Atlas level scRNAseq analysis reveals the functional landscape of cell types

Benjamin D. HARRIS

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy

Cold Spring Harbor School of Biological Sciences
Cold Spring Harbor Laboratory

October 28, 2021

"*Careful. We don't want to learn from this.*"

Bill Watterson, "Calvin and Hobbes"

Acknowledgements

First and foremost I want to thank my thesis supervisor, Jesse Gillis. His support through the highs and lows of my research was unwavering. He allowed me a lot of independence to explore my own ideas, which really helped me grow as a scientist. I need to thank the current and past members of the Gillis lab. In particular Maggie Crow, Stephan Fischer, and Sara Ballouz, who's experience and knowledge were instrumental in my success.

I would also like to thank my thesis committee David McCandlish, David Jackson, and Richard McCombie for all of the feedback and support throughout the past 4 years.

Thank you to the CSHL School of Biological Sciences staff, for all their help throughout my time here. In particular, many thanks to Alyson Kass-Eisler, Kim Geer, Kim Creteur, Monn Monn Myat, and Alex Gann. Their support makes this an such a great enviornment to grow as a scientist.

I also need to thank all my previous mentors and reccommenders who's training and support was instrumental in me succeding here. Dr. William Simonds (NIH), who allowed me to fail for 3 summers of wet lab expermients in his lab, and my mentors Dr. Mritunjay Pandey and Dr. Jian-Hua Zhang in his lab. Dr. Ahmet Ay (Colgate University) who was my academic mentor and supported me for my undergraduate thesis. And Dr. Mickey Atwal who mentored me during my summer as an URP here.

I wouldn't be here without the CSHL URP 2016 summer. I talk to many of my fellow URPs reguarly and consider some of them my closest friends and colleagues. Specically, Toby Aicher, Daniel Barabasi, Chris Giuliano, and Ajay Nadig.

I also need to recognize all the friends I made at CSHL. All the students, technicians, post docs, and faculty make this isolated research institution habitable. Activities like frisbee, volleyball, and other social events were a needed break from the diffiulty of research, and getting to know so many amazing people at these activities was such a treat. I've made it no secret that I despise Long Island, so I need to thank all the people who let me crash on their couch over the past 4 years: Ryan, Brendan, Missy, Chris, Charlie, Zach, Kamal, and of course Mother Nature (camping).

Most importantly, I need to thank my family for all their love and support. No one was prouder of me doing a PhD than my mom. And the only thing harder than this PhD was saying goodbye to her. This is work dedicated to her.

List of Figures

List of Tables

List of Abbreviations

scRNAseq	Single Cell RNA sequencing
AUROC	**A**rea **U**nder the **R**eciever **O**perating **C**haracteristic
GO	**G**ene **O**ntology
KEGG	**K**yoto **E**ncyclopedia of **G**enes and **G**enomes
LT-HSC	**L**ong **T**erm **H**ematopoietic **S**tem **C**ell
ST-HSC	**S**hort **T**erm **H**ematopoietic **S**tem **C**ell
MPP2-4	**M**ulti**p**otent **P**rogenitor
BICCN	**B**rain **I**nitiative **C**ell **C**ensus **N**etwork
EGAD	**E**xtending '**G**uilt-by-**A**ssociation' by **D**egree
UMAP	**U**niform **M**anifold **A**pproximation and **P**rojection
tSNE	**t**-distributed **S**tochastic **N**eighbor **E**mbedding
FACS	**F**luorescent **A**ctivated **C**ell **S**orting
FDR	**F**alse **D**iscovery **R**ate

Dedicated to my mom. . .

Contents

Chapter 1

Introduction

Characterizing the diversity and function of the many cells that make up living organisms is an essential aspect of cell biology. This thesis uses meta-analysis of large-scale single cell RNA sequencing (scRNAseq) data to both explore the diversity of cell types and the functional landscape that defines them. My methodology is based on meta-analysis because it serves to find robust signatures across datasets with significant technical variation, thereby determining what signatures and properties are likely to generalize to new data. After the introduction, I present two vignettes that explore the functional landscape of hematopoiesis Chapter 2 and neurons Chapter 3 using this philosophy. In Chapter 4 I summarize my results and provide a perspective on how the field can build on this work. Finally, Appendix A and B are two manuscripts I contributed to that are relevant, but not central to this work.

In this chapter, I begin by discussing the history of cell biology and how certain technological advances, up to the development of scRNAseq, influence our ability to characterize cell types in living organisms Chapter 1.1. I highlight key technologies used to study cell identity before the development of scRNAseq. This discussion is important for understanding the objectives of the field and how scRNAseq can be effectively applied to important questions in cell biology. Additionally, integrating the information learned from older technologies with scRNAseq data is critical for interpreting scRNAseq data and keeping the analysis relevant to the field.

After discussing the history of the field, I summarize the development and importance of scRNAseq as a technology (Chapter 1.2). I describe how the technology works and review the growth over the past 10 years. Additionally, I highlight the specific contributions of scRNAseq to the fields of hematopoiesis and neuroscience. In Chapter 1.3 I discuss some computational methods important to my analysis: preprocessing (Chapter 1.3.1), co-expression (Chapter 1.3.2), and pseudotime analysis (Chapter 1.3.3).

The final section of the introduction (Chapter 1.4) discusses the areas where this work makes contributions to the field by analyzing scRNAseq data through a specific philosophy of meta-analysis. Sparsity and inter-batch variation represent two major technical challenges to the fruitful analysis of single-cell data. I discuss how I handle these challenges to learn replicable and generalizable gene signatures (Chapter 1.4.1). An important aspect of my approach is meta-analysis, analyzing scRNAseq data across datasets. The most popular methods for ameliorating technical variation rely on batch correction or integration. I discuss the differences and merits of both integration and meta-analytic approaches to scRNAseq analysis in Chapter 1.4.2. For all the analyses, except for one, I rely on meta-analysis and discuss why I prefer meta-analysis, and explain exceptional instances that require integration. This leads us to the description of my work studying hematopoietic stem and progenitor cell types (Chapter 2) and co-expression in neurons (Chapter 3).

1.1 Cell Biology: Characterization of Cellular Diversity and Function

For hundreds of years, biologists have been characterizing the diversity and function of cells. Robert Hooke coined the term "cell" to describe the smallest unit in cork in 1665, and Anton Van Leeuwenhoek first visualized cells on a microscope in 1674. From these discoveries, the field of cell biology was born. In 1838, Theodore Schwann formulated the cell theory with 3 observations, of which the first two were correct: 1) The cell is the unit of structure, physiology, and organization in living organisms. 2) The cell remains a dual existence as a distinct entity and a building block in the construction of organisms. 3) Cells form by free-cell formation. A few years after Schwann's theory was published,

Rudolf Virchow said "Omnis cellula e cellula", meaning that all cells arise from preexisting cells, correctly contradicting Schwann's third point (Ribatti, 2018). By looking at the parts of a car and studying the function of each, we can learn a lot about how a car works. Cell biology studies the building blocks of life so we can explain how the individual units coalesce into a complex organism.

As technology develops, particularly with improvements in microscopy, our understanding of cellular diversity has grown. Using a state-of-the-art microscope at the end of the 19th century, Santiago Ramón y Cajal researched Golgi-stained neurons and free-hand drew the various cells he saw. These drawings remain some of the most recognizable images of individual cells today. He characterized the cells based on morphology, how they looked. Based on his observations, he expands Schwann's cell theory to develop the neuron doctrine, stating that the nervous system is built of autonomous cells connected by synapses (López-Muñoz, Boya, and Alamo, 2006). A large percentage of the work in neuroscience following Cajal's Nobel Prize-winning work attempts to further characterize the diversity of neuronal cell types and how the varying functions of the various neuron types lead to the function of the entire nervous system. There are a variety of features to study when characterizing the morphology of neurons, but the size and number of axons and dendrites, two cellular structures relevant to neuronal connectivity, are critical to understanding the function of different neuronal populations (Schubert et al., 2003; Chklovskii, 2004; Oberlaender et al., 2011).

Characterizing cells based on morphology has been exceptionally useful. Diseases such as sickle-cell anemia can be diagnosed based on abnormal morphology alone. For some systems, more recently developed technology resolves limitations with studying morphology through microscopes; specifically lower throughput and subjectivity in interpreting samples on the microscope's stage. After the invention of the microscope, the next major technological advancement in characterizing cell types was flow cytometry. Flow cytometry measures a given phenotype in thousands or millions of cells by passing them through a laser that measures the either emitted or scattered light. The first flow cytometers measured cell volume in red blood cells by measuring the amount of light scattered off the

surface of the cell passing through the laser in two dimensions. In the present day, flow cytometers are used across domains to sort cells tagged with fluorescent markers, commonly referred to as fluorescent activated cell sorting (FACS) (Fulwyler, 1965). By quantifying a phenotype in a suspension of cells, size, or fluorescence, we remove some of the subjectivity, you still have to assign thresholds, and significantly increase the throughput. FACS has been particularly impactful to the field of hematology (Orkin and Zon, 2008; Guo et al., 2013; Orkin, 2000). For example, two classes of T cells, helper T and killer T were distinguished using FACS using the proteins CD4 (helper) and CD8 (killer) (Mosmann et al., 1986).

In addition to evolving cell type identification, the development of FACS advanced the methods for evaluating the function of sorted cell populations. For hematopoietic stem and progenitor cells, their role is to produce differentiated cells. Thus researchers have defined the function of a specified progenitor cell type by the potential to differentiate into a specific lineage (Nilsson, Pronk, and Bryder, 2015). To test this, an assay will take sorted populations of cells and grow them in culture or label and transplant them into another mouse. After allowing the cells to differentiate, in either the culture or recipient mouse, cells are re-sorted with FACS to identify the cell types that the progenitors produced. For example, transplantation experiments show that the Multipotent Progenitor 4 (MPP4) is heavily biased toward differentiation into the lymphoid lineage (Adolfsson et al., 2005). Importantly, the discovery and functional annotation of cell types are dependent on the modality of data. FACS and lineage potential are not the only such methods (Markram et al., 2004; Scala et al., 2020; Liu et al., 2021). The advent of single-cell RNA sequencing (scRNAseq) allows for the classification of the hematopoietic lineage from an entirely new data modality. Using gene expression to characterize cell types gives us a new opportunity to identify the gene regulatory programs important to transcriptionally defined cell types.

1.2 The Development and Growth of scRNAseq

1.2.1 The Technology

Single-cell RNA sequencing is a tool for measuring the entire transcriptome of individual cells. Numerous protocols have been developed in a little over a decade, each with its advantages and disadvantages (Tang et al., 2009; Svensson, Vento-Tormo, and Teichmann, 2018). Despite the diversity of protocols, they all involve a method for attaching a unique DNA sequence barcode as RNA is converted to cDNA from individual cells. Methods can be separated into ones that place individual or groups of cells into 96- or 384-well plates or methods that create droplets using microfluidics (Macosko et al., 2015; Hagemann-Jensen et al., 2020; Hwang, Lee, and Bang, 2018). After segregating the cells into wells or droplets, enzymatic reactions convert RNA to cDNA and attach the barcode. Next, further enzymatic reactions amplify cDNA and prepare the samples for sequencing on Illumina sequencers. The data comes off the sequencer as paired reads, one containing the RNA sequence and the other containing the cell barcode and any other barcodes, like unique molecular identifiers (UMIs), that are used in the protocol. UMIs are used to eliminate PCR duplicates by assigning each original RNA strand a unique barcode. During sequence alignment, the cell barcodes are demultiplexed such that all reads with the same barcode are assigned to one cell, resulting in a genes-by-cells counts table, where each position stores the number of counts seen for a given gene in a given cell. While various scRNAseq methods follow this general protocol, the unique aspects of each method can greatly affect the data. Considering the tradeoffs of different protocols is a crucial step in experimental design.

When selecting a scRNAseq protocol to use there are a lot of factors to consider to conduct an experiment that is poised to answer a specific question. Methods can either capture RNA from whole cells, but only using fresh samples, or using fresh or frozen nuclei. After selecting the sample type, the next most important parameters to balance are the number of cells and sequencing depth. These are generally opposing parameters, you either sequence many cells at low depth (Cao et al., 2019) or sequence fewer cells at higher

depth (Santoro, Chien, and Sahara, 2021). The term sequencing depth is defined as the number of reads/counts per cell measured; higher depth captures a more complete picture, but is more expensive and technically more challenging. This trade-off arises from both limitations with the technology and cost. Experiments that sample many cells, at lower depths, will be more likely to capture rare cell populations while sacrificing the ability to learn more detailed signatures for each cell type. In the subsequent chapters, we analyze data from a variety of technologies, both whole-cell and nuclei, and from datasets that sample a little over 1,000 cells very deep to datasets that sample over 100,000 cells.

A major goal in applying scRNAseq to various organisms and tissue systems is to develop atlases to serve as a reference of all the cells in a system of interest. The most ambitious atlasing effort is the Human Cell Atlas consortium. The human cell atlas is a global effort to characterize all the cell types in humans tissue-by-tissue (Regev et al., 2017; Park et al., 2020; Litviňuková et al., 2020). Other efforts have produced preliminary atlases for mice, humans, and flies by comprehensively profiling multiple tissues (Li et al., 2021b; Schaum et al., 2018; Consortium and Quake, 2021). These references serve to help with labeling future datasets, as well as comparisons for disease or other perturbations to the systems (Adamson et al., 2016; Segerstolpe et al., 2016; Roberto et al., 2021). It is critical to have a high-quality reference for these disease and perturbation studies, as the quality of the signatures learned from using a reference will be constrained by the quality of the reference.

This thesis is focused on methods for developing and analyzing scRNAseq atlases. In Chapter 2 I analyze multiple scRNAseq datasets from mouse bone marrow to construct an atlas of hematopoietic differentiation. In Chapter 3 I re-analyze an existing atlas of neurons from the mouse primary motor cortex (Yao et al., 2020a). scRNAseq and cell atlases have been extremely popular in both of these fields, making robust atlasing efforts using meta-analysis very fruitful. In the next two subsections, I summarize the contributions of scRNAseq to the fields of hematopoiesis and neuroscience.

1.2.2 scRNAseq in Hematopoiesis

The hematopoietic lineage is one of the most highly studied lineages in all of developmental biology. A clear understanding of hematopoietic development is central to infectious disease, aging, and cancer research. The bias towards the development of myeloid cells instead of the lymphoid lineage is a major molecular signature of aging (Rossi et al., 2005; Kowalczyk et al., 2015; Elias, Bryder, and Park, 2017). scRNAseq expanded on an understanding of the hematopoietic signature of aging by showing an increase in CD4 T cells that occurred through clonal expansion in supercentenarians (Hashimoto et al., 2019). Additionally, some hematological cancers can be viewed as misregulation or stalled development of myeloid cells, leading to a class of therapeutics known as differentiation therapy (Nowak, Stewart, and Koeffler, 2009; Liu et al., 2012). Identifying the changes in gene regulation that cause lineage bias or developmental stalling is crucial to perturbing these systems back into a healthy state. Despite being a relatively new technology, scRNAseq is now incorporated into clinical trial protocols for myeloma to understand treatment resistance mechanisms (Cohen et al., 2021). An atlas that describes cell types involved in healthy hematopoiesis, and characterizes the function for each cell type using scRNAseq will serve as a critical reference for translational research.

We can see how well the technology has been adopted by looking at the growth in populations over the past few years (Figure 1.1A). Analogous to how transistor size serves as a metric to track the innovation of CPUs, the growth in scRNAseq dataset size shows the maturation of scRNAseq technology (Figure 1.1B) (Svensson, Vento-Tormo, and Teichmann, 2018). Hematopoietic research has largely relied on FACS as the primary readout for cell type when conducting experiments like colony-forming assays, transplantation, or drug/genetic perturbations. The maturation of scRNAseq will lead to these classical experiments switching to scRNAseq as the readout for cell type instead of FACS.

1.2.3 scRNAseq in Neuroscience

scRNAseq has potentially seen even greater adoption in neuroscience than hematology. The number of datasets and size of the datasets also are growing at a rapid pace

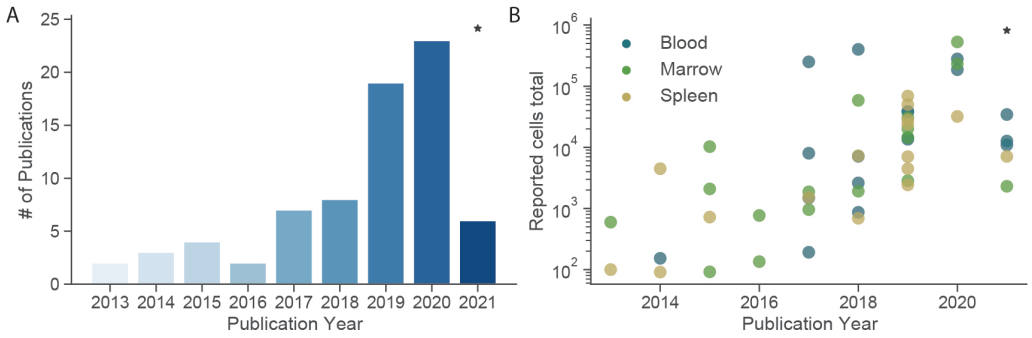


FIGURE 1.1: The popularity of scRNAseq within the hematopoietic lineage.**A)** The growing number of datasets published using blood, marrow, or spleen in mice. **B)** The growth of scRNAseq datasets over time in blood, marrow, and spleen. Panels were created using data collected in July 2021 from a curated database of scRNAseq publications (Svensson, Beltrame, and Pachter, 2020). The asterisk is noting that 2021 was only half over, with the potential for many other publications in the field. Additionally, COVID impacts in 2020 and 2021 could delay many publications because of disruptions to lab work, resulting in a deviation from the growth trend.

year over year (Figure 1.2A-B). The field uses many non-sequencing modalities for characterizing cell types; specifically electrophysiology, morphology, and connectivity. The most notable scRNAseq analysis in neuroscience comes from The Brain Initiative Cell Census Network (BICCN). The BICCN is an NIH-funded consortium that is bringing together researchers from all over to use all the modalities to characterize mice, non-human primates, and human neuronal diversity. An important goal of the consortium is to integrate data from across various sequencing modalities to better understand how individual cell types work to serve the function of the brain. In general, expression data serves as a foundation for the integration of other data modalities, providing robust signatures which can then be annotated by the data used in other modalities (Yao et al., 2020a). More targeted neuroscience research outside the BICCN includes studies capturing the development of the mouse brain at multiple timepoints and how gene regulation impacts sex differences in the brain (Bella et al., 2021; Gegenhuber et al., 2020; Trevino et al., 2021).

1.3 Analysis of scRNAseq Data

In this section, I discuss the computational methods relevant to scRNAseq analysis. Since this thesis is focused on the reanalysis of published data, and in particular data with

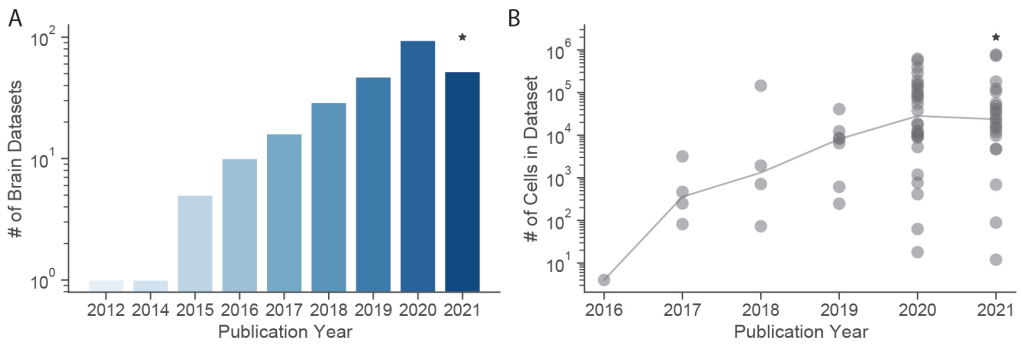


FIGURE 1.2: **Growth of scRNAseq data in the mouse brain. A)** Growth in the number of scRNAseq datasets year-over-year. **B)** Dataset size increases over time for mouse brain data. The data in these figures were collected as part of an unpublished project in the lab to re-analyze all neuronal scRNAseq data and were collected in July 2021. The asterisks note that 2021 is incomplete at the time of writing and reflects only half of the year.

labeled cell types, I begin by summarizing, in broad terms, the methods necessary to go from raw sequencing data to labeled cell types (Chapter 1.3.1). Finally, I discuss two methods, Pseudotime (Chapter 1.3.2) and Co-expression (Chapter 1.3.3), that provide more context for the later chapters.

1.3.1 Preprocessing and Labeling Cell Types

There are over 1,000 published tools for analyzing scRNAseq, positioned at all stages of the analysis pipeline (Zappia and Theis, 2021). This can become quite overwhelming and requires a fair bit of consideration to select the tools best suited to provide robust answers to specific questions (Kharchenko, 2021). Going from raw data, FASTQ files, to counts tables is the first preprocessing step. This involves aligning the reads to the genome or transcriptome, demultiplexing barcodes, and collapsing UMIs if they are used (Kaminow, Yunusov, and Dobin, 2021; Melsted et al., 2021). Not every sequenced barcode corresponds to a single cell, thousands of empty droplets and also droplets or wells containing multiple cells can be identified as a barcode. The alignment tools incorporate methods that estimate, based on the number of genes expressed and/or total counts, whether or not the barcode likely corresponds to an individual cell. These tools output a cells-by-genes matrix, with each value in the matrix containing the counts for that cell-gene combination.

From the output of the aligners, the data is read into R or python for the remaining analysis steps. These steps can be easily carried out using either Scanpy (python) or Seurat (R) (Wolf, Angerer, and Theis, 2018; Hao et al., 2021). First, normalization accounts for noise and technical differences between cells. Library size normalization is the simplest and most commonly used method, although model-based methods attempt to correct for sampling and other non-biological biases (Hafemeister and Satija, 2019). Following normalization, feature selection reduces the impact of redundant genes, and principal component analysis projects the data into a space of 20-50 dimensions. While there are 20,000 protein-coding genes in mice or humans, working in that many dimensions is both difficult and unnecessary. Most feature selection methods select highly variably expressed genes sampled across different levels of expression, to prevent all signals from coming from the highest expressed genes. Most importantly, the effective dimensionality of the data is much lower, there are few cell types/states compared to the number of genes. Constructing a nearest neighbor graph, the next step in the analysis, is particularly sensitive to the curse of dimensionality (Friedman, 1997). This graph, built from the pairwise euclidean distance of the cells in PCA space, is useful for both Louvain or Leiden clustering or projecting in a non-linear 2-D space using t-SNE, or UMAP. The clustering partitions the cells into groups. These groups are labeled with a known cell type based on the enriched expression of genes previously identified for those cell types. For our analysis in Chapter 3, all the data I use already come labeled by cell type, which I exploit in the meta-analysis. However, not all data requires labeling from scratch if the biological system already has a reference atlas. Computational methods can use reference atlases to transfer labels to new datasets (Kimmel and Kelley, 2021; Hao et al., 2021; Ge et al., 2021). As deciding the correct number of clusters can be quite subjective, it is easier to label new datasets using reliable references, as we do in Chapter 2. Methods for labeling using a reference rely on learning a signature from the training dataset and then classifying cells in a new, query dataset (Grabski and Irizarry, 2020; Kimmel and Kelley, 2021; Abdelaal et al., 2019). Selecting a high-quality reference dataset is crucial to the effectiveness of transferring labels. Using published datasets that have been thoroughly evaluated and match the field's understanding of the biological system being sampled.

1.3.2 Pseudotime

scRNAseq provides the capability to profile the cells in a developing system such that I have a sampling of progenitors and lineage-committed cells and everything in between from a single assay. To exploit this properly, methods estimate a temporal ordering, known as pseudotime, and if relevant, a tree to model the various lineages produced through development. (Kester and Oudenaarden, 2018; Lederer and Manno, 2020). There is a lot of variety in pseudotime methods. As of August 2021, there were 129 published methods. (Zappia, Phipson, and Oshlack, 2018). However, three of the most popular methods are conceptually similar (Wolf et al., 2019; Cao et al., 2019; Street et al., 2018). They all rely on placing the cells on a lower complexity graph than the nearest neighbor. This lower complexity graph is usually a minimum spanning tree of substantially fewer nodes than the number of cells. This graph represents the inferred developmental trajectory, with each cell being assigned to a node along this graph. To compute a specific pseudotime or ordering of the cells, the user identifies the potential start, or root, of the tree, and then the distance, either empirically calculated or via a stochastic process, is computed from the root to all cells on the tree. Downstream analyses that use the inferred pseudotime learn gene expression patterns associated with pseudotime progression, and lineage commitment if there is a branching trajectory (Berge et al., 2020; Qiu et al., 2017b).

Using pseudotime to analyze dynamic systems in scRNAseq is conceptually very logical. However, the methods remain far from perfect and are particularly sensitive to overfitting noise in data coming from a single dataset (Song and Li, 2020). No methods for evaluating the replicability of pseudotime methods across datasets with inferred trajectories currently exist. The methods recommend integrating (see Chapter 1.4.2) the batches together and then inferring the pseudotime from there. In Chapter 2 I compare pseudotime inferred in individual datasets to an integrated dataset and demonstrate the use of meta-analytic methods to evaluate trajectories learned across multiple datasets.

1.3.3 Co-expression

Co-expression networks characterize genes as related based on their shared expression profiles across samples. A shared profile suggests their activity is driven by the same factors or that they are functionally related (Eisen et al., 1998). Networks built from bulk gene expression data have been widely observed to recapitulate known gene functions (Figure 3)(Eisen et al., 1998; Lee et al., 2004). As a result, co-expression analysis serves many applications in genomics. For example, co-expression has been used to infer transcription factor binding and causal regulation of downstream targets (Fiers et al., 2018; Kulkarni et al., 2017; Kulkarni et al., 2017; Song et al., 2016), characterize disease (Torkamani et al., 2010), and to predict which cells will interact with each other based on ligand-receptor pairs (Torkamani et al., 2010; Efremova et al., 2020). Functional assessment tools, including Extending "Guilt-by-Association" by Degree (EGAD), use machine learning to determine whether a co-expression network predicts a reference gene functional annotation like the Gene Ontology (Ballouz, Pavlidis, and Gillis, 2017).

The tool WGCNA is the most popular software package for computing co-expression (Langfelder and Horvath, 2008; MA et al., 2012; Hartl et al., 2020; Mack et al., 2019). The package, built for microarray and bulk RNAseq analysis, provides an easy method to produce gene modules from a dataset. The network is built by computing the Pearson or Spearman correlation matrix and then raising it to a power between 1 and 10, to create a scale-free network, and then computing the topological overlap to create a distance matrix. Gene modules are identified using a dynamic tree cutting algorithm that identifies the best place to split the hierarchical clustering of the distance matrix. These modules, lists of genes, can then be evaluated for the enrichment of functional annotations, disease-associated genes, or cell type markers (Hartl et al., 2020; Kelley et al., 2018; Torkamani et al., 2010).

While many are optimistic about the use of co-expression analysis for scRNAseq data (Trapnell, 2015), networks built from bulk RNA sequencing better recapitulate known gene functions (Crow et al., 2016). The drop in performance can largely be attributed to the sparsity of scRNAseq data. Benchmarking of many association metrics used for

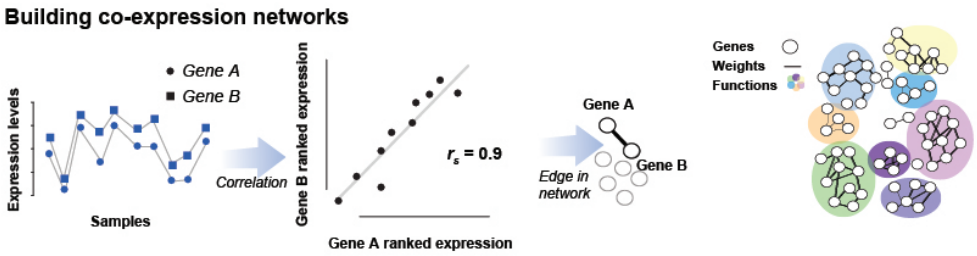


FIGURE 1.3: Co-expression measures the pairwise assocation between genes in RNA sequencing. Genes that are co-expressed with each other are more likely to share a function

building networks shows relatively consistent results regardless of the method used, with only a minor increase in performance from proportionality over more popular Pearson and Spearman correlation coefficients (Skinnider, Squair, and Foster, 2019). Both the evaluations in Crow et al 2016 and Skinnider et al 2019 were conducted across many datasets of varying cell types and tissues. By using the 7 BICCN mouse primary motor cortex datasets, we are conducting a well-powered co-expression analysis in scRNAseq using consistent cell types across the datasets (Chapter 3).

1.4 Major Challenges in the Field

In this section, I introduce two important concepts in the field of scRNAseq where this work makes significant contributions to the field. First, I explain how replicability and generalizability are major challenges within high throughput biology, and how our analysis exploits techniques from statistics and machine learning to learn generalizable signatures of gene expression (Chapter 1.4.1). I then discuss the merits of the two main options for analyzing multiple scRNAseq datasets: integration or meta-analysis (Chapter 1.4.2).

1.4.1 Generalizability

High throughput biology has revolutionized the way I conduct research, especially in hypothesis-generating projects. Analyzing data from these experiments requires careful considerations. When querying many hypotheses, like identifying cell type markers from

scRNAseq by testing all genes expressed in the data, the risk of identifying a spurious correlation increases with the number of hypotheses tested. In this thesis, I utilize multiple techniques to increase our ability to separate the real biological signals from noise and spurious correlations. Two major methods I use are multiple hypothesis corrections and meta-analysis. Using these techniques, I can learn generalizable signatures; gene lists that reflect real biological signals that can be applied to external data.

With null hypothesis significance testing, we can apply methods that correct for multiple hypotheses (Noble, 2009). This is necessary because as the number of tests increases, the number of positives identified, regardless of whether it is a true or false positive, increases. These methods convert the raw p values computed by the statistical methods to a corrected value, usually referred to as the false discovery rate (FDR). Selecting all genes with an FDR<.05 means that within the significant genes, 5% of them are expected to be false positives. This approach lets us easily determine the desired error rate, based on the null assumption that p values should be uniformly distributed. This correction, while useful on individual datasets, does not get past the bigger issue of making inferences about data that is not replicable.

Our most powerful tool for learning generalizable signatures is meta-analysis. Replicability across many datasets resolves some technical limitations of individual scRNAseq datasets, creating a more robust atlas (Hicks et al., 2017; Crow and Gillis, 2019). The most comprehensive and robust cell atlases rely on meta-analysis across many scRNAseq datasets to characterize replicable cell types (Cook and Vanderhyden, 2021; Swamy et al., 2021; McKellar et al., 2020; Yao et al., 2020a). After identifying the present cell types, characterizing their functions can be done by evaluating the signatures that define the cell types for known processes and pathways (Crow et al., 2018; Evrard et al., 2018; Cao, Wang, and Peng, 2020). Using R.A. Fisher's method for combining p values, I can compute replicable signatures across datasets by aggregating p values from individual datasets into a single value. Additionally, using cross-validation, I learn signatures from some of our datasets (training datasets) and evaluate the quality of the signatures in held-out data (testing datasets) to evaluate the generalizability of the learned signatures.

In this work, all of my analyses are applied across multiple datasets or batches with a significant technical variation. This allows me to learn signatures that do not overfit the technical variation in high throughput sequencing.

1.4.2 Integration vs. Meta-analysis

While I highlight the importance of meta-analysis in scRNAseq, how I use them is not necessarily the norm in the field. An important distinction with our methodology is that I do almost all computations on values from individual datasets and then aggregate statistics across all the datasets, whether it be clustering, pseudotime, co-expression, or differential expression analysis. This is in contrast to a lot of popular methods in both the bulk and scRNAseq space. The alternative is known as a batch correction, usually for bulk, and integration, usually for scRNAseq. The methods produce an integrated space across the datasets that can be used for clustering,

These methods come from the desire to analyze data that cannot be generated in a single experiment in a single analysis, due to limitations in either sequencing technology or experimental design. Older methods, developed for bulk RNAseq generally assume linear batch effects and rely on regression analysis to "correct" the expression values (Johnson, Li, and Rabinovic, 2007; Risso et al., 2014). These methods are significantly less complex than the scRNAseq, which makes further assumptions about both the linearity of the batch effects and also the composition of the data. All scRNAseq methods assume that the datasets being integrated share at least some cell types, and some also assume non-linearity in the batch effects (Barkas et al., 2018; Welch et al., 2019; Forcato, Romano, and Bicciato, 2020). The true effectiveness of scRNAseq methods is hard to evaluate from the biased benchmarks in the publications for the methods. Even in benchmarks done by third parties, no methods stand out as being more effective than the rest (Tran et al., 2020; Luecken et al., 2020).

It is also worth considering the utility of integration based on what downstream analysis it facilitates. The two main downstream analyses from integration are either qualitative visualization analysis and clustering/labeling cells in a joint space (Chari, Banerjee,

and Pachter, 2021; Stuart et al., 2019). The qualitative analysis is certainly useful, as reducing the integrated space to 2 dimensions, even if there is some distortion of the true distances, provides an interpretable and convenient way to view > 2-dimensional data. Visually it is more convenient to plot the expression of a gene on an integrated UMAP than on N UMAPs for each dataset/batch in your publication. More succinctly, the best use for integration methods is visualization because it is convenient and practical.

For more quantitative analysis, like clustering, an integrated space makes it easy to partition the cells into consistently labeled clusters, even if by placeholder names and not explicit cell type labels. The integrated space also makes it easy to visualize the clustering. Integration can also be used for transferring labels from a reference (Kiselev, Yiu, and Hemberg, 2018; Abdelaal et al., 2019). While less popular, you can cluster and label each dataset individually and then evaluate the replicability of the clusters using tools like MetaNeighbor (Appendix A) across the batches and provide the shared clusters with uniform labels (Appendix B).

Notably, in Chapter 2, I do not take that approach, which would be surprising given how strongly I advocate for meta-analysis. In Chapter 2, I use the tool scNym, as a semi-supervised adversarial learning method, to integrate and label 12 hematopoietic stem cell datasets using 2 labeled datasets (Kimmel and Kelley, 2021). I deviated from my normal procedure because I was less certain about the cell types present in the unlabeled data. Integrating and labeling using a reference, built by experts, does take a certain amount of trust, but after integration, I do a thorough analysis of the newly labeled data to evaluate the replicability of the cell types and the genes and functional signatures that define them. And those evaluations were conducted using a meta-analytic framework like I normally do. Ultimately, scRNAseq analysis can be properly done using either integration or meta-analysis. It is most critical that you rely on multiple datasets and thoroughly evaluate the cell types and learned signatures in a way that will be generalizable to other data.

Chapter 2

A Meta-Analytic Single-Cell Atlas of Mouse Bone Marrow Hematopoietic Development

The chapter below is the manuscript currently posted on Biorxiv https://doi.org/10.1101/2021.08.12.456098 (Harris, Lee, and Gillis, 2021). I conducted the experiments and wrote the manuscript under the supervision of my adviosr Jesse Gillis. John Lee and I created a webserver for people to use to analyze their own data gillisweb.cshl.edu/HSC_atlas.

2.1 Summary

A clear understanding of the cell types and functional programs during hematopoietic development is central to research in aging, cancer, and infectious diseases. Traditionally, cell types are identified by cell surface protein expression. Progenitor cells defined this way are assigned functions based on their lineage potential. The rapid growth of single cell RNA sequencing (scRNAseq) technologies provides a new modality for evaluating the cellular landscape of hematopoietic progenitors. Using over 300,000 cells across 12 datasets, we evaluate the classification and function of cell types based on discrete clustering, in silico FACS, and a continuous trajectory. This produces replicable cell types based

on genes and known cellular functions. We evaluate the conservation of signatures associated with erythroid and monocyte lineage development across species using co-expression networks for zebrafish and human scRNAseq data. This analysis provides a robust reference, particularly of marker genes and functional annotations, for future experiments in hematopoietic development.

2.2 Introduction

The hematopoietic lineage is one of the most highly studied lineages in all of developmental biology. Classically, cell types are identified by Fluorescent Activated Cell Sorting (FACS). For example, the Long Term Hematopoietic Stem Cell (LT-HSC) is identified by CD34low, Flt3-, and TpoR+ expression. The role of a progenitor is to produce differentiated cells, and the function of a specified progenitor cell type is defined by the potential to differentiate into a specific lineage (Nilsson, Pronk, and Bryder, 2015). The Multipotent Progenitor 4 (MPP4) is heavily biased toward differentiation into the lymphoid lineage (Adolfsson et al., 2005). Importantly, the discovery and functional annotation of cell types are dependent on the modality of data. FACS and lineage potential are not the only such methods (Markram et al., 2004; Scala et al., 2020; Liu et al., 2021). The advent of single-cell RNA sequencing (scRNAseq) allows for the classification of the hematopoietic lineage from an entirely new data modality. Using gene expression to characterize cell types gives us a new opportunity to identify the gene regulatory programs important to hematopoietic lineages.

A clear understanding of hematopoietic development is central to aging and cancer research. The bias towards the development of myeloid cells instead of the lymphoid lineage is a major molecular signature of aging (Rossi et al., 2005; Kowalczyk et al., 2015; Elias, Bryder, and Park, 2017). Additionally, some hematological cancers can be viewed as misregulation or stalled development of myeloid cells, leading to a class of therapeutics known as differentiation therapy (Nowak, Stewart, and Koeffler, 2009; Liu et al., 2012). Identifying the changes in gene regulation that cause lineage bias or developmental stalling is crucial to perturbing these systems back into a healthy state. An atlas that describes cell

types involved in healthy hematopoiesis, and characterizes the function for each cell type using scRNAseq will serve as a critical reference for translational research.

The rapid development of scRNAseq technology creates the opportunity to build a robust atlas of hematopoietic cells in the bone marrow. Multiple studies publish individual atlases of hematopoietic development, but they do not integrate information from other published datasets (Dahlin et al., 2018; Olsson et al., 2016). Replicability across many datasets resolves some technical limitations of individual scRNAseq datasets, creating a more robust atlas (Hicks et al., 2017; Crow and Gillis, 2019). The most comprehensive and robust cell atlases rely on meta-analysis across many scRNAseq datasets to characterize replicable cell types (Cook and Vanderhyden, 2021; Swamy et al., 2021; McKellar et al., 2020; Yao et al., 2020b). After identifying the present cell types, characterizing their functions can be done by evaluating the signatures that define the cell types for known processes and pathways (Crow et al., 2018; Evrard et al., 2018; Cao, Wang, and Peng, 2020).

In this work, we build a comprehensive mouse hematopoietic cell atlas by integrating and labeling over 300,000 cells from 14 datasets. We identify robust gene regulatory signatures using multiple perspectives of the data. Two bone marrow datasets from the Tabula Muris consortium and the semi-supervised machine learning algorithm scNym are used to label and integrate 12 datasets of mouse bone marrow data (Kimmel and Kelley, 2021; Schaum et al., 2018). We identify robust markers for each cell type and learn functional annotations using the Gene Ontology. Labeling cells based on genes that traditionally serve as cell surface markers identifies a latent lineage potential signature. Pseudotime analysis finds signatures associated with the development of the monocyte and erythroid lineages. Co-expression and scRNAseq from zebrafish and human samples evaluates the conservation of lineage-associated signatures in 21 species. We present a replicable view of hematopoietic development in the mouse bone marrow, that complements the FACS and lineage potential-based perspective of hematopoietic development.

2.3 Results

2.3.1 Integration and Filtering of Datasets

We collected 12 published datasets that use high throughput scRNAseq methods to profile mouse hematopoietic progenitor cells (Dahlin et al., 2018; Tikhonova et al., 2019; Weinreb et al., 2020; Rodriguez-Fraticelli et al., 2020; Giladi et al., 2018; Tusi et al., 2018; Cheng et al., 2019). Not all of the original publications label every cell and each publication has unique rules for defining cell types. These two challenges make it unclear what cell types are shared across publications by looking at the published papers and associated metadata. It is preferable to have an integrated latent space with cells from all datasets for some analyses. After identifying the shared populations across the publications, we can evaluate the discrete and continuous models of cell types (Figure 2.1). We use the tool scNym and the Tabula Muris bone marrow dataset, a high-quality reference dataset, to integrate and label the cell types from all of the studies. From 7 publications, we identified 12 sequencing batches that we refer to as datasets. Projecting individual datasets into a latent space using UMAP for Weinreb et al 2020 and Rodriguez-Fraticelli et al 2020 clearly shows technical variation between annotated batches, while Tikhonova et al 2019, despite annotating multiple batches, does not present strong batch effects (Figure 2.2). We treat each of the batches in the 3 publications as individual datasets to avoid fitting to technical variation within some datasets. Projecting all the cells into a low dimension integrated UMAP space shows the clustering of cell types based on the scNym labels and consistent overlap for most of the datasets (Figure 2.3, 2.4).

It is important to assess cell type label accuracy when transferring cell type labels from reference data using the confidence scores computed by scNym. In the reduced space, the high confidence areas are cells towards the center of clusters, while lower confidence cells are in the areas between cluster centers (Figure 2.5, 2.6). Displaying just the training Tabula Muris data in the latent space makes it clear that the high confidence scores are in the regions of the latent space occupied by Tabula Muris cells, and regions of low confidence are in between the reference cell types (Figure 2.7). The low confidence between clusters

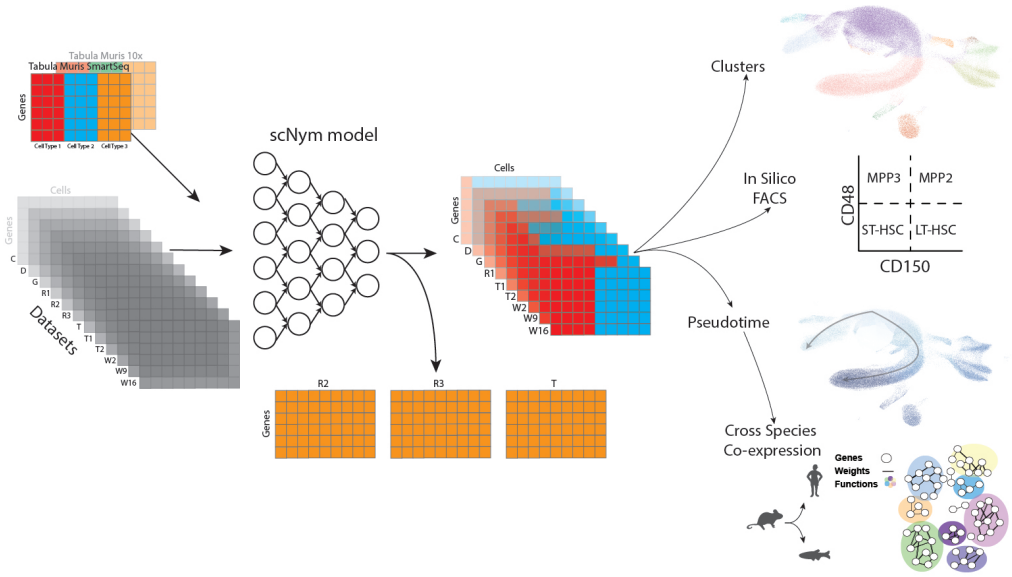


FIGURE 2.1: Two tabula muris bone marrow datasets are used as references with scNym to label 12 datasets. 3 datasets are excluded from further analysis due to poor alignment with the remaining 9 datasets. The 9 remaining datasets are evaluated using a cluster, in silico FACS, and pseudotime analysis. The results of the psuedotime analysis are evaluated across many species.

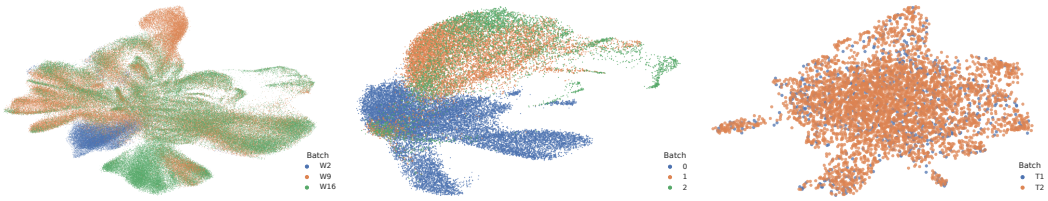


FIGURE 2.2: Projecting the datasets Weinreb et al 2020 (left), Rodriguez-Fraticelli et al 2020 (middle), and Tikhonova et al 2019 (left) using UMAP and colored by the batch label from the metadata shows strong batch effects in the left and middle datasets.

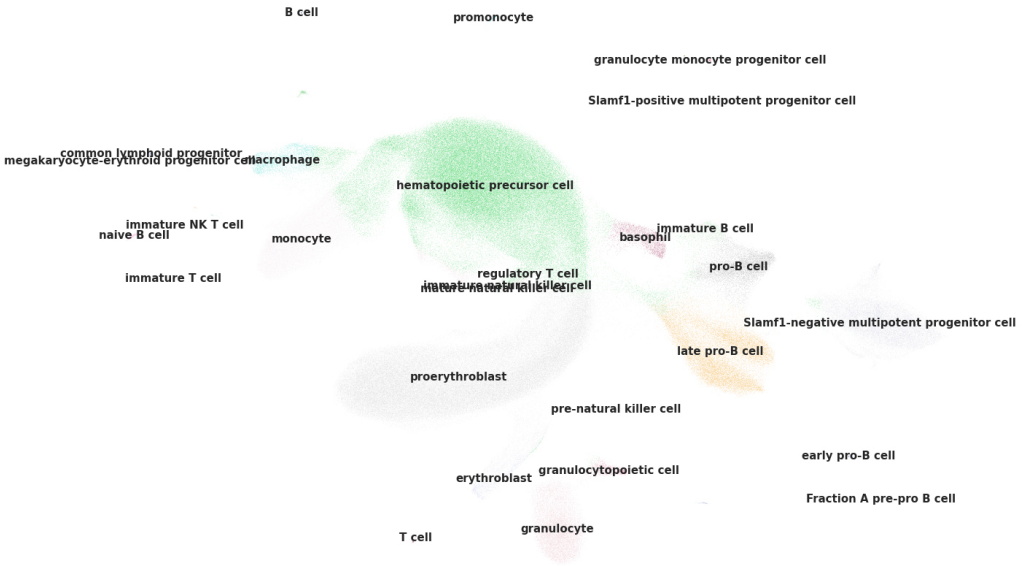


FIGURE 2.3: UMAP projection of the integrated datasets colored by cell type annotation

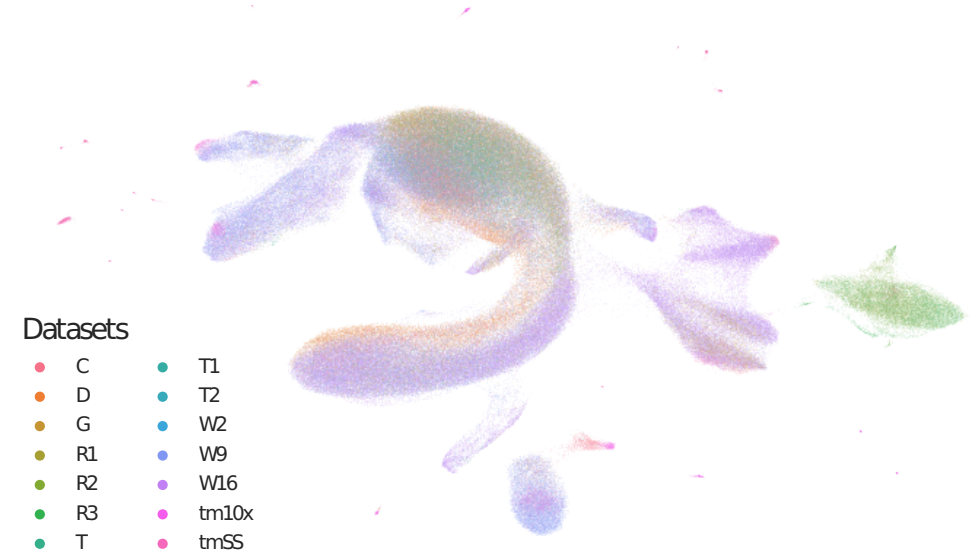


FIGURE 2.4: UMAP projection of the integrated datasets colored by dataset

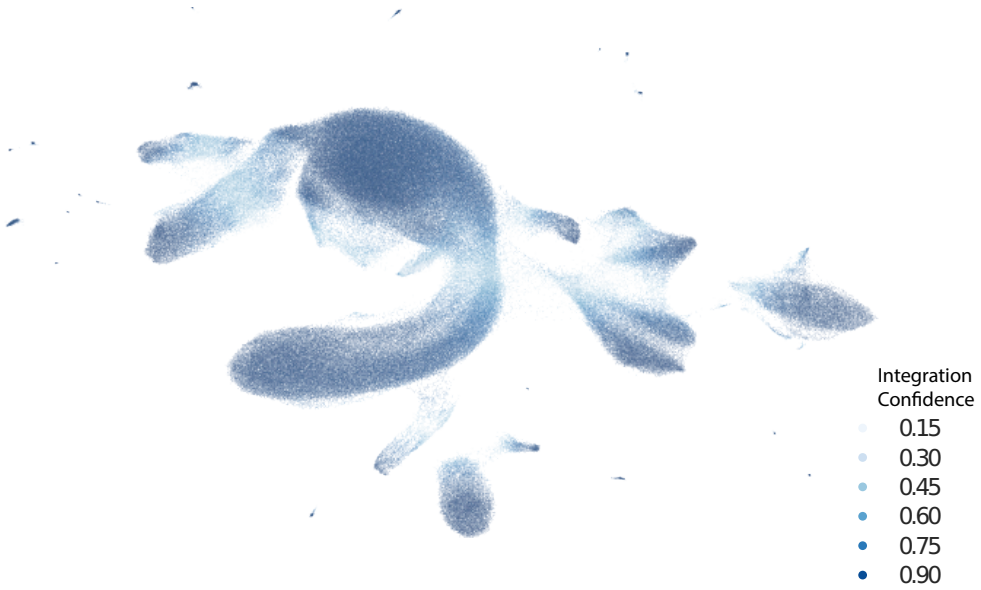


FIGURE 2.5: The confidence score for each cell type label in the UMAP projection.

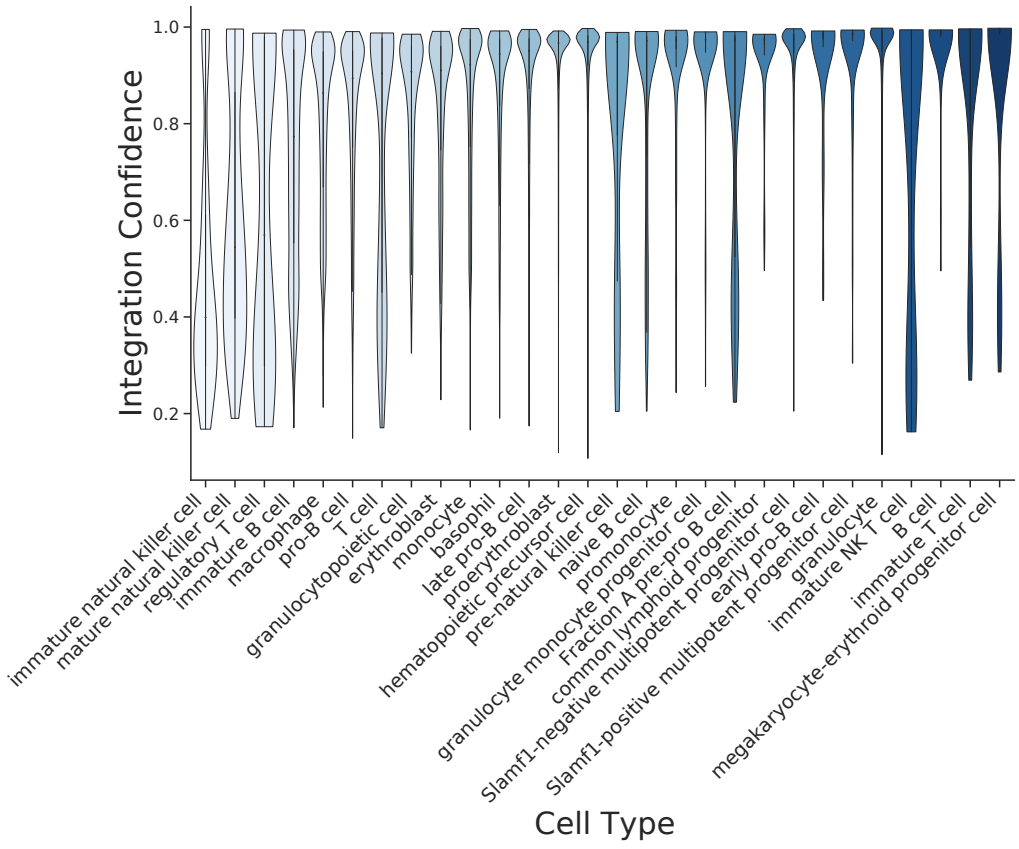


FIGURE 2.6: Confidence scores by cell type show most cells within a cell type are confidently labeled.

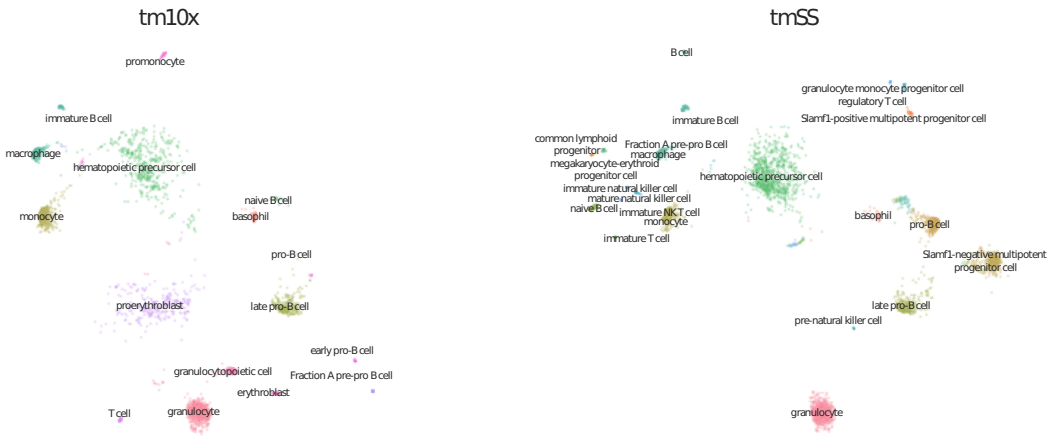


FIGURE 2.7: UMAP projection of the reference tabula muris datasets show disconnected clusters. Tabula Muris 10x (left) and Tabula Muris SmartSeq (right)

reflects the degree to which the model is extrapolating outside the training data space. Most of the Tabula Muris clusters are islands in the latent space with no adjoining neighbors. We suspect this is because the Tabula Muris cells were sorted based on cell surface markers to enrich for specific cell types before sequencing (Schaum et al., 2018). This selects for more transcriptionally homogenous populations, useful for annotation but less so for understanding lineage relationships and variability. On the other hand, the datasets labeled by scNym were only sorted to broadly include hematopoietic stem and progenitors and some lineage-committed cells. After the integration, we removed the datasets R2, R3, and T because they did not map to the other 9 datasets' cell types (Figure 2.8). We exclude the tabula muris datasets, R2, R3, and T to focus on datasets that sampled similar portions of the hematopoietic lineage for the remaining analysis.

2.3.2 Robust Clustering

The remaining 9 datasets all broadly cover the same area of the latent space (2.8, 2.9, 2.10). Most identified cell types are in all 9 datasets; 6 of the 13 clusters are shared across all datasets, and the remaining clusters are in at least 5 of the 9 datasets (Figure 2.11). Every cell type contains at least one marker with an AUROC > .8 and a large fold change (Figure 2.12, Supplementary Table 1). The top markers are very specific to the clusters they identify (Figure 2.13). Klf1 and Ermap, two genes identified as markers

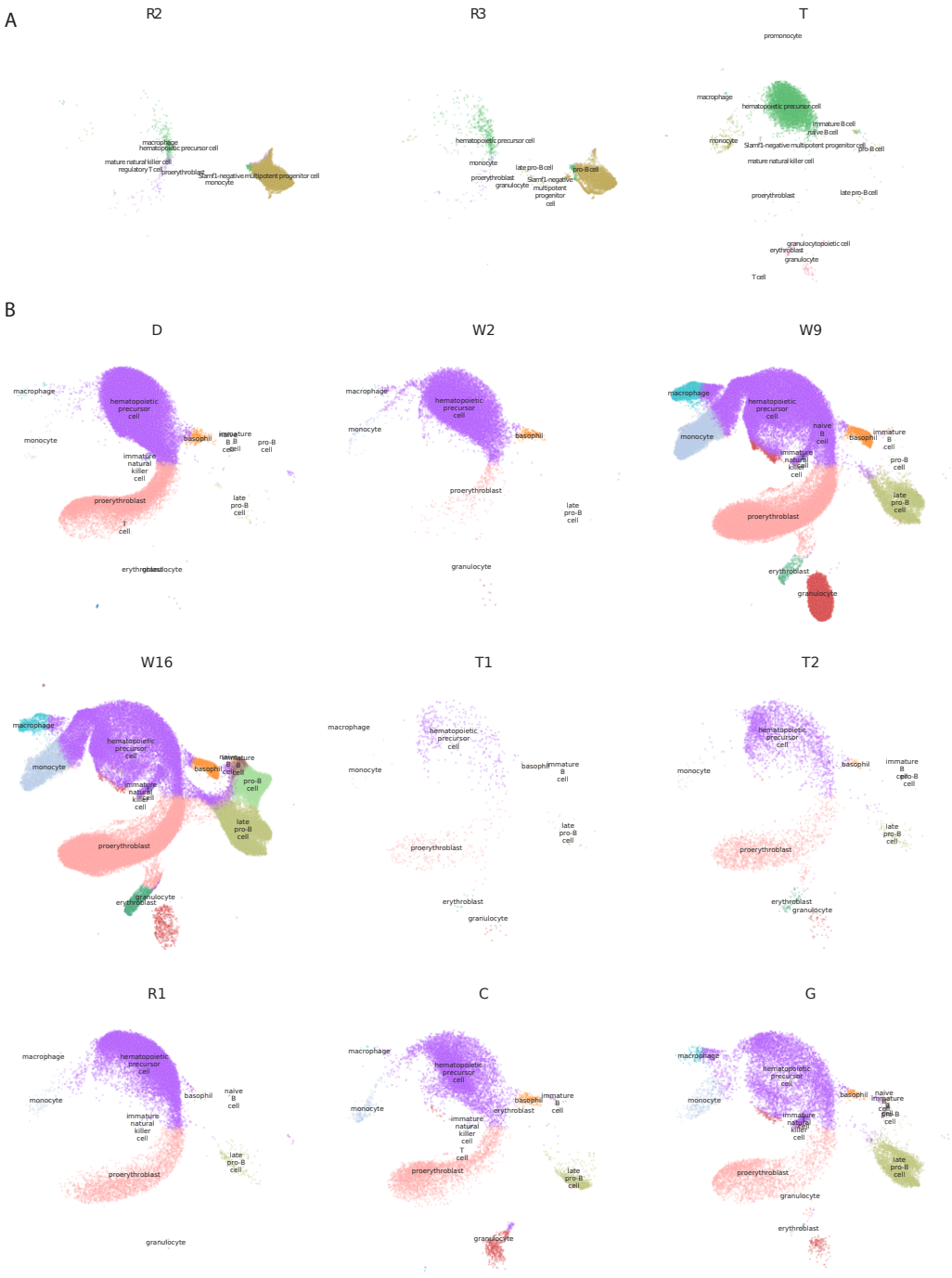


FIGURE 2.8: **Top** Projections in the integrated latent space of the 3 datasets excluded from downstream analysis colored by cell type **Bottom** Projection in the integrated latent space of the 9 datasets that share the same cell types and cover the same region of the integrated latent space

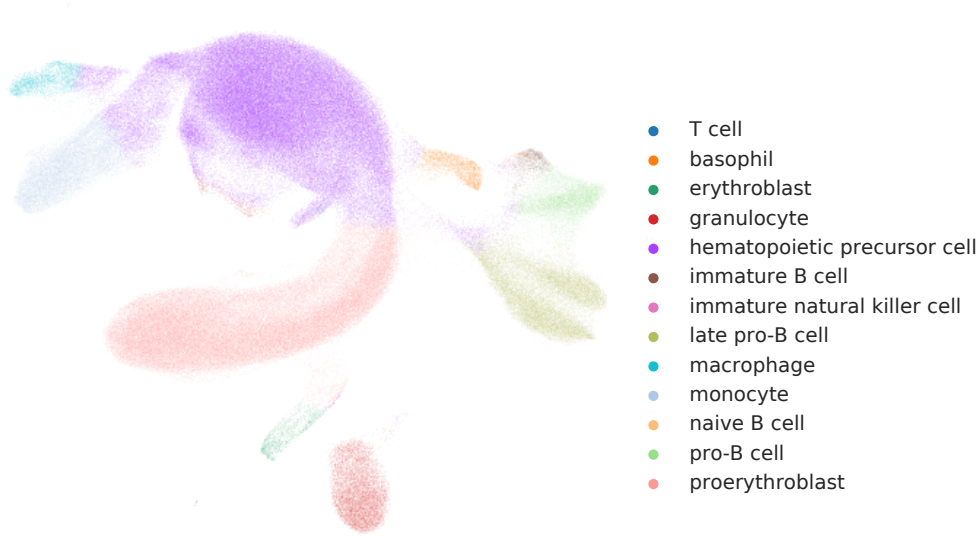


FIGURE 2.9: UMAP projection of the integrated cells colored by cell type annotation

for proerythroblast, are commonly known as erythroid markers (Kingsley et al., 2013; Siatecka and Bieker, 2011). In our dataset selection process, we focused on studies that sorted cells based on the commonly used LSK markers: Lin-, Sca1+ (some Sca1-), and cKit+. The expression of cKit distinguishes proerythroblasts from more differentiated cell types in the erythroid lineage (Munugalavadla et al., 2005). Between the selection method and the marker genes, we are confident in the identification of the proerythroblast lineage, especially over more differentiated cell types within the lineage.

We evaluate the function of each cell type by using MetaNeighbor to identify replicable functional programs associated with each cell type, as labeled by the Gene Ontology. MetaNeighbor characterizes gene sets by their ability to "barcode" particular cell types via their expression profile. Each cell type has at least 75 GO terms with an AUROC > .9 (Figure 2.14, Supplementary Table 2), meaning that the set of genes within that GO term is highly characteristic for a cell type and replicable in its expression profile. The term Embryonic Hemopoiesis (GO:00035162) has an average AUROC of .79, with moderate variation between the different cell types (Figure 2.14). We visualize that variation between the cell types with a dotplot to show the expression of the genes within the term

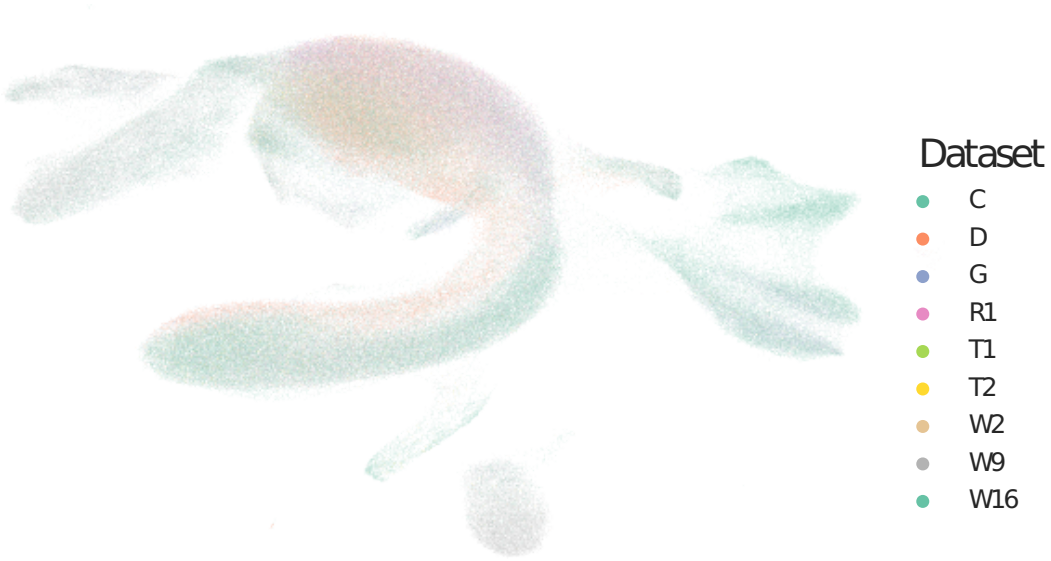


FIGURE 2.10: UMAP projection of the integrated cells colored by dataset.

in each cell type (Figure 2.14). The high performance of the term on basophils is largely driven by Runx1 expression (AUROC=.82). This is consistent with previous studies that show knockout of Runx1 reduces basophils found in bone marrow by 90% (Mukai et al., 2012). Proerythroblasts are the highest performing cluster on the term (AUROC = .92). Gata1, and two genes associated with Gata1 expression, Klf1, and Zfpm1, are enriched in the proerythroblasts. The co-expression of Lmo2 and Ldb1 in proerythroblasts is consistent with results that show their role as maintainers of erythroid progenitor states and preventing further differentiation into the erythroid lineage (Visvader et al., 1997). The marker genes and genes identified from GO enrichment show that we are predominately sampling proerythroblasts from the erythroid lineage.

