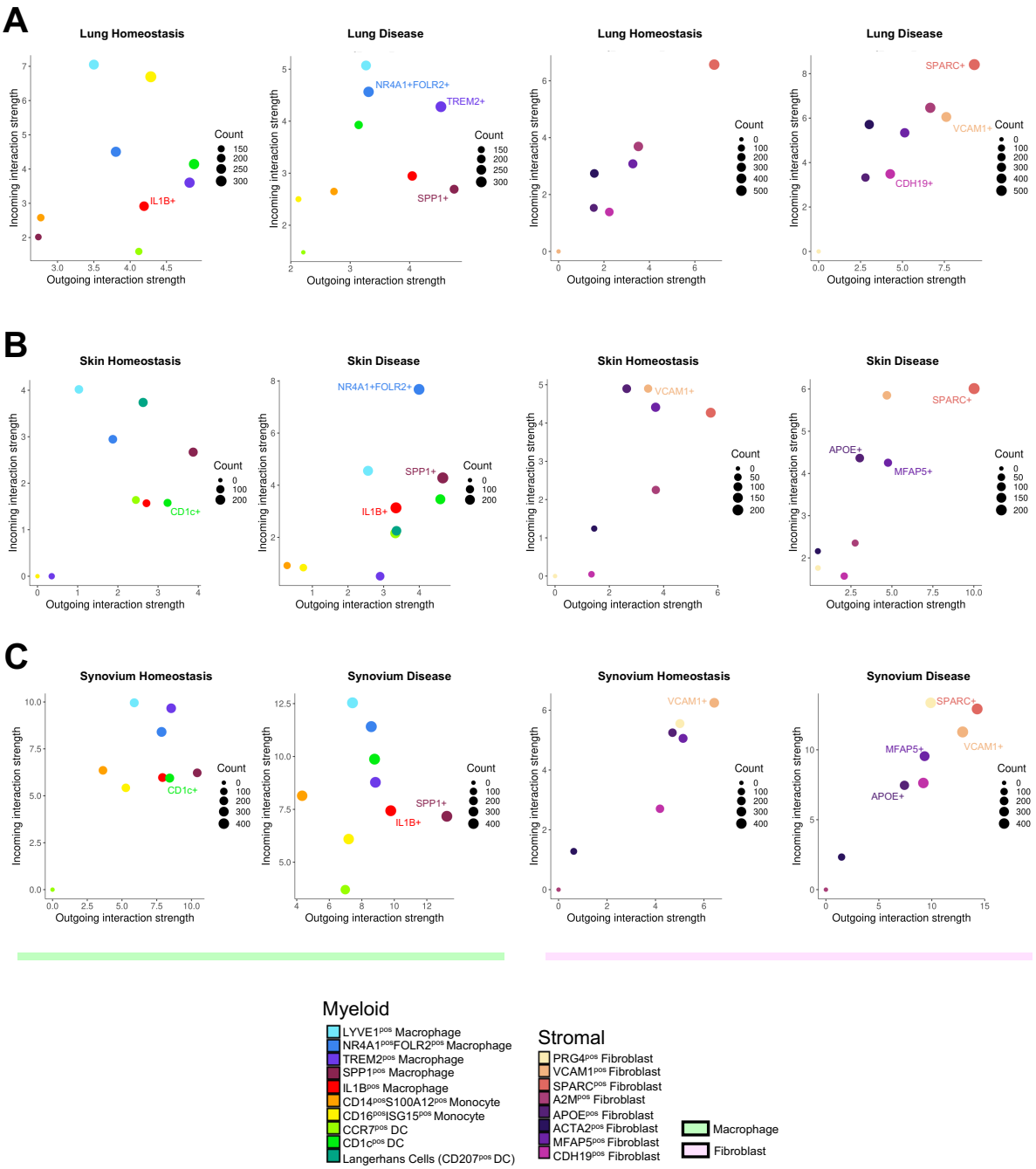


Figure 4.12: Dot plots show key drivers and receivers of communication with shared cell types labelled. Along the Y-axis of each dotplot shows the incoming interaction strength referring to the magnitude of ligands being received on the cell type. Along the x-axis of each dotplot shows the outgoing interaction strength referring to the magnitude of ligands being expressed on the cell type. Cell types that move along the Y-axis are receiving more communication and cell types that move along the X-axis are sending more communication. Note that the scale of each dotplot varies and is not uniform across all tissues so some magnitudes of change in communication are higher/lower.

In the stromal compartment we observed more organ-specific activation of cellular communication demonstrating tissue-specific stromal niches. SPARC+ fibroblasts emerged key senders in disease across all analysed tissues, known to sustain fibrosis and promote infiltration of lymphocytes to the site of inflammation[328] along with APOE+ fibroblasts, in the skin and synovium in particular. In the lung and synovium we found increase VCAM1+ fibroblast activity that have organ-specific phenotypes in inflammation with VCAM1 expression as a hallmark of RA on synovial fibroblasts, and as a cytokine induced factor on lung fibroblasts by TGF-β[329,330]. Finally, in the synovium and skin we see an increase of MFAP5+ fibroblasts both receiving and sending cellular communications. Through the activity patterns of key drivers of communication, these data suggest more pathogenic stromal-myeloid interactions in the synovium compared to lung and skin and highlight SPP1+ macrophages as disease relevant stromal activators and APOE+ fibroblasts as macrophage activators across tissues.

We then took cell types of interest based on the results of the key drivers of communication to investigate specific differentially expressed ligand receptor interactions in each tissue in disease (Figure 4.13). In the lung A2M+ fibroblasts expressed ligands associated with the collagen signalling pathway such as COL1A2, COL6A3 and COL4A1/2 that were interacting with CD44 on TREM2+ macrophages. These interactions recapitulate the pro-fibrotic function of TREM2+ macrophages in the lung[327]. SPP1+ macrophages were expressing SPP1 which was interacting with integrins expressed by MFAP5+ fibroblasts. Another notable interaction was CXCL12 on MFAP5+ fibroblasts interacting with CXCR4 on CD14+S100A12+ macrophages. This interaction is pro-inflammatory that promotes chemotaxis of immune cells to the site of inflammation and has been reported in autoinflammatory diseases such as osteoarthritis[331]. In the skin we oberved numerous collagen associated ligands (COL1A1/2, COL6A1/2/3) interacting with CD44 on SPP1+ macrophages. Similar to the lung this indicates tissue remodelling and fibrosis in skin inflammation. This is further supported by the expression of thrombospondins (THBS1/2) interacting with SPP1+ macrophages and cDC1s which promotes fibroblast migration, wound healing and tissue repair[332].

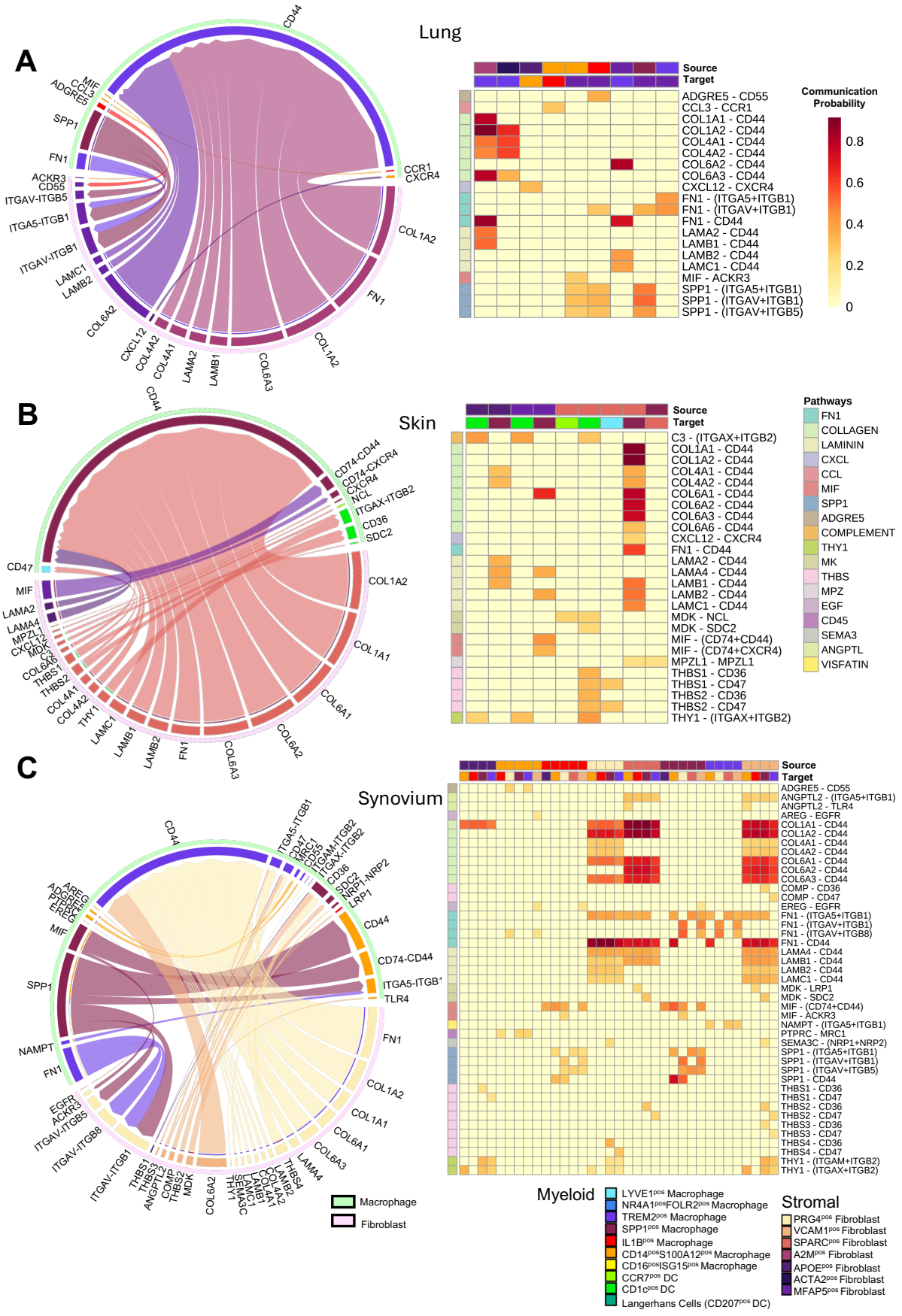


Figure 4.13: Circos plots show ligand receptor interactions that are differentially expressed in at least 20% of cells in disease, at a ligand logfold change cut-off of 0.25, p-value cut-off of 0.05. Heatmaps show the communication probability of interactions plotted in the circos plots with functional pathway annotation plotted on the side and source/target cell types along the top. A) Lung, B) Skin, C) Synovium.

Lastly, in the synovium we find expression of collagen and thrombospondin signalling as with the other tissues a shared signature of fibrosis and stromal driven inflammation. We also see SPP1+ macrophages interacting with PRG4+ lubricin producing fibroblasts, a synovial specific stromal population.

In homeostasis a prominent pathway seen across all three tissues is the collagen pathway expressed by different fibroblast subsets in each niche (Figure 4.14). In the lung this is predominantly driven by A2M+ fibroblasts to TREM2+ macrophages, in the skin it is driven by SPARC+ fibroblasts to SPP1+ macrophages and in the synovium PRG4+ fibroblasts to TREM2+ macrophages. This could be down to normal collagen processes at homeostasis that are vital for ECM organisation and maintenance in addition to normal tissue remodelling[333]. This is further supported by thrombospondin signalling seen in the skin and synovium that promotes matrix homeostasis and indirectly influences normal collagen production feeding into ECM organisation[334]. Another shared interaction that occurs across all tissues are pathways associated with cell migration such as MIF interacting with CXCR4 that promotes immune cell trafficking. Under homeostatic conditions this suggests a replenishment of immune cells at the tissue niche, or facilitation of patrolling cell types like dendritic cells. In the synovium we see a large amount of HLA-related ligands signalling to other myeloid cells that is absent in the other tissues. This could be because of the inclusion of patients of RA in remission being assigned to the 'Healthy' group that may be driving this phenotype.

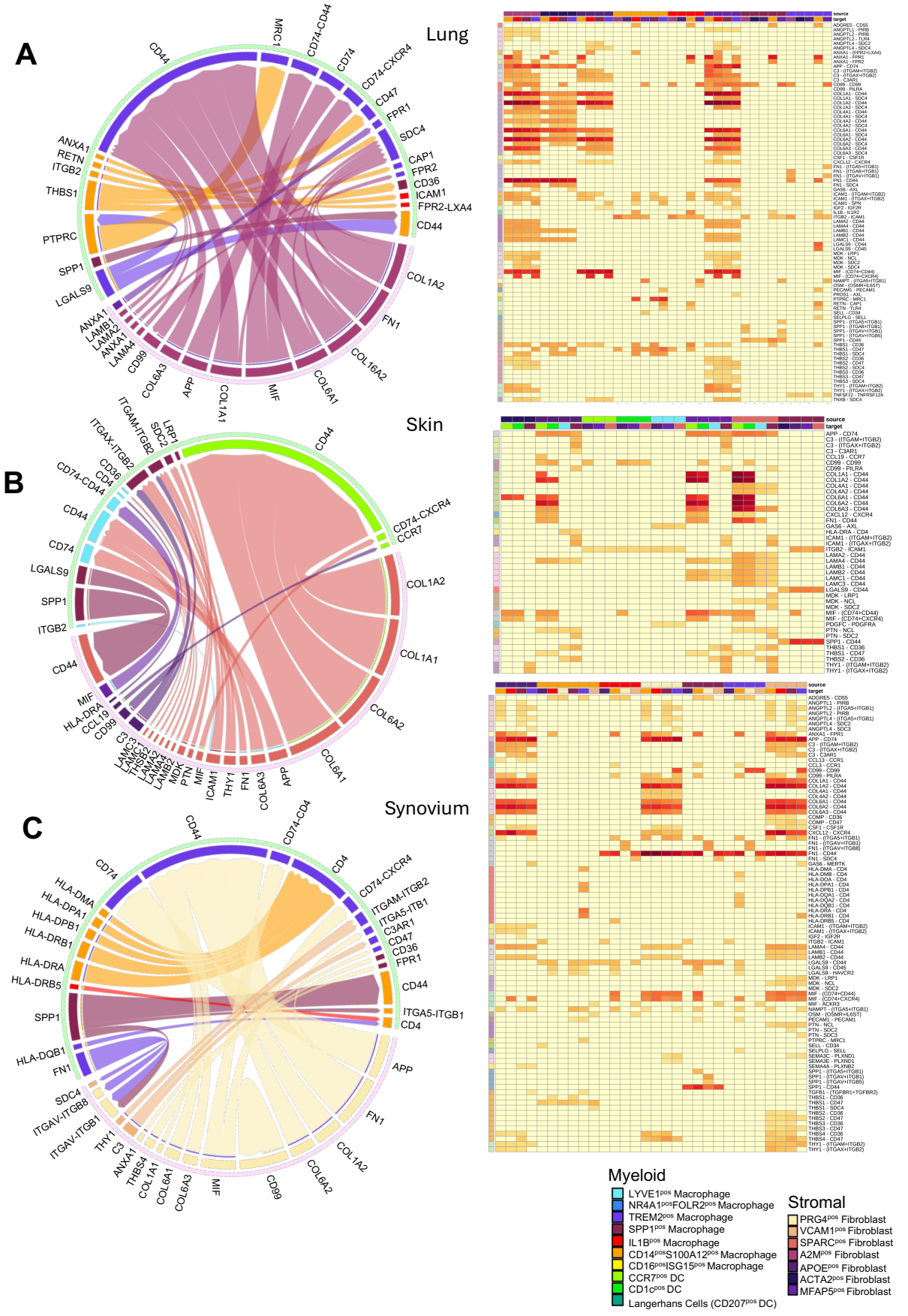


Figure 4.14: Circos plots show ligand receptor interactions that are differentially expressed in at least 20% of cells in homeostasis, at a ligand logfold change cut-off of 0.25, p-value cut-off of 0.05. Heatmaps show the communication probability of interactions plotted in the circos plots with functional pathway annotation plotted on the side and source/target cell types along the top. A) Lung, B) Skin, C) Synovium.

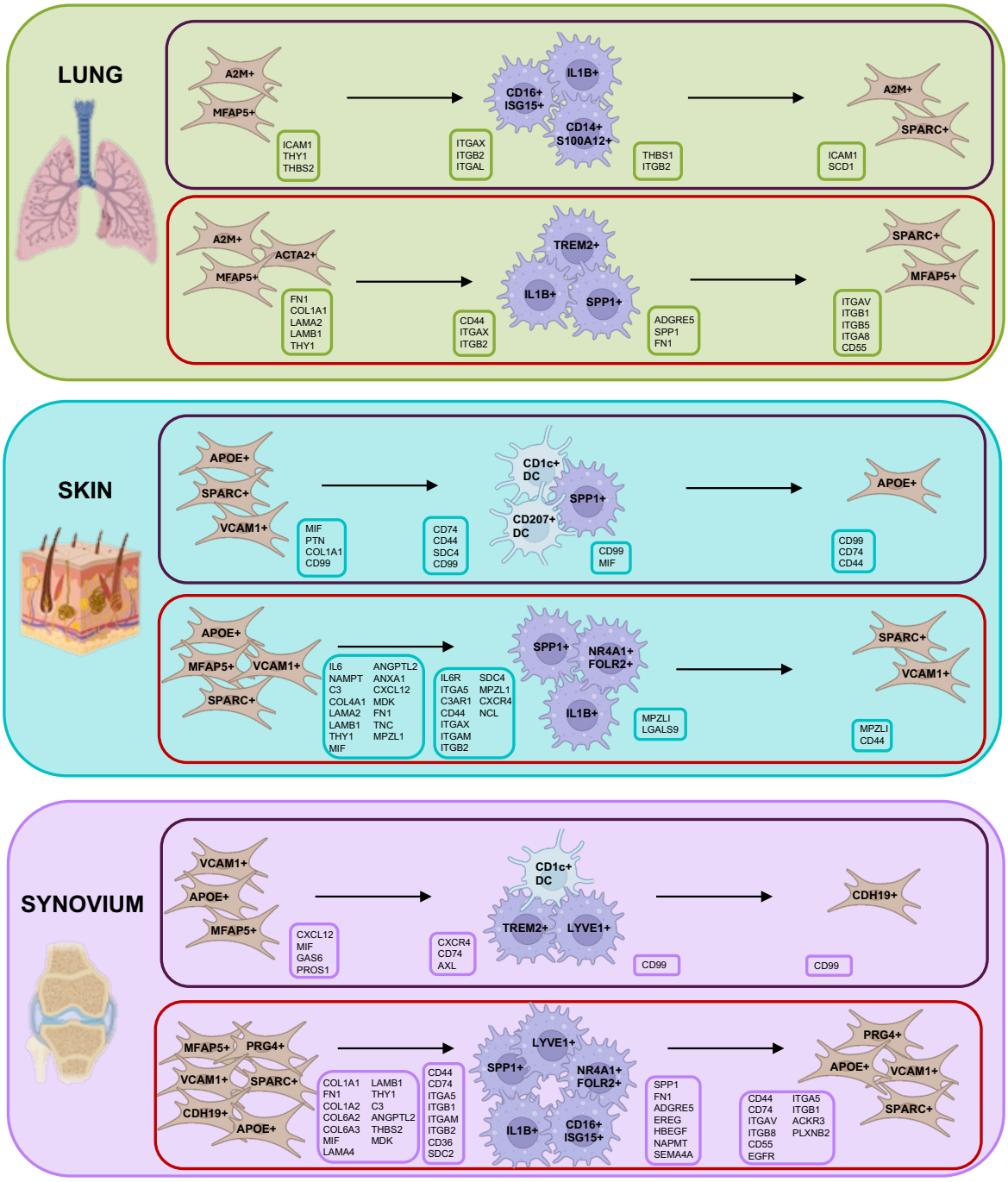


Figure 4.15: Graphical overview of the results of the Cellchat analysis of ligand receptor interactions that are differentially expressed in at least 20% of cells, at a ligand logfold change cut-off of 0.25, p-value cut-off of 0.05 in homeostasis (blue outline) and disease (red outline). Identified key drivers of communications are labelled and arrows show directionality of the interaction. The top ligand and receptors are highlighted in each tissue ordered by communication probability i.e., the likelihood of the interaction occurring according to gene expression. The key interactions in the lung are in the green box, skin in the blue box, synovium in the purple box. Cell type icons are taken from BioRender and the figure was created in PowerPoint.

