# Scalable Integrative Frameworks for Multi-Omic Systems Biology

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennesse, Knoxville

Jean Merlet

August 2026

© by Jean Merlet, 2026
All Rights Reserved.

*To my wife Diliya for her unending support and motivation, to my cats Koshka and Aurora for their fluffiness, to my friends for listening, and to my family for their belief in me.*

# Acknowledgements

I would like to thank my advisor Dr. Daniel Jacobson for giving me a second chance and believing in me when no one else would, as well as for his support and advice throughout my PhD. I would also like to thank Dr. Helen Baghdoyan, Dr. Jon Beever, and Dr. Michael Langston for serving on my committee. I would like to thank Dr. Dana Graves, Dr. Kang Ko, and Dr. John Seykora for the wonderful years of collaboration. I would also like to extend special thanks to Dr. Kyle Sullivan, Dr. Anna Vlot, and Alice Townsend, as well as all of the other Jacobson Advanced Institute for Learning (JAIL) lab members, without which none of this would have been possible.

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Advanced Scientific Computing Research programs in the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

The Atopic Dermatitis work was supported by the National Institutes of Health under award numbers R01DA051908 and R01AR082951. The SARS-CoV-2 work was supported as part of the Genomic Sciences Program DOE Systems Biology Knowledgebase (KBase) funded by the Office of Biological and Environmental Research's Genomic Science program within the US Department of Energy Office of Science under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05–00OR22725, and DE-AC02-98CH10886. This work was also supported by the U.S. National Science Foundation (EF-2133763). I gratefully acknowledge all

data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

# Abstract

Network-based systems biology and scalable computational workflows can be used to integrate and interpret increasingly large and diverse biological datasets. Here, we investigate two biological case studies through this framework, spanning single-cell transcriptomics, spatial transcriptomics, viral genomics, and epidemiological data. First, in a study on atopic dermatitis, we integrate single-cell RNA sequencing and spatial transcriptomic datasets across murine and human systems using transcriptomic clustering, spatial neighborhood graphs, and multiplex random walk with restart diffusion. Using these approaches, we identify inflammatory fibroblast populations and immune-regulatory networks associated with eosinophilic inflammation. Second, we address the computational scalability challenges underlying modern, large-scale omics analyses through the development of HALRA, a sparse-preserving and distributed-memory implementation of the ALRA single-cell imputation algorithm designed to enable atlas-scale scRNA-seq analyses. Finally, we construct a large-scale, data-driven genomic similarity network through exhaustive comparisons of millions of whole-genome SARS-CoV-2 sequences, identifying over 22,000 biologically coherent viral lineages from the network topology using Markov clustering. We integrate these lineages with two years of weekly-resolution, county-level mortality and vaccination data by creating a spatiotemporal multiplex network. We investigate relationships between viral genomic structure and sociodemographic differences through random walk with restart diffusion and show lineage-specific spatial patterns associated with epidemiological variation.

# Table of Contents


**1 Background** **1**

1.1 Introduction . . . 1

1.2 Network-based Systems Biology . . . 2

1.2.1 Networks in the Ome-ome . . . 2

1.2.2 Multiplex Networks . . . 3

1.3 Single-cell RNA Sequencing . . . 5

1.3.1 Introduction to Transcriptomics . . . 5

1.3.2 Single-cell RNA Sequencing Workflows . . . 8

1.3.3 Challenges and Future Directions . . . 25

1.4 Spatial Transcriptomics . . . 27

1.4.1 Spatial Context in Transcriptomics . . . 27

1.4.2 Spatial Transcriptomics Workflows . . . 28

1.4.3 Challenges and Future Directions . . . 32

1.5 Scalable Omics Computing . . . 34

1.5.1 Sparse and High-Dimensional Biological Data . . . 35

1.5.2 Distributed Computing in Computational Biology . . . 37

1.6 Large-Scale Viral Genomics . . . 38

1.6.1 Introduction to Viral Genomics . . . 39

1.6.2 Next-Generation Sequencing and Genomic Surveillance . . . 39

1.6.3 Scaling Challenges in Viral Genomics . . . 41

1.6.4 Underuse of Large Public Viral Genomic Datasets . . . 42

1.7 Conclusion . . . . . 42

**2 Cross-Species, Integrated Multi-Omics Reveal Immunomodulatory Fibroblast Niches in Atopic Dermatitis 44**

2.1 Introduction . . . . . 45
2.1.1 Cellular and Spatial Complexity in AD . . . . . 45
2.1.2 NF-$\kappa$B and Mouse Models . . . . . 47
2.1.3 Study Objectives . . . . . 48
2.2 Results and Discussion . . . . . 49
2.2.1 The Eosinophilic Role of Inflammatory Fibroblasts . . . . . 51
2.2.2 Human AD and Data Integration with MENTOR . . . . . 53
2.2.3 Spatial Context . . . . . 55
2.3 Methods . . . . . 58
2.3.1 Murine scRNA-seq Dataset . . . . . 58
2.3.2 Human scRNA-seq Datasets . . . . . 60
2.3.3 Single-cell Spatial Transcriptomics . . . . . 62
2.4 Conclusion . . . . . 64

**3 HALRA: a High-performance and Parallel Implementation of the ALRA Imputation Algorithm 66**

3.1 Introduction . . . . . 67
3.1.1 Imputation Algorithms . . . . . 67
3.1.2 Computational Performance . . . . . 70
3.1.3 Study Objectives . . . . . 72
3.2 Results and Discussion . . . . . 73
3.2.1 Memory Efficiency in Sparse Matrices . . . . . 73
3.2.2 Benchmarking and Validation . . . . . 77
3.2.3 Distributed Sparse Reconstruction Architecture . . . . . 80
3.2.4 Generalizable High-Performance Design Principles . . . . . 82
3.3 Methods . . . . . 85

3.3.1 The ALRA Algorithm . . . 85
3.3.2 Sparse Matrix Implementation . . . 87
3.3.3 MPI-Based Distributed Implementation . . . 87
3.4 Conclusion . . . 89

**4 Whole-Genome SARS-CoV-2 Lineage Classification and Multiplex Integration with U.S. Mortality and Vaccination Trends 91**
4.1 Introduction . . . 92
4.1.1 SARS-CoV-2 Genomes . . . 92
4.1.2 Large-scale Viral Genome Comparisons . . . 93
4.1.3 Study Objectives . . . 95
4.2 Results and Discussion . . . 97
4.2.1 High-Resolution, Data-Driven Lineages . . . 97
4.2.2 Integrating Sociodemographic Data . . . 102
4.2.3 Null Model Validation of Multiplex Diffusion . . . 105
4.3 Methods . . . 108
4.3.1 SARS-CoV-2 Genome Sequence Preprocessing . . . 108
4.3.2 Genome Vector Comparisons . . . 109
4.3.3 California Multiplex Network . . . 112
4.4 Conclusion . . . 118

**5 Conclusion 120**

**Bibliography 123**

**Appendix 168**

**Vita 173**

# List of Figures


1.1 A three-layer multiplex network constructed from genomic, transcriptomic, and proteomic data layers. . . . 4
1.2 Outline of common scRNA-seq workflows. . . . 13
1.3 UMAP visualization of scRNA-seq clustering and annotation. . . . 19
1.4 Comparison of cell type clusters with PCA, t-SNE, and UMAP projections of the cells into two dimensions. . . . 24
1.5 Overview of a typical spatial transcriptomic analysis workflow using Xenium data. . . . 29
1.6 Comparison of dense and sparse matrix representations commonly encountered in omics data analysis. . . . 35
1.7 Approximate cumulative growth of publicly available SARS-CoV-2 genome sequences deposited in GISAID from January 2020 through May 2026. . . . 40

2.1 Clustering and annotation of the WT and cKO mouse skin scRNA-seq data. . . . 50
2.2 An immune-regulatory inflammatory fibroblast population in cKO mice. 52
2.3 Conceptual model of all the genes in an immune-regulatory and ECM-related MENTOR clade of fibroblast-mediated inflammation of atopic dermatitis. . . . 54
2.4 Single-cell and spatial transcriptomic characterization of inflammatory regions in murine atopic dermatitis skin. . . . 57

3.1 Growth in the number of cells profiled by landmark scRNA-seq datasets over time (red) compared with increases in CPU transistor counts (black). 71
3.2 Computational benchmarking of HALRA and pyALRA across increasing scRNA-seq dataset sizes. . . . . 75
3.3 Distributed scaling behavior of HALRA across increasing scRNA-seq dataset sizes. . . . . 79
3.4 Overview of the ALRA imputation workflow. . . . . 86

4.1 Visual representation of cluster creation during Markov Clustering (MCL). . . . . 96
4.2 Longitudinal view of the four most sequenced SARS-CoV-2 WHO variants during the first ∼2.5 years of the COVID-19 pandemic, separated by variant. . . . . 98
4.3 Snapshot of the distribution of CoMet + HipMCL clusters (i.e., SARS-CoV-2 lineages) on August 9th, 2021, during the first peak of the Delta variant. . . . . 101
4.4 County-level mortality diffusion patterns associated with three of the largest SARS-CoV-2 HipMCL genomic clusters across California from 2020-W15 through 2022-W06, alongside statewide vaccine coverage at 2021-W37. . . . . 104
4.5 County-week heatmap of observed versus null mortality diffusion z-scores following Alpha-period genomic seeding from 2021-W01 through 2021-W30. . . . . 106

1 County-week heatmap of observed versus null mortality diffusion z-scores across the statewide SARS-CoV-2 multiplex network from 2020-W15 through 2022-W06. . . . . 170
2 UMAP visualization of the three major transcriptomic cell populations in murine single-cell spatial transcriptomic data . . . . . 171
3 Area-normalized cell type composition across murine Xenium samples. 172

# Chapter 1

# Background

## 1.1 Introduction

The rapid advancement of high-throughput sequencing technologies has generated an unprecedented volume of biological data, beginning with the genomics revolution and now encompassing transcriptomics, proteomics, metabolomics, and numerous other -omics data modalities [93, 103]. These increasingly large and diverse datasets offer extraordinary opportunities for biological discovery but also demand computational methods and analytical workflows that can operate at scale and integrate multiple data modalities. This dissertation addresses these challenges by drawing from three intersecting domains: single-cell transcriptomics, viral genomics, and the design and implementation of scalable algorithms tailored to these data types. This background chapter frames these domains through the lens of network-based systems biology and focuses in depth on single-cell RNA sequencing (scRNA-seq), which is highly relevant to two of the subsequent chapters. Forthcoming sections in this chapter will discuss spatial transcriptomics as a complementary layer in multi-omic studies, viral genomics in the context of SARS-CoV-2 evolution, and how scalable algorithm design enables analysis of petabyte-scale datasets on modern supercomputing platforms.

## 1.2 Network-based Systems Biology

Systems biology is an interdisciplinary approach that aims to understand complex biological systems as integrated wholes rather than collections of isolated components. Unlike reductionist methods, which focus on individual parts, systems biology emphasizes holistic modeling and analysis to capture the complex interactions that give rise to emergent properties. The field began to take shape in the mid-20th century and expanded rapidly with the advent of genomics, computational modeling, and high-throughput technologies in the late 20th century [267]. Its growth has continued alongside the increasing scale and complexity of biological datasets. While all modern systems biology approaches rely on mathematical and computational tools, they vary widely in implementation [288]. Among these, network-based systems biology methods offer a powerful framework for integrating diverse omics data by mapping the datasets and their interactions as networks, enabling multi-scale and cross-modal biological inference.

### 1.2.1 Networks in the Ome-ome

The rapid expansion of omics technologies has produced increasingly large, complex, and high-dimensional biological datasets spanning the genome, transcriptome, proteome, epigenome, and a growing number of omics types [56], collectively referred to as the "ome-ome". This wealth of multi-modal data presents both an opportunity and a challenge. While it enables deeper, data-driven insights into biological systems based on multiple data modalities and scales, it also requires scalable computational frameworks capable of handling issues such as high memory consumption and time complexity, data sparsity and high-dimensionality, and integration of different data modalities from separate experiments.

Biological data from different omics types are commonly modeled as networks by using graph representations, typically with molecular units (e.g., genes) as the nodes and their relationships (e.g., co-expression) as the edges between them.

These biological networks capture diverse kinds of molecular relationships, including gene regulatory networks, protein–protein interaction networks, metabolic networks, signaling pathways, co-expression networks, and gene–disease association networks [303]. Numerous network-based methods have been developed to analyze these structures, such as various community detection algorithms, network propagation algorithms, diffusion algorithms, and graph-based machine learning algorithms [34, 274, 102, 77]. Across these networks, recurring topological features such as hubs (highly connected nodes), motifs (recurring regulatory patterns), modules (densely connected subgraphs), and scale-free architectures provide valuable insights into biological robustness, hierarchy, and functional organization. For example, identifying hubs in a PPI network can reveal potential drug targets [82] and community detection in multiple, separate co-expression networks can uncover functionally related gene clusters [236]. However, when studying multi-omic systems, separate single-layer networks often fail to capture the full complexity of cross-modal relationships [3, 134], motivating the use of multiplex networks, which explicitly model both within- and between-layer interactions.

### 1.2.2 Multiplex Networks

Multiplex-network-based systems biology provides a conceptual foundation for linking and combining diverse omics datasets by representing each as a distinct yet interconnected layer in a multiplex network, where the same type of node (e.g., genes) appears across layers, connected by different types of interactions within layers (e.g., protein–protein interactions, regulatory relationships) and linked across layers when referring to the same biological entity (Fig. 1.1). These networks capture both within-layer relationships (e.g., gene co-expression) and between-layer interactions (e.g., regulatory links between the epigenome and transcriptome), enabling a more comprehensive view of biological function. Network-based methods for the analysis of biological multiplex networks can be grouped into modularity

optimization algorithms, such as InfoMap, MolTi, and Generalized Louvain, [60, 64, 61]; random-walk-based algorithms such as MultiVERSE, Multi-Net, and MENTOR [221, 12, 257]; and graph neural network algorithms such as GEAM, MALGNN, and MPXGAT [276, 48, 25]. Random-walk and neural network methods generate lower-dimensionality network embeddings of the multiplex networks as part of their workflows, while most modularity optimization algorithms do not, and only several dozen network-based methods for analysis of multiplex networks have been published [221, 186]. As multiplex networks with many layers, large layers, or many large layers can be highly computationally demanding, scalability is a major concern in the majority of these publications. In this dissertation, we apply the multiplex network embedding and distributed interpretation techniques of MENTOR in the context of multi-omic mouse and human atopic dermatitis data integration (see chapter 2).

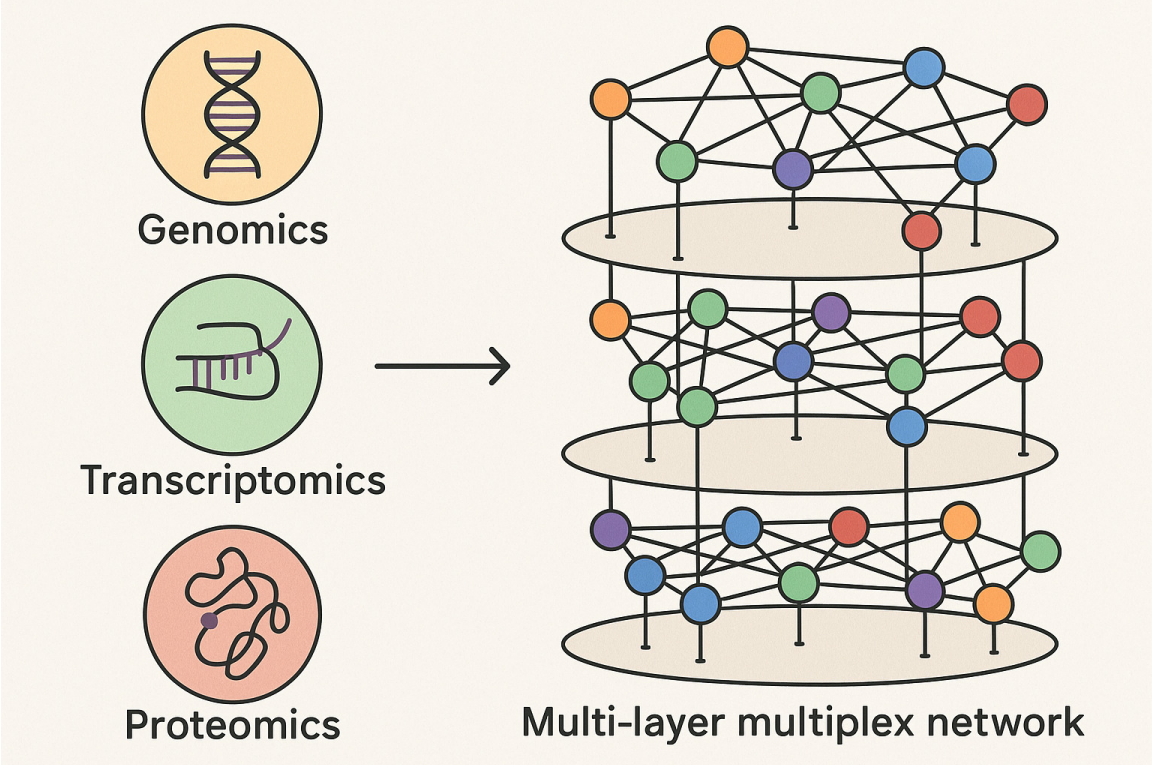



**Figure 1.1:** A three-layer multiplex network constructed from genomic, transcriptomic, and proteomic data layers.

Hundreds of thousands of omics datasets are publicly available in databases, but the majority of these datasets are only used once [218, 215]. Subsets of these omics layers can be combined to generate millions of multiplex networks that represent an integrated view of varied and complex biological systems containing biological information greater than the sum of their parts. This underscores that multiplex networks not only enable the study of complex biological systems through the integration of different data modalities and scales, but also that there exists a vast, largely untapped reservoir of biological data with which to do so.

## 1.3 Single-cell RNA Sequencing

**Note:** Although single-cell RNA sequencing (scRNA-seq) and single-nucleus RNA sequencing (snRNA-seq) differ in certain technical details, for clarity the term "scRNA-seq" will be used throughout this section to refer to both except when explaining the differences.

### 1.3.1 Introduction to Transcriptomics

Transcriptomics is the study of the transcriptome: the set of all ribonucleic acid (RNA) molecules within an organism. RNA molecules either function as blueprints for the synthesis of proteins (messenger RNA) or perform biological tasks themselves (non-coding RNA). Each time a cell expresses a gene, it uses the information encoded in a gene to create a protein (protein-coding genes) or a functional RNA molecule (non-coding genes). Messenger RNA (mRNA) is created as part of the protein creation process when protein-coding genes are expressed, while non-coding RNA is created when non-coding genes are expressed. To date, over 78,000 genes in the human genome have been identified, of which nearly 20,000 are protein-coding [203].

### A Brief History of RNA sequencing

The quantification of gene expression started in the 1970s with Northern Blotting and in-situ hybridization [5, 214], which were able to quantify one or a very small number of genes at a time. In the 1990s, Patrick Brown developed the first microarray, able to measure the gene expression of dozens and soon after hundreds of genes [240]. Next-generation sequencing (NGS) technologies, massively parallel adaptations of the original DNA sequencing technologies, were developed in the early 2000s [238]. These NGS technologies were then adapted in 2006 to detect and quantify RNA in various cell samples [13, 45, 74, 280], marking the beginning of RNA-seq, which became the widely-known bulk RNA-seq in 2008.

In 2009, Tang et al. measured the expression profile of an individual cell for the first time [261]. Since then, multiple new single-cell protocols and technologies have been developed and improved upon every year. Early years saw the introduction of multiple techniques that were later further built upon, including the introduction of reverse-transcription for 96 single cells with STRT-seq in 2011 [129], full-length mRNA sequencing of 42 individual cells with SMART-seq [230] in 2012, the introduction of single-nuclei RNA sequencing (snRNA-seq) of eight nuclei in 2013 [97], the development of in-vitro amplification with CEL-seq [112] in 2014, and the first use of the long-established fluorescence-activated cell sorting (FACS) technology to sort more than 1,500 individual cells prior to sequencing with MARS-seq [131], also in 2014.

The affordability and widespread use of scRNA-seq improved markedly with the introduction massively-parallel, droplet-based technologies, starting with Drop-seq [184] and quickly followed by inDrops [301], 10x Chromium [298], and the first snRNA-seq droplet-based technology, DroNc-seq [104]. The rapid development of these technologies has improved scRNA-seq from the transcriptomic profiling of a single cell in 2009 to the profiling of millions of individual cells in so-called cell atlases [290, 247, 263] in less than a decade. As the scale, affordability, and use of scRNA-seq

**Table 1.1:** Major technological developments in RNA sequencing and single-cell transcriptomics.

| **Technology** | **Release Year** | **Description** |
|---|---|---|
| Northern Blotting | 1970s | Early method for measuring expression of individual genes in bulk samples. |
| In-situ Hybridization | 1970s | Spatial detection of gene expression within tissues and cells. |
| Microarray | 1990s | First high-throughput transcriptomic platform capable of measuring dozens to hundreds of genes simultaneously. |
| bulk RNA-seq | 2006-2008 | Adaptation of NGS technologies for transcriptome-wide RNA quantification. |
| Tang et al. method | 2009 | First transcriptomic profiling of an individual cell. |
| STRT-seq | 2011 | Early single-cell protocol introducing reverse transcription for 96 individual cells. |
| SMART-seq | 2012 | Full-length mRNA sequencing protocol for single cells. |
| snRNA-seq | 2013 | First sequencing of RNA isolated from individual nuclei rather than whole cells. |
| CEL-seq | 2014 | Introduced in-vitro amplification for single-cell transcriptomics. |
| MARS-seq | 2014 | Combined FACS sorting with massively parallel single-cell sequencing. |
| Drop-seq | 2015 | High-throughput, droplet-based scRNA-seq platform enabling large-scale transcriptomic profiling. |
| inDrops | 2015 | High-throughput, droplet-based scRNA-seq platform enabling large-scale transcriptomic profiling. |
| 10x Chromium | 2015 | High-throughput, droplet-based scRNA-seq platform enabling large-scale transcriptomic profiling. |
| DroNc-seq | 2017 | First droplet-based massively parallel single-nuclei RNA sequencing (snRNA-seq) technology. |

has exploded from measuring a single gene in a group of cells to nearly 60,000 genes in hundreds of millions of cells, so too has the data generated from these experiments and consequent need for improved computational methods.

#### Genes and Cells as a High-Dimensional System

After alignment to a reference genome (see section 1.3.2), the raw reads produced by scRNA-seq experiments are converted to a matrix of cells by genes or genes by cells, representing the transcriptomic landscape of the cells. Because a typical dataset contains tens of thousands of cells and 35,000 or so genes after pre-processing, they are high-dimensional systems that require various computational methods to analyze. Visualization of the cells in a dataset, just like all visualizations, must be done in two to four dimensions, requiring the use of dimensionality reduction techniques such as PCA, t-SNE, and UMAP [81, 270, 194] to project the transcriptome down to two dimensions. To identify groups of cells with similar gene expression profiles (see section 1.3.2), such as for cell typing, the dimensionality of genes is often reduced to 50 or 100 dimensions before generating k nearest-neighbor graph embeddings and clustering cells together [84, 266]. On the other hand, comparisons of individual gene expression between groups of cells, such as when performing differential expression analyses [178, 187], cannot reduce the dimensions of genes and consequently are often the most computationally intensive techniques in a scRNA-seq analysis. As the size of scRNA-seq datasets has grown, so too have the number of new approaches and algorithms for use in analyses. While these are only some of the most common methods used in scRNA-seq analyses, they highlight some of the challenges in effectively extracting and conveying meaningful biological insights from the high dimensionality of single-cell transcriptomic systems.

### 1.3.2 Single-cell RNA Sequencing Workflows

Before a scRNA-seq dataset can be analyzed, it must be created. This involves experimental design, preparing cell samples, running the experiment in a laboratory or genomics core, and obtaining and transferring the large, digitized dataset to an appropriate storage medium. The subsequent computational analysis can be performed in numerous ways, ranging from end-to-end, largely automated analyses

(e.g., Cellranger [298] or ezSingleCell [242]); to extensive code libraries and tutorials (e.g., Seurat [109] or Scanpy [284]); to custom user-written scripts; or to any combination of these. ScRNA-seq workflows vary considerably from technology to technology, so the discussion here will focus on the the three most prevalent droplet-based, NGS technologies: Drop-seq [184], inDrops [301] and 10x Chromium [298] because these are the most widely-used and most relevant to this dissertation.

### Experimental Design and Sample Preparation

As in any biological experiment, scRNA-seq experiments should be designed with a minimum of three biological replicates per condition (e.g., control or treated) to be able to test whether any observed differences between conditions are statistically significant or simply due to chance. Although each cell sample will yield gene expression profiles for thousands of individual cells, the cells themselves are not biological replicates because they are not independent; they come from the same organism. Additionally, there are no true technical replicates in scRNA-seq experiments because cells are lysed during the library preparation process, but sample multiplexing and batch sequencing strategies limit technical variation.

ScRNA-seq technology can be used to generate expression profiles for a cell sample from any eukaryotic tissue and species so long as a cell suspension can be created from that cell sample, but preparing cell samples from certain types of tissues provides unique challenges, some of which have not yet been overcome. Current high-throughput, droplet-based scRNA-seq experiments require a cell or nuclei suspension for library generation, not clumps or tissue fragments. Methods exist for creating cell suspensions from a variety of sample types, such as fresh, frozen, or formalin-fixed, paraffin-embedded (FFPE) tissues; cell cultures; or fixed whole blood samples, among others. Some of these methods (e.g., Ficoll-Paque for blood cells, [75]) were developed more than 50 years ago, and methods to prepare single-cell suspensions from most soft tissues (e.g., human skin, [237]) were developed fairly quickly. But creating single-cell suspensions from certain tissues, such as the lignified and thickened

cell walls of woody plants or highly interconnected neurons in brain tissue, remains a challenge today that has only partially been solved by the development of single-nuclei methods [26, 14, 285]. Once a viable single-cell or single-nuclei suspension is obtained, the next step is to capture and process the mRNA for sequencing within each cell through library preparation.

### Library Preparation, Sequencing, and Dropout Events

ScRNA-seq libraries are prepared pools of complementary DNA (cDNA) fragments. Library preparation from single-cell suspensions with droplet-based technologies consists of: (i) using microfluidics to combine individual cells with individual hydrogel or resin beads in nanoliter-sized oil-water emulsions known as droplets, (ii) lysing cells with reagents contained in the emulsion to release their mRNA, (iii) ligating mRNA to the pre-fabricated nucleotide sequences released from the beads, (iv) reverse transcribing mRNA to cDNA, and (v) amplifying and fragmenting the cDNA prior to sequencing [184, 301, 298]. Droplet-based scRNA-seq experiments target protein-coding genes when ligating the poly-A tails of mature mRNAs, but this process also ligates many other types of genes with poly-A tails, including some non-coding genes, mitochondrial genes, ribosomal genes, and pseudogenes.

Each microscopic bead surface is covered with more than 100 million nucleotide sequences, called oligonucleotides, which are composed of four parts: (i) a primer for cDNA amplification (identical on all beads), (ii) a cell barcode for post-sequencing cell identification (identical on each bead, different between any two beads), (iii) a unique molecular identifier (UMI) to enable deduplication after cDNA amplification (different on each oligonucleotide on a bead) [156], and (iv) a short, repeat sequence of deoxythymidines known as an oligo-dT sequence to ligate the poly-A tail of mature mRNAs and prime them for reverse transcription. During the reverse transcription of mRNA to cDNA by reverse transcriptase enzymes in the emulsion, both Drop-seq and 10x Chromium employ a template-switching oligo (TSO) to add a primer for polymerase chain reaction (PCR) amplification to the end of the cDNA, while

inDrops does not. Since all of the cDNA fragments have identifying cell barcodes and UMIs, they are pooled together prior to amplification and sequencing to allow for time- and cost-cutting mass reactions.

Because there is only a small amount of mRNA in each cell, the mRNA is artificially replicated, or amplified, prior to sequencing to maximize the likelihood of detection during sequencing. The combined oligonucleotide and cDNA sequences for Drop-seq and 10x Chromium are amplified through multiple rounds of PCR amplification [205], while inDrops uses a single round of in-vitro amplification based on the CEL-seq protocol [112]. After amplification, the oligonucleotide and cDNA sequences are enzymatically fragmented into 200 to 500 nucleotide-long fragments, end-repaired, A-tailed, and ligated with adapters for sequencing using tagmentation or zinc-ion-mediated cleavage [151]. All three scRNA-seq technologies then use various Illumina [18] paired-end sequencing technologies to further amplify, tag with sample indices, and sequence the oligonucleotide and cDNA read fragments. One end of the fragment is sequenced just enough to capture the cell barcode and UMI, and the other end of the fragment is sequenced to capture the first 50 to 150 nucleotides of the partial cDNA sequence, depending on the technology.

Sequencing generates hundreds of millions of reads of the amplified and pooled fragments containing the cDNA nucleotide sequences along with their identifying cell, sample, and transcript of origin nucleotide identifiers. However, even with amplification, a transcript detected at low to moderate levels in one cell is often undetected in another cell even of the same cell type for both biological reasons, such as transcriptomic shifts due to subtle differences in cell state and the natural stochasticity of gene expression, and technical reasons such as low initial mRNA levels within individual cells and disparate reverse transcription of transcripts or inefficient capture [47]. The term "dropout event" has, confusingly, been used to refer to one or both of these sources of unobserved transcripts [164], and we will thus specify biological dropout, technical dropout, or both throughout this chapter.

### Fastq File Format

Depending on the sequencing technology used, sequencing reads are either output directly into fastq format (e.g., PacBio sequencing) or are first output into another intermediary format (e.g., Illumina's binary base call format) before being converted to fastq format with the appropriate software, such as bcl2fastq for Illumina [52, 72, 127]. The read and sequencing adapters, poly-A sequence, and TSO are removed from the reads, which is known as adapter or read trimming. Adapters are often trimmed as part of fastq file generation such as when using bcl2fastq, but can also be manually removed with various tools such as cutadapt [188] or trimmomatic [24]. The poly-A and TSO sequences can be trimmed with these tools or as part of the alignment process (see next section). In either case, read sample indices and sequencing lane indices are used during file generation to split files into separate samples while combining sequencing lanes for the same samples, known as demultiplexing. This process generates files in the plaintext fastq file format, which is composed of four newline-separated lines for each sequencing read:

```
@SRR14253415.1 K00315:218:H7NHWBBXY:5:1101:1732:1543 length=28
TGGGCGTTNACTTACTACAAATGCGGTT
+
AAFAFJJJ#JJJJJJJJJJJJJFJJJJ
```

The first line begins with "@" and contains a sequencing technology- and data source-dependent sequencing id, often containing sequencing metadata such as flowcell lane, sample id, and read number. In the above example, the first line begins with a gene expression omnibus (GEO) accession number ending in a ".1" to indicate the first read in that fastq file, followed by Illumina sequencing technology-specific ids particular to the machine that was used in sequencing and ending with the length of the nucleotide sequence in the second line, added by GEO. The second line is the raw cDNA sequence information, and is typically mostly composed of the four DNA nucleotide bases, A, C, T, and G, but can also contain the ambiguous nucleotide

code N when the sequencing technology was unable to determine the exact base. In the example, the 28 nucleotides are a concatenated 16 nucleotide cell barcode and 12 nucleotide UMI, with the other paired read containing the 50, 100, or 150 nucleotide cDNA sequence. The third line starts with a "+" but can optionally also repeat the sequencing id contained in the first line. The fourth line contains ASCII-encoded Phred quality scores [80], with one score matching each nucleotide in the 2nd line in the same order. Note that the only "N" in the example is also the only nucleotide with a Phred score of "#", which represents a very low (36.9%) base call accuracy.

Because scRNA-seq droplet-based technologies generate paired-end fastq files with two reads per sequence, the read pairs typically follow each other in the fastq file, with one read containing the cell barcode and UMI and the other read containing the cDNA nucleotide sequence. These reads represent all of the mRNA the scRNA-seq experiment was able to detect, but many of the reads correspond to the same original mRNA transcript in a particular cell. These reads are "collapsed" to a single value using the UMIs they were tagged with during library preparation as part of the alignment process.

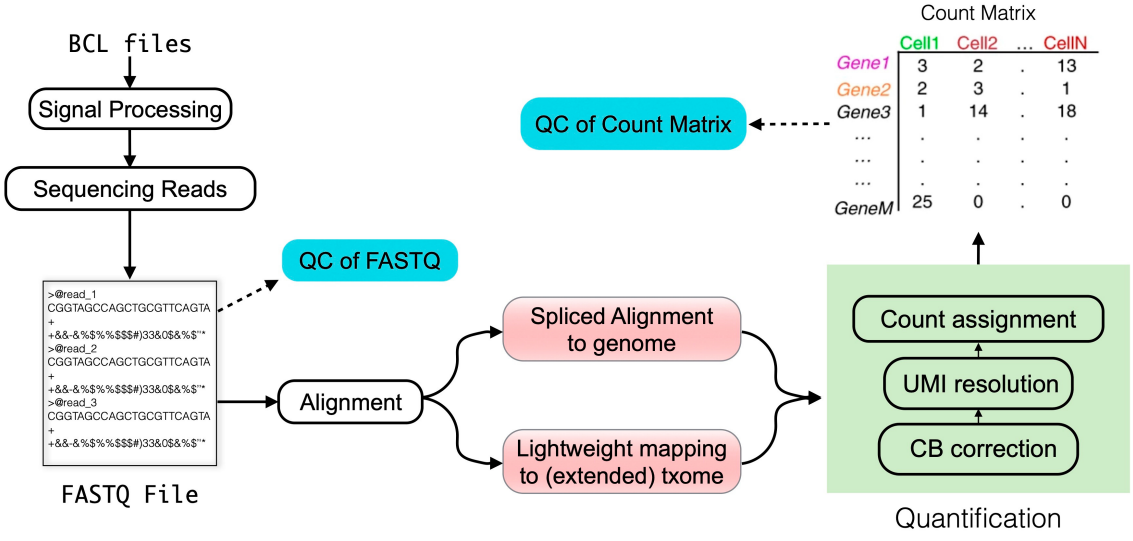



**Figure 1.2:** Outline of common scRNA-seq workflows. Here, "txome" stands for transcriptome. Adapted from [180].

### Read Alignment

Reads are either aligned to a reference genome or reference transcriptome and then counted using read cell and transcript metadata to quantify which genes each cell was expressing. Alignment is performed by an aligner or pseudo-aligner, software tools that map the short fragment of cDNA in the read to the best possible match in the genome or transcriptome, indicating which gene that read corresponds to. Most organisms have neither a reference genome nor a reference transcriptome, requiring mapping of transcripts through de novo transcriptome assembly, in which overlapping reads are stitched together to recreate the original transcriptome with tools such as Trinity or rnaSPAdes [94, 32].

Most scRNA-seq experiments either study model organisms with well-annotated reference genomes (e.g., *Mus musculus* or *Homo sapiens*) or organisms with well-defined transcriptomes but patchy or non-existent reference genomes (e.g., *Pinus elliottii*). Splice-aware aligners such as STAR or HISAT2 [66, 153] are often used to align sequencing data to reference genomes because they allow for the detection of novel splice junctions and multi-exon spanning transcripts, which is particularly important for the high proportion of intronic reads generated from the nuclear transcripts in snRNA-seq experiments. Pseudo-aligners such as Kallisto or Alevin [28, 249] are often used to align sequencing data to reference transcriptomes if no well-defined reference genome exists, if their higher specificity of mapping to known transcripts is important, or if their increased speed and computational efficiency is needed.

The scRNA-seq alignment process also involves many other steps, including soft-clipping reads, barcode and UMI error correction, filtering of low-quality reads, resolution of multi-mapped reads, and read quantification to generate gene counts. Some aligners can perform all of these additional tasks, while others are used in conjunction with additional software tools such as bustools [195]. Most modern aligners (e.g., STAR, HISAT2) and even pseudo-aligners (e.g., Kallisto, Alevin) detect

and soft-clip any adapter or other non-target sequences that may remain after read trimming as part of the alignment process. The cell barcode nucleotide sequences are often compared to a whitelist of possible cell barcodes, and both barcodes and UMIs are allowed some error when being matched, most often allowing a Hamming distance of 1 [108]. Some aligners discard multi-mapped reads (e.g., STAR) while others (e.g., Kallisto) always resolve reads based on the highest compatibility with a transcript. Certain mRNA transcripts are over-amplified during PCR amplification prior to sequencing (see previous section), due primarily to the uneven distribution of gene expression across transcripts and the stochastic nature of PCR amplification in scRNA-seq, biases which are compounded from multiple rounds of amplification [145]. However, PCR amplification bias is largely mitigated by collapsing reads to their transcript of origin through UMI deduplication [87] either as part of the alignment process (e.g., STAR) or after read mapping with specialized software (e.g., bustools).

The final product of the alignment process is a matrix of the number of unique reads of each gene in each cell, known as a count matrix (Fig. 1.2). ScRNA-seq count matrices are very sparse, with droplet-based matrices containing only 5 to 10% nonzero values because of biological and technical dropout events. Many scRNA-seq workflows employ various imputation, normalization, and dimensionality reduction methods in part to address this sparsity.

### Count Matrix Pre-processing

Count matrices are filtered, normalized and sometimes batch corrected, imputed, or both prior to any downstream analyses to account for a variety of biases not accounted for during the alignment process. Filtering is primarily done with four different filters: cell filters based on total UMI counts, cell filters based on number of unique genes, cell filters based on mitochondrial content, and gene filters based on the number of cells expressing the gene. Cells with very low UMI counts or a low number of unique genes expressed are typically omitted because they may either correspond to a droplet that only contained ambient RNA or a cell with a partially-captured transcriptome

with too little data to be biologically meaningful. These cells may either be omitted through simple lower UMI and minimum number of unique genes cutoffs (e.g., <500 UMIs or <200 unique genes) or through the use of algorithms that distinguish empty droplets containing ambient RNA from droplets containing cells such as EmptyDrops or SoupX [182, 291]. Cells with very high UMI counts or a high number of unique genes are often omitted because they may represent droplets that contained two or more cells, known as multiplets. Multiplets can also be addressed with a simple upper UMI and maximum number of unique genes cutoffs (e.g., >25,000 UMIs or >5,000 unique genes ), but these cutoffs cannot address variables such as differences in UMI counts across cell types and depend on sequencing depth. Alternatively, there exist dozens of available multiplet-detection algorithms, such as Scrublet or DoubletFinder [286, 193]. Cells are also typically filtered based on mitochondrial content as cells with an abnormally high proportion of mitochondrial gene expression can indicate a stressed, dying, or dead cell [119]. The mitochondrial filter is tissue- and species-dependent and less relevant to single-nuclei experiments because most mitochondrial transcripts originate from the mitochondria, which is located outside of the nucleus, but is often in the 5 to 15% range [213].

Certain genes are much more abundantly expressed than others in general and across cell types and states, with differences of up to 3 orders of magnitude in UMI counts. Count matrices are normalized to account for both this biological variation and various sources of technical variation. The most widely-used method divides the UMI counts in each cell by that cell's total UMI counts, scales the values in the cell by a factor of 10,000, adds a pseudo-count of 1, and log-transforms all the values. While the division addresses differences in sequencing depth across cells and the log transformation addresses differences in abundance levels across genes, highly-abundant genes can still skew some downstream analyses, particularly differential expression and some identification of fine-grained cell subtypes and states when using this method [106]. Various, more involved methods of normalization have been developed for scRNA-seq, including three of the most widely-used: sctransform, a

generalized linear model that uses total UMIs per cell as covariates [106]; SCNorm, a quantile regression that estimates the dependence of gene expression on total UMI counts [11]; and scran, a linear system for determining cell scaling factors based on pooled cell comparisons to a reference pseudo-cell [162].

Filtering and normalization are standard for all scRNA-seq analyses, but some analyses can also benefit from integration or imputation prior to downstream analyses. While sample multiplexing helps prevent batch-specific biases, integrating count matrices across experimental batches, organisms of origin, or conditions can lead to more accurate comparative analyses and boosts statistical power, and is thus often an important part of scRNA-seq workflows. More than 50 methods have been developed for scRNA-seq integration [181], but some of the most widely-used include deep-learning based methods such as scVI [174], trVAE [176], and scGen [175]; graph-based methods such as BBKNN [224] and Scanorama [117]; and various other methods such as Seurat v3's transcriptomic anchor-based integration method [255] or Harmony's linear, PCA-based approach [160]. These methods all tend to trade off biological fidelity for batch correction, but are often necessary to perform any kind of meaningful analysis when there are very strong batch effects, such as when comparing data from entirely separate experiments [181].

An analysis targeting lowly-expressed genes or using a dataset with few transcripts may benefit from imputation. There are more than 30 scRNA-seq imputation methods that can be categorized into three main approaches [121]. One set of methods uses a lower-dimensional latent space of the count matrix to rebuild an imputed version such as the low-rank matrix approximation used in ALRA [168] or the variational autoencoder used in DCA [76]. Another set of approaches use probabilistic models that typically focus on adjusting technical zeros such as the LASSO regression used by SAVER [125] or the use of an expectation-maximization algorithm by scImpute [167]. The last set of approaches are graph-based methods that adjust all values based on similar gene expression profiles such as DrImpute [92], which averages values across multiple cell clusterings, or MAGIC [65], which uses

data-diffusion across kNN-based graphs of transcriptomically similar cells to create an imputed count matrix.

While the approaches to scRNA-seq data imputation are diverse, no one method improves performance for all downstream tasks (e.g., clustering or differential expression), success rates of particular methods are highly dataset-dependent, and furthermore methods that boost particular downstream analyses may simultaneously hamper those analyses in other ways [121, 44]. For example, imputation methods that strengthen downstream differential expression analyses tend to do so by boosting the signal of highly differentially expressed genes while diluting the signal between slightly differentially expressed genes, as well as increasing the number of false positives. However, most imputation methods serve fairly well as point estimates of individual gene expression. There are hundreds of existing tools and methods for the pre-processing of scRNA-seq count matrices, and the use of a particular set of tools and techniques should be carefully considered and selected for each separate scRNA-seq analysis in order to facilitate extraction of the most accurate biological signals in all downstream analyses.

### Clustering and Cell Annotation

The transcriptomic profiles of individual cells vary considerably across a cell sample by number of unique genes expressed, the UMI counts within those genes, and the total UMI count within each cell. Perhaps the most standard part of any scRNA-seq workflow is to group cells with similar transcriptomic profiles together into clusters and label those clusters, a process known as cell annotation or cell typing. The number of clusters that cells are separated into is highly depends on the choice of clustering algorithm and hyperparameters. Furthermore, despite the development of hundreds of automated cell annotation tools, the gold standard for cell typing is still to manually annotate clusters by their top differentially expressed genes (DEGs), those most significantly upregulated or downregulated relative to other clusters [141].

Cell clusters are typically visualized by projecting genes down to two dimensions with UMAP and plotting cells in two dimensions (Fig. 1.3) [194].