2.3.3 In Silico Sorting Identifies Latent Stem Cell States

Sorting cells based on cell surface marker protein expression is the established way of defining hematopoietic stem and progenitor cell types. We use the same marker genes, Slamf1 (CD150), Slamf2 (CD48), and Flt3 to sort the hematopoietic precursor cell cluster into Long term HSCs (LT-HSC), Short-term HSC (ST-HSC), and Multipotent Progenitors (MPP2-4) based on published guides (Olsson et al., 2016; Hérault et al., 2021). Interestingly, they do not appear to spatially organize in UMAP space, even when each dataset is

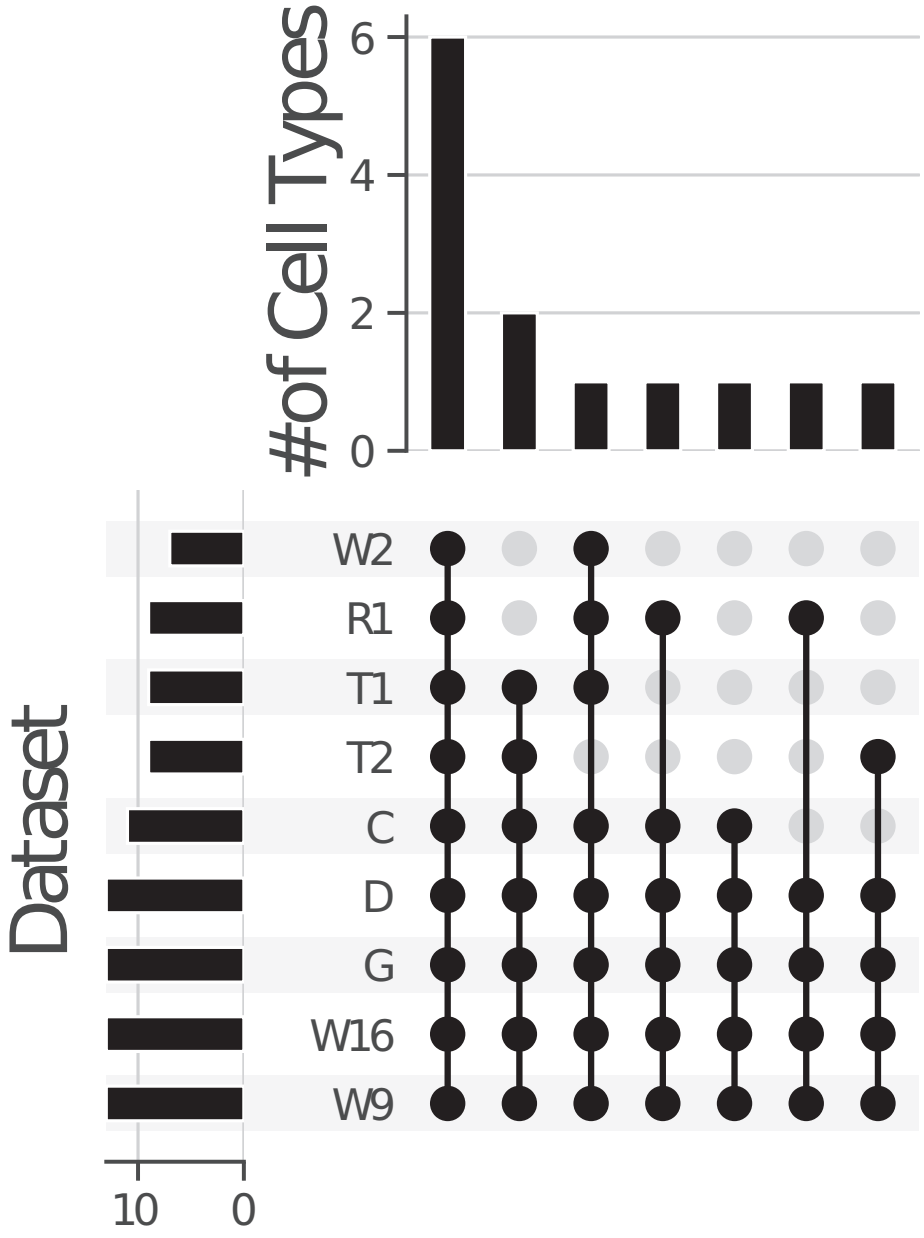


FIGURE 2.11: Upset plot depicts that most cell types are shared across datasets.

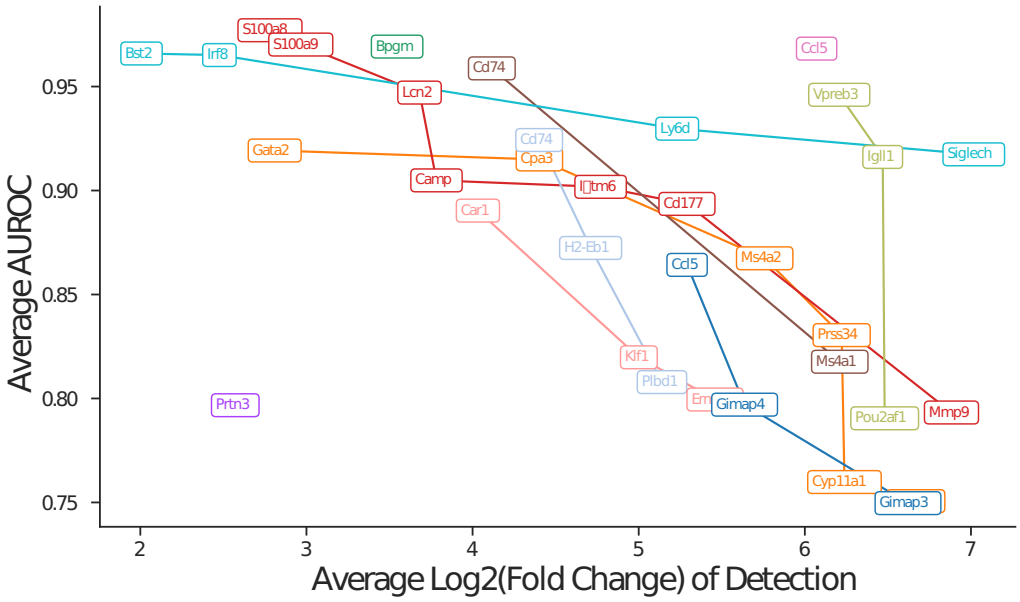


FIGURE 2.12: Each cell type has high-performing markers, as calculated using MetaMarkers. The top markers are plotted by their significance (Average AUROC) and effect size (Average Log Fold Change of Detection). They are colored by the same scheme as in 2.9.

individually projected onto a latent space (Figure 2.15). Using MetaNeighbor to evaluate the replicability of the cell states, there is moderate replicability, especially with the MPP4 and LT-HSCs (Figure 2.16). MetaNeigbhor does not identify a strong distinction between the MPP2 and MPP3 labeled cells, but they are distinct from the remaining cell states.

The top marker genes show modest cell type predictability (AUROC) and weak signal-to-noise ratios (log Fold Change) (Figure 2.17). The ST-HSCs have a near-even signal-to-noise ratio despite the highest predictability for the top markers. We test the identifiability of each cell state using the top 1-1000 markers to see if that does better than individual markers (Figure 2.18). The ST-HSCs have modest identifiability, while the other cell states have extremely low identifiability (Figure 2.18). ST-HSCs are the cell type defined by no expression of Slamf1, Slamf2 and, Flt3, given the sparsity of scRNAseq data and the low signal to noise ratio for ST-HSC marker genes, it could be possible that the cell type is a mixture of actual ST-HSCs and the other cell states incorrectly labeled. When removing the ST-HSCs, the identifiability (F1) increases to moderate levels for the MPP3 and MPP4 cell states using as few as 10 markers. LT-HSC identifiability is extremely low

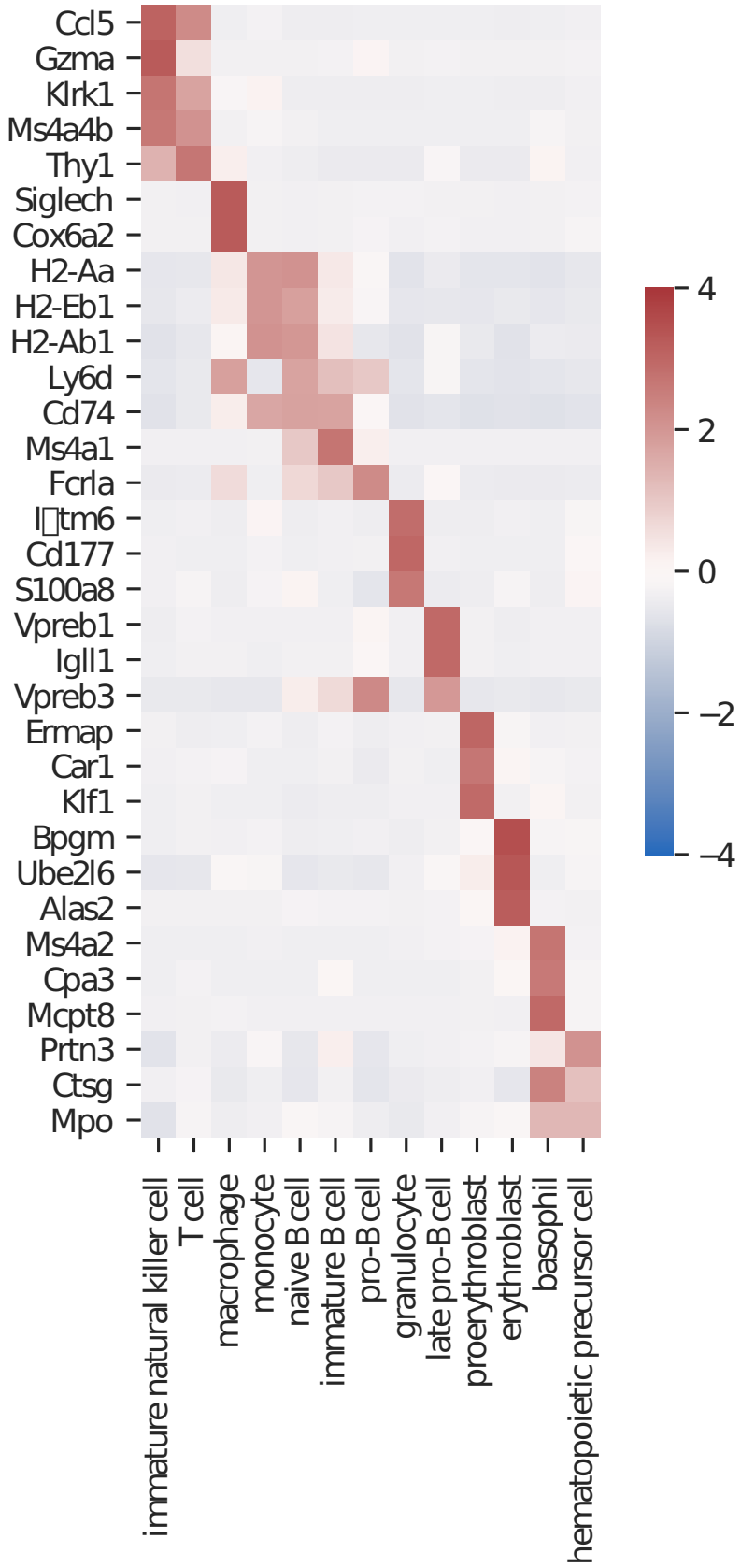


FIGURE 2.13: The top 3 markers for each cluster show high expression specificity in a heatmap of expression. Z scores were calculated within datasets and then aggregated across datasets to account for technical variation between datasets.

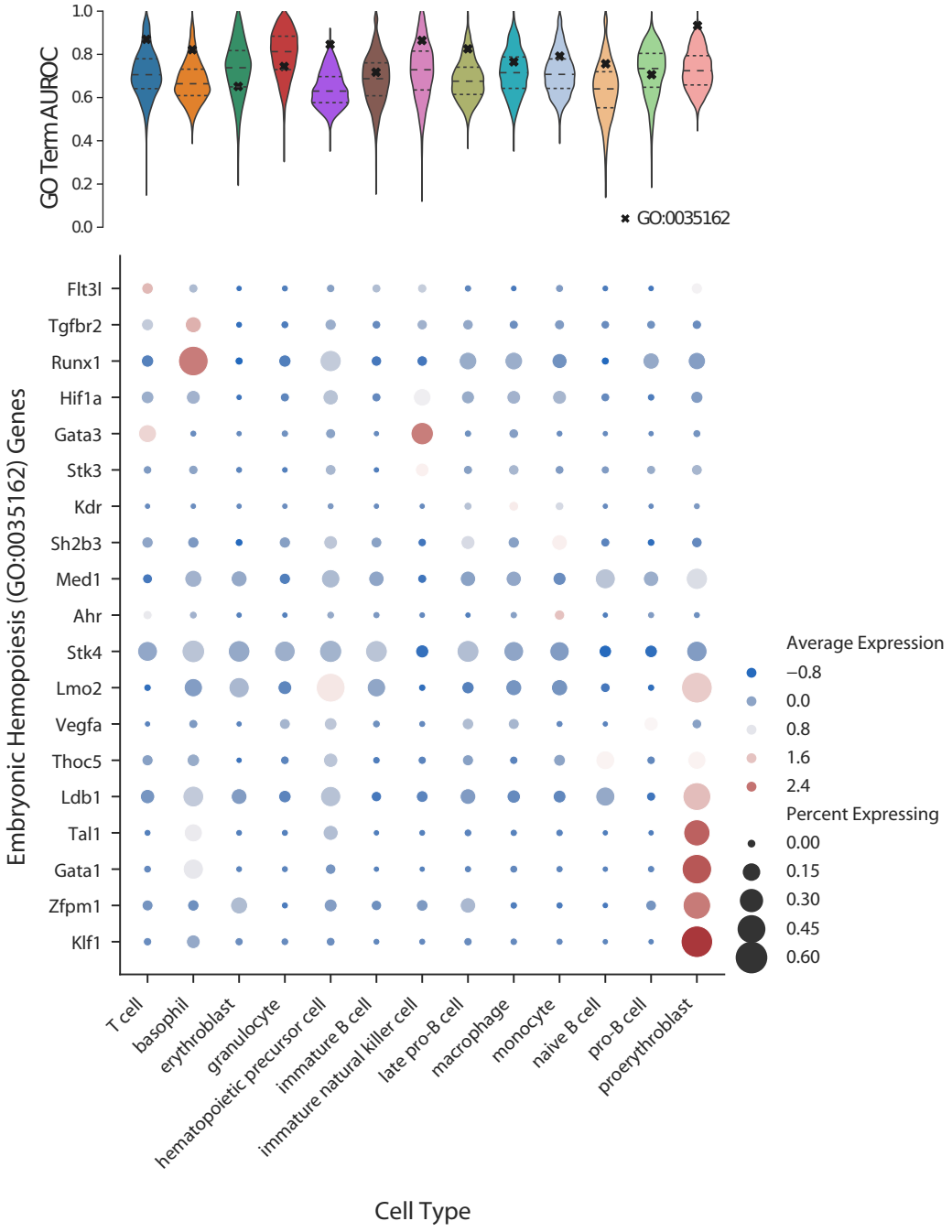


FIGURE 2.14: **Top** Violin plot of AUROCS for the MetaNeighbor results run on the Gene Ontology identifies many highly robust functional annotations that identify cell types. The AUROCs for the term Embryonic Hematopoiesis (GO:0035162) are marked for each cell type. **Bottom** Expression levels and % of cells expressed for the genes in the term Embryonic Hematopoiesis (GO:0035162) identify genes associated with the cell types that have the highest AUROCs from MetaNeighbor. The expression values are computed by Zscoring within datasets and then aggregating the values across datasets.

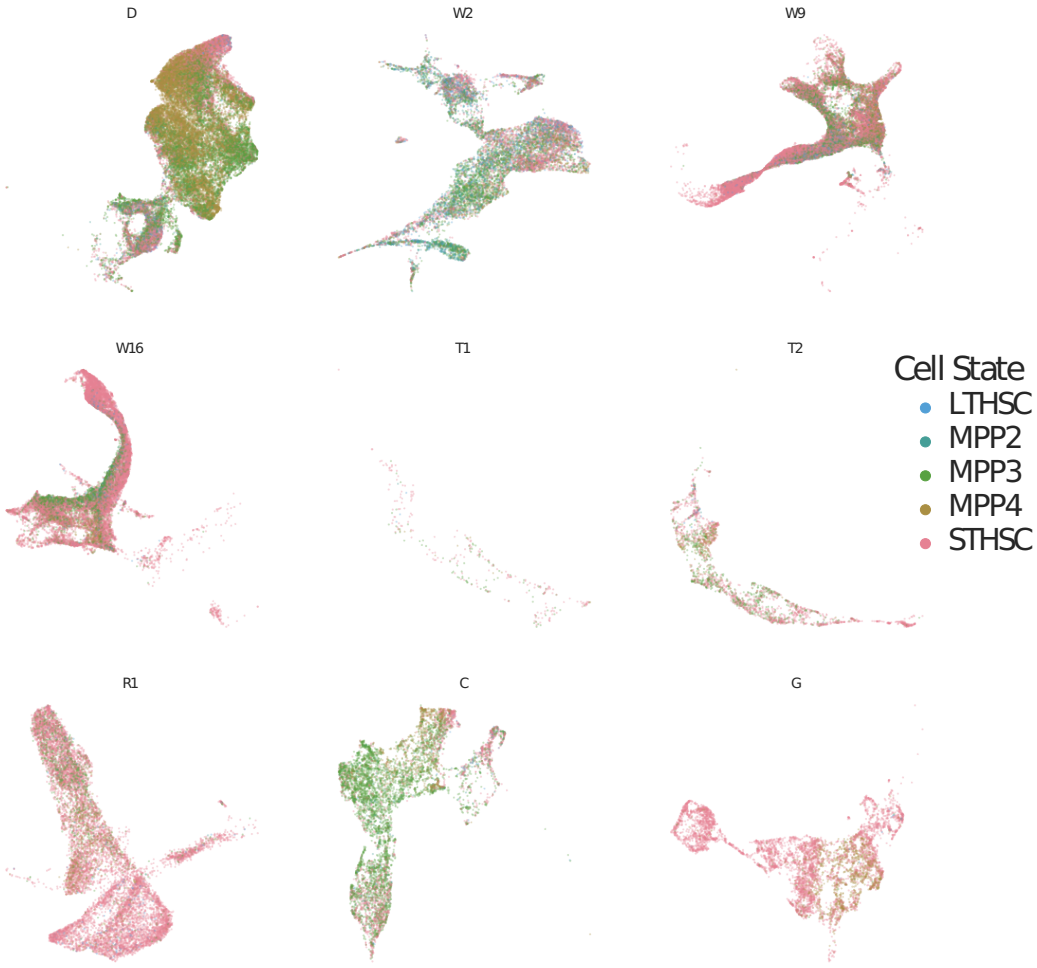


FIGURE 2.15: Cells from the hematopoietic precursor cell labeled as either LT-HSC, ST-HSC, and MPP2-4 do not cluster in UMAP space when projected by dataset.

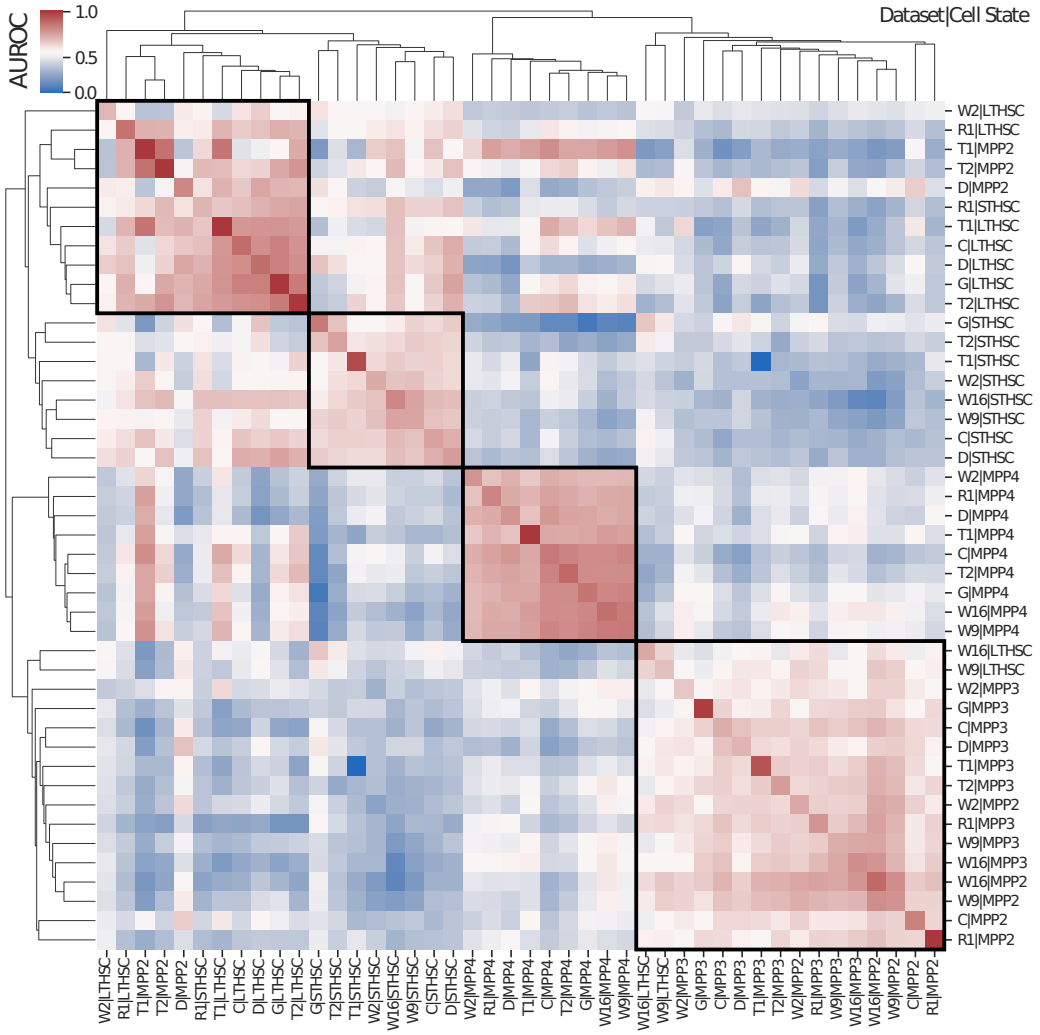


FIGURE 2.16: MetaNeigbhor unsupervised analysis shows consistency of MPP4s across datasets and moderate replicability of the other cell states.

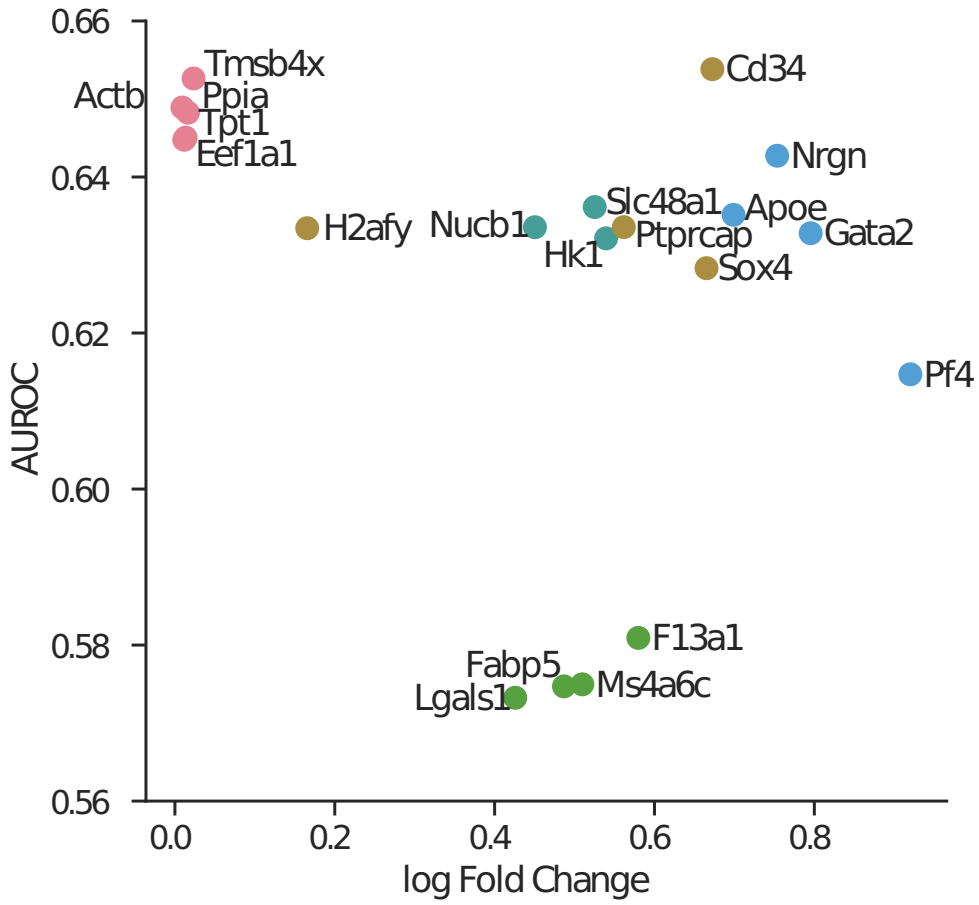


FIGURE 2.17: MetaMarkers defined cell type markers show limited significance (AUROC) and weak effect sizes (log Fold Change). The colormap follows Figure 2.15.

with 1 gene but steadily increases with the number of markers. To look at the variation across datasets we learn the top 10 markers for each cell state in 8 datasets and measure how well they classify the held out (test) dataset. The average AUROC across all the tests is .71, but with considerable variability between the different datasets and cell states (Figure 2.19). Classifying these cell states across datasets provides modest performance. The MetaNeighor and marker classification analysis identify replicable axes of variation, even if not the primary ones that would be visible in UMAP space.

We evaluate the replicability of functional connectivity of gene sets within the cell states using MetaNeighbor. Most of the 5516 tested GO terms have consistently low

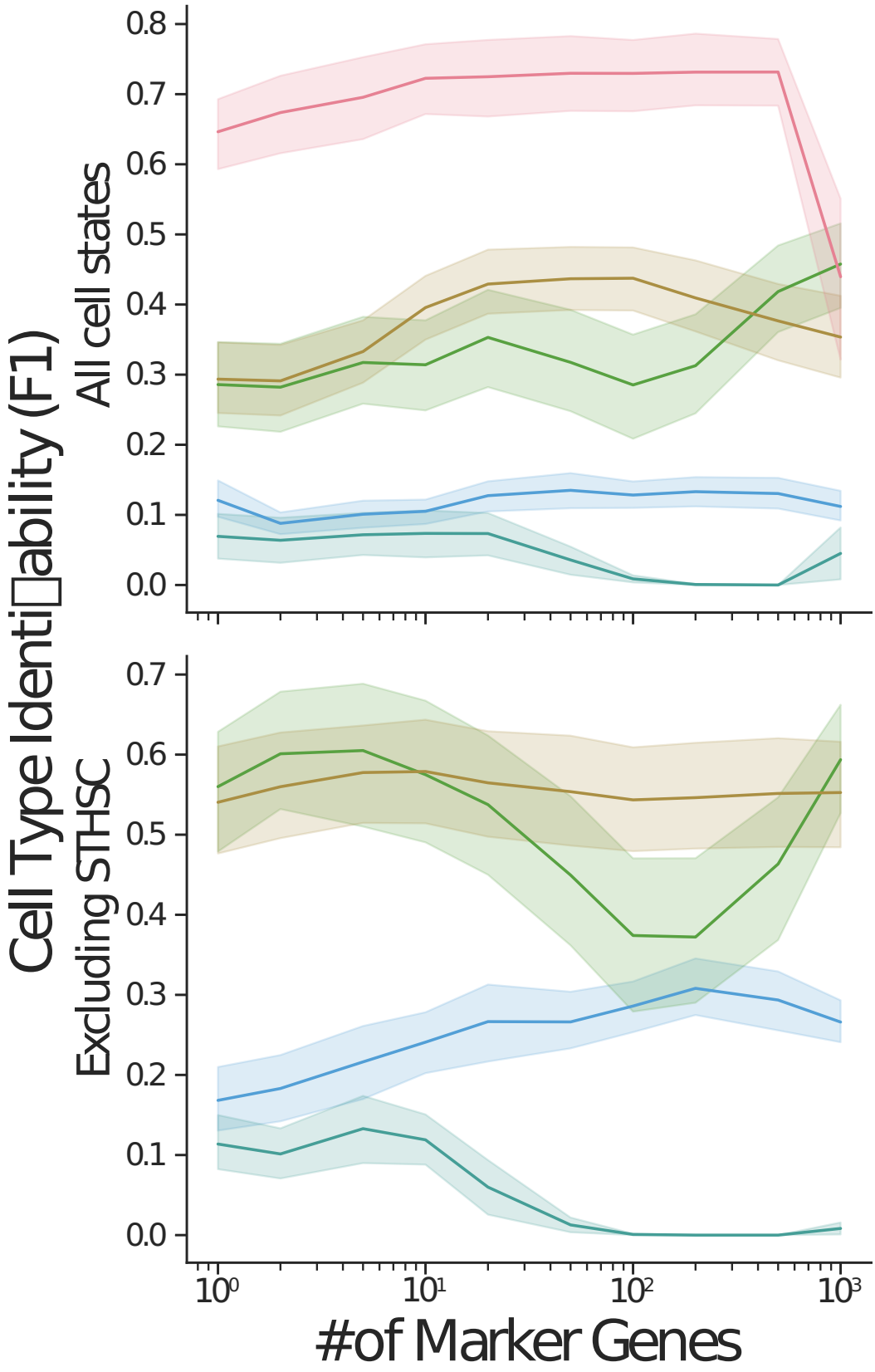


FIGURE 2.18: Evaluating the identifiability (F1 score) of cell states using 1-1000 markers in all cell states (top) and excluding the ST-HSC cell type (bottom). Computed using leave one out cross-validation. The shaded region represents 1 standard deviation. The colormap follows Figure 2.15.

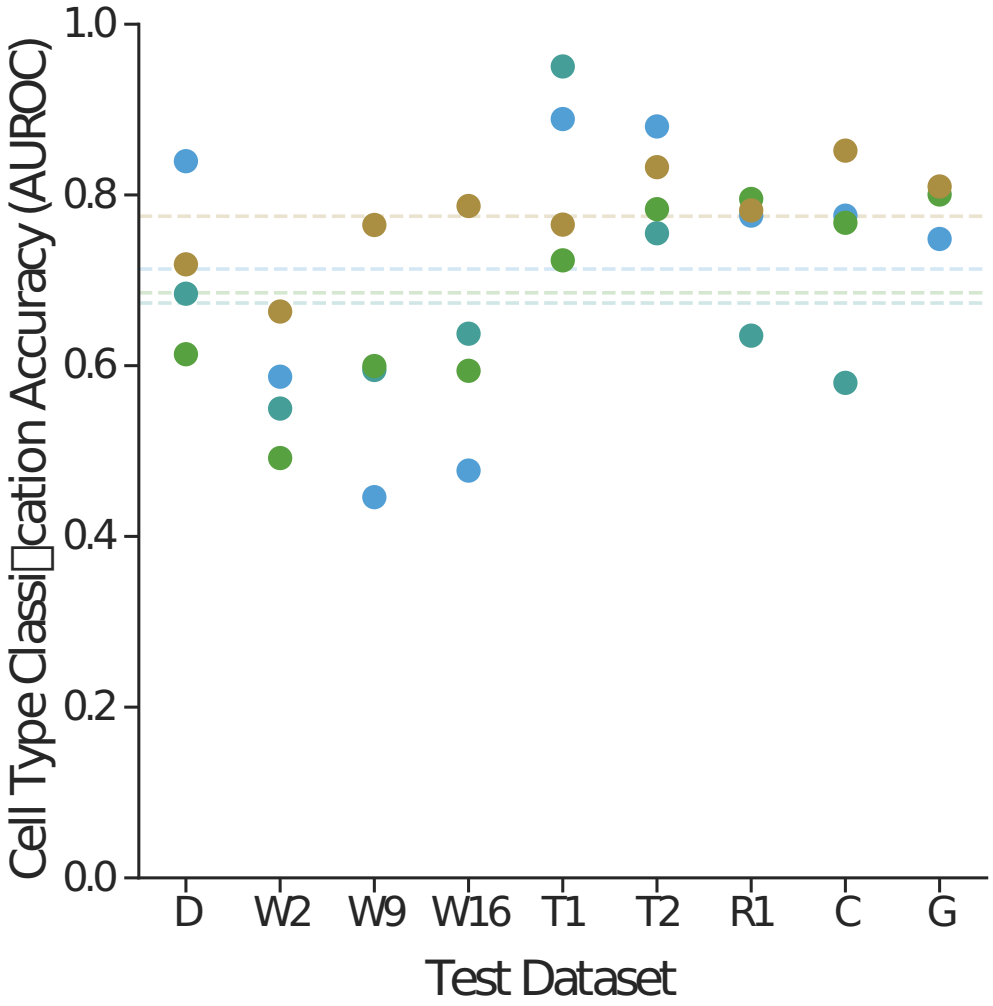


FIGURE 2.19: Classification performance (AUROC) using the top 10 marker genes. The model is trained on 9 datasets and the performance is shown for the 9th held out dataset. The dashed lines are the average across all folds. The colormap follows Figure 2.15.

AUROCs across all cell states (Figure 2.20, Supplementary Table 3). However, 81 terms have an AUROC >.9 in at least 1 cell state (Figure 2.20). Within top enriched terms, we see that many match the known differentiation bias of MPPs. The GO term "Lymphocyte Proliferation" (GO:0046651) has an AUROC of .98 in the MPP4 cluster. MPP4s are also referred to as the Lymphoid Multipotent Progenitor (LMPP) and have a significant bias towards differentiating into the lymphoid lineage (Adolfsson et al., 2005). The expression patterns for the genes in the term are displayed in a dotplot (Figure 2.21). The most variably expressed genes in the term show expression patterns consistent with bulk sorted cell populations from the Immgen Consortium (Figure 2.21, (Yoshida et al., 2019)). Rag2 and Il7r are standard markers for B and T cell development and Satb1 promotes lymphocyte differentiation (Satoh et al., 2013). The enrichment of the lymphoid proliferation term and lymphoid-associated genes could indicate that the cells in the MPP4 cell state are lymphoid primed. While not the primary axis of variability, these cell states constitute a replicable axis of variation within the hematopoietic precursor cell cluster associated with lineage potential.

2.3.4 Robust Signatures of Hematopoietic Differentiation

Modeling the cells as an ordered continuum, instead of clusters, depicts the differentiation process within the data and can identify gene regulation dynamics specific to lineage determination. We model this by computing pseudotime in individual datasets to avoid learning trajectories that are artifacts of the integration process/batch effects (Figure 2.22). The pseudotime computed on the integrated space is markedly different for each dataset (Figure 2.23). In addition to producing an ordering of the cells, the algorithm assigns all of the cells to nodes along a tree that estimates the differentiation branching within the data. We associated each end segment of the trees to either root, erythroid, or monocyte based on gene expression and label all segments in the middle as intermediate. While the clustering includes lymphocyte cells, the individual dataset projections do not connect the lymphocyte cells to the root in the latent space and we can not compute a confident trajectory through the non-linear gaps in the latent space (Figure 2.24). Evaluating the

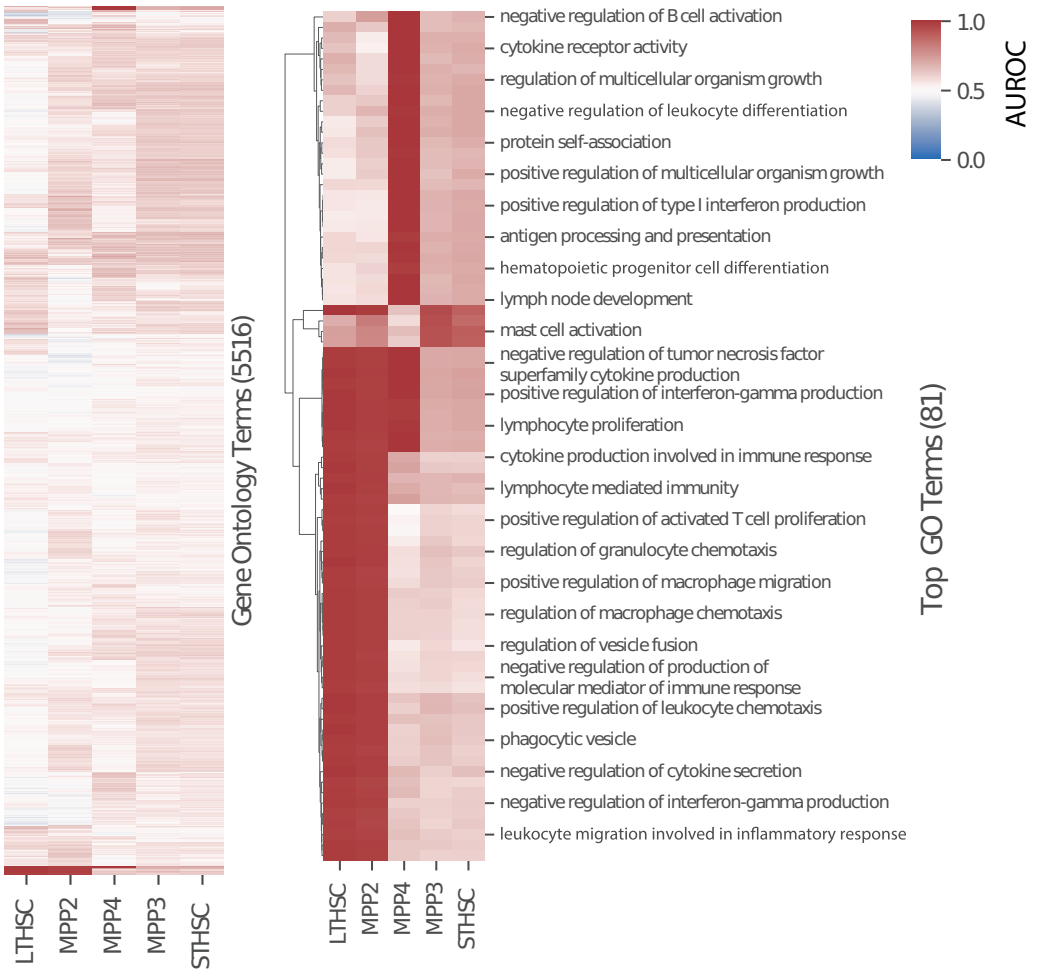


FIGURE 2.20: **Left** Metaneighbor evaluation of cell states using the whole Gene Ontology **Right** Subset of Metaneighbor results for terms with AUROC >.9 in at least 1 cell state

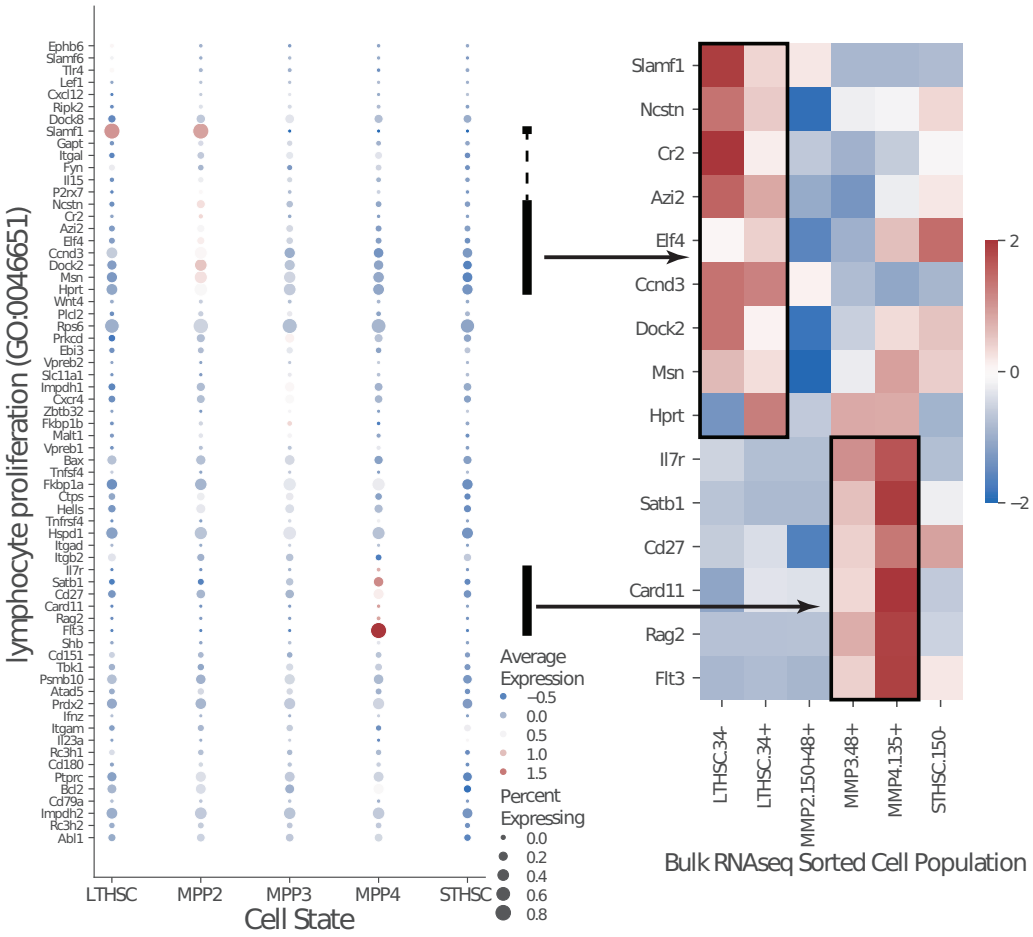


FIGURE 2.21: **Left** Expression of the term lymphocyte proliferation (GO:0046651) in each of the cell states. The Z-scores are computed within datasets and then aggregated across datasets. **Right** Bulk expression from ImmGen data for genes with notable expression in the single cell data matches single cell expression.

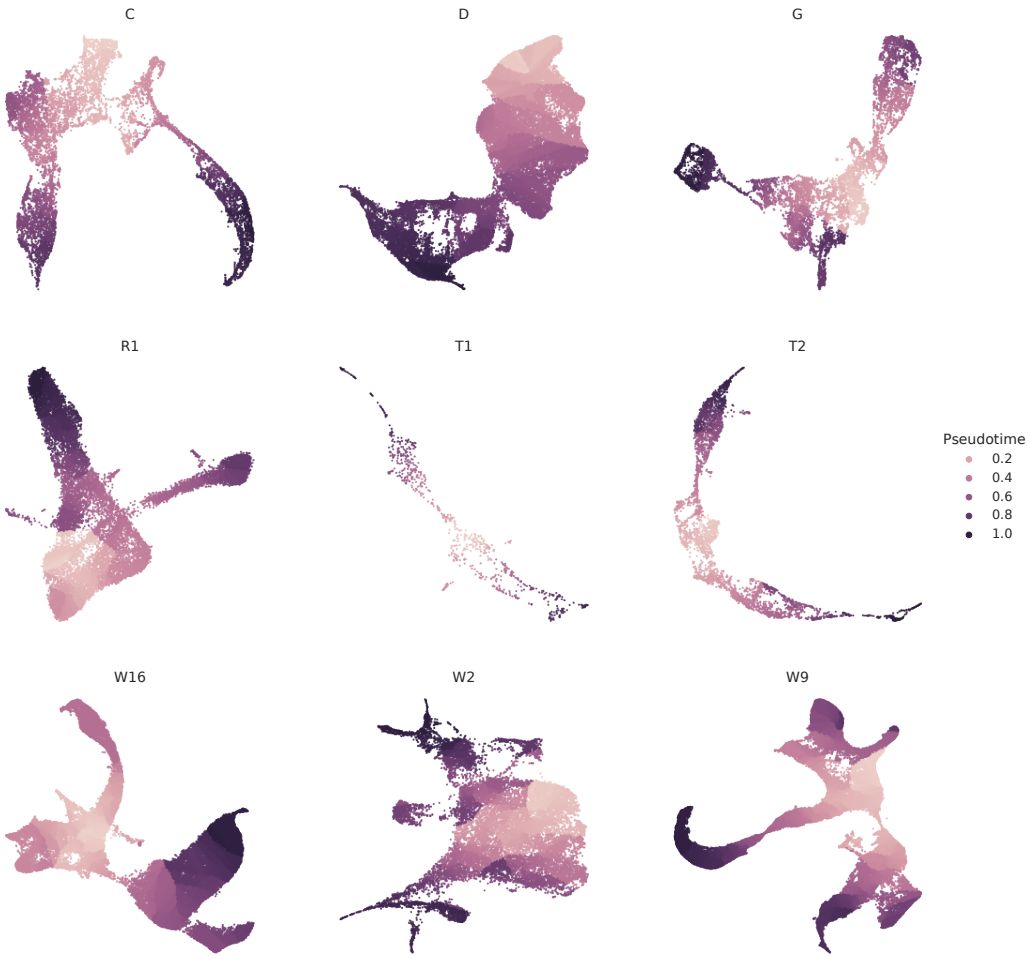


FIGURE 2.22: Individual datasets are projected into 2-dimensional space using UMAP and then Monocle3 learns a pseudotime ordering of the cells.

replicability of the segments using MetaNeigbhor shows that the root, erythroid and monocyte segments are replicable across the datasets, while the intermediate segments are not replicable across the datasets (Figure 2.25, 2.26). The inconsistency of the intermediates could be a result of the transient nature of intermediate cell types or more technical issues with scRNAseq (Haghverdi et al., 2016; Song and Li, 2020).

Using a broader approach, we fit models for every gene to each dataset and use meta-analytic statistics to identify consistent gene expression signatures associated with the erythroid and monocyte lineages (Figure 2.27, Supplementary Table 4). The top 3-5 genes for lineages are very similar to the cluster level analysis for erythroid, but not for monocyte. However, looking at the top 50 genes for each dataset shows that only 19 for

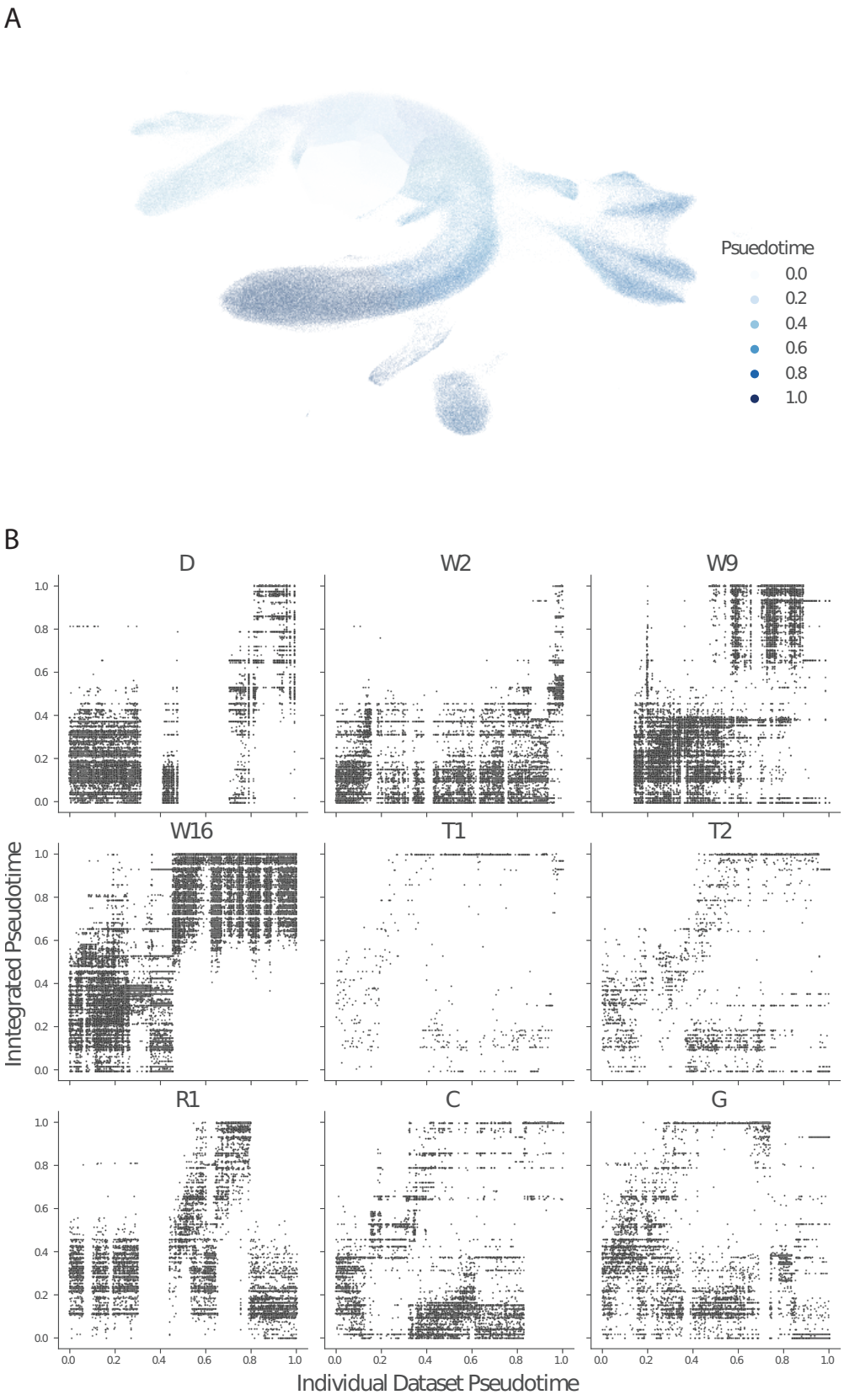


FIGURE 2.23: Pseudotime ordering of integrated space produces a different ordering of the cells than pseudotime within individual datasets. **Top** Ordering of cells computed by Monocle3 for the integrated latent space, only including cells that were used in the individual dataset pseudotime analysis. **Bottom** Pseudotime ordering of individual datasets compared is different from the integrated ordering

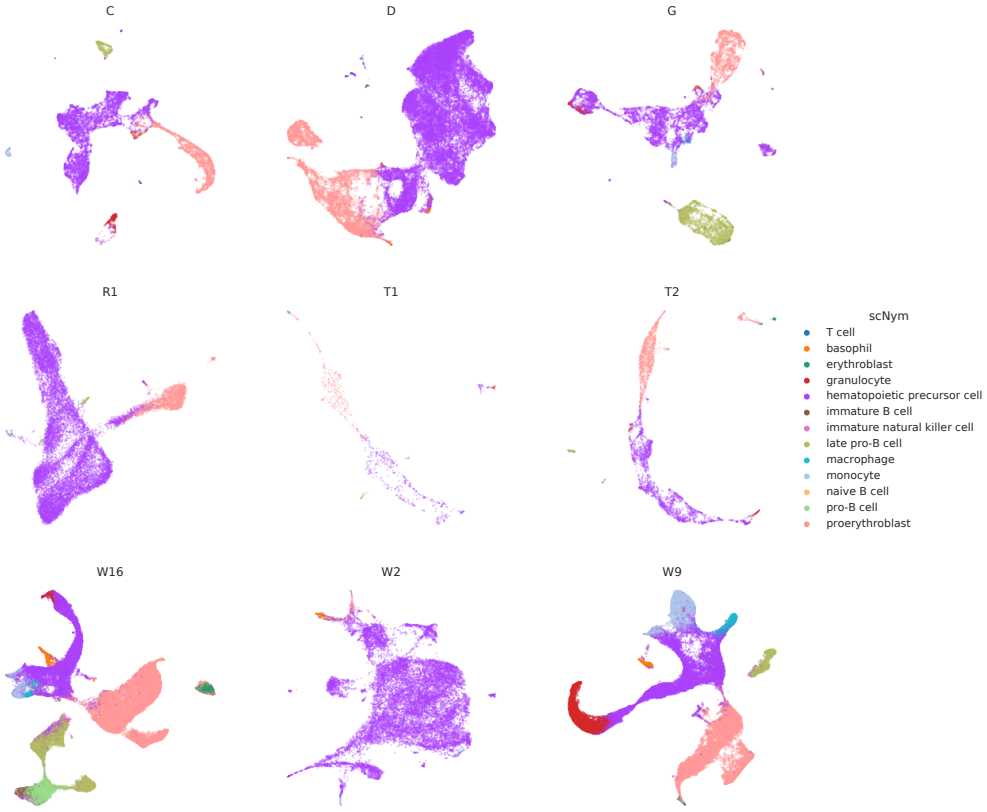


FIGURE 2.24: Projecting individual datasets using UMAP shows the lymphoid lineage (B + T cells) are disconnected from the stem cell (hematopoietic precursor cell) and other clusters. Computing a trajectory between disconnected clusters is ill-advised.

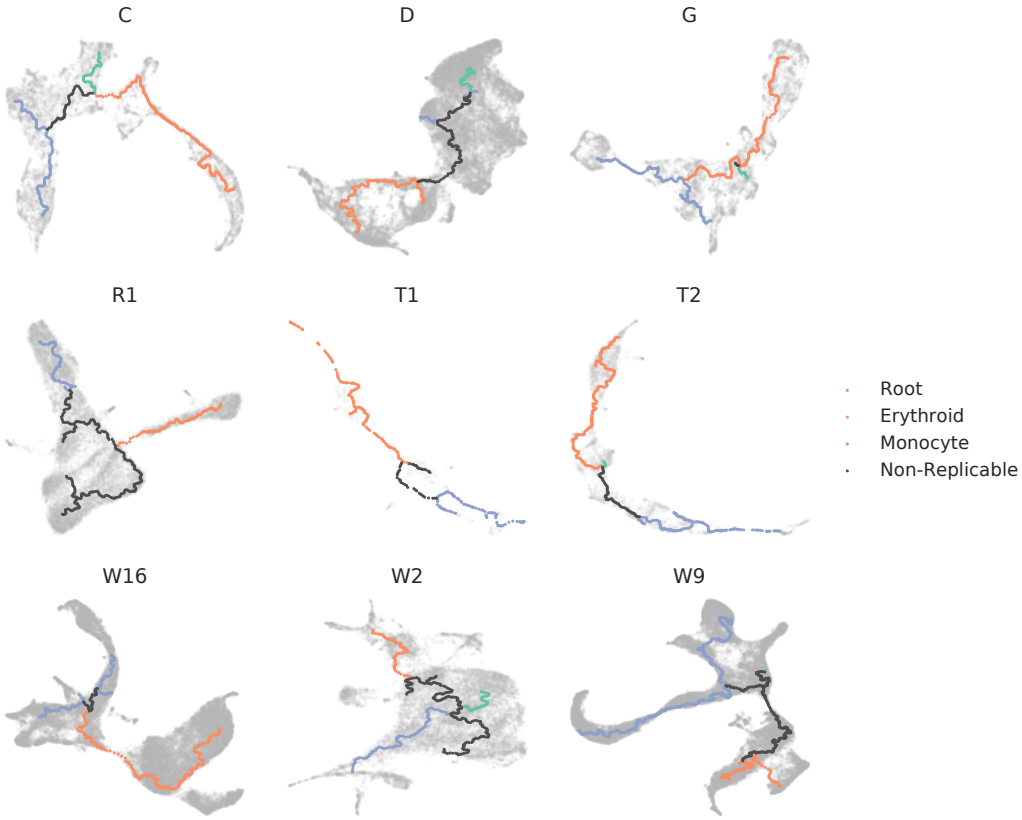


FIGURE 2.25: Branches of the pseudotime trajectories are assigned to either Root, Erythroid, Monocyte, or Non-replicable based on MetaNeighbor results 2.26. Each panel is a dataset

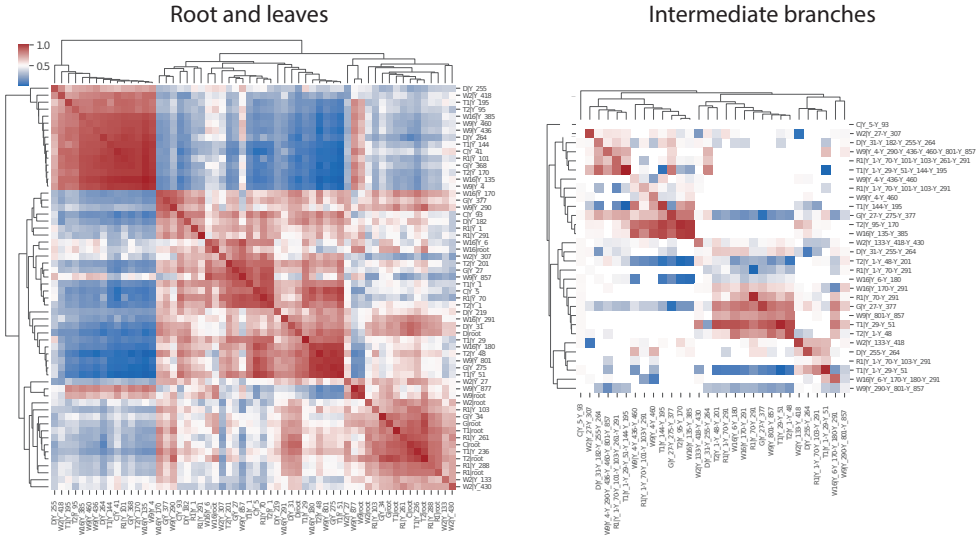


FIGURE 2.26: Unsupervised MetaNeighbor identifies replicable root and erythroid branches and non-replicable intermediate branches

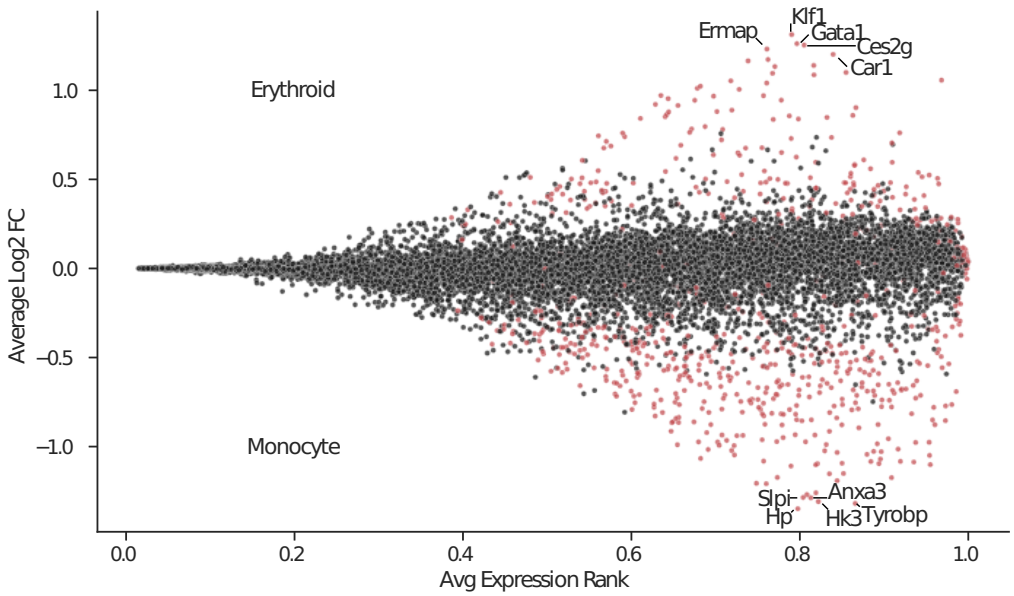


FIGURE 2.27: Meta-analytic MAplot of marker genes for Erythroid and Monocyte lineages.

erythroid and 4 for monocyte genes are shared between the cluster and pseudotime analysis (Figure 2.28). GO enrichment of the top 50 genes for each lineage identifies 11 for erythroid and 54 for monocyte significantly associated terms ($p<.05$, Figure 2.29, Supplementary Table 5). Visualization of the top 5 genes for each lineage ordered by pseudotime shows a consistent monotonic expression trend across the datasets (Figure 2.30). Despite the consistent monotonicity, each dataset has a unique inflection point where the gene expression substantially increases. The differences in timing across the datasets explain some of the replicability limitations of comparing the intermediate cells across datasets.

2.3.5 Cross-Species Co-Expression of Lineage Signatures

Co-expression networks reflect the functional landscape of gene expression (Eisen et al., 1998). Reference, bulk-RNAseq derived, co-expression networks are used to evaluate the cross-species relevance of the lineage-associated gene lists (Lee et al., 2020). We measure the connectivity (AUROC) of the erythroid and monocyte gene lists using these co-expression networks (Figure 5A). Strong connectivity, or high AUROC, of a gene set indicates shared function. As expected, the highest co-expression for both gene lists is in the

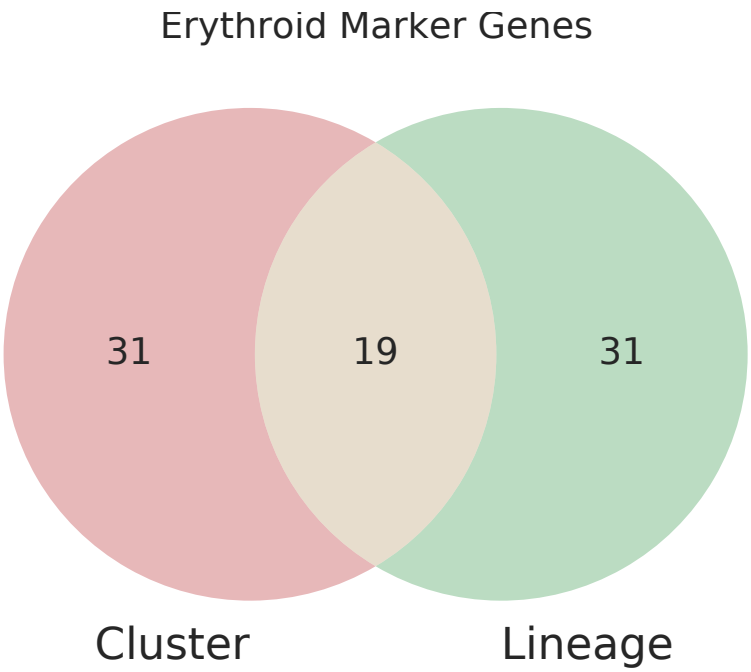


FIGURE 2.28: Modest overlap between the top 50 markers from cluster-level analysis and pseudotime analysis.

mouse network; the training species for the gene lists (monocyte AUROC=.92, erythroid AUROC=.90). Using 1-to-1 orthologs we evaluate the co-expression of the gene lists in 21 species. The monocyte gene list is more co-expressed in most species than the erythroid gene set. At the extreme is zebrafish, with near-random co-expression for erythroid (AUROC=.42) and strong co-expression for monocyte genes (AUROC=.81). Strikingly, both gene sets perform well in the human co-expression network, indicative of strong mouse-human conservation, an encouraging sign for translational research purposes (monocyte AUROC=.88, erythroid AUROC=.82).

In addition to evaluating conservation using co-expression networks, we look at the expression of the gene sets in a zebrafish hematopoietic dataset (Figure 2.32, (Xia et al., 2021)). The monocyte scores are bimodal, with the highest scoring cells matching the cells labeled as myeloid progenitors in the original study (Figure 2.33, 2.34). They mostly have very low scores for the erythroid gene set, and many of the highly scoring cells are myeloid progenitor-labeled cells. We next assessed the human bone marrow dataset from Pellin et al 2019 to evaluate the expression of the gene sets in human data (Pellin et al., 2019). We

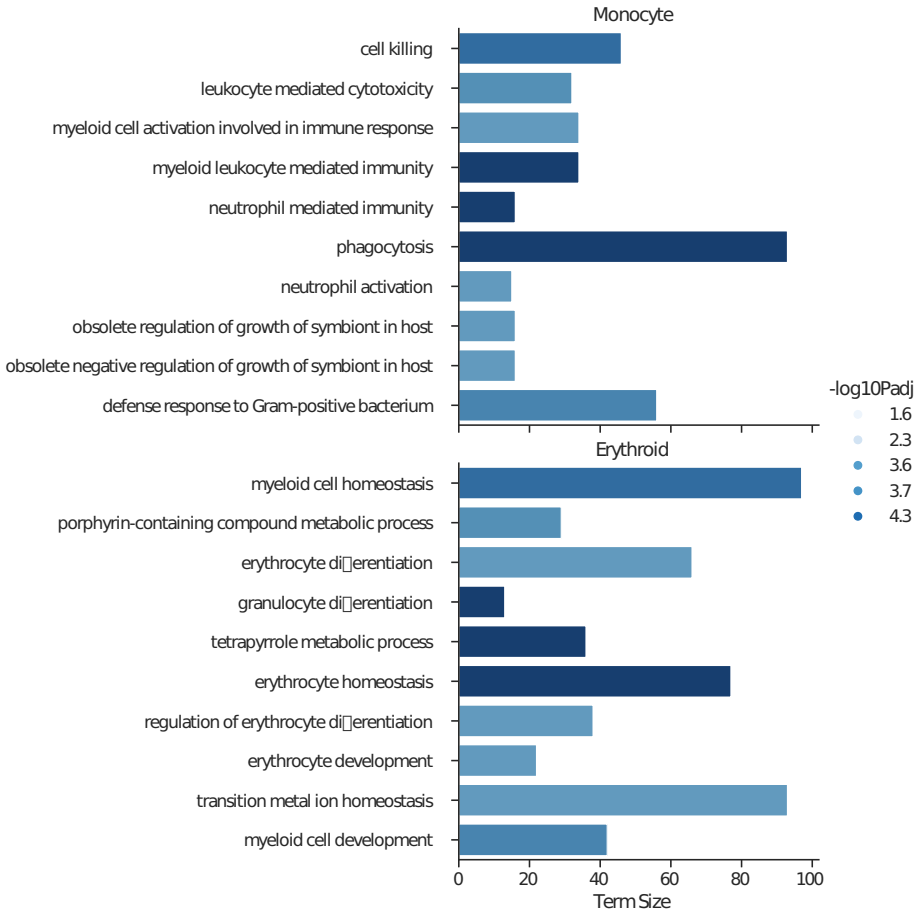


FIGURE 2.29: Top 10 terms from Gene Ontology enrichment using fisher's exact test for the top 50 makers for each lineage.