To summarise in more detail the global landscape of additional interactions we identified in the the analysis, the key findings are detailed in Figure 4.15. In the lung during homeostasis, we found that the key drivers of signalling were A2M+ and MFAP5+ fibroblasts mostly interacting with integrin complexes on CD14+S100A12+ and CD16+ISG15+ macrophages. IL1B+ macrophages were found to interact with SPARC+ fibroblasts expressing THBS1 involved in extracellular matrix processes. However, in a disease state we observed an increase of collagen related genes expressed by various stromal subsets, including the lung specific ACTA2+ fibroblasts. These interactions occurred exclusively with CD44 expressed on TREM2+ macrophages possibly reflecting matrix remodelling in lung fibrosis. Interestingly, SPP1+ macrophages were the main source of communication out of the myeloid subsets expressing SPP1 that interacts with integrin complexes on SPARC+ and MFAP5+ fibroblasts. The landscape of interactions in the skin during homeostasis involves APOE+, VCAM1+ and SPARC+ fibroblasts driving interactions with CD1c+ DCs and Langerhans cells (CD207+ DCs). In the myeloid cells the CD1c+ DCs and SPP1+ macrophages interact through MIF-CD74/CD44 complex with APOE+ fibroblasts. In disease, similarly to the predicted interactions in the lung, stromal subsets expressed various collagens that were received by CD44 expressed on ILB1+, NR4A1+FOLR2+ and SPP1+ macrophages. SPP1+ macrophages in skin disease states are predicted to be involved in the laminin pathway interacting with APOE+ fibroblasts, however unlike their role in the lung they are key receivers of communication from the stromal compartment. In the synovium we have all fibroblast subsets interacting with CD1c+ DCs through a CXCL12-CXCR4 interaction in homeostasis. We also have the APOE+ and VCAM1+ fibroblasts interacting with dendritic cells through the PROS1-AXL whereas the MFAP5+ fibroblasts send GAS6 to interact with AXL. In a disease state the synovium showed the highest number of predicted interactions that are differentially expressed compared to homeostasis. LYVE1+, N4RA1+ FOLR2+, and TREM2+ macrophages interact with SPARC+ and VCAM1+ fibroblasts expressing NAMPT-ITGA5/ITGB1. They also interact with PRG4+ and VCAM1+ fibroblasts expressing FN1 that is received by the ITGAV/ITGB8 complex, this is also shared in TREM2+ macs, SPP1+. Many interactions here between macrophage subsets and fibroblasts are with various integrin complexes. These ligands sent out by the macrophages vary, SPP1 is expressed by SPP1+, NR4A1+ and ILB1+

macrophages that are received by PRG4+, VCAM1+ and SPARC+ fibroblasts. APOE+ fibroblasts and ILB1+SPP1+ macs have a MIF-CD74/CD44 complex interaction. SPP1+ and IL1B+ macs also express EREG received by EGFR on VCAM1+ fibroblasts only. CDH19+ fibroblasts interact with various macrophage subsets expressing THY1 that are received by the same ITGAM-ITGB2 integrin complex across all macrophage subsets. They also express COMP that interacts with CD36 and CD47 on SPP1+ macrophages and on LYVE1+ TREM2+ macrophages respectively. The synovium returned various interactions that involve tissue-specific cell types such as PRG4+ fibroblasts that expressed SEMA3C that is received by the NRP1-NRP2 complex on LYVE+ and SPP1+ macrophages. PRG4+ and APOE+ fibroblasts also express HLA-DRB1 that is received by CD4 on NR4A1+FOLR2+ macrophages.

4.4.5 Validating interactions in skin and synovium

We then set out to identify the specific locations of the dominant, common fibroblast populations within the skin and synovium and to visualise the identified cell-cell interactions. Our analysis identified the shared role of SPARC+ and APOE+ fibroblasts driving communication across the lung, skin and synovium in disease along with SPP1+ macrophages being a hallmark cell type of inflammation. As the analysis with CellChat is only a prediction of cellular interactions based on gene expression of each cell type cluster in the atlas we wanted to experimentally validate whether these cell types co-localise in diseased tissue. SPARC+ fibroblasts were a common driver of interactions across all three tissues in our analysis (Figure 4.12) we wanted to stain for the presence of this population in healthy and inflamed tissue. To identify SPARC+ fibroblasts in the tissue, we used periostin (POSTN), which appeared as prominent cell-specific marker gene for this population. In addition to SPARC+ fibroblasts we also wanted to determine the presence of APOE+ fibroblasts within tissues, thus we stained for CXCL12. CXCL12, also called stromal cell-derived factor 1, is highly expressed by APOE+ fibroblasts and is the major chemokine secreted by fibroblasts. Finally, we stained Microfibrillar-associated protein

5 (MFAP5) to detect the MFAP5+ subpopulation that was also a prominent driver of expression in the skin and synovium as suggested by the cellular interaction analysis. In skin and synovium, periostin and MFAP5 expression was confined to distinct and remotely exclusive compartments (Figure 4.16). In the skin, periostin was localized in the papillary dermis, while MFAP5 was expressed in the reticular dermis. In the synovium, periostin expression was found mainly in close association to blood vessels, while MFAP5 was detected throughout the remaining sublining area. This distinct pattern of distribution was maintained in psoriatic skin and RA synovial tissue. Thus, expression of these marker proteins supported the presence of two distinct fibroblast subpopulations, creating specialised tissue niches in skin as well as synovium in health and disease. We also wanted to validate our interactions between SPP1+ macrophages as they were identified as key receivers of interactions in disease across all tissues. For this we performed a double stain of CXCL12 indicating the APOE+ fibroblasts and SPP1 for the macrophages and found indeed that in psoriatic skin and RA inflammed synovium we found co-localisation of the stains. This suggests that indeed there is likely cellular communication occurring between APOE+ fibroblasts and SPP1+ macrophages that are driving inflammation in these tissues, validating predicted interacting cell types in the single cell analysis.

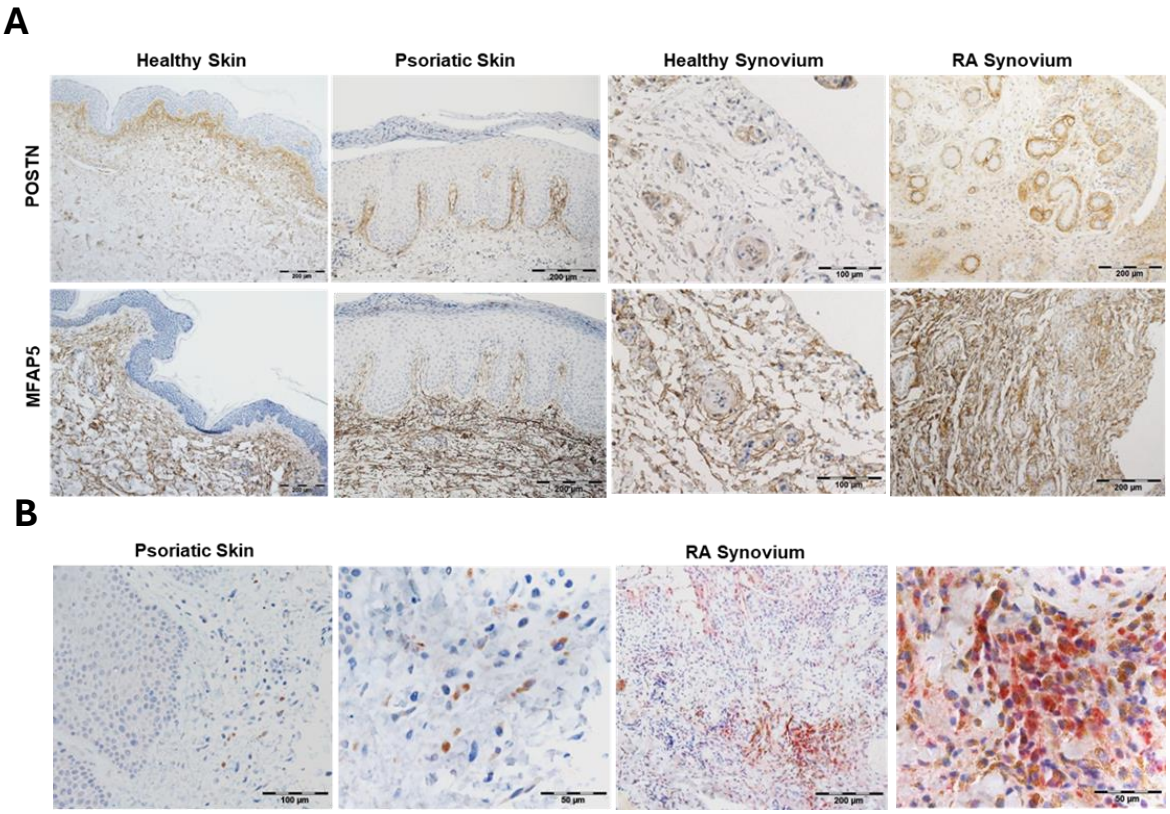


Figure 4.16: A) Immunohistochemical staining of POSTN and MFAP5 in healthy skin, psoriatic skin, healthy synovium and RA synovium. Magnification 100x for skin and RA synovium, 200x for healthy synovium. B) Double staining of CXCL12 (brown) and SPP1 (red) in psoriatic skin and RA synoivum. Brown arrows indicate CXCL12+ fibroblasts whereas red cells point out SPP1+ macrophages. Magnification 200x and 400x.

These results show that there are shared mechanisms of cellular communication across different tissue niches in homeostasis and disease, inferred from atlas-level data with complex co-variates. Understanding these shared and unique pathways can further aid our understanding of the intricate communication between tissue resident cells such as macrophages and fibroblasts and highlight the importance of cellular communication inference in single cell data. We used low-plex tissue staining to identify co-localised cell types that were predicted in the cellular interaction analysis to drive inflammation however we still cannot ascertain the validation of the ligand-interaction pairs that were reported from the

study. Thus, using newer techniques such as spatial transcriptomics where we have the whole transcriptome spatially resolved, provides us a new advent of cellular communication analysis where we can project not only cell types but also ligand-receptor pairs into space to orthogonally validate predicted interactions.

4.4.6 Section 2: Identifying cellular interactions leveraging spatial transcriptomics during *H. polygyrus* infection

This work was a collaborative effort at the University of Glasgow where the infection and Visium preparation was completed by Dr. Marta Campillo under the supervision of Professor Rick Maizels. Dataset analysis, processing, integration and differential expression analysis was jointly carried out by myself and Dr. Ross Laidlaw under the supervision of Professor Thomas Otto. Downstream analysis of cellular interactions were performed by myself and all results shown in this section were analysed independently. The complete study and details of the study can be read in our preprint here (https://www.biorxiv.org/content/10.1101/2024.02.09.579622v1) currently under review by Nature Communications, which details the comparative analysis of naive against day 7 post infection (with additional time points day 3 and day 5 added after the Biorxiv paper). This section will expand the preprint and detail analysis of additional time points and associated functional pathways over the time course of the infection.

4.4.7 Integration of *H. polygyrus* Visium datasets across time

When infective larvae of *H. polygyrus* are ingested and migrate to the small intestine, they cross the epithelial barrier and reside in the submucosa for 8 days before returning to the lumen[29]. To profile the transcriptomic landscape of the small intestine tissue during this process, we employed the Visium (10X Genomics) platform v1 to conduct spatial

transcriptomics on formalin-fixed gut rolls. Here we had full transcriptome sequencing and mapped the spatial data against the mouse genome as detailed in the methods. We had four time points during the infection with each gut roll being extracted from a separate mouse, at naive steady-state in the absence of the parasite, 3 days post-infection with *H. polygyrus*, 5 days post-infection and 7 days post-infection. Initially, we attempted to integrate the data using Harmony[101] using each time point as a co-variate which, after clustering, resulted in 9 clusters (Figure 4.17). However, when we projected these leiden clusters back to the spatial coordinates they were not representative of the underlying architecture of the tissue. As a result, this made it difficult to compare clusters from the integration across the time points as they failed to harmonise into coherent spatial niches that were comparable across the different time points. To resolve this, and to define spatial niches across the time points for comparision, Dr. Marta Campillo manually annotated using the Loupe Browser (10X Genomics), three factors that encapsulated the fundamental structure of the small intestine, the crypt zone (including the lamina propria and the crypts), the villi and the granuloma (only present in *H. polygyrus* infected mice). These annotations then represented spatial clusters that were biologically relevant and shared across time points that we could draw comparisons from. Thus, we used these histological annotations rather than the clustering from the integration analysis as this allowed us to compare distinct spatial regions across the tissue rather than incoherent clustering we generated using the leiden clustering workflow.

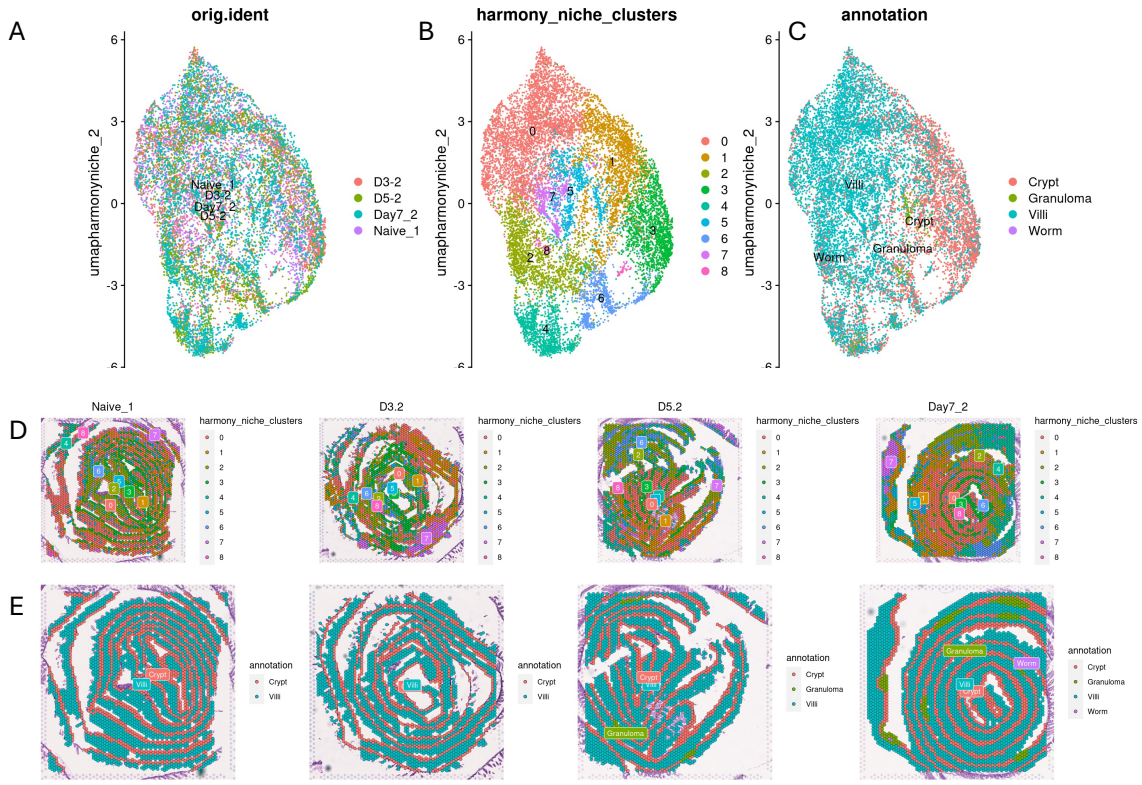


Figure 4.17: A) UMAP visualisation of the Harmony integration of naïve, day 3, day 5 and 7 days post H. polygyrus infection Visium datasets coloured by timepoint. B) UMAP visualisation of the Harmony integration of naïve, day 3, day 5 and 7 days post H. polygyrus infection Visium datasets coloured by leiden clusters at a resolution 0.5. C) UMAP visualisation of the Harmony integration of naïve, day 3, day 5 and 7 days post H. polygyrus infection Visium datasets coloured by manual histological annotation. D) Spatial plot of leiden clusters projected back to the spatial coordinates for each timepoint. E) Spatial plot of manual histological annotations projected back to the spatial coordinates for each timepoint.

4.4.8 Differential expression analysis reveals temporal gene expression patterns over infection

We next investigated temporal shifts in signalling pathways, particularly within the crypt microenvironment at days 3, 5 and 7 following *H. polygyrus* infection. First, we wanted to understand what genes were being differentially expressed at each time point, which revealed a striking temporal pattern of gene expressions on days 3 and 7 compared to naïve and day 5 (Figure 4.18). This shift matches the parasite's life cycle events: around

days 3 and 7 post-infection we have a epithelial barrier breach by the parasite and tissue penetration occurs, while naïve and day 5 represent the steady state and the period while the parasite is encapsulated inside the granuloma. We also wanted to examine the degree of overlap of differentially expressed genes between the time points and found that the naive and day 5 post-infection time points shared 951 genes with almost overlap with the other two time points. Similarly, day 3 and day 7 shared 317 differentially expressed genes recapitulating the life cycle of the parasite over time. We observe s stark temporal pattern over the time course of the infection that reflects the different stages of the parasite larvae maturation. At 3 days post infection where the larvae burrows into the epithelial wall, we see a pattern of upregulated gene expression that shares a similar profile to 7 days post infection where the mature parasite breaks out from the granuloma. Remarkably, at 5 days post infection we see an expression profile akin to the steady-state uninfected gut suggesting a mechanism of immunoregulation and suppression at the host-parasite interface.