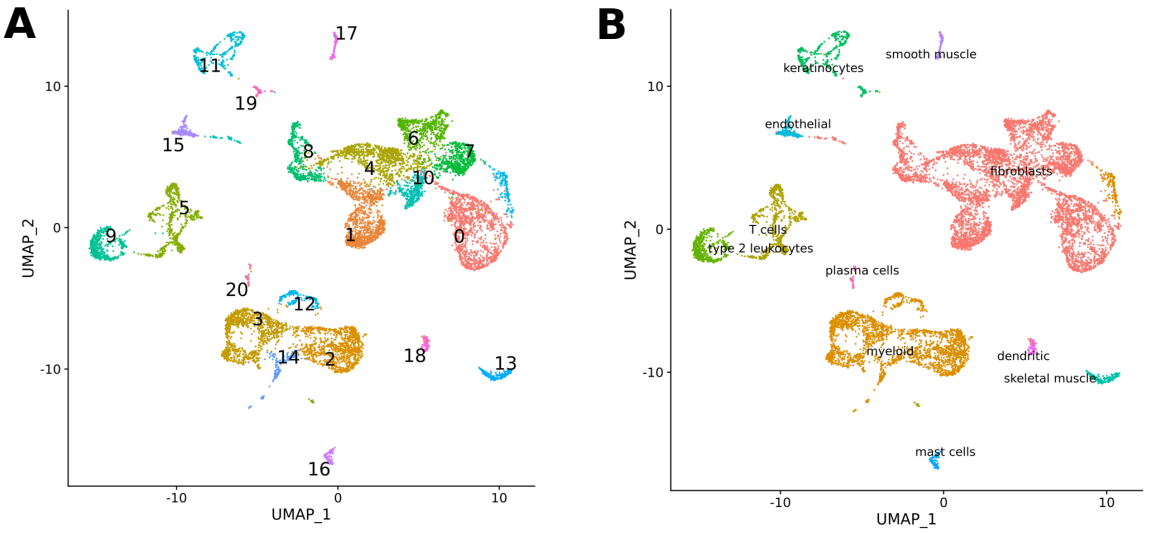



**Figure 1.3:** UMAP visualization of scRNA-seq clustering and annotation. (A) UMAP colored by unsupervised cluster assignment, with cluster numbers labeled. (B) UMAP colored by annotated cell type identities following manual cluster annotation and consolidation of transcriptionally similar clusters into shared cell type labels.

ScRNA-seq clustering algorithms can be categorized into three categories: $k$-means algorithms, hierarchical algorithms, and graph based algorithms. $K$-means clustering iteratively assigns cells to $k$ averaged cluster centers and can be applied directly through use of a $k$-means algorithm such as Lloyd's algorithm [171] or with some additional steps such as SC3's consensus of repeated $k$-means applications [155] and RaceID's additional use of outlier detection [99]. Hierarchical clustering algorithms either sequentially agglomerate cells into larger clusters or recursively divide them into smaller clusters, creating a hierarchy of progressively-sized clusters. These algorithms can be applied straightforwardly such as SINCERA [100], which uses centered Pearson's correlation and average linkage distance, or they can involve other methods such as BackSPIN and pcaReduce [295, 305], which incorporate intermediate dimensionality reduction steps between hierarchical steps. Graph-based clustering algorithms first create a graph embedding with cells as nodes and cell-cell relationships

as edges, most commonly using $k$-nearest-neighbors. Then, various graph-based clustering approaches are applied, such as the modularity function-optimizing Louvain and Leiden [23, 266] algorithms, known as community detection algorithms.

Each kind of algorithms comes with advantages and limitations. Both $k$-means and community detection methods scale linearly with the number of cells, while hierarchical clustering scales quadratically, a significant hindrance for large datasets. However, $k$-means requires the number of clusters to be chosen before clustering, assumes similar cluster sizes, and is a greedy algorithm that may fail to find the global optimum, while the Louvain and Leiden graph-based algorithms often under-perform on small datasets [85]. And hierarchical methods avoid the need to subcluster or adjust hyperparameters to obtain a desired number of clusters. The two most widely-used coding platforms, scanpy and Seurat [284, 109], both employ Leiden-based clustering, but the correct clustering strategy is an informed one that acknowledges the strategy's limitations are accounted for in any downstream analyses and visualizations.

Marker genes are genes that are consistently expressed within a particular cell type or state but not in most other cell types or states. More broadly clustered cells and cells from well-studied organisms and tissues are typically easier to identify because they are more likely to have well-known marker genes. More narrowly clustered cells and cells from rarely studied organisms and tissues may have few or no known marker genes, and in these cases clusters are commonly identified by the most enriched terms obtained from a gene ontology analysis of their marker genes [8]. Different clustering granularities correspond to differences in broader biological classifications (e.g., immune cells) or narrower (e.g., T helper cells, proliferating macrophages) classifications of cell types, subtypes, and states. At any granularity, marker genes identifying a cell population in one experiment may differ from those identifying a similar cell population in another experiment, even of the same tissue and species, both because of inconsistencies in marker gene identification methods and because of the inherent stochasticity of gene expression [225]. The ideal number of clusters to

separate cells into depends on what population of cells are important for downstream analyses. A common strategy is to broadly cluster and annotate cells and then subcluster a particular cluster of interest to further annotate it into cell subtypes or states.

Automated cell annotation methods work on a per-cell basis, while manual cell annotation is done per-cluster. Algorithmic approaches either compare cells to a reference database, such as Azimuth and Symphony [110, 144], or use a pre-trained classifier, such as CellNet and SingleCellNet [33, 260]. However, the success of both reference- and classifier-based algorithms is highly dependent on the degree to which the marker genes of the cells being annotated match those the algorithms are expecting, and there is little concordance on marker gene sets or cell type nomenclature between the most widely-used databases and classifiers [116, 2]. Manual cell annotation of a cluster by first identifying the distinguishing DEGs in a cluster and then using extant literature and domain expertise to determine an appropriate label for that cluster. This method of annotation is typically the most time-intensive and yet still the most widely-used due in part to the limitations of other methods, but is itself subject to inconsistencies in published marker genes and across DEG generation methods [141, 225].

### Differential Expression and Pseudoreplication

Differential gene expression (DGE) analyses determine which differences in gene expression between two groups of cells are statistically significant. Any two groups of cells may be compared, but the most common comparisons are between a cell cluster and all other cells to annotate that cluster or between experimental conditions within the same cell type to determine cell-type-specific responses to those experimental conditions. Cells within a biological sample are related because they have common genetic and environmental backgrounds. Therefore, the cells are pseudoreplicates with correlated transcriptomic profiles, but many DGE methods do not account for pseudoreplication at all, leading to a significant increase in false discoveries in

those methods, particularly in highly-expressed genes [248]. However, determining an appropriate DGE method also depends on various qualities of the dataset to be analyzed and on the purpose of the DGE analysis.

Approaches to differential expression analyses can be grouped into four categories: basic test statistics, pseudobulking followed by bulk RNA-seq approaches, and mixed-effects models. The Wilcoxon rank-sum test [187] and student's t-test [256] are simple, efficient, and implemented and recommended in the widely-used Seurat [109] and scanpy [284] packages. While these test statistics do not address pseudoreplication bias, they outperform most other methods when used to identify marker genes [225]. Pseudobulking refers to aggregating cell UMI counts by cell type within each biological sample. Any bulk RNA-seq approach, such as DEseq2 or edgeR [178, 234], can be used on pseudobulked cells and accounts for pseudoreplication bias, but these DGE methods are underpowered compared to methods that do not pseudobulk [302]. Mixed-effects models such as those implemented in the MAST [83] and muscat [55] frameworks can account for pseudoreplication by modeling it as a random effect and do not lose any statistical power from pseudobulking, but these approaches can be computationally intensive, especially on larger datasets. If a dataset only contains a single biological sample for each experimental condition, pseudoreplication is not an issue, but any DGE results may be indeterminably due to within-sample biological or technical variation rather than a true biological signal. DGE analyses typically not only determine which genes are differentially expressed but also calculate to what degree they are differentially expressed (e.g., $\log_2$ fold change), the statistical significance of the difference (e.g., p-value), and which proportion of each cell population expresses the gene. Several studies have noted that different DGE methods often generate results with little to no agreement, even when using different implementations of the same tools, and that these methods also tend to trade-off false positives for sensitivity and true positives for precision [248, 277, 88, 58]. These differences are due to the numerous different factors, including differences in statistical assumptions about underlying cell distributions and pseudoreplication,

in data processing, and in ranking and thresholding of results, among others. Ultimately, the best choice of DGE method depends on the purpose of an analysis and overall experimental setup, and it is common to use different approaches for cell typing and comparing experimental conditions. The differentially expressed genes generated by cell-type-specific DGE comparisons between experimental conditions are typically used to characterize cell-type-specific biological differences between those experimental conditions, often by using those gene lists as inputs into enrichment analyses.

### Genes and Enrichment Analyses

Genes quantified in scRNA-seq experiments come with a wide range of associated information, ranging from almost none (e.g., putative genes inferred from de novo transcriptome assembly) to genes with thousands of publications (e.g., human gene TP53) [67]. Annotating cell clusters and characterizing cell-type-specific biological differences between experimental conditions both often rely on determining which genes are differentially expressed and subsequently identifying biological functions or pathways that are overrepresented among those genes, known as enrichment analyses. Cell annotation and enrichment analyses are more challenging when genes are poorly defined, but they can still be performed in organisms without reference genomes by mapping unknown genes to well-studied ones using homology-based methods, such as the gene ortholog database eggNOG [115] or the k-mer–based transcriptome comparison tool Kmermaid [27].

Enrichment analyses of gene sets (e.g., differentially expressed genes) can be performed using gene ontology terms, biological pathways, transcription factor targets, disease associations, or any other category with curated gene associations. These analyses can be performed either manually by submitting gene lists to browser-based tools such as ToppGene and GOrilla [39, 70], or programmatically as part of a script using tools like iDEA [183], which integrates differential gene expression and gene set enrichment, and scGSEA [86], which combines non-negative

matrix factorization with pathway analysis. The gene sets used for enrichment analyses can be composed of up-regulated genes, down-regulated genes, or both, and any resulting terms that mention directionality (e.g, "negative regulation of cell proliferation") should be considered with this in mind. Enrichment analyses can help annotate a cluster with poorly-known or unknown marker genes by describing the biological functions that cluster is enriched for compared to the rest of the cells. Similarly, enrichment of genes differentially expressed between cells under two different experimental conditions within the same cell type cluster can reveal condition-specific biological processes, pathway-level regulation, and coordinated gene activity that may be too subtle or complex to interpret from gene-level changes alone [141].

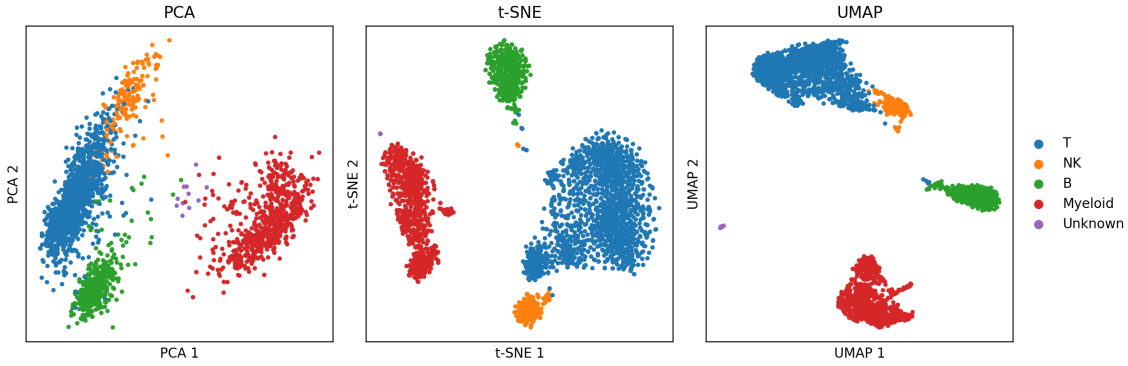



**Figure 1.4:** Comparison of cell type clusters with PCA, t-SNE, and UMAP projections of the cells into two dimensions. t-SNE and UMAP projections were after reducing to the first 50 principal components. Adapted from [19].

### Visualization

ScRNA-seq workflows involve the generation of various kinds of plots throughout an analysis to aid in visualizing cells, genes, or relationships between cells and genes. During pre-processing, histogram, density, and violin plots of per-cell quality control metrics such as total UMI count and percent mitochondrial content can be generated to visualize aberrant overall or sample-specific gene expression profiles. To attempt to

visualize both local and global relationships between cells, genes are projected down to two dimensions and cells are visualized on a scatter plot with the two dimensions as axes. This dimensionality reduction is performed in two steps, first by applying PCA [81] to obtain the first 50 to 100 principal components and then by applying methods such as t-SNE, UMAP, or PHATE [270, 194, 200] to attempt to best display the spatial relationship of the cells in two dimensions (Fig. 1.4). These two-dimensional projections can be colored by experimental batch to highlight batch effects, by cluster to display cell annotation, or by heatmap of gene expression to show gene expression patterns across all the cells. Additionally, dot plots of the top few marker genes for each cluster are commonly used as a visual aid of how cell typing was performed, and volcano plots are often used to visualize overall DGE patterns of up- and down-regulation for a particular DGE comparison. These visualizations are just a subset of some of the most widely-used in scRNA-seq analyses, but ultimately a good visual aid is one that successfully summarizes key findings or communicates biological insights.

### 1.3.3 Challenges and Future Directions

ScRNA-seq provides a snapshot of the transcriptomic landscape at single-cell resolution, enabling cell-type-specific detection of biological signals in response to experimental perturbations. However, despite numerous advances and developments in single-cell protocols, tools, and methodology, the technology still suffers from several major limitations [164, 154]. Low mRNA capture efficiency leads to sparse gene expression data, reducing the power to detect lowly expressed genes due to technical dropout events. Mass, pooled sequencing enables high-throughput of cell numbers but also incurs a trade-off in cost between the number of cells sequenced and the depth at which cells are sequenced. Droplet-based techniques require cell suspensions that cannot preserve the spatial context of the cell sample from which they originate and in which large, rare, or fragile cells may be underrepresented or lost during cell capture and isolation. Cell annotation is nontrivial, relying

on expert annotation and reference databases often containing inconsistent marker genes or lacking them altogether for rare cell populations. And the sparse, high-dimensional, highly heterogeneous, and increasingly large datasets produced by scRNA-seq experiments require complex computational methods to accurately extract meaningful biological signals. These challenges have motivated the development of multi-omic single-cell approaches that augment transcriptomic data with additional layers of molecular information.

As the number and kind of biological omics datasets has increased (see section 1.2.1), so too have the efforts to integrate scRNA-seq datasets with other datasets, in particular with other datasets also at single-cell resolution, both with other scRNA-seq datasets and with other modalities of single-cell data, such as single-cell proteomics, single-cell chromatin accessibility, and single-cell spatial transcriptomics [30, 31, 132]. Integration of multiple single-cell modalities improves downstream analysis such as cell clustering at the cost of increased expense or time [251]. Single-cell data can also be integrated with other, non single-cell datasets, with similar improvements yet increased complexity due to the increased difference in data modalities [9].

Since the first experiment that sequenced a single cell's transcriptome in 2009 [261], the field of scRNA-seq has exploded with technological developments and diverse scientific applications. As the scale of single-cell experiments has increased from one cell to millions of cells, thousands of tools have been developed to efficiently process and extract meaningful biological insights from these increasingly complex datasets [293, 294]. Early studies focused on protocol development (e.g., DropSeq) and the classification and discovery of cell types in mouse and human tissues, and these applications have since been duplicated in many other diverse model and non-model organisms such as lizards [264], fish [243], and plants [128]. However, scRNA-seq research has also expanded into numerous other biological and medical domains, including drug discovery, cancer research, neurobiology, developmental biology, and

toxicology, and the creation of hundred-million cell cell atlases [269, 123, 222, 124, 40, 235], and more than 100,000 studies have been published to date [22].

## 1.4 Spatial Transcriptomics

### 1.4.1 Spatial Context in Transcriptomics

Spatial transcriptomics is a recent extension of transcriptomics technologies that extends the field by preserving information about the physical location of RNA molecules within intact tissues. The first commercially available method, Visium, was released in 2019 by 10x Genomics [250], and several different single-cell resolution spatial transcriptomics technologies were first made commercially available in 2021 [41, 132, 114]. Whereas single-cell RNA sequencing characterizes gene expression at cellular resolution after tissue dissociation, spatial transcriptomic approaches measure gene expression in situ. This preserves the spatial context of the cells in an experiment, making it possible to determine how gene expression patterns relate to tissue architecture, cell–cell interactions, and localized spatial neighborhoods, or niches, that are lost during dissociation-based workflows.

Spatial transcriptomics emerged from a long-standing need to understand gene expression within its native tissue context. Early transcriptomic technologies such as bulk RNA-seq and, later, scRNA-seq have enabled genome-wide measurement of gene expression at up-to individual cell level but require tissue dissociation, erasing any spatial context. In response, spatially resolved transcriptomic technologies have been developed to preserve the architectural context of cellular gene expression. Early approaches such as in situ hybridization and immunohistochemistry provided targeted, low-plex spatial measurements [166, 229]. Subsequent innovations including array-based barcoding methods and highly multiplexed in situ chemistries have enabled transcriptome-scale measurements directly in tissue sections. Platforms such as Visium and high-resolution in situ technologies such as MERFISH, seqFISH, and

Xenium have collectively expanded transcriptomics from a purely dissociative to a spatially aware discipline [250, 41, 244, 114]. These advances have expanded the range of possible biological analyses from just the exploration of the transcriptomic profiles of cell populations to relationships of individual cells based on both transcriptomic identity and physical distance to one another.

Within a systems biology framework, spatial transcriptomics enables the analysis of tissues as structured, interacting cellular systems rather than as collections of isolated cell types. Cells operate within defined spatial neighborhoods, or anatomical niches, where gradients of signaling molecules, extracellular matrix composition, vascular proximity, and immune infiltration shape transcriptional states. By preserving spatial coordinates, spatial transcriptomics provides access to the data-driven relationship between molecular measurements and histological organization, allowing gene expression programs to be interpreted in the context of tissue architecture. This systems-level perspective is particularly important in complex, heterogeneous tissues, such as inflamed skin, tumors, or regenerating wounds, where emergent behavior has been shown to arise from coordinated interactions among epithelial, stromal, vascular, and immune compartments [146, 130, 198].

### 1.4.2 Spatial Transcriptomics Workflows

Workflows for modern single-cell resolution spatial transcriptomic experiments (e.g., Xenium) closely mirror workflows for modern high-throughput scRNA-seq experiments (e.g., droplet-based scRNA-seq, see 1.3.2), but also expand upon them by incorporating image-derived spatial information and additional preprocessing steps specific to in situ technologies. Because spatial transcriptomic data are generated through imaging of intact tissue sections rather than sequencing of dissociated cells, computational pipelines must integrate both transcriptomic measurements and microscopy-derived features. As a result, spatial workflows typically include image processing, cell segmentation, transcript assignment, cell annotation (clustering and

cluster identification), and spatial analysis of spatially-defined cell neighborhoods and spatially identified cell–cell relationships (Fig. 1.5).

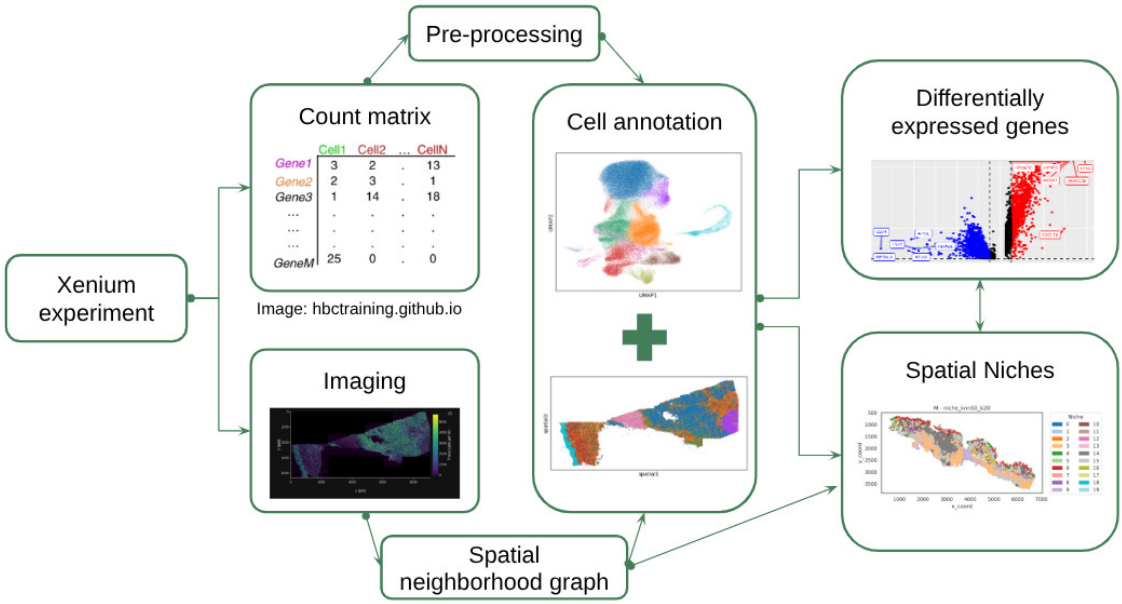



**Figure 1.5:** Overview of a typical spatial transcriptomic analysis workflow using Xenium data. Xenium experiments generate both transcriptomic count matrices and high-resolution imaging data from intact tissue sections. Following preprocessing, transcriptomic clustering and spatially resolved cell annotation are integrated to identify cell populations within their native tissue context. Spatial neighborhood graphs can then be constructed from cell coordinates to define local cellular niches and characterize tissue microenvironments. Downstream analyses include differential gene expression, spatial niche analysis, and investigation of multicellular organization and cell–cell interactions within intact tissues.

The first step in spatial transcriptomic analysis is cell segmentation, the process of identifying cell boundaries within the tissue image and assigning detected transcripts to individual cells. Segmentation typically relies on nuclear stains, membrane markers, morphological features derived from imaging data, or a combination of these approaches and, for Xenium experiments, is applied and packaged as part of the raw data. Accurately defining cell boundaries remains challenging due to variability in cell size and morphology, overlapping or densely packed cells, incomplete membrane staining, and imaging artifacts. Errors in segmentation can propagate

through downstream analyses by incorrectly assigning transcripts to neighboring cells or generating artificial cells with mixed transcriptional signals [219, 289]. Consequently, segmentation algorithms represent an active area of methodological development, with novel, optional third-party approaches ranging from classical image processing techniques to machine-learning-based segmentation models designed to improve boundary detection in complex tissues [254, 95].

Following segmentation, cells are typically annotated according to their transcriptomic identity. This step parallels cell type annotation workflows used in scRNA-seq analysis, where clusters of cells are defined based on gene expression patterns and subsequently assigned biological labels using known marker genes or reference datasets. However, spatial transcriptomic datasets present additional challenges for transcriptomic annotation. Compared with scRNA-seq experiments, spatial platforms often detect far fewer transcripts per cell and target a predefined and much reduced panel of genes (currently either approximately 500 or 5,000 genes for Xenium gene panels) rather than the full transcriptome. Further, cell segmentation methods are imperfect, and transcripts are often attributed to incorrect cells or ambient contamination instead of real biological signal. These constraints can complicate the identification of different cell populations, either by mixing cell populations or failing to identify particular cell populations. To address this issue, spatial analyses both use the spatial data of transcriptomically identified cell populations to help in annotation and often leverage external scRNA-seq reference datasets by transferring cell-type labels or probabilistic assignments to spatial cells using computational integration methods [20, 173]. However, complicating the use of spatial context in transcriptomic-based cell assignment is the fact that cells that share similar transcriptional profiles may nonetheless occupy distinct anatomical locations within a tissue, reflecting either differences in cell state or the prevalence of certain cell populations across difference tissues (e.g., leukocytes). Spatial analysis therefore frequently includes the manual identification of tissue domains or zones in each tissue sample, corresponding to known

histological structures, such as epithelial layers, immune infiltration zones, or stromal compartments.

### Spatial Neighborhoods

Finally, spatial transcriptomics enables the systematic investigation of cell–cell interactions within intact tissues. Because the physical location of each cell is preserved, researchers can analyze how the abundance, transcriptional state, and spatial proximity of neighboring cell populations influence cellular behavior. Approaches for studying these interactions include ligand–receptor co-expression analysis, spatial neighborhood enrichment tests, and graph-based models that represent tissues as networks of interacting cells [71, 206, 262]. In particular, many spatial transcriptomic workflows construct local cellular neighborhoods, or niches, by identifying the nearest neighboring cells surrounding each individual cell using distance-based or k-nearest-neighbor approaches. These neighborhoods can then be summarized according to their local cellular composition, enabling identification of recurrent spatial microenvironments enriched for particular combinations of stromal, epithelial, vascular, or immune cell populations. Subsequent clustering or graph-based analyses of these spatial niches can reveal multicellular tissue structures that may not be apparent from transcriptomic clustering alone. Methods implemented in frameworks such as scimap additionally enable quantification of neighborhood enrichment, spatial co-occurrence patterns, and niche-specific gene expression programs across tissues [206]. These analyses provide insight into how signaling relationships and localized cellular organization shape tissue architecture, immune infiltration, and disease progression. By combining transcriptomic and spatial information, spatial transcriptomics thus allows researchers to move beyond cataloging cell types and states and toward understanding the emergent behavior of multicellular systems in situ [172].

### 1.4.3 Challenges and Future Directions

Scalability has been a defining feature of recent spatial technologies. Early methods were limited in gene throughput or spatial resolution, but contemporary platforms now support thousands of genes measured across tens of thousands to millions of cells within a single experiment. However, transcript detection remains very low, with average total transcript detection per cell ranging from 20 to 100 in current Xenium experiments, as opposed to thousands in modern high-throughput scRNA-seq experiments [132, 202]. Nevertheless, improvements in probe chemistry, imaging speed, barcoding strategies, and computational pipelines have enabled increasingly comprehensive spatial atlases across organs and disease states, and advances are constant [189]. Moreover, spatial transcriptomics is rapidly integrating into a broader multi-omic landscape. Emerging approaches combine spatial RNA profiling with higher-resolution scRNA-seq profiling, protein detection, chromatin accessibility, epigenetic marks, or metabolomic signatures, enabling layered molecular characterization within the same tissue section [53, 170, 37, 231]. These developments position spatial transcriptomics as a foundational modality for integrative, high-dimensional systems biology, particularly when coupled with large-scale computational analysis pipelines and high-performance computing resources.

Despite rapid technological progress, spatial transcriptomics presents substantial computational and methodological challenges that continue to shape the field. One major limitation is the scale of the resulting datasets. High-resolution spatial experiments can generate measurements for tens of thousands to millions of cells, accompanied by large multi-channel microscopy images and detailed transcript coordinate data, resulting in GBs to TBs of data [202]. Managing, storing, and analyzing these datasets therefore often requires large-scale storage infrastructure, substantial memory resources, and efficient computational pipelines, particularly when integrating spatial transcriptomic data with other large-scale omics datasets.

Image processing also remains a critical bottleneck in spatial transcriptomic workflows. Accurate tissue registration, transcript localization, and cell segmentation depend on complex image analysis pipelines that must operate reliably across heterogeneous tissue morphologies. Errors in these steps can significantly influence downstream analyses, as transcript assignments and cell boundaries form the basis for all subsequent biological interpretation. Moreover, none of the current image segmentation techniques used in Xenium experiments account for the three-dimensional nature of cell spatial coordinates, only using two dimensions instead [132]. Continued improvements in automated segmentation algorithms, machine learning–based image analysis, and quality control procedures will be essential for improving the robustness and reproducibility of spatial transcriptomic analyses, and many such methods have been and are being developed, often from non-commercial sources [202, 254, 95].

Cell annotation presents additional challenges due to the relatively sparse nature of spatial transcriptomic data. Many spatial technologies detect only a limited number of transcripts per cell, and targeted gene panels may omit important marker genes needed to distinguish certain cell populations [202, 231]. These limitations complicate traditional clustering-based annotation approaches and often require integration with external scRNA-seq datasets or reference atlases. Consequently, computational methods for cross-modality integration have become an important component of spatial transcriptomic analysis pipelines. At the same time, the spatial structure of the data introduces new opportunities for statistical modeling. Methods that explicitly account for spatial autocorrelation, spatially variable gene expression, and domain-level organization are increasingly being developed to identify patterns that cannot be detected in dissociated transcriptomic data. Techniques for spatial microenvironment detection, neighborhood analysis, and spatially variable gene identification provide new ways to characterize tissue organization and emergent biological processes.

Looking forward, integrative computational frameworks will likely play a central role in the continued evolution and use of spatial transcriptomics. Methods such as Tangram, Seurat v5 integration workflows, SpaGCN, and probabilistic models such as gimVI enable the alignment of spatial transcriptomic data with scRNA-seq reference datasets and other omics modalities [20, 109, 122, 174]. As these approaches mature, they will facilitate the construction of increasingly comprehensive spatial atlases that integrate multi-omic (e.g., transcriptomic, proteomic, or epigenetic) datasets. Together with ongoing advances in imaging chemistry and computational analysis, these developments position spatial transcriptomics as a central technology for systems-level investigation of tissue biology and disease.

## 1.5 Scalable Omics Computing

The rapid growth of modern omics technologies has transformed biology into an increasingly data-intensive discipline. Advances in next-generation sequencing, single-cell transcriptomics, spatial transcriptomics, proteomics, most other -omics datasets, and other advances such as the recent advent of large-scale genomic surveillance have enabled experiments that routinely generate datasets containing millions of cells, billions of sequencing reads, or terabytes of image and molecular data [190, 253, 164, 202]. While these technologies have substantially expanded the scope of biological questions that can be addressed computationally, they have also introduced major challenges in data storage, memory usage, computational scalability, and algorithmic efficiency. Consequently, scalable computational methods and high-performance computing (HPC) infrastructure have become increasingly central components of modern systems biology workflows.

### 1.5.1 Sparse and High-Dimensional Biological Data

Many omics datasets are naturally represented as high-dimensional matrices in which biological entities such as genes, proteins, genomic loci, or cells correspond to rows or columns (Fig. 1.6). Single-cell RNA sequencing count matrices, for example, often contain tens of thousands of genes measured across tens of thousands to millions of cells, while spatial transcriptomic datasets additionally incorporate high-resolution imaging information and spatial coordinate data. Similarly, biological networks and multiplex systems can be represented through large adjacency matrices describing relationships between molecular or cellular entities (e.g., gene co-occurrence networks). These datasets are typically both high-dimensional and highly sparse, with the majority of matrix entries equal to zero [15, 304].

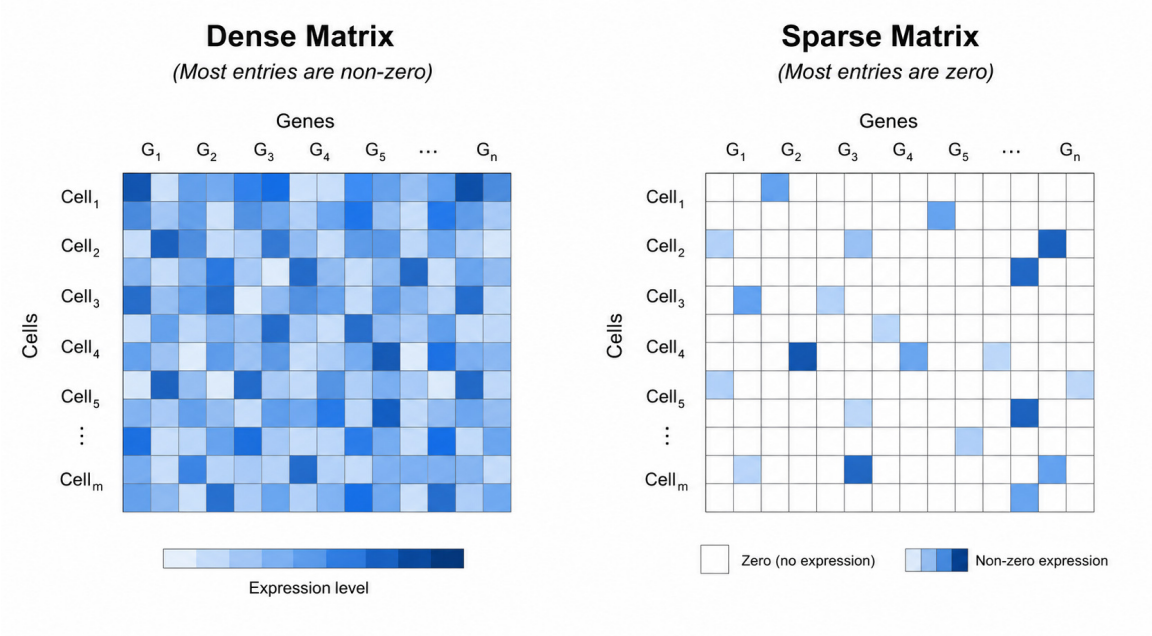



**Figure 1.6:** Comparison of dense and sparse matrix representations commonly encountered in omics data analysis. Dense matrices (left) contain mostly nonzero values and explicitly store all matrix entries, resulting in high memory usage for large datasets. In contrast, sparse matrices (right) contain predominantly zero-valued entries, as commonly observed in single-cell transcriptomic, spatial transcriptomic, and biological network data, and can therefore be stored more efficiently using sparse matrix representations that preserve only nonzero values and their indices.

Sparsity in biological datasets arises from both biological and technical sources. In scRNA-seq experiments, biological sparsity reflects the fact that many genes are not actively expressed in particular cell types or states, while technical sparsity results from low transcript capture efficiency and dropout events [47]. Spatial transcriptomic technologies often exhibit even greater sparsity due to reduced transcript detection and targeted gene panels. Similarly, biological networks are commonly sparse because most molecular entities interact with only a small subset of all possible neighbors. Sparse matrices can therefore frequently be stored more efficiently using compressed sparse row (CSR) or compressed sparse column (CSC) formats, which preserve only nonzero values and their indices rather than explicitly storing every matrix element [59]. These representations can substantially reduce memory consumption and improve computational efficiency for many operations.

However, sparse and high-dimensional biological data also introduce significant analytical challenges. Many classical statistical and machine learning methods scale poorly with increasing dimensionality, a phenomenon commonly referred to as the curse of dimensionality [4]. As dataset sizes increase, memory usage and computational complexity can rapidly become prohibitive, particularly for methods involving matrix decomposition, nearest-neighbor graph construction, or all-by-all comparisons. Furthermore, many computational operations performed on sparse matrices generate dense intermediate representations that substantially increase memory requirements. Consequently, the development of sparse-preserving computational methods has become an important area of research in systems biology and large-scale omics analysis. These approaches aim to minimize unnecessary dense matrix reconstruction while maintaining biological fidelity and computational tractability across increasingly large datasets.

### 1.5.2 Distributed Computing in Computational Biology

As biological datasets have grown beyond the memory and computational limits of individual workstations, distributed computing and HPC resources have become increasingly important for large-scale omics analyses. Traditional shared-memory systems operate within a single compute node in which multiple processor cores access the same pool of memory. While shared-memory parallelism can accelerate many computational tasks through multithreading, these systems remain fundamentally constrained by the memory capacity and processor count of a single machine. In contrast, distributed-memory systems divide computational tasks across multiple independent compute nodes connected through high-speed interconnects, enabling analyses to scale beyond the limitations of any individual node [62, 239].

#### MPI

Modern HPC systems typically combine thousands of compute nodes, each with multiple central processing unit (CPU) and graphical processing unit (GPU) cores, parallel storage systems, and high-speed communication systems to support large-scale scientific computing. Distributed-memory parallelism is commonly implemented using the Message Passing Interface (MPI), a standardized communication framework that enables processes running across multiple nodes to exchange data and coordinate computations [197]. High-level interfaces such as mpi4py extend MPI functionality to Python-based scientific workflows and have become increasingly common in computational biology applications [57]. Parallel computing strategies are also frequently combined with multithreading, GPU acceleration, and distributed linear algebra libraries to further improve scalability for computationally intensive tasks such as matrix factorization, graph analysis, and large-scale machine learning.

### Parallel I/O

Large-scale omics workflows present several computational bottlenecks that motivate the use of distributed computing strategies. Construction of nearest-neighbor graphs, dimensionality reduction, all-by-all genomic comparisons, and network-based analyses often exhibit substantial memory requirements or unfavorable computational scaling as dataset size increases. Input/output (I/O) operations also become increasingly challenging at scale, particularly for large sequencing datasets and multimodal spatial transcriptomic experiments. Parallel file formats and distributed storage frameworks such as HDF5 have therefore become important components of scalable computational pipelines [98]. Similarly, distributed graph-processing frameworks and cloud-based genomics platforms such as Apache Spark-based workflows have been developed to address the computational demands of large-scale genomic analysis [292].

## 1.6 Large-Scale Viral Genomics

The increasing affordability and throughput of next-generation sequencing technologies has transformed the study of viral evolution and infectious disease epidemiology. Viral genomics enables direct measurement of genetic variation within and between viral populations, making it possible to track transmission dynamics, identify emerging variants, reconstruct evolutionary histories, and monitor genomic adaptation in near real time [96, 89, 105]. These capabilities have fundamentally reshaped infectious disease surveillance and outbreak response, particularly for rapidly evolving RNA viruses such as influenza virus, human immunodeficiency virus (HIV), Ebola virus, and SARS-CoV-2 [16, 69]. Together with advances in computational biology and global data-sharing initiatives, viral genomics has become a central component of modern genomic epidemiology.

### 1.6.1 Introduction to Viral Genomics

RNA viruses exhibit particularly high mutation rates due to error-prone replication mechanisms and rapid population turnover, enabling fast evolutionary adaptation over relatively short timescales [69]. As viral genomes accumulate mutations during transmission, genomic sequencing can be used to infer evolutionary relationships between viral isolates and identify genetically related transmission clusters (i.e., viral lineages). Early applications of viral genomics focused on phylogenetic reconstruction of pathogens such as HIV and influenza virus to study transmission pathways, geographic spread, and viral diversification [159, 282, 17]. During the 2014 Ebola outbreak, genomic sequencing was used extensively to track viral spread and characterize outbreak dynamics in near real time, demonstrating the growing utility of genomic epidemiology during large-scale public health emergencies [91].

Viral genomic surveillance integrates sequencing data with epidemiological, temporal, and geographic metadata to characterize pathogen emergence and spread across populations. Sequencing-based surveillance can identify newly emerging variants, estimate transmission dynamics, detect introductions into new geographic regions, and monitor the evolution of mutations associated with altered transmissibility, virulence, or immune escape [89]. These approaches have become increasingly important as sequencing technologies have improved in throughput and accessibility, enabling the generation of genomic datasets at unprecedented scale.

### 1.6.2 Next-Generation Sequencing and Genomic Surveillance

The widespread adoption of next-generation sequencing has dramatically increased the scale and speed of viral genomic surveillance. Modern sequencing platforms can generate thousands to millions of sequencing reads simultaneously, enabling rapid characterization of viral genomes directly from clinical samples [93]. These advances have facilitated large international genomic surveillance efforts and the development

of public repositories that aggregate viral genomic data from laboratories around the world. Databases such as GISAID and NCBI Virus provide access to millions of viral genome sequences along with associated metadata including sampling date, geographic origin, and lineage assignment [73].

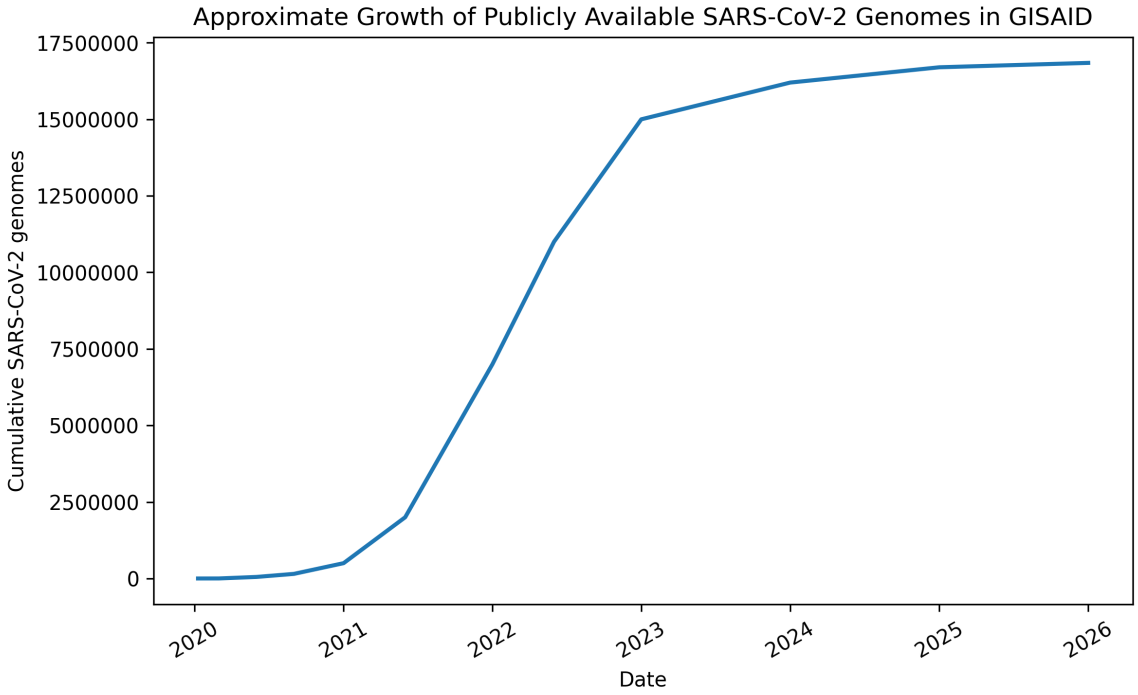



**Figure 1.7:** Approximate cumulative growth of publicly available SARS-CoV-2 genome sequences deposited in GISAID from January 2020 through May 2026. Rapid expansion of global sequencing efforts during the COVID-19 pandemic resulted in the generation of nearly 17 million publicly available viral genomes, representing the largest genomic surveillance effort in history. This unprecedented growth in viral sequencing data has enabled large-scale genomic epidemiology while simultaneously introducing substantial computational and analytical challenges for phylogenetic reconstruction, comparative genomics, and integrative population-scale analyses.

Computational frameworks for genomic epidemiology have evolved alongside these large-scale sequencing efforts. Phylogenetic reconstruction methods are commonly used to infer evolutionary relationships among viral genomes and estimate transmission dynamics across geographic regions and time periods [105]. Dynamic nomenclature systems such as the Pango lineage framework were developed during the SARS-CoV-2 pandemic to facilitate standardized classification and tracking

of emerging viral lineages [227]. These approaches enable continuous monitoring of variant emergence and dissemination while supporting integration of genomic data with epidemiological and public health analyses. More broadly, genomic surveillance has demonstrated the value of combining sequencing technologies, metadata integration, and computational modeling to characterize infectious disease dynamics at population scale.

### 1.6.3 Scaling Challenges in Viral Genomics

Although genomic surveillance has expanded rapidly, the increasing scale of viral sequencing datasets presents substantial computational and methodological challenges. Modern public repositories now contain millions of viral genome sequences (Fig. 1.7, with SARS-CoV-2 representing the largest genomic surveillance effort in history [42]. Analyses involving all-by-all genomic comparisons, pairwise distance calculations, or phylogenetic reconstruction become increasingly computationally expensive as dataset size grows, as pairwise comparisons among millions of genomes scale quadratically. Many commonly used phylogenetic methods scale poorly with increasing numbers of genomes, and exact reconstruction of large phylogenetic trees can become computationally prohibitive at pandemic-scale sequencing volumes [268].