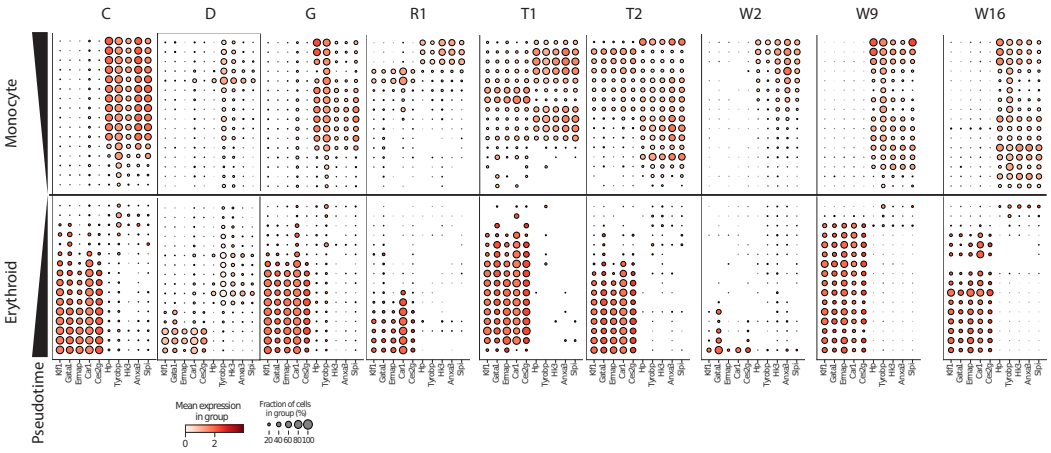


FIGURE 2.30: Expression of top 5 markers from each lineage across datasets ordered by pseudotime shows monotonic patterns but different expression profile dynamics between datasets.

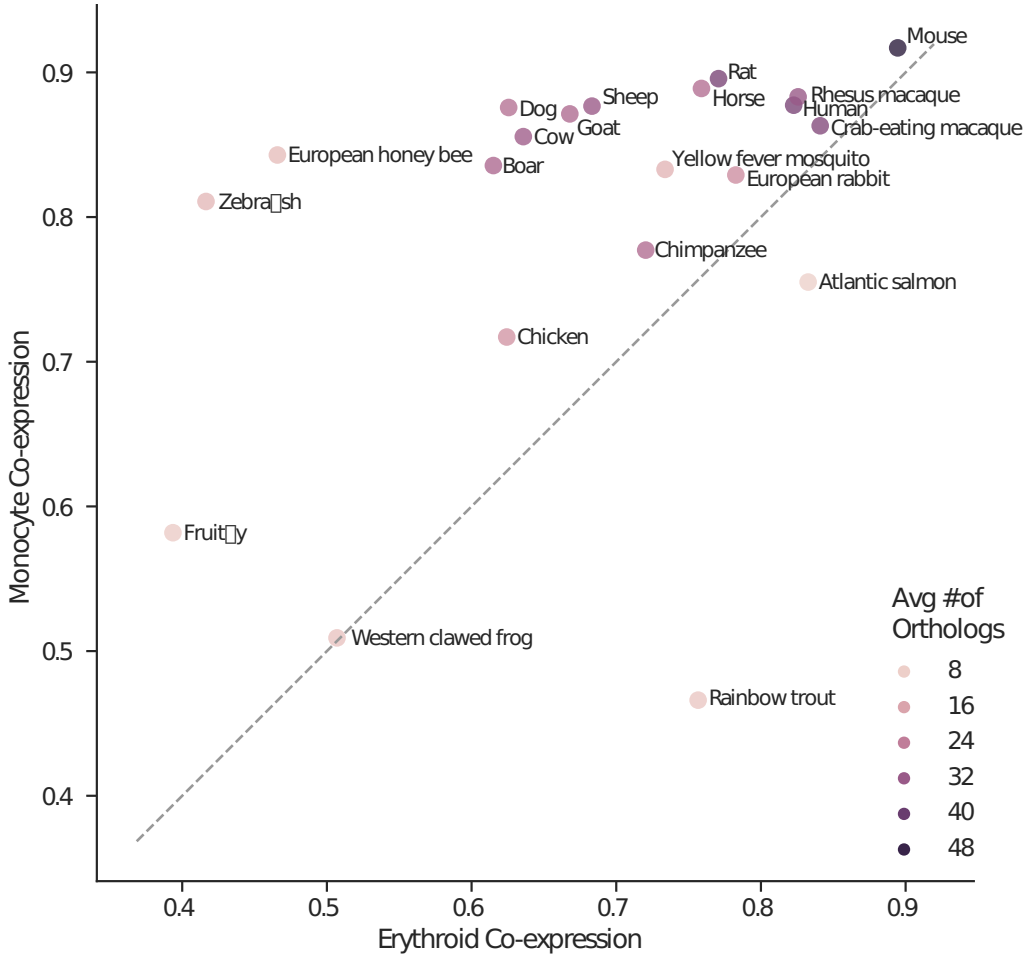


FIGURE 2.31: Co-expression of 1-to-1 orthologs across 21 species for both the erythroid and monocyte associated gene lists shows bias towards conservation of the monocyte lineage.

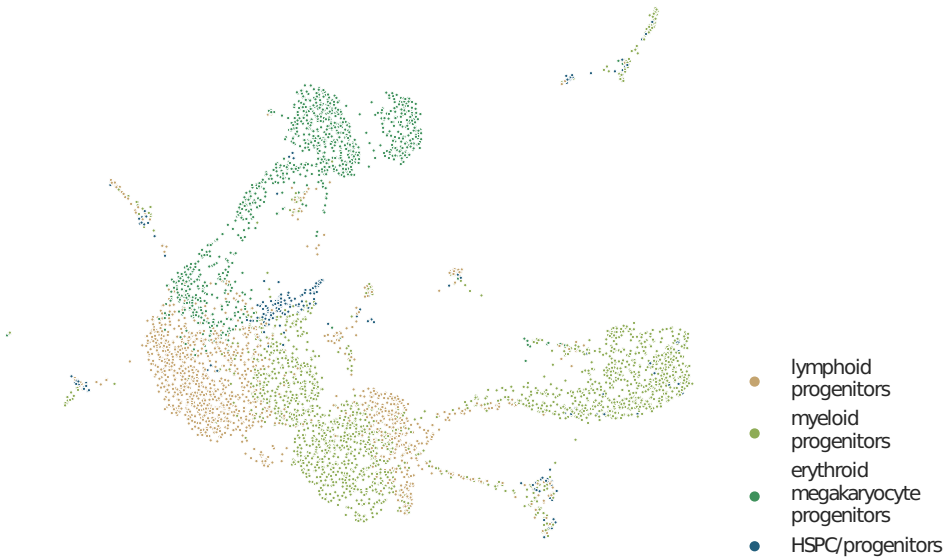


FIGURE 2.32: UMAP of Xia et al 2021 zebrafish hematopoietic dataset colored by cell type label.

rely on top markers from the publication to identify the HSCs, Monocyte, Erythroid, and Lymphocyte cell populations because discrete labels were unobtainable (Figure 2.35). The scores form increasing gradients from the HSCs to their respective lineage (Figure 2.36). The two lineage scores are orthogonal to each other, showing they serve as a good marker for the lineages (Figure 2.37). The orthogonal signatures show a binary fate decision between the erythroid and monocyte lineages; if one lineage score is upregulated, the other one remains inactivated. Between the co-expression and human scRNAseq results, it is clear that the functional relationship of the genes in the lineage-associated gene lists is conserved between mice and humans.

2.4 Discussion

Our results provide a robust evaluation of hematopoietic cell populations in mouse bone marrow. After identifying 9 hematopoietic datasets that broadly share cell types, we identified cellular populations using 3 different methods: clustering, in silico FACS sorting, and trajectory inference. These populations were characterized using both markers

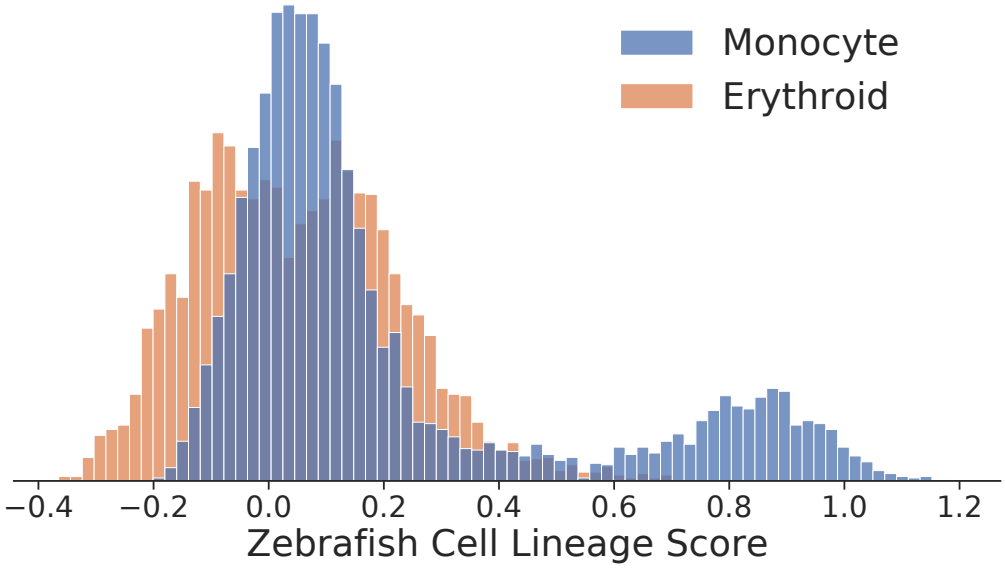


FIGURE 2.33: Histogram of lineage scores for each cell in zebrafish dataset has enriched population of cells for only the monocyte gene list

FIGURE 2.34: Lineage score for both monocyte and erythroid lineages plotted on UMAP of dataset shows only has specificity for monocyte lineage.

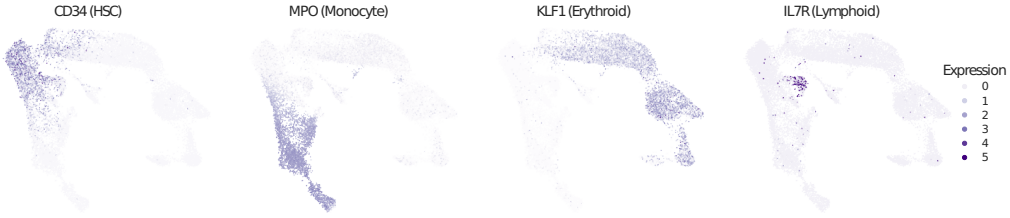


FIGURE 2.35: Expression of known markers for major hematopoietic lineages in the human hematopoietic dataset Pellin et al 2019.

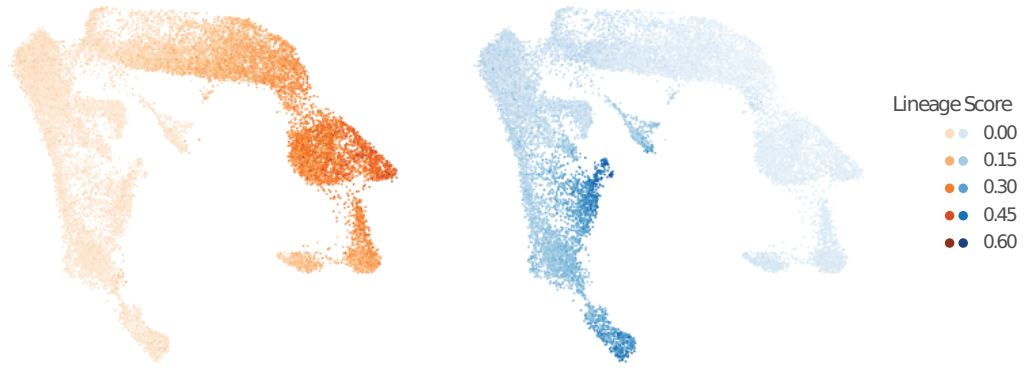


FIGURE 2.36: Scores for erythroid and monocyte lineages in the human dataset specifically identify both erythroid and monocyte cell populations

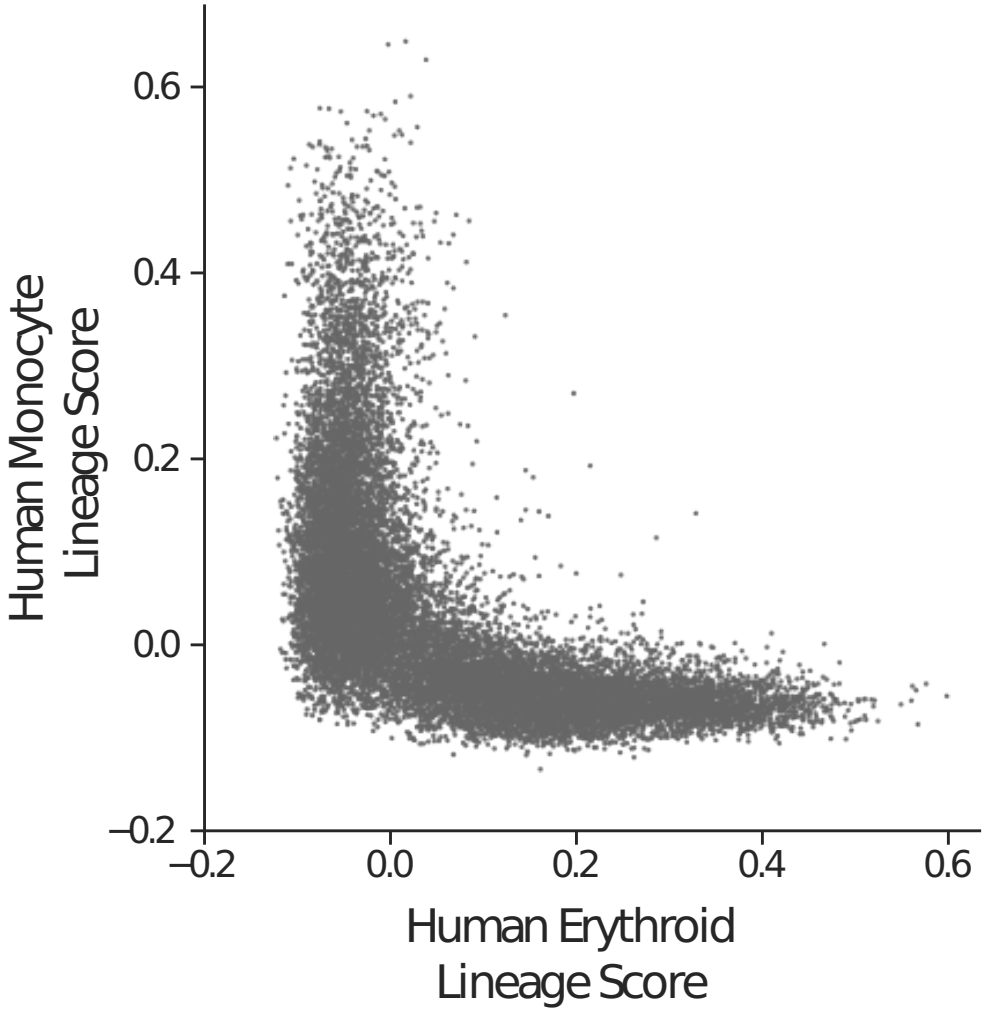


FIGURE 2.37: The lineage scores for each cell in the human dataset show the gene programs are orthogonal to each other.

and functional annotations. Furthermore, we demonstrated the conservation of lineage-associated genes using co-expression analysis across 21 species. Finally, we made the data and identified signatures accessible on our shiny webserver to compare with future experiments.

Meta-analysis serves to find robust signatures across datasets with significant technical variation (Cook and Vanderhyden, 2021), thereby determining what markers and properties are likely to generalize to new data. This meta-analytic atlas resolves technical limitations with individual batches to better represent the continuous nature of the system and provide strongly replicable signatures. The datasets sample cells unbiasedly from hematopoietic progenitors, recapitulating a developing system unlike the discrete, FACS sorting-based sampling in the Tabula Muris (Schaum et al., 2018). The most popular present resource for hematopoietic transcriptomic signatures is built from a single bulk RNAseq dataset, but has still been invaluable for basic research and studying SARS-Cov2, tuberculosis, and leukemia (Yoshida et al., 2019; Emmrich et al., 2021; Miyazaki and Miyazaki, 2021; Rothenberg, 2021; Erarslan-Uysal et al., 2020; Moreira-Teixeira et al., 2020; Blanco-Melo et al., 2020). By extending the availability of reference data to single cell and comparing across datasets, we enhance both the depth and breadth of transcriptomic signatures available to researchers.

The generalizability of our results will make them a valuable resource for translational research. An accurate reference of healthy hematopoietic stem cells is critical for identifying reliable therapeutic targets. While learning functional signatures of disease from clinical samples is often preferable, they can be difficult to acquire, and an alternative is to learn signatures associated with diseases from mouse models (Ketkar et al., 2020; Basilico et al., 2020). In order to identify disease signatures, correctly identifying cell types in healthy conditions is critical for evaluating changes in expression or abundance. Disease-associated signatures identified in single-cell data could then be evaluated as arising from changes in expression within cell types, or changes in cell type proportions (Liang et al., 2021; Zhao et al., 2021). Importantly, our cross-species analysis shows that we can evaluate the conservation of signatures identified in mice to human data, demonstrating the

atlas' utility for pre-clinical therapeutic research.

Here, we focus on the integration of one data modality, scRNAseq, but we expect additional modalities to be incorporated as data continues to be generated and robust meta-analysis can be conducted. In general, expression data serves as a foundation for the integration of other data modalities, providing robust signatures which can then be annotated by the data used in other modalities (Yao et al., 2020b). A cross-dataset, multi-modal atlas will resolve limitations and produce a more detailed picture of the gene regulatory networks driving hematopoiesis. Integrating CITE-seq data, which measures cell surface protein expression and RNA with this atlas will resolve the progenitor states better than in silico FACS sorting (Stoeckius et al., 2017). Single cell ATACseq data from mouse bone marrow will identify transcription factors and cis-regulatory elements important to lineage commitment (Ranzoni et al., 2020). CRISPR screens will test lineage-specific gene dependencies (Adamson et al., 2016; Jin et al., 2020). Cell non-autonomous signaling influences lineage commitment, either from the non-hematopoietic cells in the bone marrow or cell-cell communication between hematopoietic populations (Wang et al., 2019; Xue et al., 2019; Li et al., 2021a). Evaluating such cell-cell interactions will identify external signals that dictate lineage commitment. More data covering gaps in continuity, particularly the lymphoid lineage, will generate a more complete atlas— of great utility for studying lymphoid malignancies. Integrating other modalities with our robust scRNAseq atlas will resolve gaps in the atlas and produce a high-resolution picture of hematopoietic development.

This atlas serves as a reference for future hematopoiesis experiments that transition from FACS, the current gold standard, to RNA expression as the phenotypic measurement. In our results, we demonstrate multiple targeted analyses, made possible by a meta-analytic atlas and web server. Our analysis provides a detailed and robust evaluation of hematopoietic lineage development in mouse bone marrow. Our webserver makes it easy to evaluate the expression of any gene or known function identified in future experiments.

2.5 Methods

2.5.1 Data preprocessing

Data were downloaded for each dataset based on the info provided by their publication. For a detailed explanation, see the code for each dataset in the Github repository.

2.5.2 Integration using scNym

Data was normalized to logTPM as per the requirements for scNym. A column in the Anndata object was created that had the cell type labels from the two tabula muris datasets and the placeholder "Unlabeled" for cells from all other datasets. Additionally, we included a column in the obs data that denoted the batch. When training and testing the model we use batch as the domain. The output layer, consisting of 256 features was used as the input to UMAP. All of this was run on a server with a Nividia Tesla V100 GPU and the UMAP was done using the Nvidia rapids library.

scNym is a semi-supervised adversarial neural network. In being a semi-supervised method, it uses information from both reference (the two Tabula Muris datasets) and query (12 unlabeled datasets) to learn a representation of the cellular space. Since the tabula muris datasets were quite discrete, a result of the FACS sorting, and the query datasets are largely continuous, the method is able to learn a continuous space based on information from the query datasets while utilizing the labels in the reference data to label all cell types. By sharing batch as the domain to the tool, the adversarial network learns a representation that integrates the datasets to remove batch effects. Notably, I found 3 batches that did not integrate with the rest.

2.5.3 Marker identification and enrichment

We use the MetaMarkers package in R to compute cell type markers for both the scNym labeled cell types (Figure 2) and in silico sorted cell states (Figure 3) (Fischer and Gillis, 2021). MetaMarkers computes differential expression using the Mann-Whitney test within each batch and then computes meta-analytic statistics to aggregate the statistics

across batches. For the in silico analysis, we also use the $score\_cells$, $compute\_marker\_enrichment$, and $summarize\_precision\_recall$ functions to evaluate the identifiability and classification of cell states. Enrichment was done using the pyMN MetaNeighbor package and the mouse gene ontology.

2.5.4 Pseudotime

Pseudotime was computed using monocle3 on each dataset (Cao et al., 2019). We tuned the parameters $minimum\_branch\_length$ and $rank.k$ to balance the complexity of the trajectory and the coverage of the lineages. We used the monocle2 differentialGeneTest function to calculate the genes associated with each lineage and with branching (Qiu et al., 2017a). For GO enrichment we used custom code for Fisher's exact test (see GitHub) and the mouse gene ontology on the top 50 markers for each lineage.

2.5.5 Cross-species co-expression

We evaluated the co-expression of orthologs to the lineage-associated gene lists for every species in CoCoCoNet with at least 5 orthologs for both lineages using EGAD (Ballouz et al., 2016). In the human data Pellin et al 2019 and zebrafish dataset Xia et al 2021 we scored the expression of the orthologs using the Scanpy $score\_gene\_list$ functions.

2.5.6 Data and code availability

The code for all analysis is available in the GitHub repository `https://github.com/bharris12/hsc_paper` and processed data is available on the FTP site `ftp://gillisdata.cshl.edu/data/HSC_atlas/`. The data can also be explored in the Shiny app at `https://gillisweb.cshl.edu/HSC_atlas/`

Chapter 3

Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain

This chapter is a manuscript that has been published in Cell Systems (Harris et al., 2021). I conducted all the experiments and wrote the manuscript with help from the co-authors.

3.1 Summary

Gene-gene relationships are commonly measured via the co-variation of gene expression across samples, also known as gene co-expression. Because shared expression patterns are thought to reflect shared function, co-expression networks describe functional relationships between genes, including co-regulation. However, the heterogeneity of cell-types in bulk RNAseq samples creates connections in co-expression networks that potentially obscure co-regulatory modules. The Brain Initiative Cell Census Network (BICCN) single-cell RNA-sequencing (scRNA-seq) datasets provide an unparalleled opportunity to understand how gene-gene relationships shape cell identity. Comparison of the BICCN

data (500,000 cells/nuclei across 7 BICCN datasets) to that of bulk RNAseq networks (2,000 mouse brain samples across 52 studies) reveals a consistent topology reflecting a shared co-regulatory signal. Differential signals between broad cell classes persist in driving variation at finer levels, indicating that convergent regulatory processes affect cell phenotype at multiple scales.

3.2 Introduction

Co-expression networks characterize genes as related based on their shared expression profiles across samples. A shared profile suggests their activity is driven by the same factors or that they are functionally related (Eisen et al., 1998). Networks built from bulk gene expression data have been widely observed to recapitulate known gene functions (Eisen et al., 1998; Lee et al., 2004). As a result, co-expression analysis serves many applications in genomics. For example, co-expression has been used to infer transcription factor binding and causal regulation of downstream targets (Fiers et al., 2018; Kulkarni et al., 2017; Kulkarni et al., 2017; Song et al., 2016), characterize disease (Torkamani et al., 2010), and to predict which cells will interact with each other based on ligand-receptor pairs (Torkamani et al., 2010; Efremova et al., 2020).

Yet because cell-type composition is a major factor driving expression variation in bulk expression data, a substantial fraction of co-expression in bulk data is likely to be driven by variation in cell-type abundance, even if only indirectly through changes in abundance across other conditions (e.g., disease)(Farahbod and Pavlidis, 2020; McCall, Illei, and Halushka, 2016; Zhang et al., 2019). Although some work has been done to use deconvolution to identify cell-type specific co-expression from bulk data (Kelley et al., 2018), other analyses show that compositional differences confound co-regulatory signal (Zhang et al., 2019; Farahbod and Pavlidis, 2020). Building networks from pure cell-type data, as from single-cell RNA-seq (scRNA-seq), has the potential to identify co-regulatory relationships between genes that may be hidden due to cell-type composition in bulk (Trapnell, 2015). However, if single-cell co-expression data differs dramatically from bulk data, it could be considered as a surprise, given the longstanding utility of co-expression from

bulk data (i.e., if bulk co-expression has been useful at capturing gene-gene relationships, how different should single-cell be?). Characterizing the overlapping and distinct signals from single-cell and bulk data remains a major challenge (Crow and Gillis, 2018) and most previous research into single-cell co-expression has been limited to individual datasets or meta-analysis across unrelated biological conditions (Feregrino et al., 2019; Skinnider, Squair, and Foster, 2019; Smillie et al., 2019; Mohammadi, Davila-Velderrain, and Kellis, 2019). Further analysis using more specific and powered data will advance our understanding of both regulatory and compositional co-expression signals.

The 7 mouse primary motor cortex scRNAseq datasets from the Brain Initiative Cell Census Network (BICCN), totaling over 500,000 cells/nuclei, provide a rich opportunity to comprehensively study cell-type specific co-expression networks in scRNAseq data (Yao et al., 2020b). The BICCN data is particularly useful for studying composition and co-regulation in networks because of the diversity and specificity of cell-types available. Specifically, cell-types are annotated at multiple levels of resolution, and are replicable across datasets, enabling meta-analysis of cell-type specific co-regulatory modules.

We investigate co-expression by comparing networks built from heterogeneous data and pure cell-types. We show that there is no dichotomy between cell-type composition and co-regulatory signals in co-expression. In other words, the same gene-gene relationships that differentiate cell-types are evident at both finer and broader scales. We illustrate these conserved regulatory relationships using direct topological comparisons, reference functional annotations like the Gene Ontology and KEGG, and most importantly, marker gene lists that define cell-types. All of our analyses show overlapping connectivity in both compositional and cell-type specific networks, revealing a consistent regulatory landscape that can be defined across all BICCN cell-types. Finally, we show that finding cell-type specific co-expression relationships will require substantially more data than is currently available.

3.3 Results

The BICCN data consists of 7 datasets produced using both SmartSeq2 and 10X genomics library preparation methods. There are datasets using both whole cell and nuclei samples, and both the V2 and V3 chemistries from 10x (Figure 3.1).

Across the datasets, the clusters are labeled using a consistent hierarchical taxonomy (Supplementary Figure 3.2). Our strategy is to build co-expression networks based on the known hierarchy of cell-types within the BICCN data, and to evaluate the co-expression of cell-type markers in networks that control for this source of variation. Specifically, consider two genes α and β. Each gene can be seen as a vector of expression values over all cells C. Let $C_1, \ldots, C_K$ be K cell types forming a partition of C, such that C = C_1 U C_2 U ... U C_K. We can then split vectors α and β according to cell types, for example α[C1] is the expression of gene α over cell type 1. We compute the within-cell-type co-expression as the average Pearson correlation (rank normalization omitted for clarity, see Methods) where α_j is the expression of gene α in cell j, and is the average expression of gene α in cell type i.

$$\overline{R} = \frac{1}{K} \sum_{i=1..K} \frac{\sum_{j \in C_i} \left(\alpha_j - \overline{\alpha[C_i]}\right)\left(\beta_j - \overline{\beta[C_i]}\right)}{\sqrt{\sum_{j \in C_i} \left(\alpha_j - \overline{\alpha[C_i]}\right)^2 . \sum_{j \in Ci} \left(\beta_j - \overline{\beta[C_i]}\right)^2}}$$

The final co-expression value can only be driven by within-cell-type correlation, as cell type specific trends are effectively removed in the form of the. At K=1 (single partition with all cells), co-expression is largely driven by average cell type specific trends, while for K » 1, all these trends have been controlled for.

Thus, for example, in the absence of cell-type partitioning, two genes which are highly expressed in cell-type A relative to cell-type B will be co-expressed in a network containing both cell-types since the genes are co-variable with respect to cell-type. Historically, this is the case in bulk expression data, where the co-expression of two marker genes for cell-type A will have been calculated from samples with varying proportions of the two cell-types (Figure 3.3). The fundamental question of single-cell co-expression is the degree to which novel covariation is present in cell-type A (or B) individually, reflecting

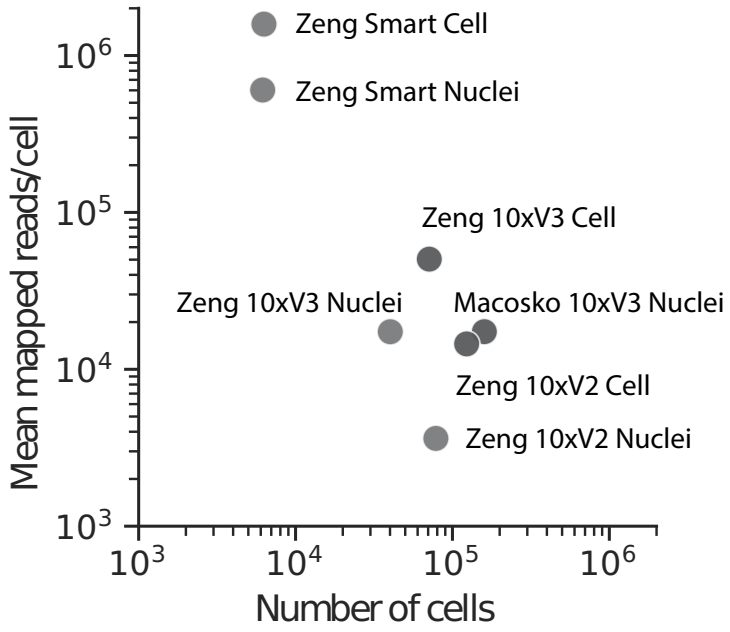


FIGURE 3.1: The single cell datasets were generated with multiple technologies, resulting in varying numbers of cells and varying read depths across datasets

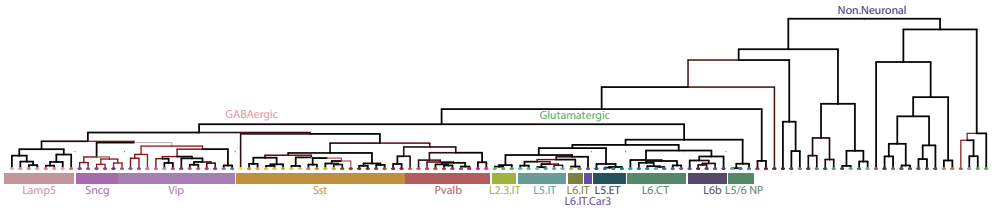


FIGURE 3.2: Dendrogram of cell-type hierarchy in scRNAseq datasets

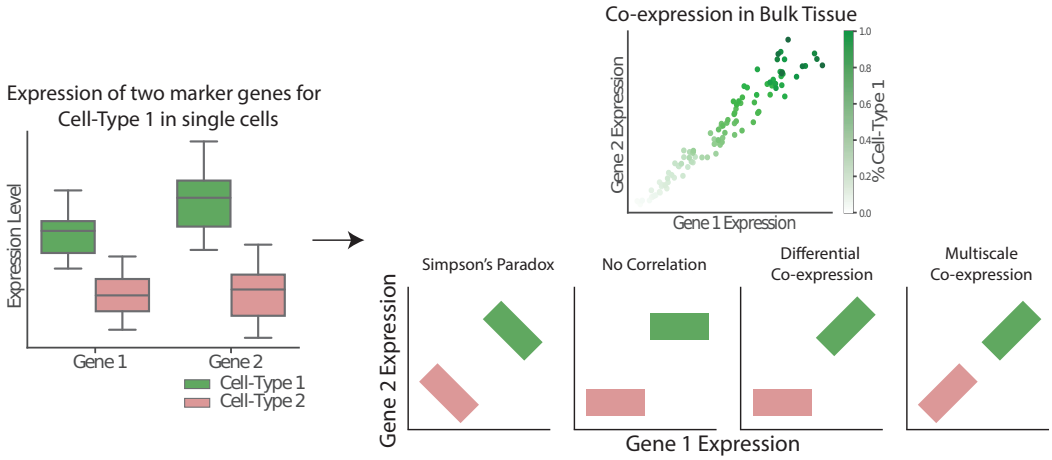


FIGURE 3.3: In bulk RNAseq data, marker genes must be co-expressed because of compositional differences in samples. Genes 1 and 2 represent hypothetical markers for Cell-type A. In bulk, the co-expression of the genes is co-linear with the percent of cell-type A within each bulk sample. Co-expression of the genes within the two cell-types could be any of the four models.

regulatory interactions rather than compositional effects.

Cell-type specific co-expression relationships can be described using at least 4 models: Simpson's paradox, no co-expression, differential co-expression, and multiscale co-expression. Single cell resolution data now makes it possible to quantify the occurrence of these models. The different models make assumptions about the relative direction of within cell-type co-variation versus that across cell-types. In the Simpson's paradox model, correlations between gene A and gene B take one sign across all cells, but reverse for subsets of cells corresponding to the cell types. Biologically, this would suggest a shared regulatory relationship (e.g. higher gene A expression is associated with lower gene B expression), which is reversed in bulk compositional data due to differential expression of the genes (e.g. expression of genes A and B are systematically higher in the first cell type). The no co-expression model is exhibited when a given gene pair is uncorrelated within each cell-type, but is co-expressed when considering both cell-types together. This would suggest that markers are not directly co-regulated within cell-types and are simply differentially expressed across cell types. A differential co-expression model, where one cell-type exhibits a significant correlation between two genes, while the other cell type has the opposite or no correlation would suggest co-regulatory network rewiring. If we

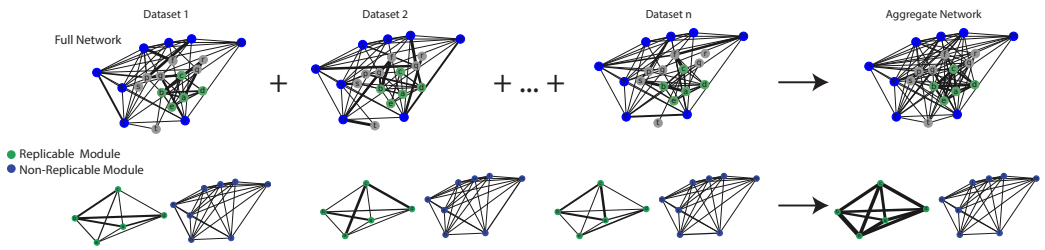


FIGURE 3.4: Meta-analysis across dataset networks identifies robust co-expression relationships. Thicker edges represent stronger co-expression. Aggregate networks give strong weight to replicable co-expression.

found this last model to be predominant, then cell-types would be defined by the creation of new gene-gene relationships. Finally, the multiscale model occurs when co-expression is similar in both bulk and single cell data. In this model, gene-gene relationships are consistent within and across cell-type, i.e. differential expression patterns align with the co-regulatory relationships, signifying modulation of the degree to which they are used.

We use the terms "co-regulatory" vs. "compositional" networks for those which do and do not control for cell-type variation, respectively. We use the term "network" to refer to the genome-wide weighted relationships between genes, and we identify robust co-expression relationships by using a meta-analytic approach (Figure 3.4). At the extreme end of defining co-regulatory gene interactions, we take advantage of "metacell" networks which measure gene-gene co-variation over statistically similar sets of cells. The metacells are smaller groups of 20-100 cells (Figure 3.5) that are significantly more homogenous than clusters (t-test between the distribution of distances for each cell to its respective metacell or cluster centroid, p 6e-23, Figure 3.5). By comparing gene-gene relationships that sample from more and more diverse cells, we incorporate increasing compositional effects across the types of cells sampled (e.g., subtypes of inhibitory cells). At the broadest level of analysis are brain-specific bulk co-expression networks, using samples made up of large numbers of cells. For our bulk analysis, we generated meta-analytic networks using 52 datasets of bulk mouse brain data from the Gemma database (Figure 3.6). Throughout, we focus on genes broadly expressed across cell-types and thus open to robust analyses of co-variation across and within cell-types.

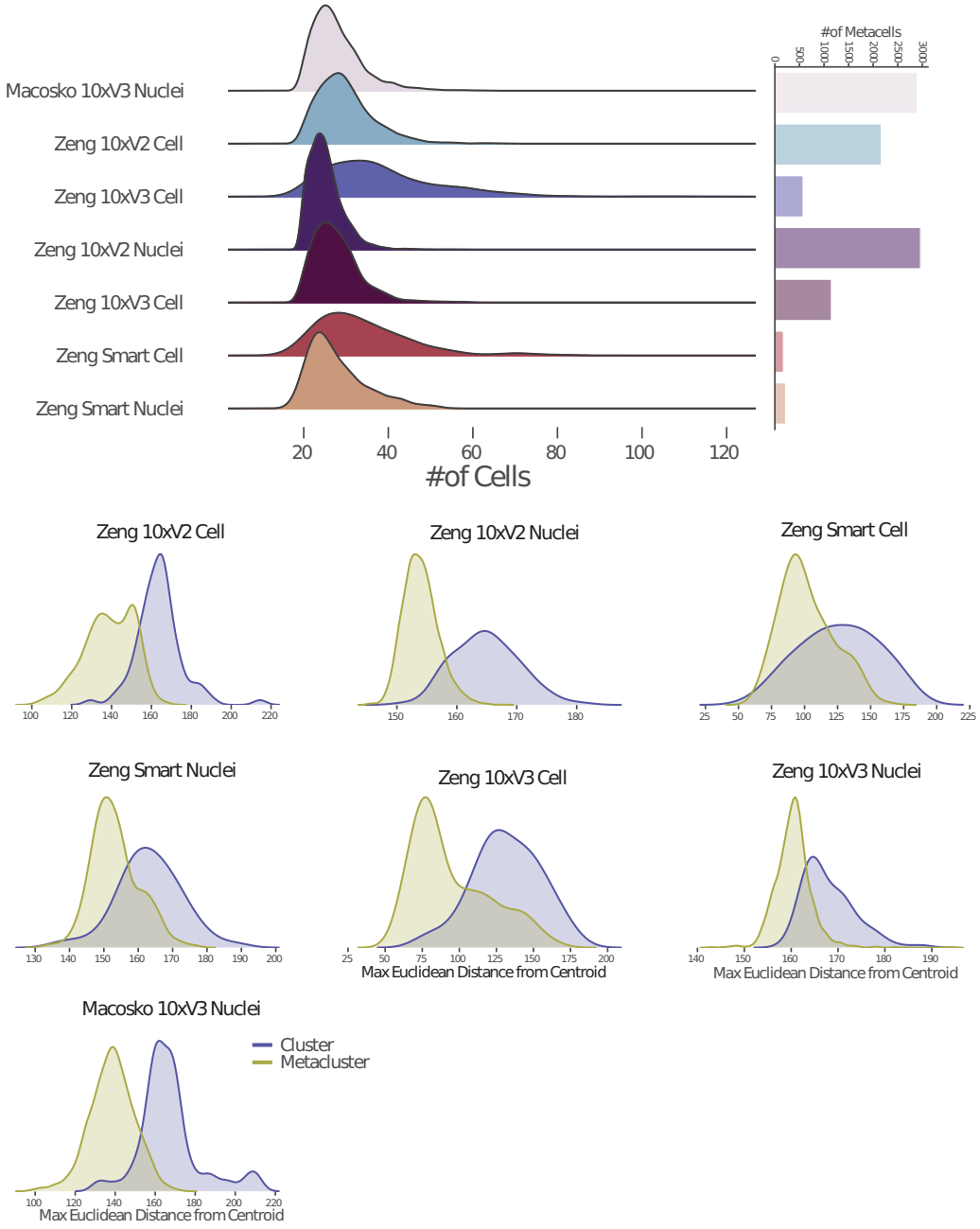


FIGURE 3.5: **Top** Number and size of metacells computed from scRNAseq datasets. **Bottom** Cells are closer to metacell cluster centroids than BICCN cluster centroids

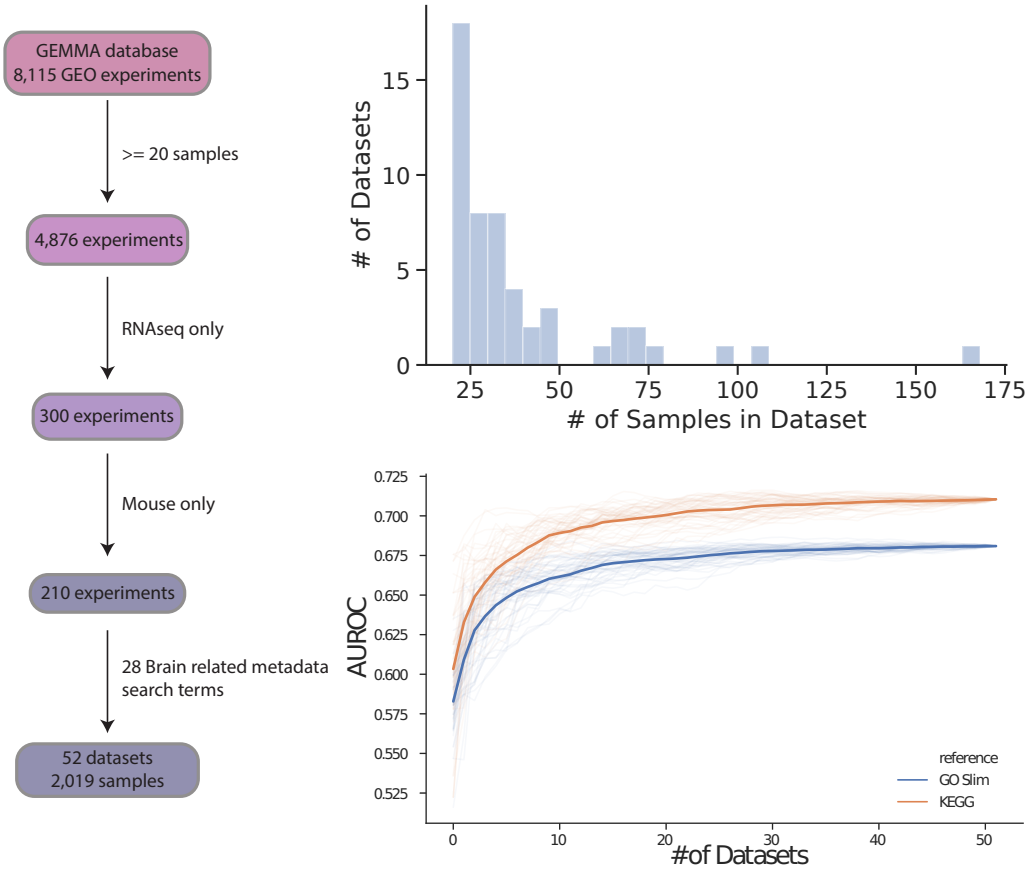


FIGURE 3.6: **Left** Co-expression of bulk RNAseq data from GEMMA. Outline of procedure for selecting datasets from GEMMA database. Distribution of dataset size (**Top right**) and performance of networks increases as networks are aggregated together (**Bottom right**).

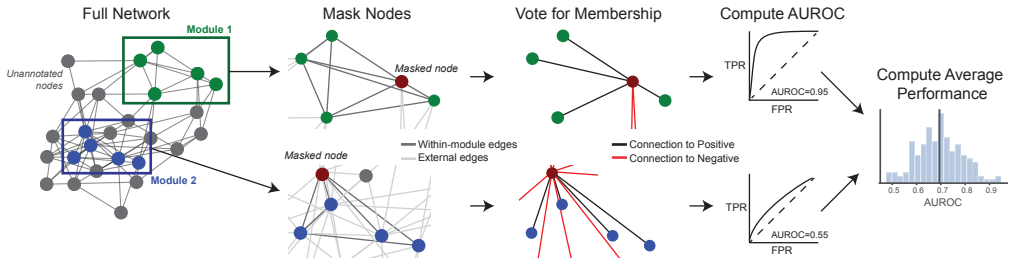


FIGURE 3.7: Guilt-By-Association algorithm assesses a network's ability to reconstruct modules.

3.3.1 A Consistent Topology Between Compositional and Co-Regulatory Networks

For our first experiment, we compared metacell to bulk RNA-seq co-expression networks in order to capture similarities and differences at the greatest range of the spectrum (see Methods for details on network construction). We first observe that both networks reflect known biology using a guilt-by-association formalism, in which each network is measured for its ability to reconstruct a partially hidden gene list from preferential connectivity within it, outputting an Area under the ROC curve (AUROC) (Figure 3.7). In the metacell network, the average AUROC across all GO slim and KEGG functional groups are 0.64 and 0.63 respectively, and similarly the average AUROC of the bulk RNAseq network is 0.67 for GO slim and 0.70 for KEGG (3.8). We also find that these networks have highly similar topologies. A comparison of coarse hierarchical clustering of both co-expression networks shows large shared modules between the two networks, visualized as a riverplot in Figure 3.9. Moreover, the average AUROC of modules drawn from the metacell network in the bulk network is 0.84, and the same is true of the reverse analysis (Figure 3.9). This indicates that modules present in one network are present in the other to a very specific degree, which is surprising since these two networks were constructed using data that capture vastly different signals, compositional versus co-regulatory.

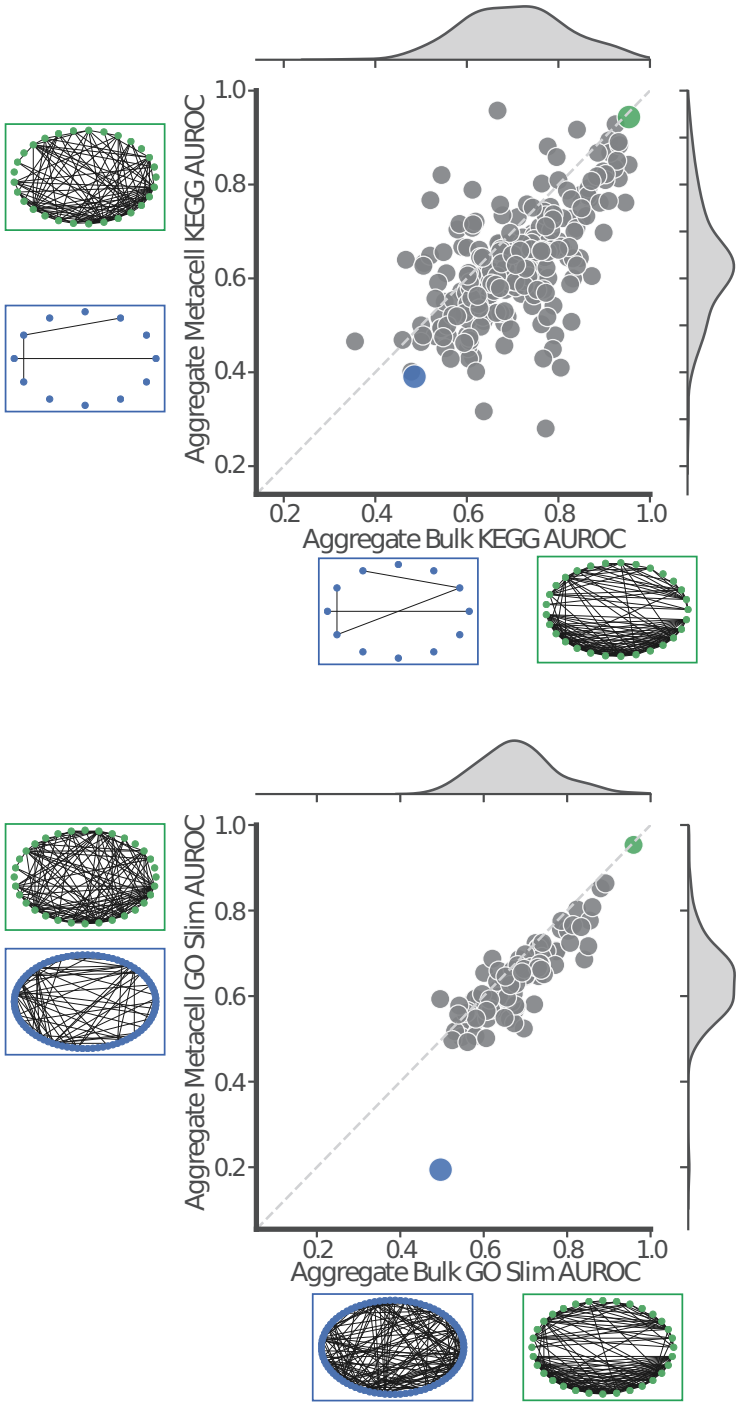


FIGURE 3.8: Performance of KEGG and GO on both metacell and bulk RNAseq co-expression networks is well correlated. Network diagrams show the best performing (green) and worst performing (blue) terms in each dataset.

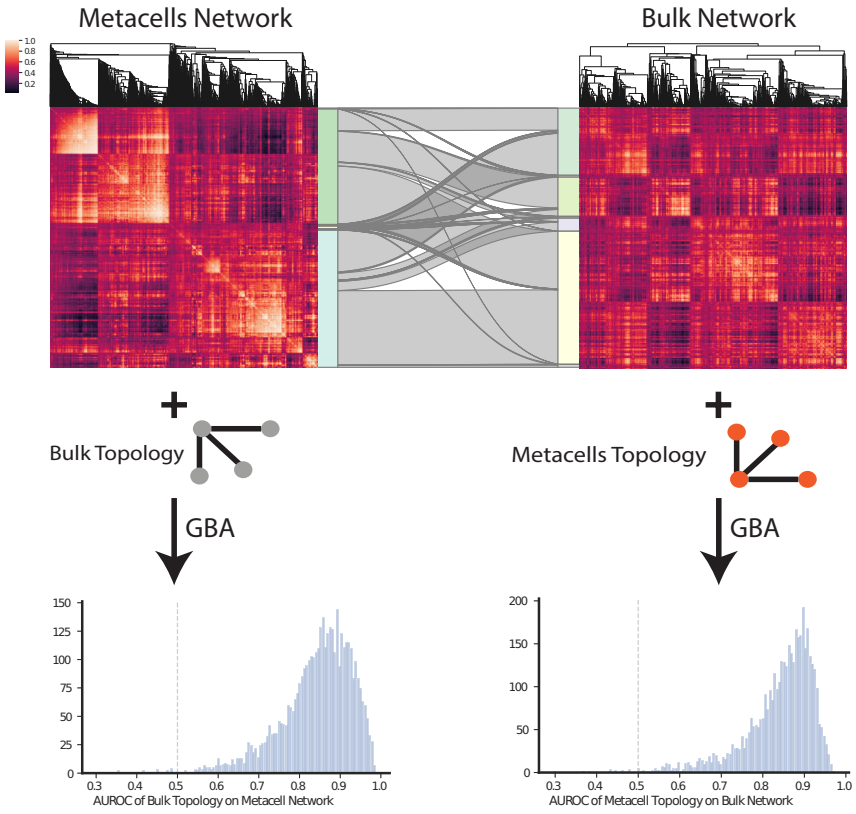


FIGURE 3.9: Clustered heatmaps for metacell and bulk networks. The riverplot joining them identifies shared genes across hierarchical clusters. Prediction of small neighborhoods in one network using the other network's topology shows shared local topology.

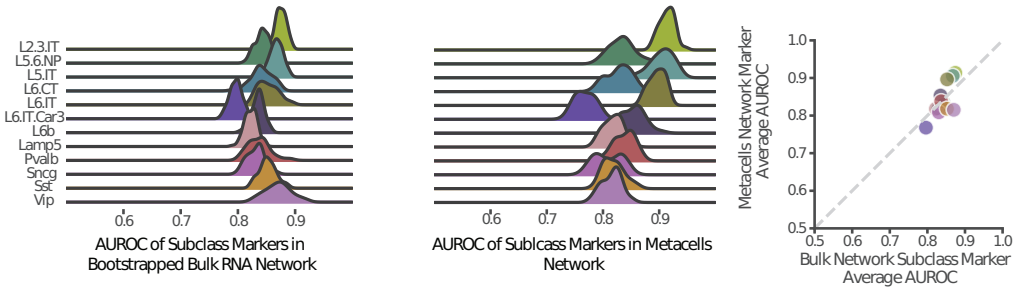


FIGURE 3.10: **Left**Markers show high performance in bulk RNAseq networks. **Middle** Subclass markers show consistent and high co-expression in metacell networks. **Right** Average performance of bulk and metacell networks is highly correlated

3.3.2 Persistent Co-expression of Cell-Type Markers in Compositional and Non-Compositional Networks

To investigate the overlap between compositional and co-regulatory variation more directly, we evaluated the modularity of neuronal subclass markers in each of these two networks, measuring how well network connectivity can reconstruct a partially hidden marker list in cross-validation. As expected, the markers are well connected in meta-analytic networks built from bulk RNA-seq (average AUROC=0.84, Figure 3.10), consistent with the notion that these networks contain cell-type signals. Surprisingly, markers are also well connected in the networks where cell-type variation has been controlled (average AUROC=0.84, Figure 3.10). The performance of the subclass markers in both networks is well correlated (r=0.73, p=0.004, Figure 3.10), in agreement with the consistent topology we find in both networks. As another comparison to the bulk data, we created pseudobulk samples from each scRNAseq dataset by randomly dividing each dataset into 20 pseudobulk samples. Networks created from pseudobulk produce comparable results for the markers and GO to the bulk RNAseq data (GO AUROC = .56, Subclass Marker AUROC = .87, Figure 3.11).This suggests that whatever regulatory factors drive differences between cell-types remain important as a source of differences within cell-types.

The BICCN data offers the unique opportunity to use consistent cell-type labels

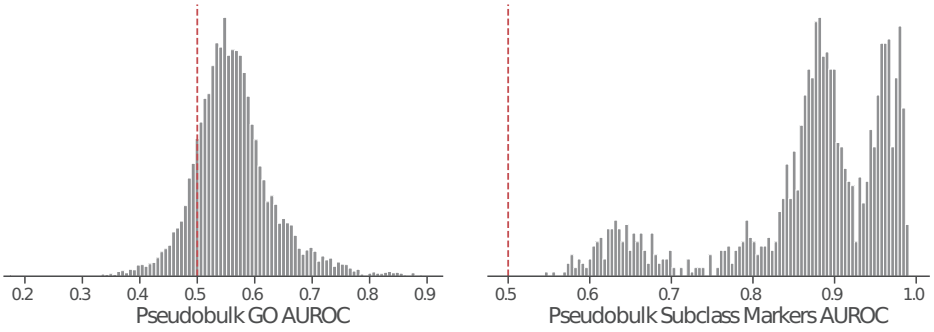


FIGURE 3.11: Pseudobulk co-expression network consistent performance. **Left** Performance of gene ontology using GBA on 100 pseudobulk aggregate networks. **Right** Performance of subclass markers using GBA on 100 pseudobulk aggregate networks. Both are consistent with the performance of bulk RNAseq co-expression networks

across independently sampled datasets so that robust analyses can be constructed at varying levels of specificity in the cell-type hierarchy across independent data. We took advantage of the known hierarchy for our next series of experiments. For each of three levels of increasing specific cell-type classification (class, subclass and cluster) we built aggregate networks to capture replicable gene-gene relationships (Figure 3.12). In each case, samples are divided into homogenous groups at the given level of specificity so that only covariation at more specific levels affects co-expression. So, for example, when we evaluate subclass markers, the class network will be compositional with respect to them, but the subclass and cluster networks will be non-compositional, and should only capture co-regulatory relationships between the same sets of genes. We find that the class network has the highest performance for subclass markers (average AUROC=0.94), but that the subclass and cluster networks still perform exceptionally well (subclass: average AUROC=0.85, cluster: average AUROC=0.83, Figure 3.12). This is also true of subsampled networks that reduce within-cluster heterogeneity, further strengthening this observation (Figure 3.13). Thus, genes which are preferentially co-expressed across cell-types remain co-expressed within cell-types: the same sets of differentially expressed genes which distinguish cells at the subclass level continue to vary across cells even when subclass is held constant.

While our focus has been on genes expressed across most cell types, a natural question is whether the same multiscale co-expression is visible for genes selected based

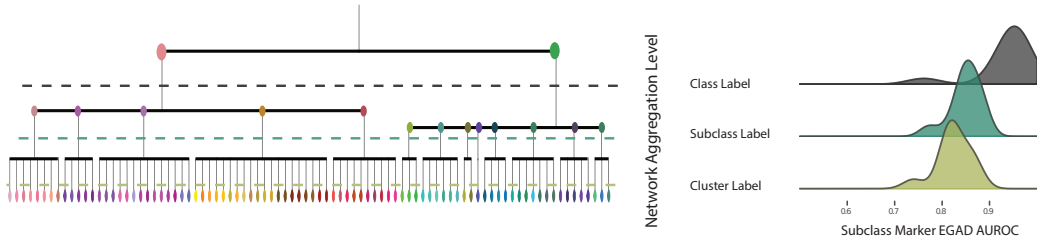


FIGURE 3.12: **Left** Dendrogram of cell-type hierarchy showing class, subclass, and clusters used to construct co-expression networks. **Right** Consistent and strong co-expression of markers in networks at each level of the cell-type hierarchy.

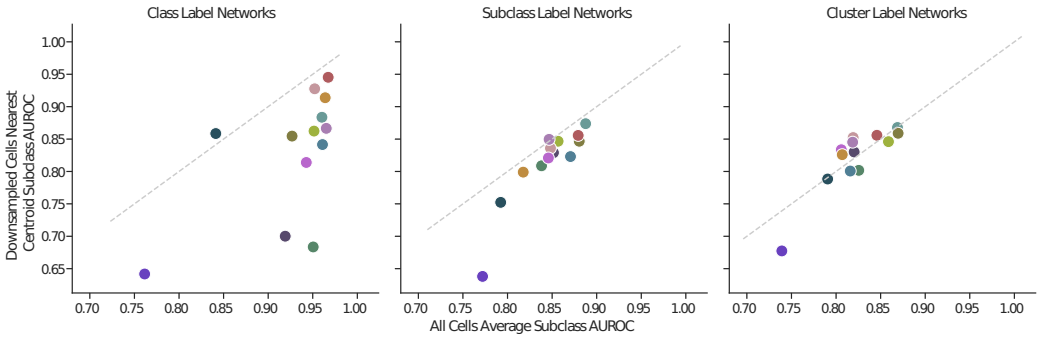


FIGURE 3.13: Pseudobulk co-expression network consistent performance. Comparison of performance of subclass level markers using full datasets and downsampling of each cluster to the 50 cells nearest the centroid of the cluster. Coloring matches subclass coloring in Figure 3.2 dendrogram.

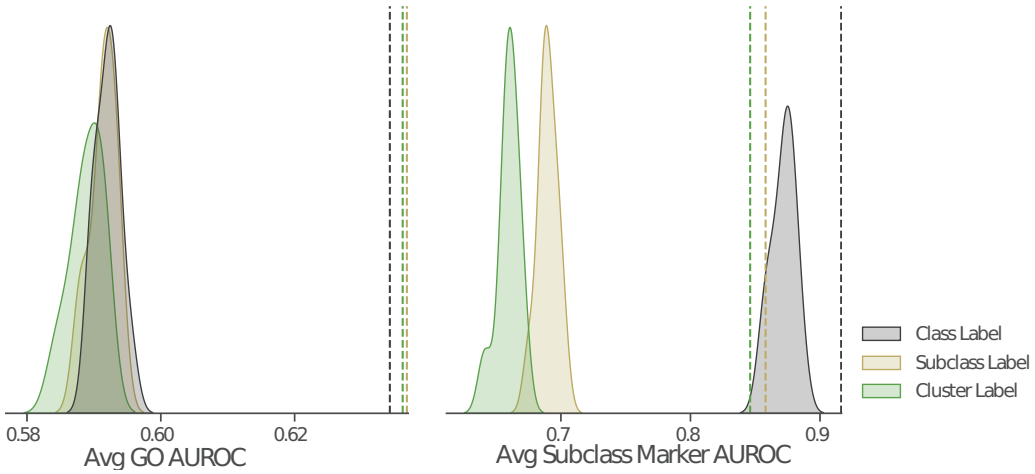


FIGURE 3.14: Consistent performance of resampling genes used in networks. **Left** Performance of gene ontology using GBA on networks computed using genes expressed in pairs of subclasses. **Right** Performance of subclass in networks computed using genes expressed in pairs of subclasses

on expression in only specific cell-types. To test this, we performed our analysis as above, but selected genes based only on their expression in pairs of subclasses. We find the same tendency for gene sets that are co-expressed across a given pair of subclasses to be co-expressed when the subclass is held constant (Figure 3.14). Global co-expression performance is lower, in this case, likely reflecting the slightly less robust gene expression of the selected genes (Figure 3.14).

Proper normalization is an important concern for all scRNAseq analysis. Most commonly, modeling methods can correct for sampling artifacts. However, most of these methods rely on the existing correlation structure, and would induce circularity in the analysis if applied in this study undefined. While we use Pearson's Correlation Coefficient (PCC) for the computational efficiency, proportionality is an association metric that is agnostic to normalization. This makes it an effective alternative to PCC. Additionally, benchmarking of the various measures of association report similar performance across the board (Skinnider, Squair, and Foster, 2019). We built aggregate networks at the class, subclass and cluster levels using proportionality and co-expression across the levels of classification comparable the PCC results (Class AUROC = .92, Subclass AUROC = .85, Cluster AUROC = .84, 3.15, Figure 3.12).

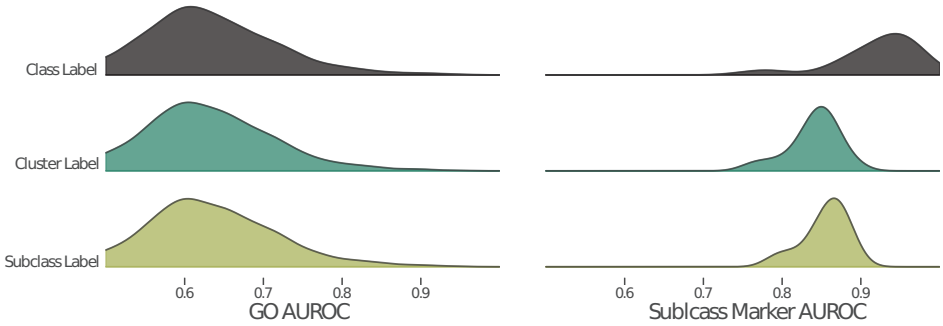


FIGURE 3.15: Consistent performance of networks computed using proportionality. **Left** Performance of gene ontology on networks computed using proportionality. **Right** Performance of subclass markers on networks computed using proportionality.

One question is the degree to which the specificity of the co-expression relationships is maintained. For example, it could be that the exact cell-type marker sets are maintained at more specific levels of the cell-type hierarchy or it could be that new types only sample from within those sets to form new marker sets, creating some novel gene-gene relationships in the process. To investigate this, we first focus on connectivity for two of the GABAergic subclasses: Vip and Sst. In the class network, the subclass markers are extremely modular, with dense connectivity within each gene list and sparser connections between Sst and Vip markers. However, for the subclass and cluster networks, the connections between the modules increase significantly. Despite this increase, the Vip and Sst modules can still be clearly discerned from each other (Figure 3.16). We quantify the change in connectivity between modules by measuring how one gene list, the training list, predicts connectivity to another gene list, the testing list. As expected, in the class network the marker lists are essentially unpredictive of one another (since they mark separate cells, class network AUROC 0.5, Figure 3.16). However, the Vip and Sst modules are more highly interconnected in the subclass and cluster networks (subclass: average AUROC=0.73, cluster: average AUROC=0.73). Testing of all pairwise combinations of subclass modules shows a consistent trend of increased modularity between subclass modules in the more homogenous subclass and cluster networks (subclass AUROC=0.65, and cluster AUROC=0.66) relative to the class network (AUROC 0.5). The modest cross-module performance of marker sets suggests that there is some "cross-talk" between modules as we

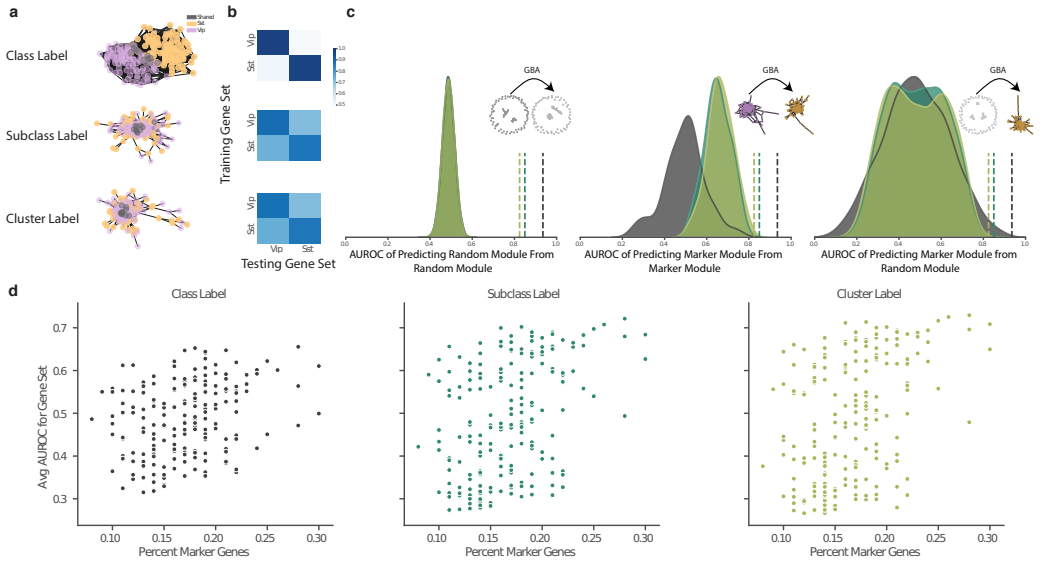


FIGURE 3.16:
a. Network diagrams show strong connectivity within both Vip and Sst modules at all levels and increased connectivity between the modules at the subclass and cluster level networks. **b.** Predicting connectivity between modules show random connections at the class level and strong connections at the subclass and cluster level for Vip and Sst modules. **c.** Random gene sets have random connectivity to other random gene sets in all networks, while subclasses have non-random connectivity between each other in only subclass and cluster level networks. Random gene sets have variable levels of connectivity to subclass modules **d.** Random gene set connectivity to subclass modules is correlated to the percent of random genes that are a marker in any subclass.

move down the hierarchy of cell-types as modules are combined in novel ways to define new cell-types.