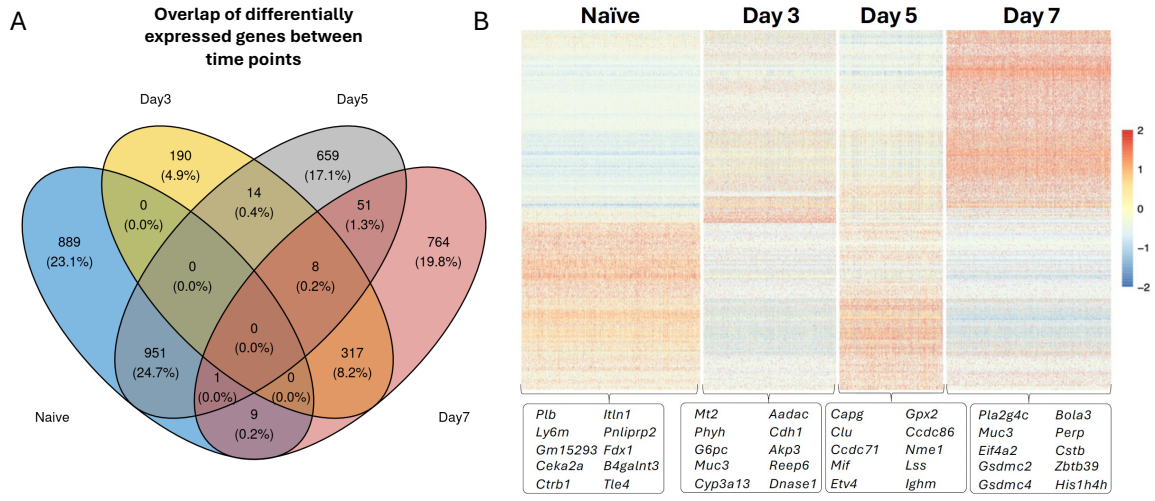


Figure 4.18: A) Venn diagram showing the overlap of up/down differentially expressed genes across each timepoint, logfold change cutoff of 0.5, p-value of 0.05. B) Heatmap of gene expression in the crypt across all timepoints. The top 10 most highly expressed genes are labelled ordered by their average log fold change values.

4.4.9 Cellular interaction analysis shows decreased Wnt signalling during *H. polygyrus* infection

To understand globally how cellular interactions were changing over time, cellular interaction inference was carried out on each time point irrespective of tissue location or cell type. Functional annotations of these interactions were then examined to identify key functional gene sets that are influenced during infection (Figure 4.19). Global immune cell genes associated with CD45 became increasingly prominent over the course of infection, coinciding with the immune cell influx characteristic of early granuloma formation, where monocytes, neutrophils, and eosinophils begin surrounding the parasite 5, 17. Notably, increased expression of CCL pathway (Ccl6, Ccl7, Ccl8) was observed at later time points, suggesting continued recruitement of immune cells to the site of the parasite. By day 7, additional signalling pathways associated with tissue remodelling became dominant, including TGF-β (Tgfb1), osteopontin (Spp1), and thrombospondins. These factors have been implicated in fibrosis and extracellular matrix remodelling in chronic helminth infections, potentially facilitating wound healing. In contrast, we observed a complete loss of Wnt pathway genes from the expression profile by day 7 post-infection. In naïve mice, crypt-to-crypt interactions were dominated by Wnt signalling, consistent with its role in maintaining intestinal stem cell proliferation and epithelial homeostasis[304]. When we examined individual Wnt pathway members, a marked reduction in their expression levels was already evident by day 3. By day 7 post-infection, Wnt gene expression was almost completely absent, while Notch signalling was strongly upregulated which mirrors findings from *H. polygyrus*-infected organoid cultures, where parasite-secreted products favour foetal-like repair phenotypes[305].

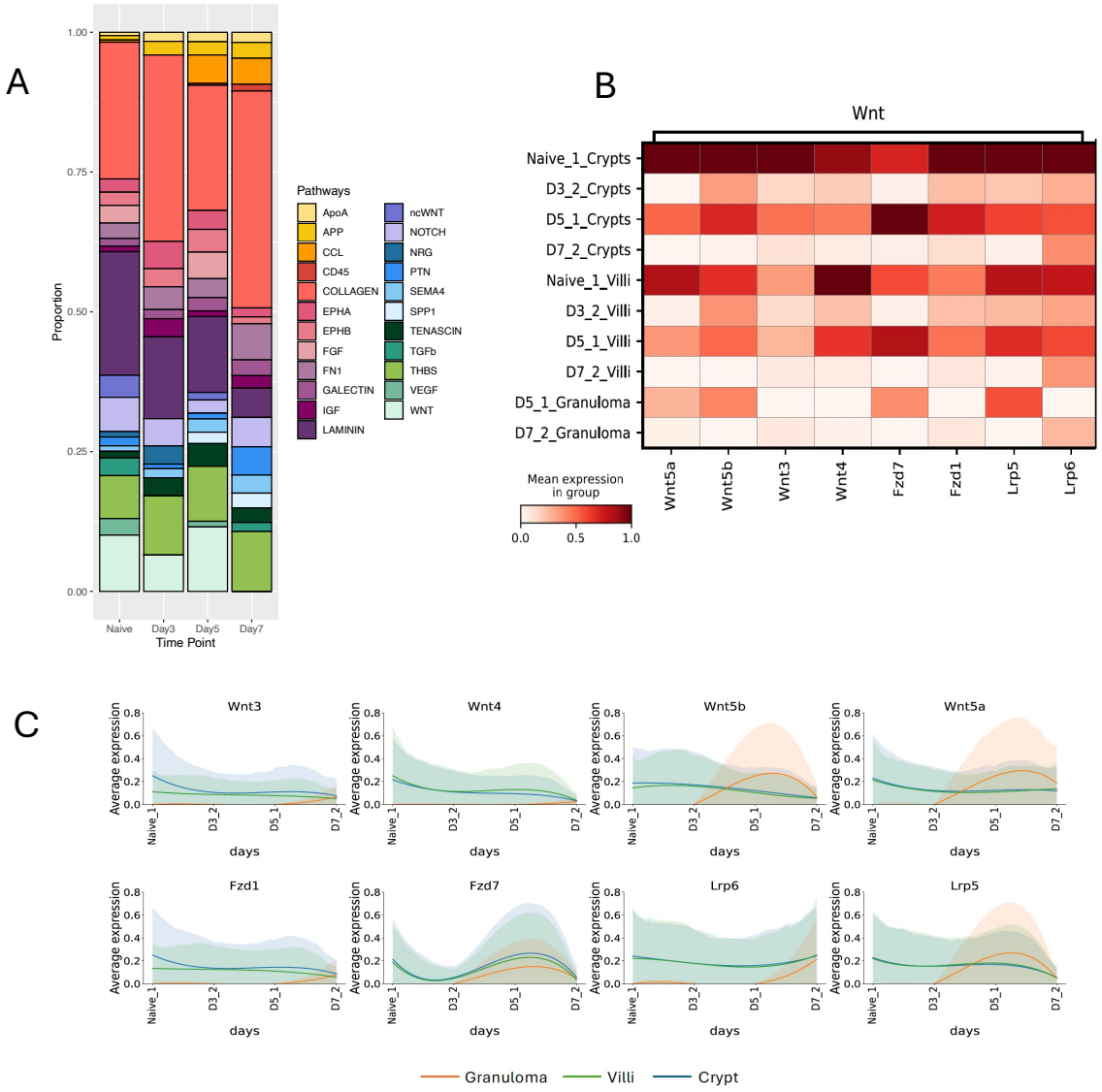


Figure 4.19: A) Stacked barplot showing the top interaction pathways at each time point during H. polygyrus infection. ApoA, Apolipoproteins; APP, Amyloid Precursor Protein; CCL, Chemokines; CD45, Leukocyte Common Antigen; EPHA/EPHB, Ephrin receptors; FGF, Fibroblast Growth Factor; FN1, Fibronectin; IGF, Insulin-like Growth Factor; ncWNT, non-canonical Wnt; NRG, Neuregulin; PTN, Pleiotophin; SEAMA4, Semaphorin 4; SPP1, Secreted Phosphoprotein/Osteopontin 1; TGFb, Transforming Growth Factor- , THBS, Thrombospondin; VEGF, Vascular Endothelial Growth Factor; WNT, Wingless/Int-1 (Integration of MMTV). B) Heatmap showing the mean expression of key ligand–receptor interactions in the Wnt signalling pathway across locations and timepoints. C) Average expression of each location of key ligand and receptors involved in the Wnt signalling pathway across time. Lines have been smoothed and fitted using polynomial regression, shaded regions represent confidence intervals for each fitted value. Crypts (blue), Villi (green), Granuloma (orange).

4.4.10 Crypts and villi show changes during *H. polygyrus* infection

We then focused on gene expression in samples taken at steady-state and 7 days post-*H. polygyrus* infection. We integrated the two time points (naive and day 7) and observed a clear separation (Figure 4.20). Leveraging the histological annotations, we examined the gene expression profiles of the the crypt and villous areas of the small intestine in naïve and infected murine tissues and identified specific genetic signatures in each of the tissue niches during infection. Specifically, after 7 days of infection, crypts show decreased expression of genes associated with intestinal homeostasis like Zg16, Muc13, Itln1 and Fcgbp. In the infected crypts there is also a reduction in factors that maintain the integrity of the intestinal barrier such as Epcam and Pls1[335]. On the other hand, a different suite of genes is upregulated in the infected crypts, with most elevated expression of cell adhesion proteins (Cldn3, Cdh17) which may indicate epithelial cell proliferation and modification during early stage of infection. We also observe the up-regulation of the phospholipase A2 family member Pla2g4c, which is involved in and required for killing of larval *H. polygyrus*[336]. Within the villi tissues, a similar down-regulation of expression is seen for pro-homeostatic genes such as Epcam, but distinct from crypt cells, the villi show down-regulation of metabolic mediators such as Slc25a5, and Cndp2, while upregulating Cldn3 and Pla2g4c within the crypt tissues. Thus, we see extensive epithelial remodelling during day 7 of the infection in particular in the crypt compartment.

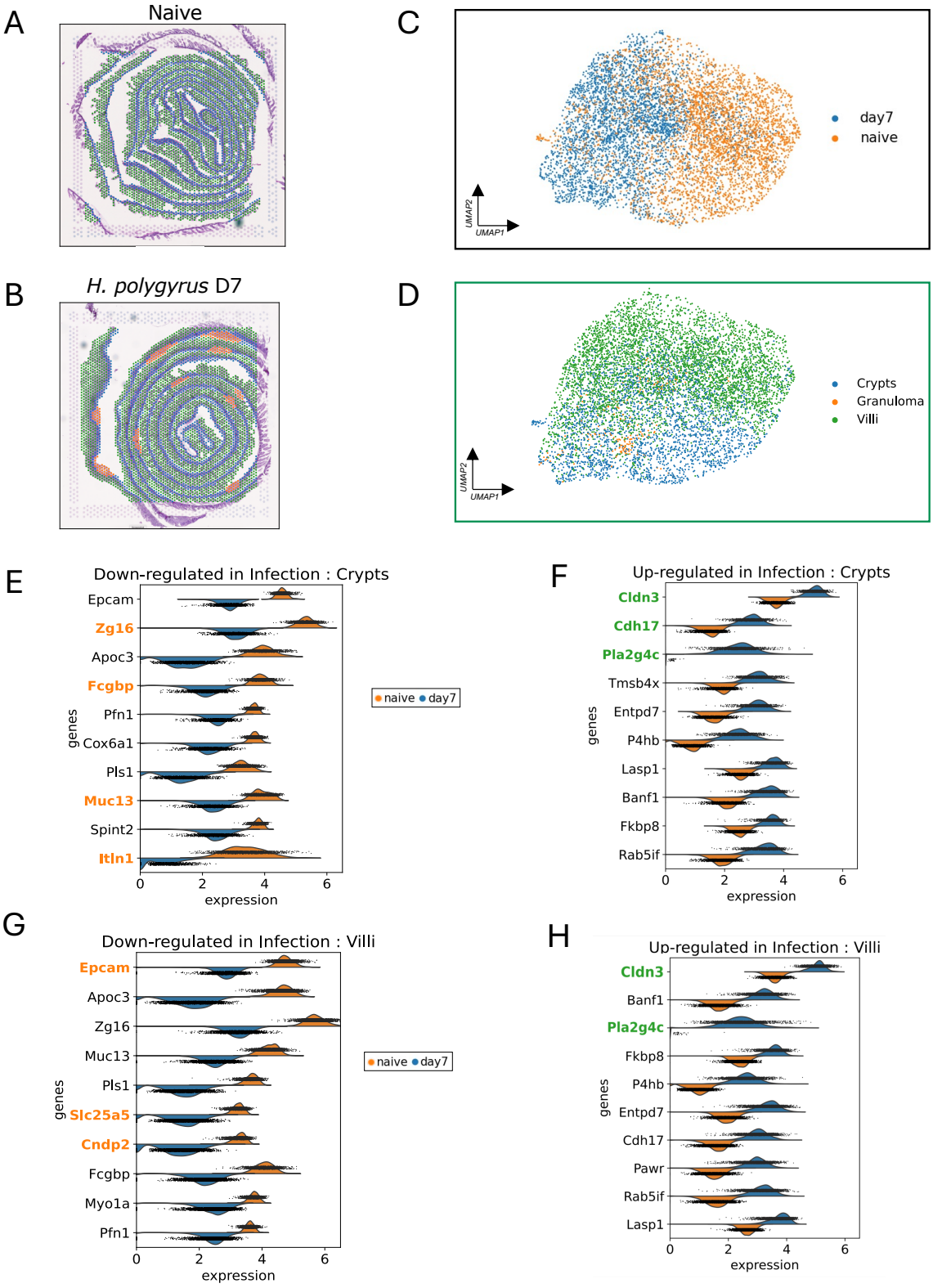


Figure 4.20: A) A, B Spatial plots of naïve (A) and day 7 H. polygyrus infection (B) highlighting tissue location clusters. C) UMAP based on Harmony integration of naïve and 7 days post H. polygyrus infection Visium datasets coloured by the sample origin of each of the dots. D) UMAP based on Harmony integration of naïve and 7 days post H. polygyrus infection Visium datasets, coloured by the tissue location of the spots. E-H) Violin plots of the expression of the top 10 differentially expressed genes in naïve and day 7 crypt (E, F) and villi (G, H). Down-regulated genes are shown in E and G, up-regulated genes in F and H, with normalized expression of naïve spots in orange, and day 7 post infection spots in blue.

4.4.11 Molecular characterisation of the *H. polygyrus* granulomas and the surrounding tissue niche

Next, we wanted to understand what is changing at the direct site of the parasite so we focused on the granulomas surrounding larval parasites in the submucosal tissue (Figure 4.21). We compared the combined transcriptomic signatures of all granulomas in comparison to the rest of the intestinal tissue. Interestingly, we found high expression of Tmsb4x, encoding thymosin beta-4, a small protein that may promote dendritic cell differentiation[337]. We also observed high expression of Arg1 (arginase-1) in the granulomas, and Retnla (encoding RELM-α), with both genes being closely associated with alternatively activated macrophages. In addition to this, the marked elevation of Ccl8 and Ccl9 is consistent with dominant infiltration by macrophages. Additional upregulated genes are involved in extracellular matrix (ECM) deposition (Fn1, Col1a1, and Ctsb) as well as antigen presentation and immune stimulation (C1qa, Tnfaip2), and lipid metabolism and oxidative stress regulation (Apoe, Cyba, Psap)[338,339].These upregulated genes suggest a coordinated response involving myeloid immune cells, tissue repair, and activation of inflammatory and remodelling processes within the granuloma microenvironment. We then wanted to know what transcriptomic differences there were between granulomas and the crypt sites. This analysis, confirmed the high levels of Arg1 and Retnla, as well as the monocyte chemoattractants Ccl8 and Basp1. In contrast, we see down-regulation of defensins and Zbtb48. The distribution of these genes was confirmed by a spatial analysis of expression relative to distance from the site of the granuloma, and by spatially mapping the distribution of expression of transcripts for Arg1, Basp1 and Lgals1 which co-localise exclusively to the granulomas.

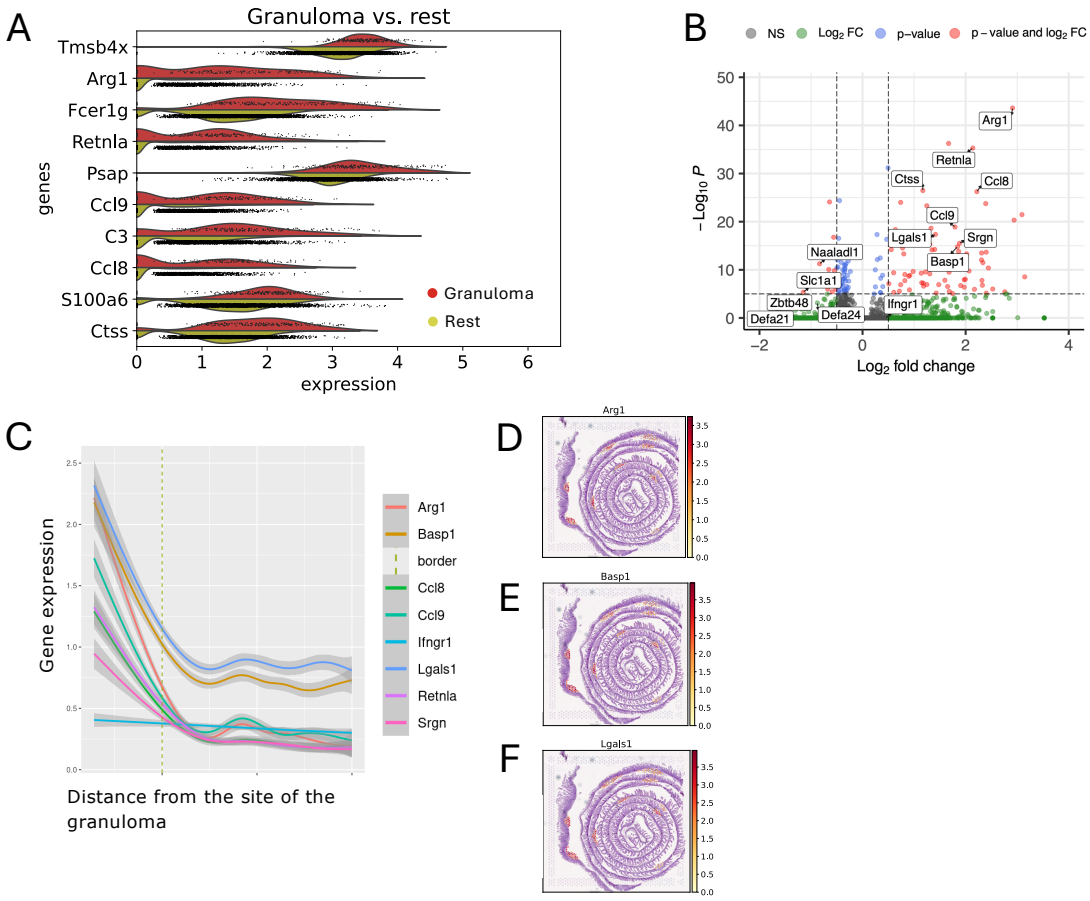


Figure 4.21: Transcriptomic landscape of the site of infection: A) Results of Scanpy marker gene analysis of the top 10 scoring genes for granuloma 7 days post H. polygyrus infection, displayed as split violin plots of normalised expression values, with red plots for granuloma spots and yellow for non-granuloma spots. B) Volcano plot showing the top 5 up/down regulated genes in the granuloma niche compared to the crypt niche. C) Spatial distribution plot showing the gene expression of the top 8 upregulated genes in the surrounding granuloma niche at day 7 post infection. The dotted line denotes the boundary of the spots that are labelled as granuloma but neighbour non-granuloma spots. D-F) Spatial plots showing the gene expression of Arg1, Basp1 and Lgals1 localised exclusively to the granuloma niches.