In addition to computational complexity, large viral genomic datasets present challenges related to metadata heterogeneity, uneven geographic sampling, incomplete metadata, and varying sequencing quality across studies and repositories. Certain geographic regions and time periods are often heavily oversampled while others remain underrepresented, introducing biases into downstream epidemiological analyses [29]. Storage and input/output requirements also become increasingly important as datasets expand to include millions of genomes and associated metadata records. These challenges have motivated development of scalable approximate

methods, compressed genomic representations, distributed workflows, and graph-based approaches capable of operating efficiently on large-scale genomic datasets [105].

### 1.6.4 Underuse of Large Public Viral Genomic Datasets

Despite the unprecedented availability of viral genomic data, comparatively few studies leverage the full scale of these datasets in downstream analyses. Many genomic epidemiology workflows rely on heavy subsampling strategies to reduce computational burden, often analyzing only a small fraction of publicly available genomes [192, 268]. While subsampling is frequently necessary for tractable phylogenetic reconstruction and visualization, it can obscure fine-grained population structure, reduce detection of rare transmission patterns, and limit characterization of geographic or temporal variation across viral populations. This highlights the growing need for scalable computational frameworks capable of integrating large viral genomic datasets with complementary epidemiological, temporal, geographic, and sociodemographic information. As genomic surveillance efforts continue to expand, methods capable of operating efficiently at global sequencing scale will likely become increasingly important for infectious disease monitoring, outbreak response, and systems-level investigation of pathogen evolution.

## 1.7 Conclusion

These domains converge on a central theme: enabling robust biological discovery from massive, and often sparse, noisy, or sparse and noisy datasets through scalable, interpretable, and integrative computational methods. The emergence of modern omics technologies has transformed biological research into a fundamentally data-intensive discipline, generating datasets that capture molecular processes across multiple biological scales and modalities. Network-based systems biology provides

a conceptual framework for integrating these diverse data sources, allowing molecular entities and their interactions to be modeled as interconnected systems rather than isolated measurements. Within this broader context, single-cell RNA sequencing has enabled transcriptomic characterization at cellular resolution, revealing the heterogeneity of biological tissues and uncovering previously unknown cell types, states, and regulatory programs. At the same time, spatial transcriptomics has extended these insights by restoring the architectural context of gene expression, enabling the study of how cellular interactions and tissue organization shape biological function. These technologies illustrate the ongoing transition from reductionist molecular measurements toward integrative, systems-level views of complex biological processes.

However, the scale, sparsity, and complexity of modern omics datasets continue to pose substantial analytical challenges. High-dimensional transcriptomic matrices, large imaging datasets, and multi-modal integration tasks require computational methods that are not only statistically robust but also capable of operating efficiently on large-scale computing infrastructure [164]. Addressing these challenges requires the development of algorithms that balance scalability with interpretability while remaining grounded in biological insight. In this dissertation, these principles guide the development and application of computational approaches for analyzing large-scale biological data, including multiplex network analysis for multi-omic integration, scalable methods for single-cell and spatial transcriptomic data analysis, and computational frameworks capable of handling massive genomic datasets such as those generated by global viral sequencing efforts. By combining advances in high-throughput molecular technologies with scalable computational methods, it becomes possible to extract meaningful biological insight from increasingly complex datasets and to move toward a more comprehensive, systems-level understanding of biological organization and disease.

# Chapter 2

# Cross-Species, Integrated Multi-Omics Reveal Immunomodulatory Fibroblast Niches in Atopic Dermatitis

## 2.1 Introduction

Atopic dermatitis (AD) is the most common chronic inflammatory skin disorder worldwide, affecting up to 20% of children and 10% of adults, with rising prevalence in both industrialized and developing countries [281]. Clinically, AD is characterized by recurrent episodes of eczematous lesions, intense pruritus (i.e., itchy skin), and xerosis (i.e., dry skin), often beginning in early childhood and persisting or recurring throughout life [207]. Pathologically, it involves genetic predisposition, epidermal barrier dysfunction, immune dysregulation, and infiltration of inflammatory cells, particularly type 2 helper T (Th2) cells and eosinophils [142]. Moreover, the disease is driven by both innate and adaptive immune responses, with contributions from multiple cell types, including keratinocytes, fibroblasts, T cells, innate lymphoid cells (particularly ILC2s), and mononuclear phagocytes such as macrophages. The highly complex pathophysiology of AD and its relapsing–remitting nature highlight the need for high-resolution approaches to understand its underlying mechanisms.

### 2.1.1 Cellular and Spatial Complexity in AD

AD can be characterized as acute or chronic [142]. Acute AD is dominated by a type 2 immune response, characterized by elevated levels of cytokines such as IL-4, IL-5, and IL-13, which drive eosinophil recruitment, B cell class switching, and disruption of the epidermal barrier. Chronic AD also suffers from a type 2 immune response but also additionally shows increased expression of Th1 and Th0 cytokines such as INF-$\gamma$, IL-6, granulocyte-macrophage colony-stimulating factor (GM-CSF), and IL-12, leading to increased production of natural killer and T cells and in turn recurrent apoptosis of keratinocytes, driving strong, recurrent skin inflammations. In both cases, cytokines originate from both adaptive immune cells, including Th2 cells, and innate sources such as group 2 innate lymphoid cells (ILC2s), contributing to complex inflammation and tissue remodeling.

Crucially, epidermal–immune cell crosstalk plays a central role in amplifying and maintaining this inflammatory environment. Keratinocytes have been shown to respond to barrier disruption and cytokine signaling by releasing alarmins and chemokines that further recruit immune cells [126]. In parallel, recent work using both mouse models of AD and human AD has uncovered a clear difference between healthy fibroblasts and fibroblasts in inflamed skin as well as multiple immune roles for the inflammatory fibroblasts [90, 120, 36, 113, 233], suggesting that specific fibroblast subpopulations contribute to disease pathogenesis through specialized roles in antigen presentation, cytokine production, and extracellular matrix remodeling. However, bulk transcriptomic approaches obscure this complexity by averaging signals across mixed cell populations, highlighting the importance of single-cell methods to disentangle the diverse cellular and molecular interactions in AD.

Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool for uncovering cell type– and cell state–specific transcriptional patterns (see 1.3). By resolving the transcriptomes of individual cells, scRNA-seq enables the identification of cell populations and subpopulations, and consequently cell-population-specific disease activation states and signaling networks that are often masked in bulk profiling. However, the dissociation of tissue required for scRNA-seq eliminates any spatial information, making it difficult to reconstruct the physical context in which cellular interactions occur. Spatial transcriptomics complements scRNA-seq by preserving tissue architecture and localizing gene expression to specific regions within intact sections but with far fewer genes quantified than in scRNA-seq [220]. Thus, integration of these two technologies into a combined workflow offers both the high-resolution transcriptomic landscape of individual cells and the context of their physical locations.

### 2.1.2 NF-$\kappa$B and Mouse Models

The nuclear factor kappa-light-chain-enhancer of activated B cells (NF-$\kappa$B) is a family of transcription factors that plays a central role in regulating immune response, inflammation, and cell survival [101]. NF-$\kappa$B, along with its canonical upstream activator, the inhibitor of nuclear factor kappa B kinase (IKK) complex, are broadly expressed across many cell types, including fibroblasts [297]. In various inflammatory skin diseases, elevated NF-$\kappa$B signaling has been closely associated with the amplification of pro-inflammatory responses [169]. Moreover, while NF-$\kappa$B is not typically the initiating driver of inflammation, it is a key mediator of downstream responses to pro-inflammatory stimuli, contributing to the persistence and propagation of pathogenic processes.

Mouse models of AD have been instrumental in elucidating the immunological and cellular mechanisms underlying the disease. These models fall into three main categories: sensitization models, in which chemical or allergen exposure induces AD-like inflammation; transgenic models, which involve targeted genetic modifications that drive skin barrier dysfunction or immune dysregulation; and spontaneous models, in which inbred mouse strains naturally develop AD-like skin lesions under specific pathogen-free or conventional conditions [135]. Also, while no model fully recapitulates the complex etiology and chronic relapsing nature of human AD, many capture multiple key features. Comparative analyses have highlighted both conserved and species-specific aspects of AD pathogenesis, making mouse models valuable tools for probing disease-relevant pathways. Their accessibility, lower cost, and amenability to experimental manipulation under controlled conditions make them particularly useful for pre-clinical testing of therapeutic interventions and for validating findings from human single-cell and spatial datasets. In this study, we use a genetically engineered mouse model in which NF-$\kappa$B activation was blocked in fibroblasts by deleting inhibitor of nuclear factor kappa-B kinase subunit beta (*Ikkb*), an NF-$\kappa$B regulator, resulting in AD-like skin lesions.

Integrating human and mouse datasets presents both technical challenges and powerful opportunities for comparative analysis in atopic dermatitis research. A key challenge lies in accurately mapping cell populations and gene expression programs across species, given differences in cell subset composition, gene regulation, and the availability of homologous genes. Additionally, experimental variation, such as differences in sample preparation, sequencing protocols, and tissue microenvironments typically lead to significant batch effects, further complicating comparisons. Nevertheless, cross-species comparison allows for the identification of disease mechanisms that are conserved between human AD and mouse models, providing insights into the translational relevance of mouse model findings [177]. The use of multiplex network frameworks and multi-omic integration algorithms across datasets have improved our ability to align and compare multiple omics layers, such as cross-species single-cell and spatial transcriptomic datasets (see 1.2.2). These approaches enhance the resolution and interpretability of complex biological signals and can accelerate the discovery of shared molecular drivers of disease.

### 2.1.3 Study Objectives

In this chapter, we aim to characterize the cell-type–specific and spatial transcriptomic alterations underlying atopic dermatitis using both our mouse model of human AD and human patient samples with the goal of identifying novel therapeutic targets. Special emphasis is placed on dissecting the roles of fibroblasts and their subpopulations in disease pathogenesis, given their emerging importance in immune modulation [35]. By integrating scRNA-seq and spatial transcriptomics data across species, we seek to validate the relevance of our mouse model to human AD and to identify conserved inflammatory pathways and intercellular signaling networks. This work also aims to demonstrate the value of combining multi-modal and cross-species data through multiplex-network-based systems biology approaches using the MENTOR multi-omic integration algorithm, which both offers a scalable framework

for multi-omic data integration and enables the systemic and data-driven study of complex biological systems.

## 2.2 Results and Discussion

To explore the role of canonical NF-$\kappa$B signaling in skin fibroblasts, we used Cre-lox recombination to generate *Prx1Cre*$^{+}$*Ikkb*$^{\mathrm{f/f}}$ mice in which *Ikkb*, the murine homolog of the human *IKBKB* gene is selectively deleted from Prx1-lineage mesenchymal cells, which are restricted to the ventral skin [157]. The conditional knockout (cKO) mice exhibited lowered NF-$\kappa$B nuclear localization levels in fibroblasts and adipocytes compared to wildtype (WT) mice, leading to ventral skin lesions marked by epidermal thickening, inflammatory infiltrates, and other features resembling those seen in human atopic dermatitis [207]. The inflammation was localized to ventral skin, preceded visible lesions and appeared in fibroblast-specific but not adipocyte-specific *Ikkb* deletion in further mouse testing, suggesting a fibroblast-driven immune dysregulation. Additional postnatal deletion experiments resulted in mild inflammation in 4-week-old mice but no inflammation in 8-week-old mice, further identifying a critical perinatal window during which fibroblast NF-$\kappa$B signaling is required to establish immune homeostasis in the skin. Histological and flow cytometry analyses revealed increased infiltration of neutrophils and myeloid cells, but not T cells, in the ventral skin of both young and adult *Ikkb*-deleted mice. Moreover, *Prx1Cre*$^{+}$*Ikkb*$^{\mathrm{f/f}}$ mice exhibited systemic effects including elevated circulating monocytes and eosinophils, splenomegaly (i.e., enlarged spleen), and myeloid expansion in peripheral lymphoid organs, despite unaltered bone marrow myelopoiesis.

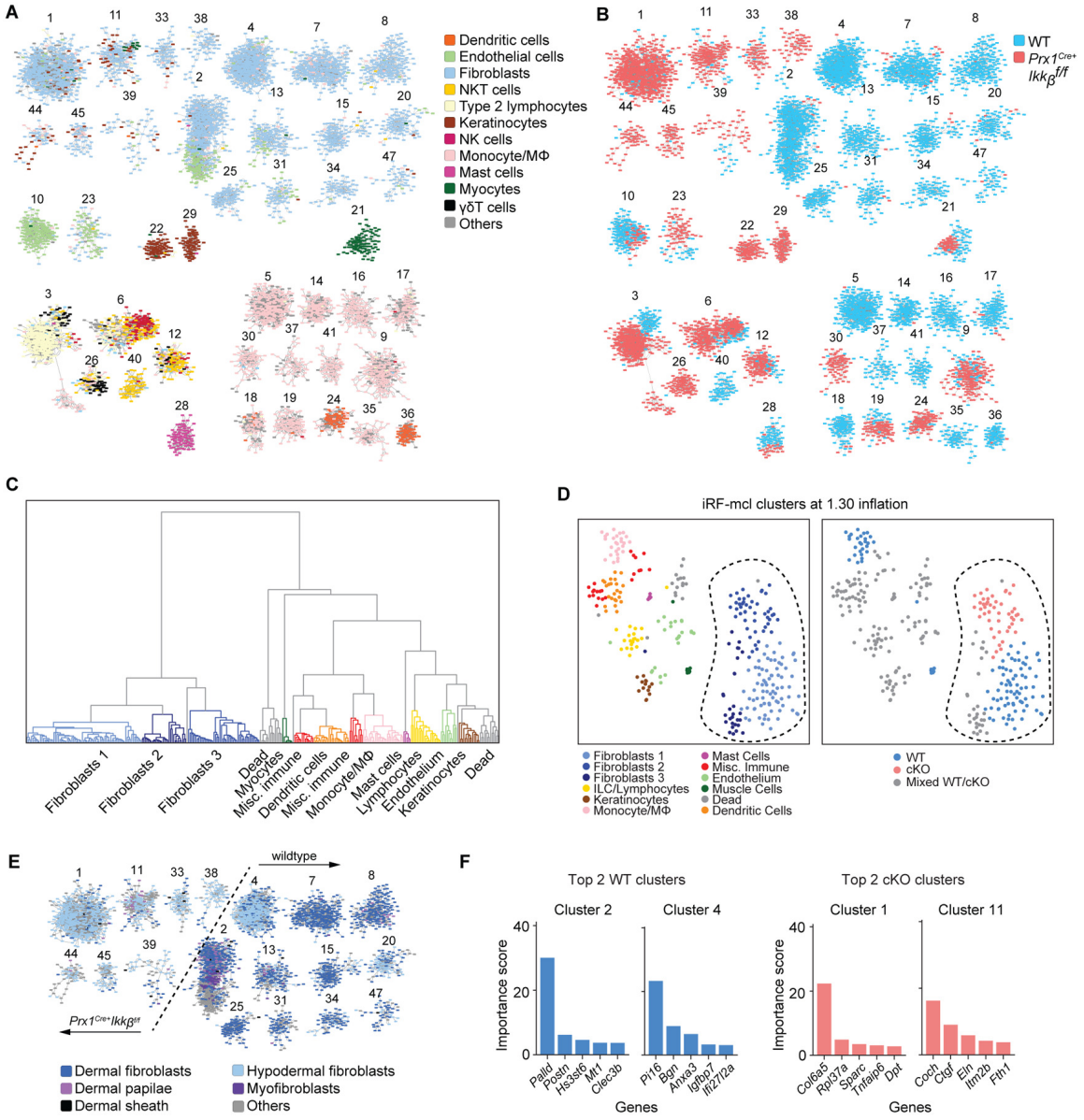



**Figure 2.1:** Clustering and annotation of the WT and cKO mouse skin scRNA-seq data. (A) Annotated iRF-LOOP + MCL clustering of 11,084 cells from WT and cKO mouse skin. Clusters with >50 cells are shown. (B) WT (blue) and cKO (red) groups demonstrating an even cell distribution (3,566 cells from WT, 2,510 cells from cKO). (C) Hierarchical partitioning of Markov clusters and cell annotation, using Ward's linkage method between Pearson correlations of mean expression vectors of the clusters. (D) UMAP plots with cell annotations (left) and WT/cKO designation (right), demonstrating clear separation of fibroblast clusters (encircled in dashed line) derived from the *Ikkb*-deleted group. Color legends are shared across A, B, C, and D. (E) Fibroblast subtyping based on [136] demonstrating a transcriptomic shift towards hypodermal fibroblasts when *Ikkb* is deleted (left of the dashed line). (F) Importance scores for genes determining the two largest WT and cKO clusters.

### 2.2.1 The Eosinophilic Role of Inflammatory Fibroblasts

To characterize the transcriptomic changes driven by *Ikkb* deletion in Prx1-lineage cells, we performed scRNA-seq on ventral skin samples from 4-week-old *Prx1Cre*$^{+}$*Ikkb*$^{f/f}$ mice (4,485 cells) and littermate controls (6,599 cells) in this same publication. We applied iterative Random Forest Leave-One-Out Prediction (iRF-LOOP) [49], an explainable machine learning approach, to construct a predictive cell–cell network using each cell's gene expression profile as a feature vector. To identify transcriptionally coherent groups, we applied Markov Clustering (MCL) [271] with an inflation parameter of 1.30 to the top 0.4% of edges in the network, resulting in 316 distinct clusters (Fig. 2.1), each comprising cells with mutually predictive expression profiles. Cell types were assigned probabilistically based on known marker genes using the CellAssign R package [296], and these annotations were mapped onto the iRF-LOOP + MCL clusters and visualized in Cytoscape [245] (Fig. 2.1 A). This high-resolution clustering approach enabled the identification of transcriptional shifts and altered cell population distributions between control and cKO samples (Fig. 2.1 B).

To assess transcriptional relationships among fibroblast subpopulations, we applied hierarchical clustering with Ward's linkage method [279] to the mean gene expression vectors of each iRF-LOOP + MCL fibroblast cluster after identification of fibroblast clusters through cell typing. These relationships were visualized using the uniform manifold approximation and projection (UMAP) method [194] of the cluster-averaged vectors. This analysis revealed three major fibroblast groups, one of which, fibroblast group 2, was predominantly composed of cells from the *Ikkb*-deleted mice, indicating a substantial shift in fibroblast transcriptomic state following *Ikkb* loss (Fig. 2.1 C–D). Fibroblasts from control mice aligned with previously defined subtypes based on the classification by Joost et al. [136], whereas fibroblasts from the *Ikkb*-deleted group largely resembled a hypodermal-like transcriptional profile (Fig. 1.1 E). To identify genes distinguishing these fibroblast states, we applied iRF using each iRF-LOOP + MCL cluster as a class label to rank genes predictive of each cluster.

Control fibroblast clusters were most strongly predicted by *Palld*, a cytoskeletal gene broadly expressed in mesenchymal cells, and *Pi16*, a known modulator of neuropathic pain (Fig. 2.1 F). In contrast, the top fibroblast clusters from *Ikkb*-deleted mice were characterized by *Col6a5*, a collagen gene upregulated in inflammatory fibroblasts in human AD [113, 198], and *Coch*, which is typically expressed in inner ear tissue. These findings suggest that the fibroblasts in *Ikkb*-deficient skin adopt functionally distinct gene expression programs from those in healthy controls and with some similarity to those found in the fibroblasts of human AD patients.

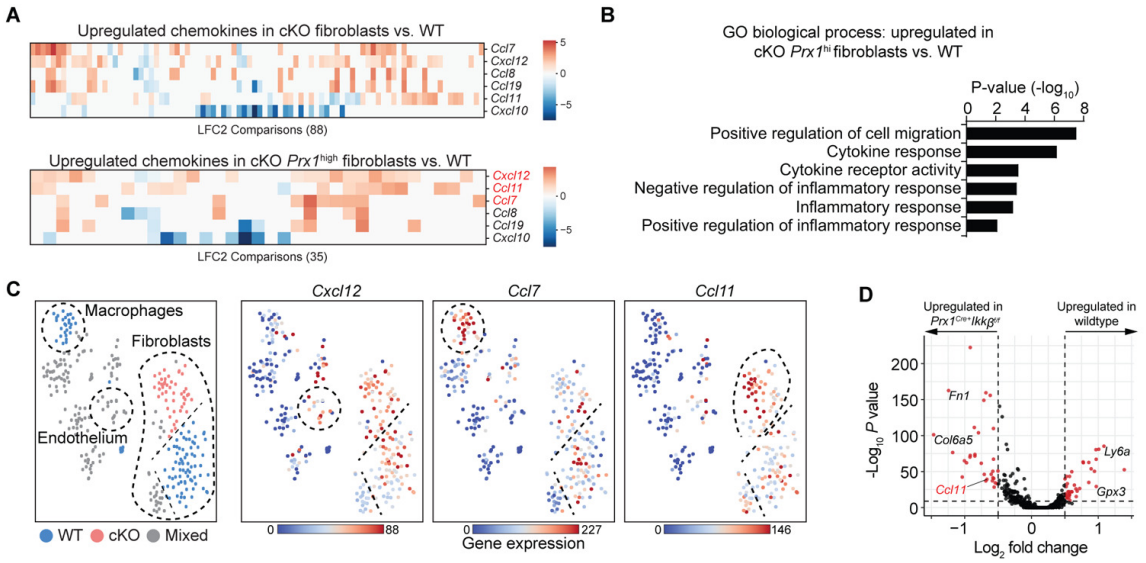



**Figure 2.2:** An immune-regulatory inflammatory fibroblast population in cKO mice. A. Pairwise WT to cKO comparisons of chemokine expression between fibroblast clusters (top) and *Prx1*-expressing fibroblasts (bottom). Notable upregulated chemokines were *Cxcl12*, *Ccl11*, and *Ccl7*. LFC2: log2 fold-change. B. Gene ontology (GO) analysis of biological processes related to chemotaxis and inflammation based on genes upregulated in *Prx1*-expressing fibroblasts when comparing cKO mice to WT. C. UMAP plots of iRF-LOOP + MCL clusters with gene expression heatmaps for *Cxcl12*, *Ccl7*, and *Ccl11*, with notable upregulation of *Ccl11* in cKO fibroblast clusters. D. Volcano plot of genes upregulated in WT and cKO groups, using each fibroblast as a datapoint.

Building on the observation of fibroblast heterogeneity, we hypothesized that *Ikkb* deletion in *Prx1*-lineage fibroblasts promotes immune cell infiltration through dysregulated chemokine expression. To investigate this, we conducted pairwise

differential gene expression analysis across all fibroblast clusters from control and *Ikkb*-deleted mice using DESeq2 [178], yielding 88 comparisons. Several chemokines were significantly upregulated in the *Ikkb*-deleted group, including *Ccl7*, *Cxcl12*, *Ccl8*, *Ccl19*, and *Ccl11* (Fig. 2.2 A). Notably, *Ccl11*, *Cxcl12*, and *Ccl7* were elevated in Prx1-expressing fibroblasts (42% of all fibroblasts in the deletion group) relative to their control counterparts (48%) (Fig. 2.2 A). Gene ontology analysis of these cells showed enrichment in pathways related to cell migration, cytokine signaling, and inflammation (Fig. 2.2 B). Visualization of *Cxcl12*, *Ccl7*, and *Ccl11* expression on the UMAP of iRF-LOOP + MCL clusters revealed that *Cxcl12* and *Ccl7* were predominantly expressed in endothelial cells and macrophages, respectively, while *Ccl11* was fibroblast-specific and markedly upregulated in *Ikkb*-deleted clusters (Fig. 2.2 C). Immunofluorescence analysis further validated that $CCL11^+$ fibroblast-like cells were enriched at the dermal–subcutaneous boundary in ventral skin of *Ikkb*-deleted mice. Given that CCL11 is a strong chemoattractant for eosinophils, we quantified eosinophil infiltration and found a significant increase in $F4/80^+$Siglec-$F^+$ eosinophils in the ventral skin of *Ikkb*-deleted mice. Co-staining with CD80, CD206, and Siglec-F revealed that over 80% of $F4/80^+$ cells in this region were eosinophils, distinguishing them from macrophages. Together, these findings indicate that *Ikkb* deletion in fibroblasts drives aberrant *Ccl11* expression, contributing to early eosinophilic infiltration and inflammation in the skin.

### 2.2.2 Human AD and Data Integration with MENTOR

In an initial publication, we compared our transcriptional findings to those from two published human AD scRNA-seq datasets [113, 233]. We used differential gene expression across control and AD patient samples of PRRX1-expressing fibroblast clusters to compare CCL11 and other gene expression levels. Both datasets exhibited upregulation of COL6A5 and CCL11 in a particular subpopulation of AD sample

fibroblasts, indicating both the existence of a population of inflammatory fibroblasts and correlating the findings of an eosinophilic role for fibroblasts in our mouse model.

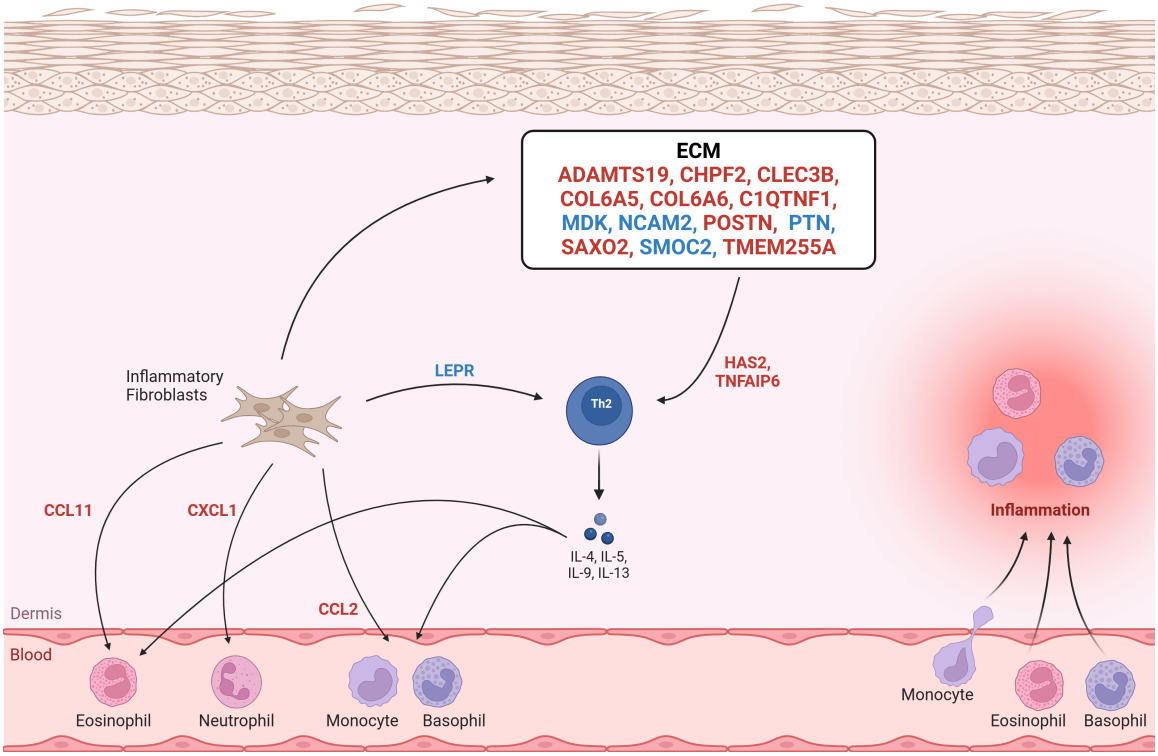



**Figure 2.3:** Conceptual model of all the genes in an immune-regulatory and ECM-related MENTOR clade of fibroblast-mediated inflammation of atopic dermatitis. Datasets are from two human AD scRNA-seq datasets and our mouse model of AD scRNA-seq dataset. Genes upregulated in AD samples versus controls are in red and those downregulated are in blue.

In a subsequent analysis, we used MENTOR [257] to generate three functionally and hierarchically related gene clades from an input set of 69 genes, with the input geneset defined by the intersection of up- and down-regulated differentially expressed genes (DEGs) identified from comparisons between inflammatory fibroblast clusters and their respective controls across our murine scRNA-seq dataset and two human scRNA-seq datasets. The largest clade (Fig. 2.3) comprised 34 genes involved in immune regulation, extracellular matrix (ECM) organization, and ECM–Th2 signaling interactions, processes central to the epidermal–immune crosstalk characteristic of atopic dermatitis. The smallest clade consisted of 9 genes,

including both CCL11 and COL6A5. Shortest-path analysis within the multiplex network used by MENTOR revealed that these two genes are connected through predictive expression layers derived from bulk RNA-seq data of fibroblasts and peripheral blood, suggesting a mechanistic link between the eosinophil-attracting chemokine CCL11 and the inflammatory fibroblast marker COL6A5.

### 2.2.3 Spatial Context

To further characterize the spatial organization of inflammatory cell populations within the tissue microenvironment, we performed Xenium single-cell spatial transcriptomics on ventral skin sections from our murine model of atopic dermatitis. In contrast to dissociative scRNA-seq approaches, Xenium preserves tissue architecture while quantifying transcript abundance at single-cell resolution, enabling direct localization of transcriptionally defined cell populations within intact tissue sections (see 1.4). Because Xenium captures a substantial proportion of transcripts outside segmented nuclei and these transcripts suffer from cell segmentation overlap and ambient transcript contamination [132, 223, 189], we restricted clustering analyses to transcripts overlapping nuclear boundaries, which represented approximately 81% of all detected transcripts. This approach reduced potential transcript assignment ambiguity, increasing clarity and confidence in cell type identification.

Leiden clustering of the nuclear transcripts resulted in 21 transcriptionally distinct cell populations spanning epithelial, mesenchymal, and immune cell compartments (Fig. 2.4A and Fig. 2). Manual annotation based on canonical marker genes (Table 2) and cluster top DEGs identified multiple keratinocyte differentiation states, inflammatory granulocyte populations, fibroblast populations, endothelial cells, lymphocytes, Schwann cells, sebocytes, and mesenchymal stromal populations. The resulting UMAP demonstrated clear separation between epidermal and stromal lineages while also revealing multiple inflammatory and activated cellular states, particularly among fibroblast and keratinocyte populations. Notably, activated

fibroblast populations separated into two transcriptionally distinct clusters, with cell type abundance comparisons between control (dorsal 20 week) and experimental (ventral 9 and 20 week) mouse populations showing a clear upregulation in one fibroblast population and a clear downregulation in the other (Fig. 3). This separation of fibroblast populations is consistent with the inflammatory fibroblast programs identified in both the murine and human scRNA-seq analyses described previously.

Projection of the annotated clusters back onto one of the tissue sections (tissue section K-2) revealed strong spatial organization of the identified cell populations (Fig. 2.4B). Keratinocyte populations localized predominantly to the epidermis, while fibroblast, adipocyte, endothelial, and immune populations occupied distinct dermal and hypodermal regions. Importantly, the spatial transcriptomic data identified a focal inflammatory region characterized by dense accumulation of inflammatory granulocytes and activated fibroblast populations. Comparison with the matched H&E section demonstrated strong concordance between histologically identified inflammatory infiltrate and transcriptomically defined immune-rich regions (Fig. 2.4C). This agreement between histology and spatial transcriptomics supports the biological validity of the identified inflammatory microenvironments and demonstrates that the transcriptomic architecture captured by Xenium accurately reflects underlying tissue pathology.

To investigate higher-order tissue organization beyond individual cell types, we next performed spatial neighborhood analysis using the scimap framework [206]. Spatial neighborhood or "niche" analyses have emerged as a powerful approach for identifying multicellular environments (i.e., spatially mixed cell populations) that cannot be resolved through cell type analysis alone [241, 130]. These methods cluster cells according to shared neighborhood transcriptomic profiles based on spatial proximity, enabling the identification of spatially organized cell neighborhoods. Using this approach, we identified 20 distinct spatial niches distributed across the tissue section (Fig. 2.4D). Several niches corresponded closely to expected anatomical

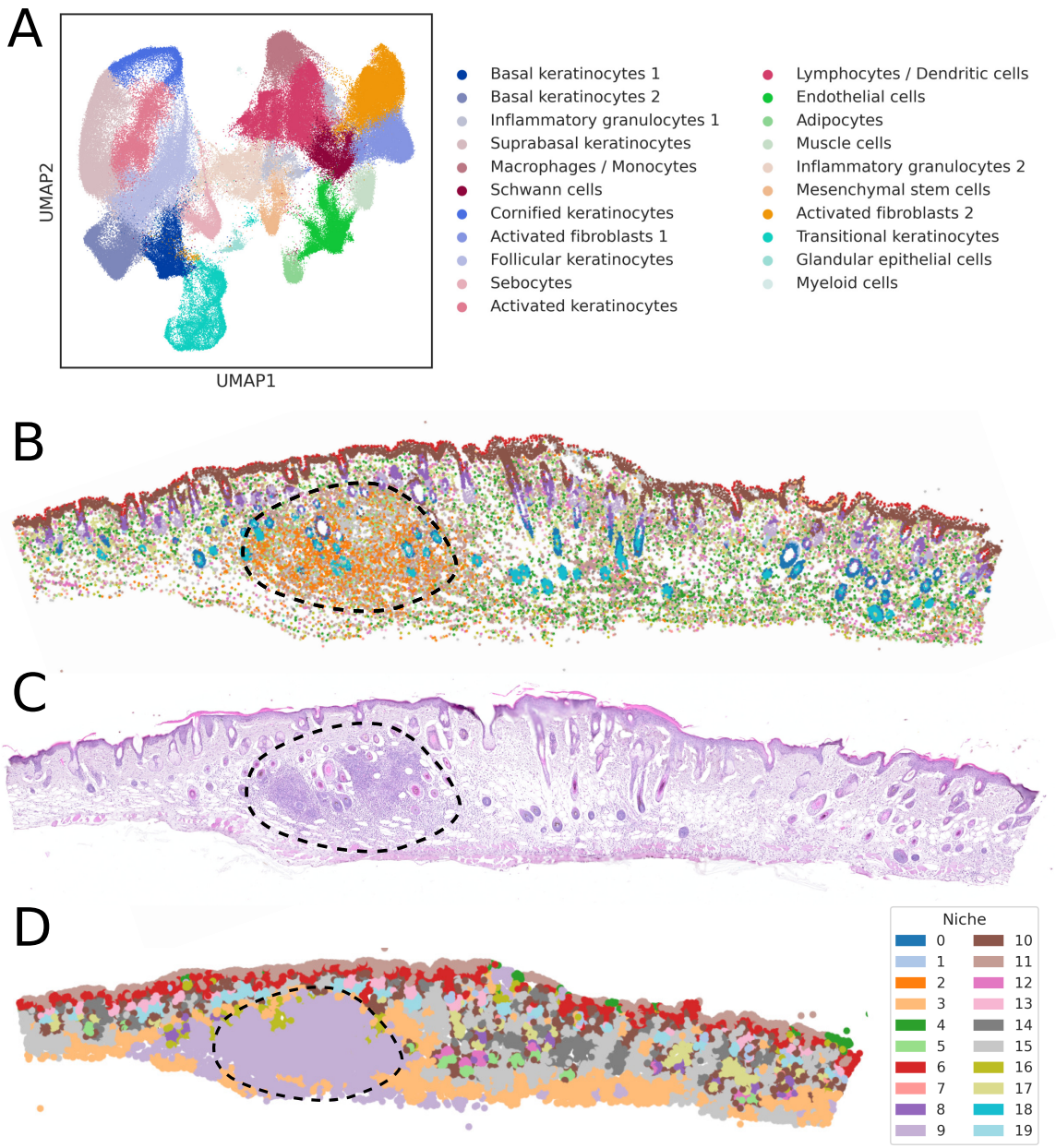



**Figure 2.4:** Single-cell and spatial transcriptomic characterization of inflammatory regions in murine atopic dermatitis skin. (A) UMAP visualization of 21 Leiden clusters (resolution = 0.6) generated from Xenium single-cell spatial transcriptomic data using only nuclear-overlapping transcripts (81% of all detected transcripts). Clusters were manually annotated based on canonical marker gene expression, identifying fine-grain epithelial, immune, and mesenchymal subpopulations. (B) Spatial projection of the annotated cell populations onto Xenium sample K-2, demonstrating preservation of tissue architecture and localization of immune and stromal cell populations. (C) Corresponding hematoxylin and eosin (H&E)-stained section of sample K-2. (D) Spatial neighborhood ("niche") assignments generated using the scimap spatial neighborhood workflow, identifying 20 transcriptionally and spatially distinct microenvironmental niches across the tissue section. The dashed black outlines in panels B–D indicate a manually identified inflammatory region characterized by dense inflammatory granulocyte infiltration in panel B and corresponding primarily to niche 9 in panel D, demonstrating concordance between histopathology, cell-type composition, and spatial neighborhood organization.

structures, including epithelial and stromal regions, while others captured various localized microenvironments enriched for immune and activated stromal populations.

Notably, the inflammatory region identified histologically and transcriptomically corresponded predominantly to niche 9 (Fig. 2.4D), suggesting the existence of a distinct inflammatory tissue microenvironment associated with granulocyte infiltration and activated stromal states. Cell type composition of this niche showed it was largely composed from various immune cell populations (63.2% combined) and from one of the two activated fibroblast populations (6.8%). This niche exhibited spatial enrichment directly beneath the epidermis within the inflamed dermal compartment and appeared compositionally distinct from surrounding stromal regions. The concordance between histological inflammation, inflammatory granulocyte accumulation, and niche-level organization suggests that the inflammatory response in this model is spatially structured rather than diffusely distributed throughout the tissue. Similar spatially organized inflammatory fibroblast and immune cell interactions have been observed in other inflammatory skin diseases and tumor microenvironments, where localized multicellular niches coordinate cytokine signaling, immune recruitment, and extracellular matrix remodeling [113, 233, 133]. The identification of analogous inflammatory niches in our model further supports the hypothesis that fibroblast-mediated immune regulation contributes to the maintenance of localized inflammatory microenvironments in AD.

## 2.3 Methods

### 2.3.1 Murine scRNA-seq Dataset

Ventral skin from WT or *Prx1Cre*$^{+}$*Ikkb*$^{f/f}$ mice (4-week-old) was enzymatically digested and immediately sorted on FACSAria II SORP instrument to exclude debris, dead cells (DAPI-positive) and aggregates. A pool of two mice per group was used to generate a single library, for a total of 5 complementary DNA (cDNA) libraries (2 for

WT and 3 for the *Ikkb* deletion group). Single cell barcode capture was performed on a GemCode instrument platform with the 10X single-cell 3' v2 chemistry (10X Genomics). cDNA libraries were prepared and sequenced on an Illumina HiSeq 4000 instrument. Single cell barcode capture, quality control using BioAnalyzer, cDNA amplification and RNA sequencing were performed at the University of Pennsylvania Next Generation Sequencing Core. Using Cell Ranger software (10X Genomics, version 1.3.0) [298] with default settings, the resulting raw reads were aligned to mouse reference genome GRCm38. Cells with an insufficient unique molecular identifier (UMI) count (<500 per cell) or an unusually high UMI count (>25,000 UMIs per cell) were discarded from all downstream analyses in order to remove any broken cells and multiplets. Data was further cleaned to remove gene vectors with expression in less than 1% of the cells and cells with unusually high mitochondrial content (>5%), resulting in a gene by cell matrix of 28,684 genes by 11,084 cells.

### Clustering and Cell Annotation

The preprocessed expression data was clustered using the iterative random forest leave-one-out-prediction (iRF-LOOP) [49] machine learning method and the R Seurat Package (version 3.0) [255]. iRF-LOOP was applied to the expression matrix by setting the cells as features and genes as samples, using five iterations, and using 1000 trees per iteration. This generated a set of approximately 14 million gene interactions (edges) between the 11,084 cells (nodes), with each edge having an iRF-LOOP assigned importance score (weighted edges). Of these, the top 0.4% (55,588) edges by importance score were kept, leaving an average of five gene interactions per cell. This network was then clustered using a high-performance implementation [10] of the Markov Clustering algorithm [271] at various inflation values between 1.2 and 2.5. The 1.3 inflation cluster set was chosen for further analyses because it contained the largest number of mid-granularity clusters (316 clusters, each with >50 cells), maximizing the likelihood that these were biologically relevant.

To determine which genes within a cluster contributed to distinguishing that cluster of cells from the rest of the cells, iRF was applied with the genes as features and all cells as the samples, in which the dependent variable was a binary vector of 1 for the cells in the cluster of interest and 0 otherwise. The iRF models used to determine the genes of importance for these clusters also used 5 iterations and 1000 trees per iteration. The resulting iRF-based cluster-defining genes were also ranked by their importance scores. For analysis with the Seurat library, cell gene expression profiles were normalized and square-root transformed, followed by principal component analysis (PCA) to the first 50 components for dimensionality reduction. The Louvain [23] algorithm, a modularity optimization-based method for clustering nearest-neighbor graph embeddings, was used to identify cell clusters. Cell clusters were visualized using the UMAP algorithm and other visualizations were generated using Seurat plotting functions such as Heatmap, Featureplot, Vlnplot, and Dotplot.

#### Differential Gene Expression

For differential expression, DESeq2 [178] was used to compare iRF-LOOP + MCL clusters in pairwise fashion by setting cells within each cluster as samples and by using a minimum of 1.5 absolute log fold 2 value and maximum adjusted p-value of 0.1. Because there are 8 fibroblast control clusters and 11 *Ikkb* deletion fibroblast clusters, the total number of pairwise differential expression analyses was 88. Another differential expression analysis was performed between Prx1+, *Ikkb*-deleted fibroblasts and Prx1+, control group fibroblasts, also using DESeq2 with the same settings as above, but with the groups dependent not on clusters but on whether or not a cell belonged to a WT or cKO sample.

### 2.3.2 Human scRNA-seq Datasets

Raw fastq files were downloaded for two publicly available human control and AD scRNA-seq datasets ([113, 233]) and processed with identical pipelines except for

cell population subsetting. For both datasets, fastq files were downloaded and aligned with STAR (v 2.7.9a) [66] to human reference genome GRCh38 using STAR's CellRanger-matching settings with STARsolo, resulting in 435,527 cells for Reynolds and 16,675 cells for He across all filtered cell matrices. Cells from the Reynolds dataset, which contained patient samples from multiple different skin diseases, were subsetted to exclude cells from patients with a condition other than AD, resulting in a new total of 26,656 cells for Reynolds. Next, using the Seurat [255] workflow, expression counts were log-normalized and scaled, and canonical correlation analysis was applied to remove batch effect for the Reynolds dataset only, which displayed a strong sample-specific batch effect. Mesenchymal cells were identified by their expression of Mesenchymal marker genes (COL1A2, COL3A1, PRRX1) and DESeq2 was applied with non-default, but single-cell recommended settings (useT=TRUE, minmu=1e-6, and minReplicatesForReplace=Inf) to the filtered count matrices in order to obtain DEGs between the PRRX1-expressing cells in the combined mesenchymal clusters by condition (healthy controls versus AD).