Performance of each subclass marker set is consistently high within any subclass specific networks (Figure 3.17). Marker sets, like the Vip interneuron markers, have extremely low variation in performance across the subclass specific networks. Diagrams of the networks show consistently dense networks (Figure 3.13). To further investigate connectivity of subclass markers in the subclass specific networks we focus on the consistency and strength of connectivity to individual genes by predicting each gene's connectivity to the rest of the genes in the subclass marker set. The strength of a gene's connection to its marker set does not depend on the data from which it was constructed, remaining high regardless of the subclass network being measured (Figure 3.17). This once again highlights the consistency of a core co-regulatory network across the cell-types. Individual

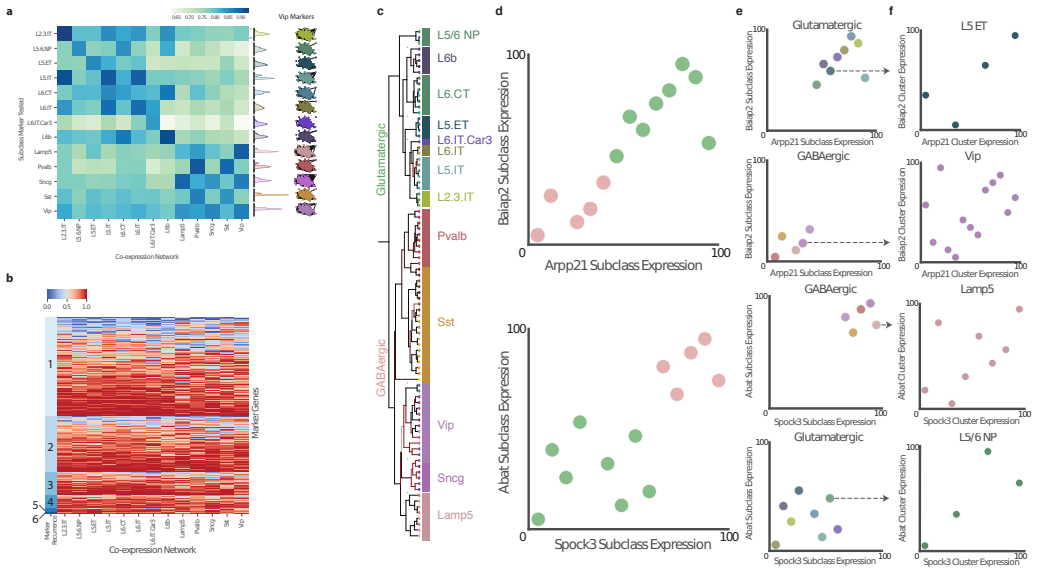


FIGURE 3.17: Multiscale performance of subclass markers and individual genes in subclass specific networks. **a.** Performance of subclass marker gene lists on aggregate subclass specific networks, with marginal distributions for each marker lists and diagram of Vip marker module in each network. **b.** Connectivity of individual genes within subclass marker lists across subclass specific networks. The recurrence of genes across networks is annotated on the left margin. For recurrent genes, the average performance across modules is shown. **c.** A dendrogram of cell-type hierarchy and colors. **d.** Expression percentiles aggregated across datasets for a pair of Glutamatergic markers, Arpp21 and Baiap2, and GABAergic marker, Spock3 and Abat. **e.** The GABAergic and Glutamatergic markers remain co-expressed when split into GABAergic and Glutamatergic subclasses. **f.** Within subclasses clusters are co-expressed in both GABAergic and Glutamatergic subclasses.

pairs of genes also exhibit multi-scale co-expression. We illustrate the co-expression of Arpp21 and Baiap2, two Glutamatergic markers, and Spock3 and Abat, two GABAergic markers (Figure 3.14). These gene pairs exhibit multiscale co-expression because they are co-expressed at the class, subclass and cluster level, even in cell-types that the genes are not markers of. The scale of the BICCN expression data and cell-type annotations cannot be matched by any other organ system. However, using 4 human pancreas datasets that are normally analyzed together, we also found consistent co-expression of cell-type markers at multiple scales (Full datasets (compositional) AUROC=.97, Cluster (non-compositional) AUROC=.97, Figure 3.18). All of these analyses show consistent co-expression of gene sets that define cell-types from broadest to finest levels of cell-type classification.

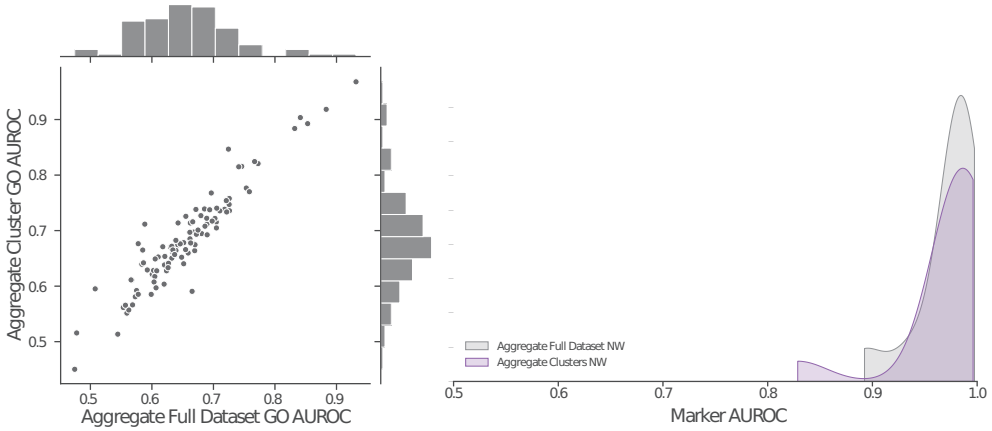


FIGURE 3.18: Consistent co-expression at multiple scales in human pancreas co-expression. **Left** Correlated co-expression of the gene ontology on the full dataset (compositional) and cluster (non-compositional) aggregate co-expression networks **Right** Consistent co-expression of cluster markers in aggregate co-expression networks at both levels of heterogeneity.

3.3.3 Differential Co-expression to Identify Novel Gene-Gene Relationships

Our results provide evidence that the multiscale model of co-expression (differential expression aligns with conserved co-regulatory relationships) plays an important role in regulatory networks. We next evaluate if we can find evidence for the differential co-expression model (change in co-regulatory relationships) by looking for cell-type specific gene-gene relationships. We take the difference between a single subclass's network and a network of the remaining subclasses to find the edges most specific to a given subclass. In a differential co-expression network between subclass A and the rest of the subclasses, the strongest connections in the network are gene pairs that are only co-expressed in subclass A. This means that if marker genes for subclass A are only co-expressed in subclass A, they will have a high AUROC. However, differential co-expression networks show minimal connectivity of subclass markers (average AUROC =0.69, Figure 3.19). These low values are particularly notable in contrast to earlier performance of the multiscale co-expression model where, using aggregates that assume a purely consistent regulatory architecture, we found strong enrichment of subclass markers (subclass network AUROC=0.84, cluster network AUROC = 0.84, Figure 3.19). GO and KEGG modules are also relatively weakly

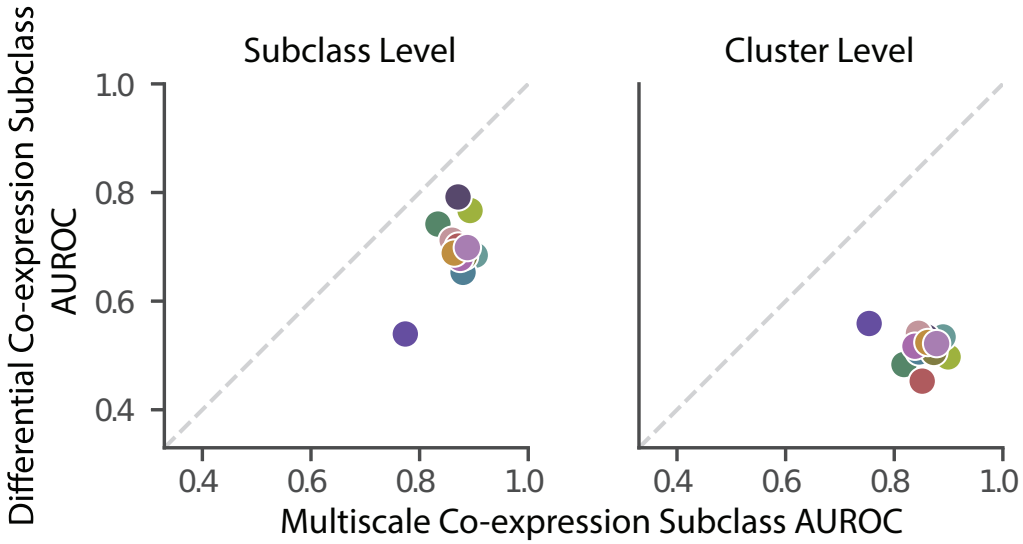


FIGURE 3.19: The multiscale model outperforms differential co-expression for recapitulating cell-type marker modules. Networks were aggregated at the subclass (left) and cluster (right) level. Differential co-expression was calculated by taking the normalized difference between a given network and all others at the same level in the cell-type hierarchy.

connected in the differential co-expression networks (Figure 3.20). These results emphasize the consistent modularity across the cell-types in co-regulatory modules.

While the multiscale model explains most of the co-expression signal within cell-types, the performance of the subclass markers in the differential co-expression networks, while lower, is non-random. This suggests the potential to identify individual edges as significantly differentially co-expressed and identify novel cell-type specific co-expression modules. We first consider how the heterogeneity of data affects our ability to confidently call connections as significantly different. When computing differential co-expression between the GABAergic and Glutamatergic cell-types we can aggregate the networks at either the class, subclass or cluster level. We find that the most heterogeneous class networks identify 10x more edges at a given false discovery rate (FDR) threshold than the subclass networks (Figure 3.21). The subclass networks also identify 10x more edges than the cluster networks. Selecting significant edges from the class network will result in 0.1% of edges being significant at an FDR<0.01, while using the cluster network has no significant

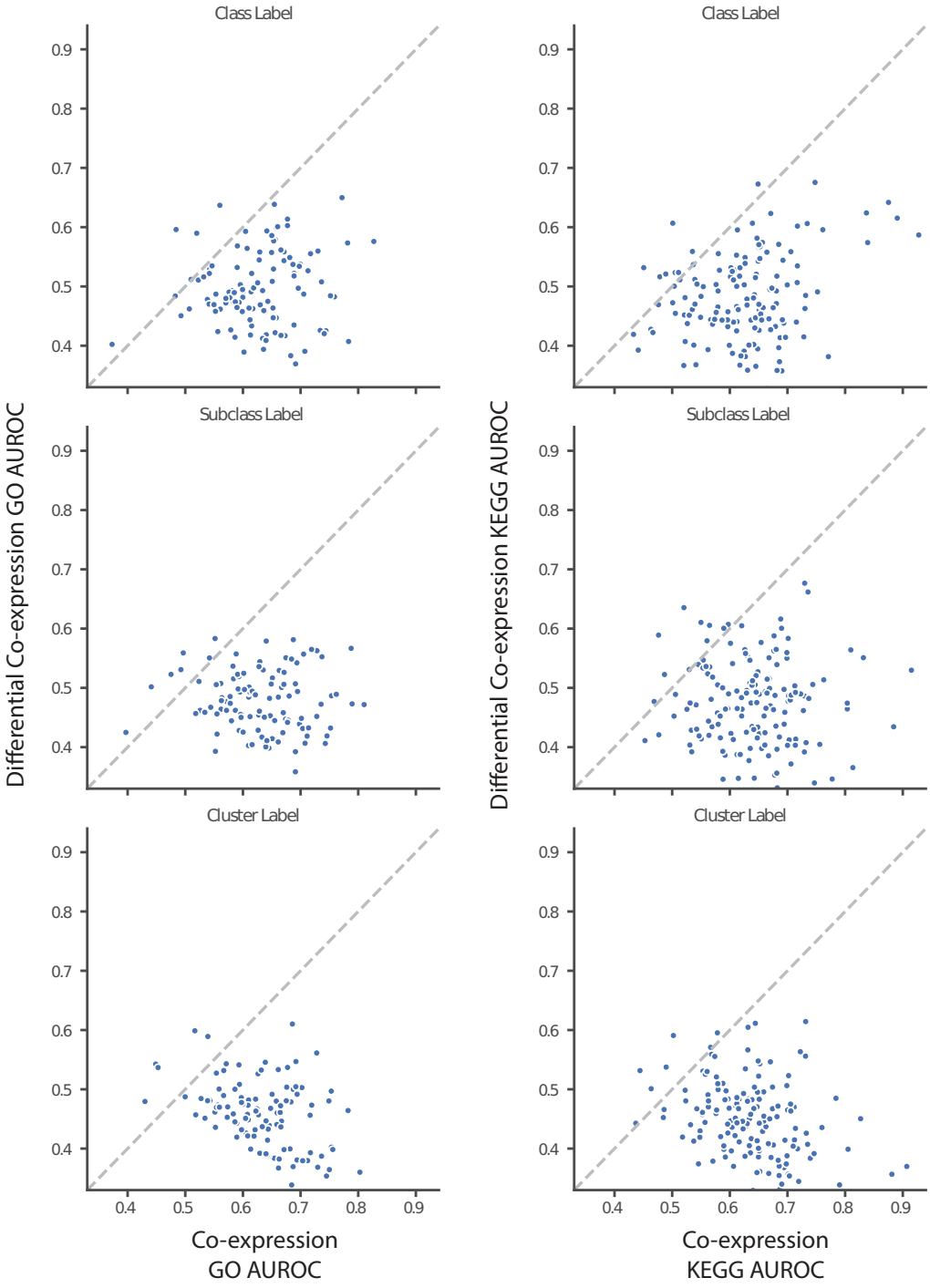


FIGURE 3.20: Differential co-expression fails to identify cell-type specific functional modules. Differential co-expression between GABAergic and Glutamatergic cells has random performance of both GO mouse slim and KEGG, while testing for multiscale co-expression identifies modules that are consistently co-expressed in both GABAergic and Glutamatergic cells.

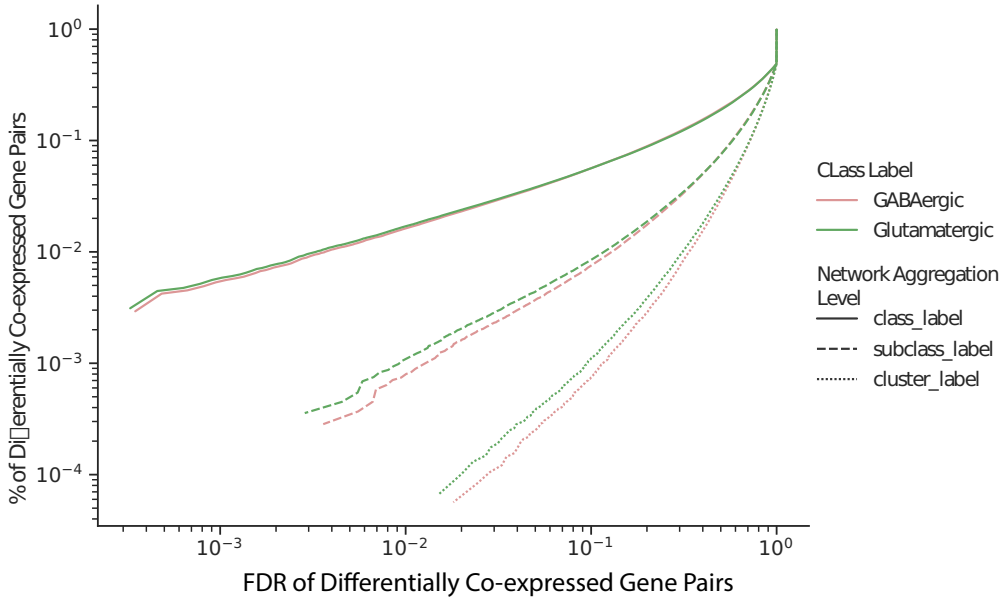


FIGURE 3.21: Differential co-expression between GABAergic and Glutamatergic cells aggregated at different levels shows limited statistical power.

edges even at a more permissive FDR<0.1. These results suggest that, even when aggregating across 7 datasets, we are underpowered to detect changes in co-regulation at the cell type level.

Incorporating more scRNA-seq datasets should provide sufficient power to confidently identify cell-type specific co-expression relationships. We show the power gained by aggregating from 2 to 7 of the existing datasets, providing an improvement in power and statistical significance on par with the improvement from the coarsest to finest cell-type definitions (Figure 3.22). Using a threshold of 1% of edges being differentially co-expressed, the class level differential co-expression network is sufficiently powered at an FDR <.01 using only 6 of the 7 datasets. Extrapolating from subclass level results, estimates 11 datasets are required to achieve the same thresholds. Even with all 7 datasets, no edges are significantly different in the cluster aggregate network so we cannot extrapolate the number of datasets required directly. (Figure 2.23). Within the edges in the class label network that are FDR <.01, we found 82% of them contained at least 1 gene that is a subclass marker (Figure 2.23). With 709 marker genes, only 31% of edges are expected to contain a marker gene. Using differential co-expression to identify novel gene-gene relationships

will require controlling for composition using networks aggregated at the finest scales. . While we are underpowered for differential co-expression, the existence of multiscale co-expression presents a powerful toehold for future analyses, potentially limiting the search space for variability to a much smaller core set of modules. In this view, co-expression is largely maintained within cell-types with a major source of variability simply being the dynamic range over each of the genes is operating (Figure 3.24).

3.4 Discussion

Our meta-analysis of bulk RNAseq data and the BICCN scRNAseq data from the mouse brain establishes the importance of a multiscale model of co-expression across neurons. We identified shared topology between compositional and cell-type specific networks using both reference functional networks, the Gene Ontology and KEGG, as well as direct comparisons of network topology. Cell-type level markers for neurons exhibit consistent topologies in networks built at all levels of the cell-type hierarchy.

Our result highlights the existence of a core co-regulatory network that is reused in all cell types of the brain. We note that this result is not likely to be brain-specific, or even cell-type specific, as previous research has also shown strong convergence in co-expression across systems. Indeed, while expression levels of genes vary across brain regions, many modules associated with cell-types replicate across brain regions and species (Hartl et al., 2020). Outside the brain, drug perturbation experiments using human iPSC-derived cardiac myocytes and fibroblasts have shown that cell identity maintenance factors are usually not tissue specific. Rather genes that play important functional roles in cell identity maintenance are broadly expressed across tissues (Mellis et al., 2020). The critical role of non-tissue specific genes to perturbation highlights the important role of core regulatory networks in contexts defined by cell-type specificity.

Given the major role of multiscale co-expression, we expect that finding differences between cell-type specific co-expression networks will be difficult. We explored the statistical power necessary to identify cell-type specific co-expression and how heterogeneity within the data influences the power. Despite the large amount of data considered,

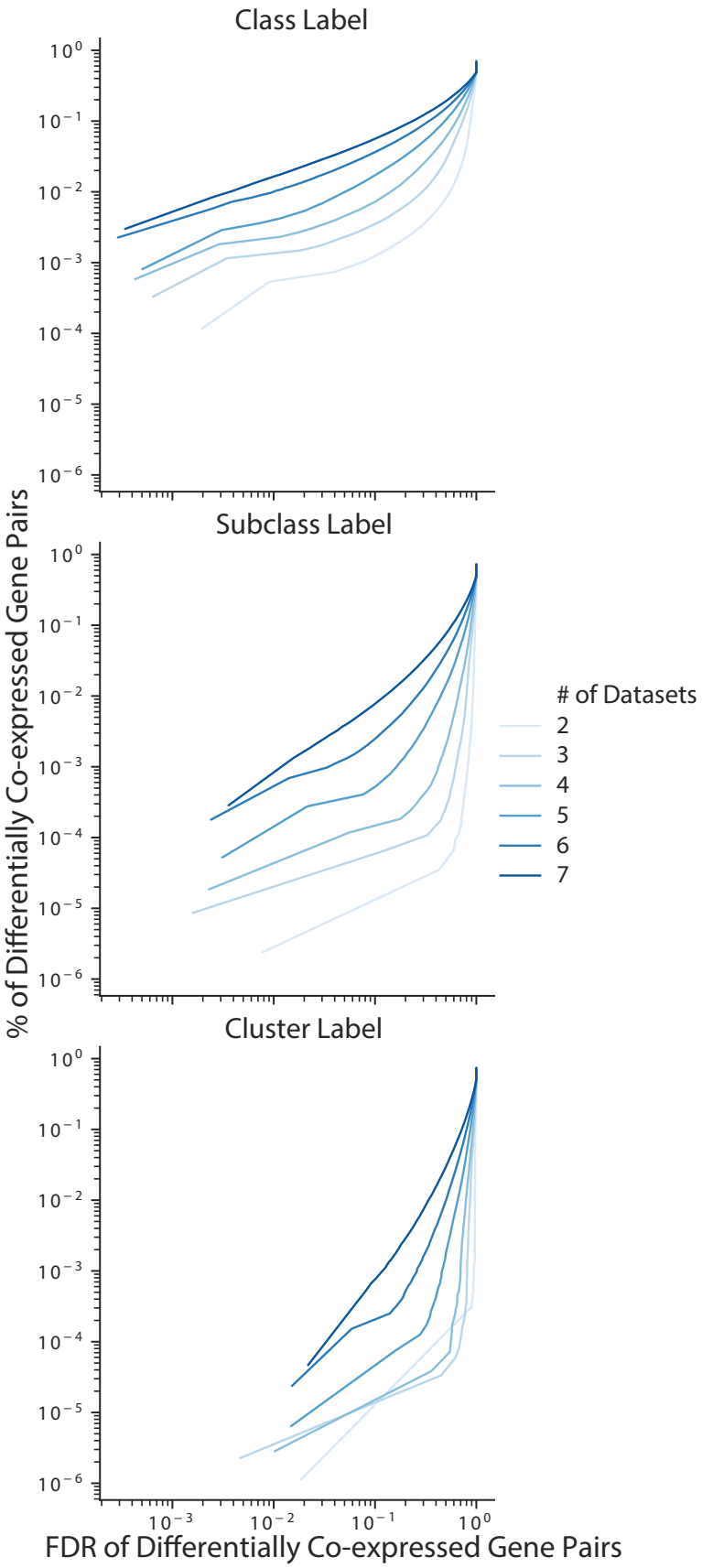


FIGURE 3.22: Statistical power increases as the number of datasets used in the meta-analysis increases

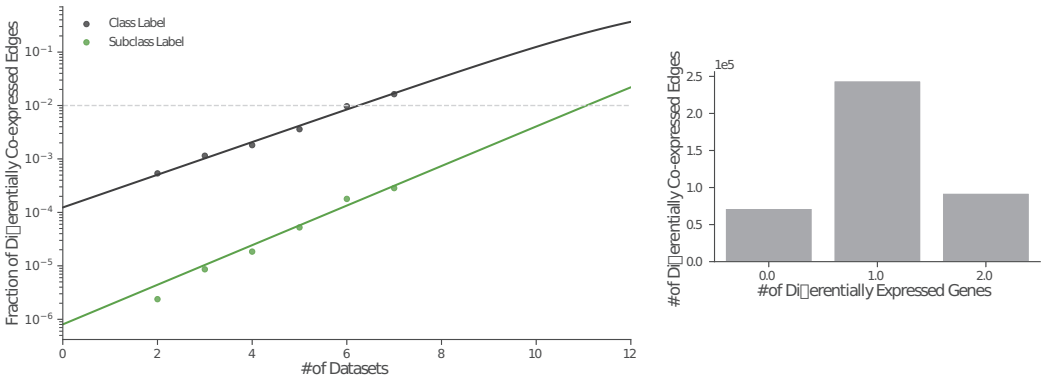


FIGURE 3.23: Meta-analysis as a path of powered and novel differential co-expression. **Left** Estimating necessary number of datasets needed for differential co-expression at an FDR <.01 at all levels. Cluster level isn't shown because even with all datasets no edges are differentially co-expressed at an FDR <.01. **Right** Most edges that are differentially co-expressed at an FDR <.01 in the class label network include at least 1 gene that is differentially expressed at the subclass level. Indicating that the differential co-expression is not identifying novel co-expression connections.

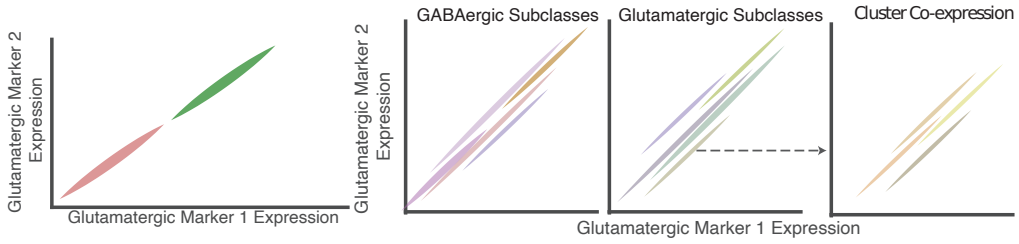


FIGURE 3.24: A multiscale model for co-expression of a pair of glutamatergic markers shows co-expression of the markers through all levels of the hierarchy of cell-types.

we are underpowered to identify cell-type specific co-expression, though networks built at the lowest resolution of cell-type classification are nearly sufficiently powered. As more scRNA-seq data becomes available, we expect the value of meta-analysis to become increasingly apparent within this data, not just as a mechanism for overcoming experiment specific biases, but in generating gold standard co-expression networks that can be used as a groundwork for exploration in data where some differences are expected (e.g. disease).

A central limitation of our study is our focus on genes that are broadly expressed across cell-types. This is a simple necessity for our analysis since co-expression is undefined if, e.g., one gene shows no variation (is unexpressed) in a given cell-type. On the other hand, it may well be that this constitutes a large fraction of cell-type variability that we do not explore. While interesting, such variation does not really reflect changes in co-expression since it can be much more easily explained through the single-gene expression. The multiscale co-expression we see may be most relevant to the growing literature on the importance of gradients in defining cell-types, particularly in the brain (Cembrowski and Menon, 2018). The relatively high cross-type marker learning performance similarly suggests a relatively simple continuous axis of co-variation between genes, at least within the well-powered BICCN data. When measuring co-variation within finer scales, such as in the cluster and metacell networks, the proportion of non-biological variance might be higher due to the smaller size and greater homogeneity of each grouping of cells compared to higher levels. We control for this by using replicable relative correlations which will tend to be insensitive to global shifts in the correlation, although more complex interactions could still affect results.

Beyond continued evaluation within the BICCN, our results open up two main directions to take future analyses: improved gene function annotation and improved cell-type specific co-expression. Computational gene function annotation is typically done using orthology or features generated from DNA sequence (Škunca, Altenhoff, and Dessimoz, 2012). Our evidence for the multi-scale model shows that well powered co-expression networks built across species should be a valuable addition to methods for computational

annotation of gene function. Improved cell-type co-expression should also be a major addition to mechanistic studies. Inferring mechanistic relationships from scRNAseq alone has proven difficult (Qiu et al., 2020), with methods that incorporate ATACseq or ChIPseq data doing only a little better Burdziak.2019. Using well powered cell-type specific co-expression networks should open up both stronger integration with other modalities (e.g., ATAC-seq) and better inference of convergent changes across conditions (Hie et al., 2020). Thus, a major source of utility of the BICCN data is simply the presence of reference data that crosses technologies, labs, and other nuisance variables to permit robust aggregation; a process which is particularly important and convenient within co-expression space. The meta-analytic use of the BICCN data sets a standard we hope can continue into the future, integrating data from outside the BICCN to obtain increasingly high-quality and useful reference co-expression networks.

3.4.1 Conclusions

The shared co-expression signal of marker genes and regulatory modules throughout the cell-type hierarchy makes it clear that co-expression is, in part, multiscale. Multiscale co-expression means that while gene expression values are significantly different between groups of cells, the core co-regulatory network remains consistent throughout the highly refined cell-type hierarchy defined within the primary cortex. The sparsity and noise in scRNAseq data often make co-expression and differential co-expression challenging. Using a meta-analytic framework, we highlight robust methods and significant use cases for co-expression and differential co-expression analysis using scRNAseq data.

3.5 Methods

3.5.1 Single Cell Datasets and preprocessing

We acquired the datasets and associated metadata directly from the BICCN. To work with the expression data, anndata objects were created by filtering the droplets to the whitelist defined by the consortium and merging with all associated metadata. All analyses were done using CPM normalized expression values. To select a shared list of

genes we ranked each gene by its average expression and selected the top 7,500 genes in each dataset. Then genes that were in the top 7,500 for at least 6 of the 7 datasets were used in all analysis, leaving us with 4,201 genes. All analyses were done with this list of genes.

3.5.2 Bulk RNA Sequencing Data from GEMMA

Metadata from the Gemma database was acquired on 11-29-19. The metadata was filtered to include only mouse bulk RNAseq datasets with at least 20 samples. Then metadata terms were filtered for relevance for the brain, leaving 29 terms (See github for terms and data info). The expression data was then downloaded using the GEMMA R API and filtered to the same genes as the scRNAseq data. Networks were built as detailed below.

3.5.3 Network Construction and Aggregation

Networks were built by rank standardizing the Pearson correlation matrix of the genes. After ranking, we replace the undefined values with the average of the network. For the bulk data networks are built using an entire dataset. For single cell datasets, a full compositional network is computed using only the labeled neurons in each dataset. When computing class, subclass, cluster, and metacell networks we partition each dataset by the metadata label and build a network for each value. After aggregating networks within each dataset, we aggregate the ranked dataset networks. Aggregating datasets occurs by summing the networks from each dataset and then ranking the sum.

When computing network performance of markers in the bulk network we bootstrap the bulk datasets 100 times to create 100 networks. In the down-sampling experiment we compute centroids for each dataset partition and select the 50 closest cells to the centroid within that partition. We exclude any partition with fewer than 100 cells to make sure it is at most 50% of the original data in the partition.

3.5.4 Computing Marker Genes

Marker gene lists are computed using the Mann-Whitney test in each dataset using a 1vsAll design. Significance is computed with a threshold of log2FC >2 and FDR <0.05. To compute markers across datasets we compute recurrence of each gene by totaling the number of datasets the gene is significantly different in. After sorting genes by recurrence, we sort by average AUROC. We used gene sets of size 100. Subclass-specific markers are computed within classes, e.g. Vip markers are extracted by finding genes that are differentially expressed with respect to all other GABAergic subclasses. For example when computing the markers for the Vip subclass, a GABAergic subclass, we only compare the expression of the Vip cells to the other GABAergic subclasses.

3.5.5 Measuring Network Performance with EGAD

A python version of the R package EGAD was created by translating the runGBA() function from the R package (Ballouz et al., 2016). It was modified to do cross validation in known splits, instead of randomly partitioning the data. We run it with 3 fold cross-validation. The algorithm uses neighbor voting to compute the sum of ranks of predictions for a given gene set within a network. Using the sum of predicted ranks we calculate an AUROC and/or a p-value as an output. When measuring performance of the subclass marker genes on the scRNAseq networks we create aggregate networks for each combination of 4 datasets and measure the performance using meta-analytic markers using the remaining 3 datasets. In the figures we report either the entire distribution or just the average of these values and in the text we report the average value.

3.5.6 Computing Metacells

Metacells are computed using the R metacells package. We set the parameters to encourage extremely small clusters(K=20, m=5,b=1000). Additionally, we used the 4,201 recurrently highly expressed genes as the gene list for the method. While the metacells method, like the original clustering method, is graph based, minor differences in the

methods allow for metacell clusters that contain multiple subclasses. To avoid any compositional effects, we filter out all metacell clusters containing cells from multiple subclasses.

3.5.7 Computing Differential Co-expression

Differential co-expression was computed by subtracting networks within datasets, then ranking the difference. Afterwards the differential networks were averaged across the datasets. To compute an FDR we used a null distribution of the average of 7 networks generated by sampling random uniformly distributed numbers.

Chapter 4

Conclusions and Perspective

In this thesis, I exploit the power of scRNAseq and meta-analysis to expand how we functionally characterize cell types.

In Chapter 2, I build an atlas of hematopoietic cell types from mouse bone marrow to explore multiple axes of variation that can define the cell types. For each class of variation, I evaluate how well they can be defined in gene and functional space. Importantly, I find replicable signatures that define cell types across each model of variation I evaluate. I identify both salient and subtle signatures that are replicable across the datasets. The subtle axes, the cell states defined from in silico sorting, are especially relevant for integrating knowledge from pre-scRNAseq studies in hematopoietic stem and progenitor cells. Through co-expression analysis, I show how to evaluate the conservation of functional signatures in one species across many other species before gathering additional scRNAseq data. Altogether, this work shows how a cell atlas built from many scRNAseq datasets can be used to define replicable and generalizable gene expression signatures. This atlas, which is easily accessible in a shiny web server, will be a critical reference for future scRNAseq experiments and will serve as an anchor to integrate with other data modalities.

In Chapter 3, I use the BICCN MOp neuron cell atlas to study the similarities and differences of cell-type-specific co-expression networks. I compare networks built from statistically homogenous populations of cells to measure correlated stochastic variation to networks built from compositionally heterogeneous bulk RNAseq data. Using both functional annotations and marker gene lists I show how co-expression of highly expressed

genes is markedly consistent at the lowest and highest levels of cell-type heterogeneity when building co-expression networks. Additionally, I demonstrate the statistical limitations of computing differential co-expression. These results show how cells are largely defined by the coordinated upregulation of specific gene programs, or network modules, as opposed to rewiring of the gene network.

4.1 Utility of analyses using both scRNAseq and bulk RNAseq

In both analyses, I rely on a combination of scRNAseq and bulk RNAseq and find that signatures, that largely are thought to only exist in scRNAseq, specifically cell-type-specific co-expression relationships, are extremely identifiable in bulk RNAseq. scRNAseq remains exceptionally popular, but in many cases, bulk RNAseq is more feasible and cost-effective. Especially in clinical studies, preserving samples for scRNAseq library preparation can be challenging. To get around this the Satija lab built a portable DropSeq device (Stephenson et al., 2018). While ingenious, they are not practical everywhere. This is not to say that scRNAseq is useless, rather we can strike a balance between scRNAseq to identify cell-type-specific signatures in a cost-effective way and then validate them at a population level using bulk RNA sequencing. For population-level studies, like GTEx, scRNAseq on every tissue sample would be impractical. Instead, they have done scRNAseq on a select few samples (Eraslan et al., 2021). Similarly, they did a gamut of epigenetic and footprinting assays on a subset of subjects as part of the Entex project (Rozowsky et al., 2021). Utilizing signatures learned across a subset of samples can be informative when comparing to many bulk samples that cannot be sequenced at the same depth. Bulk RNAseq is comparatively much easier and cheaper than scRANseq and epigenetic-based assays, making it the most suitable for population-level studies. Further methods, integrating knowledge learned from a few scRNAseq studies with population-level bulk RNAseq will be critical for conducting functional genomics at a cell type specific level.

4.2 Functional genomics at a cell type resolution

The advent of high throughput sequencing, both DNA and RNA has led to the field of functional genomics. After the completion of the first human genome draft at the turn of the century, the next herculean task was to understand the functional elements of the sequence. Two major consortiums stand out in the field of functional genomics. Using assays that measure histone marks, transcription factor binding, CpG islands, chromatin accessibility, and 3-D genome structure, the ENCODE project is the main consortium generating data for learning the functional role of DNA elements (Consortium et al., 2020). The GTEx consortium characterizes the role of genetic variation on gene expression in humans (Consortium, 2020). Nearly all of the data generated from these incredibly fruitful consortia is on either homogeneous cell lines or heterogeneous bulk primary tissue data. scRNAseq is poised to expand our understanding of gene regulation at a cell type level. The results in Chapters 2 and 3 highlight the consistency of functional programs within and across cell types but do not explain how the modulation of different functional programs occurs. Transcription factors (TF) binding to accessible cis-regulatory elements play a major role in gene regulation. Expanding on the multiscale co-expression results using scATACseq, which measures open chromatin, can help identify the TFs involved in modulating the functional programs that define cell types. scATACseq data will also identify the transcription factors that control hematopoietic development. All of the further work is best supported by high-quality reference atlases.

4.3 Cell types as defined by gene modules

In many single cell analyses, including the ones in this thesis, cell types can be identified by a single or few genes. Statistically, many cell types can be identified by one or a few genes (Missarova et al., 2021). For certain follow-up experiments using technology like in situ hybridization, you can only practically use a few genes to label cell types (He and Huang, 2018; Chen et al., 2019). The aggregate performance of a gene module can better identify a cell type than an individual gene (Fischer and Gillis, 2021). This thesis

highlights the role that functional modules play in defining cell types. In many regards, they more explicitly describe the function of cell types than individual genes. The important distinction I am making here is between identifiability, the ability to label a cell type, either within scRNAseq data or in another data modality, and definability, what functions, which can be represented as co-expression modules, define the role of the cell type. The dimensionality of genes is far greater than functional modules, which in turn are far greater dimensions than the cell types. When we use individual genes to define cell types we are skipping over the fact that cell types are largely defined by the modulation of various functional programs. This is not to say that we should not use individual genes to identify cell types, it would be impossible to do so, but when trying to understand the nature of cell types we must consider functional programs. The main limitation of taking this kind of approach is our understanding of functional programs. The most popular tools, the Gene Ontology and KEGG are far from perfect (Škunca, Altenhoff, and Dessimoz, 2012; Plessis, Škunca, and Dessimoz, 2011; Gaudet and Dessimoz, 2016). In some organisms, like maize, few genes contain experimentally determined functional annotations. This has been a major challenge for annotating cell types in our scRNAseq data from maize meristems (Appendix B). Because co-expression captures shared functional relationships between genes, without input from known functional annotations, it serves as an analytical tool for evaluating functional relationships between genes in scRNAseq data.

4.4 Replicability of Pseudotime across Datasets

As discussed in Chapter 2, the analysis of continuous trajectories across datasets remains a major open question. he documentation for most pseudotime methods recommends using batch integration methods to project the datasets into a unified latent space and then learning a pseudotime trajectory on that space (Cao et al., 2019). As I show, the pseudotime learned on an integrated space is quite different from that of the one learned on the individual datasets. A major aspect of this is that the cell type adjacency, neighbors of the cell types, are not preserved during the batch correction (results not shown). Methods like PAGA, where pseudotime is measured by clusters, will be most affected by a change in

neighbors for cell types (Wolf et al., 2019). Methods that work on individual cells will also be affected because these methods, while usually agnostic to cell type labels, largely learn trajectories that pass from one cell type to neighboring ones (Cao et al., 2019; Haghverdi et al., 2016).

In this thesis, I present two analyses that evaluate the replicability of pseudotime. First, I run MetaNeighbor on each segment of the learned trees. This showed that the root and leaf branches are replicable across datasets, while the intermediate segments were not. The other way I evaluated the replicability is by identifying genes associated with development into the monocyte or erythrocyte lineages in each dataset and using meta-analytic statistics to find the top replicable markers. Both of these methods, while imperfect, offer easily interpretable results and show that the macrostructure learned by the methods, mainly the branching, is largely consistent across the datasets used. I also attempted to adapt MetaNeighbor to evaluate the replicability of binned pseudotime to measure the replicability at a more granular level. Unfortunately, the results, while potentially promising, were not easily interpretable.

Further work evaluating replicability of pseudotime will be critical, as developmental signatures measured in scRNAseq show a lot of promise. Outside of the field of hematopoiesis, there are other systems where evaluating developmental trajectories learned across datasets would be immediately beneficial. The most obvious is looking at embryonic development single cell datasets. These datasets sample embryos at multiple developmental time points and use methods slightly different to the ones I used to evaluate development over time, instead of at a singular timepoint (Cao et al., 2019; Feregrino et al., 2019; Bella et al., 2021). Most exciting for these analyses is that scRNAseq datasets are being generated from embryos from many different species, and evaluating developmental trajectories across the species could answer important questions related to evolution and provide insights into how best to use certain animal models of human diseases. Methods for evaluating co-expression across development time and datasets currently do not evaluate replicability, but rather attempt to learn signatures across batches with significant technical variation in an integrated space (Hie et al., 2020). Combining methods from both

Chapters 2 and 3 could lead us to a way that effectively evaluates co-expression relationships through pseudotime across datasets. Additionally, methods similar to MetaNeighbor, but that use regression instead of binary classification might effectively evaluate the replicability of continuous pseudotime values.

4.5 Meta-analysis of Spatial Transcriptomics Data

Spatial transcriptomics is a popular new sequencing modality that links gene expression with a spatial location in the tissue. While less mature of technology than scRNAseq right now, it has won Method of the year from Nature methods (Marx, 2021). As the technology matures more data will be available making meta-analysis possible and in some regards necessary. Like any other technology, evaluating spatial transcriptomics using meta-analysis will be critical for learning robust signatures that are not driven by noise within and across datasets. Many of the methods used in this work can easily be applied to spatial transcriptomics data, both to analyze multiple spatial datasets on their own, or to evaluate the replicability of scRNAseq to spatial transcriptomics datasets from similar biological conditions. However, these methods are insufficient for incorporating spatial information into meta-analysis. One important aspect of this is to see if cell types have similar neighbors across datasets/samples. Because spatial measurements in their raw form are not transferable across samples, you would have to evaluate it based on relative relationships, like neighbors. Similar to scRNAseq analysis methods, existing methods that are designed to work with multiple spatial transcriptomics datasets are focused on integrating samples across datasets as opposed to evaluating the replicability of labels or signatures learned in each dataset (Zeira, Land, and Raphael, 2021).

4.6 Final thoughts

The question about cell identification and function is not a new one. The advance of scRNAseq has allowed us to explore it in an entirely new way. One thing that really stands out about scRNAseq when compared to previous technologies is the fact that the features in the data are all genes. FACS and microscopy have been limited to a few

genes and making it challenging to utilize measurements across experiments or other data modalities. While evaluating data across experiments across multiple scRANseq datasets remains challenging, my meta-analytic framework effectively identifies replicable signals across datasets. As scRNAseq continues to mature, the ability to use the shared features to integrate more scRNAseq data and other data modalities will be one of the most powerful tools. Measuring biology at the smallest unit of life, and integrating it with functional genomics data will provide us with an understanding of gene regulation that reveals so much of the mystery behind genomes.

Appendix A

Scaling up reproducible research for single cell transcriptomics using MetaNeighbor

In this appendix, I am including the manuscript from the Nature Protocol I was a co-author on. My main contribution was writing the python package (`https://github.com/gillislab/pyMN`*), and helping with desiging the procedures. The procedures are available at* `https://github.com/gillislab/MetaNeighbor-Protocol` *The code and methods from this protocol were central to my analysis in 2*

nature protocols

PROTOCOL

https://doi.org/10.1038/s41596-021-00575-5

Check for updates

Scaling up reproducible research for single-cell transcriptomics using MetaNeighbor

Stephan Fischer[1,3], Megan Crow[1,3], Benjamin D. Harris[1,2] and Jesse Gillis[1,2]✉

Single-cell RNA-sequencing data have significantly advanced the characterization of cell-type diversity and composition. However, cell-type definitions vary across data and analysis pipelines, raising concerns about cell-type validity and generalizability. With MetaNeighbor, we proposed an efficient and robust quantification of cell-type replicability that preserves dataset independence and is highly scalable compared to dataset integration. In this protocol, we show how MetaNeighbor can be used to characterize cell-type replicability by following a simple three-step procedure: gene filtering, neighbor voting and visualization. We show how these steps can be tailored to quantify cell-type replicability, determine gene sets that contribute to cell-type identity and pretrain a model on a reference taxonomy to rapidly assess newly generated data. The protocol is based on an open-source R package available from Bioconductor and GitHub, requires basic familiarity with Rstudio or the R command line and can typically be run in <5 min for millions of cells.

Introduction

The advent of single-cell technologies has enabled the molecular characterization of heterogeneous tissues at cellular resolution, complementing historical approaches based on marker genes, morphology and electrophysiology. By combining ever improving technologies, consortia efforts have published compendia totaling several hundred thousand cells over multiple modalities to provide comprehensive cell-type taxonomies and exciting new insights on the molecular basis of cell-type identity[1–6]. However, validating computationally derived cell types remains an important challenge. Single-cell data are inherently noisy and subject to laboratory-specific technical variation, which makes them difficult to normalize and combine. Moreover, putative cell types are obtained through unsupervised clustering procedures containing numerous free parameters, raising questions about their reproducibility[7].

Numerous pipelines have been proposed to combine multiple single-cell datasets to obtain a more extensive characterization of cell types[8–14]. Although they vary widely in their mathematical formalisms, these pipelines are based on the idea that data can be corrected, either directly or by embedding cells in a common space that removes unwanted technical variation. These pipelines also provide metrics that quantify how well multiple datasets have been merged. However, these metrics are applied after the correction procedure, by which point the datasets are no longer independent, thus making it difficult to assess whether data have been overcorrected. To accurately measure confidence in cross-dataset signals, we need a direct evaluation of cell-type replicability that preserves dataset independence because this is a better measure of the likelihood of rediscovering a cell type in an independent dataset.

Development of the protocol

MetaNeighbor proposes an easily interpretable cross-dataset framework that quantifies cell-type replicability while preserving dataset independence[15] (Fig. 1a). Replicability is formulated as a straightforward classification task: based on the expression profile of a cell type from a training dataset (hereafter referred to as 'reference dataset'), can I predict which cells belong to a similar cell type in an independent test dataset (hereafter referred to as 'target dataset')? In a nutshell, cells from a given reference cell type vote for their closest neighbors in an independent target dataset, effectively ranking target cells by similarity. This cell-level ranking is aggregated at the cell-type level (in the target dataset) as an area under the receiver operator characteristic curve (AUROC), which reflects

[1]Stanley Institute for Cognitive Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [2]Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA. [3]These authors contributed equally: Stephan Fischer, Megan Crow. ✉e-mail: jgillis@cshl.edu

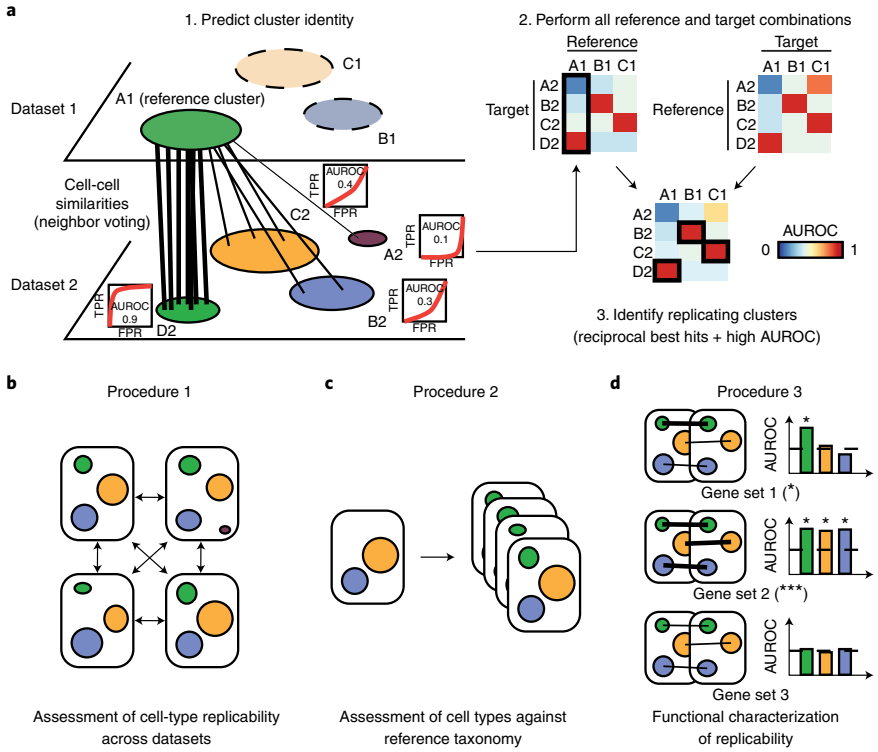


Fig. 1 | MetaNeighbor quantifies and characterizes cell-type replicability. a, Schematic of MetaNeighbor. MetaNeighbor uses a cross-dataset neighbor voting framework to compute cell-type similarities. Cells from a reference cell type (A1) vote for cells in a target dataset according to their similarity (Spearman correlation). Votes can be summarized at the cell-type level as an area under the receiver operating characteristic (AUROC) curve, reflecting the similarity of the reference and target cell types. Formally, the AUROC is computed for each pair of clusters by setting up the following classification problem: 'can cells from the reference cluster (A1) predict which cells belong to the target cluster (e.g., D2)?', where target cells are ranked according to their average similarity to A1 cells, cells from D2 are treated as positives and all other cells from the target dataset are treated as negatives. An AUROC of 1 indicates perfect performance (all D2 cells ranked at the top). This procedure is repeated for all possible reference and target combinations: replicating cell types are identified as reciprocal top hits with high average AUROC. For example, D2 was A1's top hit; reciprocally, A1 was D2's top hit; and the average AUROC of these hits exceeded 0.9. **b**–**d**, Schematic of the three MetaNeighbor procedures. Procedure 1 shows how to assess cell-type replicability by considering all possible pairs of reference and target datasets: highly replicating cell types are identified as recurrent reciprocal top hits across datasets. Procedure 2 shows how to pre-train MetaNeighbor on large reference compendia, enabling rapid identification of reference cell types that are present in a given target dataset. Procedure 3 shows how to functionally characterize replicating cell types by identifying functional gene sets (such as Gene Ontology gene sets) that contribute most to replicability. FPR, false positive rate; TPR, true positive rate.

the proximity of a target cell type to the reference cell type. For example, an AUROC of 0.9 indicates that cells are, on average, ranked in front of 90% of all other cells in the target dataset. If two cell types have a shared biological identity, we expect them to be mutual top matches (when reversing reference and target roles) with a high average AUROC score.

MetaNeighbor's framework is flexible and can be adapted to multiple applications. In its supervised mode, it evaluates the replicability of cell types that are thought or known to be matching a priori. If, as is often the case, the cross-dataset cell-type matching is unknown, MetaNeighbor also provides an unsupervised mode, which automatically identifies the strongest matching cell types and outputs the corresponding AUROCs. MetaNeighbor can also be used for the functional characterization of replicating cell types, identifying pre-defined gene sets (e.g., from the Gene Ontology or the Human Genome Organisation Gene Nomenclature Committee) that contribute to cell-type identity. Finally, the simplicity and scalability of the statistical framework facilitates the setup of computational control experiments.

To adapt to the emergence of large-scale datasets, which now routinely contain 100,000 cells or more, we improved MetaNeighbor's implementation to quickly and interactively assess replicability for data compendia containing a high number of cells and independent experiments[4]. Aside from pure speed improvements, we added the possibility of comparing a dataset to a pre-trained MetaNeighbor model, which allows the rapid evaluation of newly annotated data against comprehensive consortium data.

Applications of the method

In the original publication, we validated MetaNeighbor's ability to characterize rare and transcriptionally subtle cell types[15]. Across three early transcriptomic neuron taxonomies, MetaNeighbor identified 11 strongly replicating interneuron subtypes, along with novel robust marker genes[15]. A similar analysis was performed across seven datasets from the Brain Initiative Cell Census Network (BICCN), sampling from the mouse primary motor cortex by using various laboratories, technologies and clustering pipelines[4]. From the BICCN datasets, MetaNeighbor estimated that most (60/113) of the newly defined cell types were replicable and that these cell types were robustly identified across both sequencing technologies and clustering pipelines.

To further probe the basis of neuronal cell-type identity, MetaNeighbor's functional characterization identified gene families contributing to interneuron identity[16], as well as conserved and divergent gene families across mice, humans and marmosets[6,17]. Applied to a hand-picked selection of cardinal interneurons, MetaNeighbor showed that the identity of these cell types could be characterized by gene families related to synaptic communication, which could be further subdivided into six broad categories (cell-adhesion molecules, transmitter-modulator receptors, ion channels, signaling proteins, neuropeptides and vesicular release components and transcription factors)[16]. Independently, MetaNeighbor was used to highlight conserved expression patterns between mouse and human neurons, showing that replicable cell types shared the same characteristic gene families in the two species[17]. These results were confirmed and extended in a cross-species analysis from the BICCN, in which a MetaNeighbor analysis showed that the expression of genes relevant to neuron physiology (cadherins, ion channels and glutamate transporters) was preferentially conserved across humans and marmosets compared to mice[6].

In the future, we believe that MetaNeighbor's scalability will be further exploited to make accessible large consortium data (by querying pre-trained models) and meta-analyses (by enabling the comparison of large dataset compendia).

Overview of the protocol

We present three procedures that use MetaNeighbor to quantify and characterize cell-type replicability. In Procedure 1, we use unsupervised MetaNeighbor to identify replicable cell types across four pancreas datasets (Fig. 1b). In Procedure 2, we show how to assess newly annotated cell types against a large reference taxonomy by pre-training a MetaNeighbor model (Fig. 1c). Finally, in Procedure 3, we use supervised MetaNeighbor to investigate the molecular basis of cell-type identity by finding functional gene sets that contribute highly to cell-type replicability (Fig. 1d). All code blocks can be run in R command line, Rstudio, RMarkdown notebooks or a jupyter notebook with an R kernel.

To illustrate the procedures, we have chosen two data compendia that sample widely across laboratories and technologies. The pancreas compendium is commonly used in dataset integration assessments and contains samples from four different single-cell protocols, resulting in composition variability (e.g., rare epsilon cells are not detected in all datasets). The BICCN compendium is unique with respect to the size of the data (around half a million cells from nine independent datasets targeting the same brain region), complexity of taxonomy (over 100 cell types) and wide array of technologies used (single-cell transcriptomics, single-nuclei transcriptomics, single-nuclei methylation and single-nuclei chromatin accessibility).

Comparison with other methods

MetaNeighbor is related to four families of techniques: integrative methods designed to merge multiple datasets, methods for cell-type annotation, metrics evaluating the quality of dataset integration and metrics evaluating clustering robustness.

Integrative methods combine multiple datasets to improve cell-type characterization. Mathematically, the rationale of these methods is to find a joint space that maximally preserves shared biological variation and removes all other variation. Popular methods include mutual nearest neighbors (MNNs)-based correction[8], Seurat[9,18], LIGER[10], Harmony[11], Scanorama[12] and Conos[13],

which have been extensively reviewed[19,20] and benchmarked[21,22]. The similarity with MetaNeighbor is the idea that batch effects are effectively attenuated by identifying MNNs. However, the aim is different: in integrative methods, MNNs are used to maximally correct and align datasets, while MetaNeighbor evaluates the similarity of nearest neighbors to quantify the amount of replicable signal. Methods for cell-type annotation[23] (annotation of unlabeled cells by comparison with an annotated reference dataset), such as scmap[24] or Seurat[9], are based on the same rationale of nearest neighbors. Again, the key difference with MetaNeighbor is the intent and interpretability: annotation methods output cell-type labels, whereas MetaNeighbor outputs a statistic that lets the user evaluate how much replicable signal there is to begin with, providing complementary information about the expected robustness of these methods.

To evaluate dataset proximity, integration methods use metrics that quantify how successful the integration was. Popular metrics[21,22] include batch-mixing metrics, such as the k-nearest-neighbor batch-effect test (kBET)[25] and the local inverse Simpson's index (LISI)[11], and cell-type conservation metrics, such as the average silhouette width (ASW) and the adjusted Rand index (ARI). Batch-mixing metrics test whether datasets are well mixed in the joint dataset space, which is seen as effective attenuation of technical variation. Cell-type conservation metrics test whether biological variation is conserved; for example, ASW tests whether cell types are well separated in the joint space and ARI checks that cell types obtained by clustering the integrated data are consistent with annotations from the independent data. However, these metrics are used to assess the performance of dataset integration, rather than evaluating the amount of replicable signal. Furthermore, in cases where datasets have limited biological overlap, these mixing metrics can be difficult to interpret, particularly if they are seen as scores that methods try to optimize. By always keeping datasets and annotations independent, MetaNeighbor's focus is on rapidly identifying where data structure agrees but also differs (i.e., when cell types do not align across datasets or clustering pipelines).

In its design, MetaNeighbor is closest to methods for the validation of sample clustering, where the reproducibility of cluster structure across independent datasets is interpreted as 'biological significance'[26] and which has been used to evaluate the reproducibility of cancer subtyping from microarray data[26,27]. MetaNeighbor extends this framework to single-cell genomics data and enables direct interpretation of gene sets whose co-expression drives replicability. Consensus clustering methods, such as SC3[28] or scrattch.hicat[29], are based on a similar idea, but the focus is on the quantification of robustness to clustering parameters or methods, whereas MetaNeighbor's focus is on cross-dataset replicability, which includes variability due to clustering methods, but also laboratory-specific or conditional variability.

Experimental design

MetaNeighbor's aim is to accurately estimate cell-type replicability by preserving dataset independence. Consequently, we recommend using raw data and, if possible, cell-type labels obtained by clustering each dataset independently (such as annotations from the original publication, if external data are used), which will help evaluate the robustness of cell types to the clustering procedure. If MetaNeighbor is run on data or labels that have been obtained through an integrative clustering technique, the user must be aware that dataset independence has been broken. Practically speaking, the integrative approach has a fitting step that will make datasets artificially similar, leading to optimistic replicability estimations. Similarly, even if datasets are truly independent, but MetaNeighbor is run in its unsupervised mode, replicability estimations will be slightly inflated, because the framework will automatically match the closest cell types.

Another problem that prevents accurate replicability estimation is the confounding of technical and biological variation, in the most extreme case when each cell type has been sequenced in a different batch. MetaNeighbor works best when batches are approximately balanced in terms of cell-type composition but can be adapted to confounded experimental designs. For example, MetaNeighbor has been adapted to an extreme case of confounding by replacing cross-dataset validation with simple cross-validation[16]. Results remain interpretable biologically but must be interpreted with greater care.

Thanks to its scalability, MetaNeighbor can be used to implement carefully designed control experiments. Although we chose to output AUROCs because of their interpretability, their exact understanding depends on dataset composition and varies with cell-type rarity and subtlety of transcriptomic differences between cell types. We proposed several control experiments that are simple to implement and help pinpoint how much signal can be expected to be extracted from the datasets under investigation[15].

Expertise needed to implement the protocol

All three procedures are based on running R functions and require familiarity with the RStudio integrated development environment or the R command line.

Limitations

Classification problems, and AUROCs in particular, are known to be affected by class imbalance (cell-type composition in our case). Overall, MetaNeighbor is robust to such imbalances, but we found that scores can be distorted when class imbalance becomes extreme, in particular when there is no overlap between datasets. Benchmarking and evaluation to explore variability in performance is of continued interest and probably increasing importance if sampled data become more targeted. MetaNeighbor can be used to compare transcriptomic and epigenomic data, such as chromatin accessibility and methylation assays, but, in our experience, results are harder to interpret, in particular because there is no consensus on how to map genome-wide measurements with transcriptomic-wide measurements (see Multimodal analyses in Anticipated results). For very large datasets, MetaNeighbor can be memory intensive: when comparing several hundred thousand cells, we recommend using compute units or clusters that have a high memory capacity (>50–100 Gb).

Materials

Equipment

Hardware

- A personal computer with internet connection and ≥8 GB of random access memory, ideally 16 GB of random access memory for Procedure 3

Software

- RStudio (https://rstudio.com/products/rstudio/download/), Jupyter (https://jupyter.org/install) or R command line with R version 3.6 or higher
- Key R package: the MetaNeighbor library, available on GitHub (https://github.com/gillislab/MetaNeighbor/) and Bioconductor version 3.12 or higher (https://www.bioconductor.org/packages/release/bioc/html/MetaNeighbor.html)
- Other R packages: scRNAseq, tidyverse, org.Hs.eg.db and UpSetR, available from Bioconductor (https://www.bioconductor.org/install/) and the Comprehensive R Archive Network (https://cran.rstudio.com/)

Datasets

All procedures are based on published and publicly available datasets:

- Human pancreas datasets, accessed through the R/Bioconductor scRNAseq package (https://bioconductor.org/packages/release/data/experiment/html/scRNAseq.html), which makes available a collection of publicly available single-cell transcriptomics datasets
- The mouse primary visual cortex dataset, accessed through the scRNAseq package
- The Brain Initiative Cell Census Network (BICCN) dataset for the mouse primary motor cortex. The full dataset is available on the Neuroscience Multi-Omic archive (https://assets.nemoarchive.org/dat-ch1nqb7), and the relevant subset of the dataset is directly available on Figshare (https://doi.org/10.6084/m9.figshare.13020569.v2)

Equipment setup

This section walks through the installation process of MetaNeighbor and the packages used in the protocol. The installation process takes 1–20 min, depending on the number of dependencies already available. All code blocks can be run in R command line, Rstudio, RMarkdown notebooks or a jupyter notebook with an R kernel. ▲ **CRITICAL** The installation process may create conflicts in the notebook environment. We recommend running the installation process in a separate R shell or restarting the Rstudio R environment after the installation has completed and before starting one of the procedures. Start by installing the latest MetaNeighbor package from the Gillis laboratory GitHub page.

```
if (!require("devtools")) {
&install.packages("devtools", quiet=TRUE)
}
devtools::install_github("gillislab/MetaNeighbor")
```

Note that the latest stable version of MetaNeighbor is also available through Bioconductor by running `BiocManager::install("MetaNeighbor")`. We recommend using the latest development version from GitHub, because some of the functionalities illustrated in this protocol require Bioconductor version 3.12 or higher to work (available only with R version 4.0 or higher).

? TROUBLESHOOTING

Next, install the following packages, which are not necessary to run MetaNeighbor itself but are needed to run the protocol.

```
to_install = c("scRNAseq", "tidyverse", "org.Hs.eg.db", "UpSetR")
installed = sapply(to_install, requireNamespace)
if (sum(!installed) > 0) {
  if (!requireNamespace("BiocManager", quietly = TRUE)) {
  install.packages("BiocManager")
  BiocManager::install()
  }
  BiocManager::install(to_install[!installed])
}
```

▲ CRITICAL STEP Do not forget to restart the R session at this stage.

? TROUBLESHOOTING

Procedure 1: assessment of cell-type replicability with unsupervised MetaNeighbor

▲ CRITICAL Procedure 1 demonstrates how to compute and visualize cell-type replicability across four human pancreas datasets, detailing how to download and reformat the datasets with the SingleCellExperiment (SCE) package and how to compute and interpret MetaNeighbor AUROCs.

Creation of a merged SCE dataset ● Timing 1–2 min

1 We consider four pancreatic datasets along with their independent annotation (from the original publications). MetaNeighbor expects a gene-by-cell matrix encapsulated in a SummarizedExperiment format. We recommend the SCE package, an extension of the SummarizedExperiment class designed to efficiently store large single-cell datasets, because it is able to handle sparse matrix formats. Load the pancreas datasets by using the scRNAseq package, which provides annotated datasets that are already in the SCE format:

```
library(scRNAseq)
my_data <- list(
  baron = BaronPancreasData(),
  lawlor = LawlorPancreasData(),
  seger = SegerstolpePancreasData(),
  muraro = MuraroPancreasData()
)
```

Note that Seurat objects can easily be converted into SCE objects by using the `as.SingleCellExperiment` function for Seurat v3 objects and `Convert(from = seurat_object, to = "sce")` for Seurat v2 objects.

2 MetaNeighbor's `mergeSCE` function can be used to merge multiple SCE objects. Importantly, the output object will be restricted to genes, metadata columns and assays that are common to all datasets. Before using `mergeSCE`, make sure that gene and metadata information aligns across datasets. Start by checking if gene information aligns (stored in the `rownames` slot of the SCE object):

```
lapply(my_data, function(x) head(rownames(x), 3))
## $baron
## [1] "A1BG" "A1CF" "A2M"
##
## $lawlor
## [1] "ENSG00000229483" "ENSG00000232849" "ENSG00000229558"
```

```
##
## $seger
## [1] "SGIP1" "AZIN2" "CLIC4"
##
## $muraro
## [1] "A1BG-AS1__chr19" "A1BG__chr19" "A1CF__chr10"
```

Two datasets (Baron and Segerstolpe) use gene symbols, one dataset (Muraro) combines symbols with chromosome information (to avoid duplicate gene names) and the last dataset (Lawlor) uses Ensembl identifiers. Here, we convert all gene names to unique gene symbols. Start by converting gene names in the Muraro dataset by using the symbols stored in the `rowData` slot of the SCE object and remove all duplicated gene symbols:

```
rownames(my_data$muraro) <- rowData(my_data$muraro)$symbol
my_data$muraro <- my_data$muraro[!duplicated(rownames(my_data$mur-
aro)),]
```

Next, convert Ensembl IDs to gene symbols in the Lawlor dataset, removing all IDs with no match and all duplicated symbols:

```
library(org.Hs.eg.db)
symbols <- mapIds(org.Hs.eg.db, keys=rownames(my_data$lawlor),
  keytype="ENSEMBL", column="SYMBOL")
keep <-!is.na(symbols) &!duplicated(symbols)
my_data$lawlor <- my_data$lawlor[keep,]
rownames(my_data$lawlor) <- symbols[keep]
```

3 We now turn our attention to metadata, which are stored in the `colData` slot of the SCE objects. Here, make sure that the column that contains cell-type information is labeled identically in all datasets:

```
lapply(my_data, function(x) colnames(colData(x)))
## $baron
## [1] "donor" "label"
##
## $lawlor
## [1] "title" "age" "bmi" "cell type"
## [5] "disease" "islet unos id" "race" "Sex"
##
## $seger
## [1] "Source Name" "individual"
## [3] "single cell well quality" "cell type"
## [5] "disease" "sex"
## [7] "age" "body mass index"
##
## $muraro
## [1] "label" "donor" "plate"
```

Two datasets have the cell-type information in the 'cell type' column, whereas the other two have the cell-type information in the 'label' column. Add a 'cell type' column in the latter two datasets:

```
my_data$baron$"cell type" <- my_data$baron$label
my_data$muraro$"cell type" <- my_data$muraro$label
```

4 Last, check that count matrices, stored in the `assays` slot, have identical names:

```
lapply(my_data, function(x) names(assays(x)))
## $baron
```

```
## [1] "counts"
##
## $lawlor
## [1] "counts"
##
## $seger
## [1] "counts"
##
## $muraro
## [1] "counts"
```

The count matrices are all stored in an assay named 'counts'; no change is needed here.

5 Now that gene, cell-type and count matrix information is aligned across datasets, create a merged dataset by using `mergeSCE`, which takes a list of SCE objects as an input and outputs a single SCE object:

```
library(MetaNeighbor)
fused_data = mergeSCE(my_data)
dim(fused_data)
## [1] 15295 15793
head(colData(fused_data))
## DataFrame with 6 rows and 2 columns
## cell type study_id
## <character> <character>
## human1_lib1.final_cell_0001 acinar baron
## human1_lib1.final_cell_0002 acinar baron
## human1_lib1.final_cell_0003 acinar baron
## human1_lib1.final_cell_0004 acinar baron
## human1_lib1.final_cell_0005 acinar baron
## human1_lib1.final_cell_0006 acinar baron
```

The new dataset contains 15,295 common genes, 15,793 cells and two metadata columns: a concatenated 'cell type' column and 'study_id', a column created by `mergeSCE` containing the name of the original studies (corresponding to the names provided in the 'my_data' list).