We then wanted to know if the granulomas themselves were transcriptionally heterogenous. After subsetting the granuloma spots in day 7 post-infection, we reclustered them, which resolved into three distinct clusters (Figure 4.22). The spots in cluster 0 represent granulomas in which no larva is visible, either because the adult is already in the lumen or because the section failed to capture the worm in the histology. Analysis of differential gene expression in the 3 clusters, revealed interesting profiles of specific gene sets.

In cluster 0, there is a higher level of immune cell products including the MHC Class II antigen H2-Q2, and proteins involved in interferon responses (Ifi27l2b), and immune regulation (Clec2h). Furthermore, the upregulation of Vil1, Zg16, and Krt20 suggests the presence of epithelial cells that may contribute to the structure of the granuloma itself. Interestingly, one of the upregulated genes of cluster 0 granulomas is Zg16, which is conversely down-regulated in infected villi. The observed gene expression within cluster 1 of granulomas containing larval parasites contains features of both type 1 and type 2 immunity. The upregulation of proinflammatory genes such as Tnfaip2 and Ccl9 and Fcer1g is observed alongside genes such as Arg1, Basp1, C1qa, Fn1, Emilin1, and Psap which point towards alternative macrophage activation and associated angiogenesis, tissue repair and extracellular matrix remodelling. Cluster 2, which represents a single granuloma which like cluster 0 has no visible larva, but shows a very distinctive gene profile with a lower level of macrophage activation genes, with high Reg3b, Reg3g and Agr2 expression indicating resolution and regeneration of the cellular environment[294,340], together with Mxra7, encoding Matrix remodelling associated 7 protein implicated in wound-healing[341].

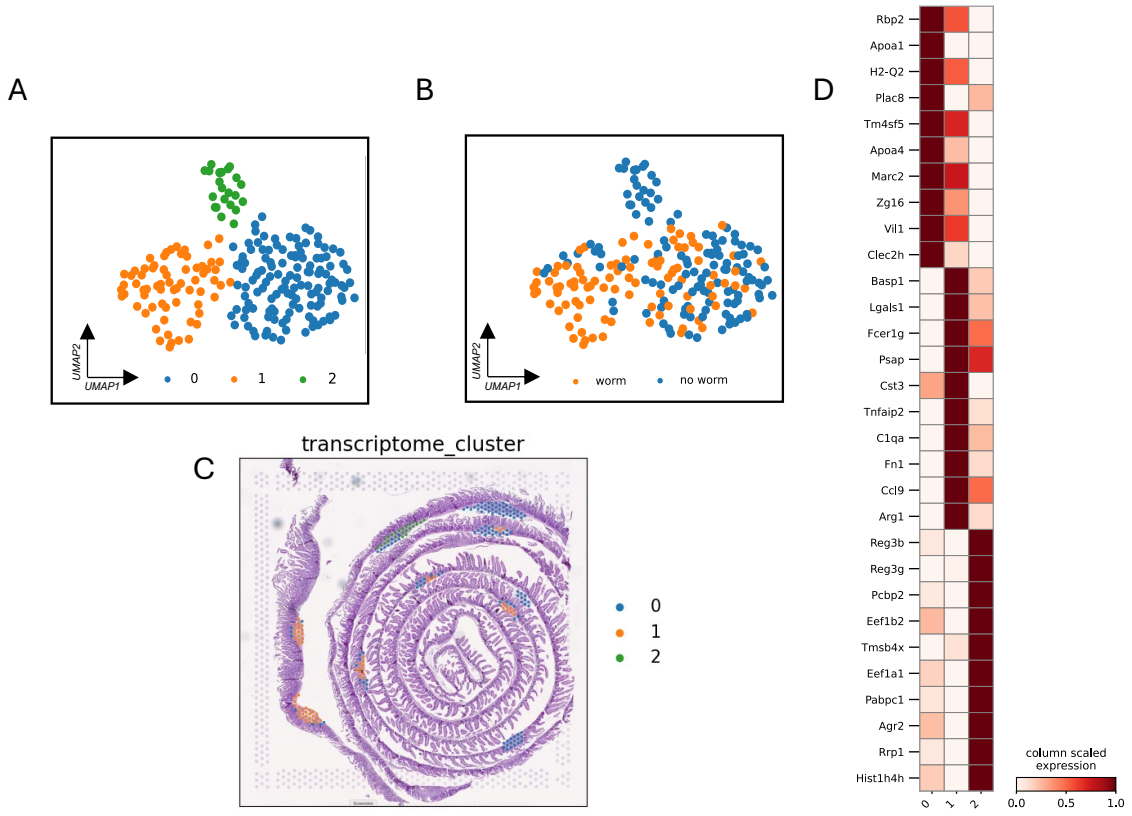


Figure 4.22: A, B) UMAP of granuloma spots from the mouse intestine 7 days post infection with H.polygyrus with spots coloured by transcriptome-based Leiden clusters (A) and by absence or presence of H. polygyrus based on histological annotation (B). C) Spatial plot of mouse intestine 7 days post infection with H. polygyrus with spots coloured by transcriptome-based Leiden clusters. D) Scaled expression of the top 10 gene markers for each granuloma transcriptome-based Leiden cluster.

In addition to defining gene expression patterns within the granulomas, we asked whether intestinal crypts adjacent to, or distant from, the sites of the granulomas showed distinct transcriptional profiles. Through manual histological annotation by Dr. Marta Campillo, sets of crypts were assigned to each category and differential gene expression was performed to identify candidate genes that are modulated by the presence of the parasite. A number of genes upregulated in the vicinity of granulomas are macrophage-associated

products also found within the granulomas such as Retnla and Fcer1g, although Arg1 was not highlighted. However, we found two genes involved in the Wnt pathway, Dact2 and Frat2 are locally down-regulated, consistent with the inital cellular interaction pathway analysis that indicated overall reduction in Wnt signalling.

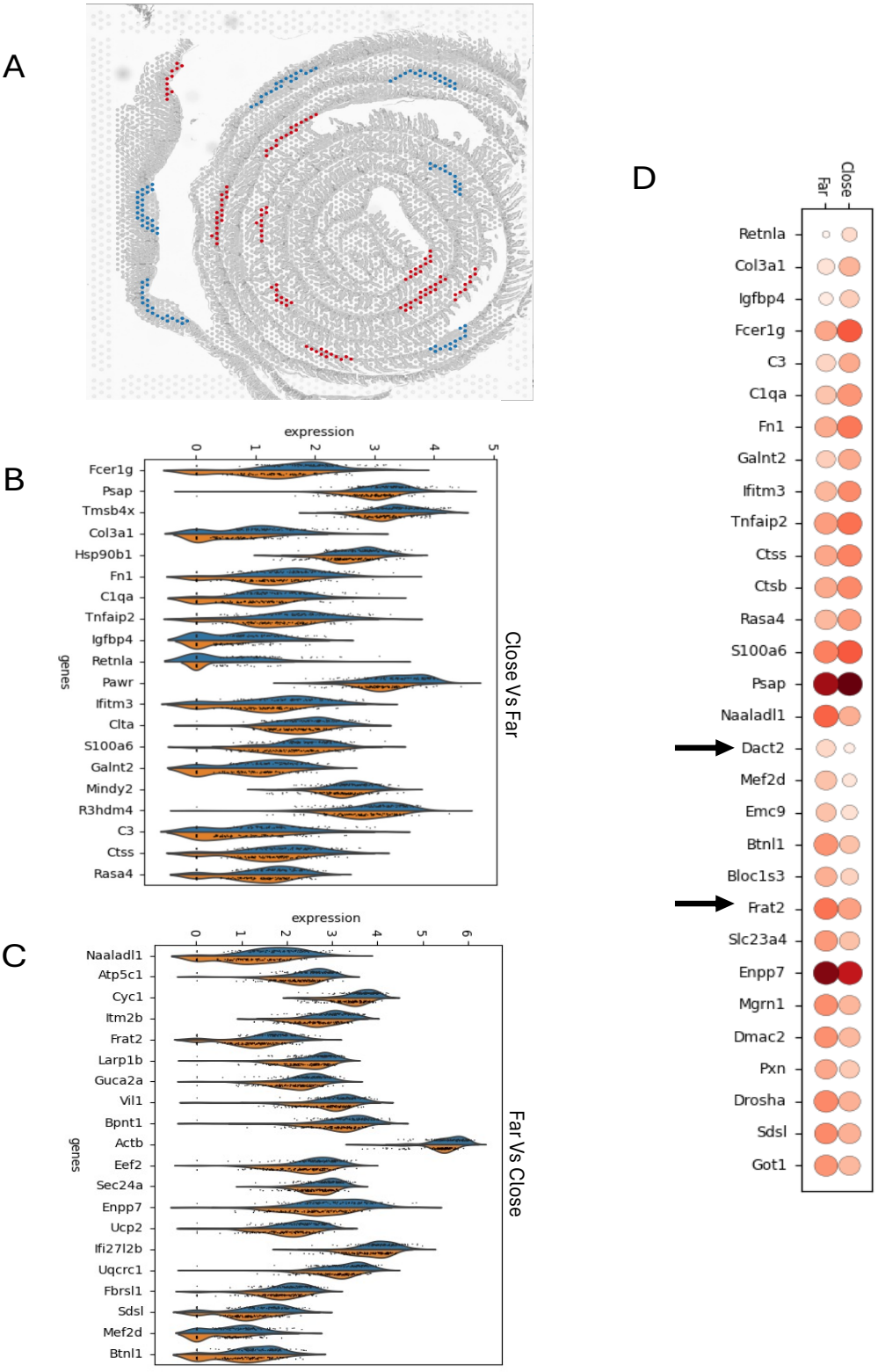


Figure 4.23: A) Assignment of crypt areas ”close” (blue) or ”far” (red) from sites of the granulomas. B-C) Violin plots of differential gene expression for top 20 genes upregulated (B) or downregulated (C) in crypts close to parasite locations. D) Dotplot showing gene expression levels for genes in B and C.

4.4.12 Using cellular deconvolution to identify spatial niches

Having recreated the original biological spatial context of the intestine using histological annotations, we wanted to next identify the cell types present and their distribution across the spatial axis. Dr. Ross Laidlaw implemented cellular deconvolution to deconvolute the pseudo-bulk Visium spot data, using cell2location and two integrated scRNA-seq datasets of immune and non-immune cells from the intestines of mice respectively from published studies[35,310] to ensure there would be representation of the immune and epithelial cell types that comprise the intestine. Leveraging these annotations, we focused on the most proximal part of the intestine, the duodenum, which is the primary site of *H. polygyrus* tissue invasion (Figure 4.24). Ensuring that we preserved spatial localisation of the tissue we used calculated spatial embeddings from a digital unrolling method proposed by Dr. Ross Laidlaw containing our unrolled length (anterior to posterior) and depth (lower crypt to villous tip) axes, and the original Visium spatial coordinates, to ensure adjacent segments in the Visium space are separated from each other. After applying non-negative matrix factorization to identify which cell types co-localise together in the same spatial niches. Within the deepest spot layer of the crypts, the lower crypt spatial niche is dominated by transit amplifying (TA) cells and intestinal stem cells, with the presence of both CD4+ and CD8+ T cell subsets. Directly above, the upper crypt is almost entirely composed of enterocyte progenitors with a small amount of lymphoid cells. Extending toward the lumen, the villous niche co-localises enterocytes, B cells, innate lymphoid cells (ILC) 1 and 2, NK cells and $\gamma\delta$ T cells. Focusing on co-localisation signatures that are specific to infected mice, an accumulation of macrophage, neutrophil, plasmacytoid DC, mast cell and lymphoid tissue inducer cells localised in the granulomas. While many of the cell types found to be co-localised around and within granuloma niches (macrophages, neutrophils, dendritic cells and CD4+ T cells) have been shown to be associated with helminth-induced granulomas, the involvement of mast cells, pDCs and lymphoid tissue inducer (LTi) cells has yet to be reported.

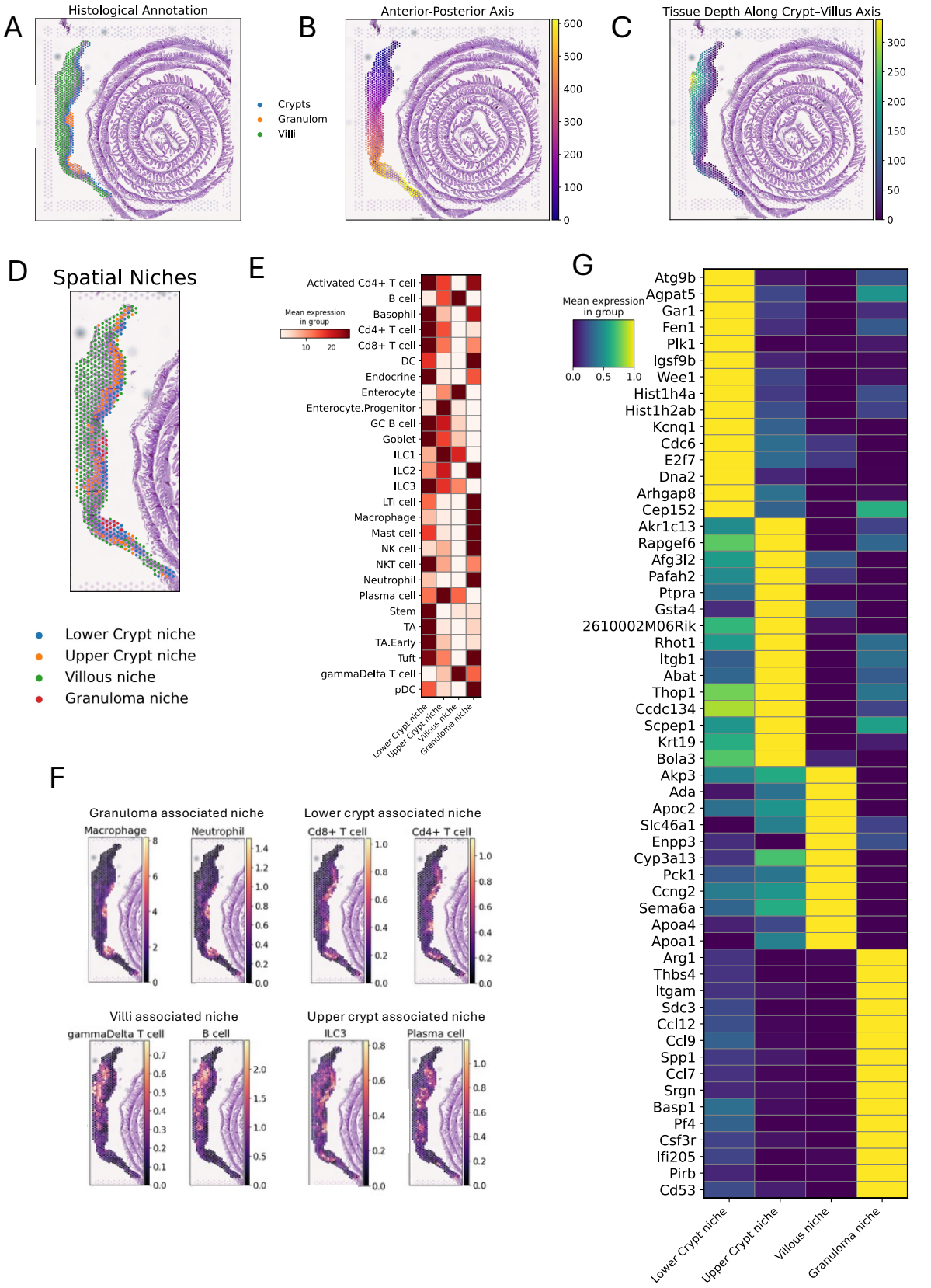


Figure 4.24: A) Visium slide of mouse intestine 7 days post H. polygyrus coloured by the histological tissue location. B-C) Visium slides of mouse intestine 7 days post H. polygyrus showing spots coloured by the recreated length (B) or depth (C) axis. D) Visium slide of mouse intestine 7 days post H. polygyrus coloured by spatial niches. E) Heatmap showing the relative mean expression of each cell type signature present in the Xu/Huber reference single cell dataset across each spatial niche. F) Spatial projects of the top 2 predicted cell types in each spatial niche coloured by normalised cell abundance. G) Top 15 highly expressed genes for each spatial niche in the infected intestine.

4.4.13 Cell-to-cell interactions and signalling pathways

Following characterisation of cell types and gene expression within each spatially resolved niche in the infected murine gut, we wanted to know what interactions are occurring between each spatial niche. Using CellChat we inferred ligand-receptor interactions between each spatial niche in the day 7 post-infection segment. From the granuloma niche, all predicted interactions were with the lower crypt, and represented the highest number of predicted interactions in the whole study (Figure 4.25). Numerous interactions were also observed between other sites, between the lower crypt and the upper crypt with the villi that were not predicted to directly interact with granuloma niche. Enriched predicted signalling between the granuloma and lower crypt niches primarily represented immune cell activation and differentiation pathways. Dominant chemokines in the granuloma profile are CCL6, CCL7, CCL8 and CCL12, with the CCL8/CCR5 pairing with the lower crypt likely indcating activated myeloid and lymphoid populations recruited to the granuloma. Interestingly, the among the strongest interactions is between the chemokine-like macrophage migration inhibitory factor (MIF) expressed in both the lower crypt and the granuloma, and its receptors CD74 and CD44 present in both niches, as this chemokine-like mediator is known to be essential for immunity to *H. polygyrus* as well as to the rat nematode *Nippostrongylus brasiliensis*[300]. As expected for the tissue remodelling involved in granuloma formation, there are prominent interactions with ligands for extracellular matrix and proteoglycan including pleiotrophin (Ptn) which binds heparin and the syndecin receptors Sdc1, Sdc3 and Sdc4 as well as nucleolin (Ncl), and the IL-6-like cytokine Oncostatin M (Osm). We also observed Spp1 which binds CD44 and integrins, which when compared to the naive 3 tissues sites, was prominent in the granuloma niche. Spp1 can also interact with integrins involved in activation of latent TGF-β. TGF-β signalling was seen to be high in granuloma-crypt interactions, primarily represented by TGF-β1 from the granuloma, binding the canonical TGF-β receptors, but also signals from the lower crypt to the upper crypt. It is known that blocking TGF-β signalling promotes the release of *H. polygyrus*[33], and more interestingly that the parasite secretes mimics of TGF-β that bind the same receptors together with CD44[342]. Finally we wanted to see if

we could spatially map the identified key ligand-receptor pairs to see co-localisation to the spatial niches. We observed that chemoattractant signalling along with TGF-β, were found largely in the sites of granulomas although MIF-CD74/CD44 appeared to be more generalised through the infected tissue.