### MENTOR Multiplex and Geneset

The multiplex for the MENTOR run was created from 30 different network layers: a bulk RNA-seq peripheral blood predictive expression network, bulk RNA-seq fibroblast predictive expression network, a transcription factors network, five layers from Hummannet [152], and 22 single-cell predictive expression networks from different skin cell populations. None of these network layers were created from any the murine or human scRNA-seq experiments analyzed in this work. The input geneset for MENTOR was defined by the intersection of up- and down-regulated DEGs of comparisons between inflammatory fibroblast clusters and their respective controls across all three scRNA-seq datasets. Shortest paths analysis of gene pairs in clades was done using RWRtoolkit [143]. Analysis of biological processes and functions of genes within clades was performed using the GOrilla [70] web-based application.

### 2.3.3 Single-cell Spatial Transcriptomics

Single-cell spatial transcriptomic profiling was performed using the Xenium platform (10x Genomics) [132] on murine dorsal and ventral skin sections collected from 9-week-old and 20-week-old mice. The dataset consisted of 21 skin sections derived from 7 mice, with each mouse contributing matched dorsal and ventral skin samples. Ventral skin sections corresponded to inflammatory lesion-bearing tissue from the *Prx1Cre*$^{+}$*Ikkb*$^{\mathrm{f/f}}$ mice, while dorsal skin sections, which did not exhibit lesions, served as internal controls. The 21 sections comprised ventral 9-week-old, ventral 20-week-old, and dorsal 20-week-old samples distributed across two Xenium slides containing 12 imaging regions labeled A, B, C, D, E, F, G, I, J, K, L, and M, with each imaging region containing between one and three skin sections. Xenium output files were processed using the Scanpy [284] and scimap [206] Python frameworks together with custom analysis scripts. Transcript counts, segmentation boundaries, and spatial coordinates were loaded into AnnData objects for downstream analysis.

#### Clustering

To reduce transcript assignment ambiguity arising from transcript spillover and segmentation overlap outside cell boundaries, analyses were restricted to transcripts overlapping segmented nuclei. Xenium transcript parquet files were filtered to retain only transcripts with `overlaps_nucleus == 1` and valid cell assignments. Nuclear-overlapping transcripts were then reconstructed into sparse cell-by-gene count matrices and stored as a separate AnnData layer. The resulting nuclear transcript matrices represented approximately 81% of all detected transcripts.

For clustering analyses, raw nuclear transcript counts were assigned to the AnnData expression matrix and cells with fewer than 10 detected genes and genes expressed in fewer than 5 cells were removed. Expression values were library-size normalized and log-transformed using Scanpy preprocessing functions. Highly variable genes were identified using a batch-aware approach with Xenium slide identity

as the batch variable. Principal component analysis (PCA) was performed for dimensionality reduction followed by nearest-neighbor graph construction and UMAP visualization [194]. Cell populations were identified using Leiden clustering [266] with a clustering resolution of 0.6.

### Cell Annotation and Composition

Cell type annotation was performed through integration of multiple sources of information, including cluster-specific differential gene expression, spatial localization patterns, and canonical marker genes represented within the Xenium panel. Cluster marker genes were identified using Wilcoxon rank-sum differential expression testing implemented in Scanpy. Canonical marker genes for keratinocytes, fibroblasts, lymphocytes and dendritic cells, macrophages and monocytes, granulocytes, endothelial cells, adipocytes, and other stromal populations were evaluated using cluster-level expression visualization and dotplot analyses. Spatial projection of the resulting clusters back onto tissue sections was additionally used to validate anatomical consistency of the annotations, including localization of keratinocyte populations to epidermal regions and stromal and immune populations to dermal and hypodermal compartments.

Cell type composition analyses were performed using area-normalized abundance estimates in order to account for variation in tissue section size and segmentation density between samples. For each sample, the total segmented cell area associated with each annotated cell type was summed and divided by the total segmented tissue area, yielding cell densities normalized by tissue area (cells/$mm^2$). Cell type abundances were compared across dorsal and ventral conditions and visualized using horizontal bar plots. Additional statistical analyses were performed using Welch's t-tests and fixed-effects linear models on logit-transformed area fractions while accounting for biological replicate identity.

### Niche Analysis

Spatial neighborhood analyses were performed using the scimap framework [206]. Raw count matrices were converted to dense matrices and spatial coordinates were extracted from Xenium spatial embeddings. Neighborhood composition matrices were generated using the `spatial_expression` workflow with a k-nearest-neighbor (kNN) approach using 50 nearest neighboring cells within each tissue section. Spatial niches were subsequently identified through k-means clustering of neighborhood composition matrices using 20 neighborhood clusters. Resulting niche assignments were projected back onto tissue sections for visualization and comparison with histological inflammatory regions. Differentially expressed genes between niches were identified using Wilcoxon rank-sum testing implemented in Scanpy.

## 2.4 Conclusion

These results reveal a previously underappreciated immune-regulatory role for fibroblasts in eosinophilic inflammation in atopic dermatitis through chemokine expression. By integrating single-cell RNA sequencing datasets from both mouse and human skin, we demonstrate that deletion of *Ikkb* in *Prx1*-lineage fibroblasts in our mouse model of human AD mimics the transcriptomic shift toward an inflammatory fibroblast phenotype marked by COL6A5 and CCL11 expression observed in human AD. Complementary Xenium spatial transcriptomic analyses further demonstrated that these inflammatory and stromal populations are spatially organized into localized multicellular niches associated with granulocyte infiltration and histologically defined inflammatory regions. The use of machine learning–based clustering, cross-species DEG comparison, spatial neighborhood analysis, and multiplex network integration via MENTOR enabled us to identify conserved gene programs and tissue microenvironments linking fibroblast signaling to immune cell recruitment across both murine and human AD datasets. These findings highlight the translational relevance of our

mouse model and establish a framework for future spatially resolved, systems-level analyses of skin inflammation. As additional spatial transcriptomic datasets and other data modalities from human and murine AD become available, they will allow us to further integrate the structural context of diverse cell populations into our multiplex-network-based framework, helping disentangle the complex multicellular interactions underlying the pathophysiology of atopic dermatitis.

# Chapter 3

# HALRA: a High-performance and Parallel Implementation of the ALRA Imputation Algorithm

## 3.1 Introduction

Single-cell RNA sequencing (scRNA-seq) has transformed transcriptomic analysis by enabling characterization of gene expression at the resolution of individual cells, facilitating the discovery of previously unrecognized cell types, transitional cellular states, and complex tissue heterogeneity [261, 299]. Advances in droplet-based technologies and large-scale consortium efforts have led to the rapid generation of increasingly massive scRNA-seq datasets, with modern cell atlases now routinely containing hundreds of thousands to millions of cells [232, 217]. While these datasets provide unprecedented opportunities to study cellular systems at scale, they also introduce substantial computational and analytical challenges due to their high dimensionality, extreme sparsity, and rapidly growing memory requirements. Consequently, scalable computational methods capable of efficiently processing and interpreting large transcriptomic matrices have become increasingly essential for modern single-cell analysis workflows.

### 3.1.1 Imputation Algorithms

Single-cell RNA sequencing (scRNA-seq) datasets are characterized by extreme sparsity, typically containing only 5–15% nonzero entries even after quality control and preprocessing [226]. This sparsity arises from a combination of true biological absence of expression and technical limitations such as low transcript capture efficiency, stochastic sampling, and shallow sequencing depth [147, 258]. As a result, many genes that are expressed at low levels in a cell may not be detected, producing so-called technical zeros or dropout events. Although recent studies have demonstrated that droplet-based scRNA-seq data are not statistically zero-inflated relative to standard count distributions [258, 47], the resulting count matrices remain highly sparse and noisy, posing substantial challenges for downstream analyses.

One effort to address these limitations has been the development of a wide range of computational imputation methods have been developed for scRNA-seq data. These

methods attempt to recover missing signal and sometimes to denoise the signal as well while preserving biologically meaningful variation. Existing approaches employ diverse strategies, including diffusion-based smoothing, Bayesian inference, graph propagation, matrix factorization, and deep learning. For example, MAGIC uses diffusion maps to smooth expression profiles across transcriptionally similar cells [65], while SAVER employs a Bayesian framework to estimate posterior distributions of true gene expression levels [125]. Deep learning approaches such as DeepImpute and DCA utilize neural network architectures to model nonlinear gene-gene relationships and reconstruct denoised count matrices [7, 76]. Collectively, these methods have demonstrated the potential to improve downstream analyses such as trajectory inference, rare cell type detection, and differential expression testing [121, 44].

Despite these successes, the role of imputation in scRNA-seq analysis remains controversial. Multiple studies have shown that overly aggressive smoothing can obscure biologically meaningful heterogeneity, artificially inflate gene-gene correlations, and generate spurious transcriptional structure [6, 120]. In particular, methods that strongly propagate information across neighboring cells may blur distinctions between closely related cell states or create false transitions in developmental trajectories. Consequently, some workflows avoid imputation altogether, relying instead on dimensionality reduction and robust statistical modeling approaches that operate directly on sparse count matrices. Nevertheless, imputation remains highly valuable for specific downstream applications, particularly those involving lowly expressed genes, gene regulatory network inference, trajectory analysis, and recovery of subtle transcriptional programs [226, 6]. The continued development of imputation algorithms therefore reflects an ongoing effort to balance noise reduction with preservation of biological fidelity.

### ALRA

Among the various classes of imputation methods, low-rank matrix approximation approaches remain particularly attractive due to their interpretability, computational

efficiency, and strong theoretical foundation. These methods are motivated by the observation that gene expression matrices often possess an underlying low-dimensional structure driven by coordinated transcriptional programs, cell types and states, and cell lineages. By approximating the original expression matrix using a limited number of latent factors, low-rank methods can effectively denoise technical variability while preserving dominant biological signal. Unlike many graph-based or deep learning approaches, low-rank approximations are comparatively straightforward to interpret mathematically and typically involve fewer hyperparameters and modeling assumptions [121, 44]. Moreover, matrix factorization methods can leverage highly optimized linear algebra routines and sparse matrix representations, making them particularly well suited for large transcriptomic datasets.

Adaptively-thresholded Low Rank Approximation (ALRA) has emerged as one of the most effective and computationally practical low-rank imputation algorithms for scRNA-seq data [168]. Broadly, ALRA uses low-rank matrix reconstruction together with adaptive thresholding procedures to recover missing expression signal while attempting to distinguish technical zeros from biological zeros, thereby reducing excessive smoothing while retaining the denoising benefits of low-rank approximation. In addition to its biological interpretability, ALRA relies primarily on linear algebra operations rather than iterative probabilistic inference or deep neural network optimization, substantially reducing implementation complexity and computational overhead. However, despite these advantages, existing ALRA implementations remain limited by dense intermediate reconstruction steps and single-node execution, resulting in rapidly increasing memory requirements as scRNA-seq datasets continue to grow. These limitations motivate the development of scalable, distributed implementations that preserve sparse matrix representations throughout the imputation pipeline while leveraging high-performance computing infrastructure.

### 3.1.2 Computational Performance

While many scRNA-seq imputation algorithms demonstrate strong performance on small benchmark datasets, computational scalability has emerged as a major limitation as scRNA-seq dataset sizes continue to increase dramatically in size (Fig. 3.1). Modern scRNA-seq studies routinely contain hundreds of thousands to millions of cells [232, 217, 1], producing count matrices with tens to hundreds of billions of potential matrix entries. Algorithms requiring fully dense representations of these large matrices quickly run into severe memory limitations. For example, a count matrix containing 1,000,000 cells and 30,000 genes contains 30 billion total entries, requiring approximately 240 GB of memory to store as a dense double-precision matrix alone, even before accounting for temporary allocations, duplicated objects, or intermediate computational steps. Consequently, algorithms that rely on dense matrix reconstruction or repeated dense linear algebra operations quickly become infeasible on standard single-node systems.

These limitations are particularly severe for imputation algorithms that rely on fully dense intermediate matrix representations, such as ALRA. Even when sufficient memory is available, computation time can increase substantially as matrix dimensions grow to atlas-scale sizes. In addition, commonly used sparse matrix implementations impose architectural constraints that further limit scalability. For example, the `dgCMatrix` sparse matrix class used extensively throughout the R ecosystem stores matrix indices as signed 32-bit integers, restricting matrices to a theoretical maximum of approximately $2^{31} - 1$ (2.1 billion) nonzero entries. Although this limit may initially appear large, it can be exceeded by modern scRNA-seq datasets after imputation; for instance, a 5% sparse matrix containing 1.5 million cells and 30,000 genes or an imputed 25% sparse matrix containing 300 thousand cells and 30,000 genes both surpass this threshold.

As a result, scalable imputation of modern scRNA-seq atlases fundamentally requires high-performance computing (HPC) approaches capable of distributing both

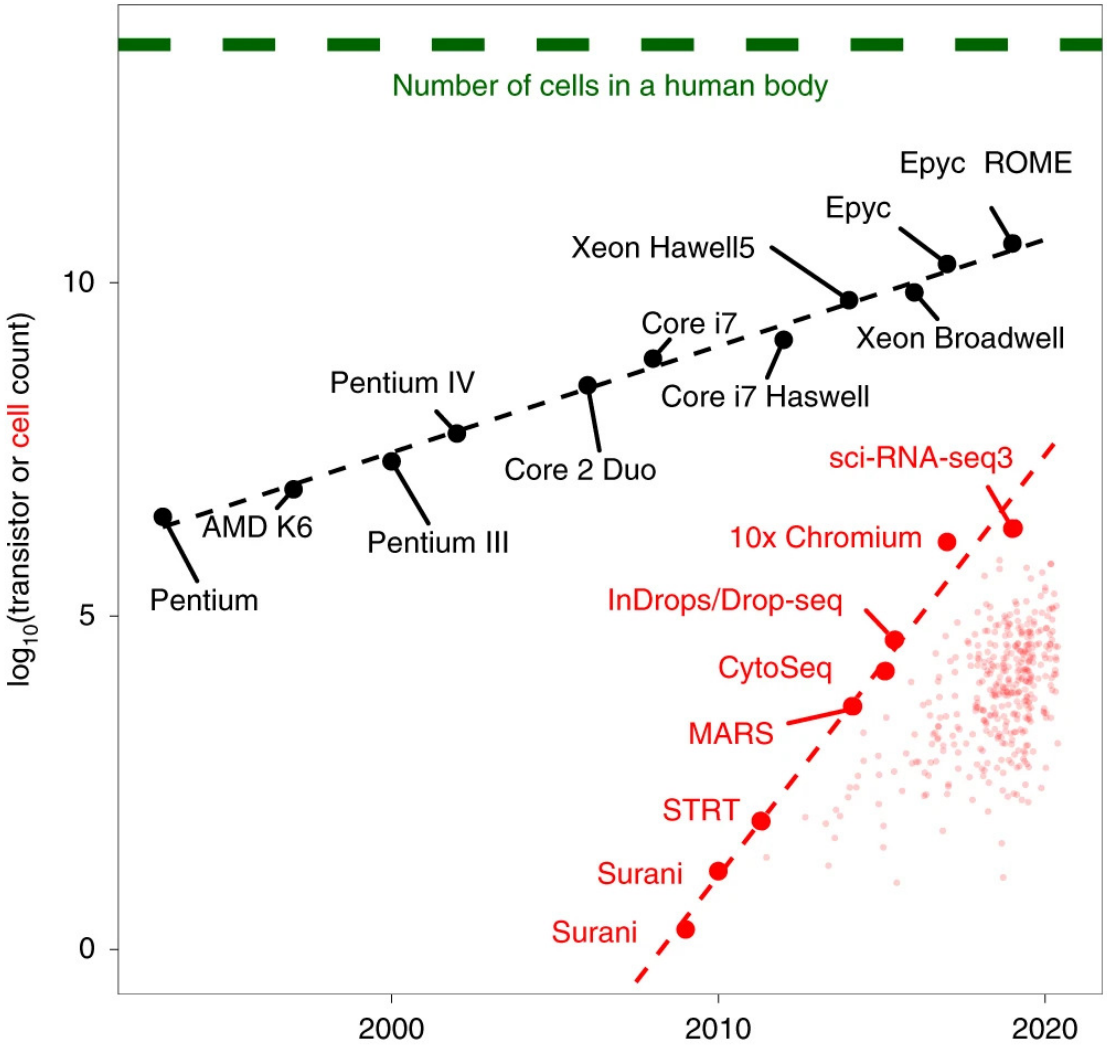



**Figure 3.1:** Growth in the number of cells profiled by landmark scRNA-seq datasets over time (red) compared with increases in CPU transistor counts (black). Small red points indicate published scRNA-seq studies included in the original 2021 analysis, while the green dashed line denotes the estimated number of cells in the human body. Adapted from [148].

memory and computation across multiple compute nodes. Distributed-memory parallelism enables large matrices to be partitioned into manageable blocks while avoiding centralization of dense intermediate representations. In addition, parallel file systems and distributed I/O strategies are increasingly necessary to efficiently read and write multi-hundred-gigabyte or even terabyte-scale scRNA-seq datasets. Despite these requirements, most currently available imputation algorithms, as

well as most scRNA-seq-specific algorithms in general, remain designed primarily for single-node execution and do not incorporate distributed-memory architectures, sparse-preserving reconstruction strategies, or parallel matrix processing workflows [259]. Consequently, there remains a substantial need for imputation frameworks and even more generally algorithmic frameworks specifically designed for modern HPC or cloud computing environments that preserve sparsity wherever possible while enabling scalable processing of modern-scale scRNA-seq datasets through distributed computation, blockwise matrix operations, and parallel input/output (I/O) methodologies.

### 3.1.3 Study Objectives

The primary objective of this study was to develop a high-performance implementation of ALRA capable of scaling to modern large-scale scRNA-seq datasets while preserving the biological and computational advantages of the original algorithm. To accomplish this, we redesigned ALRA around sparse matrix operations in order to minimize unnecessary dense matrix reconstruction and reduce memory consumption throughout the imputation pipeline. Although a Python implementation of ALRA, pyALRA, has since been developed, both pyALRA and the original R implementation still rely heavily on dense intermediate matrix representations that substantially increase memory requirements for large datasets, while the original R implementation additionally remains constrained by structural limitations associated with the `dgCMatrix` sparse matrix class [165]. In contrast, our implementation reformulates key stages of the algorithm to preserve sparsity wherever possible while maintaining compatibility with ALRA's adaptive thresholding framework for biological zero preservation [168]. We additionally developed the implementation within the Python scientific computing ecosystem in order to leverage mature sparse linear algebra libraries and high-performance computing interfaces such as mpi4py [57, 111, 273]. A major goal of this redesign was to ensure that the resulting implementation retained

comparable biological fidelity to existing ALRA implementations while substantially improving scalability and memory efficiency.

A second major objective of this work was to develop a distributed-memory implementation of ALRA capable of processing ultra-large scRNA-seq datasets across HPC systems. To achieve this, we implemented distributed normalization procedures, scalable low-rank decomposition workflows, and blockwise reconstruction strategies designed to partition computation and memory across multiple compute nodes. Particular emphasis was placed on minimizing memory duplication, avoiding centralized dense matrices, and enabling efficient distributed processing of sparse transcriptomic data. We further developed parallel I/O workflows capable of reading and writing large-scale single-cell datasets using HDF5-based formats commonly employed in modern scRNA-seq analysis pipelines [272]. Finally, beyond ALRA specifically, we aimed to establish a broader set of high-performance computational design principles for scalable single-cell algorithms, including sparse-preserving computation, distributed-memory parallelism, blockwise matrix processing, and parallel I/O strategies that may be applicable to other transcriptomic analysis methods operating at atlas scale.

## 3.2 Results and Discussion

### 3.2.1 Memory Efficiency in Sparse Matrices

A major limitation of the original ALRA implementation is its dependence on the R `dgCMatrix` sparse matrix class and the frequent use of dense intermediate matrix operations during reconstruction and scaling [168]. While effective for relatively small datasets, these design choices substantially limit scalability for modern atlas-scale or even moderately large ($\geq$ 50,000 cells) scRNA-seq datasets. During initial testing, we found that memory utilization increased rapidly during reconstruction of the low-rank approximation matrix, even when the original input count matrix

remained highly sparse. In particular, the adaptive thresholding and scaling stages of ALRA frequently required temporary dense matrix representations, producing RAM requirements that scaled approximately with the total number of matrix entries rather than the number of nonzero values. These limitations became prohibitive for matrices containing hundreds of thousands of cells, where dense intermediate representations alone could require hundreds of gigabytes of memory. Similar memory bottlenecks have been observed broadly across single-cell analysis workflows as datasets continue to expand toward atlas scale [179, 259].

To address these constraints, we developed HALRA (High-performance ALRA), a sparse-preserving reimplementation of ALRA in Python designed specifically for large-scale transcriptomic datasets and high-performance computing environments. The implementation was released as both an open-source software repository (https://github.com/jeanmerlet/halra ) and a publicly installable Python package through the Python Package Index (PyPI) (https://pypi.org/project/halra/ ), facilitating reproducibility and broader accessibility within the single-cell analysis community. HALRA was implemented using SciPy sparse matrix formats, primarily leveraging compressed sparse column (CSC) and compressed sparse row (CSR) matrix representations [273]. CSC matrices were used for operations requiring efficient column-wise access, including gene-level thresholding and variance calculations, while CSR matrices were employed for row-oriented operations such as normalization and distributed partitioning across cells. This dual-format strategy reduced the computational overhead associated with repeated sparse indexing operations and allowed different stages of the algorithm to be optimized according to matrix access patterns. In contrast to the original implementation, which relied heavily on dense matrix manipulation, HALRA preserved sparsity throughout most reconstruction and thresholding procedures by operating directly on sparse matrix structures whenever possible.

One of the most important modifications involved reformulating the adaptive thresholding stage to avoid explicit dense threshold matrices. In the original

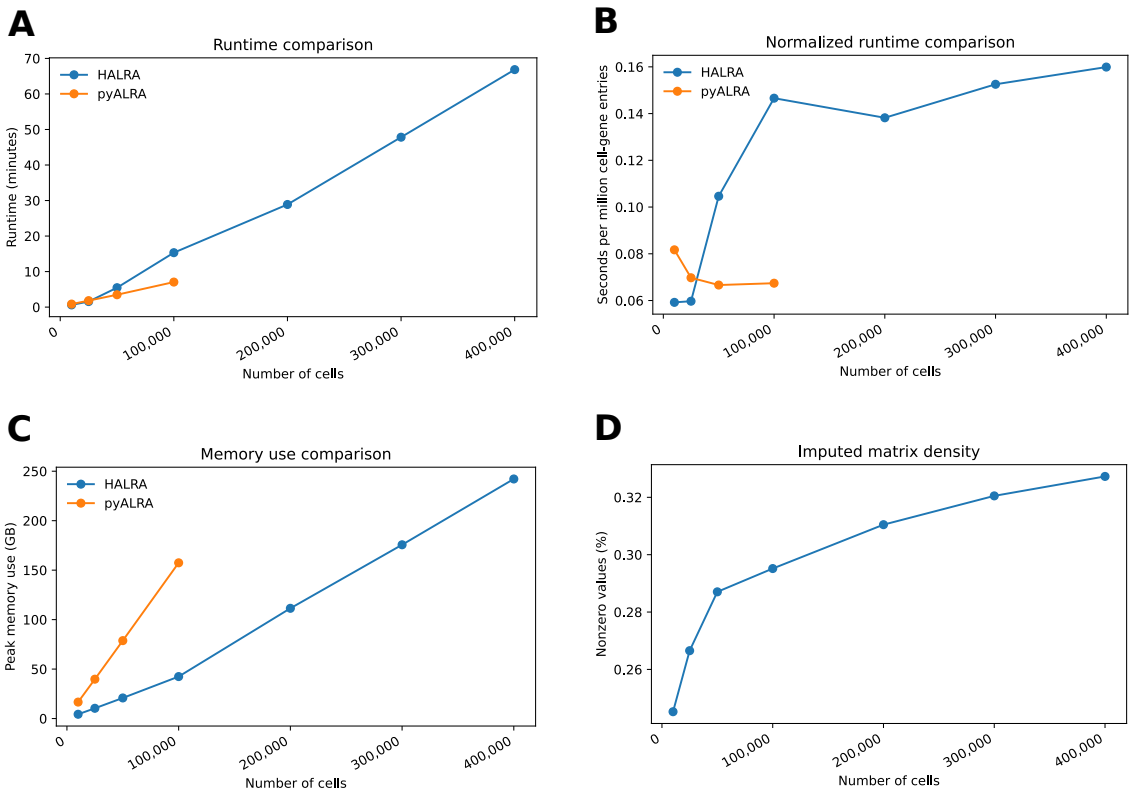



**Figure 3.2:** Computational benchmarking of HALRA and pyALRA across increasing scRNA-seq dataset sizes. (A) Total runtime comparison between HALRA and pyALRA. (B) Runtime normalized by matrix size, reported as seconds per million cell-gene entries. (C) Peak RAM utilization during imputation, demonstrating substantially reduced memory requirements for HALRA relative to pyALRA. (D) Imputed matrix density following reconstruction and adaptive thresholding. HALRA and pyALRA produced effectively identical output matrix densities across all tested dataset sizes.

ALRA workflow, thresholding is applied across the full reconstructed matrix, which can dramatically increase memory consumption when the reconstruction becomes substantially denser than the original count matrix. In HALRA, thresholding was instead performed using sparse-aware masking operations and blockwise processing approaches that eliminated many unnecessary dense allocations. Furthermore, intermediate reconstruction matrices were aggressively deleted or converted back into sparse formats immediately following thresholding and scaling steps. These changes substantially reduced peak memory utilization while preserving the overall imputation

behavior of the original algorithm. In practice, these sparse-preserving strategies proved essential for enabling analyses beyond the scale typically feasible under the original implementation.

The impact of these sparse-preserving modifications became increasingly pronounced as dataset size increased (see 3 for full statistics). For smaller datasets, runtime differences between the original ALRA implementation and HALRA were relatively modest, consistent with previous observations that sparse matrix operations can introduce additional indexing overhead for matrices that easily fit into memory [59]. However, for larger matrices, the sparse-preserving implementation enabled analyses that would otherwise exceed available RAM under the original workflow (Fig. 3.2). In particular, by avoiding fully dense reconstruction whenever possible, HALRA maintained manageable memory utilization even when imputed matrices became substantially denser following low-rank reconstruction. This observation is especially important given that imputation methods inherently increase matrix density, often converting matrices with 5–15% nonzero entries into matrices with substantially larger nonzero fractions, often ranging from 25–50% [168, 226]. Consequently, preserving sparsity during intermediate computations becomes critical for scalability.

These findings suggest that memory-efficient reformulation may be as important as algorithmic innovation itself for modern scRNA-seq analysis. While many existing imputation algorithms emphasize statistical performance or biological signal recovery, comparatively little attention has been paid to the computational consequences of dense intermediate representations at atlas scale. Our results demonstrate that sparse-preserving implementations can substantially extend the practical scalability of existing algorithms without fundamentally altering their mathematical framework. More broadly, these observations support the idea that future single-cell computational methods will increasingly require co-development of statistical methodology and high-performance computational design in order to remain feasible for rapidly growing transcriptomic datasets.

### 3.2.2 Benchmarking and Validation

Benchmarking experiments were performed in order to evaluate the computational scalability, memory utilization, and reconstruction behavior of HALRA relative to the original pyALRA implementation. Across all tested matrix sizes, HALRA consistently required substantially less peak memory than pyALRA while preserving comparable imputation behavior (Fig. 3.2). In particular, the sparse-preserving implementation reduced peak RAM utilization by approximately four-fold relative to pyALRA during large single-node reconstruction experiments. This reduction became increasingly pronounced as matrix size increased, reflecting the growing impact of dense intermediate reconstruction within the original ALRA workflow. Although HALRA generally required approximately two-fold longer runtime relative to pyALRA for equivalent datasets, this increase was offset by substantially improved memory scalability and the ability to complete analyses that would otherwise approach or exceed available system memory. These findings are consistent with previous observations that sparse matrix operations may introduce additional indexing overhead relative to dense linear algebra for moderate-sized datasets [59], but become increasingly advantageous as matrix dimensionality and sparsity increase.

#### Imputed Matrix Density

Importantly, despite substantial differences in internal computational implementation, HALRA and pyALRA produced effectively identical reconstructed matrix densities across tested datasets. The number of nonzero entries following imputation increased substantially relative to the original sparse count matrices for both implementations, consistent with the behavior of low-rank reconstruction methods [168]. Furthermore, output matrix densities increased progressively as the number of cells increased, although the rate of increase gradually plateaued at larger dataset sizes. This behavior likely reflects stabilization of the inferred low-rank transcriptional structure as increasingly large cell populations more fully sample underlying biological

programs. The near-identical output densities observed between HALRA and pyALRA indicate that sparse-preserving modifications did not substantially alter the reconstruction behavior of the original algorithm despite major differences in memory management and sparse matrix handling.

### Biological Validation

Because HALRA preserves the original ALRA reconstruction framework, biological validation was evaluated primarily through agreement with the original ALRA methodology rather than through development of a novel imputation strategy. The original ALRA study demonstrated preservation of biologically meaningful transcriptional structure, improved recovery of lowly expressed genes, and maintenance of putative biological zeros following adaptive thresholding [168]. HALRA preserved the same normalization, randomized SVD, adaptive thresholding, and scaling framework while modifying only the computational implementation. Randomized SVD procedures additionally demonstrated reproducible reconstruction behavior across repeated runs using fixed random seeds, consistent with previous evaluations of randomized low-rank approximation methods [107].

### Distributed Benchmarking

Distributed benchmarking experiments demonstrated substantial scalability improvements for HALRA across increasing MPI rank counts. Total runtime decreased approximately linearly with increasing numbers of MPI ranks for both medium- and large-scale datasets before gradually plateauing at higher rank counts (Fig. 3.3 A). This plateauing behavior became increasingly delayed as dataset size increased, indicating that larger matrices continued to benefit from additional distributed parallelism before communication and I/O overhead began to dominate runtime. In particular, the 500,000-cell and 1,000,000-cell datasets continued to exhibit

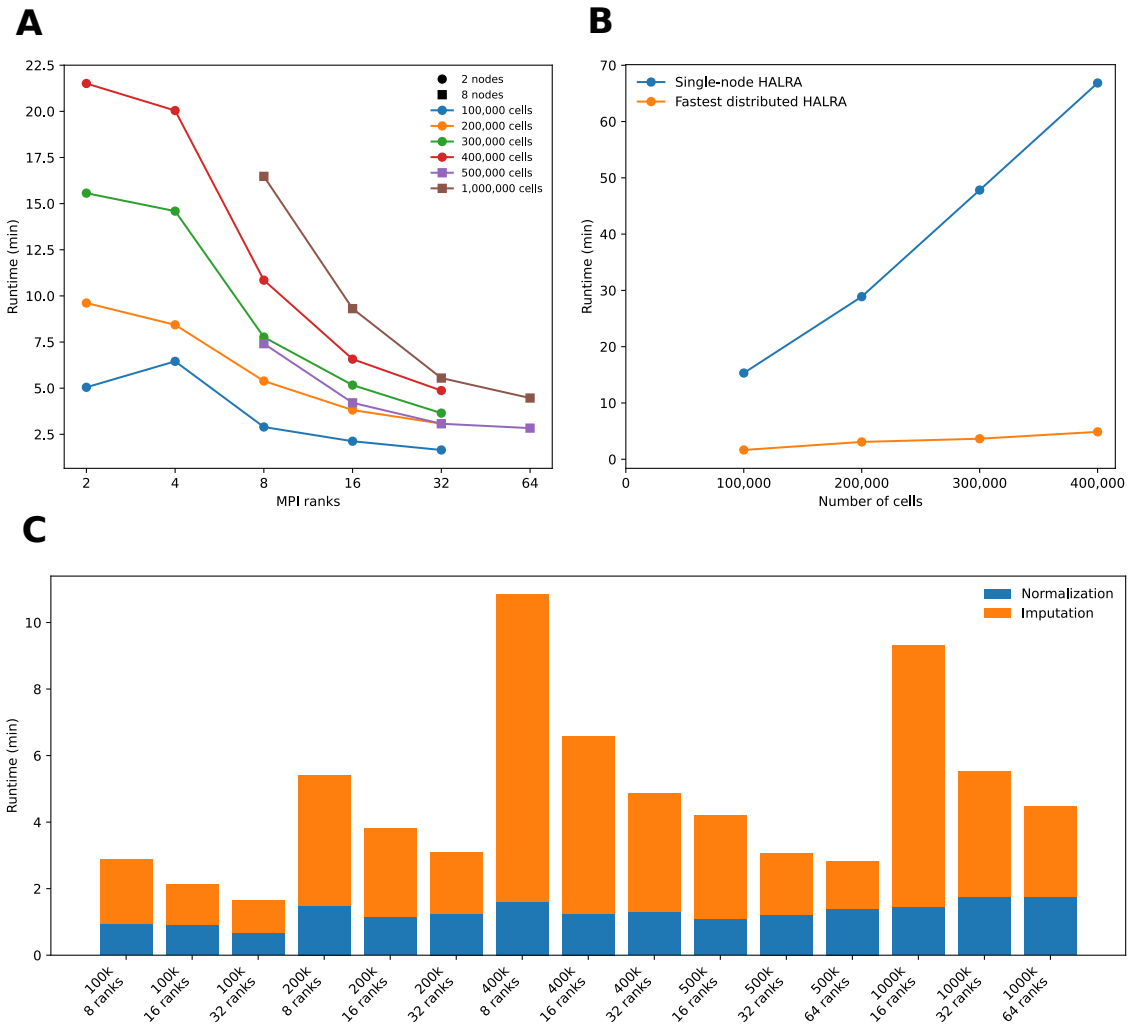



**Figure 3.3:** Distributed scaling behavior of HALRA across increasing scRNA-seq dataset sizes. (A) Total distributed HALRA runtime as a function of MPI rank count, demonstrating near-linear scaling followed by plateauing behavior at larger numbers of ranks as communication and I/O overhead increasingly dominate runtime. Circular markers denote runs performed on 2 Andes nodes, while square markers denote runs performed on 8 nodes. (B) Runtime normalized by matrix size comparing single-node HALRA with the fastest distributed HALRA configuration for each dataset size, demonstrating improved scaling efficiency for distributed execution at larger scales. (C) Distributed runtime decomposition into normalization and imputation stages across increasing MPI rank counts. Imputation runtime decreased substantially with increasing parallelism, whereas normalization exhibited comparatively limited scaling due to lower computational complexity and increasing filesystem overhead.

substantial runtime reductions at MPI rank counts where smaller datasets had already approached diminishing returns.

Comparison of single-node and distributed HALRA runtimes normalized by matrix size further demonstrated that distributed execution became increasingly computationally advantageous as dataset size increased (Fig. 3.3 B). Although distributed execution introduces additional communication and synchronization overhead, these costs were increasingly offset by reductions in per-rank computational workload and memory pressure at larger scales. Distributed execution therefore became progressively more efficient for larger matrices despite the additional overhead associated with MPI communication and distributed file operations.

Runtime decomposition analysis additionally revealed that reconstruction and imputation stages benefited substantially more from increasing MPI parallelism than normalization stages (Fig. 3.3 C). Imputation runtime decreased consistently with increasing MPI ranks due to the computational intensity of low-rank reconstruction and adaptive thresholding. In contrast, normalization exhibited comparatively modest scaling behavior, likely reflecting its lower computational complexity and greater sensitivity to filesystem overhead. Importantly, distributed and single-node implementations produced equivalent reconstruction densities and comparable reconstruction behavior across all tested datasets, indicating that distributed-memory execution preserved the underlying ALRA reconstruction framework while substantially extending computational scalability.

### 3.2.3 Distributed Sparse Reconstruction Architecture

The distributed HALRA implementation was designed to address the substantial memory and computational limitations associated with atlas-scale transcriptomic reconstruction. Matrix workloads were partitioned across MPI ranks using row-wise decomposition, allowing individual ranks to operate on contiguous subsets of cells while preserving efficient sparse CSR matrix access patterns. Row-wise partitioning

was selected because normalization and preprocessing operations are naturally cell-oriented and therefore require comparatively little inter-rank communication. This decomposition strategy additionally distributed memory utilization approximately evenly across MPI ranks, substantially reducing per-node memory requirements relative to single-node reconstruction workflows.

Distributed low-rank reconstruction presented substantially greater computational challenges due to the global nature of matrix factorization. Randomized distributed SVD was implemented using row-partitioned matrix multiplication combined with collective orthogonalization procedures based on tall-skinny QR (TSQR) factorization [63]. During randomized power iterations, individual MPI ranks independently computed local matrix products before participating in collective reduction and synchronization operations required to generate globally consistent low-rank approximations. Although row-wise decomposition simplified sparse matrix loading and preprocessing, distributed SVD remained communication-intensive due to repeated orthogonalization and singular vector aggregation steps. The plateauing runtime behavior observed at larger MPI rank counts (Fig. 3.3 A) likely reflects the growing contribution of communication overhead and synchronization costs during these collective operations.

Sparse matrix representations were preserved throughout distributed reconstruction wherever possible. Distributed normalized matrices were stored as CSR-encoded HDF5 layers, enabling efficient row-block access during reconstruction and thresholding operations. Following low-rank reconstruction and adaptive thresholding, sparse output matrices were collectively written into AnnData-compatible `.h5ad` files using MPI-enabled h5py and parallel HDF5 functionality [54, 98]. Parallel file writing allowed multiple MPI ranks to concurrently serialize sparse matrix blocks while preserving compatibility with existing Python single-cell analysis workflows. The comparatively limited scaling observed during normalization relative to reconstruction (Fig. 3.3 C) likely reflects the increasing contribution of distributed filesystem overhead during large-scale parallel reads and writes. In addition, reconstructed

matrices became progressively denser following imputation despite sparse-preserving reconstruction strategies, imposing continued scaling constraints at extreme atlas-scale sizes. Nevertheless, these results demonstrate that sparse-preserving distributed-memory reconstruction substantially extends the practical feasibility of ALRA-based imputation for large-scale scRNA-seq datasets.

### 3.2.4 Generalizable High-Performance Design Principles

Although this work focuses specifically on a scalable implementation of ALRA, many of the computational strategies developed in HALRA are broadly applicable to other large-scale single-cell analysis methods. In particular, our results suggest that sparse-preserving computation, distributed-memory parallelism, and parallelized I/O workflows represent increasingly important design principles for modern transcriptomic algorithms operating at atlas scale, or even for larger, non-atlas-scale dataset sizes. As scRNA-seq datasets continue to grow toward millions and tens of millions of cells through large consortium efforts such as the Human Cell Atlas [232], methods originally developed for comparatively small datasets will likely require substantial architectural redesign in order to remain computationally feasible.

One of the most important observations from this study is that sparse-first algorithm design can dramatically extend the practical scalability of transcriptomic workflows. Many single-cell methods conceptually operate on sparse count matrices but nevertheless generate dense intermediate representations during normalization, reconstruction, graph construction, or downstream statistical modeling. While these dense operations may be computationally acceptable for smaller datasets, they become prohibitive at modern dataset scales due to rapidly increasing memory requirements. Preserving sparsity wherever mathematically possible therefore represents a critical strategy for reducing RAM utilization and improving scalability. Importantly, this principle extends beyond imputation algorithms and applies broadly across single-cell analysis pipelines, including normalization, dimensionality

reduction, clustering, and spatial transcriptomics workflows [179, 259]. Our results demonstrate that reformulating algorithms around sparse matrix semantics can substantially improve scalability even without fundamentally altering the underlying statistical methodology.

Distributed decomposition strategies similarly represent an approach that is more broadly applicable for scaling matrix-based transcriptomic methods. Many commonly used single-cell algorithms rely heavily on matrix factorization techniques such as principal component analysis (PCA), nonnegative matrix factorization (NMF), singular value decomposition (SVD), or related low-rank approximations [265, 78]. These operations frequently dominate both runtime and memory utilization for large datasets. In HALRA, distributed randomized SVD allows large matrices to be partitioned across MPI ranks while avoiding centralized storage of dense intermediate matrices. Similar approaches could be extended to numerous matrix factorization-based workflows, particularly those involving dimensionality reduction, latent factor inference, denoising, or multimodal integration. Because randomized low-rank methods are already widely used in large-scale scientific computing [107], adapting these approaches to distributed transcriptomic analysis represents a natural extension of existing HPC methods.

Parallel I/O has also emerged as an increasingly important consideration for large-scale transcriptomic workflows. As dataset sizes approach hundreds of gigabytes or terabytes, file access itself can become a substantial computational bottleneck independent of algorithmic complexity. Traditional single-process read and write operations may introduce significant overhead during preprocessing, normalization, or intermediate matrix serialization. In HALRA, distributed HDF5-based workflows enabled multiple MPI ranks to concurrently read and write matrix blocks while preserving compatibility with AnnData-based scRNA-seq ecosystems [272]. More broadly, these findings suggest that scalable transcriptomic analysis increasingly requires co-optimization of both computation and storage infrastructure. Similar

distributed I/O strategies are likely to become increasingly important for other large-scale omics workflows involving sparse matrices, spatial transcriptomics, multimodal datasets, or large graph structures.

The sparse-preserving and distributed-memory strategies developed in this work may also be applicable to several classes of non-factorization-based imputation methods. Diffusion-based algorithms such as MAGIC [65] and graph propagation approaches frequently rely on large neighborhood graphs or affinity matrices that can themselves become computational bottlenecks for large datasets. Although these methods are often formulated in terms of dense diffusion operators, many graph representations remain intrinsically sparse and may benefit substantially from sparse-aware computation and distributed graph partitioning strategies. Similarly, iterative neighborhood smoothing operations may be amenable to blockwise or distributed implementations analogous to those developed in HALRA. These observations suggest that many computational limitations currently associated with atlas-scale imputation may arise not only from the underlying statistical methodology, but also from implementation-level assumptions regarding matrix density and memory locality.

At the same time, extending sparse-preserving HPC strategies to graph neural networks and deep learning-based transcriptomic methods presents additional challenges. Deep learning approaches such as variational autoencoders and graph neural networks frequently require repeated dense tensor operations, minibatch synchronization, and GPU-oriented memory layouts that differ substantially from traditional sparse linear algebra workflows [76, 174]. Furthermore, graph-based neural architectures often involve irregular communication patterns and dynamically constructed neighborhood relationships that may complicate efficient distributed-memory execution. While modern machine learning frameworks increasingly support distributed training, adapting these methods to extremely large sparse biological datasets remains an active area of research. Consequently, although sparse-preserving distributed design principles remain broadly valuable, additional methodological and

systems-level advances will likely be necessary to efficiently scale graph-based and deep learning-based single-cell methods to future atlas-scale datasets.

Collectively, these findings suggest that computational scalability is becoming a central methodological consideration in modern single-cell analysis. As transcriptomic datasets continue to expand in both size and complexity, future algorithm development will likely require increasingly close integration between statistical modeling, sparse numerical methods, distributed-memory computing, and scalable storage architectures. In this context, HALRA represents not only a scalable implementation of ALRA, but also a case study demonstrating how high-performance computational design principles can substantially extend the practical feasibility of existing single-cell analysis methodologies.