6 To obtain a cursory overview of cell-type composition by study, cross-tabulate cell-type annotations by study IDs:

```
table(fused_data$"cell type", fused_data$study_id)
##
## baron lawlor muraro seger
## acinar 958 0 219 0
## Acinar 0 24 0 0
## acinar cell 0 0 0 185
## activated_stellate 284 0 0 0
## alpha 2326 0 812 0
## Alpha 0 239 0 0
## alpha cell 0 0 0 886
## beta 2525 0 448 0
## Beta 0 264 0 0
## beta cell 0 0 0 270
## [Rest of output omitted]
```

Most cell types are present in all datasets, so we expect MetaNeighbor to find multiple high-confidence matches across datasets. There are slight typographic differences in cell-type annotations (e.g., ductal/Ductal), but we recommend keeping the author annotations at this stage. The only procedure that requires identical annotations across datasets is Procedure 3, where we perform functional characterization of replicating cell types.

7 To avoid having to recreate the merged object, save the R object to a file by using R's RDS format:

```
saveRDS(fused_data, "merged_pancreas.rds")
```

PAUSE POINT the remaining sections of the procedure can be run at a later time in a new R session.

Hierarchical cell-type replicability analysis ● Timing 1 min

8 Start by loading the MetaNeighbor (analysis) and the SCE (data handling) libraries, as well as the previously created pancreas dataset:

```
library(MetaNeighbor)
library(SingleCellExperiment)
pancreas_data = readRDS("merged_pancreas.rds")
```

9 To perform neighbor voting and identify replicating cell types, MetaNeighbor builds a cell-cell similarity network, which we defined as the Spearman correlation over a user-defined set of genes. We found that we obtained best results by picking genes that are highly variable across datasets, which can be done by using the `variableGenes` function. Select highly variable genes for the pancreas datasets:

```
global_hvgs = variableGenes(dat = pancreas_data,
  exp_labels = pancreas_data$study_id)
length(global_hvgs)
## [1] 600
```

The function returns a list of 600 genes that were detected as highly variable in each of the four datasets. In our experience, we obtained best performance for gene sets ranging from 200 to 1,000 variable genes. In general, using a larger number of datasets selects robustly varying genes, enabling high performance with a smaller number of genes. However, if `variableGenes` returns a gene set that is too small (in particular, when you are comparing a large number of datasets), the number of genes can be increased by setting the 'min_recurrence' parameter. For example, by setting 'min_recurrence = 2', we keep all genes that are highly variable in at least two of the four datasets. In addition, genes are sorted by relevance in the latest version of MetaNeighbor, so it is always possible to select a smaller number of genes. For example, `global_hvgs[1:500]` selects the top 500 highly variable genes that are recurrent across all four datasets. This option can be used to validate that performance is robust over gene sets of increasing size.

▲ CRITICAL STEP Variable genes are MetaNeighbor's only parameter and must be selected with care (see Anticipated results).

? TROUBLESHOOTING

10 The merged dataset and a set of biologically meaningful genes are all that is needed to run MetaNeighbor and obtain cell-type similarities. Because the dataset is large (>10,000 cells), run the fast implementation of MetaNeighbor ('fast_version=TRUE'):

```
aurocs = MetaNeighborUS(var_genes = global_hvgs,
  dat = pancreas_data,
  study_id = pancreas_data$study_id,
  cell_type = pancreas_data$"cell type",
  fast_version = TRUE)
```

`MetaNeighborUS` returns a cell-type-by-cell-type matrix containing cell-type similarities. Cell-type similarities are defined as an AUROC, which range between 0 and 1, where 0 indicates low similarity, and 1 indicates high similarity. Note that the 'fast_version = TRUE' parameter uses a slightly simplified version of MetaNeighbor that is significantly faster and more memory efficient. It should always be used on large datasets (>10,000 cells) but can also be run on smaller datasets and yields equivalent results to the original MetaNeighbor algorithm.

? TROUBLESHOOTING

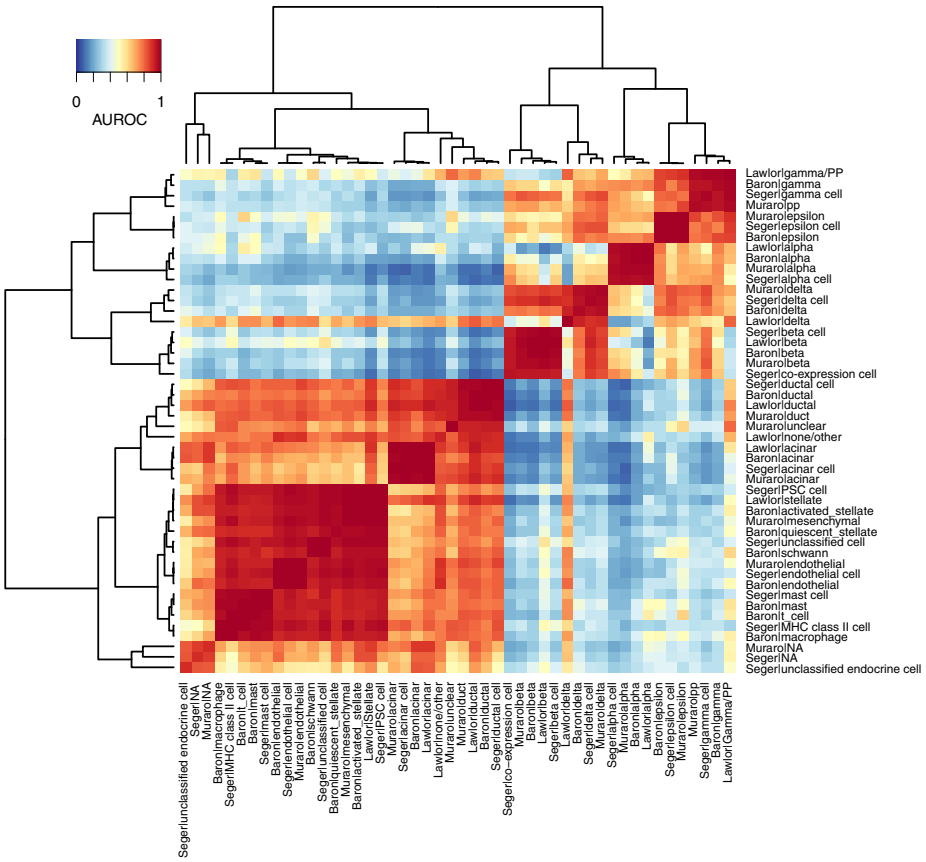


Fig. 2 | Cell types from four pancreas datasets cluster according to their biological similarity. Heatmap based on MetaNeighbor AUROCs. Red indicates high similarity, and blue indicates low similarity. By applying hierarchical clustering, replicating cell types group together (dark red squares), and biologically related cell types (e.g., endocrine cell types, such as alpha, beta and gamma cells) form secondary groups (large light red squares). MHC, major histocompatibility complex; NA, not available (missing value in R); PP, pancreatic polypeptide; PSC, pancreatic stellate cells.

11 For ease of interpretation, visualize AUROCs as a heatmap, where rows and columns are cell types from all the datasets:

```
plotHeatmap(aurocs, cex = 0.5)
```

In the heatmap (Fig. 2), the color of each square indicates the proximity of a pair of cell types, ranging from blue (low similarity) to red (high similarity). For example, 'baron|gamma' (second row) is highly similar to 'seger|gamma' (third column from the right) but very different from 'muraro|duct' (middle column). To group similar cell types together, `plotHeatmap` applies hierarchical clustering on the AUROC matrix. On the heatmap, we see two large red blocks that indicate hierarchical structure in the data, with endocrine cell types clustering together (e.g., alpha, beta and gamma) and non-endocrine cells on the other side (e.g., amacrine, ductal and endothelial). Note that each red block is composed of smaller red blocks, indicating that cell types can be matched at an even higher resolution. The presence of off-diagonal patterns (e.g., 'lawlor|Gamma/PP' and 'lawlor|Delta') suggests the presence of doublets or contamination, but the heatmap is dominated by the clear presence of red blocks, which is a strong indicator of replicability.

In the latest version of MetaNeighbor, we increased the flexibility of heatmaps. `plotHeatmap` internally relies on `gplots::heatmap.2`:. You can pass any valid `heatmap.2` parameter to `plotHeatmap`; for example, the 'ColSideColors' parameter can be used to annotate the columns

Table 1 | Reciprocal top hits with high AUROC identify replicating cell types

Study_ID|Celltype_1	Study_ID|Celltype_2	Mean_AUROC	Match_type
seger|epsilon cell	muraro|epsilon	1.00	Reciprocal_top_hit
seger|epsilon cell	baron|epsilon	1.00	Above_0.9
baron|mast	seger|mast cell	1.00	Reciprocal_top_hit
seger|endothelial cell	muraro|endothelial	1.00	Reciprocal_top_hit
lawlor|Stellate	seger|PSC cell	1.00	Reciprocal_top_hit
baron|macrophage	seger|MHC class II cell	1.00	Reciprocal_top_hit
muraro|endothelial	baron|endothelial	1.00	Above_0.9
lawlor|Stellate	baron|activated_stellate	1.00	Above_0.9
baron|acinar	lawlor|Acinar	1.00	Reciprocal_top_hit
seger|PSC cell	muraro|mesenchymal	1.00	Above_0.9
baron|alpha	lawlor|Alpha	1.00	Reciprocal_top_hit
lawlor|Acinar	seger|acinar cell	1.00	Above_0.9
baron|schwann	seger|unclassified cell	1.00	Reciprocal_top_hit
seger|acinar cell	muraro|acinar	0.99	Above_0.9
lawlor|Beta	seger|beta cell	0.99	Reciprocal_top_hit
baron|ductal	seger|ductal cell	0.99	Reciprocal_top_hit
lawlor|Beta	baron|beta	0.99	Above_0.9
baron|ductal	lawlor|Ductal	0.99	Above_0.9
seger|MHC class II cell	baron|t_cell	0.99	Above_0.9
baron|gamma	lawlor|Gamma/PP	0.99	Reciprocal_top_hit
lawlor|Beta	muraro|beta	0.98	Above_0.9
seger|ductal cell	muraro|duct	0.98	Above_0.9
lawlor|Alpha	muraro|alpha	0.98	Above_0.9
seger|PSC cell	baron|quiescent_stellate	0.98	Above_0.9
lawlor|Gamma/PP	seger|gamma cell	0.98	Above_0.9
seger|delta cell	muraro|delta	0.98	Reciprocal_top_hit
lawlor|Gamma/PP	muraro|pp	0.98	Above_0.9
muraro|alpha	seger|alpha cell	0.98	Above_0.9
muraro|delta	baron|delta	0.96	Above_0.9
baron|beta	seger|co-expression cell	0.95	Above_0.9
seger|ductal cell	muraro|unclear	0.93	Above_0.9
baron|delta	lawlor|Delta	0.92	Above_0.9
baron|ductal	lawlor|None/Other	0.91	Above_0.9

Pairs of cell types that meet the following criteria: reciprocal top hits (the cell types preferentially vote for each other in the cross-dataset voting framework) or average AUROC >0.9 (average taken by switching the reference and target dataset). Note that each study uses its own convention for cell-type annotations, resulting in differences in names and capitalization. MHC, major histocompatibility complex; PP, pancreatic polypeptide; PSC, pancreatic stellate cells.

of the heatmap (one color by dataset). Alternatively, the `MetaNeighbor::ggPlotHeatmap` function returns a customizable ggplot2 object.

? TROUBLESHOOTING

12 To identify pairs of replicable cell types, run the following function:

```
topHits(aurocs, dat = pancreas_data, study_id = pancreas_data$study_id,
  cell_type = pancreas_data$"cell type", threshold = 0.9)
```

`topHits` relies on a simple heuristic: a pair of cell types is replicable if they are reciprocal top hits (they preferentially vote for each other), and the AUROC exceeds a given threshold value (in our experience, 0.9 is a good heuristic value). We find a long list of replicable endocrine cell types (e.g., epsilon, alpha and beta cells) and non-endocrine cell types (e.g., mast, endothelial and acinar cells) (Table 1). This list provides strong evidence that these cell types are robust, because they are identified across all datasets with high AUROC.

13 In the case in which there is a clear structure in the data (here, endocrine versus non-endocrine), we can refine AUROCs by splitting the data. AUROCs have a simple interpretation: an AUROC of 0.6

indicates that cells from a given cell type are ranked in front of 60% of other target cells. However, this interpretation is outgroup dependent: because endocrine cells represent ~65% of cells, even an unrelated pair of non-endocrine cell types will have an AUROC >0.65, because non-endocrine cells will always be ranked in front of endocrine cells.

By starting with the full datasets, we uncovered the global structure in the data (endocrine versus non-endocrine). However, to evaluate replicability of endocrine cell types and reduce dataset composition effects, we can make the assessment more stringent by restricting the outgroup to close cell types (i.e., by keeping only endocrine subtypes). Split cell types in two by using the `splitClusters` function and retain only endocrine cell types:

```
level1_split = splitClusters(aurocs, k = 2)
level1_split
## [output omitted]
first_split = level1_split[[2]]
```

By outputting 'level1_split', we found that the cell types were nicely split between non-endocrine and endocrine, and that endocrine cell types were in the second element of the list. Note that `splitClusters` applies a simple hierarchical clustering algorithm to separate cell types; however, cell types can be selected manually in more complex scenarios.

14 Repeat the MetaNeighbor analysis on endocrine cells only. First, subset the data to the endocrine cell types that were previously stored in 'first_split':

```
full_labels = makeClusterName(pancreas_data$study_id,
  pancreas_data$"cell type")
subdata = pancreas_data[, full_labels %in% first_split]
dim(subdata)
## [1] 15295 9341
```

The new dataset contains the 9,341 putative endocrine cells.

15 To focus on variability that is specific to endocrine cells, re-pick highly variable genes:

```
var_genes = variableGenes(dat = subdata, exp_labels = subdata
$study_id)
```

? TROUBLESHOOTING

16 Finally, recompute cell-type similarities and visualize AUROCs:

```
aurocs = MetaNeighborUS(var_genes = var_genes,
  dat = subdata, fast_version = TRUE,
  study_id = subdata$study_id,
  cell_type = subdata$"cell type")
plotHeatmap(aurocs, cex = 0.7)
```

The resulting heatmap (Fig. 3a) illustrates an example of a strong set of replicating cell types: when the assessment becomes more stringent (restriction to closely related cell types), the similarity of replicating cell types remains strong (AUROC of ~1 for alpha, beta, gamma, delta and epsilon cells), whereas the cross-cell-type similarity decreases (shift from red to blue; e.g., similarity of alpha and beta cell types has shifted from orange/red in the global heatmap to dark blue in the endocrine heatmap) by virtue of zooming in on a subpart of the dataset.

? TROUBLESHOOTING

17 We can continue to zoom in as long as there are at least two cell types per dataset. Repeat Steps 13–16 to split the endocrine cell types:

```
level2_split = splitClusters(aurocs, k = 3)
my_split = level2_split[[3]]
subdata = pancreas_data[, full_labels %in% my_split]
var_genes = variableGenes(dat = subdata, exp_labels = subdata
$study_id)
```

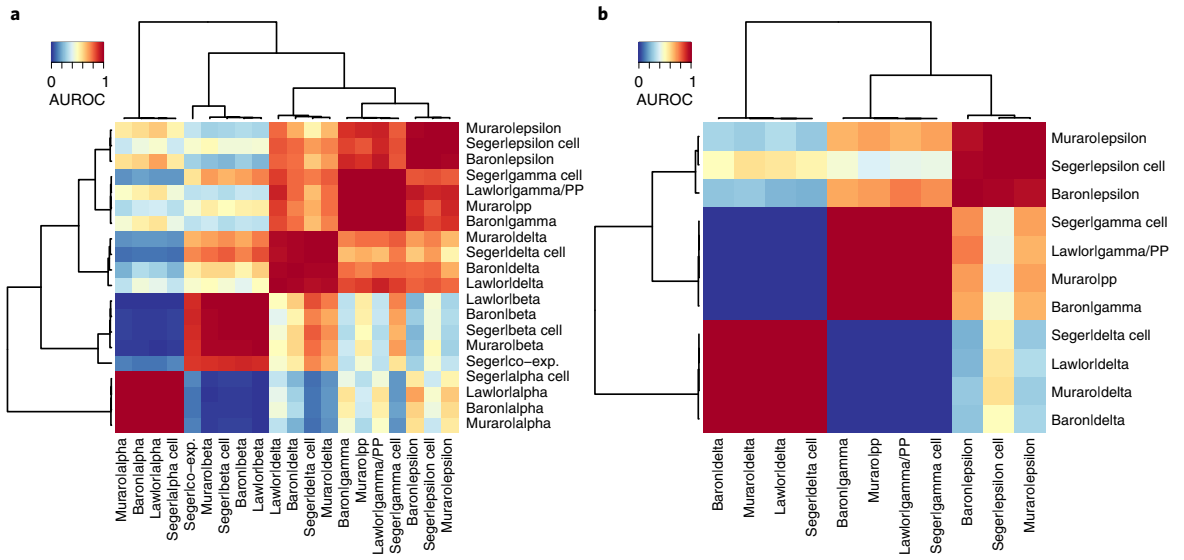


Fig. 3 | Restricting the four pancreas datasets to endocrine subtypes allows for a more stringent replicability assessment. a, Heatmap based on MetaNeighbor AUROCs applied to endocrine cell types, where cell types are grouped by applying hierarchical clustering. Red squares represent replicating cell types (alpha, beta, gamma, delta and epsilon cells). **b**, AUROCs can be refined as long as there are two cell types per dataset. Heatmap based on MetaNeighbor AUROCs applied to gamma, delta and epsilon cells, where cell types are grouped by applying hierarchical clustering. Red squares represent replicating cell types (gamma, delta and epsilon cells). PP, pancreatic polypeptide.

```
length(var_genes)
## [1] 274
aurocs = MetaNeighborUS(var_genes = var_genes,
  dat = subdata, fast_version = TRUE,
  study_id = subdata$study_id,
  cell_type = subdata$"cell type")
plotHeatmap(aurocs, cex = 1)
```

Here, we remove the alpha and beta cells (representing close to 85% of endocrine cells) and validate that, even when restricting to neighboring cell types, there is still a clear distinction between delta, gamma and epsilon cells (AUROC of ~1; Fig. 3b).

? TROUBLESHOOTING

Stringent assessment of replicability with one-vs-best AUROCs ● Timing 1 min

▲ **CRITICAL** In the previous section, we created progressively more stringent replicability assessments by selecting more and more specific subsets of related cell types. As an alternative, we provide the 'one_vs_best' parameter, which offers similar results without having to restrict the dataset manually. In this scoring mode, MetaNeighbor will automatically identify the two closest matching cell types in each target dataset and compute an AUROC based on the voting result for cells from the closest match against cells from the second-closest match. Essentially, we are asking how easily a cell type can be distinguished from its closest neighbor.

18 To obtain one-vs-best AUROCs, run the same command as before with two additional parameters: 'one_vs_best = TRUE' and 'symmetric_output = FALSE':

```
best_hits = MetaNeighborUS(var_genes = global_hvgs,
  dat = pancreas_data,
  study_id = pancreas_data$study_id,
  cell_type = pancreas_data$"cell type",
  fast_version = TRUE,
  one_vs_best = TRUE, symmetric_output = FALSE)
plotHeatmap(best_hits, cex = 0.5)
```

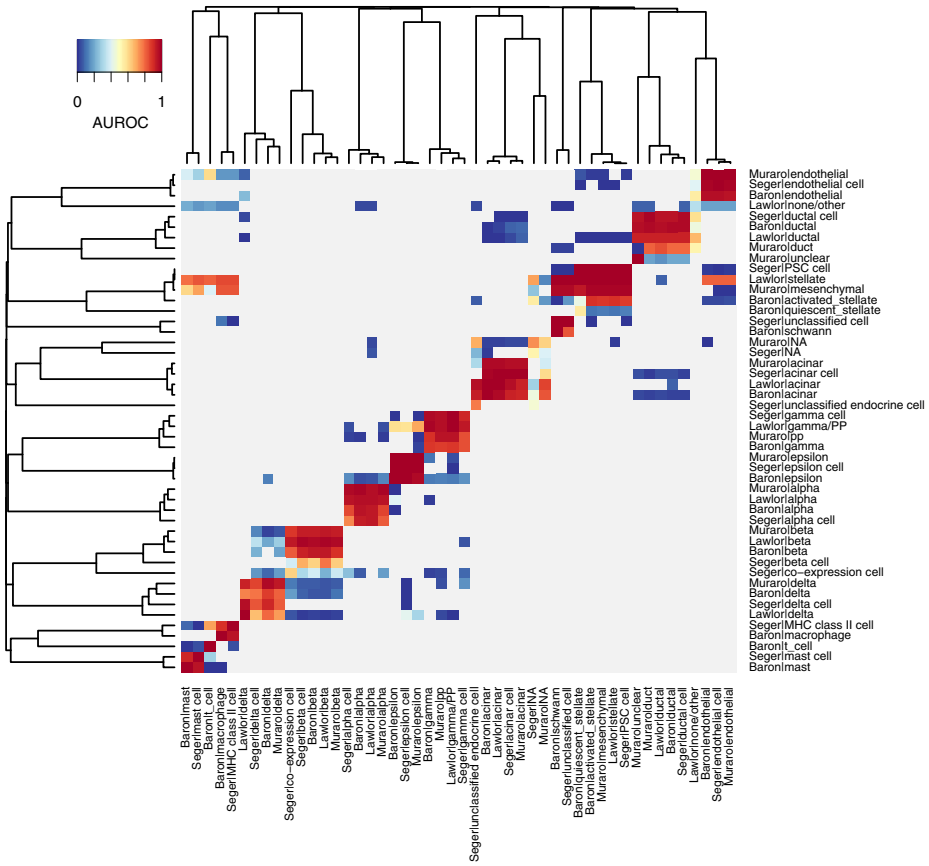


Fig. 4 | 1-vs-best AUROCs automatically identify each cell type's closest outgroup. Heatmap based on MetaNeighbor 1-vs-best AUROCs, where cell types are grouped by applying hierarchical clustering. Reference cell types are shown as columns, and target cell types are shown as rows. Red values indicate each reference cell type's best hit, and blue values indicate the closest outgroup (one value per target dataset). All other cell-type combinations are shown in gray.

The interpretation of the heatmap is slightly different compared to one-vs-all AUROCs (Fig. 4). First, because we compare only the two closest cell types, most cell-type combinations are not tested (indicated by NA (not available), shown in gray on the heatmap). Second, by setting 'symmetric_output = FALSE', we broke the symmetry of the heatmap: reference cell types are shown as columns, and target cell types are shown as rows. Because each cell type is tested against only two cell types in each target dataset (closest and second-closest match), we have eight values per column (two per dataset). This representation helps to rapidly identify a cell type's closest hits as well as its closest outgroup. For example, ductal cells (second red square from the top right) strongly match with each other (one-vs-best AUROC >0.8), and acinar cells are their closest outgroup (blue segments in the same column). The nonsymmetric view makes it clear when best hits are not reciprocal. For example, mast cells (first two columns) heavily vote for 'lawlor|Stellate' and 'muraro|mesenchymal', but this vote is not reciprocal. This pattern indicates that the mast-cell type is missing in the Lawlor and Muraro datasets: because mast cells have no natural match in these datasets, they vote for the next closest cell type (stellate cells). The lack of reciprocity in voting is an important tool to detect imbalances in dataset composition.

? TROUBLESHOOTING

19 When using one-vs-best AUROCs, we recommend extracting replicating cell types as meta-clusters. Cell types are part of the same meta-cluster if they are reciprocal best hits. Note that if cell type A is the reciprocal best hit of B and C, all three cell types are part of the same meta-cluster, even if B and C are not reciprocal best hits. To further filter for strongly replicating cell types, we specify an

AUROC threshold (in our experience, 0.7 is a strong one-vs-best AUROC threshold). To extract meta-clusters and summarize the strength of each meta-cluster, run the following functions:

```
mclusters = extractMetaClusters(best_hits, threshold = 0.7)
mcsummary = scoreMetaClusters(mclusters, best_hits)
```

The `scoreMetaClusters` function provides a good summary of meta-clusters, ordering cell types by the number of datasets in which they replicate, then by average AUROC. We find 12 cell types that have strong support across at least two datasets, with seven cell types replicating across all four datasets. Eight cell types are tagged as 'outlier', indicating that they had no strong match in any other dataset. These cell types usually contain doublets, low-quality cells or contaminated cell types. To rapidly visualize the number of robust cell types, the replicability structure can be summarized as an Upset plot with the `plotUpset` function (Fig. 5a).

```
plotUpset(mclusters)
```

To further investigate the robustness of meta-clusters, they can be visualized as heatmaps (called 'cell-type badges') with the `plotMetaClusters` function. Because the function generates one heatmap per meta-cluster, save the output to a PDF file to facilitate investigation:

```
pdf("meta_clusters.pdf")
plotMetaClusters(mclusters, best_hits)
dev.off()
```

Each badge shows an AUROC heatmap restricted to one specific meta-cluster. These badges help diagnose cases in which AUROCs are lower in a specific reference or target dataset. For example, the 'muraro|duct' cell type has systematically lower AUROCs, suggesting the presence of contaminating cells in another cell type (probably in the 'muraro|unclear' cell type) (Fig. 5b).

20 The last visualization is an alternative representation of the AUROC heatmap as a graph, which is particularly useful for large datasets. In this graph, top votes (AUROC >0.5) are shown in gray, and outgroup votes (AUROC <0.5) are shown in orange. To highlight close calls, we recommend keeping only strong outgroup votes (here, with AUROC ≥0.3). To build and plot the cluster graph, run the following functions:

```
cluster_graph = makeClusterGraph(best_hits, low_threshold = 0.3)
plotClusterGraph(cluster_graph, pancreas_data$study_id,
  pancreas_data$"cell type", size_factor=3)
```

We note that there are several orange edges, indicating that some cell types had two close matches (Fig. 5c). To investigate the origin of these close calls, we can focus on a cluster of interest (coi). Take a closer look at 'baron|epsilon', query its closest neighbors in the graph with `extendClusterSet` and then zoom in on its subgraph with `subsetClusterGraph`:

```
coi = "baron|epsilon"
coi = extendClusterSet(cluster_graph, initial_set=coi,
  max_neighbor_distance=2)
subgraph = subsetClusterGraph(cluster_graph, coi)
plotClusterGraph(subgraph, pancreas_data$study_id,
  pancreas_data$"cell type", size_factor=5)
```

In the 'baron|epsilon' case, we find that the epsilon cell type is missing in the Lawlor dataset; thus, there is no natural match for the Baron epsilon cell type (Fig. 5d). In such cases, votes are frequently nonreciprocal and equally split between two unrelated cell types (here, 'Lawlor|Gamma/PP' and 'Lawlor|Alpha'). In general, the cluster graph can be used to understand how meta-clusters are extracted and why some clusters are tagged as outliers and to diagnose problems where the resolution of cell types differs across datasets.

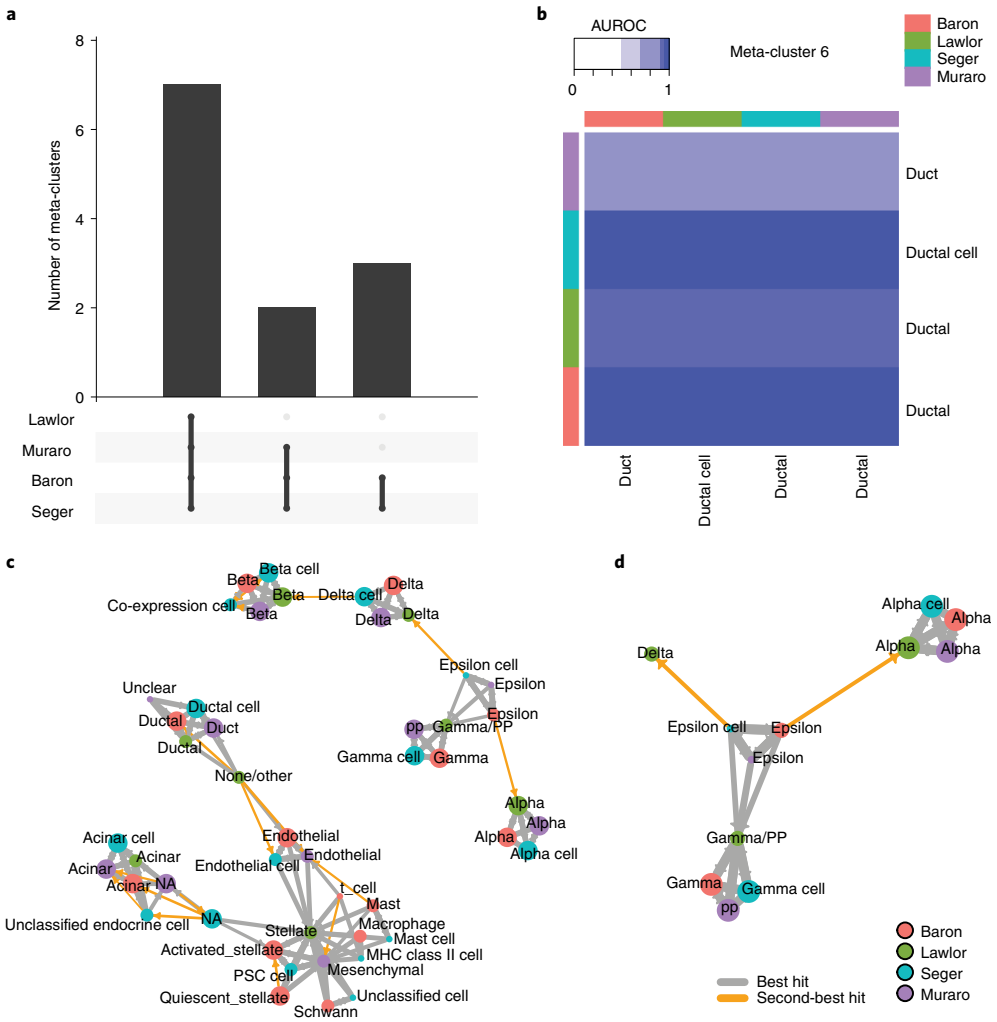


Fig. 5 | Replicating cell types can be extracted as meta-clusters. a, The Upset plot breaks down cell-type replicability by dataset. Meta-clusters (groups of replicating cell types) are organized according to the datasets in which they replicate. For example, there are two cell types that replicate in the Baron, Muraro and Seger datasets but are missing in the Lawlor dataset. **b**, 'Cell type badges' help identify datasets where cell-type replicability is weaker. 1-vs-best AUROC heatmap for meta-cluster corresponding to ductal cells. The cell type is detected across all four datasets, but AUROCs are systematically weaker when testing in the Muraro dataset, indicating that the cell type is not as clearly defined in that dataset. **c**, The cluster graph enables the rapid visualization of replicating cell types. Each node of the graph represents a cell type, colored by dataset of origin. Best hits (strong 1-vs-best AUROC) are shown by gray directed edges (oriented from reference cell type toward target cell type). Outgroups are shown by orange directed edges (reference toward target) for 1-vs-best AUROC >0.3. Ideally replicating cell types form cliques (every pair of a cell type is connected, e.g., alpha cells). **d**, Subsetting the cluster graph enables the investigation of close calls. Same representation as c, centered on the 'epsilon' cell type from the Baron dataset, which had two close matches in the Lawlor dataset ('Alpha' and 'Gamma/PP'), because the epsilon cell type is missing in the Lawlor dataset.

Procedure 2: assessing cell-type replicability against a pre-trained reference taxonomy

▲ **CRITICAL** Procedure 2 demonstrates how to assess cell types of a newly annotated dataset against a reference cell-type taxonomy. Pre-training a MetaNeighbor model provides a rigorous, fast and simple way to query a large reference dataset and obtain quantitative estimations of the replicability of newly annotated clusters. In Procedure 1, all datasets needed to be loaded simultaneously, which may be prohibitive when large datasets are involved. Pre-training a model enables the loading of large datasets only once, when the pre-trained model is generated. The pre-trained model requires only a small amount of memory, which makes it easy to share and query, particularly for large atlas taxonomies.

In this procedure, we consider the cell-type taxonomy established by the BICCN in the mouse primary motor cortex. The BICCN taxonomy was defined across a compendium of datasets sampling across multiple modalities (transcriptomics and epigenomics); it constitutes one of the richest neuronal resources currently available. When matching against a reference taxonomy, we assume that the reference is of higher resolution than the query dataset; that is, the query dataset samples the same set or a subset of cells compared to the reference.

Pre-train a reference MetaNeighbor model ● Timing 1–5 min

1 Start by loading an already merged SCE object containing the BICCN dataset. The full code for generating the dataset is available on GitHub[30]; the dataset can be downloaded directly on FigShare[31].

```
library(SingleCellExperiment)
biccn_data = readRDS("full_biccn_hvg.rds")
dim(biccn_data)
## [1] 319 482712
colnames(colData(biccn_data))
## [1] "sample_id" "cluster_id" "cluster_label"
## [4] "subclass_label" "class_label" "cluster_color"
## [7] "size" "passed_qc" "joint_cluster_id"
## [10] "joint_cluster_label" "joint_cluster_color" "joint_subclass_id"
## [13] "joint_subclass_label" "joint_subclass_color" "joint_class_id"
## [16] "joint_class_label" "joint_class_color" "joint_cl"
## [19] "joint_cluster_size" "joint_tree_order" "study_id"
table(biccn_data$study_id)
##
## scCv2 scCv3 scSS snCv2 snCv3M snCv3Z snSS
## 122641 71183 6288 76525 159738 40166 6171
```

The BICCN data contains seven datasets totaling 482,712 cells. There are multiple sets of cell-type labels depending on resolution (class, subclass and cluster) or type of labels (independent labels or labels defined from joint clustering). Note that, to reduce memory usage, we already computed and restricted the dataset to a set of 319 highly variable genes.

2 Create pre-trained models with the `trainModel` function, which has identical parameters as the `MetaNeighborUS` function used in Procedure 1. Here, we chose to focus on two sets of cell types: subclasses from the joint clustering (medium resolution; e.g., *Vip* interneurons and L2/3 intratelencephalic (IT) excitatory neurons) and clusters from the joint clustering (high resolution; e.g., Chandelier cells). Create and store pre-trained models at the subclass level, then at the cluster level:

```
library(MetaNeighbor)
pretrained_model = MetaNeighbor::trainModel(
  var_genes = rownames(biccn_data),
  dat = biccn_data,
  study_id = biccn_data$study_id,
  cell_type = biccn_data$joint_subclass_label
)
write.table(pretrained_model, "pretrained_biccn_subclasses.txt")
pretrained_model = MetaNeighbor::trainModel(
  var_genes = rownames(biccn_data),
  dat = biccn_data,
  study_id = biccn_data$study_id,
  cell_type = biccn_data$joint_cluster_label
)
write.table(pretrained_model, "pretrained_biccn_clusters.txt")
```

For simplicity of use, we store the pretrained models to file by using the `write.table` function.

PAUSE POINT The remainder of the procedure is independent and can be run in a new R session.

Compare annotations to pre-trained taxonomy ● Timing 1 min

3 Start by loading the query dataset (neurons from mouse primary visual cortex[32], available in the scRNAseq package) and the pre-trained subclass and cluster taxonomies:

```
library(scRNAseq)
tasic = TasicBrainData(ensembl = FALSE)
tasic$study_id = "tasic"
biccn_subclasses = read.table("pretrained_biccn_subclasses.txt",
  check.names = FALSE)
biccn_clusters = read.table("pretrained_biccn_clusters.txt",
  check.names = FALSE)
```

We add a 'study_id' column to the Tasic et al.[32] metadata, because this information will be needed later by MetaNeighbor. Note the 'check.names = FALSE' argument when reading a pre-trained model, which is required to preserve the correct formatting of MetaNeighbor cell-type names.

4 To run MetaNeighbor, we use the `MetaNeighborUS` function but, compared to Procedure 1, we provide a pre-trained model instead of a set of highly variable genes (which are already contained in the pre-trained model). Start by checking whether the Tasic et al.[32] cell types are consistent with the BICCN subclass resolution:

```
library(MetaNeighbor)
aurocs = MetaNeighborUS(
  trained_model = biccn_subclasses, dat = tasic,
  study_id = tasic$study_id, cell_type = tasic$primary_type,
  fast_version = TRUE
)
```

? TROUBLESHOOTING

5 Visualize AUROCs as a rectangular heatmap, with the reference taxonomy cell types as columns and query cell types as rows (Fig. 6a):

```
plotHeatmapPretrained(aurocs)
```

As in Procedure 1, we start by looking for evidence of global structure in the dataset. Here, we recognize three red blocks, which correspond to non-neurons (top left), inhibitory neurons (middle) and excitatory neurons (bottom right). The presence of sub-blocks inside the three global blocks suggests that cell types can be matched more finely. For example, inside the inhibitory block, we can recognize sub-blocks corresponding to caudal ganglionic eminence (CGE)–derived interneurons (*Vip*, *Sncg* and *Lamp5* in the BICCN taxonomy) and medial ganglionic eminence (MGE)–derived interneurons (*Pvalb* and *Sst* in the BICCN taxonomy).

6 Refine AUROCs by focusing on inhibitory neurons by using the `splitTrainClusters` and `splitTestClusters` utility functions to select the relevant cell types:

```
gabaergic_tasic = splitTestClusters(aurocs, k = 4)[[2]]
gabaergic_biccn = splitTrainClusters(aurocs[gabaergic_tasic,], k = 4)
[[4]]
full_label = makeClusterName(tasic$study_id, tasic$primary_type)
tasic_subdata = tasic[, full_label %in% gabaergic_tasic]
aurocs = MetaNeighborUS(
  trained_model = biccn_subclasses[, gabaergic_biccn],
  dat = tasic_subdata, study_id = tasic_subdata$study_id,
  cell_type = tasic_subdata$primary_type, fast_version = TRUE
)
plotHeatmapPretrained(aurocs, cex = 0.7)
```

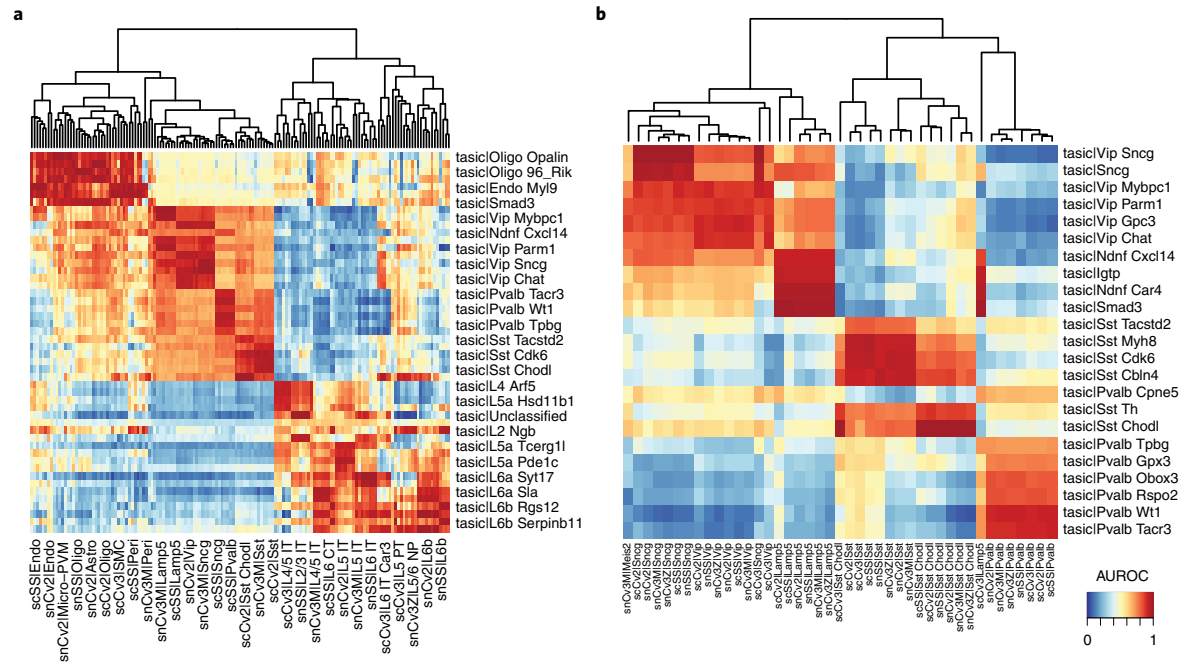


Fig. 6 | Assessment of cell-type annotations from the mouse primary visual cortex against reference neuron taxonomy from the primary motor cortex (medium resolution). a, Heatmap based on MetaNeighbor AUROCs. Reference cell types are shown as columns, and query cell types are shown as rows. Reference cell types are grouped by hierarchical clustering, and query cell types are grouped according to the strongest-matching reference cell type. **b**, Assessment of inhibitory cell types from the mouse primary visual cortex against reference inhibitory cell types (medium resolution). Same representation as **a**. Red rectangles indicate groups of related cell types: *Sncg*, *Vip*, *Lamp5*, *Sst* and *Pvalb* inhibitory neurons.

The heatmap (Fig. 6b) suggests that there is a broad agreement at the subclass level between the BICCN MOp taxonomy and the Tasic et al.[32] dataset. For example, the *Ndnf* subtypes, *Igtp* and *Smad3* cell types from the Tasic et al.[32] dataset match with the BICCN *Lamp5* subclass.

? TROUBLESHOOTING

7 The previous heatmaps suggest that all Tasic et al.[32] cell types can be matched with one BICCN subclass. We now go one step further and ask whether inhibitory cell types correspond to one of the BICCN clusters. Compute and visualize cell-type similarity:

```
aurocs = MetaNeighborUS(trained_model = biccn_clusters,
  dat = tasic_subdata,
  study_id = tasic_subdata$study_id,
  cell_type = tasic_subdata$primary_type,
  fast_version = TRUE)
plotHeatmapPretrained(aurocs, cex = 0.7)
```

Here, the heatmap is difficult to interpret because of the large number of BICCN cell types (Fig. 7a). Instead, investigate the top hits for each query cell type directly:

```
head(sort(aurocs["tasic|Sst Chodl",], decreasing = TRUE), 10)
## scCv2|Sst Chodl scCv3|Sst Chodl scSS|Sst Chodl snCv2|Sst Chodl
## 1.0000000 1.0000000 1.0000000 1.0000000
## snCv3M|Sst Chodl snCv3Z|Sst Chodl snSS|Sst Chodl scCv3|L6b Ror1
## 1.0000000 1.0000000 1.0000000 0.9960366
## scSS|L6b Ror1 snCv3M|L6b Ror1
## 0.9947832 0.9944783
head(sort(aurocs["tasic|Pvalb Cpne5",], decreasing = TRUE), 10)
## snCv2|Pvalb Vipr2_2 scCv2|Pvalb Vipr2_2 scSS|Pvalb Vipr2_2
```

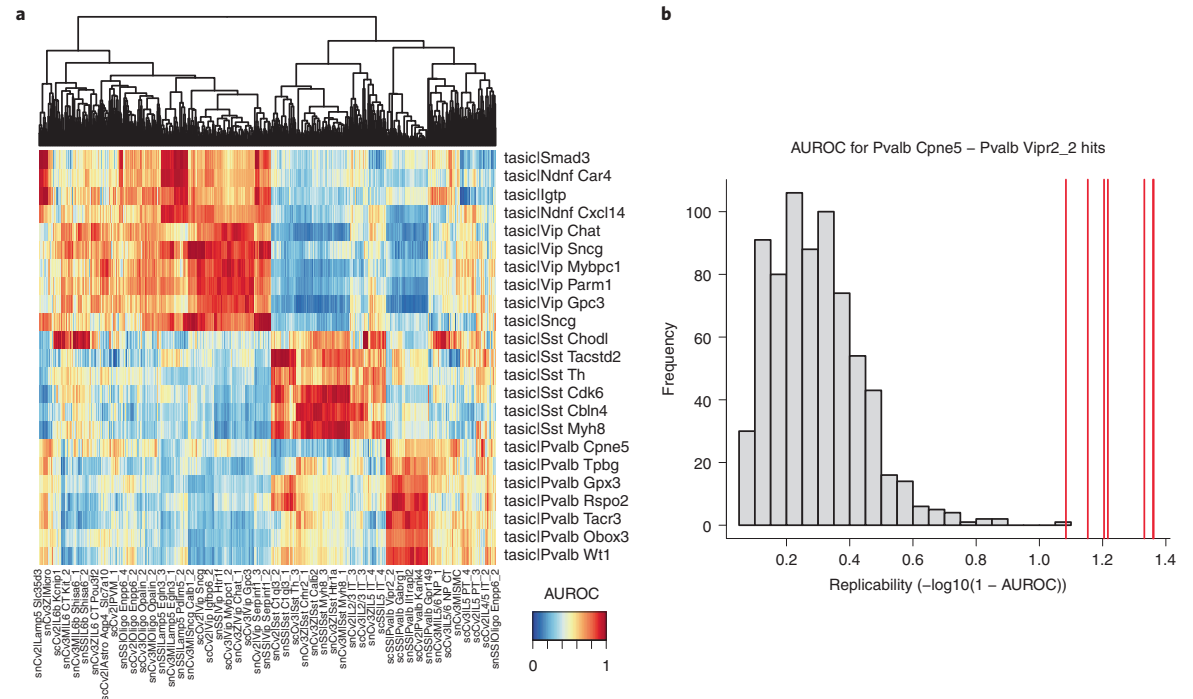


Fig. 7 | Assessment of inhibitory cell types from the mouse primary visual cortex against reference inhibitory cell types (high resolution). a, Heatmap based on MetaNeighbor AUROCs. Reference cell types are shown as columns, and query cell types are shown as rows. Global red rectangles indicate good replicability structure, suggesting replicability for *Sncg*, *Vip*, *Lamp5*, *Sst* and *Pvalb* inhibitory subtypes. **b**, Distribution of AUROC scores for the 'Pvalb Cpne5' cell type from the primary visual cortex (query cell type) against all reference cell types. Best hits (against the 'Pvalb Vipr2_2') are shown by red lines, and all other hits are shown as a gray background distribution. Replicating cell types have substantially higher AUROC scores than background cell types.

```
## 0.9564926 0.9563014 0.9534328
## snCv3Z|Pvalb Vipr2_2 snSS|Pvalb Vipr2_2 scCv3|Pvalb Vipr2_2
## 0.9392809 0.9375598 0.9297189
## snCv3Z|L4/5 IT_2 snCv3M|Pvalb Vipr2_2 scCv2|L4/5 IT_2
## 0.9177663 0.9175751 0.8719640
## snCv2|L4/5 IT_2
## 0.8676611
```

We note two properties of matching against a pre-trained reference. First, replicable cell types have a clear top match in each of the reference datasets. *Sst Chodl* (long-projecting interneurons) match to similarly named clusters in the BICCN with an AUROC >0.9999, *Pvalb Cpne5* (Chandelier cells) match with the *Pvalb Vipr2_2* cluster with AUROC >0.93. Second, we have to beware of false positives. For example, *Sst Chodl* secondarily matches with the L6b *Ror1* cell types with AUROC >0.98, an excitatory cell type only distantly related with long-projecting interneurons. When we use the pre-trained model, we compute AUROCs only with the BICCN data as the reference data; thus, we cannot identify reciprocal hits. If we had been able to use 'Tasic|Sst Chodl' as the reference cluster, its votes would have gone heavily in favor of the BICCN's *Sst Chodl*, making L6b *Ror1* a low AUROC match on average. Because of the low dimensionality of gene expression space, we expect false-positive hits to occur just by chance (e.g., cell types reusing similar pathways) when a cell type is missing in the query dataset. Here, L6b *Ror1* (an excitatory type) had no natural match with the Tasic et al.[32] inhibitory cell types and voted for its closest match, long-projecting interneurons.

There are three alternatives to separate true hits from false-positive hits. First, if a cell type is highly replicable, it will have a clear top-matching cluster in the reference dataset. Second, if the query dataset is known to be a particular subset of the reference dataset (e.g., inhibitory neurons, as is the case here), we recommend restricting the reference taxonomy to that subset. Third, if the first

two solutions do not yield clear results or cannot be performed, it is possible to go back to reciprocal testing by using the full BICCN dataset instead of the pre-trained reference.

? TROUBLESHOOTING

8 We illustrate the first solution in the case of Chandelier cells (Fig. 7b). Visualize the strength of the best hits by running the following:

```
chandelier_hits = aurocs["tasic|Pvalb Cpne5",]
is_chandelier = getCellType(names(chandelier_hits)) == "Pvalb Vipr2_2"
hist(-log10(1-chandelier_hits[!is_chandelier]), breaks = 20,
  xlab = "Replicability (-log10(1-AUROC))",
  xlim = range(-log10(1-chandelier_hits)),
  main = "AUROC for Pvalb Cpne5 - Pvalb Vipr2_2 hits")
box(bty = "L")
abline(v = -log10(1-chandelier_hits[is_chandelier]), col = "red")
```

To illustrate AUROC differences, we chose a logarithmic scaling to reflect that AUROC values do not scale linearly: when AUROCs are close to 1, a difference of 0.05 is substantial. Here, the best matching BICCN cluster ('Pvalb Vipr2_2') is orders of magnitude better than other clusters, suggesting very strong replicability.

9 The second solution to avoid false-positive hits is to subset the reference to cell types that reflect the composition of the query datasets. Because we are looking at inhibitory neurons, restrict the BICCN taxonomy to inhibitory cell types, whose names all start with 'Pvalb', 'Sst', 'Lamp5', 'Vip' or 'Sncg':

```
is_gaba = grepl("^(Pvalb|Sst|Lamp5|Vip|Sncg)",
  getCellType(colnames(biccn_clusters)))
biccn_gaba = biccn_clusters[, is_gaba]
aurocs = MetaNeighborUS(trained_model = biccn_gaba,
  dat = tasic_subdata,
  study_id = tasic_subdata$study_id,
  cell_type = tasic_subdata$primary_type,
  fast_version = TRUE)
head(sort(aurocs["tasic|Sst Chodl",], decreasing = TRUE), 10)
## scCv2|Sst Chodl scCv3|Sst Chodl scSS|Sst Chodl snCv2|Sst Chodl
## 1.0000000 1.0000000 1.0000000 1.0000000
## snCv3M|Sst Chodl snCv3Z|Sst Chodl snSS|Sst Chodl snCv2|Sst Th_3
## 1.0000000 1.0000000 1.0000000 0.8965108
## snCv3M|Sst Th_3 snCv3M|Sst Pappa
## 0.8839431 0.8721883
head(sort(aurocs["tasic|Pvalb Cpne5",], decreasing = TRUE), 10)
## snCv3Z|Pvalb Vipr2_2 snCv3M|Pvalb Vipr2_2 snCv2|Pvalb Vipr2_2
## 0.9960796 0.9959839 0.9939759
## snSS|Pvalb Vipr2_2 scSS|Pvalb Vipr2_2 scCv2|Pvalb Vipr2_2
## 0.9939759 0.9895774 0.9893861
## scCv3|Pvalb Vipr2_2 snCv3M|Pvalb Vipr2_1 scSS|Lamp5 Lhx6
## 0.9640467 0.9212086 0.8676611
## scCv3|Sncg Slc17a8
## 0.8668962
```

Again, we note that there is a significant gap between the best hit and the secondary hit, but now secondary hits are closely related cell types (*Sst* subtype for *Sst Chodl* and secondary Chandelier cell type *Pvalb Vipr2_1* for *Pvalb Cpne5*).

? TROUBLESHOOTING

10 To obtain a more stringent mapping between the query cell types and reference cell types, compute one-vs-best AUROC, which will automatically match the best hit against the best secondary hit:

```
best_hits = MetaNeighborUS(trained_model = biccn_gaba,
  dat = tasic_subdata,
```

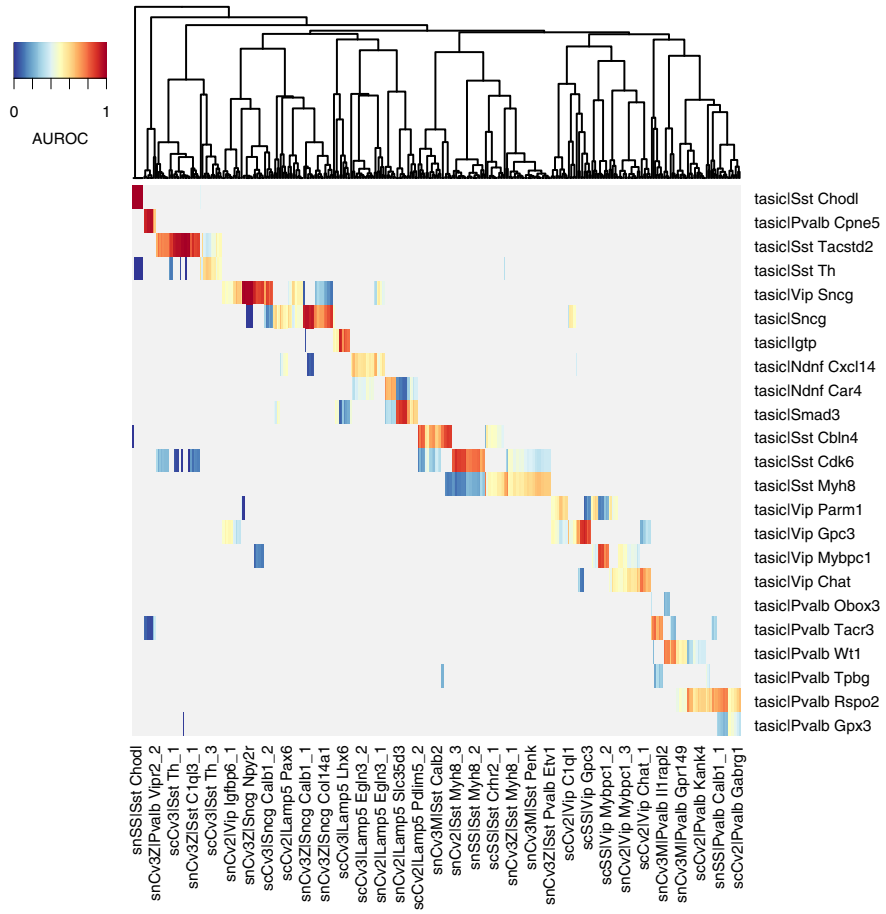


Fig. 8 | 1-vs-best AUROCs enable rapid identification of 1:1 hits and 1:*n* hits. Heatmap based on MetaNeighbor 1-vs-best AUROCs. Reference cell types are shown as columns, and query cell types are shown as rows. In this representation, the best hits are shown in red, the outgroup hit is shown in blue and all other values are gray.

```
  study_id = tasic_subdata$study_id,
  cell_type = tasic_subdata$primary_type,
  one_vs_best = TRUE,
  fast_version = TRUE)
plotHeatmapPretrained(best_hits)
```

Now, the hit structure is much sparser, which helps identify 1:1 and 1:*n* hits (Fig. 8). The heatmap suggests that most Tasic et al.[32] cell types match with one or several BICCN clusters. Inspect the top hits for three cell types from the Tasic dataset:.

```
head(sort(best_hits["tasic|Sst Chodl",], decreasing = TRUE), 10)
## scCv2|Sst Chodl scCv3|Sst Chodl scSS|Sst Chodl snCv2|Sst Chodl
## 1.0000000 1.0000000 1.0000000 1.0000000
## snCv3M|Sst Chodl snCv3Z|Sst Chodl snSS|Sst Chodl snSS|Sst Th_2
## 1.0000000 1.0000000 1.0000000 0.4094994
head(sort(best_hits["tasic|Pvalb Cpne5",], decreasing = TRUE), 10)
## snCv3M|Pvalb Vipr2_2 snCv3Z|Pvalb Vipr2_2 snSS|Pvalb Vipr2_2
## 0.9698189 0.9678068 0.9547284
## snCv2|Pvalb Vipr2_2 scSS|Pvalb Vipr2_2 scCv2|Pvalb Vipr2_2
## 0.9527163 0.9245473 0.9164990
```

```
## scCv3|Pvalb Vipr2_2 snCv3M|Pvalb Vipr2_1
## 0.7444668 0.6348089
head(sort(best_hits["tasic|Sst Tacstd2",], decreasing = TRUE), 10)
## scCv2|Sst C1ql3_1 snCv2|Sst C1ql3_1 snCv3Z|Sst C1ql3_1 snCv3M|Sst
C1ql3_1
## 0.9962406 0.9924812 0.9924812 0.9887218
##  scCv3|Sst  C1ql3_1  scSS|Sst  C1ql3_1  scCv3|Sst  C1ql3_2  scSS|Sst
C1ql3_2
## 0.9852608 0.9812030 0.9661654 0.9661654
## snSS|Sst C1ql3_1 scCv2|Sst C1ql3_2
## 0.9624060 0.9586466
```

Using this more stringent assessment, we confirm that *Sst Chodl* strongly replicates inside the BICCN (one-vs-best AUROC of ~1; best secondary hit = 0.41) and observe the same for *Pvalb Cpne5* (one-vs-best AUROC >0.74; best secondary hit = 0.63), whereas, for example, *Sst Tacstd2* corresponds to multiple BICCN subtypes (including *Sst C1ql3*_1 and *Sst C1ql3*_2; AUROC >0.95). **? TROUBLESHOOTING**

Procedure 3: functional characterization of replicating clusters

▲ **CRITICAL** Procedure 3 demonstrates how to characterize functional gene sets contributing to cell-type identity. Once replicating cell types have been identified with unsupervised MetaNeighbor (as in Procedures 1 and 2), supervised MetaNeighbor enables the functional interpretation of the biology contributing to each cell type's identity. In this procedure, we will focus on the characterization of inhibitory neuron subclasses from the mouse primary cortex as provided by the BICCN. The BICCN has shown that subclasses are strongly replicable across datasets and provided marker genes that are specific to each subclass. MetaNeighbor can be used to further quantify which pathways contribute to the subclasses' unique biological properties.

Creation of biologically relevant gene sets ● Timing 1 min

1 To compute the functional characterization of clusters, we first need an ensemble of gene sets sampling relevant biological pathways. In this procedure, we consider the Gene Ontology (GO) annotations for mouse. The scripts used to build up-to-date gene sets can be found on GitHub[30], and gene sets can be downloaded directly on FigShare[31]. Start by loading the GO sets:

```
go_sets = readRDS("go_mouse.rds")
```

Gene sets are stored as a named list, in which each element of the list corresponds to a gene set and contains a vector of gene symbols.

2 Load the dataset containing inhibitory neurons from the BICCN. The scripts used to build the dataset can be found on GitHub[30], and the dataset can be downloaded on FigShare[31].

```
library(SingleCellExperiment)
biccn_gaba = readRDS("biccn_gaba.rds")
dim(biccn_gaba)
## [1] 24140 71368
```

3 Next, restrict the gene sets to genes that are present in the dataset. Then, filter gene sets to keep gene sets of meaningful size: large enough to learn expression profiles (>10) but small enough to represent specific biological functions or processes (<100):

```
known_genes = rownames(biccn_gaba)
go_sets = lapply(go_sets, function(gene_set) {
  gene_set[gene_set %in% known_genes]
})
min_size = 10
max_size = 100
go_set_size = sapply(go_sets, length)
```

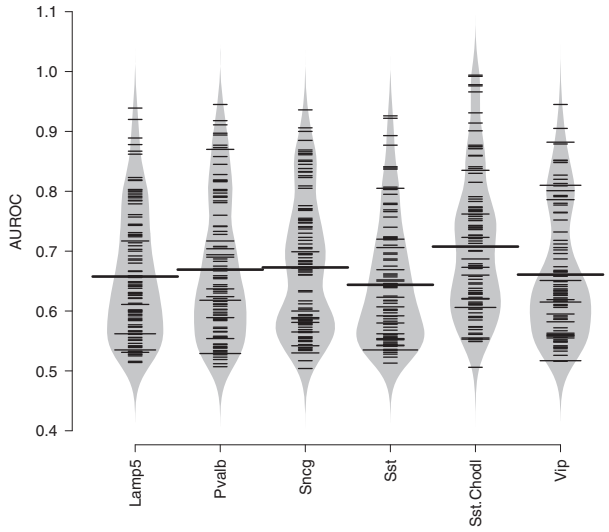


Fig. 9 | A small fraction of functional gene sets contributes highly to cell-type replicability. For each cell type, large ticks represent the average AUROC across gene sets. Each smaller tick represents an individual gene set, and the envelope is a violin-plot-style approximation of the distribution of performance across gene sets.

```
go_sets = go_sets[go_set_size >= min_size &
go_set_size <= max_size]
length(go_sets)
## [1] 6488
```

Functional characterization with supervised MetaNeighbor ● Timing 30–90 min

4 Once the gene set list is ready, run the supervised `MetaNeighbor` function. Its inputs are similar to `MetaNeighborUS`, but it assumes that cell types have already been matched across datasets (i.e., they have identical names). Here, we use joint BICCN subclasses, for which names have been normalized across datasets ('Pvalb', 'Sst', 'Sst Chodl', 'Vip', 'Lamp5' and 'Sncg'). Note that, because we are testing close to 6,500 gene sets, this step is expected to take a long time for large datasets. We recommend using this function inside a script and always saving results to a file as soon as computations are done by using the `write.table` function.

```
library(MetaNeighbor)
aurocs = MetaNeighbor(dat = biccn_gaba,
  experiment_labels = biccn_gaba$study_id,
  celltype_labels = biccn_gaba$joint_subclass_label,
  genesets = go_sets,
  fast_version = TRUE, bplot = FALSE, batch_size = 50)
write.table(aurocs, "functional_aurocs.txt")
```

Later, results can be retrieved with the *read.table* function:

```
aurocs = read.table("functional_aurocs.txt")
```

? TROUBLESHOOTING

5 Use the `plotBPlot` function on the first 100 gene sets to visualize how replicability depends on gene sets (Fig. 9).

```
plotBPlot(head(aurocs, 100))
```

In this representation, large segments represent average gene set performance, and short segments represent the performance of individual gene sets. We note that most gene sets contribute

Table 2 | Top 10 gene sets (with fewer than 100 genes) contributing to cell-type replicability

go_term	Lamp5	Pvalb	Sncg	Sst	Sst.Chodl	Vip	Average	n_genes
GO:0007215|glutamate receptor signaling pathway|BP	0.97	0.98	0.97	0.98	1.00	0.99	0.98	92
GO:0051966|regulation of synaptic transmission, glutamatergic|BP	0.96	0.97	0.98	0.96	0.99	0.97	0.97	75
GO:0060076|excitatory synapse|CC	0.96	0.97	0.99	0.96	0.99	0.96	0.97	75
GO:0033555|multicellular organismal response to stress|BP	0.95	0.98	0.98	0.95	1.00	0.98	0.97	98
GO:0098839|postsynaptic density membrane|CC	0.92	0.97	0.98	0.98	0.98	0.97	0.97	93
GO:0099565|chemical synaptic transmission, postsynaptic|BP	0.97	0.98	0.97	0.95	0.99	0.96	0.97	91
GO:0008306|associative learning|BP	0.97	0.98	0.96	0.96	0.99	0.95	0.97	100
GO:0099601|regulation of neurotransmitter receptor activity|BP	0.96	0.98	0.96	0.95	0.99	0.98	0.97	61
GO:0060079|excitatory postsynaptic potential|BP	0.97	0.98	0.97	0.95	0.99	0.95	0.97	83
GO:0010771|negative regulation of cell morphogenesis involved in differentiation|BP	0.98	0.98	0.97	0.96	0.99	0.92	0.97	98

The 'go_term' column shows the identifier, name and sub-ontology (BP, biological process; CC, cellular component) of the investigated gene set. Columns 'Lamp5' to 'Vip' show the replicability (average AUROC over cross-dataset-validation folds) for each cell-type and gene-set combination. The 'Average' column takes the average across cell types, and 'n_genes' shows the number of genes in the gene set.

moderately to replicability (AUROC of ~0.7), numerous gene sets have a performance close to random (AUROC of ~0.5–0.6), and some gene sets have exceedingly high performance (AUROC >0.8).

6 To focus on gene sets that contribute highly to cell-type specificity, create a summary table containing, for each gene set, cell-type-specific AUROCs, average AUROCs across cell types and gene set size:

```
gs_size = sapply(go_sets, length)
aurocs_df = data.frame(go_term = rownames(aurocs), aurocs)
aurocs_df$average = rowMeans(aurocs)
aurocs_df$n_genes = gs_size[rownames(aurocs)]
```

Then, order gene sets by average AUROC and look at the top-scoring gene sets (Table 2).

```
head(aurocs_df[order(aurocs_df$average, decreasing = TRUE),],10)
```

Without surprise, replicability is mainly driven by gene sets related to neuronal functions that are immediately relevant to the physiology of inhibitory neurons, such as 'glutamate receptor signaling pathway', 'regulation of synaptic transmission, glutamatergic' or 'chemical synaptic transmission, postsynaptic'. Note that most of the top-scoring gene sets have a large number of genes, because larger sets of genes make it easier to learn generalizable expression profiles.

To obtain even more specific biological functions, further filter for gene sets that have <20 genes (Table 3).

```
small_sets = aurocs_df[aurocs_df$n_genes < 20,]
head(small_sets[order(small_sets$average, decreasing = TRUE),],10)
```

Again, the top-scoring gene sets are dominated by biological functions immediately relevant to inhibitory neuron physiology, such as 'ionotropic glutamate receptor signaling pathway', 'positive regulation of synaptic transmission, GABAergic' or 'GABA-A receptor complex'.