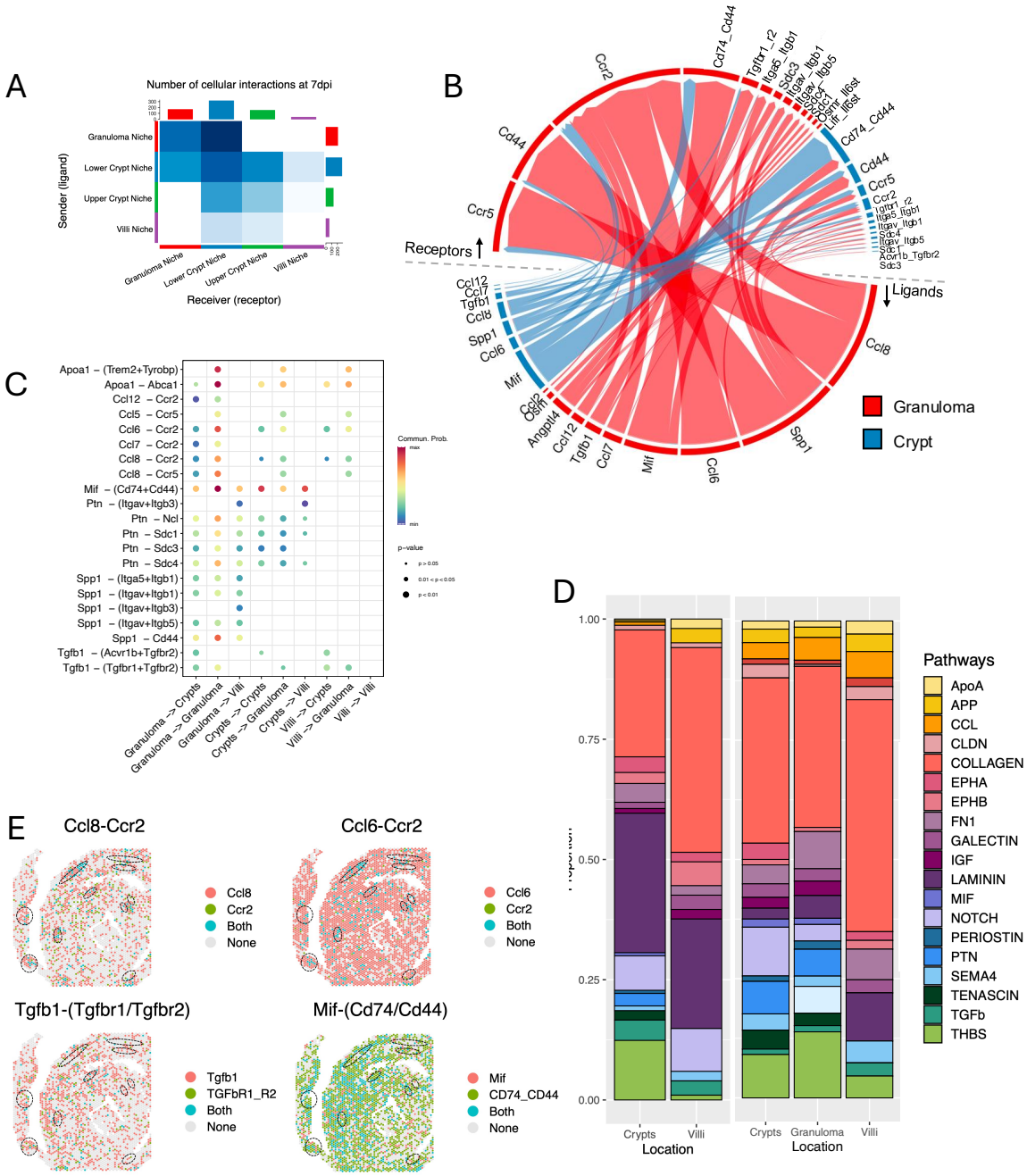


Figure 4.25: A) Heatmap showing the number of ligand-receptor (L-R) interactions inferred by CellChat for each spatial niche at 7 days post infection. Bar plots at top and side show the sum of all interactions. B) Circos plot visualisation showing upregulated ligand and receptor pairs between the granuloma (red) and the lower crypt (blue) niche. Ligands are placed in the lower half and receptors in the upper half, with arrows showing directionality of the interaction. C) Dot plot showing the communication probability of each significant ligand receptor interaction between various spatial niches in the murine gut at 7 days post infection. D) Stacked proportion bar plot showing the interaction pathways associated with each location of the murine intestine day 7 post-infection. E) Spatial projection of key interacting pairs involved between the granuloma and crypt niche during H. polygyrus infection, coloured by gene expression.

Collectively, these findings demonstrate a progressive disruption of epithelial renewal and immune-epithelial crosstalk during *H. polygyrus* infection. Early suppression of Wnt signalling (day 3) results in Notch-dominated differentiation (day 7), a pattern that has been previously linked to helminth-induced epithelial remodelling. Also we found crypt-granuloma interactions evolve from inflammatory (IL-6-driven, day 5) to a tissue remodelling state (TGF-β, SPP1, day 7), reflecting the dual nature of the host response attempting both parasite clearance and damage repair. These results reinforce the fact that *H. polygyrus* actively reshapes its host environment, leveraging immune suppression and epithelial reprogramming to establish chronic infection. Understanding these spatially distinct and time-dependent interactions provides deeper insight into how helminth parasites exploit the gut niche and highlight potential targets for therapeutic intervention.

4.5 Discussion

In this chapter we interrogate how cellular interaction inference can be adapted to atypical and challenging contexts, specifically in atlas-level single-cell datasets with complex heterogeneity and size, and spatial transcriptomics data where the tissue architecture is disrupted. In the first case study, we presented a macrophage–fibroblast atlas spanning four tissues and multiple disease states to more than 250,000 cells, to observe tissue-resident myeloid-stromal crosstalk in steady state and disease. Using a reference normalised atlas of 40,000 myeloid and stromal cells we implemented label transfer annotation to expand the full dataset to run cellular communication inference. We recapitulated cell type proportions across all tissues from the normalised atlas to the full dataset capturing representation of myeloid and stromal subtypes and preserving tissue-specific cell types such as Langerhans cells in the skin, A2M+ fibroblasts in the lung and PRG4+ fibroblasts in the synovium. After the exclusion of the heart dataset we compared cellular interactions in disease and homeostasis across the lung, skin and synovium which unveiled patterns of communication between different cell types. In disease, we found that SPP1 macrophages

acts as key communicators consistently across tissues interacting with tissue-resident fibroblasts. Osteopontin expressing macrophages have been widely reported in inflammatory disease with roles such as promoting fibrosis, remodelling of the extracellular matrix, and modulation of the immune response sustaining chronic inflammation. Studies show that SPP1+ macrophages promote fibrosis in myocardial infarction[343], contribute to reoccurrence and chemo-resistance in breast cancer[344] and have shared pathways across various diseases such as rheumatoid arthritis and COVID-19[345]. We went on to experimentally validate SPP1+ and its co-localisation with SPARC+ and APOE+ fibroblasts in healthy and disease skin and synovium. Staining revealed that at the site of inflammation in both tissues CXCL12 and SPP1 co-localise indicating that in psoriasis and RA synovial tissue interactions between SPP1+ macrophages and APOE+ fibroblasts are interacting. In tissues that are actively inflamed SPP1 acts as a stimulator of fibroblasts, promoting their activation and proliferation regulating collagen synthesis and ECM remodelling[346]. This is useful under homeostatic conditions for ECM maintenance and renewal as reported in the presence of collagen production in homeostasis across the three tissues however when this mechanism is dysregulated fibrosis exacerbates tissue damage and chronic inflammation accelerating the progression of chronic inflammatory diseases[347]. In the lung, we found that TREM2+ macrophage interactions dominated inflammatory pathways interacting with lung specific tissue resident A2M+ fibroblasts. TREM2+ fibroblasts interacting with lung resident fibroblasts have been reported to be pro-fibrotic with recent reports that inhibition of this TREM2 leading to amelioration of pulmonary inflammation and fibrosis[327,348]. Despite the popularity of single cell atlases, little progress has been made in trying to glean additional insights from these datasets such as cellular communications. Recently, an atlas of cellular communications was proposed called CellCommuNet that contains cell−cell communication networks inferred from 376 scRNA-seq datasets from human and mouse tissues in normal and disease states[26]. In the future we hope that our dataset can be integrated into multi-faceted resources such as this to facilitate the identification of shared and distinct interactions between tissue-resident cell types in inflammatory disease and steady-state.

Although we report shared signalling pathways across the lung, skin and synovium the heart dataset was excluded from our analysis due to low cell numbers. As a result, this forced us to exclude the heart from downstream cellular communication inference as most cellular interaction packages require at least 10 cells in each cell type annotation to perform inference between conditions. Since the curation of the atlas by Dr. Lucy Macdonald in 2021, there have been single cell datasets of cardiac tissue in disease published such as Koenig et al[349] in 2022 who included 6 patient samples of heart failure and more recently in 2024, the CardioAtlas curated by Jiang et al[350] which includes 12 human disease heart tissue datasets. Inclusion of these datasets would have allowed us to gain further insight into shared and distinct inflammatory and steady state signalling pathways in the heart to compare against our findings. Reports from other cardiac single cell studies have reported a SPP1+ macrophage role in promoting fibrosis in myocardial infarction[351] by interacting through various heart fibroblast subsets, suggesting that our findings presented in the lung, skin and synovium would have been complementary to inflammatory pathways in the heart.

During our analysis we also found particular interactions that may have been driven by a particular dataset such as the presence of multiple MHC-II genes in the synovium during homeostasis. By pooling the different inflammatory diseases into one umbrella of 'Disease' we fail to correct for biases introduced by interactions that may be overrepresented in a given state. This, paired with biases from the cellular interaction databases, caused the analysis to be dominated by fibrotic signalling pathways when the underlying landscape of inflammation underpinning these tissues is likely more complex. This could be mitigated by using cellular interaction packages such as multinichenetR[152] that is tailored to cellular inference of more complex experimental designs, correcting for multiple patients and conditions. In this study, we analysed cells that came from 122 different samples, across homeostasis and 10 different inflammatory diseases which a more basic cellular interaction inference tool such as CellChat may not adequately handle. However, multinichenet first runs a differential expression analysis which requires there to be sufficient cell types per sample per condition which our data did not meet in all of our myeloid-

stromal subtypes. This was due to tissue-specific expression of some cell type populations such as PRG4+ fibroblasts in the synovium and Langerhans cells in the skin which are not present in other tissues. Thus, as a workaround the analysis would have needed to implement broader cell type labels which would have affected the granularity of understanding how different sub-populations in the data are interacting. This approach would have been sufficient for investigating cellular interactions across shared cell types across the tissues, however the analysis would have omitted distinct tissue-specific interactions that may be of interest.

Finally, the analysis of cellular communications inferred from an atlas-level dataset such as the one presented in this chapter yielded over 500K interactions between different combinations of cell types and tissues. This highlights the scale of the results of a cellular interaction inference analysis can provide, requiring a clear biological question to be asked before delving into the data. Here, we chose to highlight interactions and signalling pathways that were shared in disease and followed up with experimental validation with immunohistochemistry. There are many other aspects of the data yet to be explored to further understand the complex interplay between tissue-resident myeloid and stromal populations in inflammatory disease.

Moving away from the cell atlas, we proposed a second use case where we showed that when spatial topology is warped, as in the 'swiss-rolled' intestine during *Heligmosomoides polygyrus* infection, considerations for the underlying tissue architecture is a prerequisite for cellular inference. By leveraging histological annotations and digitally unrolling and segmenting the tissue, we asked specific questions to tailor the analysis to regions of interest reconstructing the crypt–villus axis and interrogating granuloma-associated cell states. Our analysis revealed a temporal change in early Wnt-driven epithelial renewal, that is subsequently suppressed by the parasite during chronic infection, that is replaced by Notch-driven differentiation and TGF-β/SPP1 regulated tissue remodelling. When examining the specific tissue niches in the gut during *H. polygyrus* infection we found that the crypts and villi at day 7 down-regulate barrier and homeostatic genes like

Epcam while up-regulating adhesion modules such as Cldn3, Cdh17 and Pla2g4c, consistent with epithelial reprogramming and anti-helminth effector functions. Within the granuloma microenvironment, transcriptional profiles suggest alternative macrophage activation with Arg1, Retnla expression, immune recruitment through cytokine signalling, ECM deposition and tissue remodelling. We also reported transcriptional heterogeneity within granulomas determined by presence or absence of the parasite, identifying clusters reflecting active inflammation and repair, immune–epithelial crosstalk, and a resolution-like state. Cellular interactions during chronic infection revealed increased chemokine signalling to the sites of the granuloma including the expression of MIF and CD44, critical for the immunity to *H. polygyrus* infection[300]. Despite potential limitations of deconvolution approaches we leverage spatially resolved neighbourhoods with existing single cell data to predict cell-to-cell interactions of spatial niches within the small intestine. Common signatures of cell type colocalisation were identified, particularly in the crypts and villi, emphasising the coordinated organisation of various cell types within specific tissue regions. Specific colocalisation patterns in *H. polygyrus* infection, particularly in the granulomas, highlighted the dynamic cellular interactions occurring during the host response to parasitic infection. In contrast, in the steady-state there is negligible activity from immune cells and the interactome is rather associated with epithelial cell differentiation and maintaining tissue integrity of the intestine. Thus, this landscape is dramatically altered in response to parasitic infection and the formation of granulomas.

Our analysis faced some limitations, for example the initial integration of the time points proved challenging in this analysis due to high variation in sequencing depth across the tissues and time points. We noticed that there was a technical artefact in the data where the centre of the tissues contained significantly a higher number of UMIs and features when compared to the outer regions of the gut roll. This could be down to the application of the reagents during processing or that certain regions of the tissue were thicker than others, affecting the permeability of the tissue. In normal conditions computational techniques such as normalisation and batch correction should mitigate these effects, however in this case the bias still persisted and affected downstream tasks such as clustering. Thus, it

resulted in nonsensical clustering when projected back to the spatial axis providing little biological insight. Other methods that are tailored more specifically to identifying spatial niches by implementing spatially aware clustering such as PRECAST[352] also failed to provide spatial clustering that made sense in the tissue. Thus, as we had histological annotations that accurately reflected the architecture of the gut we proceeded to utilise this spatial information for the majority of the analysis.

In order to run cellular communication inference we had to address the warped spatial axis introduced by the 'swiss-rolling' technique of the intestine. We attempted to apply existing methods to digitally unroll the intestine provided by the semla package[296] as detailed in the Parigi[294] paper. However, this method required the gut roll to be intact and uninterrupted on the slide in order to accurately recreate the length of the gut. In our Visium data we did not capture the gut roll in its entirety, with the positioning of the gut roll sometimes overlapping with the fiducial border meaning it was not captured in the sequencing or through breakage of the tissue where parts of the gut were segmented. To try and overcome this, Dr. Ross Laidlaw created a method that allowed partial digital unrolling of the gut by taking uninterrupted segments and patching them together using overlapping spatial coordinates between broken sections. This was sufficient for calculating the anterior-posterior axis and length along the intestine, however for cellular communication inference the overlap of the different segments rendered the method unsuitable as the distance metric of ligand diffusivity would have been distorted. As a result, we chose to focus our cellular interaction analysis on an uninterrupted segment of the intestine during day 7 post-infection that contained multiple granulomas to understand the host-parasite interface in more detail.

Additionally, as 10X Visium captures multiple cells within each spot we attempted to deconvolute the data to try and infer which cell types were occupying the different spatial areas of the gut. For this we ran cell2location[189] and used two single cell datasets to infer our cell type signatures. At the time of the analysis there was not a single cell dataset that described both stromal and immune cells in the murine gut, as a result we concatenated

two separate datasets one encapsulating the stromal compartment by Haber et al[35] and one for the immune compartment by Xu et al[310]. After running cell2location we found that multiple cell type signatures are present in overlapping regions across the intestine making it difficult to assign a spot to a dominant cell type. Thus, we opted to instead utilise negative matrix factorisation to estimate distinct tissue niches in the gut which nicely characterised the main spatial compartments of the intestinal segment we focused our analysis on. These spatial niches became the focus of the cellular interaction analysis where we opted to use the spatial niches instead of estimated cell types to gather an understanding of which cellular interactions are occurring in which niche. This meant that we could not report which cellular interactions were originating from particular cell types in the gut and adopted a more localised view of interactions between niches. This highlights the appropriateness of using a pseudo-bulk spatial method such as Visium on such an intricate tissue like the gut. For example the villi are estimated to be 150–400μm long in mice[353] which when taking into consideration that a Visium spot measures 55μm in diameter is spaced so that the distance between the centres of each spot is 100μm it is clear that the resolution is unsuitable to capture insight within smaller tissue architectures. With the advent of higher resolution spatial technologies such as Visium HD which contains 2μm barcoded squares with no spacing and true single cell spatial platforms like Xenium and Cosmx, perhaps we can gain more granularity in cellular communication inference in this setting.

Despite this, taken together, our analyses across human tissues, disease accentuates SPP1 driven myeloid-stromal crosstalk with a conserved ECM-remodelling and pro-fibrotic signature. During helminth infection, we report temporal suppression of Wnt signalling replaced by Notch-driven differentiation, and granuloma-specific TGF-β/SPP1 immune recruitment and alternative macrophage activation, alongside epithelial interactions. These insights interestingly draw parallels in increased collagen production, ECM remodelling and collagen deposition influenced by a shared SPP1 axis during states of chronic inflammation.

Chapter 5

Discussion

In conclusion, this thesis highlights the powerful technique of cellular interaction analysis in unveiling novel and shared immunomodulatory mechanisms in diseases such as COVID-19, inflammatory disease, and parasitic infection using single cell RNA sequencing leveraging spatial technology to validate interactions. The field of cellular communication inference has boomed since the publication of the first cellular interaction tool CellPhoneDB in 2020[148], with over 100 bioinformatic tools and 50 database resources published as of 2024[354]. Core cellular interaction tools such as CellPhoneDB[148] and CellChat[146] base their calculations on the average expression of ligand and receptors across cell type clusters and compute scoring functions such as expression mean or communication probability to quantify expression but ignore the multifacted biological nature of cellular communications[137]. Nevertheless, these tools have provided insight into novel disease mechanisms such as COVID-19[4,5], developmental processes[146,355] and cancer[356]. Since then, cellular interaction tools have started to challenge these basic assumptions adopted by core tools such as SoptSC[151] which computes interactions at the single cell level thus eliminating the cluster-wide expression of a given ligand/receptor enabling the capture the single-cell nature of cellular communication and heterogeneity that may be missed by aggregating expression. Other tools aim to investigate downstream impacts of cellular communications, such as NicheNet[150] which models downstream intracellular signalling rather than intercellular signalling. Critically, more recent tools[152,357,358] have started to address mul-

tiple conditions and batch effect by implementing a differential expression step to more robustly predict interactions that are not an artefact of sample differences such as MultiNicheNet[152]. This eliminates previous approaches widely adopted by most early cellular inference tools that required the analysis to be run separately for each condition.