## 3.3 Methods

### 3.3.1 The ALRA Algorithm

HALRA was developed as a scalable implementation of Adaptively-thresholded Low Rank Approximation (ALRA), a matrix factorization-based imputation algorithm for scRNA-seq data [168]. Briefly, ALRA operates by first library-size normalizing and log-transforming the input count matrix, followed by computation of a low-rank approximation using truncated singular value decomposition (SVD). The resulting low-rank reconstruction is then adaptively thresholded on a gene-specific basis in order to distinguish likely technical zeros from biological zeros. Finally, scaling adjustments are applied to preserve gene-wise variance characteristics following reconstruction and thresholding (Fig. 3.4).

Low-rank approximation was performed using randomized singular value decomposition approaches based on implementations provided by scikit-learn and NumPy [216, 111, 107]. Randomized SVD methods were selected due to their computational efficiency for large sparse matrices relative to exact decomposition approaches. The

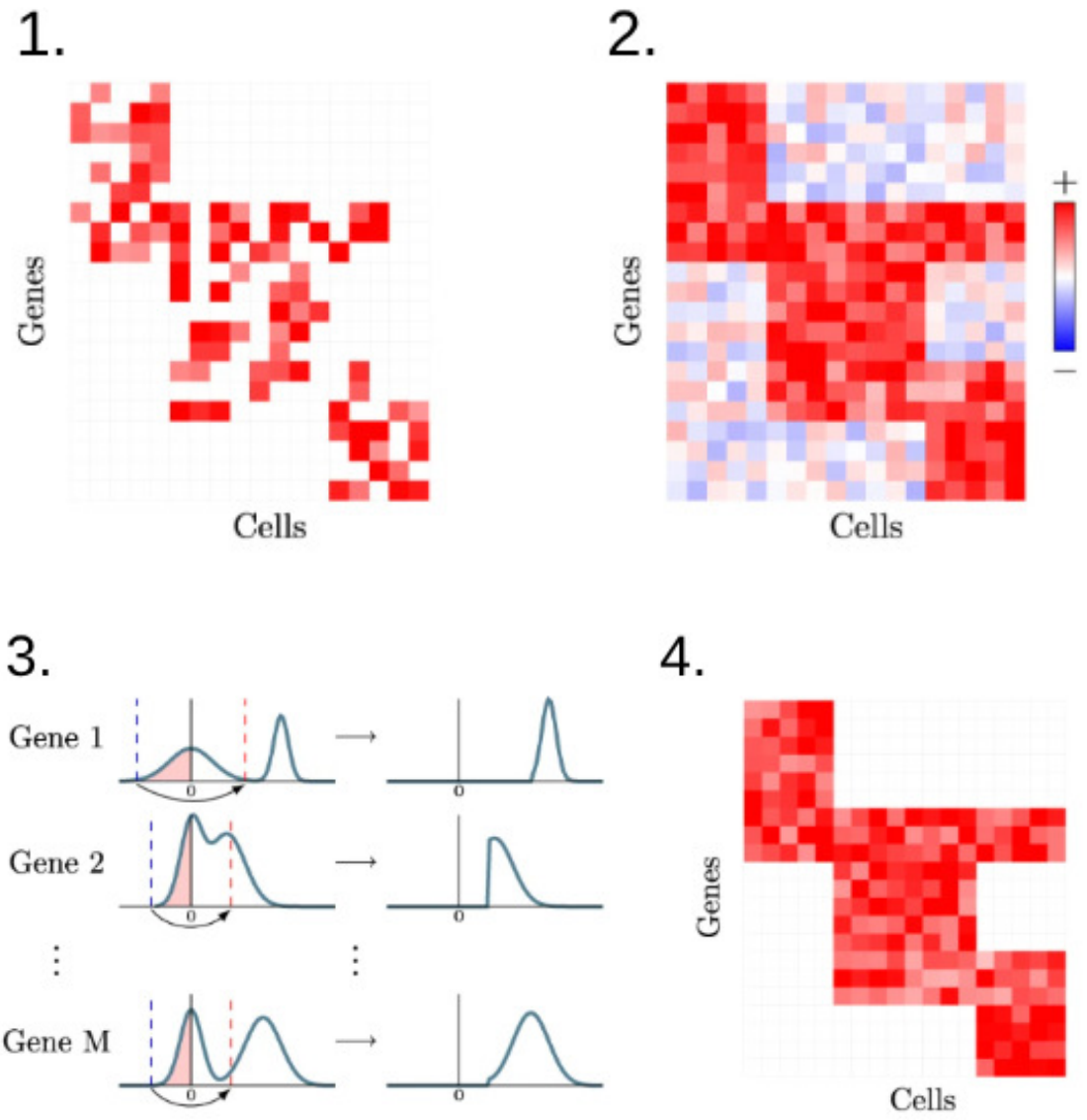



**Figure 3.4:** Overview of the ALRA imputation workflow. (1) The observed scRNA-seq expression matrix contains both technical zeros arising from dropout events and true biological zeros corresponding to absent gene expression. (2) Randomized singular value decomposition (RSVD)-based low-rank approximation reconstructs the matrix and imputes both technical and biological zeros. (3) Imputed biological zeros tend to exhibit small, approximately symmetric reconstructed values centered near zero. (4) Gene-wise adaptive thresholding removes these low-confidence values, restoring putative biological zeros while retaining imputed technical signal. Adapted from [168].

target matrix rank and number of randomized power iterations were user-configurable parameters. Adaptive thresholding was performed independently for each gene using quantile-based cutoffs following the original ALRA methodology [168]. Negative reconstructed values and low-confidence imputed values below gene-specific thresholds were set to zero in order to preserve putative biological absence of expression.

### 3.3.2 Sparse Matrix Implementation

HALRA was implemented in Python using SciPy sparse matrix representations [273]. Compressed sparse row (CSR) matrices were used for row-oriented operations including normalization, row partitioning, and distributed matrix loading, while compressed sparse column (CSC) matrices were used for column-oriented operations such as adaptive thresholding and gene-wise variance calculations. Matrix format conversions between CSR and CSC were performed selectively depending on the access patterns required by individual computational stages.

Sparse-preserving modifications were introduced throughout the reconstruction and thresholding pipeline in order to reduce peak memory utilization. Thresholding operations were implemented using sparse-aware masking procedures rather than explicit dense threshold matrices. Intermediate reconstruction matrices were converted back into sparse formats whenever possible following thresholding and scaling operations. In addition, memory-aware deletion of temporary matrices and blockwise processing strategies were used to minimize redundant allocations during reconstruction. These modifications allowed sparsity to be preserved throughout most stages of the pipeline while maintaining compatibility with the original ALRA reconstruction framework.

### 3.3.3 MPI-Based Distributed Implementation

Distributed-memory parallelization was implemented using MPI through the mpi4py Python package [57, 197]. Matrix operations were partitioned primarily across

rows (cells), allowing each MPI rank to operate on a contiguous subset of cells while minimizing inter-rank communication during normalization and reconstruction. Distributed normalization was performed by independently normalizing row blocks on each rank followed by parallel writing of normalized matrices to HDF5-based output files. Randomized distributed SVD was implemented using row-partitioned matrix multiplication and tall-skinny QR (TSQR)-based orthogonalization procedures [63]. Singular vector aggregation and randomized power iterations required periodic collective communication operations using MPI reduction and broadcast primitives.

Distributed input/output operations were implemented using parallel HDF5 functionality through h5py compiled with MPI support [54, 98]. Matrix blocks were written collectively into AnnData-compatible `.h5ad` files while preserving sparse CSC and CSR matrix encodings. Intermediate normalized matrices were stored as distributed CSR layers to facilitate efficient row-block access during subsequent reconstruction stages. Blockwise matrix writing was additionally used to reduce peak memory utilization during distributed reconstruction and serialization.

Distributed benchmarking experiments were performed on the Andes and Frontier supercomputing systems at the Oak Ridge Leadership Computing Facility (OLCF). Andes consists of dual-socket AMD EPYC CPU nodes connected through a Slingshot interconnect, while Frontier additionally provide GPU-accelerated architectures [211, 212]. The distributed HALRA implementation evaluated in this study primarily utilized CPU-based MPI parallelism on Andes. Experiments were performed using multiple node (1, 2, 4, 8) and MPI rank configurations (1, 2, 4, 8, 16, 32, 64) depending on dataset size, with workloads distributed approximately evenly across ranks. Memory utilization was monitored on a per-rank basis during normalization, decomposition, and reconstruction stages. Despite these distributed optimizations, several limitations remain, including communication overhead during randomized SVD iterations and increased I/O contention at very large scale.

## 3.4 Conclusion

In this study, we developed HALRA, a sparse-preserving and distributed-memory implementation of the ALRA imputation algorithm designed specifically for modern large-scale scRNA-seq datasets. By reformulating key stages of the reconstruction pipeline around sparse matrix operations and distributed computation, HALRA substantially reduced memory requirements relative to existing ALRA implementations while preserving comparable reconstruction behavior and biological fidelity. Benchmarking experiments demonstrated that these sparse-preserving modifications enabled analyses that would otherwise become impractical due to dense intermediate matrix reconstruction and rapidly increasing RAM utilization. Furthermore, distributed-memory parallelization through MPI allowed reconstruction workloads to scale efficiently across multiple compute nodes, substantially extending the feasible size of transcriptomic datasets that could be processed using ALRA-based imputation approaches. Collectively, these results demonstrate that high-performance computational redesign can significantly improve the practical scalability of existing single-cell analysis algorithms without fundamentally altering their underlying mathematical framework.

More broadly, this work highlights the growing importance of sparse-preserving computation, distributed-memory parallelism, and scalable I/O architectures for modern single-cell transcriptomic analysis. As scRNA-seq datasets continue to expand toward atlas and eventually organism-scale resolutions, algorithms originally developed for comparatively small datasets will increasingly require architectural redesign in order to remain computationally feasible. Although this work focused specifically on ALRA, many of the computational strategies developed in HALRA, including sparse-first reconstruction, distributed low-rank decomposition, blockwise matrix processing, and parallel HDF5-based workflows, are broadly applicable to other matrix factorization, graph-based, and diffusion-based transcriptomic methods. Consequently, HALRA represents not only a scalable implementation of ALRA, but

also a broader framework illustrating how high-performance computing principles can be integrated into future large-scale single-cell computational methodologies.

# Chapter 4

# Whole-Genome SARS-CoV-2 Lineage Classification and Multiplex Integration with U.S. Mortality and Vaccination Trends

## 4.1 Introduction

The COVID-19 pandemic caused substantial and ongoing mortality, with over seven million deaths reported globally as of June 2025 [287]. Throughout the pandemic, SARS-CoV-2 accumulated mutations across its genome, giving rise to more than 70 designated variants of concern (VOCs), so-named because they exhibited increased transmissibility, increased virulence, worsened public health outcomes, or combinations of these traits [46, 79]. In response, an unprecedented global sequencing effort has resulted in the upload of tens of millions of SARS-CoV-2 genomes into public repositories [118], accompanied by the development of multiple different lineage nomenclature and classification systems to organize and track viral evolution [158]. In parallel, a wide array of additional data types has been studied in the context of COVID-19, including data on clinical outcomes, immunological profiles, vaccination status, human mobility patterns, and numerous sociodemographic variables [161, 191, 283, 204, 185]. Integrating these diverse data modalities with SARS-CoV-2 genomic lineage classifications may offer critical insights into the biological, environmental, and societal drivers of this disease. However, the vast scale of available SARS-CoV-2 genomic data and the complexity of its lineage designations necessitate the use of robust, data-driven approaches and integrative analytical frameworks to extract meaningful patterns from analyses.

### 4.1.1 SARS-CoV-2 Genomes

Global efforts in data collection have resulted in nearly 21 million SARS-CoV-2 genome sequences uploaded to the Global Initiative on Sharing Avian Influenza Data (GISAID) database [73]. GISAID was established in 2008 to provide open access to genomic data of influenza viruses but was co-opted in light of the COVID-19 pandemic to store SARS-CoV-2 sequences and has since become the world's largest such repository. While GISAID processes all submissions, African, Asian, and South American countries generally sequence and deposit an order of magnitude fewer

COVID-19 positive samples compared to European, North American, and Oceanian countries [149]. Additionally, both testing rates and subsequent sequencing rates per positive test vary considerably from country to country and over time [29]. Almost all SARS-CoV-2 sequences in the GISAID database have associated metadata that provides sample collection location and date, as well as World Health Organization (WHO) variant classification (e.g., Alpha and Delta variants).

#### Lineage Classification

The most widely-used strategy for classifying genetic lineages of SARS-CoV-2 uses the Phylogenetic Assignment of Named Global Outbreak (PANGO) lineages nomenclature and PANGO lineage assignment tool, pangolin [227, 209]. However, PANGO lineages are manually curated classifications that rely on a relatively limited set of mutations located primarily within the spike protein and may therefore overlook other mutations across the SARS-CoV-2 genome that influence infectivity and mortality [210, 278]. Further, due to the massive scale of available genomic data and limitations in computational resources, many phylogenetic analyses rely on subsampling strategies that examine only a small fraction of available sequences, which can introduce sampling bias and substantially impact the inferred evolutionary relationships and lineage classifications [43]. Despite this, PANGO remains the dominant classification strategy, suggesting a need to develop a flexible model that uses the set of all available SARS-CoV-2 sequences to derive whole-genome classifications in a data-driven approach.

### 4.1.2 Large-scale Viral Genome Comparisons

Comparing viral genome sequences at scale presents both computational and methodological challenges. Traditional phylogenetic approaches often rely on multiple sequence alignment followed by tree-based inference methods, which are computationally intensive and increasingly infeasible as datasets grow into the

millions of sequences [201]. To manage this, many analyses resort to subsampling strategies that risk introducing bias and may obscure rare or emerging patterns [43]. Alternative strategies, such as alignment-free methods, have gained traction for their scalability and robustness to sequencing errors. These approaches typically rely on comparing fixed-length sequence features, such as k-mers, to estimate similarity between genomes without requiring full alignment. However, even alignment-free methods can struggle with the scale and heterogeneity of data generated during a global pandemic [300]. As such, novel computational frameworks capable of performing exhaustive comparisons between millions of full-length viral genomes are essential to enable unbiased discovery of genomic structure and evolution.

### Combinatorial Metrics

The Combinatorial Metrics (CoMet) codebase [137] is an application for comparing vectors on large-scale parallel accelerated computing systems such as the Frontier Supercomputer at Oak Ridge National Laboratory. CoMet requires that input vectors (i.e., SARS-CoV-2 genome sequences) be encoded in binary format, but there is no standard for the binarization process; each dataset and application requires its own formulation based on the goals of the vector comparison and downstream analyses. Moreover, the CoMet application supports a number of binary similarity metrics (e.g., Duo, CCC) [50, 51], and the choice of metric and metric parameterization, in tandem with the vector binarization scheme, are important considerations. Previous studies have used the CoMet application to conduct exhaustive vector comparisons for genomic and environmental applications [140, 138, 139, 163]. In particular, CoMet has been optimized to run efficiently on the Oak Ridge Leadership Computing Facility's (OLCF) leadership-class systems, including the (now defunct) Summit supercomputer and the OLCF's newer, exascale supercomputer Frontier, as well as the JUWELS Booster at Jülich Supercomputing Centre and Perlmutter at the National Energy Research Scientific Computing Center (NERSC). All of these implementations leverage ultra-low precision, GPU-accelerated computing

and distributed task parallelism to exhaustively compute similarity metrics for pairs or triplets of vectors at extreme scales with record-breaking (exascale) speeds and efficiency, enabling full-length, direct comparisons between millions of vectors. These capabilities make CoMet particularly well-suited for large-scale comparative analyses of viral genomes, such as those required to uncover population-level structures and evolutionary patterns in global datasets of millions of SARS-CoV-2 genome sequences.

#### High-Performance Markov Clustering

Markov clustering (MCL) is an unsupervised algorithm designed to identify dense, highly connected regions within networks based on stochastic flow simulation [271]. MCL iterates between two matrix operations: expansion and inflation. The expansion step simulates longer random walks and effectively spreads flow across the similarity network, while the inflation step favors more probably walks over less probable ones, effectively contracting flow. These operations are oscillated between, increasing edge weights between nodes of the same cluster and decreasing edge weights to nodes of different clusters, until equilibrium is reached (Fig. 4.1). While MCL is already fast and efficient, high-performance Markov clustering (HipMCL) is an even more memory-efficient and scalable implementation designed to run on large, distributed computing platforms [10]. These qualities enable the partitioning of networks composed of billions of nodes and edges, such as those derived from comparing millions of SARS-CoV-2 genomes.

### 4.1.3 Study Objectives

The aims of this chapter were twofold. First, we sought to develop a novel, data-driven lineage classification framework for SARS-CoV-2 based on whole-genome similarity relationships derived from the entirety of publicly available viral sequencing data. To accomplish this, millions of SARS-CoV-2 genome sequences were exhaustively compared pairwise using the CoMet application to construct a genome-to-genome

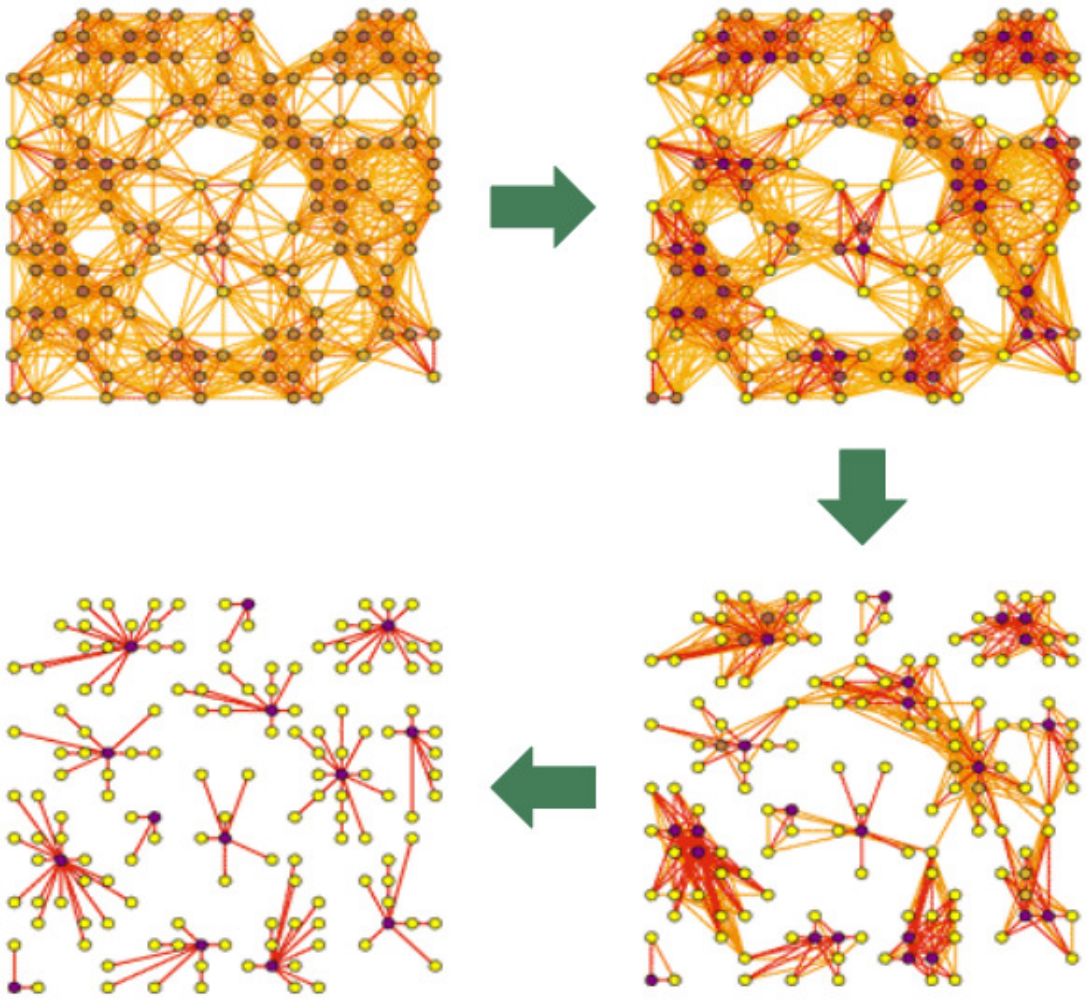


**Figure 4.1:** Visual representation of cluster creation during Markov Clustering (MCL). Cluster structure is detected by alternating inflation and expansion operations over the random walk probability matrix of the graph, resulting in the detection of densely connected subgraphs (i.e., clusters). Graphics are adapted from the MCL website [68].

similarity network, which was subsequently clustered using HipMCL to identify groups of genetically similar viral genomes [137, 10].

Second, we sought to integrate viral genomic structure with county-level epidemiological and demographic measurements through construction of a spatiotemporal multiplex network. Weekly county-level mortality rates and vaccination coverage

data from California were incorporated alongside genomic composition profiles derived from the lineage classification framework. Random walk with restart (RWR) analysis on the resulting cross-modal multiplex network was then used to investigate relationships between viral genomic substructure, mortality-associated diffusion patterns, and vaccination dynamics across space and time. Together, these analyses were intended to provide insight into how fine-scale viral genomic variation may relate to downstream epidemiological structure during the COVID-19 pandemic.

## 4.2 Results and Discussion

### 4.2.1 High-Resolution, Data-Driven Lineages

To generate a novel, data-driven SARS-CoV-2 lineage classification scheme, we created a genome-to-genome similarity network from the 11 million SARS-CoV-2 genome sequences on GISAID available from January 1, 2020 to June 2, 2022 and identified 22,040 clusters of mutationally similar genomes from the network in an initial publication [196]. We achieved this in three main steps: i) creation and application of a binarization scheme to the SARS-CoV-2 genome sequences, ii) generation of a network with genomes as nodes and their Duo similarity scores as edges by applying the CoMet codebase, and iii) use of HipMCL to extract clusters of sequentially similar genomes from the network topology.

To compare our data-driven lineage classification to existing schemes, we mapped each genome to its metadata-assigned PANGO lineage and corresponding WHO variant (which is derived from PANGO lineages). Although our clusters spanned 1,712 distinct PANGO lineages, all clusters containing at least 100 sequences were composed of genomes sharing $\geq$99% identical PANGO lineage annotations. This indicates that our method not only captures PANGO classifications but also refines them, introducing approximately 20,000 additional subclassifications that represent

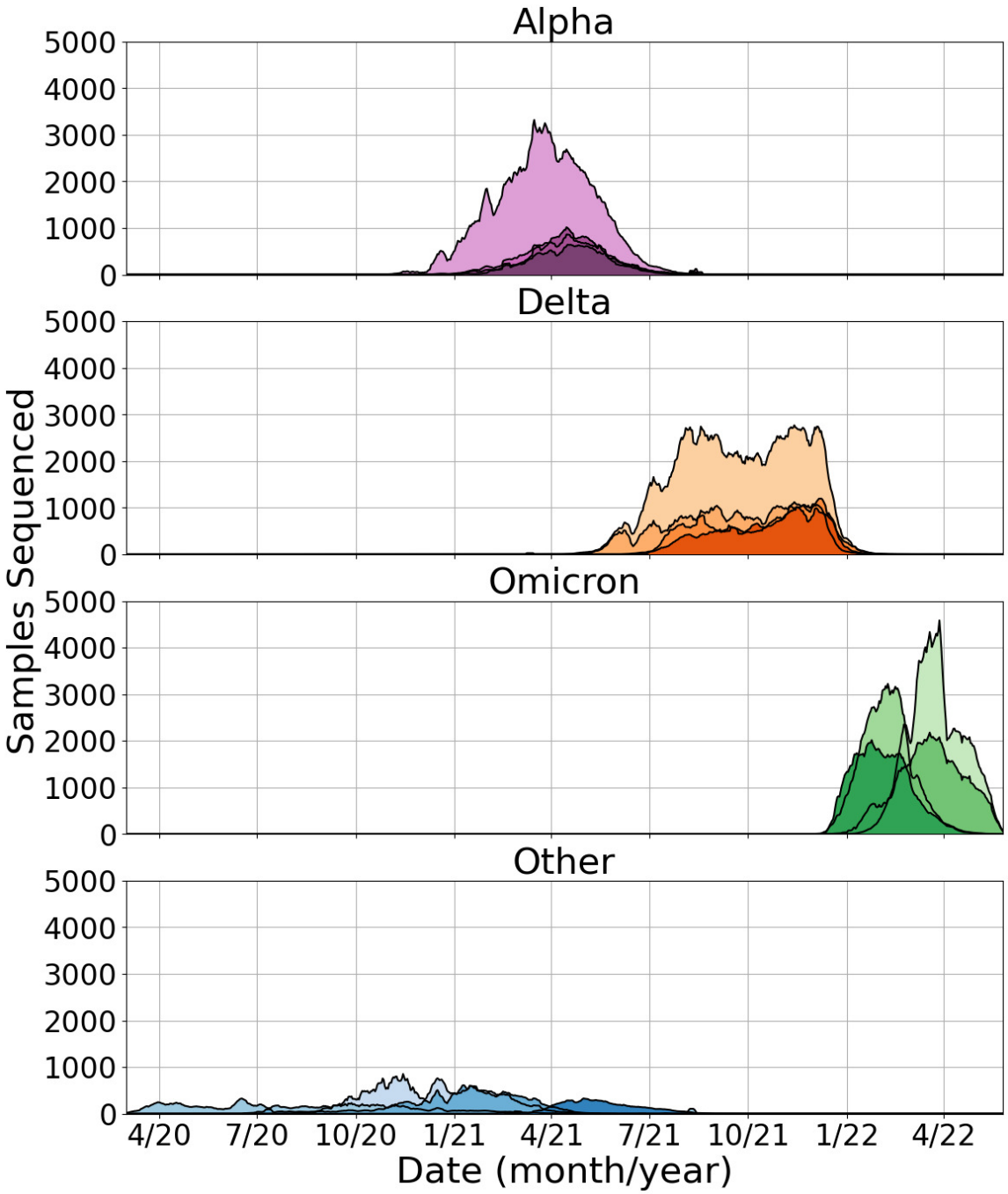



**Figure 4.2:** Longitudinal view of the four most sequenced SARS-CoV-2 WHO variants during the first ∼2.5 years of the COVID-19 pandemic, separated by variant. Each WHO variant is colored with four sub-hues, with each hue corresponding to one of the four largest CoMet + HipMCL clusters (i.e., SARS-CoV-2 lineages) for that WHO variant.

finer-grained genomic distinctions based on mutations from the entirety of the SARS-CoV-2 genome. Importantly, because there are only a handful of WHO variants but over 22,000 genome clusters, many distinct clusters were assigned to the same WHO variant: 20,856 clusters mapped to the four largest WHO variants (2,022 to Alpha, 11,301 to Delta, 2,396 to Omicron, and 5,137 to Other), while fewer than 5% of sequences (1,184 clusters) corresponded to a WHO variant other than those four. However, we retained clusters as separate groups in our analyses, as each cluster represented a unique set of mutations despite sharing the same WHO variant designation.

Using these methods, we identified high-resolution sub-groupings of WHO variants and subsequently tracked the evolution and spread of these subgroups over time and space. A longitudinal view of the four WHO variants with the most uploaded sequences (Fig. 4.2) and their four largest sub-lineages (i.e., CoMet + HipMCL clusters) demonstrates that while each WHO variant generally appears and disappears as a group, its sub-lineages may follow more complex epidemiological patterns. Note that the time between sample collection and availability on GISAID varies from days to several weeks, by which we observe a sharp decrease in the number of sampled sequences at the end of the time series (i.e., when the GISAID data was accessed). This decrease therefore does not indicate a drop in the number of genomes sequenced, but rather that the sequences had not all been uploaded to GISAID. Moreover, the vertical axes in Fig. 4.2 represent the number of samples sequenced per day, not the number of cases per day.

### Limitations

While the scale of the available SARS-CoV-2 sequences on GISAID is unprecedented, the available data come with several caveats. Metadata is available for every uploaded sequence, but it is often incomplete. For example, of the 11M sequences, 7M mark gender as unknown, with the rest composed of slightly more women than men (2.13M women and 1.98M men). Other notable categories with significant amounts of missing

or unknown data are patient age, geographical origin of the sample more detailed than country, and sample collection date more specific than month. As a result of this missing metadata, any sequence analyses requiring associated metadata (e.g. sample collection date and location for spatiotemporal analyses) must necessarily exclude these sequences, reducing the number of overall usable sequences.

Another major consideration when interpreting results stems from the disparity between infection rates and sequencing rates throughout countries globally. These differences can arise for various reasons, such as income disparity (78% of high-income but only 42% of low-income countries sequenced $\geq 0.5\%$ of COVID-19 cases) [29] or unwillingness to share sequencing data (37% of countries submitted less than half of their sequences to public databases when sequencing variants of concern) [42]. Moreover, sequencing rates overall are vastly lower than infection rates, with only 6.8% of countries sequencing $\geq 5\%$ of their total confirmed cases [29]. Lastly, our current spatiotemporal visualizations do not account for population density. One should keep these geospatial and socio-economic sequencing disparities in mind when interpreting results involving comparisons such as differences between countries or populations.

### Spatial Asynchronicity

To investigate the relationship between lineages and geospatial proximity, we visualized how often two neighboring countries (i.e., countries that share a national border) sampled SARS-CoV-2 sequences that mapped to the same 16 WHO CoMet + HipMCL subclusters each day over the first $\sim$2.5 years of the pandemic. A snapshot of this spatiotemporal view during the first Delta surge (Fig. 4.3) indicates the recurring theme we observed: neighboring countries were often dominated by different subclusters of the same WHO variant, or different variants altogether. Such spatial asynchronicity in SARS-CoV-2 lineages suggests a possibly different pattern of epidemiological risk than was understood from conventional data. The full, animated view of these snapshots is available at https://github.com/pasc-2023-anonymous/

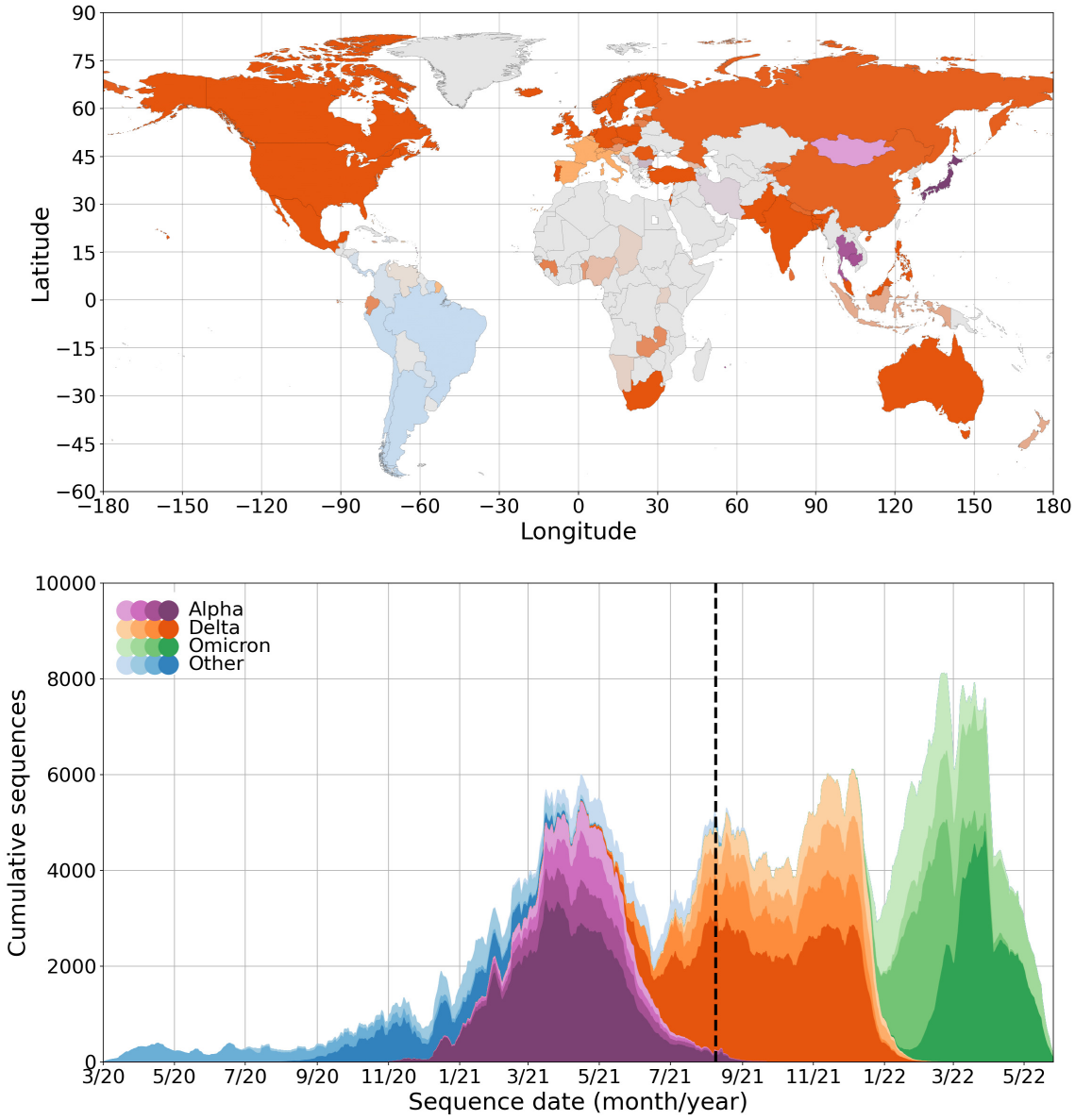



**Figure 4.3:** Snapshot of the distribution of CoMet + HipMCL clusters (i.e., SARS-CoV-2 lineages) on August 9th, 2021, during the first peak of the Delta variant.

pasc_2023_anonymous/blob/main/maps.gif. We plan on indexing border-sharing countries in order to calculate the exact disparity in geospatial proximity and dominant SARS-CoV-2 lineage designation.

### 4.2.2 Integrating Sociodemographic Data

To investigate relationships between viral genomic structure and downstream epidemiological outcomes, we integrated SARS-CoV-2 genomic composition, county-level COVID-19 mortality rates, and cumulative vaccination coverage into a unified spatiotemporal multiplex network spanning California from 2020-W15 through 2022-W06. RWR diffusion was then performed on the multiplex using genomic nodes associated with specific HipMCL clusters as seeds in order to identify mortality diffusion landscapes associated with distinct viral genomic structures.

Cluster-specific mortality diffusion maps revealed that different genomic clusters produced distinct spatial mortality topologies across California (Fig. 4.4). These diffusion landscapes emerged despite all diffusion processes operating on the same underlying multiplex framework, suggesting that differences in viral genomic composition and spatiotemporal transmission structure meaningfully altered the downstream topology of mortality-associated diffusion. Notably, the three largest genomic clusters by fraction of all sequences during the study period (Appendix Table 1) produced substantially different mortality diffusion patterns despite all representing large and geographically widespread transmission networks.

HipMCL cluster 0, a predominantly Delta-associated cluster comprising approximately 31% of all sequences, produced broad mortality diffusion across much of California, particularly throughout the Central Valley and southern portions of the state. This widespread diffusion pattern is consistent with the extensive geographic dissemination and high transmissibility associated with the Delta variant during late 2021 and early 2022 [199, 246]. In contrast, HipMCL cluster 2, which contained heterogeneous but highly internally consistent genomic structure without a dominant WHO variant designation, produced a substantially more localized mortality topology centered primarily around the San Francisco Bay Area and portions of Southern California. The reduced spatial extent of diffusion associated with cluster 2 may reflect either lower transmissibility, differing temporal persistence, or the aggregation

of multiple related but epidemiologically distinct genomic backgrounds within the cluster.

The Epsilon-associated HipMCL cluster 7 demonstrated a markedly different mortality landscape characterized by stronger diffusion into inland and southern California counties, including several counties that exhibited lower cumulative vaccination coverage relative to major coastal metropolitan regions at 2021-W37, which marks the tail end of the Epsilon variant's presence in California. This observation is particularly notable because vaccination uptake during the COVID-19 pandemic displayed strong geographic and sociodemographic heterogeneity throughout the United States [275, 252]. Counties with lower vaccination coverage often corresponded to regions with increased socioeconomic vulnerability, reduced healthcare access, and elevated COVID-19 mortality burden [150, 208]. Because diffusion within the multiplex simultaneously traversed genomic, mortality, vaccination, temporal, and interlayer edges, these mortality landscapes likely reflected not only viral genomic similarity, but also broader epidemiological and sociodemographic structure encoded within the network topology.

Importantly, mortality diffusion scores should not be interpreted as direct measurements of mortality or transmissibility. Rather, they represent the extent to which county-level mortality nodes were topologically connected within the multiplex network to genomic transmission patterns associated with a particular SARS-CoV-2 lineage. Consequently, high mortality diffusion may arise from combinations of factors including prolonged temporal persistence of a lineage, strong geographic spread, similarity in mortality trajectories between counties, and shared vaccination dynamics. The resulting diffusion spatial plots therefore capture integrated relationships between viral genomic structure and population-level epidemiological context rather than direct causal effects.

These analyses demonstrate that multiplex integration of genomic and sociodemographic data can reveal large-scale epidemiological structure associated with distinct

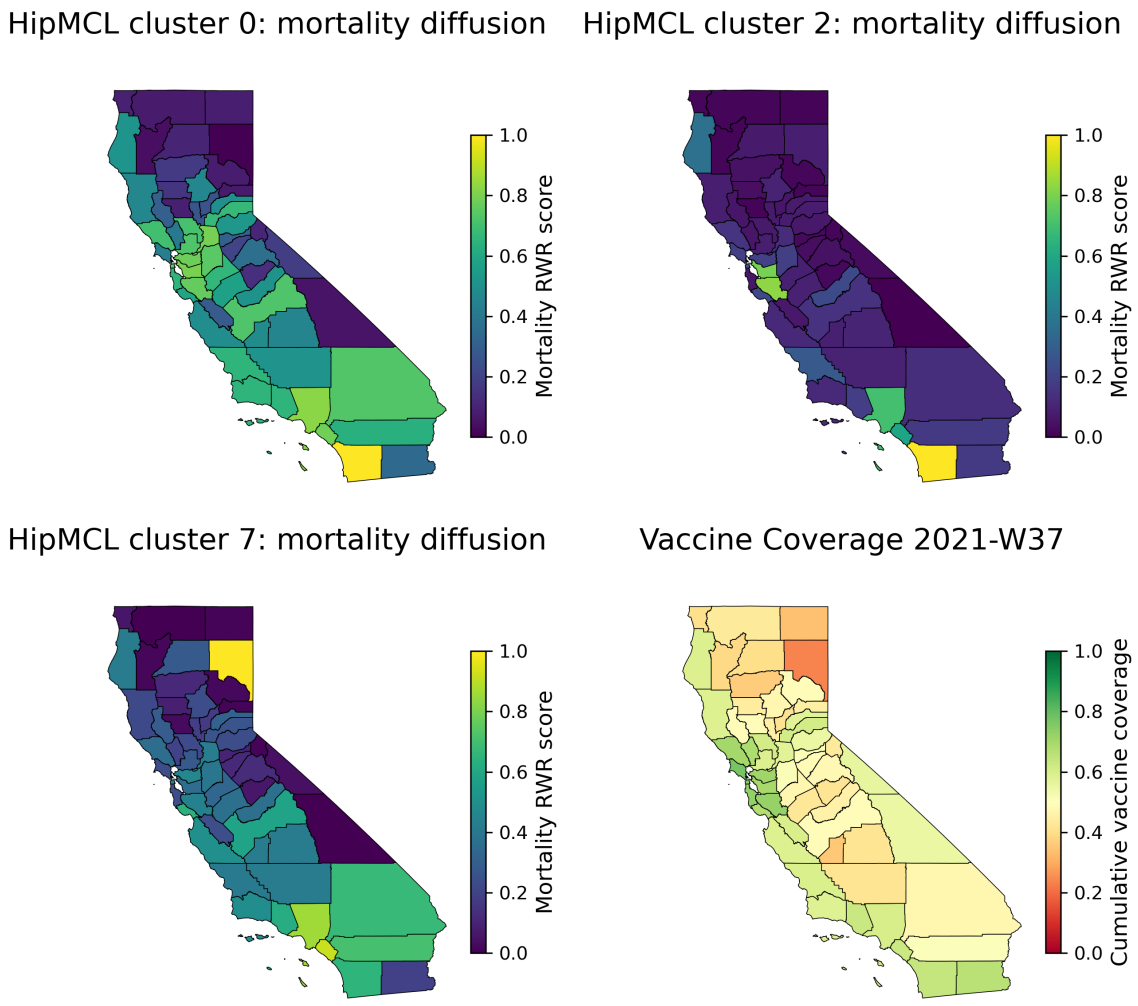



**Figure 4.4:** County-level mortality diffusion patterns associated with three of the largest SARS-CoV-2 HipMCL genomic clusters across California from 2020-W15 through 2022-W06, alongside statewide vaccine coverage at 2021-W37. Mortality diffusion maps were generated using RWR on the California multiplex network, seeded with genomic nodes corresponding to each HipMCL cluster and restricted to the weeks during which each cluster was observed. Panels show HipMCL cluster 0 (Delta-associated; 30.8% of sequences), cluster 2 (heterogeneous/non-assigned; 22.3%), and cluster 7 (Epsilon-associated; 7.6%). The bottom-right panel shows cumulative vaccine coverage by county at 2021-W37. Distinct genomic clusters produced different mortality diffusion landscapes, with the Epsilon-associated cluster showing stronger diffusion into inland and southern California counties that generally exhibited lower vaccine coverage than coastal metropolitan regions.

viral genomic backgrounds. While substantial prior work has examined SARS-CoV-2 genomic evolution, mortality dynamics, and vaccination trends individually, comparatively fewer studies have integrated these modalities within unified spatiotemporal analytical frameworks [228, 38]. However, the multiplex framework presented here enabled these modalities to be modeled simultaneously within a unified graph representation. More broadly, this approach illustrates how large-scale pathogen genomics can be integrated with spatial epidemiological data to investigate relationships between viral evolution, public health outcomes, and population-level sociodemographic structure during ongoing and future pandemics.

### 4.2.3 Null Model Validation of Multiplex Diffusion

To determine whether the observed mortality diffusion landscapes reflected meaningful multiplex organization rather than generic graph topology, we compared observed diffusion scores to those obtained from within-layer-preserving multiplex null models in which temporal, interlayer, and lagged genomic–mortality coupling edges were randomized while preserving intralayer similarity structure. Null model comparisons demonstrated that the observed multiplex network produced temporally coherent and spatially localized mortality diffusion enrichment patterns that substantially exceeded expectations under randomized cross-context coupling (Fig. 4.5).

County-week z-score heatmaps revealed that enrichment was concentrated within specific county trajectories during the Alpha-associated interval spanning 2021-W01 through 2021-W30 rather than uniformly distributed throughout the statewide network. Several counties, including Yolo, Yuba, Ventura, Solano, and Tulare counties, exhibited persistent positive mortality diffusion enrichment across contiguous temporal intervals. The temporal continuity of these enriched regions suggests that observed multiplex diffusion was not driven by random, isolated fluctuations or generic network connectivity alone, but rather by structured relationships encoded within the genomic, epidemiological, and temporal organization of the multiplex network.

Importantly, enrichment emerged near the beginning of the Alpha-associated seed interval and diminished following the end of the seeded period, indicating that the diffusion signal remained temporally aligned with the underlying genomic seeding structure.