7 To understand how individual genes contribute to gene set performance, use the `plotDotPlot` function, which shows the expression of all genes in a gene set of interest, averaged over all datasets (Fig. 10):

```
plotDotPlot(dat = biccn_gaba,
  experiment_labels = biccn_gaba$study_id,
  celltype_labels = biccn_gaba$joint_subclass_label,
```

Table 3 | Top 10 gene sets (with fewer than 20 genes) contributing to cell-type replicability

go_term	Lamp5	Pvalb	Sncg	Sst	Sst.Chodl	Vip	average	n_genes
GO:0004970|ionotropic glutamate receptor activity|MF	0.90	0.92	0.91	0.96	0.97	0.92	0.93	19
GO:0035235|ionotropic glutamate receptor signaling pathway|BP	0.82	0.82	0.91	0.93	0.94	0.87	0.88	16
GO:0032230|positive regulation of synaptic transmission, GABAergic|BP	0.84	0.86	0.82	0.92	0.98	0.83	0.88	16
GO:0007216|G protein-coupled glutamate receptor signaling pathway|BP	0.89	0.85	0.76	0.92	0.95	0.84	0.87	16
GO:1905874|regulation of postsynaptic density organization|BP	0.83	0.86	0.87	0.90	0.92	0.83	0.87	19
GO:0099150|regulation of postsynaptic specialization assembly|BP	0.83	0.89	0.86	0.91	0.91	0.80	0.87	18
GO:0150052|regulation of postsynapse assembly|BP	0.83	0.89	0.86	0.91	0.91	0.80	0.87	18
GO:0021889|olfactory bulb interneuron differentiation|BP	0.81	0.91	0.82	0.88	0.89	0.86	0.86	15
GO:0070679|inositol 1,4,5 trisphosphate binding|MF	0.92	0.94	0.79	0.81	0.86	0.85	0.86	15
GO:1902711|GABA-A receptor complex|CC	0.82	0.87	0.87	0.80	0.99	0.80	0.86	19

Same format as Table 2. MF, molecular function.

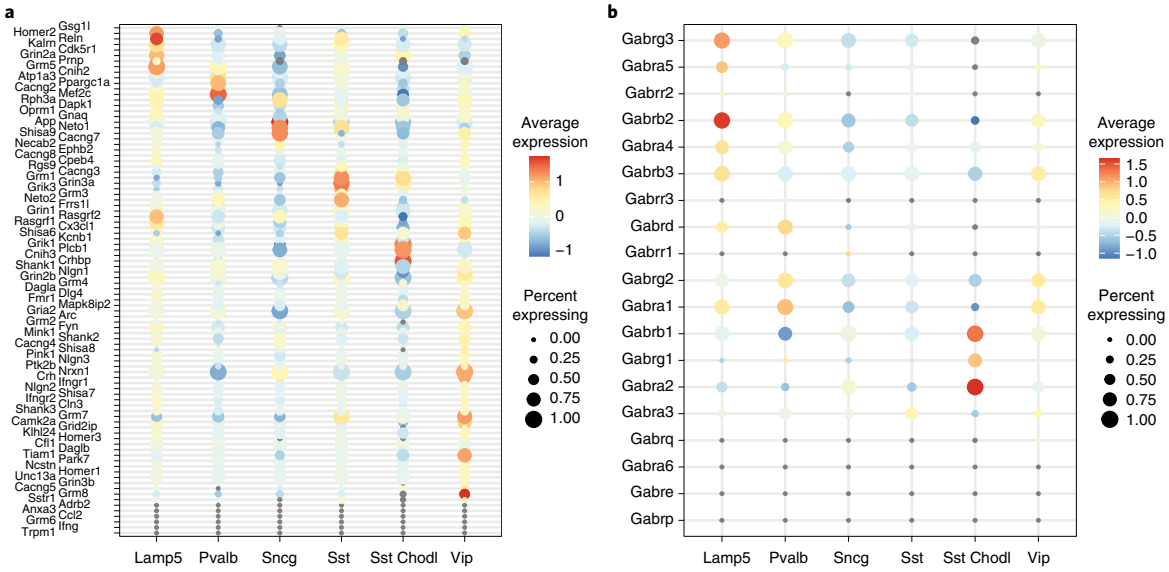


Fig. 10 | Top-scoring gene sets can be broken down into characteristic genes for each cell type. a, Dot plot of genes from the 'Glutamate receptor signalling pathway' GO term, where cell types are shown on the x-axis, and genes are shown on the y-axis. For each cell type, the dot size corresponds to the fraction of cells expressing a given gene, and the color corresponds to the z-scored average expression level, averaged across datasets. **b**, Same as **a**, for the 'GABA-A receptor complex' GO term.

```
  gene_set = go_sets[["GO:0007215|glutamate receptor signaling path-
way|BP"]])
plotDotPlot(dat = biccn_gaba,
  experiment_labels = biccn_gaba$study_id,
  celltype_labels = biccn_gaba$joint_subclass_label,
  gene_set = go_sets[["GO:1902711|GABA-A receptor complex|CC"]])
```

High-scoring gene sets are characterized by the differential usage of genes from a given gene set. For example, when looking at the γ-aminobutyric acid (GABA)-A receptor complex composition, *Lamp5* preferentially uses the *Gabrb2* and *Gabrg3* receptors; *Pvalb*, the *Gabra1* receptor; and *Sst Chodl*, the *Gabra2*, *Gabrb1* and *Gabrg1* receptors (Fig. 10b).

Troubleshooting

Troubleshooting advice can be found in Table 4.

Table 4 | Troubleshooting table

Step	Problem	Possible reason	Solution
Equipment setup	Installation failed: package could not be downloaded	Running command in the notebook fails because user input is expected	Run command directly as R command line instead of notebook
9, 15 and 17 (Procedure 1)	`variableGenes` returns 'Cholmod error "problem too large"'	Matrix is too large to be properly handled	Downsample datasets with 'downsampling_size' parameter or manually downsample datasets
10, 16, 17 and 18 (Procedure 1); 4, 6, 7, 9 and 10 (Procedure 2); and 4 (Procedure 3)	`MetaNeighbor` returns 'Cholmod error "problem too large"'	One of the datasets is too large to be properly handled	Use smaller gene sets, downsample largest dataset or run on batches of datasets; then, combine AUROC matrices
	`MetaNeighbor` returns 'Error: cannot allocate vector of size XXX Gb'	Legacy MetaNeighbor was used on a large dataset (>10,000 cells)	Use 'fast_version = TRUE'. If this does not solve the problem, see above
	`MetaNeighbor` returns rows or columns that contain only NAs	One dataset contains only one cell type	This is expected behavior (no outgroup against which to compare). Ignore NAs, use 'symmetric_output = FALSE' or make sure to keep at least two cell types when subsetting datasets
11, 16, 17 and 18 (Procedure 1)	`plotHeatmap` returns 'Error in M + t(M): non-conformable arrays'	`plotHeatmap` has been applied to non-square AUROC matrix, probably because MetaNeighbor was run on a pre-trained model	Use *plotHeatmapPretrained* instead
	In Rstudio, `plotHeatmap` causes 'Error in plot.new(): figure margins too large'	The default Rstudio resolution is too low to correctly display the heatmap	In the code block options, increase 'fig.width' and 'fig.height' until resolved

Timing

The expected timing for the procedures is as follows:
Procedure 1: 3–4 min; Procedure 2: 2–6 min
Procedure 3: 30–90 min

The first two procedures (assessment of cell-type replicability) can be run interactively: once the data are loaded, every call to MetaNeighbor returns within seconds, which enables looking at the data in different ways (e.g., by zooming in on different parts of the dataset). In contrast, Procedure 3 (functional characterization of replicability) is intended to be run as a script, allowing testing of thousands of gene sets and analyzing results within 1 d.

Note that the exact timing depends on the hardware and software used, notably the amount of memory and the BLAS (Basic Linear Algebra Subprograms) library used. MetaNeighbor relies heavily on matrix operations, leading to large speed-ups when using the Intel Math Kernel Library BLAS or openBLAS instead of R's native BLAS library.

Anticipated results

Because MetaNeighbor is nonparametric, there is no fine-tuning to be done for any of the procedures presented here. Over time, we have identified two sources of potential error: bad highly variable gene selection and coding or formatting errors, which can be easily diagnosed by looking at AUROC heatmaps. As a rule of thumb, we expect AUROCs to correctly represent global relationships between cell types, contain replicable cell types (dark red squares or rectangles on heatmaps) and generalize across studies. In the examples below, we illustrate the most common places where these errors are found by presenting a side-by-side comparison of correct and problematic code.

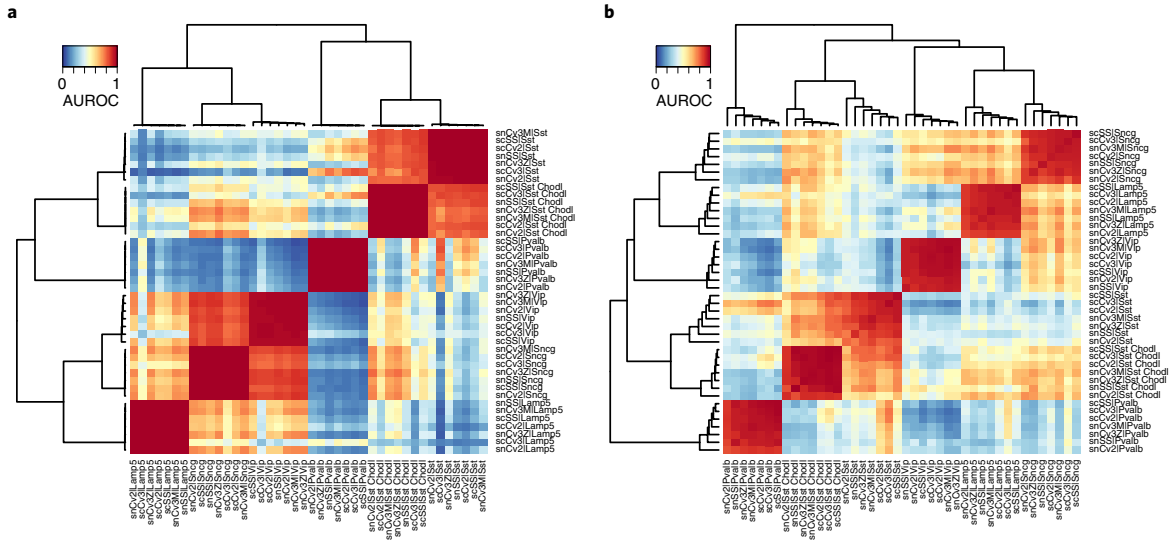


Fig. 11 | Selection of a bad highly variable gene set leads to suboptimal performance and obscures biological signal. a, Anticipated result: AUROC heatmap based on a set of highly variable genes selected by MetaNeighbor. The heatmap has clear replicating clusters (dark red squares) and known secondary biological relationships (e.g., similarity of CGE-derived interneurons *Vip, Sncg* and *Lamp5*). **b**, Possible issue: AUROC heatmap based on a set of random genes (same number of genes as the correctly selected highly variable gene set in **a**). Replicability patterns become weaker: lower performance, gradients within replicating cell types and weaker secondary relationships.

Bad gene set selection

The most common problem is to forget to select a set of highly variable genes, which is expected to dampen the impact of technical variability on neighbor voting (Procedure 1, Steps 9–11). First, we present an example of a correct analysis, where we load the BICCN GABAergic neurons, select highly variable genes and compute cluster similarities (see Procedure 1 for more details).

```
library(MetaNeighbor)
biccn_data = readRDS("biccn_gaba.rds")
biccn_hvgs = variableGenes(biccn_data, exp_labels = biccn_data
$study_id)
# GOOD
aurocs = MetaNeighborUS(var_genes = biccn_hvgs,
  dat = biccn_data,
  study_id = biccn_data$study_id,
  cell_type = biccn_data$joint_subclass_label,
  fast_version = TRUE)
plotHeatmap(aurocs, cex = 0.5)
# BAD
random_genes = sample(rownames(biccn_data), length(biccn_hvgs))
aurocs = MetaNeighborUS(var_genes = random_genes,
  dat = biccn_data,
  study_id = biccn_data$study_id,
  cell_type = biccn_data$joint_subclass_label,
  fast_version = TRUE)
plotHeatmap(aurocs, cex = 0.5)
```

We recognize strong replicability structure, evidenced by the presence of dark red blocks (Fig. 11a). When we repeat the analysis with random genes, the replicability structure is still present, but we recognize two signatures of bad gene set selection: (i) AUROCs are low overall (shift to light red and orange) and (ii) within red blocks, there is a clear gradient structure (Fig. 11b). In our experience, there are three scenarios that lead to bad gene selection: errors in gene symbol conversion,

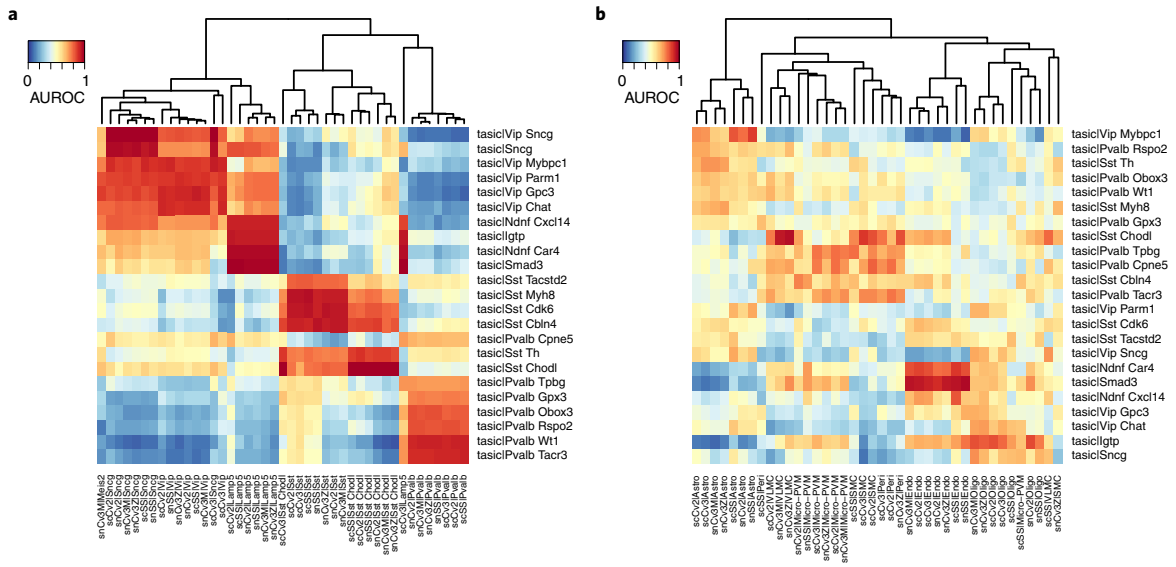


Fig. 12 | Absence of biological overlap between datasets leads to almost random performance and lack of hierarchical cell-type structure. **a**, Anticipated result: AUROC heatmap with inhibitory neuron cell types as query (rows) and inhibitory neuron cell types as reference (columns). **b**, Possible issue: same as **a**, but with non-neuronal cell types as reference (columns). The heatmap lacks clear replicating clusters (dark red rectangles) and known secondary biological relationships (e.g., similarity of CGE-derived interneurons *Vip* and *Sncg* on the query side).

errors when genes are stored as factors in R (that are implicitly converted to numerical values during indexing) and forgetting to select highly variable genes altogether.

No overlap between datasets

The second problem occurs when there is no overlap between datasets, which can be detected in Procedure 1 at Step 11 or Procedure 2 at Steps 5–7. We illustrate this problem with the data from Procedure 2, where we expect all cell types from the Tasic et al.[32] dataset to be present in the pre-trained BICCN model. According to our expectations, all cell types have strong hits with BICCN clusters, and we see a hierarchical structure that is consistent with prior biological knowledge: lighter red blocks corresponding to MGE- and CGE-derived inhibitory neurons (Fig. 12a). We compare with the same block of code, where we 'mistakenly' keep non-neurons from the BICCN taxonomy instead of inhibitory neurons. The lack of biological overlap can be deduced from three factors (Fig. 12b): (i) low AUROC values overall, (ii) almost no strong hits (contrary to expectations) and (iii) lack of expected hierarchical structure (MGE- and CGE-derived inhibitory neurons).

```
library(scRNAseq)
tasic = TasicBrainData(ensembl = FALSE)
tasic$study_id = "tasic"
biccn_subclasses = read.table("pretrained_biccn_subclasses.txt",
check.names = FALSE)
global_aurocs = MetaNeighborUS(
  trained_model = biccn_subclasses, dat = tasic,
  study_id = tasic$study_id, cell_type = tasic$primary_type,
  fast_version = TRUE
)
gabaergic_tasic = splitTestClusters(global_aurocs, k = 4)[[2]]
# GOOD
gabaergic_biccn = splitTrainClusters(global_aurocs[gabaergic_tasic,],
  k = 4)[[4]]
full_labels = makeClusterName(tasic$study_id, tasic$primary_type)
tasic_subdata = tasic[, full_labels %in% gabaergic_tasic]
```

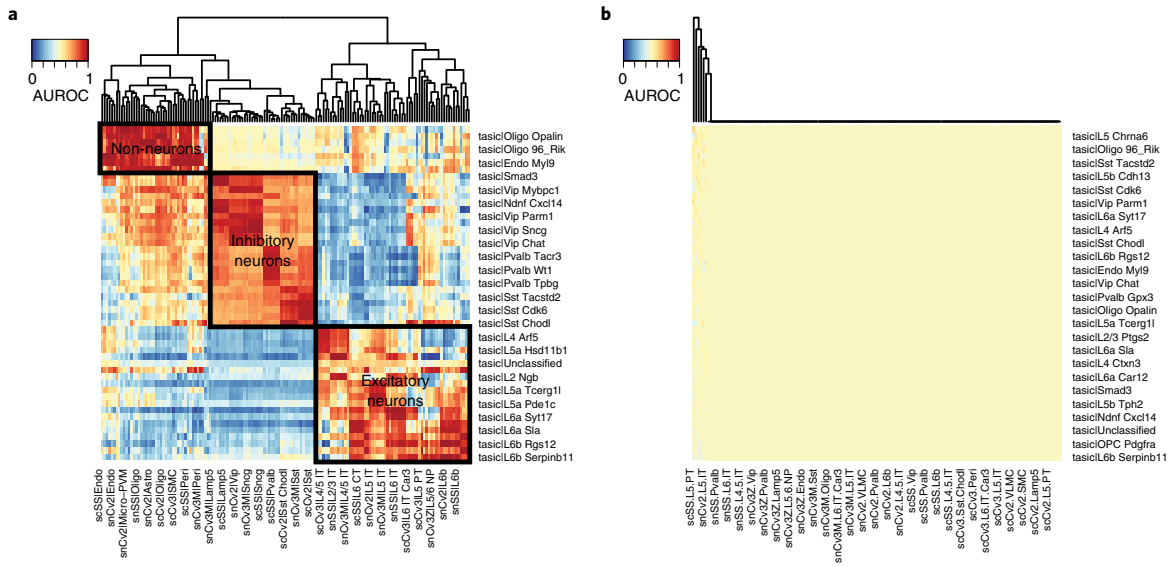


Fig. 13 | Disrupting formatting of cell type names in pre-trained models leads to random performance. a, Anticipated result: AUROC heatmap with cell types from primary visual cortex as query (rows) and cell types from primary motor cortex as reference (columns). The heatmap shows evidence of replicating cell types (dark red rectangles) and global structure (larger rectangles corresponding to non-neurons, excitatory neurons and inhibitory neurons). **b**, Possible issue: same as **a**, but with incorrect formatting of reference cell types (because of an error while reading the pre-trained model), leading to completely random performance.

```
aurocs = MetaNeighborUS(
  trained_model = biccn_subclasses[, gabaergic_biccn],
  dat = tasic_subdata, study_id = tasic_subdata$study_id,
  cell_type = tasic_subdata$primary_type, fast_version = TRUE
)
plotHeatmapPretrained(aurocs, cex = 0.7)
# BAD: non-neurons instead of GABAergic neurons
gabaergic_biccn = splitTrainClusters(global_aurocs, k = 5)[[1]]
full_labels = makeClusterName(tasic$study_id, tasic$primary_type)
tasic_subdata = tasic[, full_labels %in% gabaergic_tasic]
aurocs = MetaNeighborUS(
  trained_model = biccn_subclasses[, gabaergic_biccn],
  dat = tasic_subdata, study_id = tasic_subdata$study_id,
  cell_type = tasic_subdata$primary_type, fast_version = TRUE
)
plotHeatmapPretrained(aurocs, cex = 0.7)
```

Pretrained MetaNeighbor: bad name formatting

The third problem we have encountered is a mistake that occurs when loading a pre-trained model in Step 3 of Procedure 2 and forgetting to specify 'check.names = FALSE', which is essential to preserve correct formatting of cell-type names. Below, we present an example of correct code based on data from Procedure 2. We obtain the expected replicability structure, with evidence of strong hits across all cell types (Fig. 13a; see Procedure 2 for further details and analyses). When we forget 'check.names = FALSE', MetaNeighbor is unable to correctly recognize dataset names and cell-type names in the pre-trained model, and the similarity computations become meaningless, leading to AUROC values that are essentially 0.5 (Fig. 13b). This problem is easy to diagnose and fix but can be very confusing when it occurs.

```
# GOOD
biccn_subclasses = read.table("pretrained_biccn_subclasses.txt",
  check.names = FALSE)
```

```
aurocs = MetaNeighborUS(
  trained_model = biccn_subclasses, dat = tasic,
  study_id = tasic$study_id, cell_type = tasic$primary_type,
  fast_version = TRUE
)
plotHeatmapPretrained(aurocs)
# BAD
biccn_subclasses = read.table("pretrained_biccn_subclasses.txt")
aurocs = MetaNeighborUS(
  trained_model = biccn_subclasses, dat = tasic,
  study_id = tasic$study_id, cell_type = tasic$primary_type,
  fast_version = TRUE
)
plotHeatmapPretrained(aurocs)
```

Impact of batch effects on cell-type matching

The voting scheme used by MetaNeighbor is naturally robust to batch effects, because it relies on identifying nearest neighbors (which are approximately conserved in the presence of batch effects) rather than transcriptional similarity. Because cell-type matching is determined on the basis of reciprocal best hits (similar to BLAST (basic local alignment search tool) for the similarity between biological sequences), we expect MetaNeighbor results to be robust to a large range of batch effects and recommend using MetaNeighbor on unaligned datasets to obtain more accurate replicability values. As we show here, batch effects mainly affect the range of AUROC values and should be considered when interpreting heatmaps (Procedure 1, Step 11 and Procedure 2, Steps 5–7) and replicability strength (Procedure 1, Step 12; Procedure 2, Steps 8–10; and Procedure 3, Steps 5 and 6).

To illustrate the expected drop in AUROC with data quality, we simulated two types of batch effects in the pancreas compendium presented in the protocol: lower sensitivity and higher noise level. To simulate low sensitivity, we downsampled counts in endocrine cells of the Baron dataset and recorded the AUROCs of best hits in the three remaining datasets. AUROCs progressively declined, dropping below 0.9 around 250 unique molecular identifiers (UMIs) per cell (Fig. 14a) and stabilizing around 0.8 for nearly-empty cells. Reciprocal top hits remained perfectly conserved, except for epsilon cells (the rarest cell type), where performance started to degrade around 100 UMIs per cell, which represents exceptionally low sensitivity (Fig. 14b). In the second batch effect simulation, we subset the Baron dataset to highly variable genes, then added Gaussian noise with mean 0 and standard deviation $f \times$ average UMIs per cell, where f is the 'fraction' of noise. We observed a similar pattern to downsampling experiments: AUROC progressively declined, dropping below 0.9 when the noise level reached ~10% of the average count value and progressively declined toward 0.8 (Fig. 14c). Again, reciprocal top hits were perfectly conserved, with a slight degradation for epsilon cells beyond 25% noise (Fig. 14d). In the original MetaNeighbor publication, we further showed that AUROCs are robust to cell-type rarity and the presence of closely related cell types[15].

In practice, we found that our AUROC guidelines (AUROC >0.9 and 1-vs-best AUROC >0.7) held on datasets that spanned a wide range of quality and batch effects and should thus apply to most recently generated single-cell datasets. For example, the BICCN datasets used in this protocol include multiple types of batch effects, because it uses a large array of sequencing protocols[4]: differences in sensitivity (2,000–6,000 detected genes per cell), differences in cell-type composition (L5 pyramidal tract (PT) cells survive better in single-nucleus protocols) and systematic differences in expression profiles (PCR-amplification bias for Smart-Seq and higher expression of nuclear genes for nuclei protocols). However, if one of the datasets is known to be particularly noisy or low quality, the AUROCs for this dataset can be expected to be lower than the guidelines suggested in this article, but we recommend using AUROC >0.8 and 1-vs-best AUROC >0.6 as a minimum.

Multimodal analyses

MetaNeighbor can be applied to multimodal analyses but requires a gene-by-cell matrix for all modalities (all steps remain identical to the protocol presented here). In particular, modalities such as chromatin accessibility and methylation data require a mapping of peaks or reads to individual genes. This mapping is currently unclear, as many peaks are related to regulatory elements such as

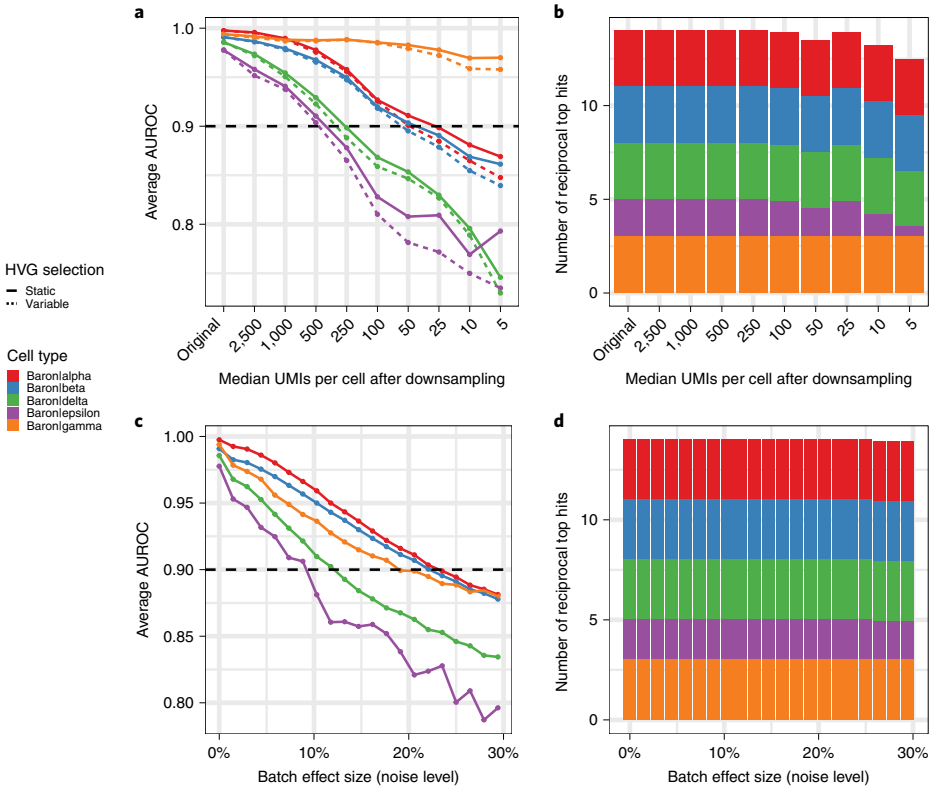


Fig. 14 | MetaNeighbor results are robust to batch effects. a, Replicability (MetaNeighbor AUROC) of endocrine cell types in the Baron pancreas dataset after downsampling the number of UMIs per cell. 'Original' corresponds to the replicability in the original dataset, without downsampling (~5,000 UMIs per cell). Line types represent the highly variable gene (HVG) selection strategy: full lines indicate that the initial set of HVG (based on the full dataset) is conserved ('static'), and dashed lines indicate that HVG are re-picked after downsampling ('variable'). **b**, Stacked barplot showing the number of reciprocal top hits for each endocrine cell type after downsampling. The height of the bars indicates the number of datasets in which the cell type was found to replicate. **c**, Replicability (MetaNeighbor AUROC) of endocrine cell types in the Baron pancreas dataset after the addition of noise to original counts. **d**, Stacked barplot showing the number of reciprocal top hits for each endocrine cell type after the addition of noise. In all panels, statistics are averaged over 10 independent experiments, and colors represent cell types.

enhancers and cannot be attributed to individual genes, resulting in an important loss of signal. As discussed in the previous section, such losses can be seen as 'batch effects' and result in lower AUROC values in some modalities (Procedure 1, Steps 11 and 12; Procedure 2, Steps 5–10; and Procedure 3, Steps 5 and 6).

We illustrate the results of a multimodal analysis in the BICCN data for the mouse primary motor cortex[4]. The full multimodal data include the seven singe-cell and single-nucleus RNA-sequencing (scRNAseq) datasets presented in the protocol, a single-nucleus assay for transposase-accessible chromatin using a sequencing dataset (ATAC-seq, 'atac' in the figure) and a single-nucleus methylation dataset ('snmc'). The initial analysis reveals a difference in resolution between modalities: although there is a single cell type for L2/3 IT and L5 IT excitatory neurons in the scRNAseq datasets (at the 'subclass' annotation level), there are multiple matching cell types in the ATAC-seq and methylation annotations (Fig. 15a). The presence of clear red blocks (high AUROC with primary match and lower AUROC with secondary match) suggests that, for example, L23.a, L23.b and L23.c in ATAC-seq all correspond to the L2/3 IT type in scRNAseq. After merging L2/3 IT and L5 IT cell types under a single annotation, we find an excellent mapping between all modalities, resulting in a clear separation of individual cell types (Fig. 15b). Almost all extracted meta-clusters span all nine datasets (Fig. 15c), with only a handful of cell types missing in one of the other modalities, such as L6b (missing in ATAC-seq), non-neuron subclasses (unannotated in the methylation data) or L6 IT Car3 (missing in several datasets). All modalities share the same range of reciprocal top hits (Fig. 15d), suggesting that the same cell types have been successfully identified in all datasets. However, AUROC values are significantly lower in the ATAC-seq data (Fig. 15e), suggesting that

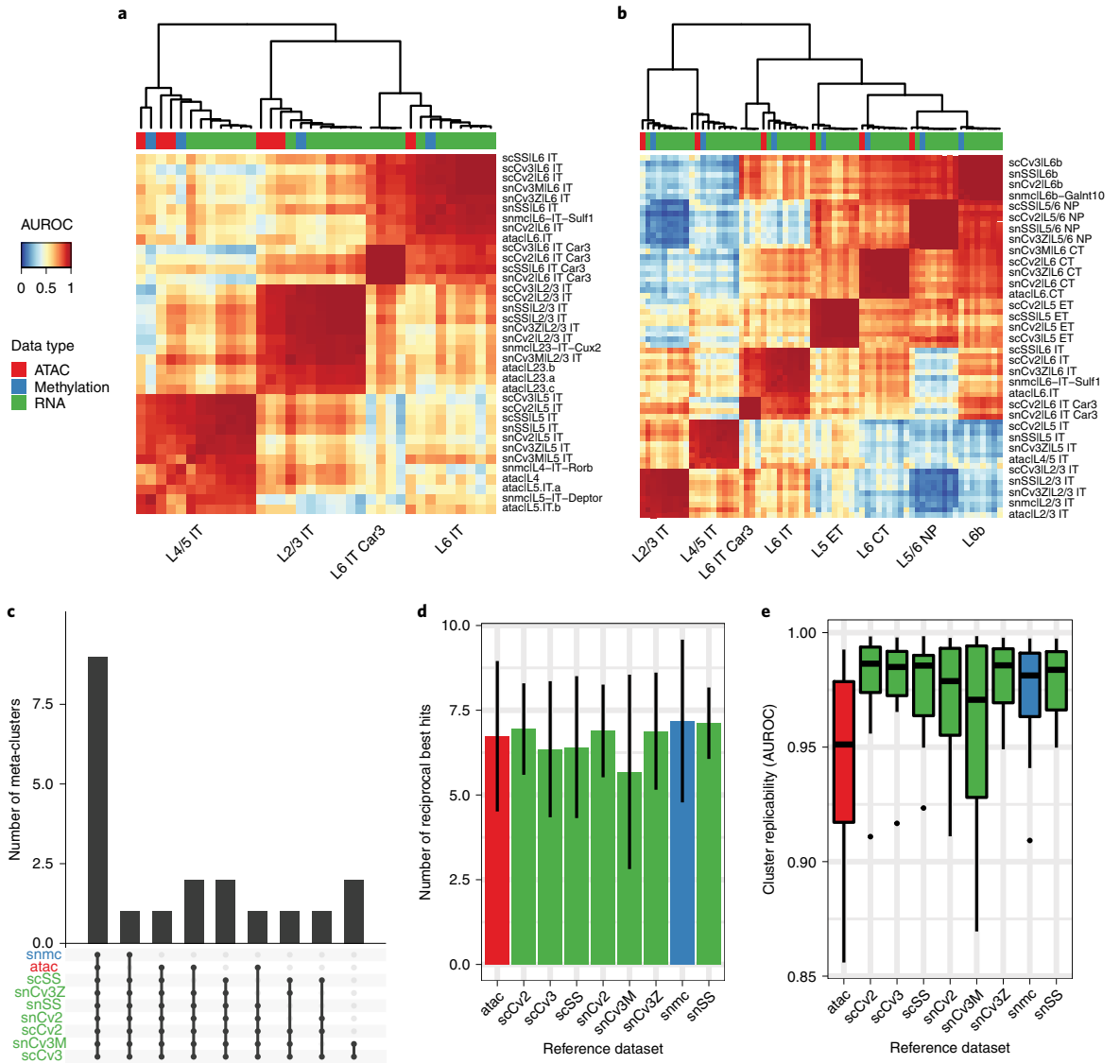


Fig. 15 | MetaNeighbor finds replicable cell types in a multimodal dataset of the mouse primary motor cortex. a, Heatmap based on MetaNeighbor AUROCs for IT projecting cell types, where cell types are grouped by applying hierarchical clustering. Column annotation colors indicate the sequencing modality (expression, chromatin accessibility or methylation). **b**, Heatmap based on MetaNeighbor AUROCs for excitatory cell types, where cell types are grouped by applying hierarchical clustering. Column annotation colors as in **a**. **c**, Upset plot showing the number of cell types that replicate across given combinations of datasets (meta-clusters). For example, nine cell types were found to replicate across all datasets. **d**, Number of reciprocal best hits for each dataset in the primary motor cortex compendium. The height of each bar indicates the average number of hits across cell types, and the line indicates the standard deviation. **e**, Boxplot showing the strength of cluster replicability (MetaNeighbor AUROC) across cell types for each dataset in the primary motor cortex compendium. The lower and upper hinges of the boxplots represent the first and third quartile, the central line represents the median, and the upper (respectively lower) whisker extends to the largest (respectively smallest) value within 1.5 interquartile range of the hinge. All points beyond 1.5 interquartile range are drawn individually.

gene-level quantification only imperfectly captures the variability of the modality and that a slightly more lenient AUROC threshold may be applied (e.g., AUROC >0.85). Note that MetaNeighbor can also be used for cross-species analyses[6] and that similar considerations may apply. When distant species are included in the analyses, expression signatures are expected to diverge, resulting in lower AUROC values overall.

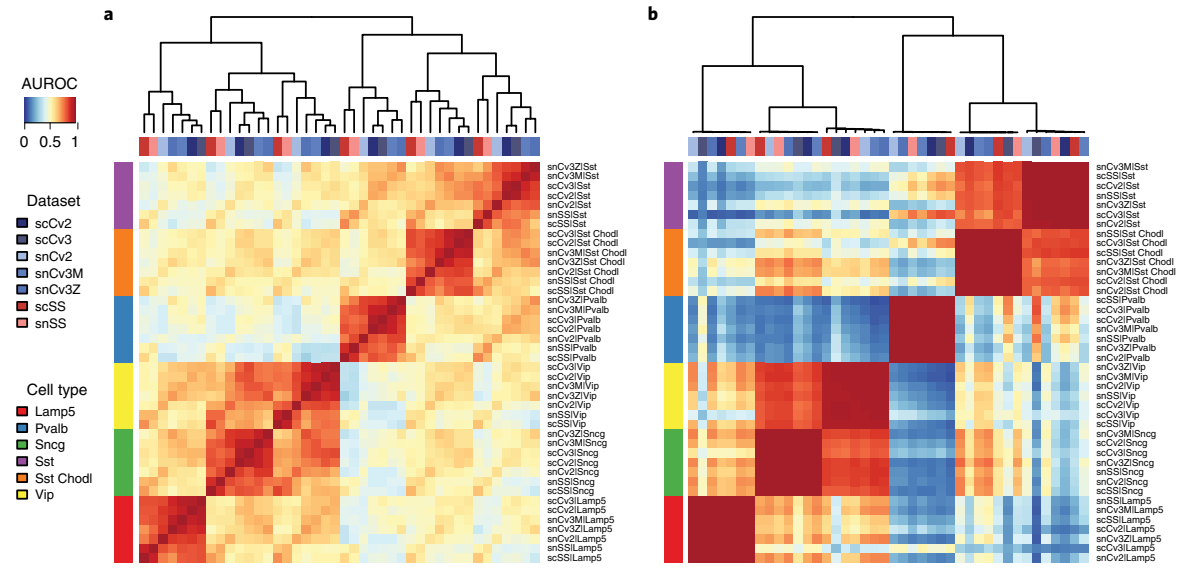


Fig. 16 | MetaNeighbor AUROCs offer a generalizable and batch-effect-free quantification of cell-type similarity. a, Possible issue: Spearman correlation of cell-type centroids is affected by technical variability. The heatmap shows some evidence of replicating cell types (light red rectangles) but is dominated by batch effects, largely obscuring secondary relationships between cell types. Red colors correspond to datasets obtained by using the Smart-Seq technology; blue colors, to datasets obtained by using the 10× technology; light colors, to single-nucleus datasets; and dark colors, to single-cell datasets. **b**, Anticipated result: MetaNeighbor AUROCs alleviate most of the concerns seen in **a**, with clear groups of replicating cell types (dark red squares, AUROC of ~1) and clear secondary relationships (e.g., similarity of CGE-derived interneurons *Vip*, *Sncg* and *Lamp5*).

Generalizable quantification of cell-type similarities

In their computation, MetaNeighbor's AUROCs are directly related to Spearman correlations. More precisely, all computations are based on average Spearman correlations between cells from two cell types but include an additional prediction step that alleviates batch effects, while keeping an interpretability power that is comparable to correlations (where AUROC = 0 maps to correlation = −1, AUROC = 0.5 maps to correlation = 0 and AUROC = 1 maps to correlation = 1).

To appreciate how the additional prediction step enables us to obtain 'batch-free correlations', we compare MetaNeighbor's output (Procedure 1, Step 11 and Procedure 2, Steps 5–7) with a more naive similarity output, where we compute the Spearman correlation between cell-type centroids (Fig. 16a). Centroid correlations display two desirable patterns: centroids cluster primarily by cell type (then by dataset), and global hierarchical structure is preserved (we can distinguish MGE-derived interneurons versus CGE-derived interneurons). However, batch effects are clearly visible throughout the heatmap. For example, within each cell type, Chromium-based datasets tend to cluster on one side, and Smart-Seq–based datasets tend to cluster on the other side. In contrast, for an equivalent computation time, all the 'good' patterns (cell types and hierarchical structure) are made pristinely clear with MetaNeighbor AUROCs (Fig. 16b), while technical substructure has been lost: technologies mix well within cell-types, homogeneous cell groupings look uniform and biological relationships between cell types are correctly displayed.

```
cell_types = as.factor(
  makeClusterName(biccn_data$study_id, biccn_data$joint_subclass_label)
)
normalization_factor = Matrix::colSums(assay(biccn_data)) / 1000000
cpm = assay(biccn_data)
cpm@x = cpm@x / rep.int(normalization_factor, diff(cpm@p))
cpm = as.matrix(cpm[biccn_hvgs,])
centroids = sapply(levels(cell_types), function(ct) {
  matrixStats::rowMeans2(log2(cpm+1), cols = cell_types == ct)
})
```

```
centroid_cor = cor(centroids, method = "spearman")
aurocs = MetaNeighborUS(var_genes = biccn_hvgs,
  dat = biccn_data,
  study_id = biccn_data$study_id,
  cell_type = biccn_data$joint_subclass_label,
  fast_version = TRUE)
plotHeatmap((1+centroid_cor)/2, cex = 0.5)
plotHeatmap(aurocs, cex = 0.5)
```

Compared to correlations, AUROCs have one additional 'parameter': the outgroup used for the prediction task. In Procedure 1, we illustrated how the outgroup can be controlled and interpreted. A deviation from AUROC = 1 can thus be interpreted as a combination of two factors: lack of similarity between cell types and choice of outgroup (difficulty of prediction task). If the outgroup is well controlled, AUROC values will generalize across studies and fundamentally indicate the quality of the clustering. For a given cell type in a given background (e.g., *Sst* cells in an unbiased sample of primary motor cortex inhibitory neurons), the similarity to *Sst* cells in another dataset (or any other inhibitory type for that matter) should be in the range of similarity observed within the BICCN.

As a robust alternative to centroid correlations, MetaNeighbor AUROCs can be applied to simple preprocessing tasks, such as identifying and selecting cell types that overlap between datasets before applying a merging framework. However, beyond the purely applicative viewpoint, we believe that MetaNeighbor-style AUROCs are a stepping stone toward a generalizable formalization of cell-type similarity.

Data availability

The datasets analyzed in the protocol are all previously published and publicly available. Human pancreas datasets were from Baron et al.[33] (Gene Expression Omnibus (GEO) accession code GSE84133), Lawlor et al.[34] (GEO accession code GSE86473), Muraro et al.[35] (GEO accession code GSE85241) and Segerstolpe et al.[36] (ArrayExpress accession code E-MTAB-5061). These datasets are accessed through the Bioconductor scRNAseq library in the protocol. The mouse primary visual cortex dataset was from Tasic et al.[32] (GEO accession code GSE71585), accessed through the Bioconductor scRNAseq library. The BICCN dataset for the mouse primary motor cortex from Yao et al.[4] is available on the Neuroscience Multi-Omic archive (https://assets.nemoarchive.org/dat-ch1nqb7). The subset of the BICCN data necessary to run the protocol is also available on FigShare at https://doi.org/10.6084/m9.figshare.13020569 (R version) and https://doi.org/10.6084/m9.figshare.13034171 (Python version).

Code availability

The code for the procedures (including all figures) is freely available on GitHub at https://github.com/gillislab/MetaNeighbor-Protocol in multiple formats (Rmd, PDF and jupyter notebook for R and Python). The scripts used to generate the protocol data are available in the same repository. The stable R version of MetaNeighbor is available through Bioconductor (https://www.bioconductor.org/install/) at https://www.bioconductor.org/packages/release/bioc/html/MetaNeighbor.html (the protocol was generated by using version 3.12), and the development versions are available on GitHub at https://github.com/gillislab/MetaNeighbor (R version) and https://github.com/gillislab/pyMN (Python version).

References

1. Hay, S. B., Ferchen, K., Chetal, K., Grimes, H. L. & Salomonis, N. The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp. Hematol.* **68**, 51–61 (2018).
2. Schaum, N. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
3. Almanzar, N. et al. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).
4. Yao, Z. et al. An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *Nature* (in the press).
5. Yao, Z. et al. A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. *Cell* (in the press).

6. Bakken, T. E. et al. Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. *Nature* (in the press).
7. Duò, A., Robinson, M. D. & Soneson, C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Res.* **7**, 1141 (2018).
8. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
9. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
10. Welch, J. D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887.e17 (2019).
11. Korsunsky, I. et al. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
12. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. *Nat. Biotechnol.* **37**, 685–691 (2019).
13. Barkas, N. et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. *Nat. Methods* **16**, 695–698 (2019).
14. Luo, C. et al. Single nucleus multi-omics links human cortical cell regulatory genome diversity to disease risk variants. Preprint at *bioRxiv* https://doi.org/10.1101/2019.12.11.873398 (2019).
15. Crow, M., Paul, A., Ballouz, S., Huang, Z. J. & Gillis, J. Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. *Nat. Commun.* **9**, 884 (2018).
16. Paul, A. et al. Transcriptional architecture of synaptic communication delineates GABAergic neuron identity. *Cell* **171**, 522–539.e20 (2017).
17. Hodge, R. D. et al. Conserved cell types with divergent features in human versus mouse cortex. *Nature* **573**, 61–68 (2019).
18. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
19. Forcato, M., Romano, O. & Bicciato, S. Computational methods for the integrative analysis of single-cell data. *Brief. Bioinform.* **22**, 20–29 (2020).
20. Hie, B. et al. Computational methods for single-cell RNA sequencing. *Annu. Rev. Biomed. Data Sci.* **3**, 339–364 (2020).
21. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).
22. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.22.111161 (2020).
23. Abdelaal, T. et al. A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019).
24. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
25. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat. Methods* **16**, 43–49 (2019).
26. Kapp, A. V. & Tibshirani, R. Are clusters found in one dataset present in another dataset? *Biostatistics* **8**, 9–31 (2007).
27. Dudoit, S., Fridlyand, J. & Speed, T. P. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **97**, 77–87 (2002).
28. Kiselev, V. Y. et al. SC3: consensus clustering of single-cell RNA-seq data. *Nat. Methods* **14**, 483–486 (2017).
29. Tasic, B. et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature* **563**, 72–78 (2018).
30. gillislab/MetaNeighbor-Protocol. https://github.com/gillislab/MetaNeighbor (2020).
31. Protocol data (R version). https://doi.org/10.6084/m9.figshare.13020569.v2 (2020).
32. Tasic, B. et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. *Nat. Neurosci.* **19**, 335–346 (2016).
33. Baron, M. et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Syst.* **3**, 346–360.e4 (2016).
34. Lawlor, N. et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.* **27**, 208–222 (2017).
35. Muraro, M. J. et al. A single-cell transcriptome atlas of the human pancreas. *Cell Syst.* **3**, 385–394.e3 (2016).
36. Segerstolpe, Å. et al. Single-cell transcriptome profiling of human pancreatic islets in health and type 2 diabetes. *Cell Metab.* **24**, 593–607 (2016).

Acknowledgements

J.G. was supported by NIH grants R01MH113005 and R01LM012736. S.F. was supported by NIH grant U19MH114821. B.D.H. was supported by the CSHL Crick Cray Fellowship. M.C. was supported by NIH grant K99MH120050.

Author contributions

S.F., M.C., B.D.H. and J.G. designed the experiments, performed the data analysis and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.G.

Peer review information *Nature Protocols* thanks Praneet Chaturvedi, Guoji Guo, Ahmed Mahfouz, Nathan Salomonis and Daniel Schnell for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 17 October 2020; Accepted: 25 May 2021;
Published online: 07 July 2021

Related links

Key references using this protocol

Crow, M. et al. *Nat. Commun.* **9**, 884 (2018): https://doi.org/10.1038/s41467-018-03282-0
Paul, A. et al. *Cell* **171**, 522–539.e20 (2017): https://doi.org/10.1016/j.cell.2017.08.032
Yao, Z. et al. Preprint at *bioRxiv* (2020): https://doi.org/10.1101/2020.02.29.970558
Bakken, T. E. et al. Preprint at *bioRxiv* (2020): https://doi.org/10.1101/2020.03.31.016972

Key data used in this protocol

Yao, Z. et al. Preprint at *bioRxiv* (2020) https://doi.org/10.1101/2020.02.29.970558
Baron, M. et al. *Cell Syst.* **3**, 346–360.e4 (2016) https://doi.org/10.1016/j.cels.2016.08.011
Lawlor, N. et al. *Genome Res.* **27**, 208–222 (2017) https://doi.org/10.1101/gr.212720.116
Muraro, M. J. et al. *Cell Syst.* **3**, 385–394.e3 (2016) https://doi.org/10.1016/j.cels.2016.09.002
Segerstolpe, Å. et al. *Cell Metab.* **24**, 593–607 (2016) https://doi.org/10.1016/j.cmet.2016.08.020
Tasic, B. et al. *Nat. Neurosci.* **19**, 335–346 (2016) https://doi.org/10.1038/nn.4216

Appendix B

Single cell RNA sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery in maize

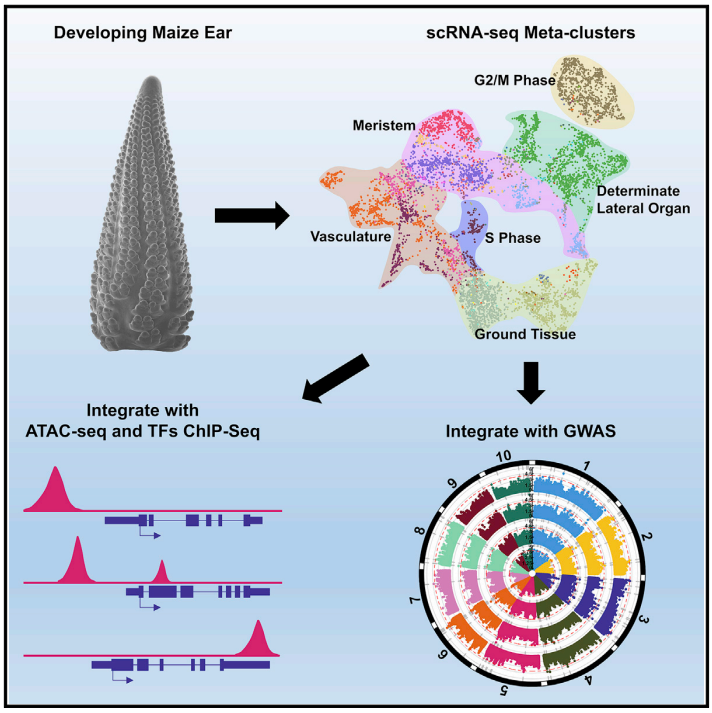
In this chapter I include the manuscript from my collaboration with Xiaosa (Jack) Xu and Dave Jackson. I did the co-expression analysis and worked with my labmates, Megan Crow and Nathan Fox to cluster the cells, annotate them, and evaluate the replicability of the cell types using MetaNeighbor A. My contributions were specifically in figures 1, 2 and Supplementary Figures 1,2 and, 4. At the time of thesis submission, this is an ongiong collaboration building of this work and also the multiscale co-expression work from 2. The reference genome for maize has poor functional annotations so I am developing methods that rely on known annotations an reference bulk co-expression networks to find co-expression modules in bulk that might be informative of the cell types in new scRNAseq data Xiaosa is generating.

Resource

Developmental Cell

Single-cell RNA sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery

Graphical Abstract



Highlights

- scRNA-seq of developing maize ears reveals major cell types and developmental markers
- scRNA-seq co-expression networks predict genetic redundancy
- Integration of scRNA-seq and ChIP-seq/ATAC-seq helps build transcriptional networks
- Integration of scRNA-seq and GWAS identifies candidate maize yield-associated genes

Authors

Xiaosa Xu, Megan Crow, Brian R. Rice, ..., Alexander E. Lipka, Jesse Gillis, David Jackson

Correspondence

jacksond@cshl.edu

In Brief

Xu et al. construct and validate a single-cell transcriptomic atlas of developing maize ears. Their single-cell gene co-expression networks will facilitate developmental genetics studies by predicting genetic redundancy and revealing transcriptional regulatory networks. Their results also inform maize breeding by identifying candidate trait-associated genes.

Xu et al., 2021, Developmental Cell *56*, 557–568
February 22, 2021
https://doi.org/10.1016/j.devcel.2020.12.015

CellPress

Developmental Cell

CellPress

Resource

Single-cell RNA sequencing of developing maize ears facilitates functional analysis and trait candidate gene discovery

Xiaosa Xu,[1] Megan Crow,[1] Brian R. Rice,[2] Forrest Li,[1] Benjamin Harris,[1] Lei Liu,[1] Edgar Demesa-Arevalo,[1] Zefu Lu,[3] Liya Wang,[1] Nathan Fox,[1] Xiaofei Wang,[1] Jorg Drenkow,[1] Anding Luo,[4] Si Nian Char,[5] Bing Yang,[5,6] Anne W. Sylvester,[4] Thomas R. Gingeras,[1] Robert J. Schmitz,[3] Doreen Ware,[1,7] Alexander E. Lipka,[2] Jesse Gillis,[1] and David Jackson[1,8,*]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
[2]Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[3]Department of Genetics, University of Georgia, Athens, GA 30602, USA
[4]Department of Molecular Biology, University of Wyoming, Laramie, WY 82071, USA
[5]Division of Plant Sciences, Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA
[6]Donald Danforth Plant Science Center, St. Louis, MO 63132, USA
[7]USDA-ARS, Robert W. Holley Center, Ithaca, NY 14853, USA
[8]Lead contact
*Correspondence: jacksond@cshl.edu
https://doi.org/10.1016/j.devcel.2020.12.015

SUMMARY

Crop productivity depends on activity of meristems that produce optimized plant architectures, including that of the maize ear. A comprehensive understanding of development requires insight into the full diversity of cell types and developmental domains and the gene networks required to specify them. Until now, these were identified primarily by morphology and insights from classical genetics, which are limited by genetic redundancy and pleiotropy. Here, we investigated the transcriptional profiles of 12,525 single cells from developing maize ears. The resulting developmental atlas provides a single-cell RNA sequencing (scRNA-seq) map of an inflorescence. We validated our results by mRNA *in situ* hybridization and by fluorescence-activated cell sorting (FACS) RNA-seq, and we show how these data may facilitate genetic studies by predicting genetic redundancy, integrating transcriptional networks, and identifying candidate genes associated with crop yield traits.

INTRODUCTION

Plant architecture is initiated by meristems made up of pluripotent stem cells and their descendants that are organized in distinct cell types and domains with specific functions. In developing maize ears, a series of meristems build inflorescence architecture, including spikelet pair meristems (SPMs) formed from the inflorescence meristem (IM) and spikelet meristems (SMs) made from the branching of SPMs (Irish, 1997). Mutant studies have identified key cell type or domain-specific regulators that orchestrate inflorescence architecture by specifying different developmental domains (Vollbrecht and Schmidt, 2009). For example, the homeodomain transcription factor encoded by *KNOTTED1* (*KN1*) is critical for meristem establishment and maintenance and is expressed throughout shoot meristems (Jackson et al., 1994). The production of axillary meristems to elaborate branching architecture depends on expression of a basic helix-loop-helix transcription factor encoded by *BARREN STALK1* (*BA1*), expressed specifically in the adaxial meristem periphery where axillary meristems initiate (Gallavotti et al., 2004). Another transcription factor, *BRANCHED SILKLESS1* (*BD1*), is expressed at the boundary of meristems and glumes to control spikelet architecture by promoting meristem determinacy (Chuck et al., 2002), whereas *RAMOSA* genes, such as *RA1* and *RA3*, are expressed in an arc of cells at the base of meristems, to impose determinacy on spikelet branches (Satoh-Nagasawa et al., 2006). Many of these key regulators have reshaped inflorescence architecture during evolution or domestication, and their discovery was enabled by the availability of mutants that block specific aspects of development. However, such insights are limited by genetic redundancy and pleiotropy, so a high-resolution expression atlas of specific cell types and domains is needed to gain further insights into the gene networks that control development.

Single-cell RNA sequencing (scRNA-seq) offers the opportunity to assay gene expression with high resolution and to construct developmental maps of complex organs or organisms (Kulkarni et al., 2019; Potter, 2018). Recently, the 10x Genomics Chromium scRNA-seq platform has been used extensively to identify cell type or domain markers in *Arabidopsis* roots (Rich-Griffin et al., 2020), but the application of this technology to shoot tissues has been limited. As well as providing expression

CellPress
Developmental Cell
Resource

information, scRNA-seq data can be integrated with other genomic datasets, such as ChIP-seq identification of targets of transcription factors, or surveys of chromatin status. Such datasets have been generated from developing maize inflorescences (Bolduc et al., 2012; Eveland et al., 2014; Pautler et al., 2015), but single-cell data have not yet been integrated.

Productivity of maize depends on development of the inflorescences, in particular the seed-bearing ear. Genome-wide association studies have identified candidate genes associated with yield-related traits (Liu et al., 2020) that can guide breeding or trait engineering. Regulatory genes functioning in early stages of ear development show significant association with ear yield traits (Vollbrecht and Schmidt, 2009; Bommert et al., 2013; Je et al., 2016; Liu et al., 2020), yet it remains challenging to identify and validate such regulators on a genome-wide scale. To fill these gaps, we optimized a protocol using 10x Genomics scRNA-seq technology to generate a high-resolution transcriptome atlas of the developing maize ear inflorescence. We illustrate how these data can enhance maize genetics by predicting genetic redundancy, build transcriptional regulatory networks at cell-type resolution, and identify candidate loci associated with ear yield traits.

RESULTS

Construction of a single-cell transcriptome atlas of the developing maize ear

To generate a single-cell atlas from developing maize ears, we used the 5–10 mm stage, where major developmental and architectural decisions, including meristem initiation, maintenance and determinacy, organ specification, and differentiation of vascular and ground tissues, are being made (Irish, 1997; Vollbrecht and Schmidt, 2009). We first optimized a cell wall digestion method, taking into account the different composition of grass cell walls (Ortiz-Ramírez et al., 2018), that allowed us to isolate ear protoplasts within ~45 min (see STAR methods). However, developing ear protoplasts were fragile, and we removed small debris and organelles from broken cells (Figure S1A) by filtration followed by FACS (see STAR methods; Figure 1A) before loading into the 10x Genomics Chromium Controller. Then, scRNA-seq libraries were generated and sequenced on the Illumina platform (Table S1). In total, we profiled 12,525 individual cells from three independent replicates (Table S1) and detected expression from 28,899 genes using maize V3 reference, comparable to the number detected by bulk RNA-seq of the same tissue (Eveland et al., 2014; Pautler et al., 2015).

Technical variation and sparse data in scRNA-seq make it challenging to identify reproducible clusters (groups of cells) that represent homogeneous cell types across technical replicates (Crow et al., 2018). We used MetaNeighbor to ask how well the identity of cells in a given cluster of one replicate can be predicted based on their similarity to a cluster from another replicate (Crow et al., 2018), reported as the average area under the receiver operating characteristic curve (AUROC). All cluster pairs with AUROCs > 0.9 in both directions, across at least two replicates, were used to merge and identify 12 replicable cell identity clusters (hereafter referred to as meta-clusters) (Figure 1B). As some genes may be affected by protoplasting (the process to isolate protoplasts), we used mRNA-seq to compare total ear protoplasts with freshly dissected, intact developing ear tissue, and we identified 713 protoplasting-responsive genes (FDR < 0.05, $|\log_2 FC| > 2$; Table S1). After excluding these genes, > 97% of the highly variable genes used for generating clusters were unchanged, as were the meta-clusters (Figures S1B and S1C). Thus, protoplasting-responsive genes did not affect clustering, as reported previously (Ma et al., 2020).

Prediction and validation of meta-cluster identities

We visualized meta-clusters using a uniform manifold approximation and projection (UMAP) plot, where we could track the distribution of genes of interest (Figure 1C). Next, to predict identities for each meta-cluster, we compiled a list of known or predicted inflorescence development marker genes, whose expression patterns have been studied in maize or *Arabidopsis* (Table S1). Among them, 74 are functionally characterized by their mutant phenotype in maize. Importantly, we detected the expression of 73 of these genes, and each had enriched expression in one or more of the 12 meta-clusters (Figures 1D–1N and S2; Table S1). For example, to identify meristem cell types, we used *KN1*, which is expressed throughout the meristem as well as the developing stem and vascular tissues, but strictly excluded from the epidermis and determinate lateral organs (Jackson et al., 1994). *KN1* was expressed in multiple (10 out of 12) meta-clusters, as expected (Figure 1D). Ear architecture is governed by branching events that are controlled by genes expressed in different meristem domains. To identify meta-clusters representing these domains, we searched for expression of characterized marker genes, including *BD1*, which is expressed at the boundary of spikelet meristems (Figures 1E and 1O; Chuck et al., 2002), and found it to be uniquely expressed in meta-cluster 9, identifying this as a meristem boundary meta-cluster (Figure 1E). Another well-characterized marker, *BA1*, is expressed in a distinct adaxial meristem periphery domain (Figures 1F and 1O; Gallavotti et al., 2004) and was expressed in meta-cluster 11 (Figure 1F). A third cellular domain that controls branching is marked by an arc of expression of *RAMOSA* genes at the meristem base (Figures 1G and 1O) that partially overlaps with *BA1* (Satoh-Nagasawa et al., 2006; Vollbrecht and Schmidt, 2009). Correspondingly, we found expression of *RAMOSA* genes in meta-cluster 11 (Satoh-Nagasawa et al., 2006), similar to *BA1*, as well as in meta-cluster 10 (Figure 1G). As such, we could identify three of the *KN1* expressing meta-clusters as distinct branching domains.

KN1 expression is excluded from the meristem epidermis and determinate lateral organs (Jackson et al., 1994). We found epidermis marker gene *ZmHOMEODOMAIN LEUCINE ZIPPER IV8* (*ZmHDZIV8*) (Javelle et al., 2011) was highly enriched in the two meta-clusters, 3 and 6, that did not express *KN1* (Figure 1H), while determinate lateral organ marker gene *ZmYABBY14* (*ZmYAB14*) (Strable et al., 2017) was expressed throughout meta-cluster 3 (Figure 1I), suggesting that meta-cluster 3 was a determinate lateral organ meta-cluster while meta-cluster 6 corresponded to meristem epidermis cells (Figure 1O). Consistently, meta-cluster 3 was significantly enriched for organ initiation genes from a maize shoot laser capture microdissection (LCM) study (q < 0.001) (Table S1; Knauer et al., 2019). Using additional markers, we identified four distinct sub-clusters of

Developmental Cell
Resource

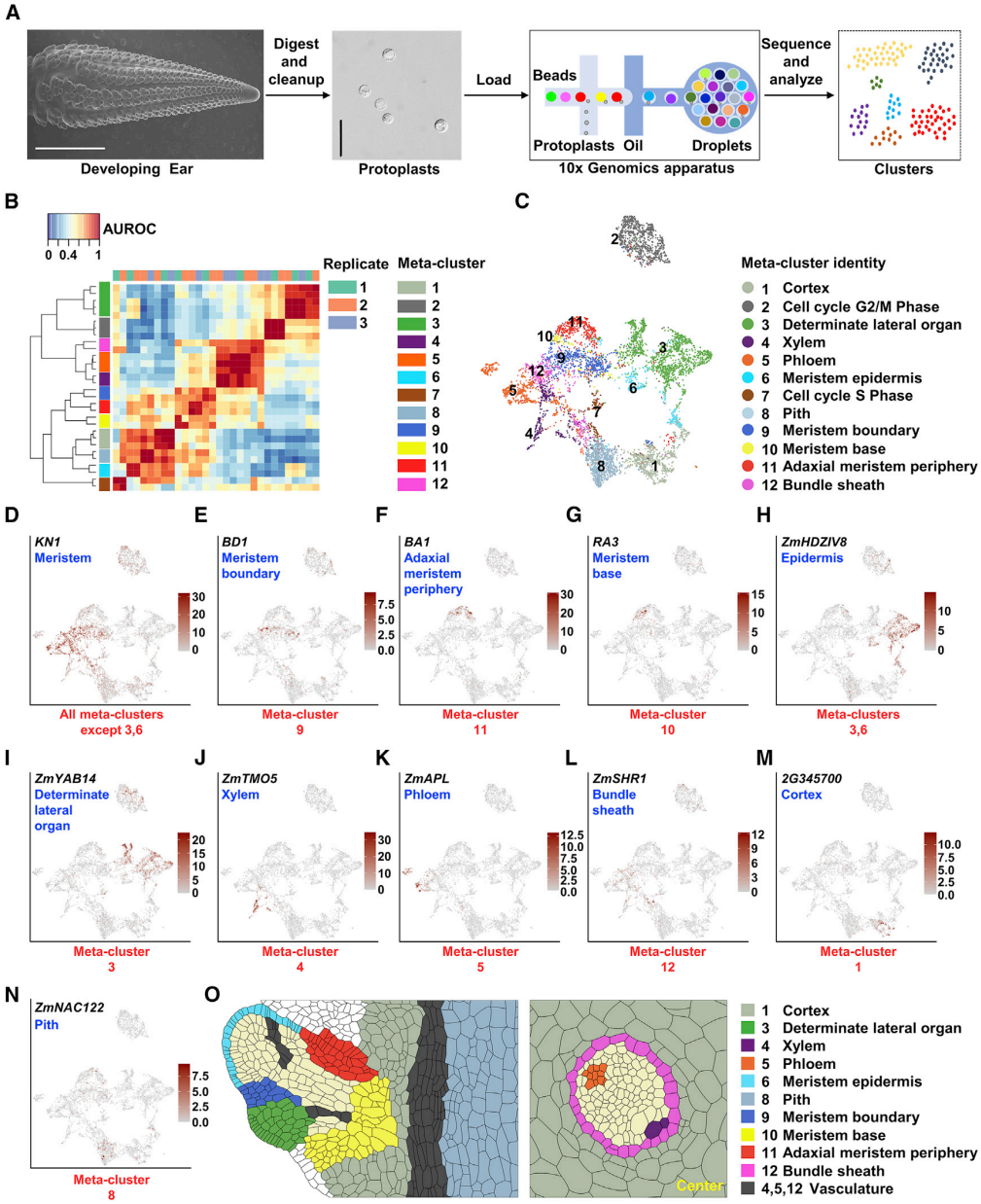


Figure 1. Isolation of maize ear protoplasts to construct a single-cell transcriptomic atlas

(A) Experimental design, the first panel shows a scanning electron microscope image of a 5–10 mm developing ear (scale bar = 2 mm), second panel image of ear protoplasts, scale bar = 50 μm.

(B) MetaNeighbor identifies 12 reproducible meta-clusters (left color blocks) across three biological replicates (top color blocks) of single-cell RNA-seq datasets.

(C) 12 meta-clusters displayed by an integrated uniform manifold approximation and projection (UMAP) plot in two dimensions, with each dot representing a cell.

(D–N) UMAP plots of marker genes predicting the identities of meta-clusters, with color scale indicating normalized expression level. (D) *KN1*, meristem, all meta-clusters except 3 and 6; (E) *BD1*, meristem boundary, meta-cluster 9; (F) *BA1*, adaxial meristem periphery, meta-cluster 11; (G) *RA3*, meristem base, meta-cluster 10; (H) *ZmHDZIV8*, epidermis, meta-clusters 6 and part of 3; (I) *ZmYAB14*, determinate lateral organ, meta-cluster 3; (J) *ZmTMO5*, xylem, meta-cluster 4; (K) *ZmAPL*, phloem, meta-cluster 5; (L) *ZmSHR1*, bundle sheath, meta-cluster 12; (M) *GRMZM2G345700* (*2G345700*), cortex, meta-cluster 1; (N) *ZmNAC122*, pith, meta-cluster 8.