The major downfall of cellular communication inference tools is the absence of a ground-truth to validate the tool and the identified cellular interactions. Thus, a degree of orthogonal validation is necessary to eliminate false positives from predicted cellular interactions inferred from transcriptomic data. In this thesis we demonstrated how using MultiNicheNet[152] allowed us to correct for patient biases and identify IFNGR1 being differentially expressed on alveolar macrophages in lethal COVID-19. We then implemented RNAscope alongside an integrative analysis with low-plex proteomic imaging mass cytometry data to validate this which resulted in increased expression of this receptor supporting the validity of the inferred interaction. Similarly, in Chapter 3 we applied CellChat[146] to an atlas-level dataset and identified APOE+ and SPARC+ fibroblasts being key drivers of inflammation across multiple tissues along with SPP1+ macrophages, which was supported by immunohistochemical staining of cell type markers showing co-localisation in diseased tissues. However, experimental validation of this nature is often limited to a few ligand-receptor pairs and/or few cell types per experiment lacking the power to simultaneously spatially resolve multiple cellular interactions. With the advent of spatial transcriptomics, comes an attractive alternative for indirectly validating cellular interactions as we can profile cells at either full-transcriptome resolution, for sequencing-based approaches such as 10X Visium, or targeted panels, for probe-based methods such as 10X Xenium, in their native spatial context. This technology drove the development of next generation cellular interaction inference methods that considered a spatial axis in their models for predicting spatially aware ligand-receptor interactions[180,181,194,197,198]. Furthermore, the development of spatial transcriptomics inspired the development for core cellular interaction tools to introduce spatial constraints as cells in close proximity are more likely to interact with each other. CellChat v2 offers functionality to spatially

visualise cellular interactions as shown in Chapter 3 visualising chemokine and TGF-β signalling (Figure 4.25 E) such as thresholds for intercellular distance[146]. On the other hand, CellPhoneDB v5 requires a user input of manually defined spatial niches that it utilises to filter interactions that do not occur within the same niche[144].

Although the power of cellular interaction analysis is evident, leveraging both single cell and spatial transcriptomic data, interpretation and visualisation of the results still remains a computational burden[354]. Due to the high-dimensional nature of the output interactive visualisation tools are lacking in the field to facilitate navigation of results. This was addressed in Chapter 2 presenting cellXplore that allows interrogation and visualisation of cellular interaction results through a user-friendly web interface. Although some interactive visualisation tools exist[205,206], a tool that appropriately visualise both single cell and spatial data simultaneously in a uniform view is yet to be seen, mandating the development of cellXplore. In particular, as high-dimensional single cell and spatial data continue to expand the need for easy visualisation without the hindrance of prior bioinformatic knowledge is becoming increasingly important to extract biologically meaningful interactions from cellular communication data. In addition to this, appropriate visualisation of multi-dimensional cellular interaction data is an open challenge. Most tools adopt common plotting functionalities such as dot plots, heatmaps, circos plots and spatial plots however CellChat amongst all tools provides the most extensive plotting library to comprehensibly and coherently plot cellular interactions. Thus, throughout this thesis, CellChat both version 1/2 has been implemented to generate a majority of the figures allowing unprecedented flexibility of customisation and control over the data displayed. These plotting visualisations inspired the interactive plots that are implemented in cellXplore and continue to outperform other visualisations provided by most cellular interaction packages. However, cellular interaction inference tools are continuously evolving. More recently, the database of CellPhoneDB v5 is the most comprehensive, including signalling metabolites and small molecules and has been used by myself to infer cellular interactions between T cells and DC subsets in different spatial niches in rheumatoid arthritis[210]. Therefore, as

the field progresses, it is important to select the optimum tool to answer the biological question, considering the ligand-receptor database the tool adopts and, most importantly, the visualisation options provided to extract the most information from the dataset at hand.

Despite the plethora of cellular interaction tools that are currently available and the multi-faceted nature of complex multimodal datasets that emerge in the fast developing field of transcriptomics, the work presented in this thesis provides validated interactions that pertain to mechanisms of disease. Through the application of spatially aware cellular interaction inference we report a type II interferon response in lethal COVID-19 driven by alveolar macrophages, a mechanism of action distinct from represented cohorts in the Northern Hemisphere, validated using imaging mass cytometry and RNAscope. This then inspired the development of cellXplore to facilitate interactive visualisation of cellular interactions inferred from single cell RNA sequencing data leveraging spatial transcriptomic data to indirectly validate interactions of interest. Finally we apply cellular interaction inference to uncover a shared SPARC+/APOE+ fibroblast and SPP1+ macrophage driven axis of communication in active inflammation, experimentally validated by immunohistochemistry. Additionally, we provide insight into the mechanisms of gut epithelial remodelling and immunoregulation during *Heligmosomoides polygyrus* infection over time using spatial transcriptomics to identify cellular interactions occurring in distinct spatial niches in the gut. Collectively, these findings exercise the power of cellular interaction inference, critically utilising a spatial axis, to uncover novel mechanisms of action across immunology.

Bibliography

1. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases* **20,** 533–534 (2020).
2. Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nature medicine* **26,** 842–844 (2020).
3. Delorey, T. M. *et al.* COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. *Nature* **595,** 107–113 (2021).
4. Melms, J. C. *et al.* A molecular single-cell lung atlas of lethal COVID-19. *Nature* **595,** 114–119 (2021).
5. Yang, A. C. *et al.* Dysregulation of brain and choroid plexus cell types in severe COVID-19. *Nature* **595,** 565–571 (2021).
6. Pita-Juarez, Y. *et al.* A single-nucleus and spatial transcriptomic atlas of the COVID-19 liver reveals topological, functional, and regenerative organ disruption in patients. *Genome Biology* **26,** 56 (2025).
7. Rendeiro, A. F. *et al.* The spatial landscape of lung pathology during COVID-19 progression. *Nature* **593,** 564–569 (2021).
8. Cross, A. R. *et al.* Spatial transcriptomic characterization of COVID-19 pneumonitis identifies immune circuits related to tissue injury. *JCI insight* **8,** e157837 (2023).
9. Lee, J. T. H. *et al.* Integrated histopathology, spatial and single cell transcriptomics resolve cellular drivers of early and late alveolar damage in COVID-19. *Nature Communications* **16,** 1979 (2025).
10. Regev, A. *et al.* The human cell atlas. *elife* **6,** e27041 (2017).

11. Consortium, T. M., coordination Schaum Nicholas 1 Karkanias Jim 2 Neff Norma F. 2 May Andrew P. 2 Quake Stephen R. quake@ stanford. edu 2 3 f Wyss-Coray Tony twc@ stanford. edu 4 5 6 g Darmanis Spyros spyros. darmanis@ czbiohub. org 2 h, O., coordination Batson Joshua 2 Botvinnik Olga 2 Chen Michelle B. 3 Chen Steven 2 Green Foad 2 Jones Robert C. 3 Maynard Ashley 2 Penland Lolita 2 Pisco Angela Oliveira 2 Sit Rene V. 2 Stanley Geoffrey M. 3 Webber James T. 2 Zanini Fabio 3, L. & data analysis Batson Joshua 2 Botvinnik Olga 2 Castro Paola 2 Croote Derek 3 Darmanis Spyros 2 DeRisi Joseph L. 2 27 Karkanias Jim 2 Pisco Angela Oliveira 2 Stanley Geoffrey M. 3 Webber James T. 2 Zanini Fabio 3, C. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562,** 367–372 (2018).
12. Schaum, N. *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris: The Tabula Muris Consortium. *Nature* **562,** 367 (2018).
13. Li, H. *et al.* Fly Cell Atlas: A single-nucleus transcriptomic atlas of the adult fruit fly. *Science* **375,** eabk2432 (2022).
14. Sikkema, L. *et al.* An integrated cell atlas of the lung in health and disease. *Nature medicine* **29,** 1563–1577 (2023).
15. Rozenblatt-Rosen, O. *et al.* The human tumor atlas network: charting tumor transitions across space and time at single-cell resolution. *Cell* **181,** 236–249 (2020).
16. Pijuan-Sala, B. *et al.* A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* **566,** 490–495 (2019).
17. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370,** eaba7721 (2020).
18. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184,** 3573–3587 (2021).
19. Novella-Rausell, C., Grudniewska, M., Peters, D. J. & Mahfouz, A. A comprehensive mouse kidney atlas enables rare cell population characterization and robust marker discovery. *IScience* **26** (2023).
20. Huynh, K. L. *et al.* Deconvolution of cell types and states in spatial multiomics utilizing TACIT. *Nature Communications* **16,** 3747 (2025).

21. Hickey, J. W. *et al.* Organization of the human intestine at single-cell resolution. *Nature* **619,** 572–584 (2023).

22. Shi, Q. *et al.* Cross-tissue multicellular coordination and its rewiring in cancer. *Nature,* 1–10 (2025).

23. Lagger, C. *et al.* scDiffCom: a tool for differential analysis of cell–cell interactions provides a mouse atlas of aging changes in intercellular communication. *Nature Aging* **3,** 1446–1461 (2023).

24. Consortium, T. T. M. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583,** 590–595 (2020).

25. Kimmel, J. C. *et al.* Murine single-cell RNA-seq reveals cell-identity-and tissue-specific trajectories of aging. *Genome research* **29,** 2088–2103 (2019).

26. Ma, Q., Li, Q., Zheng, X. & Pan, J. CellCommuNet: an atlas of cell–cell communication networks from single-cell RNA sequencing of human and mouse tissues in normal and disease states. *Nucleic Acids Research* **52,** D597–D606 (2024).

27. Hotez, P. J. *et al.* Helminth infections: the great neglected tropical diseases. *The Journal of clinical investigation* **118,** 1311–1321 (2008).

28. Maizels, R. M. *et al.* Helminth parasites–masters of regulation. *Immunological reviews* **201,** 89–116 (2004).

29. Reynolds, L. A., Filbey, K. J. & Maizels, R. M. *Immunity to the model intestinal helminth parasite Heligmosomoides polygyrus* in *Seminars in immunopathology* **34** (2012), 829–846.

30. Phythian-Adams, A. T. *et al.* CD11c depletion severely disrupts Th2 induction and development in vivo. *Journal of Experimental Medicine* **207,** 2089–2096 (2010).

31. Maizels, R. M. & Gause, W. C. Targeting helminths: The expanding world of type 2 immune effector mechanisms. *Journal of Experimental Medicine* **220,** e20221381 (2023).

32. Shang, K. *et al.* Regulation of the tuft cell-ILC2 circuit in intestinal mucosal immunity. *Frontiers in Immunology* **16,** 1568062 (2025).

33. Grainger, J. R. *et al.* Helminth secretions induce de novo T cell Foxp3 expression and regulatory function through the TGF-β pathway. *Journal of Experimental Medicine* **207,** 2331–2341 (2010).

34. Johnston, C. J. *et al.* A structurally distinct TGF-β mimic from an intestinal helminth parasite potently induces regulatory T cells. *Nature communications* **8,** 1741 (2017).

35. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551,** 333–339 (2017).

36. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome medicine* **9,** 75 (2017).

37. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5,** 621–628 (2008).

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

39. Li, L. & Clevers, H. Coexistence of quiescent and active adult stem cells in mammals. *science* **327,** 542–545 (2010).

40. Huang, S. Non-genetic heterogeneity of cells in development: more than just noise. *Development* **136,** 3853–3862 (2009).

41. Chen, G., Ning, B. & Shi, T. Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in genetics* **10,** 317 (2019).

42. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* **6,** 377–382 (2009).

43. Heumos, L. *et al.* Best practices for single-cell analysis across modalities. *Nature Reviews Genetics* **24,** 550–572 (2023).

44. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature biotechnology* **30,** 777–782 (2012).

45. Picelli, S. *et al.* Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nature methods* **10,** 1096–1098 (2013).

46. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nature biotechnology* **38,** 708–714 (2020).

47. Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161,** 1202–1214 (2015).

48. Zheng, G. X. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nature communications* **8,** 14049 (2017).

49. Kharchenko, P. V. The triumphs and limitations of computational methods for scRNA-seq. *Nature methods* **18,** 723–732 (2021).

50. Ding, J. *et al.* Systematic comparison of single-cell and single-nucleus RNA-sequencing methods. *Nature biotechnology* **38,** 737–746 (2020).

51. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nature methods* **11,** 41–46 (2014).

52. Brennecke, P. *et al.* Accounting for technical noise in single-cell RNA-seq experiments. *Nature methods* **10,** 1093–1095 (2013).

53. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature biotechnology* **32,** 896–902 (2014).

54. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nature methods* **14,** 381–387 (2017).

55. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology* **15,** e8746 (2019).

56. Suter, D. M. *et al.* Mammalian genes are transcribed with widely different bursting kinetics. *science* **332,** 472–474 (2011).

57. Kempe, H., Schwabe, A., Crémazy, F., Verschure, P. J. & Bruggeman, F. J. The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. *Molecular biology of the cell* **26,** 797–804 (2015).

58. Barron, M. & Li, J. Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data. *Scientific reports* **6,** 33892 (2016).

59. Griffiths, J. A., Scialdone, A. & Marioni, J. C. Using single-cell genomics to understand developmental processes and cell fate decisions. *Molecular systems biology* **14,** e8046 (2018).

60. Ilicic, T. *et al.* Classification of low quality cells from single-cell RNA-seq data. *Genome biology* **17,** 29 (2016).

61. Cheung, T. H. & Rando, T. A. Molecular regulation of stem cell quiescence. *Nature reviews Molecular cell biology* **14,** 329–340 (2013).

62. Monaco, G. *et al.* RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell reports* **26,** 1627–1640 (2019).

63. Osorio, D. & Cai, J. J. Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control. *Bioinformatics* **37,** 963–967 (2021).

64. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9,** giaa151 (2020).

65. Fleming, S. J. *et al.* Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nature methods* **20,** 1323–1335 (2023).

66. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell systems* **8,** 329–337 (2019).

67. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems* **8,** 281–291 (2019).

68. Germain, P.-L., Lun, A., Meixide, C. G., Macnair, W. & Robinson, M. D. Doublet identification in single-cell sequencing data using scDblFinder. *f1000research* **10,** 979 (2022).

69. DePasquale, E. A. *et al.* DoubletDecon: deconvoluting doublets from single-cell RNA-sequencing data. *Cell reports* **29,** 1718–1727 (2019).

70. Heaton, H. *et al.* Souporcell: robust clustering of single-cell RNA-seq data by genotype without reference genotypes. *Nature methods* **17,** 615–620 (2020).

71. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome biology* **20,** 273 (2019).

72. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature biotechnology* **36,** 89–94 (2018).

73. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature methods* **14,** 865–868 (2017).

74. Maier, T., Güell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. *FEBS letters* **583,** 3966–3973 (2009).

75. Villani, A.-C. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356,** eaah4573 (2017).

76. Abdelmohsen, K. *et al.* Identification of senescent cell subpopulations by CITE-seq analysis. *Aging Cell* **23,** e14297 (2024).

77. Pombo Antunes, A. R. *et al.* Single-cell profiling of myeloid cells in glioblastoma across species and disease stage reveals macrophage competition and specialization. *Nature neuroscience* **24,** 595–610 (2021).

78. Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome biology* **19,** 224 (2018).

79. Boggy, G. J. *et al.* BFF and cellhashR: analysis tools for accurate demultiplexing of cell hashing data. *Bioinformatics* **38,** 2791–2801 (2022).

80. Mulè, M. P., Martins, A. J. & Tsang, J. S. Normalizing and denoising protein expression data from droplet-based single cell profiling. *Nature communications* **13,** 2099 (2022).

81. Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S. & Marioni, J. C. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature methods* **14,** 565–571 (2017).

82. L. Lun, A. T., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome biology* **17,** 75 (2016).

83. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome biology* **20,** 296 (2019).

84. Booeshaghi, A. S., Hallgrímsdóttir, I. B., Gálvez-Merchán, Á. & Pachter, L. Depth normalization for single-cell genomics count data. *BioRxiv,* 2022–05 (2022).

85. Ahlmann-Eltze, C. & Huber, W. Comparison of transformations for single-cell RNA-seq data. *Nature Methods* **20,** 665–672 (2023).

86. Aggarwal, C. C., Hinneburg, A. & Keim, D. A. *On the surprising behavior of distance metrics in high dimensional space* in *International conference on database theory* (2001), 420–434.

87. Imoto, Y. *et al.* Resolution of the curse of dimensionality in single-cell RNA sequencing data analysis. *Life Science Alliance* **5** (2022).

88. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low dimensionality in gene expression data enables the accurate extraction of transcriptional programs from shallow sequencing. *Cell systems* **2,** 239–250 (2016).

89. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* **2,** 559–572 (1901).

90. Maaten, L. v. d. & Hinton, G. Visualizing data using t-SNE. *Journal of machine learning research* **9,** 2579–2605 (2008).

91. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

92. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome biology* **19,** 1–5 (2018).

93. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* **37,** 38–44 (2019).

94. Chari, T. & Pachter, L. The specious art of single-cell genomics. *PLOS Computational Biology* **19,** e1011288 (2023).

95. McQueen, J. B. *Some methods of classification and analysis of multivariate observations* in *Proc. of 5th Berkeley Symposium on Math. Stat. and Prob.* (1967), 281–297.

96. Ester, M., Kriegel, H.-P., Sander, J., Xu, X. *et al. A density-based algorithm for discovering clusters in large spatial databases with noise* in *kdd* **96** (1996), 226–231.

97. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nature biotechnology* **33,** 495–502 (2015).

98. Traag, V. A., Waltman, L. & Van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific reports* **9,** 1–12 (2019).

99. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology* **34,** 1145–1160 (2016).

100. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology* **36,** 411–420 (2018).

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

102. Tran, H. T. N. *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome biology* **21,** 12 (2020).

103. Stuart, T. *et al.* Comprehensive integration of single-cell data. *cell* **177,** 1888–1902 (2019).

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

105. Xu, C. *et al.* Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology* **17,** e9620 (2021).

106. Andreatta, M. *et al.* Semi-supervised integration of single-cell transcriptomics data. *Nature Communications* **15,** 872 (2024).

107. Luecken, M. D. *et al.* Benchmarking atlas-level data integration in single-cell genomics. *Nature methods* **19,** 41–50 (2022).

108. Chazarra-Gil, R., van Dongen, S., Kiselev, V. Y. & Hemberg, M. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. *Nucleic acids research* **49,** e42–e42 (2021).

109. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology* **20,** 163–172 (2019).

110. Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q. & Powell, J. E. scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome biology* **20,** 264 (2019).

111. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32,** 381–386 (2014).

112. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566,** 496–502 (2019).

113. Street, K. *et al.* Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics* **19,** 477 (2018).

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

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

116. Lange, M. *et al.* CellRank for directed single-cell fate mapping. *Nature methods* **19,** 159–170 (2022).

117. Weiler, P., Lange, M., Klein, M., Pe'er, D. & Theis, F. CellRank 2: unified fate mapping in multiview single-cell data. *Nature Methods* **21,** 1196–1205 (2024).

118. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome biology* **16,** 1–13 (2015).

119. Crowell, H. L. *et al.* Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature communications* **11,** 6077 (2020).

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

121. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *bioinformatics* **26,** 139–140 (2010).

122. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43,** e47–e47 (2015).

123. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nature methods* **15,** 255–261 (2018).

124. Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC bioinformatics* **20,** 40 (2019).

125. Das, S., Rai, A., Merchant, M. L., Cave, M. C. & Rai, S. N. A comprehensive survey of statistical approaches for differential expression analysis in single-cell RNA sequencing studies. *Genes* **12,** 1947 (2021).

126. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nature communications* **12,** 738 (2021).

127. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The innovation* **2** (2021).

128. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102,** 15545–15550 (2005).

129. Aibar, S. *et al.* SCENIC: single-cell regulatory network inference and clustering. *Nature methods* **14,** 1083–1086 (2017).

130. Andreatta, M. & Carmona, S. J. UCell: Robust and scalable single-cell gene signature scoring. *Computational and structural biotechnology journal* **19,** 3796–3798 (2021).

131. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC bioinformatics* **14,** 7 (2013).

132. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352,** 189–196 (2016).

133. Schubert, M. *et al.* Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature communications* **9,** 20 (2018).

134. Garcia-Alonso, L., Holland, C. H., Ibrahim, M. M., Turei, D. & Saez-Rodriguez, J. Benchmark and integration of resources for the estimation of human transcription factor activities. *Genome research* **29,** 1363–1375 (2019).

135. Holland, C. H. *et al.* Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome biology* **21,** 36 (2020).

136. Badia-i Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics advances* **2,** vbac016 (2022).

137. Armingol, E., Officer, A., Harismendy, O. & Lewis, N. E. Deciphering cell–cell interactions and communication from gene expression. *Nature Reviews Genetics* **22,** 71–88 (2021).

138. Almet, A. A., Cang, Z., Jin, S. & Nie, Q. The landscape of cell–cell communication through single-cell transcriptomics. *Current opinion in systems biology* **26,** 12–23 (2021).

139. Dimitrov, D. *et al.* Comparison of methods and resources for cell-cell communication inference from single-cell RNA-Seq data. *Nature communications* **13,** 3224 (2022).

140. Fabregat, A. *et al.* The reactome pathway knowledgebase. *Nucleic acids research* **46,** D649–D655 (2018).

141. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28,** 27–30 (2000).

142. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic acids research* **49,** D545–D551 (2021).

143. Szklarczyk, D. *et al.* The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research,* gkw937 (2016).

144. Troulé, K. *et al.* CellPhoneDB v5: inferring cell–cell communication from single-cell multiomics data. *Nature Protocols,* 1–29 (2025).

145. Ramilowski, J. A. *et al.* A draft network of ligand–receptor-mediated multicellular signalling in human. *Nature communications* **6,** 7866 (2015).

146. Jin, S. *et al.* Inference and analysis of cell-cell communication using CellChat. *Nature communications* **12,** 1088 (2021).

147. Türei, D., Korcsmáros, T. & Saez-Rodriguez, J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nature methods* **13,** 966–967 (2016).

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

149. Noël, F. *et al.* Dissection of intercellular communication using the transcriptome-based framework ICELLNET. *Nature communications* **12,** 1089 (2021).

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

151. Wang, S., Karikomi, M., MacLean, A. L. & Nie, Q. Cell lineage and communication network inference via optimization for single-cell transcriptomics. *Nucleic acids research* **47,** e66–e66 (2019).

152. Browaeys, R. *et al.* MultiNicheNet: a flexible framework for differential cell-cell communication analysis from multi-sample multi-condition single-cell transcriptomics data. *BioRxiv,* 2023–06 (2023).

153. Tsuyuzaki, K., Ishii, M. & Nikaido, I. scTensor detects many-to-many cell–cell interactions from single cell RNA-sequencing data. *BMC bioinformatics* **24,** 420 (2023).

154. Li, D. *et al.* TraSig: inferring cell-cell interactions from pseudotime ordering of scRNA-Seq data. *Genome biology* **23,** 73 (2022).

155. Dimitrov, D. *et al.* LIANA+ provides an all-in-one framework for cell–cell communication inference. *Nature Cell Biology* **26,** 1613–1622 (2024).

156. Wang, F. *et al.* RNAscope: a novel in situ RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *The Journal of molecular diagnostics* **14,** 22–29 (2012).

157. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nature methods* **11,** 417–422 (2014).

158. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353,** 78–82 (2016).

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

160. Merritt, C. R. *et al.* Multiplex digital spatial profiling of proteins and RNA in fixed tissue. *Nature biotechnology* **38,** 586–599 (2020).

161. He, S. *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nature biotechnology* **40,** 1794–1806 (2022).

162. Janesick, A. *et al.* High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature communications* **14,** 8353 (2023).

163. Karimi, E. *et al.* Method of the Year 2024: spatial proteomics. *Nat Methods* **21,** 2195–2196 (2024).

164. Lin, J.-R., Fallahi-Sichani, M., Chen, J.-Y. & Sorger, P. K. Cyclic immunofluorescence (CycIF), a highly multiplexed method for single-cell imaging. *Current protocols in chemical biology* **8,** 251–264 (2016).

165. Black, S. *et al.* CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nature protocols* **16,** 3802–3835 (2021).

166. Keren, L. *et al.* MIBI-TOF: A multiplexed imaging platform relates cellular phenotypes and tissue structure. *Science advances* **5,** eaax5851 (2019).

167. Gyau, B. B. *et al.* Multiplex Imaging Mass Cytometry Reveals Prognostic Immunosuppressive Subpopulations and Macrophage-Driven Metastasis in Osteosarcoma. *Cancers* **17,** 2780 (2025).

168. Eder, L. *et al.* Imaging mass cytometry in psoriatic disease reveals immune profile heterogeneity in skin and synovial tissue. *Journal of Investigative Dermatology* **145,** 1361–1370 (2025).

169. Böttcher, C. *et al.* Human microglia regional heterogeneity and phenotypes determined by multiplexed single-cell mass cytometry. *Nature neuroscience* **22,** 78–90 (2019).

170. Bollhagen, A. *et al.* High-resolution imaging mass cytometry to map subcellular structures. *Nature Methods,* 1–8 (2025).

171. Hu, B. *et al.* High-resolution spatially resolved proteomics of complex tissues based on microfluidics and transfer learning. *Cell* **188,** 734–748 (2025).

172. Tan, W. C. C. *et al.* Overview of multiplex immunohistochemistry/immunofluorescence techniques in the era of cancer immunotherapy. *Cancer Communications* **40,** 135–153 (2020).

173. Stringer, C., Wang, T., Michaelos, M. & Pachitariu, M. Cellpose: a generalist algorithm for cellular segmentation. *Nature methods* **18,** 100–106 (2021).

174. Pachitariu, M. & Stringer, C. Cellpose 2.0: how to train your own model. *Nature methods* **19,** 1634–1641 (2022).

175. Stringer, C. & Pachitariu, M. Cellpose3: one-click image restoration for improved cellular segmentation. *Nature methods* **22,** 592–599 (2025).

176. Greenwald, N. F. *et al.* Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning. *Nature biotechnology* **40,** 555–565 (2022).

177. Petukhov, V. *et al.* Cell segmentation in imaging-based spatial transcriptomics. *Nature biotechnology* **40,** 345–354 (2022).

178. Heidari, E. *et al.* Segger: Fast and accurate cell segmentation of imaging-based spatial transcriptomics data. *bioRxiv* (2025).

179. 10x Genomics. *Xenium In Situ Multimodal Cell Segmentation: Workflow and Data Highlights* tech. rep. CG000750. Technical Note, Rev B (10x Genomics, Pleasanton, CA, 2025). https://cdn.10xgenomics.com/image/upload/v1754601291/support-documents/CG000750_XeniumInSitu_CellSegmentation_TechNote_RevB.pdf (2025).

180. Palla, G. *et al.* Squidpy: a scalable framework for spatial omics analysis. *Nature methods* **19,** 171–178 (2022).

181. Dries, R. *et al.* Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology* **22,** 1–31 (2021).

182. Varrone, M., Tavernari, D., Santamaria-Martínez, A., Walsh, L. A. & Ciriello, G. CellCharter reveals spatial cell niches associated with tissue remodeling and cell plasticity. *Nature genetics* **56,** 74–84 (2024).

183. Singhal, V. *et al.* BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis. *Nature genetics* **56,** 431–441 (2024).

184. Long, Y. *et al.* Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications* **14,** 1155 (2023).

185. Qian, J. *et al.* Identification and characterization of cell niches in tissue from spatial omics data at single-cell resolution. *Nature Communications* **16,** 1693 (2025).

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

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

188. Shengquan, C., Boheng, Z., Xiaoyang, C., Xuegong, Z. & Rui, J. stPlus: a reference-based method for the accurate enhancement of spatial transcriptomics. *Bioinformatics* **37,** i299–i307 (2021).

189. Kleshchevnikov, V. *et al.* Cell2location maps fine-grained cell types in spatial transcriptomics. *Nature biotechnology* **40,** 661–671 (2022).

190. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature biotechnology* **40,** 517–526 (2022).

191. Tsoucas, D. *et al.* Accurate estimation of cell-type composition from gene expression data. *Nature communications* **10,** 2975 (2019).

192. Biancalani, T. *et al.* Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram. *Nature methods* **18,** 1352–1362 (2021).

193. Pham, D. *et al.* Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. *Nature communications* **14,** 7739 (2023).

194. Li, Z., Wang, T., Liu, P. & Huang, Y. SpatialDM for rapid identification of spatially co-expressed ligand–receptor and revealing cell–cell communication patterns. *Nature communications* **14,** 3995 (2023).

195. Arnol, D., Schapiro, D., Bodenmiller, B., Saez-Rodriguez, J. & Stegle, O. Modeling cell-cell interactions from spatial molecular data with spatial variance component analysis. *Cell reports* **29,** 202–211 (2019).

196. Schrod, S. *et al.* Spatial Cellular Networks from omics data with SpaCeNet. *Genome Research* **34,** 1371–1383 (2024).

197. Tang, Z., Zhang, T., Yang, B., Su, J. & Song, Q. spaCI: deciphering spatial cellular communications through adaptive graph model. *Briefings in Bioinformatics* **24,** bbac563 (2023).

198. Birk, S. *et al.* Quantitative characterization of cell niches in spatially resolved omics data. *Nature Genetics,* 1–13 (2025).

199. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS computational biology* **14,** e1006245 (2018).

200. Davis, S. *et al. seandavi/awesome-single-cell: 2018-06-20-1* version 2018-06-20-1. June 2018. https://doi.org/10.5281/zenodo.1294021.

201. Zappia, L. & Theis, F. J. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. *Genome biology* **22,** 301 (2021).

202. Köster, J. & Rahmann, S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* **28,** 2520–2522 (2012).

203. Mölder, F. *et al.* Sustainable data analysis with Snakemake. *F1000Research* **10,** 33 (2021).

204. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nature biotechnology* **35,** 316–319 (2017).

205. Interlandi, M., Kerl, K. & Dugas, M. InterCellar enables interactive analysis and exploration of cell- cell communication in single-cell transcriptomic data. *Communications biology* **5,** 21 (2022).

206. Sethi, R. *et al.* ezSingleCell: an integrated one-stop single-cell and spatial omics analysis platform for bench scientists. *Nature Communications* **15,** 5600 (2024).

207. Nyirenda, J. *et al.* Spatially resolved single-cell atlas unveils a distinct cellular signature of fatal lung COVID-19 in a Malawian population. *Nature Medicine,* 1–13 (2024).

208. MacDonald, L. *et al.* Human Fibroblast-Myeloid cell tissue atlas across lung, synovium, skin and heart. *bioRxiv,* 2025–03 (2025).

209. Campillo Poveda, M., Hardy, O., Laidlaw, R. F., Otto, T. D. & Maizels, R. M. Spatial transcriptomics reveals recasting of signalling networks in the small intestine following tissue invasion by the helminth parasite Heligmosomoides polygyrus. *BioRxiv,* 2024–02 (2024).

210. MacDonald, L. *et al.* Synovial tissue myeloid dendritic cell subsets exhibit distinct tissue-niche localization and function in health and rheumatoid arthritis. *Immunity* **57,** 2843–2862 (2024).

211. Lindeboom, R. G., Regev, A. & Teichmann, S. A. Towards a human cell atlas: taking notes from the past. *Trends in Genetics* **37,** 625–630 (2021).

212. Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A. & Regev, A. Impact of the Human Cell Atlas on medicine. *Nature medicine* **28,** 2486–2496 (2022).

213. Consortium*, T. T. S. *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376,** eabl4896 (2022).

214. Ahern, D. J. *et al.* A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* **185,** 916–938 (2022).

215. Majumder, P. P., Mhlanga, M. M. & Shalek, A. K. The Human Cell Atlas and equity: lessons learned. *Nature Medicine* **26,** 1509–1511 (2020).

216. Divangahi, M. *et al.* Trained immunity, tolerance, priming and differentiation: distinct immunological processes. *Nature immunology* **22,** 2–6 (2021).

217. Mangino, M., Roederer, M., Beddall, M. H., Nestle, F. O. & Spector, T. D. Innate and adaptive immune traits are differentially affected by genetic and environmental factors. *Nature communications* **8,** 13850 (2017).

218. Ilieva, M., Tschaikowski, M., Vandin, A. & Uchida, S. The current status of gene expression profilings in COVID-19 patients. *Clinical and Translational Discovery* **2,** e104 (2022).

219. Granados, A. A. *et al.* Single-nuclei characterization of pervasive transcriptional signatures across organs in response to COVID-19. *Elife* **12,** e81090 (2023).

220. Da Silva Filho, J. *et al.* A spatially resolved single-cell lung atlas integrated with clinical and blood signatures distinguishes COVID-19 disease trajectories. *Science Translational Medicine* **16,** eadk9149 (2024).

221. Alijotas-Reig, J. *et al.* Immunomodulatory therapy for the management of severe COVID-19. Beyond the anti-viral therapy: A comprehensive review. *Autoimmunity reviews* **19,** 102569 (2020).

222. Hall, M. W., Joshi, I., Leal, L. & Ooi, E. E. Immune immunomodulation in coronavirus disease 2019 (COVID-19): strategic considerations for personalized therapeutic intervention. *Clinical Infectious Diseases* **74,** 144–148 (2022).

223. De Jong, S. E. *et al.* Systems analysis and controlled malaria infection in Europeans and Africans elucidate naturally acquired immunity. *Nature Immunology* **22,** 654–665 (2021).

224. Franklin, B. S. *et al.* Malaria primes the innate immune response due to interferon-γ induced enhancement of toll-like receptor expression and function. *Proceedings of the National Academy of Sciences* **106,** 5789–5794 (2009).

225. Guha, R. *et al.* Plasmodium falciparum malaria drives epigenetic reprogramming of human monocytes toward a regulatory phenotype. *PLoS pathogens* **17,** e1009430 (2021).

226. Morton, B. *et al.* Distinct clinical and immunological profiles of patients with evidence of SARS-CoV-2 infection in sub-Saharan Africa. *Nature communications* **12,** 3554 (2021).

227. Breiman, R. F. *et al.* Postmortem investigations and identification of multiple causes of child deaths: An analysis of findings from the Child Health and Mortality Prevention Surveillance (CHAMPS) network. *PLoS medicine* **18,** e1003814 (2021).

228. Chawana, R. *et al.* Potential of minimally invasive tissue sampling for attributing specific causes of childhood deaths in South Africa: a pilot, epidemiological study. *Clinical Infectious Diseases* **69,** S361–S373 (2019).

229. Taylor, A. W. *et al.* Initial findings from a novel population-based child mortality surveillance approach: a descriptive study. *The Lancet Global Health* **8,** e909–e919 (2020).

230. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32,** 2847–2849 (2016).

231. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2021). https://www.R-project.org/.

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

233. Lab, S. *Azimuth Reference - Human PBMC* version 1.0.0. Feb. 2021. https://doi.org/10.5281/zenodo.4546839.

234. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *biorxiv,* 060012 (2016).

235. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27,** 1739–1740 (2011).

236. Liberzon, A. *et al.* The molecular signatures database hallmark gene set collection. *Cell systems* **1,** 417–425 (2015).

237. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature genetics* **25,** 25–29 (2000).

238. Aleksander, S. A. *et al.* The gene ontology knowledgebase in 2023. *Genetics* **224,** iyad031 (2023).

239. Carbon, S. *et al.* AmiGO: online access to ontology and annotation data. *Bioinformatics* **25,** 288–289 (2009).

240. Roelli, P., bbimber, Flynn, B., santiagorevale & Gui, G. *Hoohm/CITE-seq-Count: 1.4.2* version 1.4.2. Mar. 2019. https://doi.org/10.5281/zenodo.2590196.

241. Xin, H. *et al.* GMM-Demux: sample demultiplexing, multiplet detection, experiment planning, and novel cell-type verification in single cell sequencing. *Genome biology* **21,** 1–35 (2020).

242. Lun, A. T. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome biology* **20,** 1–9 (2019).

243. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. & Marioni, J. C. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nature communications* **9,** 1–6 (2018).

244. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29,** 24–26 (2011).

245. Thorvaldsdóttir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14,** 178–192 (2013).

246. Schiller, H. B. *et al.* The human lung cell atlas: a high-resolution reference map of the human lung in health and disease. *American journal of respiratory cell and molecular biology* **61,** 31–41 (2019).

247. Megill, C. *et al.* Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *bioRxiv,* 2021–04 (2021).

248. Brosseau, C., Colas, L., Magnan, A. & Brouard, S. CD9 tetraspanin: a new pathway for the regulation of inflammation? *Frontiers in immunology* **9,** 2316 (2018).

249. Chen, S. *et al.* Integration of spatial and single-cell data across modalities with weakly linked features. *Nature biotechnology* **42,** 1096–1106 (2024).

250. Li, K. *et al.* cellxgene VIP unleashes full power of interactive visualization, plotting and analysis of scRNA-seq data in the scale of millions of cells. *BioRxiv* **10,** 28–270652 (2020).

251. Su, H. C., Jing, H., Zhang, Y. & Casanova, J.-L. Interfering with interferons: A critical mechanism for critical COVID-19 pneumonia. *Annual review of immunology* **41,** 561–585 (2023).