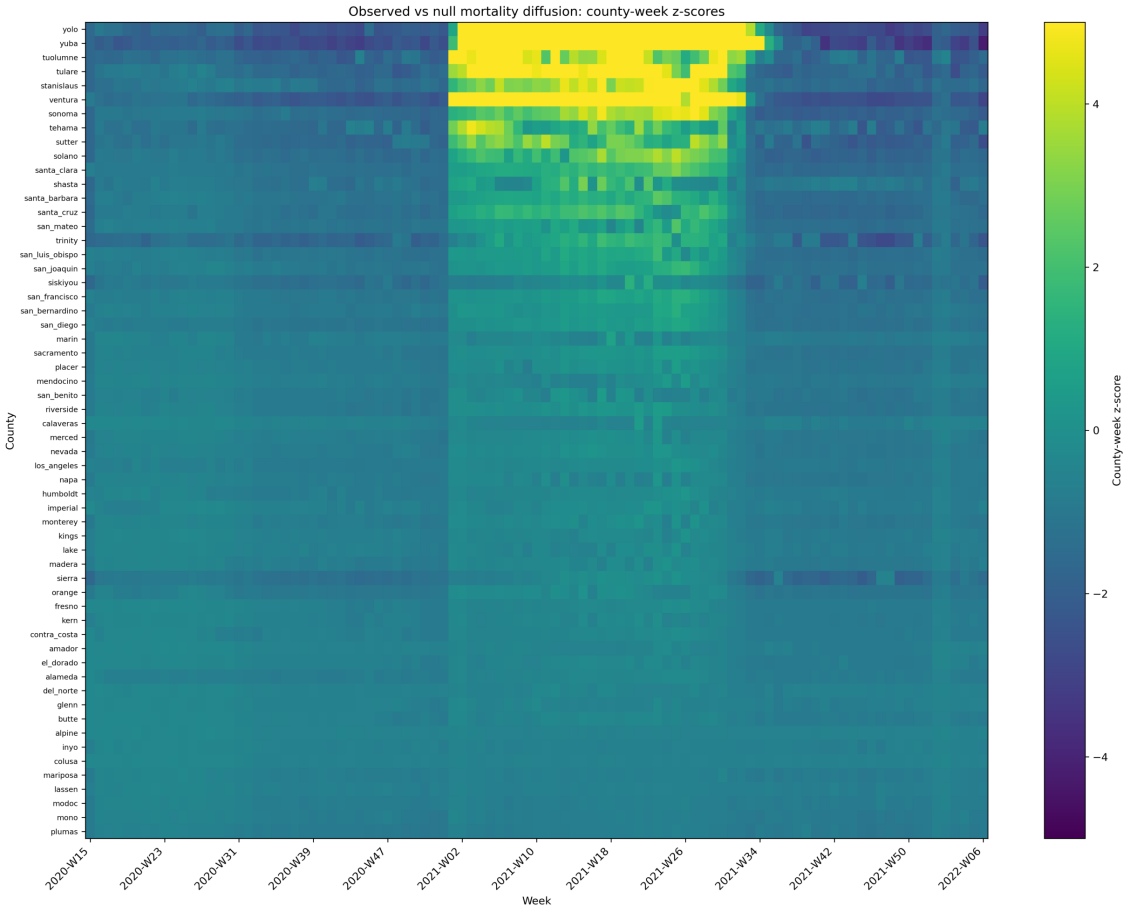



**Figure 4.5:** County-week heatmap of observed versus null mortality diffusion z-scores following Alpha-period genomic seeding from 2021-W01 through 2021-W30. Positive z-scores indicate county-weeks in which the observed multiplex structure concentrated mortality diffusion more strongly than expected under within-layer-preserving multiplex null models. Enrichment was spatially localized and temporally coherent, with persistent elevated diffusion observed across subsets of counties during the Alpha-associated interval.

These results are notable because the null framework preserved the underlying spatial similarity topology within the genomic, mortality, and vaccination layers.

Consequently, the observed enrichment cannot be attributed solely to broad county-level epidemiological similarity, population size effects, or general graph smoothness. Instead, the increased diffusion observed within specific county-week trajectories emerged only when the observed cross-context coupling structure between modalities and across time was retained. This suggests that the temporal alignment between viral genomic composition, mortality trajectories, and vaccination structure contributed meaningfully to the resulting mortality diffusion landscapes.

The observed spatial localization of enrichment is also consistent with the heterogeneous regional dynamics that characterized SARS-CoV-2 transmission throughout California and the broader United States during the Alpha wave [161, 21]. Prior studies demonstrated that COVID-19 burden varied substantially across counties due to differences in socioeconomic status, population density, mobility patterns, healthcare access, and vaccination uptake [150, 275]. Because diffusion within the multiplex simultaneously traversed genomic, mortality, vaccine, temporal, and lagged coupling edges, the resulting enrichment patterns likely reflected integrated epidemiological structure arising from combinations of viral transmission dynamics and population-level contextual factors.

More broadly, these analyses demonstrate that multiplex diffusion patterns associated with SARS-CoV-2 genomic structure were not reproducible under randomized coupling organization despite preservation of the underlying modality-specific topology. The null model framework therefore supports the interpretation that the observed mortality diffusion landscapes emerged from biologically and temporally coherent multiplex structure rather than from generic within-layer similarity relationships alone. These findings further illustrate the utility of multiplex-network-based diffusion analyses for investigating relationships between pathogen genomic variation and downstream epidemiological structure across space and time.

## 4.3 Methods

### 4.3.1 SARS-CoV-2 Genome Sequence Preprocessing

All sequences available on the GISAID database through June 2nd, 2022 were downloaded using GISAID's "high coverage" (< 1% unassigned nucleotides) and "collection date compl" (has a sample collection date) filters, resulting in an initial dataset of approximately 11 million SARS-CoV-2 sequences. The metadata for all of these sequences were also downloaded from GISAID and used to exclude sequences with non-human hosts, sequences with a collection date only containing the month or year of collection (as opposed to the day), and sequences with no sample collection location or sample collection location less specific than country. The remaining sequences were then aligned to the Wuhan reference SARS-CoV-2 genome (NCBI RefSeq "NC_045512.2") with the Multiple Sequence Alignment using Fast Fourier Transform (MAFFT) software package (version 7) with settings "–auto" (selects an appropriate strategy from L-INS-i, FFT-NS-i, or FFT-NS-2 algorithms, according to data size), "–addfragments" (adds unaligned fragmentary sequence(s) into an existing alignment), and "–keeplength" (keeps alignment length unchanged). Finally, any genome sequences that contained invalid nucleotide codes (e.g. W), greater than 1000 deletions, or greater than 1% unassigned nucleotides (nucleotide code N) were omitted, resulting in 7,722,980 SARS-CoV-2 sequences.

Each sequence originally contained the same 29,903 nucleotide positions after alignment. For each sequence, nucleotide positions before the first Artic primer overlap position (nucleotide position 342) and after the last Artic primer overlap position (nucleotide position 29,665) were omitted from all genome vectors. Then, the total number of mutations to each of A, C, G, T, and - (i.e., deletions) were summed across all sequences at each nucleotide position in the genome, and any nucleotide position that did not have at least one of these five mutation counts greater than 100 was omitted from all sequences. Any nucleotide positions that did not pass

this mutation filter were assigned the reference genome value at the corresponding position. Finally, all nucleotide codes other than A, C, G, T, and - (e.g., R) were treated as missing values by replacing the nucleotide codes with Ns, resulting in each of the 7.7 million SARS-CoV-2 sequences containing the same set of 21,433 nucleotide positions. All preprocessing steps were implemented using Python (version 3.9.2), in scripts that are publicly available at https://github.com/jeanmerlet/covid_comet.

### 4.3.2 Genome Vector Comparisons

To prepare each genome sequence for exhaustive vector comparison with CoMet, each vector was converted to binary using the following strategy. Nucleotides corresponding to 'A', 'C', 'G', 'T', and '-' were replaced with one-hot vectors with five bits. For example, 'A' corresponded to the vector '10000', 'C' to '01000', and '-' to '00001'. Missing values (i.e., 'N') were replaced by vectors of all zeros: '00000'. This process increased the genome vector length from 21,433 nucleotides to $21,433 \times 5 = 107,165$ bits.

To compare binary vectors, we modified the Duo similarity metric to account for our binarization scheme. Duo is a similarity metric that is used to compute the frequency of co-occurring binary elements in vector pairs. Duo is used to group binary vector values into high (pairs of 1's), low (pairs of 0's), and anti-correlated (pairs of 10 or 01) categories, which can be modified depending on the domain application. The Duo metric is represented by the equation

$$\text{Duo}_{ij} = mD_{ij}(1 - qf_i)(1 - qf_j) \tag{4.1}$$

where $i$ and $j$ are binary vectors, $D$ computes the frequency of a specified binary relationship (i.e., 11, 10, 01, or 00) between $i$ and $j$, $f$ computes the frequency of binary elements in vector $i$ or $j$, and $m = 4$ and $q = 2/3$ are scaling constants. This metric results in four values for each pair of input vectors based on the different bit

comparisons: high-high, high-low, low-high, and low-low, with each value computing the frequency of all such comparisons across the two vectors. For our binary vector comparisons, two genome sequences were similar if they shared nucleotides in corresponding positions (e.g., A's that lined up with A's, C's that lined up with C's, etc.). Or, if using the one-hot representation, two nucleotides were the same if they shared a 1 in the same position. We modified Equation 4.1 by: (i) setting $q = 0$ in order to disable the frequency terms, $f_i$ and $f_j$, and (ii) only considering the 'high-high' correlation. This reformulated Duo metric measured the frequency of co-occurrences of 1's between any two SARS-CoV-2 binary vectors, thereby measuring the co-occurrences of identical nucleotides.

Exhaustive pairwise vector comparisons were then computed using CoMet compiled with software versions GCC 6.4.0, Spectrum MPI 10.3.1.2-20200121, and CUDA 10.1.243. The full computational run was launched on the Summit supercomputer with IBM's Job Step Manager (JSM) software, which provided the 'jsrun' tool using six MPI ranks per node, with each MPI rank allocated 1 NVIDIA V100 GPU and 7 POWER9 CPU cores mapped via OpenMP threads. The following environment variables were set prior to the CoMet run:

- `OMP_NUM_THREADS=7`
- `PAMI_IBV_ENABLE_DCT=1`
- `PAMI_ENABLE_STRIPING=1`
- `PAMI_IBV_ADAPTER_AFFINITY=0`
- `PAMI_IBV_QP_SERVICE_LEVEL=8`
- `PAMI_IBV_ENABLE_OOO_AR=1`

The CoMet vector comparisons were thresholded at our modified Duo score of 0.19991 (i.e., ∼99.9% of the maximum theoretical score, 0.2) in order to retain the top ∼0.1% of most similar vector pairs. This process resulted in a sparse

sequence-to-sequence similarity network containing 5,290,386 sequences (nodes) and 777,928,464,646 sequence pair similarity scores (edges). This all-against-all comparison of SARS-CoV-2 genomes resulted in approximately $3 \times 10^{13}$ vector comparisons, and each sequence shared an edge with ∼147,000 other nodes on average in the network.

### Genomic Lineages

We applied Markov clustering to the sparse sequence-to-sequence similarity network to group sets of highly sequentially similar SARS-CoV-2 sequences together. Clustering was performed using HipMCL [10], a high-performance implementation of Markov clustering designed for use on distributed parallel computing systems. We considered a range of Markov clustering inflation values: 1.2, 1.5, 2.0, 4.0, and 8.0, which produced up to 80,000 novel lineages, and subsequently chose 1.2 inflation (which produced the smallest number of clusters) for further downstream analyses. Markov clustering at this inflation yielded 22,040 clusters, with the top 20 largest clusters containing 26% of all edges in the underlying sequence-to-sequence similarity network.

### Visualization

SARS-CoV-2 clusters were visualized using several geospatial and temporal views. Using the GISAID metadata for each sequence, we assigned sequences to dates and locations. For dates, any sequence with only the month and year of collection was automatically assigned the 15th day of the month. For locations, we extracted both the country and continent of collection. Any sequences missing such date or location information were removed during sequence preprocessing. Further, we used the PANGO lineage labels of each sequence to map them to the corresponding World Health Organization (WHO) variant, or "Other" if there was no corresponding WHO variant (e.g., "AY.1" was mapped to "Delta"). Then, for each cluster, we assigned a

WHO variant label if there was a clear majority variant (e.g., $\geq 90\%$ of the sequences within that cluster) and "Other" otherwise. We then chose the top four largest clusters belonging to each of the three largest strains through June 2022: "Alpha", "Delta", "Omicron", and "Other", resulting in a total of 16 data-driven lineages (i.e., 4 clusters $\times$ 4 WHO variants). The WHO clusters with majority variants not included in visualizations (e.g. Mu) comprised a small subset of all sequences (2.7% of sequences) and were omitted. To construct global spatiotemporal views, each country was chosen to display a unique color corresponding to the cluster from the set of top 16 clusters with the most sequences at that time point. Finally, we applied a 7-day moving average to each time series in order to smooth the noisy signal and improve the quality of the visualizations.

### 4.3.3 California Multiplex Network

To integrate whole-genome SARS-CoV-2 genomic structure with epidemiological and demographic measurements across California, a spatiotemporal multiplex network was constructed in which nodes represented county-week observations across three data modalities. The multiplex was designed to capture relationships within and between these data modalities while preserving temporal continuity and delayed epidemiological effects, resulting in five different edge types.

The multiplex network consisted of three primary modality-specific layers: genomic composition, COVID-19 mortality rates, and cumulative vaccination coverage. Each node in the multiplex was represented as:

`modality|county|week`

where `modality` corresponded to one of the three network layers, `county` represented a California county, and `week` corresponded to an ISO-8601 epidemiological week. For example:

`genomic|los_angeles|2021-W12`

represented the genomic composition of SARS-CoV-2 sequences sampled in Los Angeles County during epidemiological week 12 of 2021.

### Genomic Layer Construction

Whole-genome SARS-CoV-2 sequences originating from California were assigned to genomic clusters generated previously using HipMCL [10] clustering of exhaustive pairwise whole-genome sequence similarity relationships. For each county-week observation, the relative abundance of each HipMCL cluster was computed as:

$$p_{c,w,k} = \frac{n_{c,w,k}}{\sum_j n_{c,w,j}}$$

where $n_{c,w,k}$ represented the number of sequences belonging to genomic cluster $k$ observed in county $c$ during week $w$. County-week genomic composition vectors were then constructed from these normalized cluster abundance profiles. Pairwise similarity between county-week genomic profiles within each week was computed using Pearson correlation and assigned as edges between these nodes.

### Mortality Layer Construction

County-level COVID-19 mortality data were obtained from the California COVID-19 Time-Series Metrics by County and State dataset. Daily mortality counts were aggregated to epidemiological weeks and normalized by county population:

$$\text{mortality rate}_{c,w} = \frac{\text{weekly deaths}_{c,w}}{\text{population}_c} \times 100,000$$

Weekly mortality vectors were then compared across counties within each week using Pearson correlation, and KNN sparsification was applied to retain only the strongest local mortality relationships.

### Vaccination Layer Construction

County-level vaccination data were obtained from the California Statewide COVID-19 Vaccines Administered by County dataset. Weekly cumulative vaccination coverage was computed using the final reported cumulative fully vaccinated count for each county-week:

$$\text{vaccine coverage}_{c,w} = \frac{\text{cumulative fully vaccinated}_{c,w}}{\text{population}_c}$$

Pairwise similarity between county vaccination trajectories within each week was computed using Pearson correlation followed by KNN sparsification.

### Interlayer Edges

To integrate the three modalities into a single multiplex representation, interlayer edges were introduced between nodes corresponding to the same county-week observation across modalities:

$$\texttt{genomic|county|week} \leftrightarrow \texttt{mortality|county|week}$$

$$\texttt{genomic|county|week} \leftrightarrow \texttt{vaccine|county|week}$$

$$\texttt{mortality|county|week} \leftrightarrow \texttt{vaccine|county|week}$$

These interlayer edges encoded the assumption that genomic composition, mortality outcomes, and vaccination coverage jointly contribute to the epidemiological state of a county during a given week.

### Temporal Edges

Temporal continuity within each modality was modeled by connecting county observations across adjacent epidemiological weeks:

$$\texttt{modality|county|week}_t \leftrightarrow \texttt{modality|county|week}_{t+1}$$

Additional lagged interlayer edges were introduced to model delayed relationships between viral genomic structure and downstream mortality outcomes:

$$\texttt{genomic|county|week}_t \leftrightarrow \texttt{mortality|county|week}_{t+\Delta}$$

where $\Delta$ represented a biologically motivated mortality lag measured in weeks. The final multiplex network therefore jointly encoded:

- spatial similarity within genomic, mortality, and vaccination modalities
- cross-modality coupling between county-week epidemiological states
- temporal continuity within modalities
- delayed genomic-to-mortality relationships

### Random Walk with Restart Diffusion

To investigate relationships between viral genomic structure and downstream mortality topology, random walk with restart (RWR) diffusion was performed on the multiplex network. For a selected HipMCL genomic cluster, county-week genomic nodes were seeded proportionally to the relative abundance of the cluster:

$$s_{c,w,k} = p_{c,w,k} \cdot \log(1 + n_{c,w})$$

where $p_{c,w,k}$ represented the genomic cluster fraction and $n_{c,w}$ represented the total number of sequences sampled in county $c$ during week $w$. Logarithmic scaling was

applied to reduce the influence of extremely heavily sampled county-week observations while preserving sequencing depth information.

The RWR process iteratively propagated probability mass across the multiplex graph while repeatedly restarting from the seeded genomic nodes with probability $r$:

$$p^{(t+1)} = (1 - r)Wp^{(t)} + rp^{(0)}$$

where $W$ represented the normalized multiplex transition matrix and $p^{(0)}$ represented the initial seeded distribution.

After convergence, steady-state diffusion scores were extracted from mortality-layer nodes and aggregated spatially to generate county-level mortality diffusion maps associated with specific SARS-CoV-2 genomic clusters. For cluster-specific analyses, the multiplex graph was temporally restricted to the interval spanning the first and last observed weeks of each HipMCL cluster in order to reduce diffusion into unrelated epidemiological periods and later-emerging variants. Genomic seed nodes consisted of all county-week genomic nodes assigned to the selected HipMCL cluster within the restricted temporal interval. Mortality diffusion scores were then summed across all mortality-layer week nodes for each county and min-max normalized for visualization. Higher diffusion scores therefore reflected counties more strongly connected, through the multiplex topology, to the spatiotemporal transmission patterns associated with a given genomic cluster. Because diffusion occurred across genomic, mortality, vaccine, temporal, and interlayer edges simultaneously, the resulting mortality landscapes captured integrated relationships between viral genomic composition, geographic spread, temporal persistence, mortality burden, and vaccination context throughout the study period.

### Null Model

To evaluate whether observed mortality diffusion patterns arose from meaningful multiplex coupling structure rather than generic graph topology, a within-layer-preserving multiplex null model was constructed. The null model preserved all intralayer similarity relationships within the genomic, mortality, and vaccination layers while randomizing cross-context coupling structure, including temporal, interlayer, and lagged genomic–mortality edges. This design preserved the overall topology and local connectivity structure of each modality-specific layer while disrupting the biologically and temporally aligned relationships encoded between modalities and across time.

Edges within the genomic, mortality, and vaccination layers were retained unchanged during each permutation. In contrast, temporal edges connecting adjacent county-week observations within modalities, interlayer edges connecting modalities within county-week observations, and lagged genomic-to-mortality edges were shuffled through random permutation of target nodes within each edge class. Edge weights and the total number of edges within each class were preserved during shuffling. This generated randomized multiplex networks in which the underlying spatial similarity structure of each modality remained intact, but the cross-context coupling between layers was disrupted.

Random walk with restart diffusion was then repeated on each shuffled multiplex network using the same Alpha-associated genomic seed interval spanning 2021-W01 through 2021-W30. For each county-week mortality node, diffusion scores from the observed multiplex were compared to the distribution of diffusion scores obtained from 500 independently shuffled null multiplex networks. Standardized county-week enrichment scores were computed as:

$$z_{c,w} = \frac{D_{c,w}^{\text{obs}} - \mu_{c,w}^{\text{null}}}{\sigma_{c,w}^{\text{null}}}$$

where $D_{c,w}^{\text{obs}}$ represented the observed mortality diffusion score for county $c$ during week $w$, and $\mu_{c,w}^{\text{null}}$ and $\sigma_{c,w}^{\text{null}}$ represented the mean and standard deviation of diffusion scores across the rewired null multiplex networks, respectively.

Consequently, positive z-scores indicated county-weeks in which the multiplex structure concentrated mortality diffusion more strongly than expected under randomized cross-context coupling. This null framework enabled evaluation of whether observed diffusion landscapes reflected biologically and temporally coherent multiplex organization beyond expectations from generic within-layer epidemiological similarity structure alone.

## 4.4 Conclusion

In this chapter, we presented a novel, high-resolution, and data-driven classification of SARS-CoV-2 genomes generated from full-genome comparisons of 11 million sequences and its integration of sociodemographic data via multiplex. By applying the CoMet vector comparison codebase and HipMCL clustering to exhaustively compare and group viral genomes, we generated over 22,000 mutationally similar lineages that captured substantially more fine-scale genomic structure than conventional lineage designation systems such as PANGO lineages and WHO variants. These analyses demonstrated that whole-genome similarity relationships can reveal complex spatiotemporal lineage dynamics and geographically asynchronous transmission structure that may be obscured by broader variant classifications alone.

We further integrated these genomic lineages with county-level mortality and vaccination data through construction of a spatiotemporal California multiplex network. Random walk with restart diffusion on the multiplex revealed that distinct genomic clusters produced different mortality diffusion landscapes across California, suggesting that viral genomic structure, epidemiological dynamics, vaccination patterns, and temporal persistence collectively shaped downstream mortality-associated diffusion topology. Null model analyses demonstrated that these diffusion patterns were

not fully reproducible under randomized cross-context coupling despite preservation of within-layer epidemiological structure, supporting the interpretation that the observed mortality landscapes emerged from biologically and temporally coherent multiplex organization rather than generic graph topology alone. More broadly, this work illustrates how scalable graph-based and multiplex-network frameworks can integrate viral genomics with large-scale spatial epidemiological data to investigate relationships between viral evolution, transmission dynamics, and public health outcomes during ongoing and future pandemics.

# Chapter 5

# Conclusion

The studies presented in this dissertation collectively demonstrate how scalable computational methodologies can be applied to the increasing size and multi-faceted nature of modern biological datasets in order to better characterize disease-associated molecular systems across multiple biological scales. Advances in sequencing technologies, spatial profiling methods, and public genomic surveillance efforts have dramatically expanded both the scale and complexity of modern biological data. At the same time, these developments increasingly blur the boundaries between traditionally distinct domains such as transcriptomics, spatial biology, network science, epidemiology, and high-performance computing. Consequently, future biological discovery will likely depend not only on continued advances in experimental technologies, but also on the development of integrative computational frameworks capable of unifying diverse data modalities while remaining scalable to rapidly increasing dataset sizes.

In chapter 2, we applied integrated single-cell and spatial transcriptomic analyses to investigate the role of fibroblast populations in atopic dermatitis and eosinophilic inflammation. By combining mouse and human scRNA-seq datasets with multiplex network integration and spatial transcriptomic analysis, we identified conserved inflammatory fibroblast programs associated with chemokine signaling and immune

recruitment. These analyses demonstrated that *Ikkb* deletion in *Prx1*-lineage fibroblasts recapitulates transcriptional features observed in human atopic dermatitis, supporting the translational relevance of the murine model. Importantly, integration of spatial transcriptomic data further contextualized these transcriptional programs within intact tissue architecture, illustrating how spatial organization and localized cellular neighborhoods contribute to inflammatory signaling within diseased skin.

In chapter 3, we addressed the growing computational challenges associated with modern atlas-scale single-cell transcriptomic datasets through development of HALRA, a sparse-preserving and distributed-memory implementation of the ALRA imputation algorithm. By reformulating key stages of the reconstruction pipeline around sparse matrix operations, distributed decomposition strategies, and parallel I/O workflows, HALRA substantially reduced memory requirements while preserving comparable reconstruction behavior to existing ALRA implementations. Distributed benchmarking demonstrated efficient scaling across multiple compute nodes, enabling analyses of transcriptomic datasets substantially larger than those feasible under conventional single-node implementations. Beyond ALRA specifically, this work highlighted the increasing importance of sparse-preserving computation, distributed-memory architectures, and scalable storage strategies for future large-scale transcriptomic analysis workflows. As omics datasets continue to expand toward atlas and eventually perhaps whole-organism-scale resolutions, these computational design principles will likely become increasingly necessary across a broad range of systems biology methodologies.

In chapter 4, we extended systems-level analysis into the domain of viral genomic epidemiology through large-scale multiplex network modeling of SARS-CoV-2 genomic, mortality, and vaccination data across California counties and epidemiological time windows. By constructing integrated multiplex representations of these heterogeneous datasets and applying diffusion-based analyses, we demonstrated that the combined network structure captured temporally and geographically coherent patterns associated with SARS-CoV-2 lineage spread and downstream mortality

dynamics. These findings further illustrated how network-based integration of heterogeneous public health and genomic datasets can reveal relationships that are difficult to observe when modalities are analyzed independently. More broadly, this work demonstrated the utility of multiplex network methodologies for integrating diverse biological and epidemiological data streams into unified analytical frameworks capable of capturing complex population-scale dynamics.

Several future directions naturally emerge from this work. Added integration of other omics datasets (e.g., proteomics, metabolomics) with our single-cell RNA sequencing and spatial transcriptomics and multiplex network methodologies may provide increasingly comprehensive systems-level views of tissue organization and disease progression. Continued development of sparse-preserving and distributed-memory computational methodologies will likely become essential not only for transcriptomics, but also for emerging multimodal and imaging-intensive omics technologies that increasingly generate terabyte-scale datasets. Finally, extending multiplex epidemiological modeling frameworks to incorporate additional modalities such as mobility data, clinical metadata, or longitudinal host immune profiling may further improve interpretation of population-scale infectious disease dynamics. Ultimately, this dissertation demonstrates that the integration of scalable computational design with systems-level biological analysis enables new forms of biological inference across diverse domains of modern omics research. By combining network-based integration, single-cell and spatial transcriptomics, large-scale viral genomic analysis, and high-performance computational methodologies, this work contributes toward the broader goal of developing computational frameworks capable of extracting biologically meaningful insight from increasingly large, diverse, and complex biological datasets.

# Bibliography


[1] 10x Genomics. *1.3 Million Brain Cells from E18 Mice (v1.3.0)*. https://www.10xgenomics.com/datasets/1-3-million-brain-cells-from-e-18-mice-2-standard-1-3-0. Accessed June 11, 2025. 2017.

[2] Tamim Abdelaal et al. "A comparison of automatic cell identification methods for single-cell RNA sequencing data". In: *Genome Biology* 20.1 (Sept. 2019). ISSN: 1474-760X. DOI: 10.1186/s13059-019-1795-z. URL: http://dx.doi.org/10.1186/s13059-019-1795-z.

[3] Francis E. Agamah et al. "Computational approaches for network-based integrative multi-omics analysis". In: *Frontiers in Molecular Biosciences* 9 (Nov. 2022). ISSN: 2296-889X. DOI: 10.3389/fmolb.2022.967205. URL: http://dx.doi.org/10.3389/fmolb.2022.967205.

[4] Naomi Altman and Martin Krzywinski. "The curse(s) of dimensionality". In: *Nature Methods* 15.6 (May 2018), pp. 399–400. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0019-x. URL: http://dx.doi.org/10.1038/s41592-018-0019-x.

[5] J C Alwine, D J Kemp, and G R Stark. "Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 74.12 (Dec. 1977), pp. 5350–5354.

[6] Tallulah S. Andrews et al. "Tutorial: guidelines for the computational analysis of single-cell RNA sequencing data". In: *Nature Protocols* 16.1 (Dec. 2020), pp. 1–9. ISSN: 1750-2799. DOI: 10.1038/s41596-020-00409-w. URL: http://dx.doi.org/10.1038/s41596-020-00409-w.

[7] Cédric Arisdakessian et al. "DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data". In: *Genome Biology* 20.1 (Oct. 2019). ISSN: 1474-760X. DOI: 10.1186/s13059-019-1837-6. URL: http://dx.doi.org/10.1186/s13059-019-1837-6.

[8] Michael Ashburner et al. "Gene ontology: tool for the unification of biology". In: *Nature genetics* 25.1 (2000), pp. 25–29.

[9] Lyla Atta and Jean Fan. "Computational challenges and opportunities in spatially resolved transcriptomic data analysis". In: *Nature Communications* 12.1 (2021), p. 5283.

[10] Ariful Azad et al. "HipMCL: a high-performance parallel implementation of the Markov clustering algorithm for large-scale networks". In: *Nucleic Acids Research* 46.6 (Jan. 2018), e33–e33. ISSN: 1362-4962. DOI: 10.1093/nar/gkx1313. URL: http://dx.doi.org/10.1093/nar/gkx1313.

[11] Rhonda Bacher et al. "SCnorm: robust normalization of single-cell RNA-seq data". In: *Nature Methods* 14.6 (Apr. 2017), pp. 584–586. ISSN: 1548-7105. DOI: 10.1038/nmeth.4263. URL: http://dx.doi.org/10.1038/nmeth.4263.

[12] Arunkumar Bagavathi and Siddharth Krishnan. "Multi-Net: a scalable multiplex network embedding framework". In: *International conference on complex networks and their applications*. Springer. 2018, pp. 119–131.

[13] Matthew N Bainbridge et al. "Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach". In: *BMC Genomics* 7.1 (Sept. 2006). ISSN: 1471-2164. DOI: 10.1186/1471-2164-7-246. URL: http://dx.doi.org/10.1186/1471-2164-7-246.

[14] Trygve E. Bakken et al. “Single-nucleus and single-cell transcriptomes compared in matched cortical cell types”. In: *PLOS ONE* 13.12 (Dec. 2018). Ed. by Eduardo Soriano, e0209648. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0209648. URL: http://dx.doi.org/10.1371/journal.pone.0209648.

[15] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. “Network medicine: a network-based approach to human disease”. In: *Nature Reviews Genetics* 12.1 (Dec. 2010), pp. 56–68. ISSN: 1471-0064. DOI: 10.1038/nrg2918. URL: http://dx.doi.org/10.1038/nrg2918.

[16] Trevor Bedford et al. “Global circulation patterns of seasonal influenza viruses vary with antigenic drift”. In: *Nature* 523.7559 (June 2015), pp. 217–220. ISSN: 1476-4687. DOI: 10.1038/nature14460. URL: http://dx.doi.org/10.1038/nature14460.

[17] Trevor Bedford et al. “Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2)”. In: *PLoS Pathogens* 6.5 (May 2010). Ed. by Neil M. Ferguson, e1000918. ISSN: 1553-7374. DOI: 10.1371/journal.ppat.1000918. URL: http://dx.doi.org/10.1371/journal.ppat.1000918.

[18] David R. Bentley et al. “Accurate whole human genome sequencing using reversible terminator chemistry”. In: *Nature* 456.7218 (Nov. 2008), pp. 53–59. ISSN: 1476-4687. DOI: 10.1038/nature07517. URL: http://dx.doi.org/10.1038/nature07517.

[19] Max Bernstein. *Dimensionality Reduction: A Field Guide*. Accessed: 2025-06-13. 2023. URL: https://mbernste.github.io/posts/dim_reduc/.

[20] Tommaso Biancalani et al. “Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram”. In: *Nature Methods* 18.11 (Oct. 2021), pp. 1352–1362. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01264-7. URL: http://dx.doi.org/10.1038/s41592-021-01264-7.

[21] Usama Bilal et al. "Racial/ethnic and neighbourhood social vulnerability disparities in COVID-19 testing positivity, hospitalization, and in-hospital mortality in a large hospital system in Pennsylvania: A prospective study of electronic health records". In: *The Lancet Regional Health - Americas* 10 (June 2022), p. 100220. ISSN: 2667-193X. DOI: 10.1016/j.lana.2022.100220. URL: http://dx.doi.org/10.1016/j.lana.2022.100220.

[22] National Center for Biotechnology Information. *PubMed search results for "single cell sequencing" (2009–2025)*. https://pubmed.ncbi.nlm.nih.gov/?term=single+cell+sequencing&filter=years.2009-2025. Accessed: 2025-04-28. 2025.

[23] Vincent D Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/p10008. URL: http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

[24] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics* 30.15 (Apr. 2014), pp. 2114–2120. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu170. URL: http://dx.doi.org/10.1093/bioinformatics/btu170.

[25] Marco Bongiovanni et al. *MPXGAT: An Attention based Deep Learning Model for Multiplex Graphs Embedding*. 2024. DOI: 10.48550/ARXIV.2403.19246. URL: https://arxiv.org/abs/2403.19246.

[26] Dulal Borthakur et al. "Current status and trends in forest genomics". In: *Forestry Research* 2.11 (2022). ISSN: 2767-3812. DOI: 10.48130/fr-2022-0011. URL: http://dx.doi.org/10.48130/FR-2022-0011.

[27] Olga Borisovna Botvinnik et al. "Single-cell transcriptomics for the 99.9% of species without reference genomes". In: *bioRxiv* (July 2021). DOI: 10.1101/2021.07.09.450799. URL: http://dx.doi.org/10.1101/2021.07.09.450799.

[28] Nicolas L Bray et al. “Near-optimal probabilistic RNA-seq quantification”. In: *Nature Biotechnology* 34.5 (Apr. 2016), pp. 525–527. ISSN: 1546-1696. DOI: 10.1038/nbt.3519. URL: http://dx.doi.org/10.1038/nbt.3519.

[29] Anderson F. Brito et al. “Global disparities in SARS-CoV-2 genomic surveillance”. In: *Nature Communications* 13.1 (Nov. 2022). ISSN: 2041-1723. DOI: 10.1038/s41467-022-33713-y. URL: http://dx.doi.org/10.1038/s41467-022-33713-y.

[30] Bogdan Budnik et al. “SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation”. In: *Genome biology* 19 (2018), pp. 1–12.

[31] Jason D Buenrostro et al. “Single-cell chromatin accessibility reveals principles of regulatory variation”. In: *Nature* 523.7561 (2015), pp. 486–490.

[32] Elena Bushmanova et al. “rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data”. In: *GigaScience* 8.9 (Sept. 2019). ISSN: 2047-217X. DOI: 10.1093/gigascience/giz100. URL: http://dx.doi.org/10.1093/gigascience/giz100.

[33] Patrick Cahan et al. “CellNet: network biology applied to stem cell engineering”. In: *Cell* 158.4 (2014), pp. 903–915.

[34] Genís Calderer and Marieke L. Kuijjer. “Community Detection in Large-Scale Bipartite Biological Networks”. In: *Frontiers in Genetics* 12 (Apr. 2021). ISSN: 1664-8021. DOI: 10.3389/fgene.2021.649440. URL: http://dx.doi.org/10.3389/fgene.2021.649440.

[35] Kellen J. Cavagnero and Richard L. Gallo. “Essential immune functions of fibroblasts in innate host defense”. In: *Frontiers in Immunology* 13 (Dec. 2022). ISSN: 1664-3224. DOI: 10.3389/fimmu.2022.1058862. URL: http://dx.doi.org/10.3389/fimmu.2022.1058862.

[36] Ben C. L. Chan et al. "IL33: Roles in Allergic Inflammation and Therapeutic Perspectives". In: *Frontiers in Immunology* 10 (Mar. 2019). ISSN: 1664-3224. DOI: 10.3389/fimmu.2019.00364. URL: http://dx.doi.org/10.3389/fimmu.2019.00364.

[37] Ao Chen et al. "Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays". In: *Cell* 185.10 (May 2022), 1777–1792.e21. ISSN: 0092-8674. DOI: 10.1016/j.cell.2022.04.003. URL: http://dx.doi.org/10.1016/j.cell.2022.04.003.

[38] Chaoran Chen et al. "CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants". In: *Bioinformatics* 38.6 (Dec. 2021). Ed. by Can Alkan, pp. 1735–1737. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btab856. URL: http://dx.doi.org/10.1093/bioinformatics/btab856.

[39] J. Chen et al. "ToppGene Suite for gene list enrichment analysis and candidate gene prioritization". In: *Nucleic Acids Research* 37.Web Server (May 2009), W305–W311. ISSN: 1362-4962. DOI: 10.1093/nar/gkp427. URL: http://dx.doi.org/10.1093/nar/gkp427.

[40] Junhui Chen et al. "Recent advances of single-cell RNA sequencing in toxicology research: Insight into hepatotoxicity and nephrotoxicity". In: *Current Opinion in Toxicology* 37 (Mar. 2024), p. 100462. ISSN: 2468-2020. DOI: 10.1016/j.cotox.2024.100462. URL: http://dx.doi.org/10.1016/j.cotox.2024.100462.

[41] Kok Hao Chen et al. "Spatially resolved, highly multiplexed RNA profiling in single cells". In: *Science* 348.6233 (Apr. 2015). ISSN: 1095-9203. DOI: 10.1126/science.aaa6090. URL: http://dx.doi.org/10.1126/science.aaa6090.

[42] Zhiyuan Chen et al. "Global landscape of SARS-CoV-2 genomic surveillance and data sharing". In: *Nature Genetics* 54.4 (Mar. 2022), pp. 499–507. ISSN:

1546-1718. DOI: 10.1038/s41588-022-01033-y. URL: http://dx.doi.org/10.1038/s41588-022-01033-y.

[43] Yexiao Cheng et al. "covSampler: A subsampling method with balanced genetic diversity for large-scale SARS-CoV-2 genome data sets". In: *Virus Evolution* 8.2 (July 2022). ISSN: 2057-1577. DOI: 10.1093/ve/veac071. URL: http://dx.doi.org/10.1093/ve/veac071.

[44] Yi Cheng et al. "Evaluating imputation methods for single-cell RNA-seq data". In: *BMC Bioinformatics* 24.1 (July 2023). ISSN: 1471-2105. DOI: 10.1186/s12859-023-05417-7. URL: http://dx.doi.org/10.1186/s12859-023-05417-7.

[45] Foo Cheung et al. "Sequencing MedarticleEmrich2006, title = Gene discovery and annotation using LCM-454 transcriptome sequencing, volume = 17, ISSN = 1088-9051, url = http://dx.doi.org/10.1101/gr.5145806, DOI = 10.1101/gr.5145806, number = 1, journal = Genome Research, publisher = Cold Spring Harbor Laboratory, author = Emrich, Scott J. and Barbazuk, W. Brad and Li, Li and Schnable, Patrick S., year = 2006, month = nov, pages = 69–73 icago truncatula expressed sequenced tags using 454 Life Sciences technology". In: *BMC Genomics* 7.1 (Oct. 2006). ISSN: 1471-2164. DOI: 10.1186/1471-2164-7-272. URL: http://dx.doi.org/10.1186/1471-2164-7-272.

[46] Jun Yong Choi and Davey M. Smith. "SARS-CoV-2 Variants of Concern". In: *Yonsei Medical Journal* 62.11 (2021), p. 961. ISSN: 1976-2437. DOI: 10.3349/ymj.2021.62.11.961. URL: http://dx.doi.org/10.3349/ymj.2021.62.11.961.

[47] Kwangbom Choi et al. "Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics". In: *Genome Biology* 21.1 (July 2020). ISSN: 1474-760X. DOI: 10.1186/s13059-020-02103-2. URL: http://dx.doi.org/10.1186/s13059-020-02103-2.

[48] Pietro Cinaglia. "Multilayer biological network alignment based on similarity computation via Graph Neural Networks". In: *Journal of Computational Science* 78 (June 2024), p. 102259. ISSN: 1877-7503. DOI: 10.1016/j.jocs.2024.102259. URL: http://dx.doi.org/10.1016/j.jocs.2024.102259.

[49] Ashley Cliff et al. "A High-Performance Computing Implementation of Iterative Random Forest for the Creation of Predictive Expression Networks". In: *Genes* 10.12 (Dec. 2019), p. 996. ISSN: 2073-4425. DOI: 10.3390/genes10120996. URL: http://dx.doi.org/10.3390/genes10120996.

[50] Sharlee Climer et al. "A Custom Correlation Coefficient (CCC) Approach for Fast Identification of Multi-SNP Association Patterns in Genome-Wide SNPs Data". In: *Genetic Epidemiology* 38.7 (Aug. 2014), pp. 610–621. DOI: 10.1002/gepi.21833. URL: https://doi.org/10.1002/gepi.21833.

[51] Sharlee Climer et al. "Synchronized genetic activities in Alzheimer's brains revealed by heterogeneity-capturing network analysis". In: *bioRxiv* (2020).

[52] Peter J. A. Cock et al. "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants". In: *Nucleic Acids Research* 38.6 (Dec. 2009), pp. 1767–1771. ISSN: 1362-4962. DOI: 10.1093/nar/gkp1137. URL: http://dx.doi.org/10.1093/nar/gkp1137.

[53] Simone Codeluppi et al. "Spatial organization of the somatosensory cortex revealed by osmFISH". In: *Nature Methods* 15.11 (Oct. 2018), pp. 932–935. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0175-z. URL: http://dx.doi.org/10.1038/s41592-018-0175-z.

[54] Andrew Collette et al. *h5py/h5py: 3.7.0*. 2022. DOI: 10.5281/ZENODO.6575970. URL: https://zenodo.org/record/6575970.

[55] Helena L. Crowell et al. "muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data". In: *Nature Communications* 11.1 (Nov. 2020). ISSN: 2041-1723. DOI: 10.1038/

s41467-020-19894-4. URL: http://dx.doi.org/10.1038/s41467-020-19894-4.

[56] Xiaofeng Dai and Li Shen. "Advances and Trends in Omics Technology Development". In: *Frontiers in Medicine* 9 (July 2022). ISSN: 2296-858X. DOI: 10.3389/fmed.2022.911861. URL: http://dx.doi.org/10.3389/fmed.2022.911861.

[57] Lisandro Dalcín, Rodrigo Paz, and Mario Storti. "MPI for Python". In: *Journal of Parallel and Distributed Computing* 65.9 (Sept. 2005), pp. 1108–1115. ISSN: 0743-7315. DOI: 10.1016/j.jpdc.2005.03.010. URL: http://dx.doi.org/10.1016/j.jpdc.2005.03.010.

[58] Samarendra Das et al. "A Comprehensive Survey of Statistical Approaches for Differential Expression Analysis in Single-Cell RNA Sequencing Studies". In: *Genes* 12.12 (Dec. 2021), p. 1947. ISSN: 2073-4425. DOI: 10.3390/genes12121947. URL: http://dx.doi.org/10.3390/genes12121947.

[59] Timothy A. Davis. *Direct Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, Jan. 2006. ISBN: 9780898718881. DOI: 10.1137/1.9780898718881. URL: http://dx.doi.org/10.1137/1.9780898718881.

[60] Manlio De Domenico et al. "Identifying Modular Flows on Multilayer Networks Reveals Highly Overlapping Organization in Interconnected Systems". In: *Physical Review X* 5.1 (Mar. 2015). ISSN: 2160-3308. DOI: 10.1103/physrevx.5.011027. URL: http://dx.doi.org/10.1103/PhysRevX.5.011027.