(O) Sketches of longitudinal section of a spikelet meristem (left panel) and transverse section of vascular bundle (right panel) showing cell/domain identities in scRNA-seq meta-clusters.

CellPress
Developmental Cell
Resource

meta-cluster 3 (Figure S2A), including determinate lateral organ epidermis domain (marker = *LIPID TRANSFER PROTEIN1* [*LTP1*]) (Takacs et al., 2012) and non-epidermis domain (marker = *ZmRIBULOSE BISPHOSPHATE CARBOXYLASE SMALL SUBUNIT 1A* [*ZmRBCS1A*]), as well as adaxial domain (marker = *DROOPING LEAF2/ZmYABBY7* [*DRL2/ZmYAB7*]) and abaxial domain (marker = Homolog of *Arabidopsis AUXIN RESPONSE FACTOR3/4* [*ZmARF3/4-LIKE1*]) (Chitwood et al., 2007).

KN1 is also expressed in developing vascular tissues (Jackson et al., 1994). To identify distinct vascular meta-clusters, we used maize homologs of *Arabidopsis* genes (De Rybel et al., 2016). For example, xylem marker *ZmTARGET OF MONOPTEROS 5* (*ZmTMO5*) and its paralogs *ZmTMO5-LIKE1* and *2* were expressed in meta-cluster 4 (Figures 1J, 1O, and S2B). We also found a sub-cluster of predicted maturing xylem cells in meta-cluster 4 (Figure S2B), using homologs of *Arabidopsis* marker genes for secondary cell walls and tracheary elements, including *ZmMYB DOMAIN PROTEIN 46* (*ZmMYB46*) (Zhong et al., 2007) and *ZmXYLEM CYSTEINE PEPTIDASE 2* (*ZmXCP2*) (Funk et al., 2002). In contrast, we found that meta-cluster 5 represented phloem cells, as shown by specific expression of a maize homolog of *Arabidopsis ALTERED PHLOEM DEVELOPMENT* (Figures 1K and 1O; De Rybel et al., 2016). Phloem tissues include distinctive sieve element and companion cells, which were reflected in sub-clusters marked by maize homologs of *Arabidopsis* protophloem sieve element marker *PHLOEM EARLY DOF 1* (*PEAR1*)/*PEAR2* (Miyashima et al., 2019) and companion cell marker *PHLOEM PROTEIN 2-LIKE A1* (Figure S2C; Guo et al., 2018). Therefore, meta-clusters 4 and 5 correspond to xylem and phloem cells, respectively, and we found significant enrichment of genes in these meta-clusters with vascular markers in a maize LCM study ($q < 0.001$) (Table S1; Knauer et al., 2019). In addition to meta-cluster 4 and 5, cells in meta-cluster 12 also expressed vascular marker genes ($q < 0.001$) (Table S1; Knauer et al., 2019), including bundle sheath markers, such as *ZmSHORT-ROOT* (*ZmSHR1*) (Figures 1L and 1O; Chang et al., 2012); thus, meta-cluster 12 was predicted to be bundle sheath.

The remainder of the developing ear corresponds to ground tissue, including the outer cortex and inner pith tissues (Figures 1M–1O and S2D). A maize homolog, *GRMZM2G345700,* of the most unique *Arabidopsis* root cortex marker, *AT1G62510,* a bifunctional lipid-transfer/2S albumin superfamily gene (Denyer et al., 2019), had restricted expression in meta-cluster 1 (Figure 1M), as did stem cortex marker homolog *ZmNITRATE TRANSPORTER 1/PEPTIDE TRANSPORTER FAMILY 6.4-LIKE 2* (*ZmNPF6.4-LIKE2*) (Figure S2D; Tong et al., 2016). We thus predicted this meta-cluster to be cortex. In contrast, we predicted that meta-cluster 8 was comprised of pith cells, by specific expression of homologs of sorghum or *Arabidopsis* markers, such as *ZmNO APICAL MERISTEM DOMAIN CONTAINING* (*NAC*) *TRANSCRIPTION FACTOR 122* (*ZmNAC122*), *GRMZM2G430849*, a homolog of *Sobic.006G147400* (Figure 1N; Fujimoto et al., 2018), or *GRMZM2G039074*, a homolog of *AT2G3830*, an MYB transcriptional regulator (Figure S2D; Schürholz et al., 2018). Finally, meta-clusters 2 and 7 were highly enriched for expression of cell cycle genes, such as *ZmCYCLINB1;2* (*ZmCYCB1;2*) and *ZmHISTONE2A12* (*ZmHIS2A12*), indicating that these two meta-clusters contained dividing cells at different phases of the cell cycle (Figure S2D); similar cell cycle clusters are found in root scRNA-seq studies (Denyer et al., 2019; Rich-Griffin et al., 2020). We calculated the percentage of cells in each meta-cluster (Figure S2E). 23% of cells were from meristem domains (meta-clusters 6, 9, 10, and 11), 21% from ground tissues (meta-clusters 1 and 8), 20% from vascular tissues (meta-clusters 4, 5, and 12), and 19% from determinate lateral organ tissues (meta-cluster 3). In summary, using maize inflorescence development markers and homologs of markers from other plants, we predicted the cell or domain identities of all 12 meta-clusters (Figure 1C) and in several cases sub-divided them into more specific cell types or developmental stages.

To validate our predicted meta-cluster identities, we first used differential expression (DE) analysis to identify marker genes with AUROCs $\geq$ 0.7 in at least one replicate, and we identified 813 candidate markers (Table S1). The top markers of each meta-cluster were further selected based on the percentage of cells expressing the marker and showed highly enriched expression, as expected (Figure 2A). We prioritized a set of these markers by predicted developmental roles and validated them using *in situ* hybridization (Figures 2B–2M; Table S1). For example, marker genes for meta-cluster 9, *GRMZM2G004528*, annotated as *ZmMYO-INOSITOL PHOSPHATE SYNTHASE2* (*ZmMIPS2*), predicted to act in auxin signaling and transport (Chen and Xiong, 2010), and *GRMZM2G097989*, annotated as *ZmGLUTATHIONE TRANSFERASE 41* (*ZmGST41*), involved in meristem size control (Horváth et al., 2019), showed specific expression in the meristem boundary, similar to *BD1* (Figures 2B and 2C; Chuck et al., 2002). *ZmGST41* was also a DE marker for meta-cluster 12 and consistently showed vascular trace expression (Figure 2C, arrow). Markers of a second meristem meta-cluster, 11, included *GRMZM2G038284,* which was expressed in the adaxial meristem periphery similar to *BA1* (Figure 2D), and encodes a homolog of *Arabidopsis DROUGHT INDUCED19*, of interest because maize ear development is especially sensitive to drought stress (Nuccio et al., 2015). Two additional markers of meta-cluster 11, *GRMZM2G034152*, which encodes a *ZmPOLYAMINE OXIDASE 1* (*ZmPAO1*) (Figure 2E), and *GRMZM2G430522*, a homolog of *Arabidopsis CUP-SHAPED COTYLEDON 3* (*ZmCUC3-LIKE*) (Figure S3A), also showed restricted expression at the adaxial meristem periphery. Meta-clusters 3 and 6 were predicted to have an epidermal identity, and specific expression was observed as expected for markers such as *EF517601.1_FG016*, annotated as *MALE FLOWER SPECIFIC 18* (Figure 2F), and *GRMZM2G126397*, a *ZmPHOSPHOLIPID TRANSFER PROTEIN3* (*ZmPLTP3*) gene (Figure 2G). Moving away from the meristem, marker genes for meta-cluster 3, predicted to be determinate lateral organ, included *GRMZM2G019686*, annotated as *ZmFLOWERING PROMOTING FACTOR 1* (*ZmFPF1*) (Figure 2H), and *GRMZM2G075255*, annotated as *ZmECERIFERUM1* (*ZmCER1*) (Figure 2I), and showed expected expression patterns.

We also identified candidate vascular markers in meta-clusters 4 and 5. Predicted xylem markers included *ZmTARGET OF MONOPTEROS5-LIKE3* (*ZmTMO5-LIKE3*), *GRMZM2G176141* (Figure 2J), *ZmTRANSMEMBRANE AMINO ACID TRANSPORTER FAMILY PROTEIN* (*ZmTMAAT*), *GRMZM2G109865* (Figure S3B), and *ZmWALLS ARE THIN 1* (*ZmWAT1*), *GRMZM2G007953* (Figure S3C), and all showed specific expression identified by the distinctive cell walls of xylem vessels. Interestingly, some xylem

Developmental Cell
Resource
CellPress

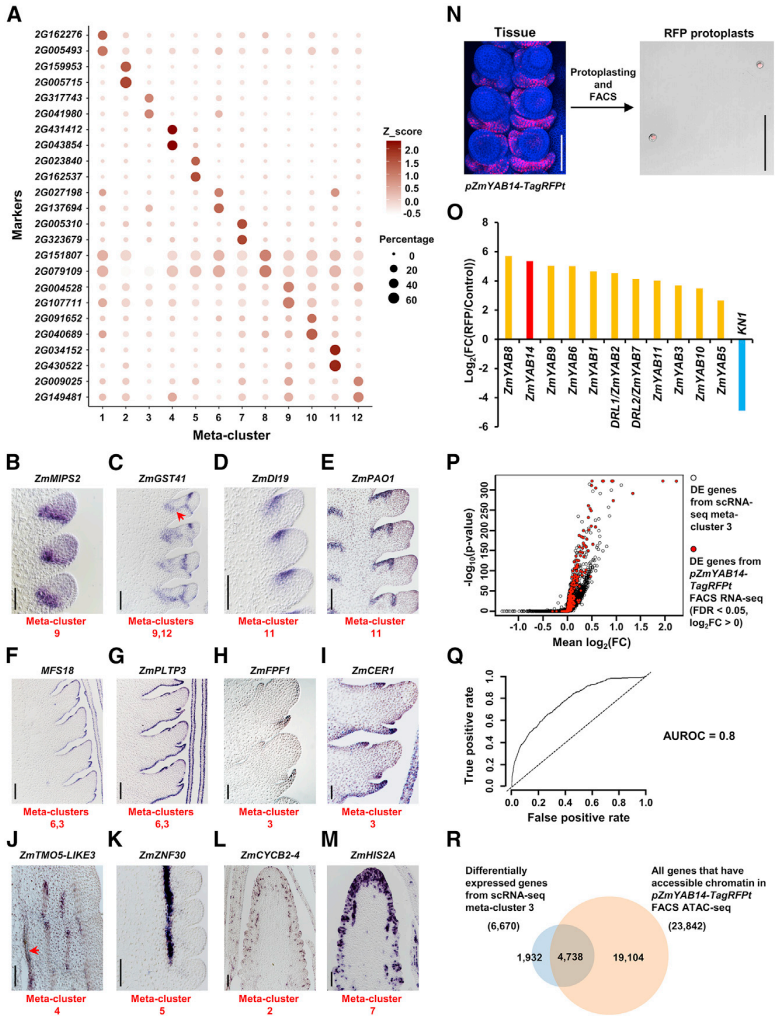


Figure 2. Validation of scRNA-seq by mRNA *in situ* and FACS RNA-seq

(A) The top two marker genes of each meta-cluster are shown in dot plots with circle size indicating the percentage of cells expressing the marker and color representing Z_scored expression value.

(B–M) mRNA *in situ* of meta-cluster marker genes validates the predicted identities: (B) *ZmMIPS2*, meristem boundary; (C) *ZmGST41*, meristem boundary (meta-cluster 9) and bundle sheath (meta-cluster 12, red arrow); (D and E) *ZmDI19* (D) and *ZmPAO1* (E), adaxial meristem periphery; (F and G) *MFS18* (F) and *ZmPLTP3* (G), meristem epidermis (meta-cluster 6) and determinate lateral organ (meta-cluster 3) epidermis; (H and I) *ZmFPF1* (H) and *ZmCER1* (I), determinate lateral organ; (J) *ZmTMO5-LIKE3*, xylem (red arrow indicates xylem vessels); (K) *ZmZNF30*, phloem. (L) *ZmCYCB2-4*, cell cycle G2/M phase; (M) *ZmHIS2A*, cell cycle S phase. Scale bar = 100 μm.

(N) Collection of RFP protoplasts from *pZmYAB14-TagRFPt* reporter line using FACS. Scale bar = 100 μm. Three biological replicates were collected for FACS RNA-seq. One biological replicate was collected for FACS ATAC-seq.

(O) Log_2(fold change(FC)) of determinate lateral organ domain enriched markers, *ZmYAB* genes, and depleted marker, *KN1*, between RFP and total control protoplasts (Control) in FACS RNA-seq.

(P) Volcano plot with 1-sided test positions the hits of enriched markers from *pZmYAB14-TagRFPt* FACS RNA-seq (red dots) on the ranked list of scRNA-seq differentially expressed (DE) genes from meta-cluster 3 (black circles). x axis indicates the mean log_2(FC) of DE genes between meta-cluster 3 and all other meta-clusters. y axis indicates corresponding -log_{10}(p-value).

(Q) *pZmYAB14-TagRFPt* FACS RNA-seq and scRNA-seq meta-cluster 3 have concordant differential gene expression patterns with area under the receiver operating characteristics (AUROC) score = 0.8 (indicated by curved line; dashed line indicates the null (AUROC score = 0.5)). Axes indicates the true and false positive rate, the proportion of *pZmYAB14-TagRFPt* FACS RNA-seq enriched markers that do or do not match to scRNA-seq meta-cluster 3 enriched markers, respectively.

(R) Meta-cluster 3 DE genes are enriched in open chromatin in *pZmYAB14-TagRFPt* FACS sorted cells, see text for details.

markers were also expressed in meristem tips, mostly enriched in central zone (Figures S3D–S3F). We also confirmed the specific expression of meta-cluster 5 (phloem) marker, *GRMZM2G116079*, which encodes Zinc Finger Protein 30 (ZmZNF30), whose *Arabidopsis* homolog, *AT3G15680*, is predicted to be involved in RNA regulation (Figure 2K; Gipson et al., 2020). Lastly, candidate markers from meta-clusters 2 and 7, predicted to be dividing cells, included several cyclin and histone encoding genes, such as *ZmCYCLINB2-4* (*ZmCYCB2-4*), *GRMZM2G061287* (Figure 2L), and *ZmHISTONE2A* (*ZmHIS2A*), *GRMZM2G305046* (Figure 2M), and had punctate expression, as expected.

FACS RNA-seq has been used to validate scRNA-seq data in *Arabidopsis* roots (Rich-Griffin et al., 2020). Few marker lines are available in maize, but one, *pZmYAB14-TagRFPt* (Je et al., 2016), is specifically expressed in determinate lateral organs (Figure 2N). We introgressed this reporter into a *bd1;Tunicate* (*bd1;Tu*) double mutant background, which produces highly proliferative ears, to generate large amounts of ear tissue. We made protoplasts from this tissue and used FACS to sort RFP-positive cells, followed by RNA-seq to identify lateral organ domain-specific genes (Figure 2N; Table S2). We found highly enriched expression of *ZmYAB14* and other *YAB* genes (Strable et al., 2017) in the FACS-sorted cells, while negative control markers such as *KN1* (Jackson et al., 1994) were significantly depleted (Figure 2O), as expected. We identified 2,040 differentially expressed genes (FDR < 0.05) (Table S2), and as we expected the majority were differentially expressed (AUROC score 0.8) in scRNA-seq meta-cluster 3, with predicted lateral organ identity, validating our scRNA-seq data (Figures 2P and 2Q).

CellPress
Developmental Cell
Resource

We also used FACS-sorted *pZmYAB14-TagRFPt*-expressing cells in an ATAC-seq experiment to investigate how chromatin accessibility changes during differentiation of determinate lateral organs in the ear. Genome-wide analysis of accessible chromatin regions (ACRs) found that 31% mapped within 10 kb upstream of the transcription start site (TSS), and 18% localized to transcription termination sites (TTS), untranslated regions (UTRs), exons, or introns, comparable to whole ear tissue ATAC-seq results (Figure S3G; Table S2; Ricci et al., 2019). 60% of all maize genes had ACRs in *pZmYAB14-TagRFPt* FACS-sorted cells (Figure 2R; Table S2), and this value was significantly enriched for DE genes from scRNA-seq meta-cluster 3 (71%; $p < 0.001$, chi-square test, Figure 2R; Table S2). As expected, several *ZmYAB* genes that were significantly enriched in scRNA-seq meta-cluster 3 had accessible chromatin (Figure S3H), and we validated the expression of two meta-cluster 3 marker genes with ACRs by *in situ* hybridization (Figures S3I and S3J), including *GRMZM2G026556*, a homolog of *Arabidopsis BLADE ON PETIOLE2*, which controls lateral organ fate (Ha et al., 2007), and *GRMZM2G004012*, homolog of *Arabidopsis PLANTACYANIN*, that plays a role in the development of reproductive organs (Dong et al., 2005). To gain further insight into these data, we subtracted the determinate lateral organ-specific ATAC-seq peaks from whole developing ear ATAC-seq peaks (Ricci et al., 2019) to predict ACRs specific to indeterminate meristem tissues, as well as vascular and ground tissues (Figure S3K; Table S2). scRNA-seq marker genes corresponding to these domains (Figure S3K; Table S1) were significantly enriched with these ACR-associated genes ($p < 0.001$, chi-square test; Figure S3K; Table S2). Thus, the integration of ATAC-seq and scRNA-seq may provide insights into chromatin accessibility and its effect on gene expression in specific cell types or developmental contexts.

scRNA-seq networks predict redundancy

Gene redundancy often masks the phenotype of single-gene knockouts (Lloyd and Meinke, 2012); however, distinguishing redundant from non-redundant paralogs can be challenging. In a recent study of the maize branching mutant *ramosa3* (*ra3*), we identified a *ra3* enhancer as its paralog, *ZmTREHALOSE PHOSPHATE PHOSPHATASE 4* (*ZmTPP4*) (Claeys et al., 2019). Among 12 maize *ZmTPP* genes, two of them, *ZmTPP4* and *ZmTPP12,* are upregulated in *ra3* mutants (Eveland et al., 2014), a common predictor of compensating redundant paralogs (Rodriguez-Leal et al., 2019). However, CRISPR knockouts of *ZmTPP12* do not affect ear development, nor do they enhance *ra3* (Claeys et al., 2019). Neither of the two paralogs is more similar in sequence to RA3, so to ask why *ZmTPP4*, and not *ZmTPP12,* acts as a redundant compensator, we queried their co-expression. In an aggregate network across 89 maize bulk tissue RNA-seq datasets (Lee et al., 2020), *RA3* and its two paralogs had similar co-expression scores (Figure S4A). In contrast, *ZmTPP4* was highly co-expressed with *RA3* in our single-cell data, similar to a *RA3-RA1*-positive control (Satoh-Nagasawa et al., 2006), whereas *ZmTPP12* was not (Figure S4B). Thus, functional redundancy in maize ear branching could be predicted by co-expression in scRNA-seq, but not in bulk tissue RNA-seq networks. To test this idea further, we identified a small gene family of *VASCULAR PLANT ONE-ZINC-FINGER* (*ZmVOZ*) genes, whose homologs regulate flowering time in *Arabidopsis* (Yasui et al., 2012). Two of them, *ZmVOZ4* and *ZmVOZ5*, exhibited highly similar co-expression within our scRNA-seq data (Spearman correlation 0.88, Figures S4C–S4H), and we identified a common set of high confidence genes that showed consistent co-expression with both genes (FDR < 0.05; Table S3). To ask whether these paralogs acted redundantly, we made CRISPR-Cas9 knockouts of all *ZmVOZ* members, including *ZmVOZ1* and *2* that were not detected in our scRNA-seq dataset, possibly due to their low expression in ears. As predicted, single *Zmvoz* mutants had no obvious phenotype, but *Zmvoz1,2,4,5* quadruple mutants were severely delayed in the floral transition, reminiscent of *voz1,2* double mutants in *Arabidopsis* (Figure 3A; Yasui et al., 2012). Therefore, these two examples highlight the utility of maize ear scRNA-seq data in predicting genetic redundancy.

Using scRNA-seq to build transcriptional regulatory networks

We next asked whether scRNA-seq might aid in building transcriptional regulatory networks, given that directly modulated targets of a transcription factor (TF) should be co-expressed in the same cell types. We used our scRNA-seq data to calculate co-expression of *KN1* with its published directly modulated targets (Table S3; Bolduc et al., 2012) and found that it was significantly higher than expected compared to a control using all maize genes ($p < 0.01$), supporting our hypothesis (Figure 3B). Thus, we next generated two additional ChIP-seq datasets, for ZmHOMEODOMAIN LEUCINE ZIPPER IV6 (ZmHDZIV6) (Javelle et al., 2011), which was uniquely expressed in the epidermis (Figures 3C and S4I), and ZmMADS16 (ZmM16) (Bartlett et al., 2015), which was expressed in specific floral organs (Figure 3D). Biological replicates for each TF ChIP-seq had significant overlap (Figure S4J; Table S3), and we identified 907 high-confidence peaks for ZmHDZIV6 and 1,155 for ZmM16 (Table S3). ~60% of these peaks mapped to gene regions, with a preference for promoters (Figures S4K and S4L), similar to other maize ChIP-seq studies (Bolduc et al., 2012). Members of the homeodomain leucine zipper IV family bind a GCAT TAAATGC consensus sequence (Nakamura et al., 2006), and we found a similar sequence in motif analysis of ZmHDZIV6 bound peaks (Figure 3E). Similarly, motif analysis of ZmM16 bound peaks found an expected MADS binding motif, $CC(A/T)_6GG$ (Figure 3F; Aerts et al., 2018). We were thus confident in our ZmHDZIV6 and ZmM16 bound target predictions (Table S3).

Modulated TF targets are often inferred by comparison of ChIP-seq bound targets and expression changes in mutant RNA-seq. However, since ZmHDZIV6 and ZmM16 are members of large gene families and mutant RNA-seq of them was not available, we asked whether we could predict modulated targets based on scRNA-seq co-expression with each TF (Table S3). We therefore identified 79 and 55 candidate modulated targets for ZmHDZIV6 and ZmM16, respectively, using a Jaccard index co-expression cutoff of ≥ 0.05 (Table S3). Among the predicted modulated targets of ZmHDZIV6, we identified five additional members of the *ZmHDZIV* family (Figures S4M and S4N; Table S3), some of which are similarly expressed in the maize SAM epidermis (Javelle et al., 2011) and might form transcriptional

Developmental Cell
Resource
CellPress

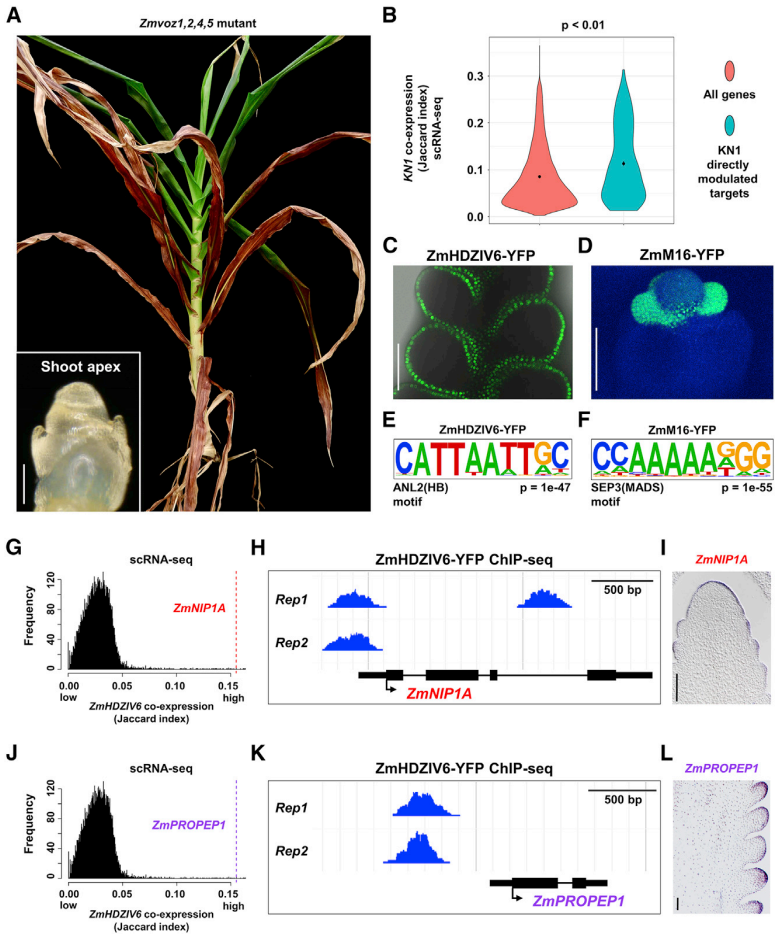


Figure 3. scRNA-seq can predict genetic redundancy and aid in predicting transcriptional regulatory networks

(A) Maize plant with CRISPR-Cas9 knockout of four *ZmVOZ* paralogs fails to transition to flowering, as shown by 2-month-old shoot apex (left bottom panel, scale bar = 100 μm) and a 6-month-old plant that lacks ears or tassel.

(B) Directly modulated transcriptional targets of *KN1* are significantly co-expressed with *KN1* at the single-cell level; all maize genes are used as control ($p < 0.01$, one-way ANOVA with Tukey's HSD).

(C and D) Expression of TF translational fusion lines, ZmHDZIV6-YFP (C, merge of YFP channel and bright field) and ZmM16-YFP (D, merge of YFP and DAPI channels), used for two biological replicates of ChIP-seq. Scale bar = 100 μm.

(E and F) Expected motifs are significantly over-represented in bound peaks of ZmHDZIV6 (E) or ZmM16 (F). p = 1e-47 (E) and p = 1e-55 (F).

(G and J) ZmHDZIV6 candidate modulated targets, *ZmNIP1A* (G) and *ZmPROPEP1* (J) are highly co-expressed with *ZmHDZIV6* in scRNA-seq (Jaccard index = 0.155 for both targets).

(H and K) ZmHDZIV6 bound peaks in *ZmNIPA1* (H) and *ZmPROPEP1* (K). Scale bar = 500 bp.

(I and L) *ZmNIP1A* (I) and *ZmPROPEP1* (L) are specifically expressed in the epidermis, by mRNA *in situ*. Scale bar = 100 μm.

cascades to regulate epidermal differentiation. We validated the epidermal expression of additional candidate modulated targets, including *ZmNOD26-LIKE MEMBRANE INTRINSIC PROTEIN1* (*ZmNIP1A*, *GRMZM2G041980*) and *ZmPRECURSOR ELICITOR PEPTIDE1* (*ZmPROPEP1*, *GRMZM5G899080*) (Figures 3G–3L). Homologs of these genes in *Arabidopsis* function as transporters (Liu et al., 2009), or in pathogen defense (Huffaker et al., 2011), suggesting similar roles in the maize ear epidermis. Similarly, among the 55 co-expressed ChIP-seq targets identified for ZmM16, we found additional members of the MADS family, such as *ZmSEP3/ZmMADS7* and *ZmAGAMOUS-LIKE 8* (*ZmAGL8*) (Figures S4O and S4P; Table S3), indicating that these genes might act downstream of ZmM16 to form gene regulatory networks controlling inflorescence development, analogous to MADS networks in *Arabidopsis* (Chen et al., 2018).

scRNA-seq identifies genes associated with maize yield traits

Maize ear morphology is associated with yield traits (Je et al., 2016; Liu et al., 2020). To ask whether the cell- or domain-specific genes identified in our scRNA-seq overlapped with candidate regulators of maize yield, we used a targeted GWAS approach, by comparing our scRNA-seq marker genes from meristem, determinate lateral organ, and vascular meta-clusters against a GWAS panel of 281 maize lines phenotyped for ear morphology traits related to yield (Figures 4A–4D; Table S3; Rice et al., 2020). Using SNPs in or within 2 kb of scRNA-seq marker genes, we found the meta-cluster 3 marker gene *ZmYABBY9* (*ZmYAB9*) had two significant SNPs (at 5% FDR) for cob weight (CW) (Figure 4B). We also found two significant SNPs (at 10% FDR) associated with ear diameter (ED), with minor additive effects (Figures 4C and 4D). One was associated with *GRMZM2G361210,* a marker of meristem branching related meta-clusters 9, 10, and 11 (Figure 4C), that encodes a C2H2-type zinc finger transcription factor related to *RAMOSA1* (*RA1*), a major player in maize domestication that controls branching and grain yield traits (Sigmon and Vollbrecht, 2010). A second significant SNP for ear diameter (ED) was associated with *ZmTMO5*, *GRMZM2G043854* (Figure 4D), a xylem meta-cluster marker (Figure 1J), whose *Arabidopsis* homolog controls periclinal cell divisions during vascular development (De Rybel et al., 2016).

We also conducted lambda analysis (Parvathaneni et al., 2020) to ask whether scRNA-seq marker SNPs were more significantly associated with yield traits compared to a random subset of maize genes (Table S3; see STAR methods). Using SNPs in or within 2 kb of genes (2 kb partition), we indeed found that ear diameter (ED) was significantly associated (Figures 4A and 4E; Table S3), suggesting that scRNA-seq marker genes preferentially control this trait (Figure 4E). Given that natural variation in distal regulatory elements also controls maize domestication

CellPress
Developmental Cell
Resource

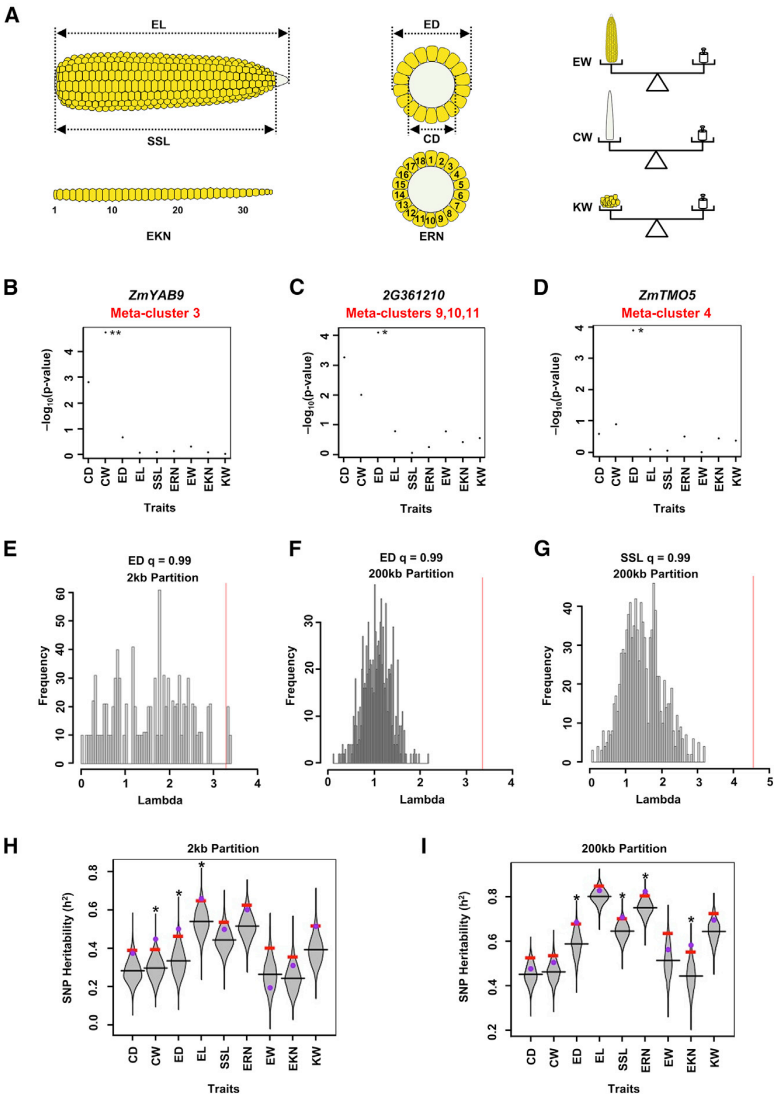


Figure 4. scRNA-seq marker genes are associated with maize ear traits

(A) Diagrams of nine different ear traits measured for GWAS analysis: ear length (EL), seed set length (SSL), ear rank number (EKN), ear diameter (ED), cob diameter (CD), ear row number (ERN), ear weight (EW), cob weight (CW), kernel weight (20 Seeds) (KW).

(B–D) Targeted GWAS using SNPs in or within 2 kb of genes reveals that scRNA-seq marker gene *ZmYAB9* has significant SNPs for ear weight (B), and two marker genes *GRMZM2G361210* (*2G361210*) (C) and *ZmTMO5* (D) have significant SNPs for ear diameter. ** FDR threshold of 0.05, * FDR threshold of 0.1. y axis indicates the $-\log_{10}$(p-value) (Table S3).

(E–G) Lambda values of scRNA-seq marker genes (red lines) are greater than two standard deviations from mean lambda values of 1,000 random gene sets (histogram distributions) for ear diameter (2 kb partition, E, 200 kb partition, F), and for seed set length trait (200 kb partition, G). Lambda values are reported in Table S3.

(H and I) Distributions of SNP heritability (h^2) using 2 kb (H) or 200 kb (I) partitions; h^2 values for scRNA-seq marker genes (purple dots) for the given traits (*) are greater than the top 5% permuted h^2 values (red bars) using 1,000 random subsets of maize genes (gray violin plots). h^2 values are reported in Table S3.

and yield traits (Liu et al., 2020), we also considered SNPs within 200 kb of gene coding regions (200 kb partition) and found again that ED was significantly associated (Figure 4F), as was an additional yield-related trait, seed set length (SSL) (Figures 4A and 4G).

Although lambda analysis detects significance for a target set of markers, it does not quantify the level of trait variability that is explained by those markers. Therefore, we next estimated narrow-sense heritability (h^2) for scRNA-seq markers, compared to a distribution of h^2 estimates from random subsets of markers (Table S3; see STAR methods). We found that SNP heritability for scRNA-seq markers was consistently greater than the 95th percentile of h^2 estimates from random markers for ED, CW, and SSL traits (Figures 4A, 4H, and 4I), validating our targeted GWAS and lambda analysis. Similar findings were obtained for ear length (EL) in the 2 kb partition (Figure 4H), as well as for ear row number (ERN) and ear rank number (EKN) in the 200 kb partition (Figure 4I).

To ask whether scRNA-seq data was required for these insights, we also calculated a list of "whole ear"-specific genes from bulk tissue RNA-seq data (Table S3; Walley et al., 2016) and found a low overlap with our scRNA-seq markers (4%, Table S3), indicating that the bulk tissue RNA-seq lacked cell- or developmental domain-specific information, as expected. We performed GWAS analysis using the whole ear-specific genes, and they were also enriched for association with ear morphology traits, but in most cases for different traits and genes (Table S3). For instance, this analysis identified *ZmYAB9*, that was also identified using scRNA-seq markers (Figure 4B). However, the ear diameter (ED) trait was associated with different genes in scRNA-seq (Figures 4C and 4D) and bulk tissue RNA-seq (Table S3) datasets. Furthermore, in heritability (h^2) analysis at a 200 kb partition, no traits were associated with whole ear-specific genes (Table S3), whereas four of them were significantly associated with scRNA-seq marker genes (Figure 4I). In summary, scRNA-seq markers revealed associations with multiple ear traits that were not found using bulk RNA-seq data, suggesting a unique application of this approach in identifying candidates to improve crop yield traits.

DISCUSSION

Development requires programmed cell- or domain-specific expression of regulatory and effector genes that together

Developmental Cell
Resource

orchestrate stereotypical patterns of morphogenesis. The maize ear has a complex morphology with multiple indeterminate meristem and determinate organ types, and optimization of morphology is important for maize yield. To identify spatial regulators of ear development, we performed scRNA-seq of ~12,500 single cells from developing ears and predicted 12 meta-clusters that were identified using known markers. As expected, many meta-clusters were meristem associated, including for discrete domains that control branching, and we also found distinct vasculature meta-clusters, including xylem, phloem, and bundle sheath. We identified meta-clusters from determinate lateral organs and ground tissues, and in several cases, sub-clusters could be identified. Our method was sensitive enough to detect most maize genes, though we failed to detect the expression of *CLV3* and *WUS* orthologs, possibly due to their low expression, or to a relatively low representation of central zone and organizing center cells in our experiments using whole developing ears. Another possibility is that these cell types were not recovered with our current protocol, and further improvements or profiling of cells from more finely dissected meristem tissues may address this issue. It is also intriguing that some xylem markers were expressed in the tips of meristems. In *Arabidopsis*, *class III HD-ZIP* and *KANADI* genes, which are expressed in xylem and phloem, respectively, are expressed in complementary patterns in adaxial and abaxial sides of lateral organs to specify their polarity (Emery et al., 2003). This polarity is pre-patterned by their corresponding central and peripheral expression in the shoot meristem (Caggiano et al., 2017). Thus, our finding of maize ear xylem markers expressed in the meristem tip suggests that additional vascular genes specify a pre-pattern in meristems.

We validated our scRNA-seq results by mRNA *in situ* hybridization and by comparing to a FACS RNA-seq dataset, thus constructing a robust single-cell transcriptome atlas of a developing inflorescence. We also provided three applications showing how this atlas can enhance functional studies. First, we highlight how the cellular resolution of scRNA-seq data can accurately predict redundancy, a major obstacle in genetic analyses. We could predict redundancy in a family of maize TPPs that control inflorescence branching, and we identified a family of redundant maize *VOZ* genes that produced a delayed floral transition phenotype through multiplex CRISPR-Cas9 mutagenesis. Flowering time is a major target of maize breeding (Liu et al., 2020), so our findings provide candidates to fine tune flowering for crop improvement. In some cases, redundant paralogs may show only partial co-expression, as observed for SQUAMOSA PROMOTER BINDING (SBP)-box transcription factors UNBRANCHED2 (UB2) and UB3 which control initiation of lateral organ (Chuck et al., 2014; Du et al., 2020), or *AINTEGUMENTA* (*ANT*) and *ANT-LIKE6*, whose expression overlaps partially to redundantly regulate floral organ patterning and growth (Krizek, 2009). Therefore, care should be taken in use of scRNA-seq data in discerning such partially overlapping expression patterns.

In a second example, we hypothesized that the resolution of scRNA-seq could be combined with ChIP-seq to predict directly modulated targets of TFs. We created ChIP-seq datasets for two TFs that are likely to act redundantly, precluding the use of single mutants to find their modulated targets. We also integrated scRNA-seq with FACS ATAC-seq to provide evidence of spatially regulated accessible chromatin. The limited availability of maize reporter lines for FACS may be overcome in the future by application of single-cell ATAC-seq (Rich-Griffin et al., 2020). Lastly, we hypothesized that scRNA-seq marker genes with spatially restricted expression in developing ears are enriched for regulators of ear morphology traits important for crop yields. Indeed, the scRNA-seq marker genes were significantly associated with ear morphology trait SNPs in a GWAS panel. These marker candidates could be selected in breeding programs or genetically modified to test their effects on yield.

In summary, scRNA-seq allowed valuable insights into maize ear development. The atlas can inform developmental genetics studies and breeding, and the methods we developed can be applied to studies of other complex shoot systems. As more plant scRNA-seq datasets are generated, a cross-species (e.g., between maize and *Arabidopsis*) or cross-tissue (e.g., between shoot and root) comparative analysis at single-cell resolution will inform how gene signatures were selected during evolution to shape the diverse morphologies that are critical to reproductive success and agricultural production.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Protoplast preparation and 10x Genomics library construction and sequencing
  - Anatomy and confocal microscopy
  - mRNA *in situ* hybridization
  - FACS and bulk RNA-seq, ATAC-seq library preparation
  - ChIP-seq library preparation
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - scRNA-seq analysis, clustering, and selection of marker genes
  - FACS and bulk RNA-seq, ATAC-seq analysis
  - ChIP-seq analysis
  - scRNA-seq co-expression analysis
  - Integration of GWAS with scRNA-seq or ear tissue bulk RNA-seq

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.devcel.2020.12.015.

ACKNOWLEDGMENTS

We thank Drs. Kenneth Birnbaum, Carlos Ortiz Ramírez, and Bruno Guillotin for advice on scRNA-seq and FACS RNA-seq experiments, and Drs. Bert De Rybel and Joyce Chery for discussions on vascular biology. We thank support from Dr. Jon Preall and the CSHL single-cell core facility, flow cytometry facility, microscopy facility, NGS facility, and farm management team. We thank

CellPress

Developmental Cell
Resource

Cassidy Danyko for assisting in FACS RNA-seq library preparation, Richelle Chen for assisting in bioinformatic analyses, and Michael Okoro for assisting in field work and sample preparation. We acknowledge funding support from NSF (IOS-1833182, IOS-1445025, IOS-1027445, IOS-1934388, IOS-1755141). R.J.S. acknowledges support from NSF (IOS-1856627). A.E.L. acknowledges support from NSF (IOS-1733606). J.G. acknowledges support from NIH R01 LM012736 and R01 MH113005. M.C. acknowledges support from K99 MH120050. B.H. acknowledges support from The Robertson Research Fund, CSHL School for Biological Sciences. We thank Drs. Munenori Kitagawa and Penelope Lindsay for their feedback on the manuscript.

AUTHOR CONTRIBUTIONS

X.X. performed experimental procedures and data analysis, except for those listed below, and wrote the draft of the manuscript. M.C., B.H., and N.F. performed scRNA-seq analysis. B.R.R. performed GWAS analysis. F.L. performed ChIP-seq analysis and assisted ATAC-seq analysis. L.L. performed FACS RNA-seq analysis and assisted in mRNA *in situ*. E.D.-A. and S.N.C. generated *Zmvoz* CRISPR mutants. Z.L. assisted preparation of FACS ATAC-seq library and performed analysis. L.W. assisted both FACS RNA-seq and ChIP-seq analyses. X.W. assisted ChIP-seq analysis. J.D. assisted preparation of FACS RNA-seq libraries. A.L. generated ZmHDZIV6-YFP transgenic lines. D.J., J.G., D.W., A.E.L., R.J.S., T.R.G., A.W.S., and B.Y. supervised the research. D.J. co-wrote the manuscript, and all authors edited.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: September 4, 2020
Revised: October 31, 2020
Accepted: December 15, 2020
Published: January 4, 2021

REFERENCES

Aerts, N., de Bruijn, S., van Mourik, H., Angenent, G.C., and van Dijk, A.D.J. (2018). Comparative analysis of binding patterns of MADS-domain proteins in Arabidopsis thaliana. BMC Plant Biol. *18*, 131–131.

Ballouz, S., Weber, M., Pavlidis, P., and Gillis, J. (2017). EGAD: ultra-fast functional analysis of gene networks. Bioinformatics *33*, 612–614.

Bartlett, M.E., Williams, S.K., Taylor, Z., DeBlasio, S., Goldshmidt, A., Hall, D.H., Schmidt, R.J., Jackson, D.P., and Whipple, C.J. (2015). The Maize PI/GLO Ortholog Zmm16/sterile tassel silky ear1 Interacts with the Zygomorphy and Sex Determination Pathways in Flower Development. Plant Cell *27*, 3081–3098.

Bolduc, N., Yilmaz, A., Mejia-Guerra, M.K., Morohashi, K., O'Connor, D., Grotewold, E., and Hake, S. (2012). Unraveling the KNOTTED1 regulatory network in maize meristems. Genes Dev. *26*, 1685–1690.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

Bommert, P., Nagasawa, N.S., and Jackson, D. (2013). Quantitative variation in maize kernel row number is controlled by the FASCIATED EAR2 locus. Nat. Genet. *45*, 334–337.

Caggiano, M.P., Yu, X., Bhatia, N., Larsson, A., Ram, H., Ohno, C.K., Sappl, P., Meyerowitz, E.M., Jönsson, H., and Heisler, M.G. (2017). Cell type boundaries organize plant development. eLife *6*, e27421.

Chang, Y.-M., Liu, W.-Y., Shih, A.C.-C., Shen, M.-N., Lu, C.-H., Lu, M.-Y.J., Yang, H.-W., Wang, T.-Y., Chen, S.C.C., Chen, S.M., et al. (2012). Characterizing regulatory and functional differentiation between maize mesophyll and bundle sheath cells by transcriptomic analysis. Plant Physiol. *160*, 165–177.

Chen, H., and Xiong, L. (2010). myo-Inositol-1-phosphate synthase is required for polar auxin transport and organ development. J. Biol. Chem. *285*, 24238–24247.

Chen, D., Yan, W., Fu, L.-Y., and Kaufmann, K. (2018). Architecture of gene regulatory networks controlling flower development in *Arabidopsis thaliana*. Nat. Commun. *9*, 4534.

Chitwood, D.H., Guo, M., Nogueira, F.T.S., and Timmermans, M.C.P. (2007). Establishing leaf polarity: the role of small RNAs and positional signals in the shoot apex. Development *134*, 813–823.

Chuck, G., Muszynski, M., Kellogg, E., Hake, S., and Schmidt, R.J. (2002). The control of spikelet meristem identity by the branched silkless1 gene in maize. Science *298*, 1238–1241.

Chuck, G.S., Brown, P.J., Meeley, R., and Hake, S. (2014). Maize SBP-box transcription factors unbranched2 and unbranched3 affect yield traits by regulating the rate of lateral primordia initiation. Proc. Natl. Acad. Sci. USA *111*, 18775–18780.

Claeys, H., Vi, S.L., Xu, X., Satoh-Nagasawa, N., Eveland, A.L., Goldshmidt, A., Feil, R., Beggs, G.A., Sakai, H., Brennan, R.G., et al. (2019). Control of meristem determinacy by trehalose 6-phosphate phosphatases is uncoupled from enzymatic activity. Nat. Plants *5*, 352–357.

Crow, M., Paul, A., Ballouz, S., Huang, Z.J., and Gillis, J. (2018). Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor. Nat. Commun. *9*, 884.

Csardi, G., and Nepusz, T. (2006). "The igraph software package for complex network research." InterJournal, Complex Systems, 1695. https://igraph.org.

De Rybel, B., Mähönen, A.P., Helariutta, Y., and Weijers, D. (2016). Plant vascular development: from early specification to differentiation. Nat. Rev. Mol. Cell Biol. *17*, 30–40.

Denyer, T., Ma, X., Klesen, S., Scacchi, E., Nieselt, K., and Timmermans, M.C.P. (2019). Spatiotemporal developmental trajectories in the *Arabidopsis* root revealed using high-throughput single-cell RNA sequencing. Dev. Cell *48*, 840–852.e5.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21.

Dong, J., Kim, S.T., and Lord, E.M. (2005). Plantacyanin plays a role in reproduction in *Arabidopsis*. Plant Physiol. *138*, 778–789.

Du, Y., Liu, L., Peng, Y., Li, M., Li, Y., Liu, D., Li, X., and Zhang, Z. (2020). UNBRANCHED3 Expression and Inflorescence Development is Mediated by UNBRANCHED2 and the Distal Enhancer, KRN4, in Maize. PLoS Genet. *16*, e1008764.

Emery, J.F., Floyd, S.K., Alvarez, J., Eshed, Y., Hawker, N.P., Izhaki, A., Baum, S.F., and Bowman, J.L. (2003). Radial patterning of *Arabidopsis* shoots by class III HD-ZIP and KANADI genes. Curr. Biol. *13*, 1768–1774.

Erichson, N.B., Voronin, S., Brunton, S.L., and Kutz, J.N. (2019). Randomized Matrix Decompositions Using R. J. Stat. Softw. *89*, 1–48, https://doi.org/10.18637/jss.v089.i11.

Eveland, A.L., Goldshmidt, A., Pautler, M., Morohashi, K., Liseron-Monfils, C., Lewis, M.W., Kumari, S., Hiraga, S., Yang, F., Unger-Wallace, E., et al. (2014). Regulatory modules controlling maize inflorescence architecture. Genome Res. *24*, 431–443.

Fujimoto, M., Sazuka, T., Oda, Y., Kawahigashi, H., Wu, J., Takanashi, H., Ohnishi, T., Yoneda, J.I., Ishimori, M., Kajiya-Kanegae, H., et al. (2018). Transcriptional switch for programmed cell death in pith parenchyma of sorghum stems. Proc. Natl. Acad. Sci. USA *115*, E8783–E8792.

Funk, V., Kositsup, B., Zhao, C., and Beers, E.P. (2002). The *Arabidopsis* xylem peptidase XCP1 is a tracheary element vacuolar protein that may be a papain ortholog. Plant Physiol. *128*, 84–94.

Gallavotti, A., Zhao, Q., Kyozuka, J., Meeley, R.B., Ritter, M.K., Doebley, J.F., Pè, M.E., and Schmidt, R.J. (2004). The role of barren stalk1 in the architecture of maize. Nature *432*, 630–635.

Gipson, A.B., Giloteaux, L., Hanson, M.R., and Bentolila, S. (2020). *Arabidopsis* RanBP2-Type Zinc Finger Proteins Related to Chloroplast RNA Editing Factor OZ1. Plants (Basel) *9*, 307.

Guo, P., Zheng, Y., Peng, D., Liu, L., Dai, L., Chen, C., and Wang, B. (2018). Identification and expression characterization of the Phloem Protein 2 (PP2) genes in ramie (Boehmeria nivea L. Gaudich). Sci. Rep. *8*, 10734.

Developmental Cell
Resource
CellPress

Ha, C.M., Jun, J.H., Nam, H.G., and Fletcher, J.C. (2007). BLADE-ON-PETIOLE 1 and 2 control *Arabidopsis* lateral organ fate through regulation of LOB domain and adaxial-abaxial polarity genes. Plant Cell *19*, 1809–1825.

Han, J.-J., Jackson, D., and Martienssen, R. (2012). Pod corn is caused by rearrangement at the Tunicate1 locus. Plant Cell *24*, 2733–2744.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Herzeel, C., Costanza, P., Decap, D., Fostier, J., and Reumers, J. (2015). elPrep: High-Performance Preparation of Sequence Alignment/Map Files for Variant Calling. PLoS ONE *10*, e0132868.

Horváth, E., Bela, K., Holinka, B., Riyazuddin, R., Gallé, Á., Hajnal, Á., Hurton, Á., Fehér, A., and Csiszár, J. (2019). The *Arabidopsis* glutathione transferases, AtGSTF8 and AtGSTU19 are involved in the maintenance of root redox homeostasis affecting meristem size and salt stress sensitivity. Plant Sci. *283*, 366–374.

Huffaker, A., Dafoe, N.J., and Schmelz, E.A. (2011). ZmPep1, an ortholog of *Arabidopsis* elicitor peptide 1, regulates maize innate immunity and enhances disease resistance. Plant Physiol. *155*, 1325–1338.

Irish, E. (1997). Class II tassel seed mutations provide evidence for multiple types of inflorescence meristems in maize (Poaceae). Am. J. Bot. *84*, 1502–1515.

Jackson, D., Veit, B., and Hake, S. (1994). Expression of maize KNOTTED1 related homeobox genes in the shoot apical meristem predicts patterns of morphogenesis in the vegetative shoot. Development *120*, 405–413.

Javelle, M., Klein-Cosson, C., Vernoud, V., Boltz, V., Maher, C., Timmermans, M., Depège-Fargeix, N., and Rogowsky, P.M. (2011). Genome-wide characterization of the HD-ZIP IV transcription factor family in maize: preferential expression in the epidermis. Plant Physiol. *157*, 790–803.

Je, B.I., Gruel, J., Lee, Y.K., Bommert, P., Arevalo, E.D., Eveland, A.L., Wu, Q., Goldshmidt, A., Meeley, R., Bartlett, M., et al. (2016). Signaling from maize organ primordia via FASCIATED EAR3 regulates stem cell proliferation and yield traits. Nat. Genet. *48*, 785–791.

Knauer, S., Javelle, M., Li, L., Li, X., Ma, X., Wimalanathan, K., Kumari, S., Johnston, R., Leiboff, S., Meeley, R., et al. (2019). A high-resolution gene expression atlas links dedicated meristem genes to key architectural traits. Genome Res. *29*, 1962–1973.

Konopka, T. (2020). umap: Uniform Manifold Approximation and Projection. R package version 0.2.6.0. https://CRAN.R-project.org/package=umap.

Krizek, B. (2009). AINTEGUMENTA and AINTEGUMENTA-LIKE6 act redundantly to regulate *Arabidopsis* floral growth and patterning. Plant Physiol. *150*, 1916–1929.

Kulkarni, A., Anderson, A.G., Merullo, D.P., and Konopka, G. (2019). Beyond bulk: a review of single cell transcriptomics methodologies and applications. Curr. Opin. Biotechnol. *58*, 129–136.

Lee, J., Shah, M., Ballouz, S., Crow, M., and Gillis, J. (2020). CoCoCoNet: conserved and comparative co-expression across a diverse set of species. Nucleic Acids Res. *48* (W1), W566–W571.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760.

Liu, Q., Wang, H., Zhang, Z., Wu, J., Feng, Y., and Zhu, Z. (2009). Divergence in function and expression of the NOD26-like intrinsic proteins in plants. BMC Genomics *10*, 313.

Liu, J., Fernie, A.R., and Yan, J. (2020). The Past, Present, and Future of Maize Improvement: Domestication, Genomics, and Functional Genomic Routes toward Crop Enhancement. Plant Communications *1*, 100010.

Lloyd, J., and Meinke, D. (2012). A comprehensive dataset of genes with a loss-of-function mutant phenotype in *Arabidopsis*. Plant Physiol. *158*, 1115–1129.

Lu, Z., Hofmeister, B.T., Vollmers, C., DuBois, R.M., and Schmitz, R.J. (2017). Combining ATAC-seq with nuclei sorting for discovery of cis-regulatory regions in plant genomes. Nucleic Acids Res. *45*, e41.

Lun, A.T., McCarthy, D.J., and Marioni, J.C. (2016). A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. *5*, 2122.

Lun, A.T.L., Riesenfeld, S., Andrews, T., Dao, T.P., Gomes, T., and Marioni, J.C.; participants in the 1st Human Cell Atlas Jamboree (2019). EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. Genome Biol. *20*, 63.

Ma, X., Denyer, T., and Timmermans, M.C. (2020). PscB: A Browser to Explore Plant Single Cell RNA-Sequencing Datasets. Plant Physiology, pp.00250.02020.

Marchette, D.J. (2015). cccd: Class Cover Catch Digraphs. R package version 1.5. https://CRAN.R-project.org/package=cccd.

McCarthy, D.J., Campbell, K.R., Lun, A.T., and Wills, Q.F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinformatics *33*, 1179–1186.

McGinnis, C.S., Murrow, L.M., and Gartner, Z.J. (2019). DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors. Cell Syst. *8*, 329–337.e4.

Miyashima, S., Roszak, P., Sevilem, I., Toyokura, K., Blob, B., Heo, J.O., Mellor, N., Help-Rinta-Rahko, H., Otero, S., Smet, W., et al. (2019). Mobile PEAR transcription factors integrate positional cues to prime cambial growth. Nature *565*, 490–494.

Nakamura, M., Katsumata, H., Abe, M., Yabe, N., Komeda, Y., Yamamoto, K.T., and Takahashi, T. (2006). Characterization of the class IV homeodomain-Leucine Zipper gene family in *Arabidopsis*. Plant Physiol. *141*, 1363–1375.

Nuccio, M.L., Wu, J., Mowers, R., Zhou, H.P., Meghji, M., Primavesi, L.F., Paul, M.J., Chen, X., Gao, Y., Haque, E., et al. (2015). Expression of trehalose-6-phosphate phosphatase in maize ears improves yield in well-watered and drought conditions. Nat. Biotechnol. *33*, 862–869.

Ortiz-Ramírez, C., Arevalo, E.D., Xu, X., Jackson, D.P., and Birnbaum, K.D. (2018). An Efficient Cell Sorting Protocol for Maize Protoplasts. Curr. Protoc. Plant Biol. *3*, e20072.

Parvathaneni, R.K., Bertolini, E., Shamimuzzaman, M., Vera, D.L., Lung, P.-Y., Rice, B.R., Zhang, J., Brown, P.J., Lipka, A.E., Bass, H.W., and Eveland, A.L. (2020). The regulatory landscape of early maize inflorescence development. Genome Biol. *21*, 165.

Pautler, M., Eveland, A.L., LaRue, T., Yang, F., Weeks, R., Lunde, C., Je, B.I., Meeley, R., Komatsu, M., Vollbrecht, E., et al. (2015). FASCIATED EAR4 encodes a bZIP transcription factor that regulates shoot meristem size in maize. Plant Cell *27*, 104–120.

Potter, S.S. (2018). Single-cell RNA sequencing for the study of development, physiology and disease. Nat. Rev. Nephrol. *14*, 479–492.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing (Vienna, Austria). http://www.R-project.org/.

Ricci, W.A., Lu, Z., Ji, L., Marand, A.P., Ethridge, C.L., Murphy, N.G., Noshay, J.M., Galli, M., Mejía-Guerra, M.K., Colomé-Tatché, M., et al. (2019). Widespread long-range cis-regulatory elements in the maize genome. Nat. Plants *5*, 1237–1249.

Rice, B.R., Fernandes, S.B., and Lipka, A.E. (2020). Multi-Trait Genome-Wide Association Studies Reveal Loci Associated with Maize Inflorescence and Leaf Architecture. Plant Cell Physiol. *61*, 1427–1437.

Rich-Griffin, C., Stechemesser, A., Finch, J., Lucas, E., Ott, S., and Schäfer, P. (2020). Single-Cell Transcriptomics: A High-Resolution Avenue for Plant Functional Genomics. Trends Plant Sci. *25*, 186–197.

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

Rodriguez-Leal, D., Xu, C., Kwon, C.-T., Soyars, C., Demesa-Arevalo, E., Man, J., Liu, L., Lemmon, Z.H., Jones, D.S., Van Eck, J., et al. (2019). Evolution of buffering in a genetic circuit controlling plant stem cell proliferation. Nat. Genet. *51*, 786–792.

CellPress

Developmental Cell
Resource

Rosvall, M., and Bergstrom, C.T. (2008). Maps of random walks on complex networks reveal community structure. Proc. Natl. Acad. Sci. USA *105*, 1118–1123.

Satoh-Nagasawa, N., Nagasawa, N., Malcomber, S., Sakai, H., and Jackson, D. (2006). A trehalose metabolic enzyme controls inflorescence architecture in maize. Nature *441*, 227–230.

Schürholz, A.-K., López-Salmerón, V., Li, Z., Forner, J., Wenzl, C., Gaillochet, C., Augustin, S., Barro, A.V., Fuchs, M., Gebert, M., et al. (2018). A Comprehensive Toolkit for Inducible, Cell Type-Specific Gene Expression in *Arabidopsis*. Plant Physiol. *178*, 40–53.

Sigmon, B., and Vollbrecht, E. (2010). Evidence of selection at the ramosa1 locus during maize domestication. Mol. Ecol. *19*, 1296–1311.

Speed, D., Hemani, G., Johnson, M.R., and Balding, D.J. (2012). Improved heritability estimation from genome-wide SNPs. Am. J. Hum. Genet. *91*, 1011–1021.

Strable, J., Wallace, J.G., Unger-Wallace, E., Briggs, S., Bradbury, P.J., Buckler, E.S., and Vollbrecht, E. (2017). Maize *YABBY* Genes *drooping leaf1* and *drooping leaf2* Regulate Plant Architecture. Plant Cell *29*, 1622–1641.

Takacs, E.M., Li, J., Du, C., Ponnala, L., Janick-Buckner, D., Yu, J., Muehlbauer, G.J., Schnable, P.S., Timmermans, M.C., Sun, Q., et al. (2012). Ontogeny of the maize shoot apical meristem. Plant Cell *24*, 3219–3234.

Tong, W., Imai, A., Tabata, R., Shigenobu, S., Yamaguchi, K., Yamada, M., Hasebe, M., Sawa, S., Motose, H., and Takahashi, T. (2016). Polyamine Resistance Is Increased by Mutations in a Nitrate Transporter Gene NRT1.3 (AtNPF6.4) in *Arabidopsis thaliana*. Front. Plant Sci. *7*, 834.

Vollbrecht, E., and Schmidt, R.J. (2009). Development of the Inflorescences. In Handbook of Maize: Its Biology, J.L. Bennetzen and S.C. Hake, eds. (New York, NY: Springer), pp. 13–40.

Walley, J.W., Sartor, R.C., Shen, Z., Schmitz, R.J., Wu, K.J., Urich, M.A., Nery, J.R., Smith, L.G., Schnable, J.C., Ecker, J.R., and Briggs, S.P. (2016). Integration of omic networks in a developmental atlas of maize. Science *353*, 814–818.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80–84.

Wang, L., Lu, Z., delaBastide, M., Van Buren, P., Wang, X., Ghiban, C., Regulski, M., Drenkow, J., Xu, X., Ortiz-Ramirez, C., et al. (2020). Management, Analyses, and Distribution of the MaizeCODE Data on the Cloud. Front. Plant Sci. *11*, 289.

Wu, Q., Xu, F., Liu, L., Char, S.N., Ding, Y., Je, B.I., Schmelz, E., Yang, B., and Jackson, D. (2020). The maize heterotrimeric G protein β subunit controls shoot meristem development and immune responses. Proc. Natl. Acad. Sci. USA *117*, 1799–1805.

Yasui, Y., Mukougawa, K., Uemoto, M., Yokofuji, A., Suzuri, R., Nishitani, A., and Kohchi, T. (2012). The phytochrome-interacting vascular plant one-zinc finger1 and VOZ2 redundantly regulate flowering in *Arabidopsis*. Plant Cell *24*, 3248–3263.

Zhong, R., Richardson, E.A., and Ye, Z.-H. (2007). The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in *Arabidopsis*. Plant Cell *19*, 2776–2792.

Developmental Cell
Resource

CellPress

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-GFP antibody (GFP-Trap magnetic agarose)	ChromoTek	Cat# gtma-20; RRID: AB_2631358
Bacterial and virus strains		
Agrobacterium	N/A	EHA101
E.coli	N/A	DH5α
Biological samples		
Zea mays B73	Maize Genetics COOP Stock Center	N/A
Zea mays pZmYAB14-TagRFPt reporter line	Je et al., 2016	N/A
Zea mays ZmM16-YFP translational fusion line	Bartlett et al., 2015	N/A
Zea mays ZmHDZIV6-YFP translational fusion line	This paper	N/A
Zea mays CRISPR/Cas9 knock out mutants of *ZmVOZs*	This paper	N/A
Zea mays branched silkless1;*Tunicate* (*bd1*;*Tu*) mutants	This paper	N/A
Chemicals, peptides, and recombinant proteins		
Mannitol	Sigma-Aldrich	Cat# M4125
Bovine serum albumin	Sigma-Aldrich	Cat# A7907-50G
T7 RNA polymerase	Sigma-Aldrich	Cat# 10881775001
Calcofluor white stain	Sigma-Aldrich	Cat# 18909
Toluidine blue	Sigma-Aldrich	Cat# T3260
CelLytic™ PN Isolation/Extraction Kit	Sigma-Aldrich	Cat# CELLYTPN1
Paraformaldehyde	Electron Microscopy Sciences	Cat# 15714 s
Glutaraldehyde	Electron Microscopy Sciences	Cat# 16537-16
Cacodylate buffer	Electron Microscopy Sciences	Cat# 11652
LR white resin	Electron Microscopy Sciences	Cat# 905072
Cellulase RS	Onozuka	N/A
Cellulase R-10	Onozuka	N/A
Macerozyme R-10	Onozuka	N/A
Pectolyase Y-23	Duchefa Biochem.	Cat# P8OO4.0001
Trypan blue	Thermo Fisher Scientific	Cat# 15250061
Paraplast	McCormick Scientific	Cat# 39503002
ProbeOn Plus™ Slides	Fisher Scientific	Cat# 22-230-900
NBT/BCIP Ready-to-Use Tablets	Roche	Cat# 11697471001
Critical commercial assays		
Chromium i7 Multiplex Kit	10X Genomics	Cat# PN-120262
Chromium Single Cell 3¢ Library & Gel Bead Kit v2	10X Genomics	Cat# PN-120237
Chromium Single Cell A Chip Kit v2	10X Genomics	Cat# PN-1000009
Dynabeads™ MyOne™ Silane Beads	Thermo Fisher Scientific	Cat# 37002D
Arcturus PicoPure RNA Isolation Kit	Thermo Fisher Scientific	Cat# KIT0204
RNA Bioanalyzer kit	Agilent	Cat# 5067-1513

(*Continued on next page*)

CellPress
Developmental Cell
Resource

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
DNA High Sensitivity Bioanalyzer kit	Agilent	Cat# 5067-4626
SMART-Seq™ v4 Ultra™ Low Input RNA Kit	Takara Bio USA, Inc.	Cat# 634890
Nextera XT DNA Library Prep Kit	Illumina	Cat# FC-131-1024
AMPure XP Beads	Beckman Coulter	Cat# A63880
KAPA Library Quantification Kits	Roche	Cat# KK4824
NEXTflex ChIP-seq Kit	PerkinElmer Applied Genomics	Cat# NOVA-5143-02
Deposited data		
B73 whole ear scRNA-seq_replicate 1	This paper	PRJNA646989
B73 whole ear scRNA-seq_replicate 2	This paper	PRJNA646996
B73 whole ear scRNA-seq_replicate 3	This paper	PRJNA647001
Protoplasting-response bulk RNA-seq	This paper	PRJNA647196
pZmYAB14-TagRFPt FACS RNA-seq	This paper	PRJNA647195
pZmYAB14-TagRFPt FACS ATAC-seq	This paper	PRJNA647197
ZmHDZIV6-YFP ChIP-seq	This paper	PRJNA647198
ZmM16-YFP ChIP-seq	This paper	PRJNA647200
Oligonucleotides		
See Table S4 for primers or sgRNAs sequences for mRNA *in situ*, *ZmVOZs* crispr and genotyping, ZmHDZIV6-YFP transgene, and *bd1* genotyping.	N/A	N/A
Recombinant DNA		
pGW-Cas9 vector	Wang et al., 2014	Addgene Plasmid # 50661; RRID: Addgene_50661
pTF101 Gateway-compatible vector	Je et al., 2016	N/A
Software and algorithms		
STAR	Dobin et al., 2013	https://github.com/alexdobin/STAR/wiki
R	R Core Team, 2013	https://www.r-project.org/
EmptyDrops	Lun et al., 2019	https://rdrr.io/github/MarioniLab/DropletUtils/man/emptyDrops.html
DoubletFinder	McGinnis et al., 2019	https://github.com/chris-mcginnis-ucsf/DoubletFinder
Scater	McCarthy et al., 2017	https://github.com/Alanocallaghan/scater
Scran	Lun et al., 2016	https://github.com/MarioniLab/scran
Rsvd package	Erichson et al., 2019	https://github.com/erichson/rSVD
Cccd package	Marchette, 2015	https://github.com/cran/cccd
InfoMap	Csardi and Nepusz, 2006	https://github.com/mapequation/infomap
MetaNeighbor	Crow et al., 2018	https://github.com/maggiecrow/MetaNeighbor
UMAP package	Konopka, 2020	https://github.com/tkonopka/umap
EGAD	Ballouz et al., 2017	https://github.com/sarbal/EGAD
Trimmomatic - version 0.36	Bolger et al., 2014	http://www.usadellab.org/cms/?page=trimmomatic
edgeR	Robinson et al., 2010	https://bioconductor.riken.jp/packages/devel/bioc/html/edgeR.html
elprep	Herzeel et al., 2015	https://github.com/ExaScience/elprep
BWA-MEM	Li and Durbin, 2009	https://github.com/lh3/bwa
HOMER	Heinz et al., 2010	http://homer.ucsd.edu/homer/
LDAK software v5.0	Speed et al., 2012	http://www.ldak.org
Other		
Cell Strainer	pluriStrainer	Cat# 43-50030-50

Developmental Cell
Resource

CellPress

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, David Jackson (jacksond@cshl.edu).