252. Taks, E. J., Moorlag, S. J., Netea, M. G. & van der Meer, J. W. Shifting the immune memory paradigm: trained immunity in viral infections. *Annual Review of Virology* **9,** 469–489 (2022).

253. McCune, J. M. The dynamics of CD4+ T-cell depletion in HIV disease. *Nature* **410,** 974–979 (2001).

254. Yuan, C. *et al.* The role of cell death in SARS-CoV-2 infection. *Signal transduction and targeted therapy* **8,** 357 (2023).

255. Vannan, A. *et al.* Spatial transcriptomics identifies molecular niche dysregulation associated with distal lung remodeling in pulmonary fibrosis. *Nature genetics* **57,** 647–658 (2025).

256. Kong, L. *et al.* Single-cell and spatial transcriptomics of stricturing Crohn's disease highlights a fibrosis-associated network. *Nature Genetics,* 1–12 (2025).

257. Keller, M. S. *et al.* Vitessce: integrative visualization of multimodal and spatially resolved single-cell data. *Nature Methods* **22,** 63–67 (2025).

258. Quintana, J. F. *et al.* Single cell and spatial transcriptomic analyses reveal microglia-plasma cell crosstalk in the brain during Trypanosoma brucei infection. *Nature communications* **13,** 5752 (2022).

259. Grinberg, M. *Flask web development* (" O'Reilly Media, Inc.", 2018).

260. Bostock, M., Ogievetsky, V. & Heer, J. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics* **17,** 2301–2309 (2011).

261. Karlsson, M. *et al.* A single–cell type transcriptomics map of human tissues. *Science advances* **7,** eabh2169 (2021).

262. Dong, R. & Yuan, G.-C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome biology* **22,** 145 (2021).

263. Türei, D. *et al.* Integrated intra-and intercellular signaling knowledge for multicellular omics analysis. *Molecular systems biology* **17,** e9923 (2021).

264. Lundkvist, G. B., Kristensson, K. & Bentivoglio, M. Why trypanosomes cause sleeping sickness. *Physiology* **19,** 198–206 (2004).

265. Chiang, W. *et al.* Carcinoembryonic antigen-related cell adhesion molecule 6 (CEACAM6) promotes EGF receptor signaling of oral squamous cell carcinoma metastasis via the complex N-glycosylation. *Oncogene* **37,** 116–127 (2018).

266. Kobayashi, M *et al.* Carcinoembryonic antigen-related cell adhesion molecules as surrogate markers for EGFR inhibitor sensitivity in human lung adenocarcinoma. *British journal of cancer* **107,** 1745–1753 (2012).

267. Zheng, M.-Y. *et al.* Cytokine and epigenetic regulation of CEACAM6 mediates EGFR-driven signaling and drug response in lung adenocarcinoma. *NPJ Precision Oncology* **9,** 115 (2025).

268. Sikora, J., Harzer, K. & Elleder, M. Neurolysosomal pathology in human prosaposin deficiency suggests essential neurotrophic function of prosaposin. *Acta neuropathologica* **113,** 163–175 (2007).

269. He, Y. *et al.* Prosaposin maintains lipid homeostasis in dopamine neurons and counteracts experimental parkinsonism in rodents. *Nature communications* **14,** 5804 (2023).

270. Leng, L. *et al.* MIF signal transduction initiated by binding to CD74. *The Journal of experimental medicine* **197,** 1467–1476 (2003).

271. Matejuk, A. *et al.* MIF contribution to progressive brain diseases. *Journal of Neuroinflammation* **21,** 8 (2024).

272. Gil-Yarom, N. *et al.* CD74 is a novel transcription regulator. *Proceedings of the National Academy of Sciences* **114,** 562–567 (2017).

273. Heneka, M. T. *et al.* Neuroinflammation in Alzheimer's disease. *The Lancet Neurology* **14,** 388–405 (2015).

274. Lian, H. *et al.* Astrocyte-microglia cross talk through complement activation modulates amyloid pathology in mouse models of Alzheimer's disease. *Journal of Neuroscience* **36,** 577–589 (2016).

275. Wu, G. *et al.* The emerging roles of CEACAM6 in human cancer. *International journal of oncology* **64,** 27 (2024).

276. Russo, G. C. *et al.* E-cadherin interacts with EGFR resulting in hyper-activation of ERK in multiple models of breast cancer. *Oncogene* **43,** 1445–1462 (2024).

277. Yao, Z. *et al.* A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624,** 317–332 (2023).

278. Zhang, B. *et al.* A human embryonic limb cell atlas resolved in space and time. *Nature* **635,** 668–678 (2024).

279. Zeng, A. G. *et al.* Single-cell transcriptional atlas of human hematopoiesis reveals genetic and hierarchy-based determinants of aberrant AML differentiation. *Blood Cancer Discovery* **6,** 307–324 (2025).

280. Chu, Y. *et al.* Pan-cancer T cell atlas links a cellular stress response state to immunotherapy resistance. *Nature medicine* **29,** 1550–1562 (2023).

281. Liu, K. *et al.* Fibroblast atlas: Shared and specific cell types across tissues. *Science Advances* **11,** eado0173 (2025).

282. Wynn, T. A., Chawla, A. & Pollard, J. W. Macrophage biology in development, homeostasis and disease. *Nature* **496,** 445–455 (2013).

283. Yona, S. *et al.* Fate mapping reveals origins and dynamics of monocytes and tissue macrophages under homeostasis. *Immunity* **38,** 79–91 (2013).

284. Gomez Perdiguero, E. *et al.* Tissue-resident macrophages originate from yolk-sac-derived erythro-myeloid progenitors. *Nature* **518,** 547–551 (2015).

285. Blériot, C., Chakarov, S. & Ginhoux, F. Determinants of resident tissue macrophage identity and function. *Immunity* **52,** 957–970 (2020).

286. Okabe, Y. & Medzhitov, R. Tissue biology perspective on macrophages. *Nature immunology* **17,** 9–17 (2016).

287. Bain, C. C. *et al.* Constant replenishment from circulating monocytes maintains the macrophage pool in the intestine of adult mice. *Nature immunology* **15,** 929–937 (2014).

288. Bonnardel, J. *et al.* Stellate cells, hepatocytes, and endothelial cells imprint the Kupffer cell identity on monocytes colonizing the liver macrophage niche. *Immunity* **51,** 638–654 (2019).

289. Mulder, K. *et al.* Cross-tissue single-cell landscape of human monocytes and macrophages in health and disease. *Immunity* **54,** 1883–1900 (2021).

290. Gao, Y. *et al.* Cross-tissue human fibroblast atlas reveals myofibroblast subtypes with distinct roles in immune modulation. *Cancer Cell* **42,** 1764–1783 (2024).

291. Korsunsky, I *et al.* Cross-tissue, single-cell stromal atlas identifies shared pathological fibroblast phenotypes in four chronic inflammatory diseases. Med (New York, NY). 2022; 3 (7): 481. doi: 10.1016. *J. MEDJ* **2** (2022).

292. Qi, J. *et al.* Single-cell and spatial analysis reveal interaction of FAP+ fibroblasts and SPP1+ macrophages in colorectal cancer. *Nature communications* **13,** 1742 (2022).

293. Ke, D. *et al.* Macrophage and fibroblast trajectory inference and crosstalk analysis during myocardial infarction using integrated single-cell transcriptomic datasets. *Journal of Translational Medicine* **22,** 560 (2024).

294. Parigi, S. *et al. The spatial transcriptomic landscape of the healing mouse intestine following damage. Nat Commun 13: 828* 2022.

295. Moolenbeek, C & Ruitenberg, E. The 'Swiss roll': a simple technique for histological studies of the rodent intestine. *Laboratory animals* **15,** 57–60 (1981).

296. Larsson, L., Franzén, L., Ståhl, P. L. & Lundeberg, J. Semla: a versatile toolkit for spatially resolved transcriptomics analysis and visualization. *Bioinformatics* **39,** btad626 (2023).

297. Reina-Campos, M. *et al.* Tissue-resident memory CD8 T cell diversity is spatiotemporally imprinted. *Nature* **639,** 483–492 (2025).

298. Anthony, R. M., Rutitzky, L. I., Urban Jr, J. F., Stadecker, M. J. & Gause, W. C. Protective immune mechanisms in helminth infection. *Nature Reviews Immunology* **7,** 975–987 (2007).

299. Gieseck, R. L., Wilson, M. S. & Wynn, T. A. Type 2 immunity in tissue repair and fibrosis. *Nature Reviews Immunology* **18,** 62–76 (2018).

300. Filbey, K. J. *et al.* Innate and adaptive type 2 immune cell responses in genetically controlled resistance to intestinal helminth infection. *Immunology and cell biology* **92,** 436–448 (2014).

301. Sanjabi, S, Oh, S. & Li, M. *Regulation of the immune response by TGF-β: From conception to autoimmunity and infection. Cold Spring Harb. Persp. Biol. 9, a022236* 2017.

302. Maizels, R. M. Regulation of immunity and allergy by helminth parasites. *Allergy* **75,** 524–534 (2020).

303. Artis, D & Grencis, R. The intestinal epithelium: sensors to effectors in nematode infection. *Mucosal immunology* **1,** 252–264 (2008).

304. Nusse, Y. M. *et al.* Parasitic helminths induce fetal-like reversion in the intestinal stem cell niche. *Nature* **559,** 109–113 (2018).

305. Drurey, C. *et al.* Intestinal epithelial tuft cell induction is negated by a murine helminth and its secreted products. *Journal of Experimental Medicine* **219,** e20211140 (2021).

306. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12.** ISSN: 1471-2105. https://doi.org/10.1186/1471-2105-12-35 (2011).

307. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33,** 2938–2940 (2017).

308. Gu, Z., Gu, L., Eils, R., Schlesner, M. & Brors, B. " Circlize" implements and enhances circular visualization in R (2014).

309. Heumos, L. & Ansari, M. *sc-toolbox: Project templates and useful functions for single-cell analysis with Scanpy* https://github.com/schillerlab/sc-toolbox. GitHub repository. Version 0.12.3. MIT License. Archived on 2024-01-06. Accessed 2025-09-16. Schillerlab and Theislab, 2023.

310. Xu, H. *et al.* Transcriptional atlas of intestinal immune cells reveals that neuropeptide α-CGRP modulates group 2 innate lymphoid cell responses. *Immunity* **51,** 696–708 (2019).

311. Litviňuková, M. *et al.* Cells of the adult human heart. *Nature* **588,** 466–472 (2020).

312. Tucker, N. R. *et al.* Transcriptional and cellular diversity of the human heart. *Circulation* **142,** 466–482 (2020).

313. Wang, L. *et al.* Single-cell reconstruction of the adult human heart during heart failure and recovery reveals the cellular landscape underlying cardiac function. *Nature cell biology* **22,** 108–119 (2020).

314. Reyfman, P. A. *et al.* Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. *American journal of respiratory and critical care medicine* **199,** 1517–1536 (2019).

315. Madissoon, E. *et al.* scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome biology* **21,** 1 (2019).

316. Valenzi, E. *et al.* Single-cell analysis reveals fibroblast heterogeneity and myofibroblasts in systemic sclerosis-associated interstitial lung disease. *Annals of the rheumatic diseases* **78,** 1379–1387 (2019).

317. He, H. *et al.* Single-cell transcriptome analysis of human skin identifies novel fibroblast subpopulation and enrichment of immune subsets in atopic dermatitis. *Journal of Allergy and Clinical Immunology* **145,** 1615–1628 (2020).

318. Hughes, T. K. *et al.* Second-strand synthesis-based massively parallel scRNA-seq reveals cellular states and molecular features of human inflammatory skin pathologies. *Immunity* **53,** 878–894 (2020).

319. Solé-Boldo, L. *et al.* Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. *Communications biology* **3,** 188 (2020).

320. Alivernini, S. *et al.* Distinct synovial tissue macrophage subsets regulate inflammation and remission in rheumatoid arthritis. *Nature medicine* **26,** 1295–1306 (2020).

321. Micheroli, R. *et al.* Role of synovial fibroblast subsets across synovial pathotypes in rheumatoid arthritis: a deconvolution analysis. *RMD open* **8** (2022).

322. Stephenson, W. *et al.* Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation. *Nature communications* **9,** 791 (2018).

323. Zhang, F. *et al.* Defining inflammatory cell states in rheumatoid arthritis joint synovial tissues by integrating single-cell transcriptomics and mass cytometry. *Nature immunology* **20,** 928–942 (2019).

324. Yang, X. *et al.* SPP1 promotes the polarization of M2 macrophages through the Jak2/Stat3 signaling pathway and accelerates the progression of idiopathic pulmonary fibrosis. *International Journal of Molecular Medicine* **54,** 89 (2024).

325. King, E. M. *et al.* Gpnmb and Spp1 mark a conserved macrophage injury response masking fibrosis-specific programming in the lung. *JCI insight* **9,** e182700 (2024).

326. Matsubara, E. *et al.* The significance of SPP1 in lung cancers and its impact as a marker for protumor tumor-associated macrophages. *Cancers* **15,** 2250 (2023).

327. Cui, H. *et al.* TREM2 promotes lung fibrosis via controlling alveolar macrophage survival and pro-fibrotic activity. *Nature communications* **16,** 1761 (2025).

328. Gauthier, V. *et al.* Fibroblast heterogeneity: Keystone of tissue homeostasis and pathology in inflammation and ageing. *Frontiers in Immunology* **14,** 1137659 (2023).

329. Li, P., Sanz, I., O'Keefe, R. J. & Schwarz, E. M. NF-κB regulates VCAM-1 expression on fibroblast-like synoviocytes. *The journal of Immunology* **164,** 5990–5997 (2000).

330. Agassandian, M. *et al.* VCAM-1 is a TGF-β1 inducible gene upregulated in idiopathic pulmonary fibrosis. *Cellular signalling* **27,** 2467–2473 (2015).

331. Li, J, Chen, H, Zhang, D, Xie, J & Zhou, X. The role of stromal cell-derived factor 1 on cartilage development and disease. *Osteoarthritis and Cartilage* **29,** 313–322 (2021).

332. Cárdenas-León, C. G. *et al.* Matricellular proteins in cutaneous wound healing. *Frontiers in Cell and Developmental Biology* **10,** 1073320 (2022).

333. Jürgensen, H. J. *et al.* Cellular uptake of collagens and implications for immune cell regulation in disease. *Cellular and Molecular Life Sciences* **77,** 3161–3176 (2020).

334. Rosini, S. *et al.* Thrombospondin-1 promotes matrix homeostasis by interacting with collagen and lysyl oxidase precursors and collagen cross-linking sites. *Science signaling* **11,** eaar2566 (2018).

335. Chen, G. *et al.* EpCAM is essential for maintenance of the small intestinal epithelium architecture via regulation of the expression and localization of proteins that compose adherens junctions. *International Journal of Molecular Medicine* **47,** 621–632 (2021).

336. Entwistle, L. J. *et al.* Epithelial-cell-derived phospholipase A2 group 1B is an endogenous anthelmintic. *Cell host & microbe* **22,** 484–493 (2017).

337. Liao, Y., Xiao, N., Wang, X., Dai, S. & Wang, G. Promoting effect of Tmsb4x on the differentiation of peripheral blood mononuclear cells to dendritic cells during septicemia. *International Immunopharmacology* **111,** 109002 (2022).

338. Mahley, R. W. Apolipoprotein E: cholesterol transport protein with expanding role in cell biology. *Science* **240,** 622–630 (1988).

339. Lee, M. Y. & Griendling, K. K. Redox signaling, vascular function, and hypertension. *Antioxidants & redox signaling* **10,** 1045–1059 (2008).

340. Qu, M. *et al.* Establishment of intestinal organoid cultures modeling injury-associated epithelial regeneration. *Cell research* **31,** 259–271 (2021).

341. Shen, Y. *et al.* Matrix remodeling associated 7 proteins promote cutaneous wound healing through vimentin in coordinating fibroblast functions. *Inflammation and Regeneration* **43,** 5 (2023).

342. Van Dinther, M. *et al.* CD44 acts as a coreceptor for cell-specific enhancement of signaling and regulatory T cell induction by TGM1, a parasite TGF-β mimic. *Proceedings of the National Academy of Sciences* **120,** e2302370120 (2023).

343. Hoeft, K. *et al.* Platelet-instructed SPP1+ macrophages drive myofibroblast activation in fibrosis in a CXCL4-dependent manner. *Cell reports* **42** (2023).

344. Gu, Y. *et al.* Osteopontin is a therapeutic target that drives breast cancer recurrence. *Nature Communications* **15,** 9174 (2024).

345. MacDonald, L. *et al.* COVID-19 and RA share an SPP1 myeloid pathway that drives PD-L1+ neutrophils and CD14+ monocytes. *JCI insight* **6,** e147413 (2021).

346. Plikus, M. V. *et al.* Fibroblasts: Origins, definitions, and functions in health and disease. *Cell* **184,** 3852–3872 (2021).

347. Lang, F., Li, Y., Yao, R. & Jiang, M. Osteopontin in Chronic Inflammatory Diseases: Mechanisms, Biomarker Potential, and Therapeutic Strategies. *Biology* **14,** 428 (2025).

348. Gu, X., Kang, H., Cao, S., Tong, Z. & Song, N. Blockade of TREM2 ameliorates pulmonary inflammation and fibrosis by modulating sphingolipid metabolism. *Translational Research* **275,** 1–17 (2025).

349. Koenig, A. L. *et al.* Single-cell transcriptomics reveals cell-type-specific diversification in human heart failure. *Nature cardiovascular research* **1,** 263–280 (2022).

350. Jiang, T. *et al.* CardioAtlas: deciphering the single-cell transcriptome landscape in cardiovascular tissues and diseases. *Biomarker Research* **12,** 149 (2024).

351. Kuppe, C. *et al.* Spatial multi-omic map of human myocardial infarction. *Nature* **608,** 766–777 (2022).

352. Liu, W. *et al.* Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with PRECAST. *Nature communications* **14,** 296 (2023).

353. Wright, N., Carter, J & Irwin, M. The measurement of villus cell population size in the mouse small intestine in normal and abnormal states: a comparison of absolute measurements with morphometric estimators in sectioned immersion-fixed material. *Cell Proliferation* **22,** 425–450 (1989).

354. Cesaro, G. *et al.* Advances and challenges in cell–cell communication inference: a comprehensive review of tools, resources, and future directions. *Briefings in Bioinformatics* **26,** bbaf280 (2025).

355. Vento-Tormo, R. *et al.* Single-cell reconstruction of the early maternal–fetal interface in humans. *Nature* **563,** 347–353 (2018).

356. Abdelfattah, N. *et al.* Single-cell analysis of human glioma and immune cells identifies S100A4 as an immunotherapy target. *Nature communications* **13,** 767 (2022).

357. Armingol, E. *et al.* Context-aware deconvolution of cell–cell communication with Tensor-cell2cell. *Nature communications* **13,** 3665 (2022).

358. Jerby-Arnon, L. & Regev, A. DIALOGUE maps multicellular programs in tissue from single-cell or spatial transcriptomics data. *Nature biotechnology* **40,** 1467–1477 (2022).