[61] Pasquale De Meo et al. "Generalized Louvain method for community detection in large networks". In: *2011 11th International Conference on Intelligent Systems Design and Applications*. IEEE, Nov. 2011, pp. 88–93. DOI: 10.1109/isda.2011.6121636. URL: http://dx.doi.org/10.1109/ISDA.2011.6121636.

[62] Jeffrey Dean and Sanjay Ghemawat. "MapReduce: simplified data processing on large clusters". In: *Communications of the ACM* 51.1 (Jan. 2008), pp. 107–113. ISSN: 1557-7317. DOI: 10.1145/1327452.1327492. URL: http://dx.doi.org/10.1145/1327452.1327492.

[63] James Demmel et al. "Communication-optimal Parallel and Sequential QR and LU Factorizations". In: *SIAM Journal on Scientific Computing* 34.1 (Jan. 2012), A206–A239. ISSN: 1095-7197. DOI: 10.1137/080731992. URL: http://dx.doi.org/10.1137/080731992.

[64] Gilles Didier, Alberto Valdeolivas, and Anaïs Baudot. "Identifying communities from multiplex biological networks by randomized optimization of modularity". In: *F1000Research* 7 (Nov. 2018), p. 1042. ISSN: 2046-1402. DOI: 10.12688/f1000research.15486.2. URL: http://dx.doi.org/10.12688/f1000research.15486.2.

[65] David van Dijk et al. "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion". In: *Cell* 174.3 (July 2018), 716–729.e27. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.05.061. URL: http://dx.doi.org/10.1016/j.cell.2018.05.061.

[66] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (Oct. 2012), pp. 15–21. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts635. URL: http://dx.doi.org/10.1093/bioinformatics/bts635.

[67] Elie Dolgin. "The most popular genes in the human genome". In: *Nature* 551.7681 (Nov. 2017), pp. 427–431. ISSN: 1476-4687. DOI: 10.1038/d41586-017-07291-9. URL: http://dx.doi.org/10.1038/d41586-017-07291-9.

[68] Stijn van Dongen. *MCL – a cluster algorithm for graphs*. Accessed: June 5, 2025. 2021. URL: https://micans.org/mcl/.

[69] Siobain Duffy. "Why are RNA virus mutation rates so damn high?" In: *PLOS Biology* 16.8 (Aug. 2018), e3000003. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.3000003. URL: http://dx.doi.org/10.1371/journal.pbio.3000003.

[70] Eran Eden et al. "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists". In: *BMC Bioinformatics* 10.1 (Feb. 2009). ISSN: 1471-2105. DOI: 10.1186/1471-2105-10-48. URL: http://dx.doi.org/10.1186/1471-2105-10-48.

[71] Mirjana Efremova et al. "CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes". In: *Nature Protocols* 15.4 (Feb. 2020), pp. 1484–1506. ISSN: 1750-2799. DOI: 10.1038/s41596-020-0292-x. URL: http://dx.doi.org/10.1038/s41596-020-0292-x.

[72] John Eid et al. "Real-Time DNA Sequencing from Single Polymerase Molecules". In: *Science* 323.5910 (Jan. 2009), pp. 133–138. ISSN: 1095-9203. DOI: 10.1126/science.1162986. URL: http://dx.doi.org/10.1126/science.1162986.

[73] Stefan Elbe and Gemma Buckland-Merrett. "Data, disease and diplomacy: GISAID's innovative contribution to global health". In: *Global Challenges* 1.1 (Jan. 2017), pp. 33–46. ISSN: 2056-6646. DOI: 10.1002/gch2.1018. URL: http://dx.doi.org/10.1002/gch2.1018.

[74] Scott J. Emrich et al. "Gene discovery and annotation using LCM-454 transcriptome sequencing". In: *Genome Research* 17.1 (Nov. 2006), pp. 69–73. ISSN: 1088-9051. DOI: 10.1101/gr.5145806. URL: http://dx.doi.org/10.1101/gr.5145806.

[75] Denis English and Burton R. Andersen. "Single-step separation of red blood cells, granulocytes and mononuclear leukocytes on discontinuous density gradients of Ficoll-Hypaque". In: *Journal of Immunological Methods* 5.3 (Aug. 1974), pp. 249–252. ISSN: 0022-1759. DOI: 10.1016/0022-1759(74)90109-4. URL: http://dx.doi.org/10.1016/0022-1759(74)90109-4.

[76] Gökcen Eraslan et al. "Single-cell RNA-seq denoising using a deep count autoencoder". In: *Nature Communications* 10.1 (Jan. 2019). ISSN: 2041-1723. DOI: 10.1038/s41467-018-07931-2. URL: http://dx.doi.org/10.1038/s41467-018-07931-2.

[77] Rossin Erbe et al. "The use of machine learning to discover regulatory networks controlling biological systems". In: *Molecular Cell* 82.2 (Jan. 2022), pp. 260–273. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2021.12.011. URL: http://dx.doi.org/10.1016/j.molcel.2021.12.011.

[78] N. Benjamin Erichson et al. "Randomized Matrix Decompositions Using R". In: *Journal of Statistical Software* 89.11 (2019). ISSN: 1548-7660. DOI: 10.18637/jss.v089.i11. URL: http://dx.doi.org/10.18637/jss.v089.i11.

[79] European Centre for Disease Prevention and Control. *SARS-CoV-2 variants of concern*. https://www.ecdc.europa.eu/en/covid-19/variants-concern. Accessed: 2025-06-05. 2024.

[80] Brent Ewing et al. "Base-Calling of Automated Sequencer Traces Using *Phred*. I. Accuracy Assessment". In: *Genome Research* 8.3 (Mar. 1998), pp. 175–185. ISSN: 1549-5469. DOI: 10.1101/gr.8.3.175. URL: http://dx.doi.org/10.1101/gr.8.3.175.

[81] Karl Pearson F.R.S. "LIII. On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572. DOI: 10.1080/14786440109462720.

[82] Yanghe Feng, Qi Wang, and Tengjiao Wang. "Drug Target Protein-Protein Interaction Networks: A Systematic Perspective". In: *BioMed Research International* 2017 (2017), pp. 1–13. ISSN: 2314-6141. DOI: 10.1155/2017/1289259. URL: http://dx.doi.org/10.1155/2017/1289259.

[83] Greg Finak et al. "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data". In: *Genome Biology* 16.1 (Dec. 2015). ISSN: 1474-760X. DOI: 10.1186/s13059-015-0844-5. URL: http://dx.doi.org/10.1186/s13059-015-0844-5.

[84] E Fix and J Hodges. "An important contribution to nonparametric discriminant analysis and density estimation". In: *International Statistical Review* 3.57 (1951), pp. 233–238.

[85] Santo Fortunato and Marc Barthélemy. "Resolution limit in community detection". In: *Proceedings of the National Academy of Sciences* 104.1 (Jan. 2007), pp. 36–41. ISSN: 1091-6490. DOI: 10.1073/pnas.0605965104. URL: http://dx.doi.org/10.1073/pnas.0605965104.

[86] Melania Franchini et al. "Single-cell gene set enrichment analysis and transfer learning for functional annotation of scRNA-seq data". In: *NAR Genomics and Bioinformatics* 5.1 (Jan. 2023). ISSN: 2631-9268. DOI: 10.1093/nargab/lqad024. URL: http://dx.doi.org/10.1093/nargab/lqad024.

[87] Yu Fu et al. "Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers". In: *BMC Genomics* 19.1 (July 2018). ISSN: 1471-2164. DOI: 10.1186/s12864-018-4933-1. URL: http://dx.doi.org/10.1186/s12864-018-4933-1.

[88] Jake Gagnon et al. "Recommendations of scRNA-seq Differential Gene Expression Analysis Based on Comprehensive Benchmarking". In: *Life* 12.6 (June 2022), p. 850. ISSN: 2075-1729. DOI: 10.3390/life12060850. URL: http://dx.doi.org/10.3390/life12060850.

[89] Jennifer L. Gardy and Nicholas J. Loman. "Towards a genomics-informed, real-time, global pathogen surveillance system". In: *Nature Reviews Genetics* 19.1 (Nov. 2017), pp. 9–20. ISSN: 1471-0064. DOI: 10.1038/nrg.2017.88. URL: http://dx.doi.org/10.1038/nrg.2017.88.

[90] Debajyoti Ghosh et al. "Leveraging Multilayered "Omics" Data for Atopic Dermatitis: A Road Map to Precision Medicine". In: *Frontiers in Immunology* 9 (Dec. 2018). ISSN: 1664-3224. DOI: 10.3389/fimmu.2018.02727. URL: http://dx.doi.org/10.3389/fimmu.2018.02727.

[91] Stephen K. Gire et al. "Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak". In: *Science* 345.6202 (Sept. 2014), pp. 1369–1372. ISSN: 1095-9203. DOI: 10.1126/science.1259657. URL: http://dx.doi.org/10.1126/science.1259657.

[92] Wuming Gong et al. "DrImpute: imputing dropout events in single cell RNA sequencing data". In: *BMC Bioinformatics* 19.1 (June 2018). ISSN: 1471-2105. DOI: 10.1186/s12859-018-2226-y. URL: http://dx.doi.org/10.1186/s12859-018-2226-y.

[93] Sara Goodwin, John D. McPherson, and W. Richard McCombie. "Coming of age: ten years of next-generation sequencing technologies". In: *Nature Reviews Genetics* 17.6 (May 2016), pp. 333–351. ISSN: 1471-0064. DOI: 10.1038/nrg.2016.49. URL: http://dx.doi.org/10.1038/nrg.2016.49.

[94] Manfred G Grabherr et al. "Full-length transcriptome assembly from RNA-Seq data without a reference genome". In: *Nature Biotechnology* 29.7 (May 2011), pp. 644–652. ISSN: 1546-1696. DOI: 10.1038/nbt.1883. URL: http://dx.doi.org/10.1038/nbt.1883.

[95] Noah F. Greenwald et al. "Whole-cell segmentation of tissue images with human-level performance using large-scale data annotation and deep learning". In: *Nature Biotechnology* 40.4 (Nov. 2021), pp. 555–565. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01094-0. URL: http://dx.doi.org/10.1038/s41587-021-01094-0.

[96] Alexander L. Greninger. "The challenge of diagnostic metagenomics". In: *Expert Review of Molecular Diagnostics* 18.7 (June 2018), pp. 605–615. ISSN:

1744-8352. DOI: 10.1080/14737159.2018.1487292. URL: http://dx.doi.org/10.1080/14737159.2018.1487292.

[97] Rashel V. Grindberg et al. "RNA-sequencing from single nuclei". In: *Proceedings of the National Academy of Sciences* 110.49 (Nov. 2013), pp. 19802–19807. ISSN: 1091-6490. DOI: 10.1073/pnas.1319700110. URL: http://dx.doi.org/10.1073/pnas.1319700110.

[98] The HDF Group. *Hierarchical Data Format, version 5*. URL: https://github.com/HDFGroup/hdf5.

[99] Dominic Grün et al. "Single-cell messenger RNA sequencing reveals rare intestinal cell types". In: *Nature* 525.7568 (Aug. 2015), pp. 251–255. ISSN: 1476-4687. DOI: 10.1038/nature14966. URL: http://dx.doi.org/10.1038/nature14966.

[100] Minzhe Guo et al. "SINCERA: A Pipeline for Single-Cell RNA-Seq Profiling Analysis". In: *PLOS Computational Biology* 11.11 (Nov. 2015). Ed. by Andreas Prlic, e1004575. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1004575. URL: http://dx.doi.org/10.1371/journal.pcbi.1004575.

[101] Qing Guo et al. "NF-$\kappa$B in biology and targeted therapy: new insights and translational implications". In: *Signal Transduction and Targeted Therapy* 9.1 (Mar. 2024). ISSN: 2059-3635. DOI: 10.1038/s41392-024-01757-9. URL: http://dx.doi.org/10.1038/s41392-024-01757-9.

[102] Zhiye Guo et al. "Diffusion models in bioinformatics and computational biology". In: *Nature Reviews Bioengineering* 2.2 (Oct. 2023), pp. 136–154. ISSN: 2731-6092. DOI: 10.1038/s44222-023-00114-9. URL: http://dx.doi.org/10.1038/s44222-023-00114-9.

[103] Cristian D. Gutierrez Reyes et al. "Multi Omics Applications in Biological Systems". In: *Current Issues in Molecular Biology* 46.6 (June 2024), pp. 5777–

5793. ISSN: 1467-3045. DOI: 10.3390/cimb46060345. URL: http://dx.doi.org/10.3390/cimb46060345.

[104] Naomi Habib et al. "Massively parallel single-nucleus RNA-seq with DroNc-seq". In: *Nature Methods* 14.10 (Aug. 2017), pp. 955–958. ISSN: 1548-7105. DOI: 10.1038/nmeth.4407. URL: http://dx.doi.org/10.1038/nmeth.4407.

[105] James Hadfield et al. "Nextstrain: real-time tracking of pathogen evolution". In: *Bioinformatics* 34.23 (May 2018). Ed. by Janet Kelso, pp. 4121–4123. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/bty407. URL: http://dx.doi.org/10.1093/bioinformatics/bty407.

[106] Christoph Hafemeister and Rahul Satija. "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome Biology* 20.1 (Dec. 2019). ISSN: 1474-760X. DOI: 10.1186/s13059-019-1874-1. URL: http://dx.doi.org/10.1186/s13059-019-1874-1.

[107] N. Halko, P. G. Martinsson, and J. A. Tropp. "Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions". In: *SIAM Review* 53.2 (Jan. 2011), pp. 217–288. ISSN: 1095-7200. DOI: 10.1137/090771806. URL: http://dx.doi.org/10.1137/090771806.

[108] R. W. Hamming. "Error Detecting and Error Correcting Codes". In: *Bell System Technical Journal* 29.2 (Apr. 1950), pp. 147–160. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1950.tb00463.x. URL: http://dx.doi.org/10.1002/j.1538-7305.1950.tb00463.x.

[109] Yuhan Hao et al. "Dictionary learning for integrative, multimodal and scalable single-cell analysis". In: *Nature Biotechnology* 42.2 (May 2023), pp. 293–304. ISSN: 1546-1696. DOI: 10.1038/s41587-023-01767-y. URL: http://dx.doi.org/10.1038/s41587-023-01767-y.

[110] Yuhan Hao et al. "Integrated analysis of multimodal single-cell data". In: *Cell* 184.13 (June 2021), 3573–3587.e29. ISSN: 0092-8674. DOI: 10.1016/j.cell.2021.04.048. URL: http://dx.doi.org/10.1016/j.cell.2021.04.048.

[111] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2649-2. URL: http://dx.doi.org/10.1038/s41586-020-2649-2.

[112] Tamar Hashimshony et al. "CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification". In: *Cell Reports* 2.3 (Sept. 2012), pp. 666–673. ISSN: 2211-1247. DOI: 10.1016/j.celrep.2012.08.003. URL: http://dx.doi.org/10.1016/j.celrep.2012.08.003.

[113] Helen He et al. "Single-cell transcriptome analysis of human skin identifies novel fibroblast subpopulation and enrichment of immune subsets in atopic dermatitis". In: *Journal of Allergy and Clinical Immunology* 145.6 (June 2020), pp. 1615–1628. ISSN: 0091-6749. DOI: 10.1016/j.jaci.2020.01.042. URL: http://dx.doi.org/10.1016/j.jaci.2020.01.042.

[114] Shanshan He et al. "High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging". In: *Nature Biotechnology* 40.12 (Oct. 2022), pp. 1794–1806. ISSN: 1546-1696. DOI: 10.1038/s41587-022-01483-z. URL: http://dx.doi.org/10.1038/s41587-022-01483-z.

[115] Ana Hernández-Plaza et al. "eggNOG 6.0: enabling comparative genomics across 12,535 organisms". In: *Nucleic Acids Research* 51.D1 (Nov. 2022), pp. D389–D394. ISSN: 1362-4962. DOI: 10.1093/nar/gkac1022. URL: http://dx.doi.org/10.1093/nar/gkac1022.

[116] Lukas Heumos et al. "Best practices for single-cell analysis across modalities". In: *Nature Reviews Genetics* 24.8 (Mar. 2023), pp. 550–572. ISSN: 1471-0064. DOI: 10.1038/s41576-023-00586-w. URL: http://dx.doi.org/10.1038/s41576-023-00586-w.

[117] Brian Hie, Bryan Bryson, and Bonnie Berger. "Efficient integration of heterogeneous single-cell transcriptomes using Scanorama". In: *Nature Biotechnology* 37.6 (May 2019), pp. 685–691. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0113-3. URL: http://dx.doi.org/10.1038/s41587-019-0113-3.

[118] Verity Hill et al. "Toward a global virus genomic surveillance network". In: *Cell Host & Microbe* 31.6 (June 2023), pp. 861–873. ISSN: 1931-3128. DOI: 10.1016/j.chom.2023.03.003. URL: http://dx.doi.org/10.1016/j.chom.2023.03.003.

[119] Rui Hong et al. "Comprehensive generation, visualization, and reporting of quality control metrics for single-cell RNA sequencing data". In: *Nature Communications* 13.1 (Mar. 2022). ISSN: 2041-1723. DOI: 10.1038/s41467-022-29212-9. URL: http://dx.doi.org/10.1038/s41467-022-29212-9.

[120] Tianheng Hou et al. "IL-37 Ameliorating Allergic Inflammation in Atopic Dermatitis Through Regulating Microbiota and AMPK-mTOR Signaling Pathway-Modulated Autophagy Mechanism". In: *Frontiers in Immunology* 11 (Apr. 2020). ISSN: 1664-3224. DOI: 10.3389/fimmu.2020.00752. URL: http://dx.doi.org/10.3389/fimmu.2020.00752.

[121] Wenpin Hou et al. "A systematic evaluation of single-cell RNA-sequencing imputation methods". In: *Genome Biology* 21.1 (Aug. 2020). ISSN: 1474-760X. DOI: 10.1186/s13059-020-02132-x. URL: http://dx.doi.org/10.1186/s13059-020-02132-x.

[122] Jian Hu et al. "SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network". In: *Nature Methods* 18.11 (Oct. 2021), pp. 1342–1351. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01255-8. URL: http://dx.doi.org/10.1038/s41592-021-01255-8.

[123] Dezhi Huang et al. "Advances in single-cell RNA sequencing and its applications in cancer research". In: *Journal of Hematology & Oncology* 16.1 (Aug. 2023). ISSN: 1756-8722. DOI: 10.1186/s13045-023-01494-6. URL: http://dx.doi.org/10.1186/s13045-023-01494-6.

[124] Kongwei Huang et al. "The Advancement and Application of the Single-Cell Transcriptome in Biological and Medical Research". In: *Biology* 13.6 (June 2024), p. 451. ISSN: 2079-7737. DOI: 10.3390/biology13060451. URL: http://dx.doi.org/10.3390/biology13060451.

[125] Mo Huang et al. "SAVER: gene expression recovery for single-cell RNA sequencing". In: *Nature Methods* 15.7 (June 2018), pp. 539–542. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0033-z. URL: http://dx.doi.org/10.1038/s41592-018-0033-z.

[126] Mélanie Humeau, Katia Boniface, and Charles Bodet. "Cytokine-Mediated Crosstalk Between Keratinocytes and T Cells in Atopic Dermatitis". In: *Frontiers in Immunology* 13 (Apr. 2022). ISSN: 1664-3224. DOI: 10.3389/fimmu.2022.801579. URL: http://dx.doi.org/10.3389/fimmu.2022.801579.

[127] Inc. Illumina. *bcl2fastq Conversion Software v2.20*. Accessed April 10, 2025. 2017. URL: https://support.illumina.com/downloads/bcl2fastq-conversion-software-v2-20.html.

[128] Md Torikul Islam et al. "Advances in the Application of Single-Cell Transcriptomics in Plant Systems and Synthetic Biology". In: *BioDesign Research* 6 (2024), p. 0029. ISSN: 2693-1257. DOI: 10.34133/bdr.0029. URL: http://dx.doi.org/10.34133/bdr.0029.

[129] Saiful Islam et al. "Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq". In: *Genome Research* 21.7 (May 2011), pp. 1160–1167. ISSN: 1088-9051. DOI: 10.1101/gr.110882.110. URL: http://dx.doi.org/10.1101/gr.110882.110.

[130] Hartland W. Jackson et al. “The single-cell pathology landscape of breast cancer”. In: *Nature* 578.7796 (Jan. 2020), pp. 615–620. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1876-x. URL: http://dx.doi.org/10.1038/s41586-019-1876-x.

[131] Diego Adhemar Jaitin et al. “Massively Parallel Single-Cell RNA-Seq for Marker-Free Decomposition of Tissues into Cell Types”. In: *Science* 343.6172 (Feb. 2014), pp. 776–779. ISSN: 1095-9203. DOI: 10.1126/science.1247651. URL: http://dx.doi.org/10.1126/science.1247651.

[132] Amanda Janesick et al. “High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis”. In: *Nature Communications* 14.1 (Dec. 2023). ISSN: 2041-1723. DOI: 10.1038/s41467-023-43458-x. URL: http://dx.doi.org/10.1038/s41467-023-43458-x.

[133] Andrew L. Ji et al. “Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma”. In: *Cell* 182.2 (July 2020), 497–514.e22. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.05.039. URL: http://dx.doi.org/10.1016/j.cell.2020.05.039.

[134] Wei Jiang et al. “Network-based multi-omics integrative analysis methods in drug discovery: a systematic review”. In: *BioData Mining* 18.1 (Mar. 2025). ISSN: 1756-0381. DOI: 10.1186/s13040-025-00442-z. URL: http://dx.doi.org/10.1186/s13040-025-00442-z.

[135] Haoli Jin et al. “Animal Models of Atopic Dermatitis”. In: *Journal of Investigative Dermatology* 129.1 (Jan. 2009), pp. 31–40. ISSN: 0022-202X. DOI: 10.1038/jid.2008.106. URL: http://dx.doi.org/10.1038/jid.2008.106.

[136] Simon Joost et al. “Single-Cell Transcriptomics Reveals that Differentiation and Spatial Signatures Shape Epidermal and Hair Follicle Heterogeneity”. In: *Cell Systems* 3.3 (Sept. 2016), 221–237.e9. ISSN: 2405-4712. DOI: 10.1016/j.cels.2016.08.010. URL: http://dx.doi.org/10.1016/j.cels.2016.08.010.

[137] Wayne Joubert. *Combinatorial Metrics (CoMet)*. https://github.com/wdj/comet. Accessed: 2025-06-05. 2024.

[138] Wayne Joubert et al. "Attacking the opioid epidemic: determining the epistatic and pleiotropic genetic architectures for chronic pain and opioid addiction". In: *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. SC '18. Dallas, Texas: IEEE Press, 2018.

[139] Wayne Joubert et al. "Parallel accelerated Custom Correlation Coefficient calculations for genomics applications". In: *Parallel Computing* 84 (May 2019), pp. 15–23. ISSN: 0167-8191. DOI: 10.1016/j.parco.2019.02.003. URL: http://dx.doi.org/10.1016/j.parco.2019.02.003.

[140] Wayne Joubert et al. "Parallel accelerated vector similarity calculations for genomics applications". In: *Parallel Computing* 75 (July 2018), pp. 130–145. ISSN: 0167-8191. DOI: 10.1016/j.parco.2018.03.009. URL: http://dx.doi.org/10.1016/j.parco.2018.03.009.

[141] Dragomirka Jovic et al. "Single-cell RNA sequencing technologies and applications: A brief overview". In: *Clinical and Translational Medicine* 12.3 (Mar. 2022). ISSN: 2001-1326. DOI: 10.1002/ctm2.694. URL: http://dx.doi.org/10.1002/ctm2.694.

[142] Hidaya A. Kader et al. "Current Insights into Immunology and Novel Therapeutics of Atopic Dermatitis". In: *Cells* 10.6 (June 2021), p. 1392. ISSN: 2073-4409. DOI: 10.3390/cells10061392. URL: http://dx.doi.org/10.3390/cells10061392.

[143] David Kainer et al. "RWRtoolkit: multi-omic network analysis using random walks on multiplex networks in any species". In: *GigaScience* 14 (2025). ISSN: 2047-217X. DOI: 10.1093/gigascience/giaf028. URL: http://dx.doi.org/10.1093/gigascience/giaf028.

[144] Joyce B. Kang et al. “Efficient and precise single-cell reference atlas mapping with Symphony”. In: *Nature Communications* 12.1 (Oct. 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-25957-x. URL: http://dx.doi.org/10.1038/s41467-021-25957-x.

[145] Justus M. Kebschull and Anthony M. Zador. “Sources of PCR-induced distortions in high-throughput sequencing data sets”. In: *Nucleic Acids Research* (July 2015), gkv717. ISSN: 1362-4962. DOI: 10.1093/nar/gkv717. URL: http://dx.doi.org/10.1093/nar/gkv717.

[146] Leeat Keren et al. “A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging”. In: *Cell* 174.6 (Sept. 2018), 1373–1387.e19. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.08.039. URL: http://dx.doi.org/10.1016/j.cell.2018.08.039.

[147] Peter V Kharchenko, Lev Silberstein, and David T Scadden. “Bayesian approach to single-cell differential expression analysis”. In: *Nature Methods* 11.7 (May 2014), pp. 740–742. ISSN: 1548-7105. DOI: 10.1038/nmeth.2967. URL: http://dx.doi.org/10.1038/nmeth.2967.

[148] Peter V. Kharchenko. “The triumphs and limitations of computational methods for scRNA-seq”. In: *Nature Methods* 18.7 (June 2021), pp. 723–732. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01171-x. URL: http://dx.doi.org/10.1038/s41592-021-01171-x.

[149] Shruti Khare et al. “GISAID’s Role in Pandemic Response”. In: *China CDC Weekly* 3.49 (2021), pp. 1049–1051. ISSN: 2096-7071. DOI: 10.46234/ccdcw2021.255. URL: http://dx.doi.org/10.46234/ccdcw2021.255.

[150] Rohan Khazanchi et al. “County-Level Association of Social Vulnerability with COVID-19 Cases and Deaths in the USA”. In: *Journal of General Internal Medicine* 35.9 (June 2020), pp. 2784–2787. ISSN: 1525-1497. DOI: 10.1007/s11606-020-05882-3. URL: http://dx.doi.org/10.1007/s11606-020-05882-3.

[151] Nadiya Khyzha, Steven Henikoff, and Kami Ahmad. "Profiling RNA at chromatin targets in situ by antibody-targeted tagmentation". In: *Nature Methods* 19.11 (Oct. 2022), pp. 1383–1392. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01618-9. URL: http://dx.doi.org/10.1038/s41592-022-01618-9.

[152] Chan Yeong Kim et al. "HumanNet v3: an improved database of human gene networks for disease research". In: *Nucleic Acids Research* 50.D1 (Nov. 2021), pp. D632–D639. ISSN: 1362-4962. DOI: 10.1093/nar/gkab1048. URL: http://dx.doi.org/10.1093/nar/gkab1048.

[153] Daehwan Kim et al. "Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype". In: *Nature Biotechnology* 37.8 (Aug. 2019), pp. 907–915. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0201-4. URL: http://dx.doi.org/10.1038/s41587-019-0201-4.

[154] Vladimir Yu Kiselev, Tallulah S. Andrews, and Martin Hemberg. "Challenges in unsupervised clustering of single-cell RNA-seq data". In: *Nature Reviews Genetics* 20.5 (Jan. 2019), pp. 273–282. ISSN: 1471-0064. DOI: 10.1038/s41576-018-0088-9. URL: http://dx.doi.org/10.1038/s41576-018-0088-9.

[155] Vladimir Yu Kiselev et al. "SC3: consensus clustering of single-cell RNA-seq data". In: *Nature Methods* 14.5 (Mar. 2017), pp. 483–486. ISSN: 1548-7105. DOI: 10.1038/nmeth.4236. URL: http://dx.doi.org/10.1038/nmeth.4236.

[156] Teemu Kivioja et al. "Counting absolute numbers of molecules using unique molecular identifiers". In: *Nature Methods* 9.1 (Nov. 2011), pp. 72–74. ISSN: 1548-7105. DOI: 10.1038/nmeth.1778. URL: http://dx.doi.org/10.1038/nmeth.1778.

[157] Kang I. Ko et al. "NF-$\kappa$B perturbation reveals unique immunomodulatory functions in Prx1+ fibroblasts that promote development of atopic dermatitis". In: *Science Translational Medicine* 14.630 (Feb. 2022). ISSN: 1946-6242. DOI:

10.1126/scitranslmed.abj0324. URL: http://dx.doi.org/10.1126/scitranslmed.abj0324.

[158] Frank Konings et al. "SARS-CoV-2 Variants of Interest and Concern naming scheme conducive for global discourse". In: *Nature Microbiology* 6.7 (June 2021), pp. 821–823. ISSN: 2058-5276. DOI: 10.1038/s41564-021-00932-w. URL: http://dx.doi.org/10.1038/s41564-021-00932-w.

[159] B. Korber et al. "Timing the Ancestor of the HIV-1 Pandemic Strains". In: *Science* 288.5472 (June 2000), pp. 1789–1796. ISSN: 1095-9203. DOI: 10.1126/science.288.5472.1789. URL: http://dx.doi.org/10.1126/science.288.5472.1789.

[160] Ilya Korsunsky et al. "Fast, sensitive and accurate integration of single-cell data with Harmony". In: *Nature Methods* 16.12 (Nov. 2019), pp. 1289–1296. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0619-0. URL: http://dx.doi.org/10.1038/s41592-019-0619-0.

[161] Moritz U. G. Kraemer et al. "The effect of human mobility and control measures on the COVID-19 epidemic in China". In: *Science* 368.6490 (May 2020), pp. 493–497. ISSN: 1095-9203. DOI: 10.1126/science.abb4218. URL: http://dx.doi.org/10.1126/science.abb4218.

[162] Aaron T. L. Lun, Karsten Bach, and John C. Marioni. "Pooling across cells to normalize single-cell RNA sequencing data with many zero counts". In: *Genome Biology* 17.1 (Apr. 2016). ISSN: 1474-760X. DOI: 10.1186/s13059-016-0947-7. URL: http://dx.doi.org/10.1186/s13059-016-0947-7.

[163] John Lagergren et al. "Climatic Clustering and Longitudinal Analysis with Impacts on Food, Bioenergy, and Pandemics". In: *Phytobiomes Journal* 7.1 (June 2023), pp. 65–77. ISSN: 2471-2906. DOI: 10.1094/pbiomes-02-22-0007-r. URL: http://dx.doi.org/10.1094/PBIOMES-02-22-0007-R.

[164] David Lähnemann et al. “Eleven grand challenges in single-cell data science”. In: *Genome Biology* 21.1 (Feb. 2020). ISSN: 1474-760X. DOI: 10.1186/s13059-020-1926-6. URL: http://dx.doi.org/10.1186/s13059-020-1926-6.

[165] Alexandre Lanau and Joshua J Waterfall. “pyALRA: python implementation of low-rank zero-preserving approximation of single cell RNA-seq”. In: *Bioinformatics Advances* 5.1 (Dec. 2024). Ed. by Shanfeng Zhu. ISSN: 2635-0041. DOI: 10.1093/bioadv/vbaf279. URL: http://dx.doi.org/10.1093/bioadv/vbaf279.

[166] Jeffrey M. Levsky and Robert H. Singer. “Fluorescence in situ hybridization: past, present and future”. In: *Journal of Cell Science* 116.14 (July 2003), pp. 2833–2838. ISSN: 0021-9533. DOI: 10.1242/jcs.00633. URL: http://dx.doi.org/10.1242/jcs.00633.

[167] Wei Vivian Li and Jingyi Jessica Li. “An accurate and robust imputation method scImpute for single-cell RNA-seq data”. In: *Nature Communications* 9.1 (Mar. 2018). ISSN: 2041-1723. DOI: 10.1038/s41467-018-03405-7. URL: http://dx.doi.org/10.1038/s41467-018-03405-7.

[168] George C. Linderman et al. “Zero-preserving imputation of single-cell RNA-seq data”. In: *Nature Communications* 13.1 (Jan. 2022). ISSN: 2041-1723. DOI: 10.1038/s41467-021-27729-z. URL: http://dx.doi.org/10.1038/s41467-021-27729-z.

[169] Ting Liu et al. “NF-$\kappa$B signaling in inflammation”. In: *Signal Transduction and Targeted Therapy* 2.1 (July 2017). ISSN: 2059-3635. DOI: 10.1038/sigtrans.2017.23. URL: http://dx.doi.org/10.1038/sigtrans.2017.23.

[170] Yang Liu et al. “High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue”. In: *Cell* 183.6 (Dec. 2020), 1665–1681.e18. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.10.026. URL: http://dx.doi.org/10.1016/j.cell.2020.10.026.

[171] Stuart Lloyd. "Least squares quantization in PCM". In: *IEEE transactions on information theory* 28.2 (1982), pp. 129–137.

[172] Sophia K. Longo et al. "Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics". In: *Nature Reviews Genetics* 22.10 (June 2021), pp. 627–644. ISSN: 1471-0064. DOI: 10.1038/s41576-021-00370-8. URL: http://dx.doi.org/10.1038/s41576-021-00370-8.

[173] Romain Lopez et al. *A joint model of unpaired data from scRNA-seq and spatial transcriptomics for imputing missing gene expression measurements*. 2019. DOI: 10.48550/ARXIV.1905.02269. URL: https://arxiv.org/abs/1905.02269.

[174] Romain Lopez et al. "Deep generative modeling for single-cell transcriptomics". In: *Nature Methods* 15.12 (Nov. 2018), pp. 1053–1058. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0229-2. URL: http://dx.doi.org/10.1038/s41592-018-0229-2.

[175] Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. "scGen predicts single-cell perturbation responses". In: *Nature Methods* 16.8 (July 2019), pp. 715–721. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0494-8. URL: http://dx.doi.org/10.1038/s41592-019-0494-8.

[176] Mohammad Lotfollahi et al. "Conditional out-of-distribution generation for unpaired data using transfer VAE". In: *Bioinformatics* 36.Supplement_2 (Dec. 2020), pp. i610–i617. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btaa800. URL: http://dx.doi.org/10.1093/bioinformatics/btaa800.

[177] Mohammad Lotfollahi et al. "The future of rapid and automated single-cell data analysis using reference mapping". In: *Cell* 187.10 (May 2024), pp. 2343–2358. ISSN: 0092-8674. DOI: 10.1016/j.cell.2024.03.009. URL: http://dx.doi.org/10.1016/j.cell.2024.03.009.

[178] Michael I Love, Wolfgang Huber, and Simon Anders. "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12 (Dec. 2014). ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8. URL: http://dx.doi.org/10.1186/s13059-014-0550-8.

[179] Malte D Luecken and Fabian J Theis. "Current best practices in single-cell RNA-seq analysis: a tutorial". In: *Molecular Systems Biology* 15.6 (June 2019). ISSN: 1744-4292. DOI: 10.15252/msb.20188746. URL: http://dx.doi.org/10.15252/msb.20188746.

[180] Malte D. Luecken, Fabian J. Theis, et al. *Raw Data Processing — Single-Cell Best Practices*. https://www.sc-best-practices.org/introduction/raw_data_processing.html. Accessed: 2025-06-13. 2023.

[181] Malte D. Luecken et al. "Benchmarking atlas-level data integration in single-cell genomics". In: *Nature Methods* 19.1 (Dec. 2021), pp. 41–50. ISSN: 1548-7105. DOI: 10.1038/s41592-021-01336-8. URL: http://dx.doi.org/10.1038/s41592-021-01336-8.

[182] Aaron T. L. Lun et al. "EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data". In: *Genome Biology* 20.1 (Mar. 2019). ISSN: 1474-760X. DOI: 10.1186/s13059-019-1662-y. URL: http://dx.doi.org/10.1186/s13059-019-1662-y.

[183] Ying Ma et al. "Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies". In: *Nature Communications* 11.1 (Mar. 2020). ISSN: 2041-1723. DOI: 10.1038/s41467-020-15298-6. URL: http://dx.doi.org/10.1038/s41467-020-15298-6.

[184] Evan Z. Macosko et al. "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". In: *Cell* 161.5 (May 2015), pp. 1202–1214. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.05.002. URL: http://dx.doi.org/10.1016/j.cell.2015.05.002.

[185] Shruti Magesh et al. "Disparities in COVID-19 Outcomes by Race, Ethnicity, and Socioeconomic Status: A Systematic Review and Meta-analysis". In: *JAMA Network Open* 4.11 (Nov. 2021), e2134147. ISSN: 2574-3805. DOI: 10.1001/jamanetworkopen.2021.34147. URL: http://dx.doi.org/10.1001/jamanetworkopen.2021.34147.

[186] Matteo Magnani et al. *Community Detection in Multiplex Networks*. 2019. DOI: 10.48550/ARXIV.1910.07646. URL: https://arxiv.org/abs/1910.07646.

[187] H. B. Mann and D. R. Whitney. "On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other". In: *The Annals of Mathematical Statistics* 18.1 (Mar. 1947), pp. 50–60. ISSN: 0003-4851. DOI: 10.1214/aoms/1177730491. URL: http://dx.doi.org/10.1214/aoms/1177730491.

[188] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet.journal* 17.1 (May 2011), p. 10. ISSN: 2226-6089. DOI: 10.14806/ej.17.1.200. URL: http://dx.doi.org/10.14806/ej.17.1.200.

[189] Vivien Marx. "Method of the Year: spatially resolved transcriptomics". In: *Nature Methods* 18.1 (Jan. 2021), pp. 9–14. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01033-y. URL: http://dx.doi.org/10.1038/s41592-020-01033-y.

[190] Vivien Marx. "The big challenges of big data". In: *Nature* 498.7453 (June 2013), pp. 255–260. ISSN: 1476-4687. DOI: 10.1038/498255a. URL: http://dx.doi.org/10.1038/498255a.

[191] Divij Mathew et al. "Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications". In: *Science* 369.6508 (Sept. 2020). ISSN: 1095-9203. DOI: 10.1126/science.abc8511. URL: http://dx.doi.org/10.1126/science.abc8511.

[192] Jakob McBroome et al. “A Daily-Updated Database and Tools for Comprehensive SARS-CoV-2 Mutation-Annotated Trees”. In: *Molecular Biology and Evolution* 38.12 (Sept. 2021). Ed. by Jian Lu, pp. 5819–5824. ISSN: 1537-1719. DOI: 10.1093/molbev/msab264. URL: http://dx.doi.org/10.1093/molbev/msab264.

[193] Christopher S. McGinnis, Lyndsay M. Murrow, and Zev J. Gartner. “DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors”. In: *Cell Systems* 8.4 (Apr. 2019), 329–337.e4. ISSN: 2405-4712. DOI: 10.1016/j.cels.2019.03.003. URL: http://dx.doi.org/10.1016/j.cels.2019.03.003.

[194] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[195] Páll Melsted et al. “Modular, efficient and constant-memory single-cell RNA-seq preprocessing”. In: *Nature Biotechnology* 39.7 (Apr. 2021), pp. 813–818. ISSN: 1546-1696. DOI: 10.1038/s41587-021-00870-2. URL: http://dx.doi.org/10.1038/s41587-021-00870-2.

[196] Jean Merlet et al. “Data-Driven Whole-Genome Clustering to Detect Geospatial, Temporal, and Functional Trends in SARS-CoV-2 Evolution”. In: *Proceedings of the Platform for Advanced Scientific Computing Conference*. PASC ’23. ACM, June 2023, pp. 1–7. DOI: 10.1145/3592979.3593425. URL: http://dx.doi.org/10.1145/3592979.3593425.

[197] Message Passing Interface Forum. *MPI: A Message-Passing Interface Standard Version 4.0*. https://www.mpi-forum.org/docs/mpi-4.0/mpi40-report.pdf. 2021.

[198] Yasutaka Mitamura et al. “Spatial transcriptomics combined with single-cell RNA-sequencing unravels the complex inflammatory cell network in atopic

dermatitis". In: *Allergy* 78.8 (June 2023), pp. 2215–2231. ISSN: 1398-9995. DOI: 10.1111/all.15781. URL: http://dx.doi.org/10.1111/all.15781.

[199] Petra Mlcochova et al. "SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion". In: *Nature* 599.7883 (Sept. 2021), pp. 114–119. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03944-y. URL: http://dx.doi.org/10.1038/s41586-021-03944-y.

[200] Kevin R. Moon et al. "Visualizing structure and transitions in high-dimensional biological data". In: *Nature Biotechnology* 37.12 (Dec. 2019), pp. 1482–1492. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0336-3. URL: http://dx.doi.org/10.1038/s41587-019-0336-3.

[201] Benoit Morel et al. "Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult". In: *Molecular Biology and Evolution* 38.5 (Dec. 2020). Ed. by Harmit Malik, pp. 1777–1791. ISSN: 1537-1719. DOI: 10.1093/molbev/msaa314. URL: http://dx.doi.org/10.1093/molbev/msaa314.

[202] Lambda Moses and Lior Pachter. "Museum of spatial transcriptomics". In: *Nature Methods* 19.5 (Mar. 2022), pp. 534–546. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01409-2. URL: http://dx.doi.org/10.1038/s41592-022-01409-2.

[203] Jonathan M Mudge et al. "GENCODE 2025: reference gene annotation for human and mouse". en. In: *Nucleic Acids Res.* 53.D1 (Jan. 2025), pp. D966–D975.

[204] Alexander Muik et al. "Neutralization of SARS-CoV-2 lineage B.1.1.7 pseudovirus by BNT162b2 vaccine–elicited human sera". In: *Science* 371.6534 (Mar. 2021), pp. 1152–1153. ISSN: 1095-9203. DOI: 10.1126/science.abg6105. URL: http://dx.doi.org/10.1126/science.abg6105.

[205] K. Mullis et al. "Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction". In: *Cold Spring Harbor Symposia on Quantitative*

*Biology* 51.0 (Jan. 1986), pp. 263–273. ISSN: 1943-4456. DOI: 10.1101/sqb.1986.051.01.032. URL: http://dx.doi.org/10.1101/sqb.1986.051.01.032.

[206] Ajit J. Nirmal and Peter K. Sorger. "SCIMAP: A Python Toolkit for Integrated Spatial Analysis of Multiplexed Imaging Data". In: *Journal of Open Source Software* 9.97 (May 2024), p. 6604. ISSN: 2475-9066. DOI: 10.21105/joss.06604. URL: http://dx.doi.org/10.21105/joss.06604.

[207] Roman Nowicki et al. "Atopic dermatitis. Interdisciplinary diagnostic and therapeutic recommendations of the Polish Dermatological Society, Polish Society of Allergology, Polish Pediatric Society and Polish Society of Family Medicine. Part I. Prophylaxis, topical treatment and phototherapy". In: *Advances in Dermatology and Allergology* 37.1 (2020), pp. 1–10. ISSN: 1642-395X. DOI: 10.5114/ada.2020.93423. URL: http://dx.doi.org/10.5114/ada.2020.93423.

[208] Kathryn M. Nowotny, Zinzi Bailey, and Lauren Brinkley-Rubinstein. "The Contribution of Prisons and Jails to US Racial Disparities During COVID-19". In: *American Journal of Public Health* 111.2 (Feb. 2021), pp. 197–199. ISSN: 1541-0048. DOI: 10.2105/ajph.2020.306040. URL: http://dx.doi.org/10.2105/AJPH.2020.306040.