Materials availability

Requests for materials should be directed to Lead Contact, David Jackson (jacksond@cshl.edu). Requests for transgenic plant materials will require a Materials Transfer Agreement (MTA).

Data and code availability

The accession numbers for the raw scRNA-seq data reported in this paper are NCBI's Sequence Read Archive (SRA) BioProjects: PRJNA646989, PRJNA646996, and PRJNA647001. The accession numbers for the raw protoplasting-response bulk RNA-seq data reported in this paper is NCBI's SRA BioProject: PRJNA647196. The accession numbers for the raw *pZmYAB14-TagRFPt* FACS RNA-seq data reported in this paper is NCBI's SRA BioProject: PRJNA647195. The accession numbers for the raw *pZmYAB14-TagRFPt* FACS ATAC-seq data reported in this paper is NCBI's SRA BioProject: PRJNA647197. The accession numbers for the raw ZmHDZIV6-YFP ChIP-seq data reported in this paper is NCBI's SRA BioProject: PRJNA647198. The accession numbers for the raw ZmM16-YFP ChIP-seq data reported in this paper is NCBI's SRA BioProject: PRJNA647200. SRA BioProject IDs were also listed in Key resources table. This study used codes from published software described in Quantification and statistical analysis.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All analyses were performed with *Zea mays* (Maize). Maize plants were grown in the summer field (June – October) of Uplands Farm Agricultural Station at Cold Spring Harbor, New York or in the greenhouse with 16 h daytime, 26-28°C, and 8 h night, 22-24°C (Wu et al., 2020). Reference B73 inbred plants were used for single-cell experiments. The *pZmYAB14-TagRFPt* reporter line (Je et al., 2016) and ZmM16-YFP translational fusion line (Bartlett et al., 2015) were obtained from previous studies. The ZmHDZIV6-YFP translational fusion line was constructed using *ZmHDZIV6* native promoter and coding sequence as previously described in the pTF101 Gateway-compatible vector (Je et al., 2016), primer sequences were listed in Table S4. The *pZmYAB14-TagRFPt* reporter line and ZmHDZIV6-YFP translational fusion line were introgressed into a proliferative cauliflower-like double mutant line, *bd1;Tu*, to generate a large amount of ear meristem tissue for FACS RNA-seq and ChIP-seq respectively. Primer sequences for genotyping *bd1* are listed in Table S4. *Tu* genotyping was performed as previously described (Han et al., 2012).

CRISPR/Cas9 was used to knockout *ZmVOZs* genes following *Agrobacterium*-mediated transformation of Hi-II embryos (Je et al., 2016). Guide RNAs (sgRNAs) were designed based on B73 V3 reference genome, one pair targeting *ZmVOZ1* and *ZmVOZ2*, and a second pair targeting *ZmVOZ4* and *ZmVOZ5*, Table S4. The sgRNAs were introduced by Gateway Recombination into pGW-Cas9 vector (Addgene plasmid # 50661; RRID: Addgene_50661) (Wang et al., 2014) and transferred to *Agrobacterium* (EHA101) for maize transformation. 29 plants from 8 transformation events were obtained and analyzed by PCR amplification and Sanger sequencing to identify mutations in the targeted regions. The Cas9 transgene was segregated away by crossing with B73 to recover stable mutant alleles. Primer sequences and PCR assays for genotyping were listed in Table S4.

METHOD DETAILS

Protoplast preparation and 10x Genomics library construction and sequencing

For the three biological replicates of wild type (B73) whole ear samples, protoplasts were prepared as previously described but without L-cysteine pretreatment (Ortiz-Ramírez et al., 2018). 5-10mm developing ears were dissected into protoplast washing buffer and diced with a razor blade to 0.5-1mm pieces, then washed three times with protoplast washing buffer before adding enzyme solution. Tissues were protoplasted for ~45min at room temperature with gentle shaking. The mixture was filtered through a 30μm cell strainer (pluriStrainer, 43-50030-50) then collected by centrifugation at 500 g for 3 min at 4°C. The supernatant was gently removed without disturbing the protoplast pellet, which was washed by gentle resuspension in protoplast washing buffer. Protoplasts were then filtered as before and purified by FACS sorting (FACSAria II SORP with 100-micron setup, purity precision, yield mask at 32, purity mask at 32, plates voltage at 2,500, voltage centering at 20, sheath pressure at 20, and target gap at 12). Protoplasts were sorted into 1 × PBS with 0.1% BSA and 0.4M mannitol, and pelleted at 400 g for 2 min at 4°C. The supernatant was carefully removed, leaving 20-40μl to gently resuspend the pellet. The protoplasts were stained with trypan blue (Thermo Fisher Scientific, 15250061) to check concentration and viability with a hemocytometer under a light microscope, and good quality protoplasts with viability ≥ 70% were immediately loaded into the 10x Genomics Chromium System using V2 chemistry kits. scRNA-seq libraries were sequenced by Illumina short reads with ~400M paired end reads per library (read1 = 28bp, read2 = 56bp) (Table S1). Raw sequencing data were deposited in NCBI's Sequence Read Archive (SRA). SRA IDs were listed in Table S1.

CellPress
Developmental Cell
Resource

Anatomy and confocal microscopy

For anatomy, wild type (B73) developing ears from fresh plants were hand dissected and fixed in 4% paraformaldehyde (Electron Microscopy Sciences, 15714 s) (Jackson et al., 1994). The fixed ear tissue was dehydrated through a graded alcohol series (50%, 70%, 85%, 95%, and 100%) and a histoclear series, then embedded in paraplast (McCormick Scientific, 39503002) (Jackson et al., 1994). 5μm sections were cut using a Leica microtome, then mounted on ProbeOn Plus Slides (Fisher Scientific, 22-230-900) and rehydrated and stained with Calcofluor white stain (Sigma-Aldrich, 18909). Images were taken on a ZEISS LSM 710 confocal microscope using DAPI channel.

For vascular bundle anatomy, wild type (B73) developing ears were fixed overnight at 4°C in 0.1 M cacodylate buffer pH 7.4 (Electron Microscopy Sciences, 11652) with 4% paraformaldehyde (Electron Microscopy Sciences, 15714 s) and 1% glutaraldehyde (Electron Microscopy Sciences, 16537-16), washed three times with 0.1M cacodylate buffer and then dehydrated in a graded ethanol series (50%, 60%, 70%, 80%, 90%, 95%, two times 100%). The dehydrated ear samples were transferred to LR white resin / ethanol in 1:3, 1:1, and 3:1 ratios, and twice overnight in 100% LR white resin (Electron Microscopy Sciences, 905072). Tissues in LR white resin were then polymerized in a 60°C oven for 48 h, and 0.5-1mm thick sections were cut using 45 Diamond DiATOME Histo Knife on a Reichert Ultracut E microtome. Sections were collected and stained by toluidine blue (Sigma-Aldrich, T3260). Images were taken by Nikon DS-Ri2 microscope.

pZmYAB14-TagRFPt, ZmM16-YFP, and ZmHDZIV6-YFP lines were imaged using ZEISS LSM 710 or 780 confocal microscopes. The DAPI channel was used for capturing autofluorescence (blue color) in Figures 2N and 3D.

mRNA *in situ* hybridization

mRNA *in situs* were conducted as previously described (Jackson et al., 1994). Briefly, wild type (B73) developing ears were freshly collected, fixed in 4% paraformaldehyde (Electron Microscopy Sciences, 15714 s) (Jackson et al., 1994). The fixed ear tissue was dehydrated through a graded alcohol series (50%, 70%, 85%, 95%, and 100%) and a histoclear series, then embedded in paraplast (McCormick Scientific, 39503002) (Jackson et al., 1994). 10μm sections were cut using a Leica microtome, then mounted on ProbeOn Plus Slides (Fisher Scientific, 22-230-900). To prepare probes for marker genes, we added T7 promoter sequences GAGTAATACGACTCACTATAGGGAGA into reverse primers used to amplify gene-specific PCR products from cDNA templates. Then probes were synthesized by *in vitro* transcription using T7 RNA Polymerase (Sigma-Aldrich, 10881775001). Primer sequences for all genes were listed in Table S4. Probes were then applied on tissue sections and incubated at 50°C overnight. To detect the hybridization signal, we applied freshly dissolved NBT/BCIP Ready-to-Use Tablets (Roche, 11697471001) in the alkaline phosphatase reaction solution. Images were taken using a Nikon DS-Ri2 DIC microscope.

FACS and bulk RNA-seq, ATAC-seq library preparation

For three biological replicates of bulk RNA-seq to identify protoplasting-responsive genes, wild type B73 background ear tissue was used. RNA was isolated from protoplasted and equivalent non-protoplasted tissue to compare side by side. For three biological replicates of *pZmYAB14-TagRFPt* FACS RNA-seq, ear tissue of *pZmYAB14-TagRFPt/bd1;Tu* and *bd1;Tu* negative control plants were collected and digested as described earlier. Protoplasts were gently washed, filtered, and resuspended as before. *bd1;Tu* negative control protoplasts were first loaded into FACSAria II SORP to set up the gate for identifying autofluorescence signals. *pZmYAB14-TagRFPt/bd1;Tu* protoplasts were then loaded using the same settings. RFP cells were collected based on specific signals from the mStrawberry channel, and examined under a ZEISS LSM 710 confocal microscope. RNA for RFP positive protoplast samples and control samples (total protoplasts without sorting) was extracted using Arcturus PicoPure RNA Isolation Kit (Thermo Fisher Scientific, KIT0204). RNA was examined by a RNA Bioanalyzer kit (Agilent, 5067-1513). RNA-seq libraries were built using SMART-Seq v4 Ultra Low Input RNA Kit (Takara Bio USA, Inc., 634890) and Nextera XT DNA Library Prep Kit (Illumina, FC-131-1024). Library quality and size was examined by a DNA High Sensitivity Bioanalyzer chip (Agilent, 5067-4626), and quantified using the KAPA Library Quantification Kit (Roche, KK4824) before Illumina sequencing. Raw sequencing data were deposited into NCBI's Sequence Read Archive (SRA). SRA IDs were listed in Tables S1 and S2. For one biological replicate of *pZmYAB14-TagRFPt* FACS ATAC-seq, RFP protoplasts were collected as described above, and nuclei isolated for library construction and sequencing as previously (Lu et al., 2017). Raw sequencing data were deposited in NCBI's Sequence Read Archive (SRA). SRA ID was listed in Table S2.

ChIP-seq library preparation

ChIP experiments were conducted as previously described (Pautler et al., 2015) with some modifications. Briefly, two biological replicates of freshly harvested ear tissues of ZmHDZIV6-YFP/*bd1;Tu* and ZmM16-YFP were immediately cross-linked in buffer containing 1% formaldehyde, 10mM HEPES-NaOH PH7.4, 0.4 M sucrose, 1 mM EDTA, and 1 mM PMSF, for 20min under vacuum. Glycine was then added to a concentration of 0.1 M to for another 5 min under vacuum. Nuclei extraction was conducted using CelLytic PN Isolation/Extraction Kit (Sigma-Aldrich, CELLYTPN1). For immunoprecipitation, we used high-affinity GFP-Trap magnetic agarose (ChromoTek, gtma-20; RRID: AB_2631358). For building ChIP-seq libraries, we used NEXTflex ChIP-seq Kit (PerkinElmer Applied Genomics, NOVA-5143-02) with AMPure XP beads (Beckman Coulter, A63880). ChIP-seq libraries were quantified by KAPA Library Quantification Kits (Roche, KK4824) before Illumina sequencing. Raw sequencing data were deposited in NCBI's Sequence Read Archive (SRA). SRA IDs were listed in Table S3.

Developmental Cell
Resource

QUANTIFICATION AND STATISTICAL ANALYSIS

scRNA-seq analysis, clustering, and selection of marker genes

Sequencing reads of three biological replicates of wild type (B73) whole ear scRNA-seq samples were aligned to the maize v3 reference genome using STARsolo v2.7.0.f (Dobin et al., 2013). We updated the v3.31 GTF annotation file by adding four maize *CLAV ATA3/EMBRYO SURROUNDING REGION-RELATED (ZmCLE)* genes, including *ZmCLE7* (*GRMZM2G372364*), *ZmCLE14* (*AC191109.3_FG001*), *ZmCLE25* (*GRMZM2G525788*), and *ZmCLEug-2* (*GRMZM2G054501*) (Table S4; MaizeGDB), before using it to build a STAR genome index with default parameters (Dobin et al., 2013). All downstream processing was performed in R (R Core Team, 2013), with each dataset analyzed separately. In brief, we removed droplets that lacked a protoplast using EmptyDrops with a minimum threshold of 800 UMIs (Lun et al., 2019), and removed probable doublets using DoubletFinder (McGinnis et al., 2019). Expression data was log2 normalized using scater (McCarthy et al., 2017). We identified highly variable genes using the trendVar function in scran (Lun et al., 2016), selecting genes at FDR < 0.05, and we used the rsvd package (Erichson et al., 2019) to calculate approximate principal components for all cells after subsetting to highly variable genes. We then generated a nearest neighbor graph for cells with the cccd package (Marchette, 2015) using the Euclidean distance across the top 20 PCs, with k = 100. To find clusters, we used the InfoMap algorithm implemented in the igraph R package (Csardi and Nepusz, 2006), resampling 100 times (Rosvall and Bergstrom, 2008). MetaNeighbor analysis was performed as previously described (Crow et al., 2018). We used the umap package to generate embeddings for visualization (Konopka, 2020). Within each UMAP, every dot represents a cell, and the color scale indicates the normalized expression level by adding a constant 75% of each cell's expression value to nearest neighbors for clear visualization. Differential expression statistics were calculated with an AUROC test on log counts using the auc multifunc function from EGAD (Ballouz et al., 2017). Differential expression genes with AUROC scores of $\geq$ 0.7 in at least one replicate were considered as meta-cluster marker genes (Table S1).

FACS and bulk RNA-seq, ATAC-seq analysis

Three biological replicates of FACS and bulk RNA-seq analyses were performed as previously (Wang et al., 2020) with some modifications. Bulk RNA-seq datasets for extracting ear tissue specific genes were downloaded from a previous study (Walley et al., 2016). Raw sequencing reads were first trimmed with Trimmomatic (Bolger et al., 2014), and then mapped with STAR (Dobin et al., 2013) using the same updated maize V3 reference as scRNA-seq analysis. edgeR (Robinson et al., 2010) was used to perform differential expression analysis. We calculated similarity between scRNA-seq and FACS RNA-seq data using the auroc_analytic function in EGAD (Ballouz et al., 2017), with ranked scRNA-seq meta-cluster 3 p-values as the "scores" and FACS RNA-seq DE genes (log2 FC > 0 and FDR < 0.05, Table S2) as the "labels." Whole ear ATAC-seq datasets were downloaded from a previous study (Ricci et al., 2019). One biological replicate of FACS ATAC-seq analysis was performed as previously (Ricci et al., 2019), using the same reference as scRNA-seq analysis.

ChIP-seq analysis

Two biological replicates of ChIP-seq reads for both ZmHDZIV6-YFP and ZmM16-YFP datasets were trimmed using sickle (https://github.com/najoshi/sickle). Duplicated reads were further removed by elprep (Herzeel et al., 2015), before aligning to the same updated maize V3 reference used for scRNA-seq analysis with BWA-MEM (Li and Durbin, 2009). Alignment reads were filtered for those above a MAPping Quality (MAPQ) threshold above 40. Peak calling, peak annotation, and motif enrichment were performed with HOMER (Heinz et al., 2010), with peak calling parameters as the following: -F = 8: Fold enrichment threshold of IP tag count over input tag count, -L = 2: Fold enrichment of a putative peak's tag count over surrounding (local) genomic region, and -LP = 0.0001: p-value cutoff for local fold enrichment to a peak to be considered. High confidence peaks between two biological replicates were determined by finding midpoints of peaks positioned within 300bp of each other (Pautler et al., 2015).

scRNA-seq co-expression analysis

89 maize bulk tissue RNA-seq datasets were used for calculating co-expression values as previously described (Lee et al., 2020). Briefly, co-expression networks were constructed using Spearman's correlation. Three gene pairs including *RA3-ZmTPP4*, *RA3-ZmTPP12*, and *ZmTPP4-ZmTPP12* were then extracted from the datasets. The aggregated co-expression value of each gene pair was calculated and reported in Figure S4. For scRNA-seq co-expression to predict *RA3-ZmTPPs* redundancy, *ZmVOZ* co-expression, and TF ChIP-Seq directly modulated targets, the average Jaccard index was calculated for B73 whole ear datasets, using genes expressed in > 1% of cells. Genes with at least 1 UMI were assigned a value of 1, and all others were assigned a value of 0. The Jaccard index was calculated for each ear dataset separately and then averaged. To calculate significance (Figure S4) we took a non-parametric approach with the null hypothesis that there is no replicable co-expression across batches, calculated by convolving the uniform distributions obtained for each batch after ranking. To compare overlaps in co-expression between two genes, we re-ranked and assessed similarity against the null using the same approach. FDRs were calculated by dividing cumulative nulls by cumulative empiricals.

Integration of GWAS with scRNA-seq or ear tissue bulk RNA-seq

To integrate GWAS with scRNA-seq, we first selected unique scRNA-seq marker genes from ear meristem, determinate lateral organ, and vasculature meta-clusters (3, 4, 5, 6, 9, 10, 11, 12) using two different cutoffs. The stringent cutoff was AUROC $\geq$ 0.7 across

CellPress
Developmental Cell
Resource

all replicates, which gave 68 scRNA-seq marker genes (Table S3). The less stringent cutoff was AUROC $\geq$ 0.7 in at least one replicate, which gave 241 scRNA-seq marker genes (Table S3). For targeted GWAS analysis: we used the best linear unbiased predictors (BLUPs) from nine ear phenotypes to perform a targeted GWAS using the unified mixed linear model. BLUPs estimation, GWAS model details, and genomic marker filtering procedures were described in (Rice et al., 2020). The SNPs in or within 2kb of scRNA-seq marker genes from lists of both stringent and less stringent cutoffs were used (Table S3).

For lambda analysis, the procedure has been previously described (Parvathaneni et al., 2020). Briefly, lambda was a ratio of the FDR adjusted p-values for a given set of markers when included with the full marker set compared to when considered on their own (equation below).

$$\text{Lambda} = \frac{qth\_Percentile(-\log(Reduced\ FDR\ Adjusted\ p-values))}{qth\_Percentile(-\log(Genomewide\ FDR\ Adjusted\ p-values))}$$

We looked at the significance of 99th percentile ($q = 99$) of FDR adjusted p-values subset of markers (SNPs in the region of genes of interest). We performed 1,000 replicates (random subsets) and estimated the lambda distributions for each trait. Lambda values of 241 unique markers from scRNA-seq meta-clusters (3, 4, 5, 6, 9, 10, 11, 12) with less stringent cutoff (Table S3) were compared against random subsets lambda distributions to determine significance. Traits with $\lambda \geq$ mean ± 2SD were considered to be biologically significant.

For SNP Heritability analysis, we estimated narrow-sense heritability (h^2) from the subsets of the scRNA-seq SNPs considered in the lambda analysis using the LDAK software v5.0 (Speed et al., 2012). Thus, the resulting estimate of h^2 provides an estimate of the additive genetic variance explained by genes of interest. To determine if the resulting heritability for a given trait was greater than chance, the heritability for 1000 permutations using a random subset of maize genes was estimated. For a given permutation, genes with at least one SNPs within the genic region were randomly selected. Enough genes were selected to ensure the total number in a permuted subset was ± 5 compared to the target set. A target set was declared significant for a given trait if its heritability was greater than the top 5% of permuted values.

To integrate GWAS with ear tissue bulk RNA-seq, we first extracted ear tissue specific genes from a previous study (Table S3) (Walley et al., 2016). Then we used these genes to perform same analyses, including targeted GWAS, lambda, and SNP Heritability as mentioned above for scRNA-seq (Table S3).

Appendix C

Differential co-expression between scRNAseq and bulk RNAseq

C.1 Main

In Chapter 3, I focus on the comparison of co-expression networks between scRNAseq and bulk. I find that co-expression relationships seen in bulk are consistent with co-expression relationships seen in networks built at a variety of levels of heterogeneity in scRNAseq data. However, I largely focused on this analysis at the module/geneset level. This ignores the possibility that scRNAseq is able to find individual cell type specific co-expression relationships that are obscured by bulk RNAseq data (Trapnell, 2015). In this appendix, I follow up on the work in chapter 3 to look at edges identified in cell type specific co-expression networks that are absent in networks in bulk RNAseq data.

First, I look at the difference in edges between the GABAergic (a class level) network and the bulk brain network. To threshold for edges that are only in the single cell network, I took the difference in the edge between the single cell network and the bulk and selected edges with a difference greater than .95. At this threshold, I identify 1,321 edges. About 33% of the genes found in these edges are markers in at least 1 subclass. Marker genes constitute 14% of all genes in the networks, so they are overrepresented in the differentially co-expressed edges. 66% of the differentially co-expressed edges contain at least one marker. This analysis is looking at markers for cell type labels for one level of

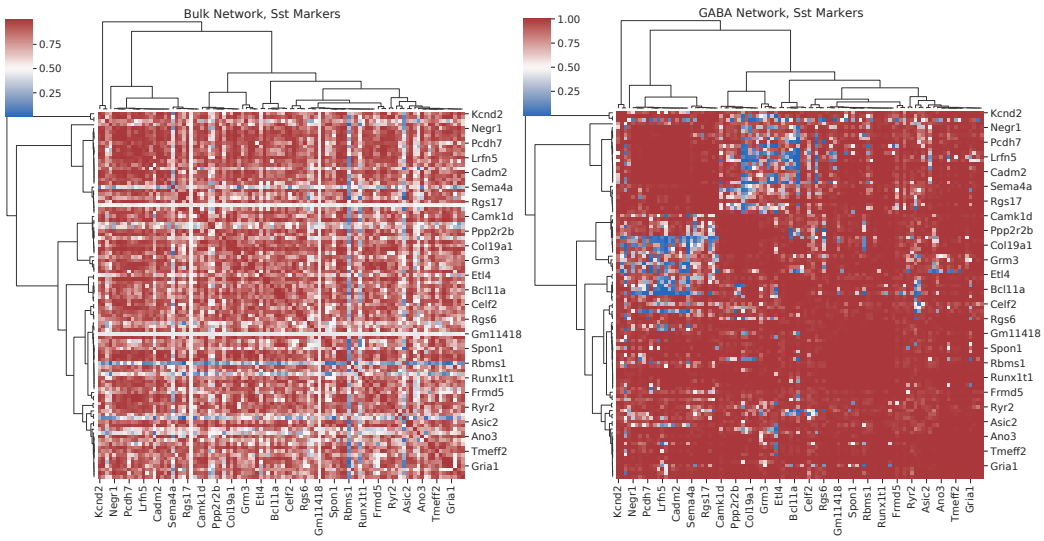


FIGURE C.1: Sst marker co-expression in bulk (left) and GABAergic neuron (right) networks

cell type classification lower than the network built. So edges for markers are expected to be really strong (Figure 3.3). When we evaluated the markers (Figures 3.10-11) as whole modules we noted that they are remarkably similar. It is possible that the single cell network here is doing a better job at prioritizing the most relevant edges to the cell types, but it also could just be noise. A heatmap of the edges for the Sst markers in both the GABAergic and bulk networks shows they are very similar Figure C.1. It is possible that the differences between the networks could constitute signals only visible in the single cell network, but more likely is just noise.

Next, I evaluate the networks built from each of the 13 subclasses. Unlike the analysis using the GABAergic network, these networks are constructed at the same level of markers that are used to evaluate them. Using the .95 threshold again, I find about 2,000 edges per network C.2. The L6 IT Car3 subclass is quite rare (the subclass is not in all datasets) so the increase in edges in that network is likely drive-by noise. Overall, the number of edges that are different from the bulk is largely consistent with the difference seen in the GABAergic network.

Most edges that are differential co-expressed with the bulk are unique to 1 or a few of the subclass networks Figure C.3. There are 13 edges that are recurrent across 12 or more of the subclasses Table C.1. These edges could be interesting if there are some shared

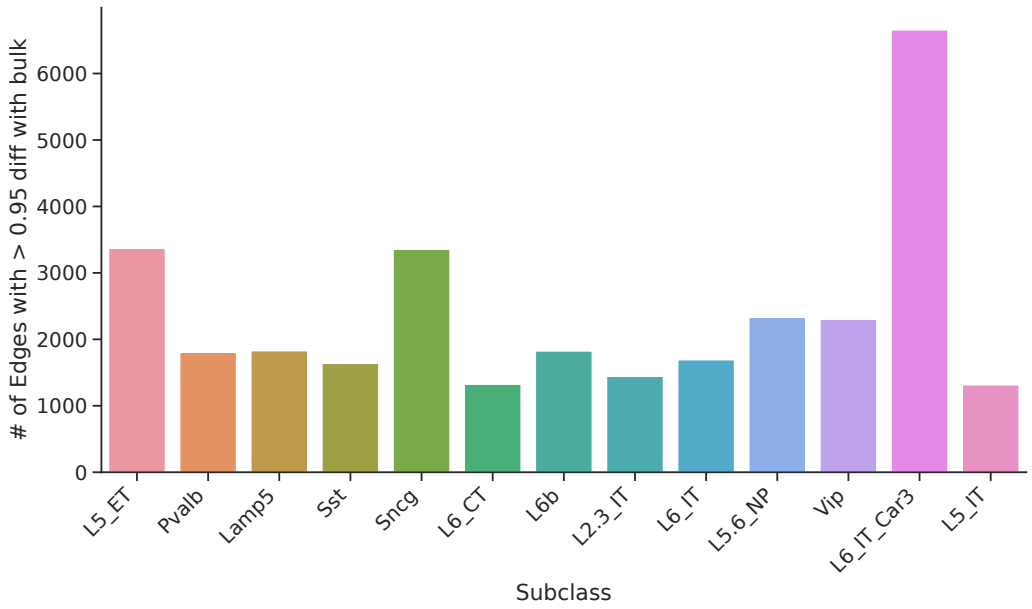


FIGURE C.2: Number of differentially co-expressed edges per subclass-specific co-expression network. The threshold is at .95 for the difference between the subclass-specific network and bulk network.

functional relationships across the subclasses that are being obscured in the bulk data. However, the genes that constitute these edges are only shared across high-level GO terms, so if there is some interesting functional relationship it would require some alternative analysis to identify Table C.2.

To characterize the genes that make the edges I looked at the percent of differentially co-expressed edges in each subclass network that contain a marker for a subclass. Most subclasses have a bias towards genes that are for markers of themselves, but also contain many edges with markers from other subclasses (Figure C.4. GO enrichment of edges unique to each subclass mainly finds terms that are quite general processes. Part of this could be the fact that the genes used in building the networks, being highly expressed across the datasets, are biased towards more general functions (Table C.3. Alternatively, it could mean that a major aspect of cell type specificity is small changes to the co-expression relationships within core functions that, when analyzed at a module level appear more consistent across cell types.

This analysis finds edges specific to subclass networks that are connecting genes with

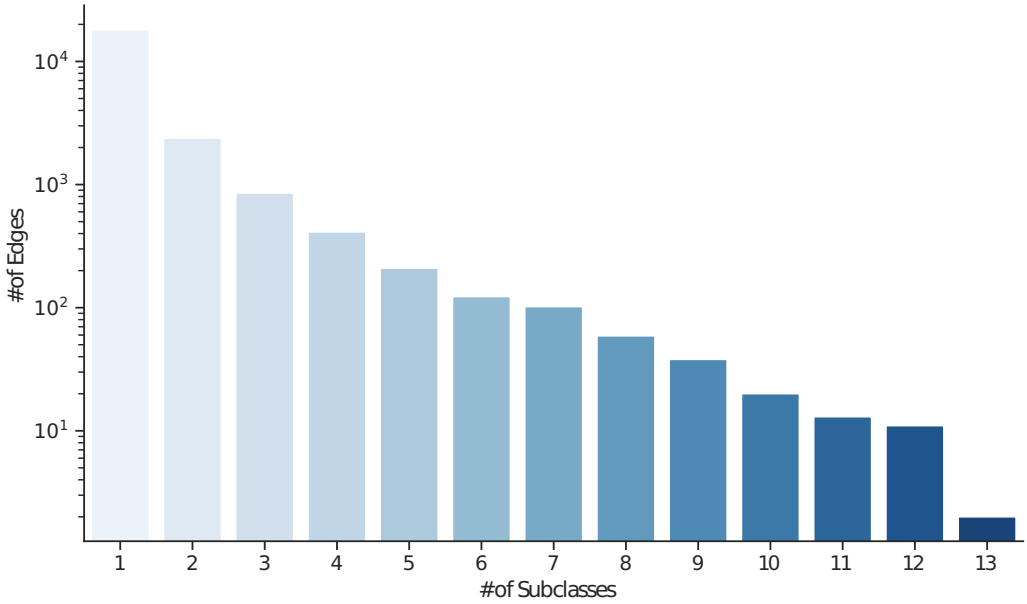


FIGURE C.3: Recurrence of differentially co-expressed edges with bulk

Gene1	Gene2	recurrence
Clstn1	Hspa5	13
Atp1a3	Hspa5	13
Hsp90b1	Slc22a17	12
Calm1	Tpt1	12
Atp5d	Hsp90aa1	12
Hnrnpk	Sult4a1	12
Sncb	Tpt1	12
Hspa5	Tmem151a	12
Hspa5	Sult4a1	12
Aplp1	Hsp90b1	12
Hspa5	Igsf8	12
Sult4a1	Ube2d3	12
Hsp90b1	Tmem59l	12

TABLE C.1: Edges that are differentially co-expressed with the bulk network in at least 12 of 13 subclasses

GO ID	Overlap	Annot
GO:0003674	17.0	molecular_function
GO:0008150	17.0	biological_process
GO:0005575	17.0	cellular_component
GO:0005623	16.0	cell
GO:0044464	16.0	obsolete cell part
GO:0005622	15.0	intracellular
GO:0005488	15.0	binding
GO:0044424	15.0	obsolete intracellular part
GO:0005737	15.0	cytoplasm
GO:0044444	14.0	obsolete cytoplasmic part
GO:0005515	14.0	protein binding
GO:0043227	14.0	membrane-bounded organelle
GO:0043229	14.0	intracellular organelle
GO:0043226	14.0	organelle
GO:0043231	13.0	intracellular membrane-bounded organelle
GO:0016020	13.0	membrane
GO:0065007	12.0	biological regulation
GO:0044425	12.0	obsolete membrane part
GO:0009987	12.0	cellular process
GO:0050789	10.0	regulation of biological process

TABLE C.2: GO terms shared by the 17 genes that are nodes in the differentially co-expressed edges with the bulk network in at least 12 of 13 subclasses

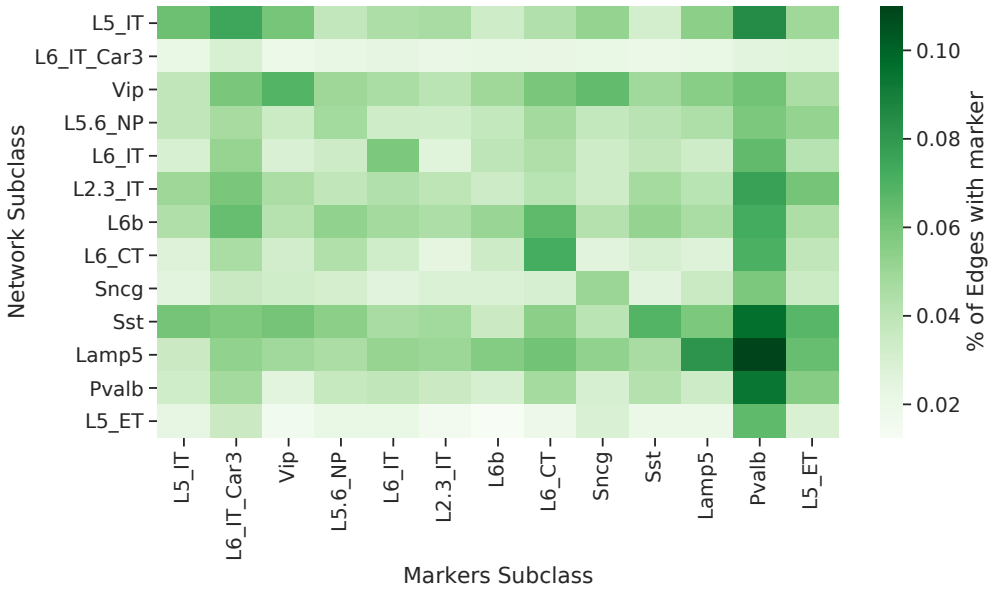


FIGURE C.4: Differentially co-expressed edges in each subclass specific network that contain a marker.

	GO ID	Network Subclass	padj	Annot	Term Size
0	GO:0009123	L5_ET	0.000134	nucleoside monophosphate metabolic process	63.0
1	GO:0009161	L5_ET	0.000134	ribonucleoside monophosphate metabolic process	60.0
2	GO:0009126	L5_ET	0.000134	purine nucleoside monophosphate metabolic process	60.0
3	GO:0009167	L5_ET	0.000134	purine ribonucleoside monophosphate metabolic ...	60.0
4	GO:0046034	L5_ET	0.000376	ATP metabolic process	54.0
5	GO:0044455	L5_ET	0.000376	obsolete mitochondrial membrane part	77.0
6	GO:0009144	L5_ET	0.001806	purine nucleoside triphosphate metabolic process	64.0
7	GO:0006412	L5_ET	0.001806	translation	96.0
8	GO:0009205	L5_ET	0.001844	purine ribonucleoside triphosphate metabolic p...	62.0
9	GO:0009199	L5_ET	0.002947	ribonucleoside triphosphate metabolic process	63.0
10	GO:0070469	L5_ET	0.003953	respirasome	32.0
11	GO:0009141	L5_ET	0.004419	nucleoside triphosphate metabolic process	69.0
12	GO:0099060	L6b	0.005762	integral component of postsynaptic specializat...	52.0
13	GO:0009156	L5_ET	0.007679	ribonucleoside monophosphate biosynthetic process	23.0
14	GO:0009168	L5_ET	0.007679	purine ribonucleoside monophosphate biosynthet...	23.0
15	GO:0009127	L5_ET	0.007679	purine nucleoside monophosphate biosynthetic p...	23.0
16	GO:0030141	Sst	0.009110	secretory granule	87.0
17	GO:0003735	L5_ET	0.012424	structural constituent of ribosome	32.0
18	GO:0099634	L6b	0.013554	postsynaptic specialization membrane	68.0
19	GO:0098948	L6b	0.013761	intrinsic component of postsynaptic specializa...	57.0
20	GO:0098803	L5_ET	0.014519	respiratory chain complex	28.0
21	GO:0098800	L5_ET	0.014519	inner mitochondrial membrane protein complex	53.0
22	GO:0042773	L5_ET	0.015080	ATP synthesis coupled electron transport	24.0
23	GO:0009124	L5_ET	0.015080	nucleoside monophosphate biosynthetic process	24.0
24	GO:0006119	L5_ET	0.015080	oxidative phosphorylation	24.0
25	GO:0098609	Sst	0.023362	cell-cell adhesion	92.0
26	GO:0005840	L5_ET	0.024743	ribosome	62.0
27	GO:0006413	L5_ET	0.029647	translational initiation	25.0
28	GO:0042775	L5_ET	0.031569	mitochondrial ATP synthesis coupled electron t...	23.0
29	GO:0043297	Sst	0.032990	apical junction assembly	7.0
30	GO:0022626	L5_ET	0.033635	cytosolic ribosome	19.0
31	GO:0022904	L5_ET	0.044029	respiratory electron transport chain	30.0
32	GO:0098798	L5_ET	0.044862	mitochondrial protein complex	95.0
33	GO:0099055	L6b	0.047252	integral component of postsynaptic membrane	79.0
34	GO:0005746	L5_ET	0.047915	mitochondrial respirasome	28.0
35	GO:0006091	L5_ET	0.048927	generation of precursor metabolites and energy	85.0
36	GO:0008135	L5_ET	0.049724	translation factor activity; RNA binding	35.0

TABLE C.3: GO enrichment for genes in edges that are uniquely differentially co-expressed between the bulk. Edges included are ones with a recurrence of 1 in Figure C.3 and a p-value of less than 0.05. Enrichment was done using R.A. Fisher's exact test with a Benjamini-Hochberg correction.

largely high level and more general processes. It is conceivable that an important aspect to defining cell types is small changes in co-expression, just a few edges, within a cellular process or pathways that are broadly important to all cell types. However, it is hard to make strong conclusions about the importance of the individual edges identified here without further follow-up experiments or other modalities of data.

Bibliography

Abdelaal, Tamim et al. (2019). "A comparison of automatic cell identification methods for single-cell RNA sequencing data". In: *Genome Biology* 20.1. Publisher: BioMed Central, p. 194.

Adamson, Britt et al. (2016). "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response". In: *Cell* 167.7, 1867–1882.e21.

Adolfsson, Jörgen et al. (2005). "Identification of Flt3+ Lympho-Myeloid Stem Cells Lacking Erythro-Megakaryocytic Potential A Revised Road Map for Adult Blood Lineage Commitment". In: *Cell* 121.2.

Ballouz, Sara, Paul Pavlidis, and Jesse Gillis (2017). "Using predictive specificity to determine when gene set analysis is biologically meaningful". In: *Nucleic Acids Research* 45.4, e20–e20.

Ballouz, Sara et al. (2016). "EGAD: ultra-fast functional analysis of gene networks". In: *Bioinformatics*, btw695.

Barkas, Nikolas et al. (2018). "Wiring together large single-cell RNA-seq sample collections". en. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 460246.

Basilico, Silvia et al. (2020). "Dissecting the early steps of MLL induced leukaemogenic transformation using a mouse model of AML". In: *Nature Communications* 11.1, p. 1407.

Bella, Daniela J. Di et al. (2021). "Molecular logic of cellular diversification in the mouse cerebral cortex". In: *Nature*, pp. 1–6.

Berge, Koen Van den et al. (2020). “Trajectory-based differential expression analysis for single-cell sequencing data”. In: *Nature Communications* 11.1, p. 1201.

Blanco-Melo, Daniel et al. (2020). “Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19”. In: *Cell* 181.5, 1036–1045.e9.

Cao, Junyue et al. (2019). “The single-cell transcriptional landscape of mammalian organogenesis”. In: *Nature* 566.7745, pp. 496–502.

Cao, Yinghao, Xiaoyue Wang, and Gongxin Peng (2020). “SCSA: A Cell Type Annotation Tool for Single-Cell RNA-seq Data”. In: *Frontiers in Genetics* 11, p. 490.

Cembrowski, Mark S. and Vilas Menon (2018). “Continuous Variation within Cell Types of the Nervous System”. In: *Trends in Neurosciences* 41.6, pp. 337–348.

Chari, Tara, Joeyta Banerjee, and Lior Pachter (2021). “The Specious Art of Single-Cell Genomics”. In: *bioRxiv*, p. 2021.08.25.457696.

Chen, Xiaoyin et al. (2019). “High-Throughput Mapping of Long-Range Neuronal Projection Using In Situ Sequencing”. In: *Cell* 179.3, 772–786.e19.

Cheng, Yuanming et al. (2019). “m6A RNA Methylation Maintains Hematopoietic Stem Cell Identity and Symmetric Commitment”. In: *Cell Reports* 28.7, 1703–1716.e6.

Chklovskii, Dmitri B. (2004). “Synaptic Connectivity and Neuronal Morphology Two Sides of the Same Coin”. In: *Neuron* 43.5, pp. 609–617.

Cohen, Yael C. et al. (2021). “Identification of resistance pathways and therapeutic targets in relapsed multiple myeloma patients through single-cell sequencing”. In: *Nature Medicine* 27.3, pp. 491–503.

Consortium, The ENCODE Project et al. (2020). “Expanded encyclopaedias of DNA elements in the human and mouse genomes”. In: *Nature* 583.7818, pp. 699–710.

Consortium, The GTEx (2020). “The GTEx Consortium atlas of genetic regulatory effects across human tissues”. In: *Science* 369.6509, pp. 1318–1330.

Consortium, The Tabula Sapiens and Stephen R Quake (2021). “The Tabula Sapiens: a single cell transcriptomic atlas of multiple organs from individual human donors”. In: *bioRxiv*, p. 2021.07.19.452956.

Cook, David P. and Barbara C. Vanderhyden (2021). “Transcriptional census of epithelial-mesenchymal plasticity in cancer”. In: *bioRxiv*, p. 2021.03.05.434142.

Crow, Megan and Jesse Gillis (2018). "Co-expression in Single-Cell Analysis: Saving Grace or Original Sin?" In: *Trends in Genetics* 34.Nat. Protoc. 13 2018, pp. 823–831.

— (2019). "Single cell RNA-sequencing: replicability of cell types". In: *Current Opinion in Neurobiology* 56, pp. 69–77.

Crow, Megan et al. (2016). "Exploiting single-cell expression to characterize co-expression replicability". In: *Genome Biology* 17.1, p. 101.

— (2018). "Characterizing the replicability of cell types defined by single cell RNA-sequencing data using MetaNeighbor". In: *Nature Communications* 9.1. Publisher: Nature Publishing Group ISBN: 2041-1723, p. 884.

Dahlin, Joakim S et al. (2018). "A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in Kit mutant mice". In: *Blood* 131.21, e1–e11.

Efremova, Mirjana et al. (2020). "CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes". In: *Nature Protocols*, pp. 1–23.

Eisen, M B et al. (1998). "Cluster analysis and display of genome-wide expression patterns". In: *Proceedings of the National Academy of Sciences* 95.25, pp. 14863–14868.

Elias, Harold K., David Bryder, and Christopher Y. Park (2017). "Molecular mechanisms underlying lineage bias in aging hematopoiesis". In: *Seminars in Hematology* 54.1, pp. 4–11.

Emmrich, Stephan et al. (2021). "Naked Mole-Rat Hematopoietic Stem and Progenitors are Highly Quiescent with an Inherent Myeloid Bias". In: *bioRxiv*, p. 2021.08.01.454652.

Erarslan-Uysal, Büşra et al. (2020). "Chromatin accessibility landscape of pediatric T-lymphoblastic leukemia and human T-cell precursors". In: *EMBO Molecular Medicine* 12.9, e12104.

Eraslan, Gokcen et al. (2021). "Single-nucleus cross-tissue molecular reference maps to decipher disease gene function". In: *bioRxiv*, p. 2021.07.19.452954.

Evrard, Maximilien et al. (2018). "Developmental Analysis of Bone Marrow Neutrophils Reveals Populations Specialized in Expansion, Trafficking, and Effector Functions". In: *Immunity* 48.2, 364–379.e8.

Farahbod, Marjan and Paul Pavlidis (2020). "Untangling the effects of cellular composition on coexpression analysis". In: *Genome Research* 30.6, pp. 849–859.

Feregrino, Christian et al. (2019). "A single-cell transcriptomic atlas of the developing chicken limb". In: *BMC Genomics* 20.1, p. 401.

Fiers, Mark W E J et al. (2018). "Mapping gene regulatory networks from single-cell omics data". In: *Briefings in Functional Genomics* 17.4, pp. 246–254.

Fischer, Stephan and Jesse Gillis (2021). "How many markers are needed to robustly determine a cell's type?" In: *bioRxiv*, p. 2021.04.16.439807.

Forcato, Mattia, Oriana Romano, and Silvio Bicciato (2020). "Computational methods for the integrative analysis of single-cell data". en. In: *Briefings in Bioinformatics* 22.3, bbaa042.

Friedman, Jerome H. (1997). "On Bias, Variance, 0/1—Loss, and the Curse-of-Dimensionality". In: *Data Mining and Knowledge Discovery* 1.1, pp. 55–77.

Fulwyler, M. J. (1965). "Electronic Separation of Biological Cells by Volume". In: *Science* 150.3698, pp. 910–911.

Gaudet, Pascale and Christophe Dessimoz (2016). "The Gene Ontology Handbook". In: pp. 189–205.

Ge, Songwei et al. (2021). "Supervised Adversarial Alignment of Single-Cell RNA-seq Data". In: *Journal of Computational Biology* 28.5, pp. 501–513.

Gegenhuber, Bruno et al. (2020). "Regulation of neural gene expression by estrogen receptor alpha". In: *bioRxiv*, p. 2020.10.21.349290.

Giladi, Amir et al. (June 2018). "Single-cell characterization of haematopoietic progenitors and their trajectories in homeostasis and perturbed haematopoiesis". In: *Nature Cell Biology* 20.7, pp. 836–846.

Grabski, Isabella N. and Rafael A. Irizarry (2020). "A probabilistic gene expression barcode for annotation of cell-types from single cell RNA-seq data". In: *bioRxiv*, p. 2020.01.05.895441.

Guo, Guoji et al. (Sept. 2013). "Mapping Cellular Hierarchy by Single-Cell Analysis of the Cell Surface Repertoire". In: *Cell Stem Cell* 13.4, pp. 492–505.

Hafemeister, Christoph and Rahul Satija (2019). "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome Biology* 20.1, p. 296.

Hagemann-Jensen, Michael et al. (2020). "Single-cell RNA counting at allele and isoform resolution using Smart-seq3". In: *Nature Biotechnology* 38.6, pp. 708–714.

Haghverdi, Laleh et al. (Aug. 2016). "Diffusion pseudotime robustly reconstructs lineage branching". In: *Nature Methods* 13.10, pp. 845–848.

Hao, Yuhan et al. (2021). "Integrated analysis of multimodal single-cell data". In: *Cell* 184.13, 3573–3587.e29.

Harris, Benjamin D., John Lee, and Jesse Gillis (2021). "A Meta-Analytic Single-Cell Atlas of Mouse Bone Marrow Hematopoietic Development". In: *bioRxiv*, p. 2021.08.12.456098.

Harris, Benjamin D. et al. (2021). "Single-cell co-expression analysis reveals that transcriptional modules are shared across cell types in the brain". In: *Cell Systems*.

Hartl, Christopher L et al. (2020). "The architecture of brain co-expression reveals the brain-wide basis of disease susceptibility". In: *bioRxiv*, p. 2020.03.05.965749.

Hashimoto, Kosuke et al. (2019). "Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians". In: *Proceedings of the National Academy of Sciences* 116.48, pp. 24242–24251.

He, Miao and Z Josh Huang (2018). "Genetic approaches to access cell types in mammalian nervous systems". In: *Current Opinion in Neurobiology* 50.Annu Rev Neurosci 36 2013.

Hicks, Stephanie C et al. (2017). "Missing data and technical variability in single-cell RNA-sequencing experiments". In: *Biostatistics* 19.4, pp. 562–578.

Hie, Brian et al. (2020). "Coexpression enables multi-study cellular trajectories of development and disease". In: *bioRxiv*, p. 719088.

Hwang, Byungjin, Ji Hyun Lee, and Duhee Bang (2018). "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental & Molecular Medicine* 50.8, pp. 1–14.

Hérault, Léonard et al. (2021). "Single-cell RNA-seq reveals a concomitant delay in differentiation and cell cycle of aged hematopoietic stem cells". In: *BMC Biology* 19.1, p. 19.

Jin, Xin et al. (2020). "In vivo Perturb-Seq reveals neuronal and glial abnormalities associated with autism risk genes". In: *Science* 370.6520, eaaz6063.

Johnson, W. Evan, Cheng Li, and Ariel Rabinovic (2007). "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1, pp. 118–127.

Kaminow, Benjamin, Dinar Yunusov, and Alexander Dobin (2021). "STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data". In: *bioRxiv*, p. 2021.05.05.442755.

Kelley, Kevin W. et al. (2018). "Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes". In: *Nature Neuroscience* 21.9, pp. 1171–1184.

Kester, Lennart and Alexander van Oudenaarden (2018). "Single-Cell Transcriptomics Meets Lineage Tracing". In: *Cell Stem Cell* 23.2, pp. 166–179.

Ketkar, Shamika et al. (2020). "Remethylation of Dnmt3a-/- hematopoietic cells is associated with partial correction of gene dysregulation and reduced myeloid skewing". In: *Proceedings of the National Academy of Sciences* 117.6, pp. 3123–3134.

Kharchenko, Peter V. (2021). "The triumphs and limitations of computational methods for scRNA-seq". In: *Nature Methods*, pp. 1–10.

Kimmel, Jacob C. and David R Kelley (2021). "Semi-supervised adversarial neural networks for single-cell classification". In: *Genome Research*, gr.268581.120.

Kingsley, Paul D et al. (2013). "Ontogeny of erythroid gene expression". In: *Blood* 121.6, e5–e13.

Kiselev, Vladimir Yu, Andrew Yiu, and Martin Hemberg (2018). "scmap: projection of single-cell RNA-seq data across data sets". In: *Nature Methods* 15.5. Publisher: Nature Publishing Group, pp. 359–362.

Kowalczyk, Monika S et al. (2015). "Single-cell RNA-seq reveals changes in cell cycle and differentiation programs upon aging of hematopoietic stem cells." In: *Genome research* 25.12, pp. 1860–72.

Kulkarni, Shubhada R et al. (2017). "TF2Network: predicting transcription factor regulators and gene regulatory networks in Arabidopsis using publicly available binding site information". In: *Nucleic Acids Research* 46.6, gkx1279–.

Langfelder, Peter and Steve Horvath (2008). "WGCNA: an R package for weighted correlation network analysis". In: *BMC Bioinformatics* 9.1, p. 559.

Lederer, Alex R and Gioele La Manno (2020). "The emergence and promise of single-cell temporal-omics approaches". In: *Current Opinion in Biotechnology* 63, pp. 70–78.

Lee, Homin K. et al. (2004). "Coexpression Analysis of Human Genes Across Many Microarray Data Sets". In: *Genome Research* 14.6, pp. 1085–1094.

Lee, John et al. (2020). "CoCoCoNet: conserved and comparative co-expression across a diverse set of species". In: *Nucleic Acids Research* 48.W1, gkaa348–.

Li, Dongshunyi et al. (2021a). "Inferring cell-cell interactions from pseudotime ordering of scRNA-Seq data". In: *bioRxiv*, p. 2021.07.28.454054.

Li, Hongjie et al. (2021b). "Fly Cell Atlas: a single-cell transcriptomic atlas of the adult fruit fly". In: *bioRxiv*, p. 2021.07.04.451050.

Liang, Shaoheng et al. (2021). "Stratified Test Accurately Identifies Differentially Expressed Genes Under Batch Effects in Single-Cell Data". In: *bioRxiv*, p. 2021.06.08.447617.

Litviňuková, Monika et al. (2020). "Cells of the adult human heart". In: *Nature* 588.7838, pp. 466–472.

Liu, Chen et al. (2021). "T Cell Development: Old Tales Retold By Single-Cell RNA Sequencing". In: *Trends in Immunology*.

Liu, Chuan-Xu et al. (2012). "Adenanthin targets peroxiredoxin I and II to induce differentiation of leukemic cells". In: *Nature Chemical Biology* 8.5, pp. 486–493.

Luecken, M. D. et al. (2020). "Benchmarking atlas-level data integration in single-cell genomics". en. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2020.05.22.111161.

López-Muñoz, Francisco, Jesús Boya, and Cecilio Alamo (2006). "Neuron theory, the cornerstone of neuroscience, on the centenary of the Nobel Prize award to Santiago Ramón y Cajal". In: *Brain Research Bulletin* 70.4-6, pp. 391–405.

MA, SHUANGGE et al. (2012). "Gene network-based cancer prognosis analysis with sparse boosting". In: *Genetics Research* 94.4, pp. 205–221.

Mack, Katya et al. (2019). "Gene Expression Networks Across Multiple Tissues Are Associated with Rates of Molecular Evolution in Wild House Mice". In: *Genes* 10.3, p. 225.

Macosko, Evan Z. et al. (2015). "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". In: *Cell* 161.5, pp. 1202–1214.

Markram, Henry et al. (2004). "Interneurons of the neocortical inhibitory system". In: *Nature Reviews Neuroscience* 5.10, pp. 793–807.

Marx, Vivien (2021). "Method of the Year: spatially resolved transcriptomics". In: *Nature Methods* 18.1, pp. 9–14.

McCall, Matthew N., Peter B. Illei, and Marc K. Halushka (2016). "Complex Sources of Variation in Tissue Expression Data: Analysis of the GTEx Lung Transcriptome". In: *The American Journal of Human Genetics* 99.3, pp. 624–635.

McKellar, David W. et al. (2020). "Strength in numbers: Large-scale integration of single-cell transcriptomic data reveals rare, transient muscle progenitor cell states in muscle regeneration". In: *bioRxiv*, p. 2020.12.01.407460.

Mellis, Ian A et al. (2020). "Responsiveness to perturbations is a hallmark of transcription factors that maintain cell identity". In: *bioRxiv*, p. 2020.06.11.147207.

Melsted, Páll et al. (2021). "Modular, efficient and constant-memory single-cell RNA-seq preprocessing". In: *Nature Biotechnology* 39.7, pp. 813–818.

Missarova, Alsu et al. (2021). "geneBasis: an iterative approach for unsupervised selection of targeted gene panels from scRNA-seq". In: *bioRxiv*, p. 2021.08.10.455720.

Miyazaki, Kazuko and Masaki Miyazaki (2021). "The Interplay Between Chromatin Architecture and Lineage-Specific Transcription Factors and the Regulation of Rag Gene Expression". In: *Frontiers in Immunology* 12, p. 659761.

Mohammadi, Shahin, Jose Davila-Velderrain, and Manolis Kellis (2019). "Reconstruction of Cell-type-Specific Interactomes at Single-Cell Resolution." In: *Cell systems* 9.6, 559–568.e4.

Moreira-Teixeira, Lúcia et al. (2020). "Mouse transcriptome reveals potential signatures of protection and pathogenesis in human tuberculosis". In: *Nature Immunology* 21.4, pp. 464–476.

Mosmann, T R et al. (1986). "Two types of murine helper T cell clone. I. Definition according to profiles of lymphokine activities and secreted proteins." In: *The Journal of Immunology* 136.7, pp. 2348–2357. eprint: https://www.jimmunol.org/content/136/7/2348.full.pdf.

Mukai, Kaori et al. (2012). "Critical role of P1-Runx1 in mouse basophil development". In: *Blood* 120.1, pp. 76–85.

Munugalavadla, Veerendra et al. (2005). "Repression of c-Kit and Its Downstream Substrates by GATA-1 Inhibits Cell Proliferation during Erythroid Maturation". In: *Molecular and Cellular Biology* 25.15, pp. 6747–6759.

Nilsson, Alexandra Rundberg, Cornelis J.H. Pronk, and David Bryder (2015). "Probing hematopoietic stem cell function using serial transplantation: Seeding characteristics and the impact of stem cell purification". In: *Experimental Hematology* 43.9, 812–817.e1.

Noble, William S (2009). "How does multiple testing correction work?" In: *Nature Biotechnology* 27.12, pp. 1135–1137.

Nowak, Daniel, Daphne Stewart, and H Phillip Koeffler (2009). "Differentiation therapy of leukemia: 3 decades of development". In: *Blood* 113.16, pp. 3655–3665.

Oberlaender, Marcel et al. (2011). "Three-dimensional axon morphologies of individual layer 5 neurons indicate cell type-specific intracortical pathways for whisker motion and touch". In: *Proceedings of the National Academy of Sciences* 108.10, pp. 4188–4193.

Olsson, Andre et al. (2016). "Single-cell analysis of mixed-lineage states leading to a binary cell fate choice". In: *Nature* 537.7622, pp. 698–702.

Orkin, Stuart H. (2000). "Diversification of haematopoietic stem cells to specific lineages". In: *Nature Reviews Genetics* 1.1, pp. 57–64.

Orkin, Stuart H. and Leonard I. Zon (2008). "Hematopoiesis: An Evolving Paradigm for Stem Cell Biology". In: *Cell* 132.4.

Park, Jong-Eun et al. (2020). "A cell atlas of human thymic development defines T cell repertoire formation". In: *Science* 367.6480, eaay3224.

Pellin, Danilo et al. (2019). "A comprehensive single cell transcriptional landscape of human hematopoietic progenitors". In: *Nature Communications* 10.1, p. 2395.

Plessis, Louis du, Nives Škunca, and Christophe Dessimoz (2011). "The what, where, how and why of gene ontology—a primer for bioinformaticians". In: *Briefings in Bioinformatics* 12.6.

Qiu, Xiaojie et al. (2017a). "Reversed graph embedding resolves complex single-cell trajectories". In: *Nature Methods* 14.10, pp. 979–982.

Qiu, Xiaojie et al. (2017b). "Single-cell mRNA quantification and differential analysis with Census". In: *Nature Methods* 14.3, pp. 309–315.

Qiu, Xiaojie et al. (2020). "Inferring Causal Gene Regulatory Networks from Coupled Single-Cell Expression Dynamics Using Scribe". In: *Cell Systems*.

Ranzoni, Anna Maria et al. (2020). "Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis". In: *Cell Stem Cell*.

Regev, Aviv et al. (2017). "The Human Cell Atlas". In: *eLife* 6, e27041.

Ribatti, Domenico (2018). "An historical note on the cell theory". In: *Experimental Cell Research* 364.1, pp. 1–4.

Risso, Davide et al. (2014). "Normalization of RNA-seq data using factor analysis of control genes or samples". In: *Nature Biotechnology* 32.9, pp. 896–902.

Roberto, Raphaël B. Di et al. (2021). "speedingCARs: accelerating the engineering of CAR T cells by signaling domain shuffling and single-cell sequencing". In: *bioRxiv*, p. 2021.08.23.457342.

Rodriguez-Fraticelli, Alejo E. et al. (2020). "Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis". In: *Nature* 583.7817, pp. 585–589.

Rossi, Derrick J. et al. (2005). "Cell intrinsic alterations underlie hematopoietic stem cell aging". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.26, pp. 9194–9199.

Rothenberg, Ellen V. (2021). "Single-cell insights into the hematopoietic generation of T lymphocyte precursors in mouse and man". In: *Experimental Hematology* 95, pp. 1–12.

Rozowsky, Joel et al. (2021). "Multi-tissue integrative analysis of personal epigenomes". In: *bioRxiv*, p. 2021.04.26.441442.

Santoro, Federica, Kenneth R. Chien, and Makoto Sahara (2021). "Isolation of human ESC-derived cardiac derivatives and embryonic heart cells for population and single-cell RNA-seq analysis". In: *STAR Protocols* 2.1, p. 100339.

Satoh, Yusuke et al. (2013). "The Satb1 Protein Directs Hematopoietic Stem Cell Differentiation toward Lymphoid Lineages". In: *Immunity* 38.6, pp. 1105–1115.

Scala, Federico et al. (2020). "Phenotypic variation of transcriptomic cell types in mouse motor cortex". In: *Nature*, pp. 1–7.

Schaum, Nicholas et al. (2018). "Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris". en. In: *Nature* 562.7727. Number: 7727 Publisher: Nature Publishing Group, pp. 367–372.

Schubert, Dirk et al. (2003). "Cell Type-Specific Circuits of Cortical Layer IV Spiny Neurons". In: *Journal of Neuroscience* 23.7, pp. 2961–2970.

Segerstolpe, Åsa et al. (2016). "Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes". English. In: *Cell Metabolism* 24.4. Publisher: Elsevier, pp. 593–607.

Siatecka, Miroslawa and James J. Bieker (2011). "The multifunctional role of EKLF/KLF1 during erythropoiesis". In: *Blood* 118.8, pp. 2044–2054.

Skinnider, Michael A., Jordan W. Squair, and Leonard J. Foster (2019). "Evaluating measures of association for single-cell transcriptomics". In: *Nature Methods* 16.5, pp. 381–386.

Smillie, Christopher S. et al. (2019). "Intra- and Inter-cellular Rewiring of the Human Colon during Ulcerative Colitis". In: *Cell* 178.3, 714–730.e22.

Song, Dongyuan and Jingyi Jessica Li (2020). "PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data". In: *bioRxiv*, p. 2020.11.17.387779.

Song, Liang et al. (2016). "A transcription factor hierarchy defines an environmental stress response network". In: *Science* 354.6312, aag1550.

Stephenson, William et al. (2018). "Single-cell RNA-seq of rheumatoid arthritis synovial tissue using low-cost microfluidic instrumentation". In: *Nature Communications* 9.1, p. 791.

Stoeckius, Marlon et al. (2017). "Simultaneous epitope and transcriptome measurement in single cells". In: *Nature Methods* 14.9, pp. 865–868.

Street, Kelly et al. (2018). "Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics". In: *BMC Genomics* 19.1, p. 477.

Stuart, Tim et al. (2019). "Comprehensive Integration of Single-Cell Data". English. In: *Cell* 177.7. Publisher: Elsevier, 1888–1902.e21.

Svensson, Valentine, Eduardo da Veiga Beltrame, and Lior Pachter (2020). "A curated database reveals trends in single-cell transcriptomics". In: *Database* 2020, baaa073–.

Svensson, Valentine, Roser Vento-Tormo, and Sarah A Teichmann (2018). "Exponential scaling of single-cell RNA-seq in the past decade". In: *Nature Protocols* 13.4, pp. 599–604.

Swamy, Vinay S et al. (2021). "Building the Mega Single Cell Transcriptome Ocular Meta-Atlas". In: *bioRxiv*, p. 2021.03.26.437190.

Tang, Fuchou et al. (2009). "mRNA-Seq whole-transcriptome analysis of a single cell". In: *Nature Methods* 6.5, pp. 377–382.

Tikhonova, Anastasia N. et al. (2019). "The bone marrow microenvironment at single-cell resolution". In: *Nature* 569.7755, pp. 222–228.

Torkamani, Ali et al. (2010). "Coexpression network analysis of neural tissue reveals perturbations in developmental processes in schizophrenia". In: *Genome Research* 20.4, pp. 403–412.

Tran, Hoa Thi Nhu et al. (2020). "A benchmark of batch-effect correction methods for single-cell RNA sequencing data". In: *Genome Biology* 21.1, p. 12.

Trapnell, Cole (2015). "Defining cell types and states with single-cell genomics". In: *Genome Research* 25.10, pp. 1491–1498.

Trevino, Alexandro E. et al. (2021). "Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution". In: *Cell*.

Tusi, Betsabeh Khoramian et al. (2018). "Population snapshots predict early haematopoietic and erythroid hierarchies". In: *Nature* 555.7694, pp. 54–60.

Visvader, Jane E. et al. (1997). "The LIM-domain binding protein Ldb1 and its partner LMO2 act as negative regulators of erythroid differentiation". In: *Proceedings of the National Academy of Sciences* 94.25, pp. 13707–13712.

Wang, Shuxiong et al. (2019). "Cell lineage and communication network inference via optimization for single-cell transcriptomics". In: *Nucleic Acids Research* 47.11, gkz204–.

Weinreb, Caleb et al. (2020). "Lineage tracing on transcriptional landscapes links state to fate during differentiation". In: *Science* 367.6479, eaaw3381.

Welch, Joshua D. et al. (2019). "Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity". English. In: *Cell* 177.7. Publisher: Elsevier, 1873–1887.e17.

Wolf, F Alexander, Philipp Angerer, and Fabian J Theis (2018). "SCANPY: large-scale single-cell gene expression data analysis". In: *Genome Biology* 19.1, p. 15.

Wolf, F. Alexander et al. (2019). "PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells". In: *Genome Biology* 20.1, p. 59.

Xia, Jun et al. (2021). "A single-cell resolution developmental atlas of hematopoietic stem and progenitor cell expansion in zebrafish". In: *Proceedings of the National Academy of Sciences* 118.14, e2015748118.

Xue, Yuanyuan et al. (2019). "A 3D Atlas of Hematopoietic Stem and Progenitor Cell Expansion by Multi-dimensional RNA-Seq Analysis". In: *Cell Reports* 27.5, 1567–1578.e5.

Yao, Zizhen et al. (2020a). "A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation". In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory, p. 2020.03.30.015214.

Yao, Zizhen et al. (2020b). "An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types". en. In: *bioRxiv*. Publisher: Cold Spring Harbor Laboratory Section: New Results, p. 2020.02.29.970558.

Yoshida, Hideyuki et al. (2019). "The cis-Regulatory Atlas of the Mouse Immune System". In: *Cell* 176.4, 897–912.e20.

Zappia, Luke, Belinda Phipson, and Alicia Oshlack (2018). "Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database". In: *PLOS Computational Biology* 14.6, e1006245.

Zappia, Luke and Fabian J Theis (2021). "Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape". In:

Zeira, Ron, Max Land, and Benjamin J. Raphael (2021). "Alignment and Integration of Spatial Transcriptomics Data". In: *bioRxiv*, p. 2021.03.16.435604.

Zhang, Yun et al. (2019). "The effect of tissue composition on gene co-expression". In: *Briefings in Bioinformatics*.

Zhao, Jun et al. (2021). "Detection of differentially abundant cell subpopulations in scRNA-seq data". In: *Proceedings of the National Academy of Sciences* 118.22, e2100293118.

Škunca, Nives, Adrian Altenhoff, and Christophe Dessimoz (2012). "Quality of Computationally Inferred Gene Ontology Annotations". In: *PLoS Computational Biology* 8.5, e1002533.