[209] Áine O'Toole et al. "Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool". In: *Virus Evolution* 7.2 (July 2021). ISSN: 2057-1577. DOI: 10.1093/ve/veab064. URL: http://dx.doi.org/10.1093/ve/veab064.

[210] Áine O'Toole et al. "Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences". In: *BMC Genomics* 23.1 (Feb. 2022). ISSN: 1471-2164. DOI: 10.1186/s12864-022-08358-2. URL: http://dx.doi.org/10.1186/s12864-022-08358-2.

[211] Oak Ridge Leadership Computing Facility. *Andes User Guide*. https://docs.olcf.ornl.gov/systems/andes_user_guide.html. 2020.

[212] Oak Ridge Leadership Computing Facility. *Frontier User Guide*. https://docs.olcf.ornl.gov/systems/frontier_user_guide.html. 2024.

[213] Daniel Osorio and James J Cai. "Systematic determination of the mitochondrial proportion in human and mice tissues for single-cell RNA-sequencing data quality control". In: *Bioinformatics* 37.7 (Aug. 2020). Ed. by Anthony Mathelier, pp. 963–967. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btaa751. URL: http://dx.doi.org/10.1093/bioinformatics/btaa751.

[214] Mary Lou Pardue and Joseph G. Gall. "MOLECULAR HYBRIDIZATION OF RADIOACTIVE DNA TO THE DNA OF CYTOLOGICAL PREPARATIONS". In: *Proceedings of the National Academy of Sciences* 64.2 (Oct. 1969), pp. 600–604. ISSN: 1091-6490. DOI: 10.1073/pnas.64.2.600. URL: http://dx.doi.org/10.1073/pnas.64.2.600.

[215] Braja Gopal Patra et al. "A content-based literature recommendation system for datasets to improve data reusability – A case study on Gene Expression Omnibus (GEO) datasets". In: *Journal of Biomedical Informatics* 104 (Apr. 2020), p. 103399. ISSN: 1532-0464. DOI: 10.1016/j.jbi.2020.103399. URL: http://dx.doi.org/10.1016/j.jbi.2020.103399.

[216] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[217] Richard K. Perez et al. "Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus". In: *Science* 376.6589 (Apr. 2022). ISSN: 1095-9203. DOI: 10.1126/science.abf1970. URL: http://dx.doi.org/10.1126/science.abf1970.

[218] Yasset Perez-Riverol et al. "Quantifying the impact of public omics data". In: *Nature Communications* 10.1 (Aug. 2019). ISSN: 2041-1723. DOI: 10.1038/

s41467-019-11461-w. URL: http://dx.doi.org/10.1038/s41467-019-11461-w.

[219] Viktor Petukhov et al. "Cell segmentation in imaging-based spatial transcriptomics". In: *Nature Biotechnology* 40.3 (Oct. 2021), pp. 345–354. ISSN: 1546-1696. DOI: 10.1038/s41587-021-01044-w. URL: http://dx.doi.org/10.1038/s41587-021-01044-w.

[220] Arianna J. Piñeiro, Aubrey E. Houser, and Andrew L. Ji. "Research Techniques Made Simple: Spatial Transcriptomics". In: *Journal of Investigative Dermatology* 142.4 (Apr. 2022), 993–1001.e1. ISSN: 0022-202X. DOI: 10.1016/j.jid.2021.12.014. URL: http://dx.doi.org/10.1016/j.jid.2021.12.014.

[221] Léo Pio-Lopez et al. *MultiVERSE: a multiplex and multiplex-heterogeneous network embedding approach*. 2020. DOI: 10.48550/ARXIV.2008.10085. URL: https://arxiv.org/abs/2008.10085.

[222] Monika Piwecka, Nikolaus Rajewsky, and Agnieszka Rybak-Wolf. "Single-cell and spatial transcriptomics: deciphering brain complexity in health and disease". In: *Nature Reviews Neurology* 19.6 (May 2023), pp. 346–362. ISSN: 1759-4766. DOI: 10.1038/s41582-023-00809-y. URL: http://dx.doi.org/10.1038/s41582-023-00809-y.

[223] Jasmine T. Plummer et al. "Standardized metrics for assessment and reproducibility of imaging-based spatial transcriptomics datasets". In: *Nature Biotechnology* (Dec. 2025). ISSN: 1546-1696. DOI: 10.1038/s41587-025-02811-9. URL: http://dx.doi.org/10.1038/s41587-025-02811-9.

[224] Krzysztof Polański et al. "BBKNN: fast batch alignment of single cell transcriptomes". In: *Bioinformatics* 36.3 (Aug. 2019). Ed. by Bonnie Berger, pp. 964–965. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btz625. URL: http://dx.doi.org/10.1093/bioinformatics/btz625.

[225] Jeffrey M. Pullin and Davis J. McCarthy. "A comparison of marker gene selection methods for single-cell RNA sequencing data". In: *Genome Biology* 25.1 (Feb. 2024). ISSN: 1474-760X. DOI: 10.1186/s13059-024-03183-0. URL: http://dx.doi.org/10.1186/s13059-024-03183-0.

[226] Peng Qiu. "Embracing the dropouts in single-cell RNA-seq analysis". In: *Nature Communications* 11.1 (Mar. 2020). ISSN: 2041-1723. DOI: 10.1038/s41467-020-14976-9. URL: http://dx.doi.org/10.1038/s41467-020-14976-9.

[227] Andrew Rambaut et al. "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology". In: *Nature Microbiology* 5.11 (July 2020), pp. 1403–1407. ISSN: 2058-5276. DOI: 10.1038/s41564-020-0770-5. URL: http://dx.doi.org/10.1038/s41564-020-0770-5.

[228] Sindhu Ramesh et al. "Emerging SARS-CoV-2 Variants: A Review of Its Mutations, Its Implications and Vaccine Efficacy". In: *Vaccines* 9.10 (Oct. 2021), p. 1195. ISSN: 2076-393X. DOI: 10.3390/vaccines9101195. URL: http://dx.doi.org/10.3390/vaccines9101195.

[229] José A. Ramos-Vara. "Principles and Methods of Immunohistochemistry". In: *Drug Safety Evaluation*. Springer New York, 2017, pp. 115–128. ISBN: 9781493971725. DOI: 10.1007/978-1-4939-7172-5_5. URL: http://dx.doi.org/10.1007/978-1-4939-7172-5_5.

[230] Daniel Ramsköld et al. "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells". In: *Nature Biotechnology* 30.8 (July 2012), pp. 777–782. ISSN: 1546-1696. DOI: 10.1038/nbt.2282. URL: http://dx.doi.org/10.1038/nbt.2282.

[231] Anjali Rao et al. "Exploring tissue architecture using spatial transcriptomics". In: *Nature* 596.7871 (Aug. 2021), pp. 211–220. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03634-9. URL: http://dx.doi.org/10.1038/s41586-021-03634-9.

[232] Aviv Regev et al. "The Human Cell Atlas". In: *eLife* 6 (Dec. 2017). ISSN: 2050-084X. DOI: 10.7554/elife.27041. URL: http://dx.doi.org/10.7554/elife.27041.

[233] Gary Reynolds et al. "Developmental cell programs are co-opted in inflammatory skin disease". In: *Science* 371.6527 (Jan. 2021). ISSN: 1095-9203. DOI: 10.1126/science.aba6500. URL: http://dx.doi.org/10.1126/science.aba6500.

[234] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". In: *Bioinformatics* 26.1 (Nov. 2009), pp. 139–140. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp616. URL: http://dx.doi.org/10.1093/bioinformatics/btp616.

[235] Jennifer E. Rood et al. "The Human Cell Atlas from a cell census to a unified foundation model". In: *Nature* 637.8048 (Nov. 2024), pp. 1065–1071. ISSN: 1476-4687. DOI: 10.1038/s41586-024-08338-4. URL: http://dx.doi.org/10.1038/s41586-024-08338-4.

[236] Madison Russell et al. "Gene communities in co-expression networks across different tissues". In: *PLOS Computational Biology* 19.11 (Nov. 2023). Ed. by Feng Fu, e1011616. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1011616. URL: http://dx.doi.org/10.1371/journal.pcbi.1011616.

[237] Simona Saluzzo et al. "Processing human skin samples for single-cell assays". In: *STAR Protocols* 3.3 (Sept. 2022), p. 101470. ISSN: 2666-1667. DOI: 10.1016/j.xpro.2022.101470. URL: http://dx.doi.org/10.1016/j.xpro.2022.101470.

[238] F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain-terminating inhibitors". In: *Proceedings of the National Academy of Sciences* 74.12 (Dec. 1977), pp. 5463–5467. ISSN: 1091-6490. DOI: 10.1073/pnas.74.12.5463. URL: http://dx.doi.org/10.1073/pnas.74.12.5463.

[239] Michael C Schatz, Ben Langmead, and Steven L Salzberg. "Cloud computing and the DNA data race". In: *Nature Biotechnology* 28.7 (July 2010), pp. 691–693. ISSN: 1546-1696. DOI: 10.1038/nbt0710-691. URL: http://dx.doi.org/10.1038/nbt0710-691.

[240] Mark Schena et al. "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray". In: *Science* 270.5235 (Oct. 1995), pp. 467–470. ISSN: 1095-9203. DOI: 10.1126/science.270.5235.467. URL: http://dx.doi.org/10.1126/science.270.5235.467.

[241] Christian M. Schürch et al. "Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front". In: *Cell* 182.5 (Sept. 2020), 1341–1359.e19. ISSN: 0092-8674. DOI: 10.1016/j.cell.2020.07.005. URL: http://dx.doi.org/10.1016/j.cell.2020.07.005.

[242] Raman Sethi et al. "ezSingleCell: an integrated one-stop single-cell and spatial omics analysis platform for bench scientists". In: *Nature Communications* 15.1 (July 2024). ISSN: 2041-1723. DOI: 10.1038/s41467-024-48188-2. URL: http://dx.doi.org/10.1038/s41467-024-48188-2.

[243] Maxwell ER Shafer, Ahilya N Sawh, and Alexander F Schier. "Gene family evolution underlies cell-type diversification in the hypothalamus of teleosts". In: *Nature ecology & evolution* 6.1 (2022), pp. 63–76.

[244] Sheel Shah et al. "seqFISH Accurately Detects Transcripts in Single Cells and Reveals Robust Spatial Organization in the Hippocampus". In: *Neuron* 94.4 (May 2017), 752–758.e1. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2017.05.008. URL: http://dx.doi.org/10.1016/j.neuron.2017.05.008.

[245] Paul Shannon et al. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks". In: *Genome Research* 13.11 (Nov. 2003), pp. 2498–2504. ISSN: 1088-9051. DOI: 10.1101/gr.1239303. URL: http://dx.doi.org/10.1101/gr.1239303.

[246] Aziz Sheikh et al. "SARS-CoV-2 Delta VOC in Scotland: demographics, risk of hospital admission, and vaccine effectiveness". In: *The Lancet* 397.10293 (June 2021), pp. 2461–2462. ISSN: 0140-6736. DOI: 10.1016/s0140-6736(21)01358-1. URL: http://dx.doi.org/10.1016/S0140-6736(21)01358-1.

[247] Kimberly Siletti et al. "Transcriptomic diversity of cell types across the adult human brain". In: *Science* 382.6667 (Oct. 2023). ISSN: 1095-9203. DOI: 10.1126/science.add7046. URL: http://dx.doi.org/10.1126/science.add7046.

[248] Jordan W. Squair et al. "Confronting false discoveries in single-cell differential expression". In: *Nature Communications* 12.1 (Sept. 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-25960-2. URL: http://dx.doi.org/10.1038/s41467-021-25960-2.

[249] Avi Srivastava et al. "Alevin efficiently estimates accurate gene abundances from dscRNA-seq data". In: *Genome Biology* 20.1 (Mar. 2019). ISSN: 1474-760X. DOI: 10.1186/s13059-019-1670-y. URL: http://dx.doi.org/10.1186/s13059-019-1670-y.

[250] Patrik L. Ståhl et al. "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics". In: *Science* 353.6294 (July 2016), pp. 78–82. ISSN: 1095-9203. DOI: 10.1126/science.aaf2403. URL: http://dx.doi.org/10.1126/science.aaf2403.

[251] Stefan Stanojevic et al. "Computational methods for single-cell multi-omics integration and alignment". In: *Genomics, Proteomics & Bioinformatics* 20.5 (2022), pp. 836–849.

[252] Gillian K. SteelFisher, Robert J. Blendon, and Hannah Caporello. "An Uncertain Public — Encouraging Acceptance of Covid-19 Vaccines". In: *New England Journal of Medicine* 384.16 (Apr. 2021), pp. 1483–1487. ISSN: 1533-4406. DOI: 10.1056/nejmp2100351. URL: http://dx.doi.org/10.1056/NEJMp2100351.

[253] Zachary D. Stephens et al. “Big Data: Astronomical or Genomical?” In: *PLOS Biology* 13.7 (July 2015), e1002195. ISSN: 1545-7885. DOI: 10.1371/journal.pbio.1002195. URL: http://dx.doi.org/10.1371/journal.pbio.1002195.

[254] Carsen Stringer et al. “Cellpose: a generalist algorithm for cellular segmentation”. In: *Nature Methods* 18.1 (Dec. 2020), pp. 100–106. ISSN: 1548-7105. DOI: 10.1038/s41592-020-01018-x. URL: http://dx.doi.org/10.1038/s41592-020-01018-x.

[255] Tim Stuart et al. “Comprehensive Integration of Single-Cell Data”. In: *Cell* 177.7 (June 2019), 1888–1902.e21. ISSN: 0092-8674. DOI: 10.1016/j.cell.2019.05.031. URL: http://dx.doi.org/10.1016/j.cell.2019.05.031.

[256] Student. “The Probable Error of a Mean”. In: *Biometrika* 6.1 (Mar. 1908), p. 1. ISSN: 0006-3444. DOI: 10.2307/2331554. URL: http://dx.doi.org/10.2307/2331554.

[257] Kyle A. Sullivan et al. *MENTOR: Multiplex Embedding of Networks for Team-Based Omics Research*. July 2024. DOI: 10.1101/2024.07.17.603821. URL: http://dx.doi.org/10.1101/2024.07.17.603821.

[258] Valentine Svensson. “Droplet scRNA-seq is not zero-inflated”. In: *Nature Biotechnology* 38.2 (Jan. 2020), pp. 147–150. ISSN: 1546-1696. DOI: 10.1038/s41587-019-0379-5. URL: http://dx.doi.org/10.1038/s41587-019-0379-5.

[259] Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. “A curated database reveals trends in single-cell transcriptomics”. In: *Database* 2020 (2020). ISSN: 1758-0463. DOI: 10.1093/database/baaa073. URL: http://dx.doi.org/10.1093/database/baaa073.

[260] Yuqi Tan and Patrick Cahan. “SingleCellNet: a computational tool to classify single cell RNA-Seq data across platforms and across species”. In: *Cell systems* 9.2 (2019), pp. 207–213.

[261] Fuchou Tang et al. “mRNA-Seq whole-transcriptome analysis of a single cell”. In: *Nature Methods* 6.5 (Apr. 2009), pp. 377–382. ISSN: 1548-7105. DOI: 10.1038/nmeth.1315. URL: http://dx.doi.org/10.1038/nmeth.1315.

[262] Alejandro Tejada-Lapuerta et al. “Nicheformer: a foundation model for single-cell and spatial omics”. In: *Nature Methods* 22.12 (Oct. 2025), pp. 2525–2538. ISSN: 1548-7105. DOI: 10.1038/s41592-025-02814-z. URL: http://dx.doi.org/10.1038/s41592-025-02814-z.

[263] Marina Terekhova et al. “Single-cell atlas of healthy human blood unveils age-related loss of NKG2C+GZMB-CD8+ memory T cells and accumulation of type 2 memory T cells”. In: *Immunity* 56.12 (Dec. 2023), 2836–2854.e9. ISSN: 1074-7613. DOI: 10.1016/j.immuni.2023.10.013. URL: http://dx.doi.org/10.1016/j.immuni.2023.10.013.

[264] Maria Antonietta Tosches et al. “Evolution of pallium, hippocampus, and cortical cell types revealed by single-cell transcriptomics in reptiles”. In: *Science* 360.6391 (2018), pp. 881–888.

[265] F. William Townes et al. “Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model”. In: *Genome Biology* 20.1 (Dec. 2019). ISSN: 1474-760X. DOI: 10.1186/s13059-019-1861-6. URL: http://dx.doi.org/10.1186/s13059-019-1861-6.

[266] Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. “From Louvain to Leiden: guaranteeing well-connected communities”. In: *Scientific reports* 9.1 (2019), pp. 1–12.

[267] Anthony Trewavas. “A Brief History of Systems Biology: “Every object that biology studies is a system of systems.” Francois Jacob (1974).” In: *The Plant Cell* 18.10 (Oct. 2006), pp. 2420–2430. ISSN: 1532-298X. DOI: 10.1105/tpc.106.042267. URL: http://dx.doi.org/10.1105/tpc.106.042267.

[268] Yatish Turakhia et al. “Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic”. In: *Nature Genetics* 53.6 (May 2021), pp. 809–816. ISSN: 1546-1718. DOI: 10.1038/s41588-021-00862-7. URL: http://dx.doi.org/10.1038/s41588-021-00862-7.

[269] Bram Van de Sande et al. “Applications of single-cell RNA sequencing in drug discovery and development”. In: *Nature Reviews Drug Discovery* 22.6 (Apr. 2023), pp. 496–520. ISSN: 1474-1784. DOI: 10.1038/s41573-023-00688-4. URL: http://dx.doi.org/10.1038/s41573-023-00688-4.

[270] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[271] Stijn Van Dongen. “Graph Clustering Via a Discrete Uncoupling Process”. In: *SIAM Journal on Matrix Analysis and Applications* 30.1 (Jan. 2008), pp. 121–141. ISSN: 1095-7162. DOI: 10.1137/040608635. URL: http://dx.doi.org/10.1137/040608635.

[272] Isaac Virshup et al. “anndata: Access and store annotated data matrices”. In: *Journal of Open Source Software* 9.101 (Sept. 2024), p. 4371. ISSN: 2475-9066. DOI: 10.21105/joss.04371. URL: http://dx.doi.org/10.21105/joss.04371.

[273] Pauli Virtanen et al. “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”. In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[274] Giovanni Visonà et al. “Network propagation for GWAS analysis: a practical guide to leveraging molecular networks for disease gene discovery”. In: *Briefings in Bioinformatics* 25.2 (Jan. 2024). ISSN: 1477-4054. DOI: 10.1093/bib/bbae014. URL: http://dx.doi.org/10.1093/bib/bbae014.

[275] Megan Wallace et al. "The Advisory Committee on Immunization Practices' Interim Recommendation for Use of Pfizer-BioNTech COVID-19 Vaccine in Adolescents Aged 12–15 Years — United States, May 2021". In: *MMWR. Morbidity and Mortality Weekly Report* 70.20 (May 2021). ISSN: 1545-861X. DOI: 10.15585/mmwr.mm7020e1. URL: http://dx.doi.org/10.15585/mmwr.mm7020e1.

[276] Bang Wang et al. "A graph-enhanced attention model for community detection in multiplex networks". In: *Expert Systems with Applications* 230 (Nov. 2023), p. 120552. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2023.120552. URL: http://dx.doi.org/10.1016/j.eswa.2023.120552.

[277] Tianyu Wang et al. "Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data". In: *BMC Bioinformatics* 20.1 (Jan. 2019). ISSN: 1471-2105. DOI: 10.1186/s12859-019-2599-6. URL: http://dx.doi.org/10.1186/s12859-019-2599-6.

[278] Yi Wang, Bingqing Xia, and Zhaobing Gao. "A comprehensive review of current insights into the virulence factors of SARS-CoV-2". In: *Journal of Virology* 99.2 (Feb. 2025). Ed. by Herman W. Favoreel. ISSN: 1098-5514. DOI: 10.1128/jvi.02049-24. URL: http://dx.doi.org/10.1128/jvi.02049-24.

[279] Joe H. Ward. "Hierarchical Grouping to Optimize an Objective Function". In: *Journal of the American Statistical Association* 58.301 (Mar. 1963), pp. 236–244. ISSN: 1537-274X. DOI: 10.1080/01621459.1963.10500845. URL: http://dx.doi.org/10.1080/01621459.1963.10500845.

[280] Andreas P.M. Weber et al. "Sampling the Arabidopsis Transcriptome with Massively Parallel Pyrosequencing". In: *Plant Physiology* 144.1 (Mar. 2007), pp. 32–42. ISSN: 1532-2548. DOI: 10.1104/pp.107.096677. URL: http://dx.doi.org/10.1104/pp.107.096677.

[281] Stephan Weidinger et al. "Atopic dermatitis". In: *Nature Reviews Disease Primers* 4.1 (June 2018). ISSN: 2056-676X. DOI: 10.1038/s41572-018-0001-z. URL: http://dx.doi.org/10.1038/s41572-018-0001-z.

[282] Joel O. Wertheim et al. "The Global Transmission Network of HIV-1". In: *The Journal of Infectious Diseases* 209.2 (Oct. 2013), pp. 304–313. ISSN: 0022-1899. DOI: 10.1093/infdis/jit524. URL: http://dx.doi.org/10.1093/infdis/jit524.

[283] Elizabeth J. Williamson et al. "Factors associated with COVID-19-related death using OpenSAFELY". In: *Nature* 584.7821 (July 2020), pp. 430–436. ISSN: 1476-4687. DOI: 10.1038/s41586-020-2521-4. URL: http://dx.doi.org/10.1038/s41586-020-2521-4.

[284] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome Biology* 19.1 (Feb. 2018). ISSN: 1474-760X. DOI: 10.1186/s13059-017-1382-0. URL: http://dx.doi.org/10.1186/s13059-017-1382-0.

[285] Markus Wolfien et al. "Single-Nucleus Sequencing of an Entire Mammalian Heart: Cell Type Composition and Velocity". In: *Cells* 9.2 (Jan. 2020), p. 318. ISSN: 2073-4409. DOI: 10.3390/cells9020318. URL: http://dx.doi.org/10.3390/cells9020318.

[286] Samuel L. Wolock, Romain Lopez, and Allon M. Klein. "Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data". In: *Cell Systems* 8.4 (Apr. 2019), 281–291.e9. ISSN: 2405-4712. DOI: 10.1016/j.cels.2018.11.005. URL: http://dx.doi.org/10.1016/j.cels.2018.11.005.

[287] World Health Organization. *WHO COVID-19 Dashboard: Cumulative Reported Deaths*. https://data.who.int/dashboards/covid19/deaths. Accessed: 2025-06-05. 2024.

[288] Jingwen Yan et al. "Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data". In: *Briefings in Bioinformatics* 19.6 (June 2017), pp. 1370–1381. ISSN: 1477-4054. DOI: 10.1093/bib/bbx066. URL: https://doi.org/10.1093/bib/bbx066.

[289] Rongrong Yang et al. "Systematic evaluation of computational methods for cell segmentation". In: *Briefings in Bioinformatics* 27.1 (Jan. 2026). ISSN: 1477-4054. DOI: 10.1093/bib/bbag066. URL: http://dx.doi.org/10.1093/bib/bbag066.

[290] Zizhen Yao et al. "A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain". In: *Nature* 624.7991 (Dec. 2023), pp. 317–332. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06812-z. URL: http://dx.doi.org/10.1038/s41586-023-06812-z.

[291] Matthew D Young and Sam Behjati. "SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data". In: *GigaScience* 9.12 (Dec. 2020). ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa151. URL: http://dx.doi.org/10.1093/gigascience/giaa151.

[292] Matei Zaharia et al. "Apache Spark: a unified engine for big data processing". In: *Communications of the ACM* 59.11 (Oct. 2016), pp. 56–65. ISSN: 1557-7317. DOI: 10.1145/2934664. URL: http://dx.doi.org/10.1145/2934664.

[293] Luke Zappia, Belinda Phipson, and Alicia Oshlack. "Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database". In: *PLOS Computational Biology* 14.6 (June 2018). Ed. by Dina Schneidman, e1006245. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006245. URL: http://dx.doi.org/10.1371/journal.pcbi.1006245.

[294] Luke Zappia and Fabian J. Theis. "Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape". In: *Genome Biology* 22.1 (Oct. 2021). ISSN: 1474-760X. DOI: 10.1186/s13059-021-02519-4. URL: http://dx.doi.org/10.1186/s13059-021-02519-4.

[295] Amit Zeisel et al. "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq". In: *Science* 347.6226 (Mar. 2015), pp. 1138–1142. ISSN: 1095-9203. DOI: 10.1126/science.aaa1934. URL: http://dx.doi.org/10.1126/science.aaa1934.

[296] Allen W. Zhang et al. "Probabilistic cell-type assignment of single-cell RNA-seq for tumor microenvironment profiling". In: *Nature Methods* 16.10 (Sept. 2019), pp. 1007–1015. ISSN: 1548-7105. DOI: 10.1038/s41592-019-0529-1. URL: http://dx.doi.org/10.1038/s41592-019-0529-1.

[297] Qian Zhang, Michael J Lenardo, and David Baltimore. "30 years of NF-$\kappa$B: a blossoming of relevance to human pathobiology". In: *Cell* 168.1 (2017), pp. 37–57.

[298] Grace X. Y. Zheng et al. "Massively parallel digital transcriptional profiling of single cells". In: *Nature Communications* 8.1 (Jan. 2017). ISSN: 2041-1723. DOI: 10.1038/ncomms14049. URL: http://dx.doi.org/10.1038/ncomms14049.

[299] Grace X. Y. Zheng et al. "Massively parallel digital transcriptional profiling of single cells". In: *Nature Communications* 8.1 (Jan. 2017). ISSN: 2041-1723. DOI: 10.1038/ncomms14049. URL: http://dx.doi.org/10.1038/ncomms14049.

[300] Andrzej Zielezinski et al. "Alignment-free sequence comparison: benefits, applications, and tools". In: *Genome Biology* 18.1 (Oct. 2017). ISSN: 1474-760X. DOI: 10.1186/s13059-017-1319-7. URL: http://dx.doi.org/10.1186/s13059-017-1319-7.

[301] Rapolas Zilionis et al. "Single-cell barcoding and sequencing using droplet microfluidics". In: *Nature Protocols* 12.1 (Dec. 2016), pp. 44–73. ISSN: 1750-2799. DOI: 10.1038/nprot.2016.154. URL: http://dx.doi.org/10.1038/nprot.2016.154.

[302] Kip D. Zimmerman, Mark A. Espeland, and Carl D. Langefeld. "A practical solution to pseudoreplication bias in single-cell studies". In: *Nature Communications* 12.1 (Feb. 2021). ISSN: 2041-1723. DOI: 10.1038/s41467-021-21038-1. URL: http://dx.doi.org/10.1038/s41467-021-21038-1.

[303] Marinka Zitnik et al. "Current and future directions in network biology". In: *Bioinformatics Advances* 4.1 (2024). Ed. by Thomas Lengauer. ISSN: 2635-0041. DOI: 10.1093/bioadv/vbae099. URL: http://dx.doi.org/10.1093/bioadv/vbae099.

[304] Marinka Zitnik et al. "Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities". In: *Information Fusion* 50 (Oct. 2019), pp. 71–91. ISSN: 1566-2535. DOI: 10.1016/j.inffus.2018.09.012. URL: http://dx.doi.org/10.1016/j.inffus.2018.09.012.

[305] Justina žurauskienė and Christopher Yau. "pcaReduce: hierarchical clustering of single cell transcriptional profiles". In: *BMC Bioinformatics* 17.1 (Mar. 2016). ISSN: 1471-2105. DOI: 10.1186/s12859-016-0984-y. URL: http://dx.doi.org/10.1186/s12859-016-0984-y.

# Appendix

**Table 1:** Summary statistics for the ten largest SARS-CoV-2 HipMCL genomic clusters identified across California from 2020-W15 through 2022-W06.

| Cluster | Seqs. | Counties | Wks. | First Wk | Last Wk | Frac. | Variant |
|---|---|---|---|---|---|---|---|
| 0 | 47730 | 57 | 50 | 2020-W41 | 2022-W11 | 0.31 | Delta |
| 2 | 34606 | 54 | 85 | 2020-W04 | 2022-W04 | 0.22 | none |
| 4 | 27773 | 48 | 33 | 2020-W53 | 2022-W18 | 0.18 | Omicron |
| 7 | 11832 | 55 | 47 | 2020-W04 | 2021-W37 | 0.08 | Epsilon |
| 1 | 6579 | 53 | 43 | 2020-W15 | 2021-W37 | 0.04 | Epsilon |
| 3 | 6509 | 54 | 39 | 2021-W14 | 2021-W52 | 0.04 | Delta |
| 10 | 6255 | 52 | 41 | 2020-W15 | 2021-W37 | 0.04 | Alpha |
| 6 | 5570 | 54 | 43 | 2021-W13 | 2022-W03 | 0.04 | Delta |
| 8 | 2715 | 41 | 21 | 2021-W52 | 2022-W20 | 0.02 | Omicron |
| 9 | 1517 | 50 | 24 | 2021-W05 | 2021-W31 | 0.01 | Alpha |

**Table 2:** Canonical marker genes used contributing to Xenium cell type annotation. Cell type assignments were based on the combined expression patterns of canonical lineage-associated genes together with spatial localization and top DEGs between clusters.

| **Cell type** | **Representative marker genes** |
|---|---|
| Keratinocytes | *Krt1*, *Krt5*, *Krt15*, *Krt6b*, *Dsc1*, *Dsg1a*, *Dsg1b*, *Dsg3* |
| Fibroblasts | *Pdgfra*, *Col5a2*, *Col6a1* |
| Lymphocytes / Dendritic cells | *Cd3d*, *Cd3e*, *Cd8a*, *Cd8b1*, *Blk*, *Gzma*, *Prf1*, *Gzmb*, *Cd69*, *Clec9a* |
| Macrophages / Monocytes | *Cd68*, *C1qa*, *C5ar1*, *Fcer1g* |
| Granulocytes | *Ear6*, *Ccl24*, *Cxcr2*, *Il1b*, *Cxcl2*, *Cxcl3*, *Alox15* |

**Table 3:** Summary benchmarking results comparing HALRA and pyALRA across increasing scRNA-seq dataset sizes. Runtime (RT) is reported in minutes and peak memory utilization in gigabytes (GB). Nnz in and Nnz out correspond to input and output nonzero entry density, respectively.

| Method | Cells | Peak Mem. (GB) | RT (min) | Nnz in (%) | Nnz out (%) |
|---|---|---|---|---|---|
| HALRA | 10,000 | 4.33 | 0.62 | 2.27 | 24.53 |
| pyALRA | 10,000 | 16.64 | 0.85 | 2.27 | 24.53 |
| HALRA | 25,000 | 10.37 | 1.56 | 2.28 | 26.65 |
| pyALRA | 25,000 | 39.78 | 1.82 | 2.28 | 26.65 |
| HALRA | 50,000 | 20.82 | 5.47 | 2.32 | 28.70 |
| pyALRA | 50,000 | 80.48 | 3.32 | 2.32 | 28.70 |
| HALRA | 100,000 | 48.66 | 14.85 | 2.39 | 30.48 |
| pyALRA | 100,000 | 186.16 | 7.58 | 2.39 | 30.48 |
| HALRA | 200,000 | 112.51 | 40.79 | 2.51 | 31.93 |
| pyALRA | 200,000 | 428.94 | 20.84 | 2.51 | 31.93 |
| HALRA | 300,000 | 177.48 | 69.39 | 2.65 | 32.62 |
| pyALRA | 300,000 | 681.38 | 36.83 | 2.65 | 32.62 |
| HALRA | 400,000 | 241.66 | 101.93 | 2.80 | 33.09 |
| pyALRA | 400,000 | 924.69 | 59.17 | 2.80 | 33.09 |

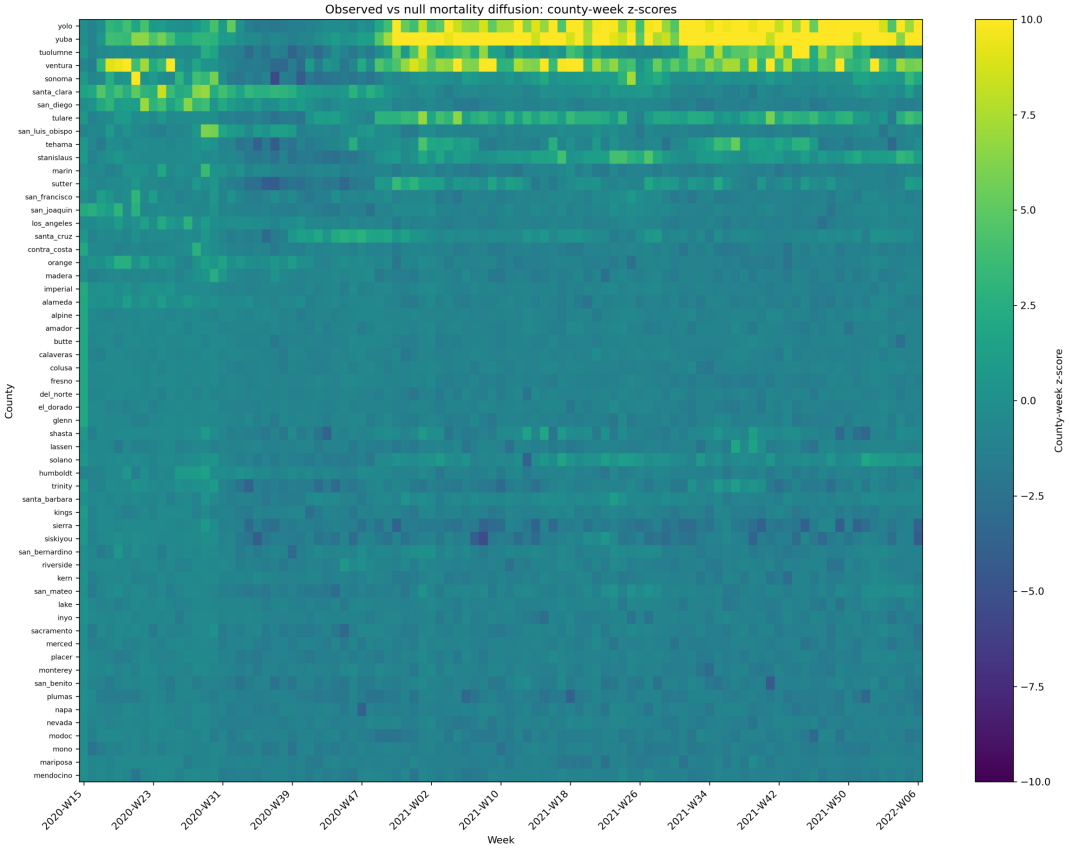



**Figure 1:** County-week heatmap of observed versus null mortality diffusion z-scores across the statewide SARS-CoV-2 multiplex network from 2020-W15 through 2022-W06. Each cell represents the standardized difference between observed mortality diffusion and the mean diffusion obtained from within-layer-preserving multiplex null models, in which temporal, interlayer, and lagged genomic–mortality coupling edges were randomized while preserving intralayer topology. Positive z-scores indicate county-weeks in which the observed multiplex structure concentrated mortality diffusion more strongly than expected under randomized cross-context coupling. Enrichment was spatially and temporally localized rather than uniformly distributed statewide, with persistent elevated regions observed in specific counties and pandemic intervals.

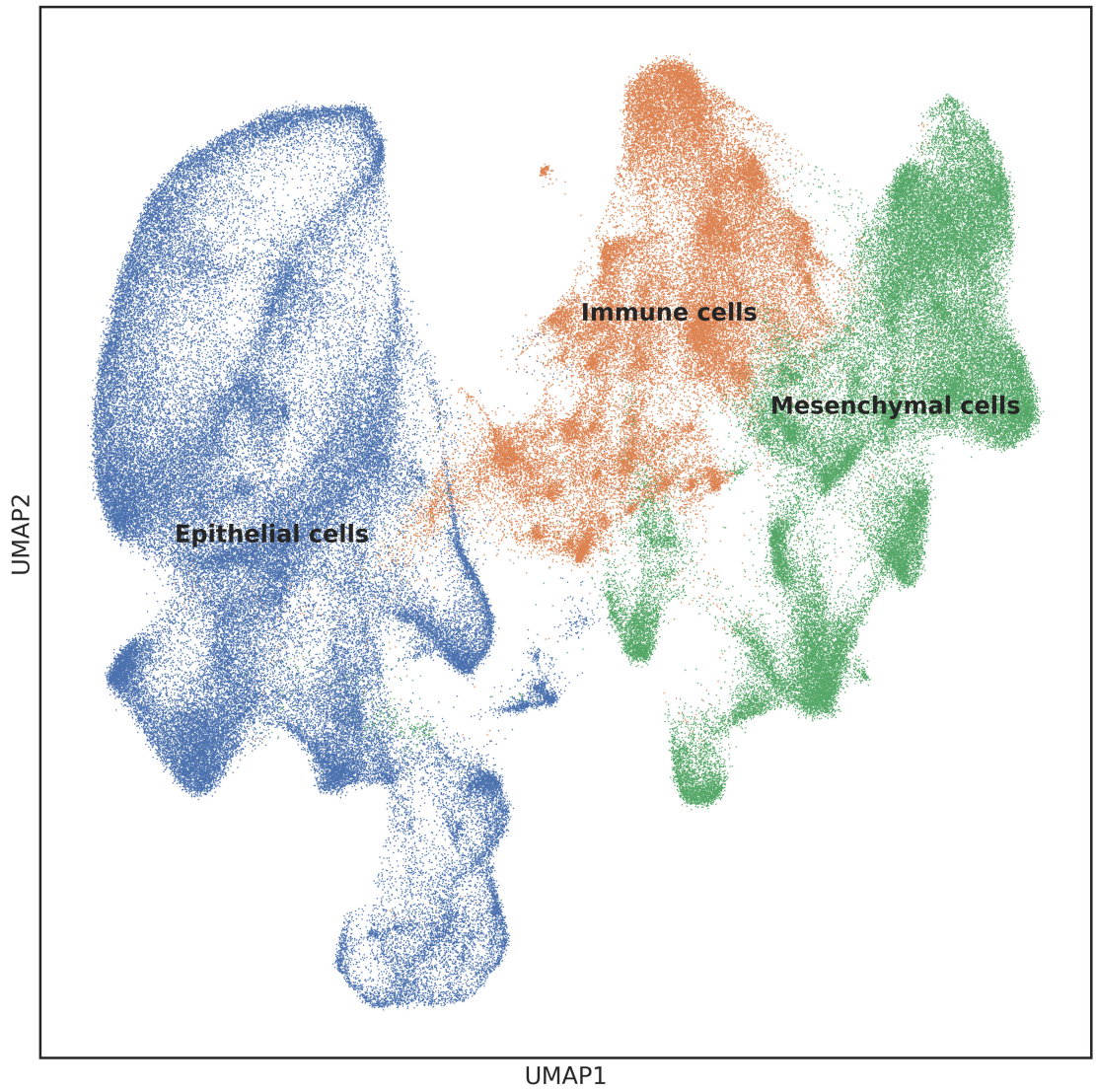



**Figure 2:** UMAP visualization of the three major transcriptomic cell populations in murine single-cell spatial transcriptomic data. Leiden clustering (resolution = 0.6) resulted in 21 clusters, generated from Xenium single-cell spatial transcriptomic data using only nuclear-overlapping transcripts (81% of all detected transcripts). These clusters were manually annotated based on canonical marker gene expression and merged into three broad cell types, identifying the epithelial, immune, and mesenchymal populations.

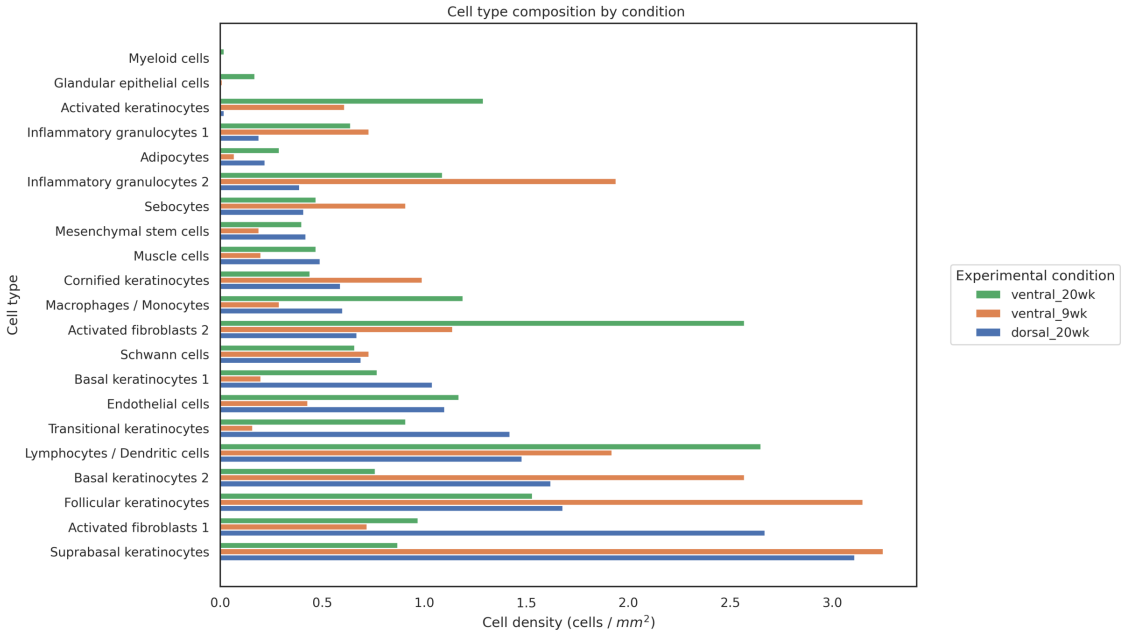



**Figure 3:** Area-normalized cell type composition across murine Xenium samples. Bar plots show the density of annotated cell populations across dorsal (control) and ventral (experimental) skin samples at 9 and 20 weeks of age. Cell type abundances were normalized by total segmented tissue area (cells/mm$^2$) to account for differences in tissue size between sections. Inflammatory and activated stromal populations were enriched in ventral samples relative to dorsal controls, consistent with localized inflammatory remodeling in the murine atopic dermatitis model.

# Vita

Jean Merlet was born in France, where he lived until the age of six before moving to Massachusetts. He attended high school in Newton, Massachusetts, and later earned a Bachelor of Science degree in Mathematics from Wake Forest University. He subsequently joined the Bredesen Center for Interdisciplinary Research and Graduate Education at University of Tennessee, Knoxville, where he pursued a Doctor of Philosophy degree in Data Science and Engineering under the advisement of Daniel Jacobson.

During his doctoral studies, Merlet conducted research at Oak Ridge National Laboratory focused on scalable computational biology, high-performance computing, and multi-omic data integration. His dissertation, Scalable Integrative Frameworks for Multi-Omic Systems Biology, explored distributed methods for large-scale biological network analysis, transcriptomic data integration, and scalable analysis of high-dimensional omic datasets. His work included the development of distributed workflows for single-cell and spatial transcriptomics as well as large-scale comparative viral genomics.

Merlet presented his research at the 2023 Platform for Advanced Scientific Computing Conference and contributed to multiple collaborative biomedical research projects and peer-reviewed publications during his graduate studies. Following completion of his doctoral degree, he plans to continue working at the intersection of computational biology, scalable machine learning, and translational multi-omic research.