

Exploring Neurodevelopmental Origins of Pediatric Brain Cancer Using Transferable Topic Modeling

Ashmitha Rajendran

A dissertation submitted
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Siobhan S. Pattwell, Chair

John H. Gennari, Chair

Jay Shendure

Ali Shojaie,

Program Authorized to Offer Degree:

Biomedical Informatics and Medical Education

@Copyright 2026

Ashmitha Rajendran

University of Washington

Abstract

Exploring Neurodevelopmental Origins of Pediatric Brain Cancer Using Transferable Topic Modeling

Ashmitha Rajendran

Co-Chairs of the Supervisory Committee:

Siobhan S. Pattwell

Ben Towne Center of Childhood Cancer and Blood Disorders Research

John H. Gennari

Department of Biomedical Informatics and Medical Education

Brain cancers are the leading cause of cancer-related death in children and adolescents, with many childhood central nervous system (CNS) tumors originating during prenatal development. Understanding the developmental origins of pediatric brain cancers requires identifying transcriptional programs that persist from normal development into disease states. Despite remarkable advances in single-cell genomics, fundamental challenges remain in comparing cellular heterogeneity across datasets, particularly the inability of conventional clustering approaches to capture transitional cell states and the difficulty of integrating data across different sequencing technologies, timepoints, and biological contexts.

Objectives. This dissertation develops and validates a transferable topic modeling framework to identify conserved transcriptional programs across developmental and disease contexts. The central hypotheses are that (1) transcriptional signatures present in neurodevelopment persist in pediatric brain cancers, and (2) these signatures can be uncovered using topic modeling while enabling systematic comparison across diverse datasets.

Methods. We use Latent Dirichlet Allocation (LDA), a probabilistic framework from natural language processing, to single-cell transcriptomics by treating cells as "documents" and genes as "words." This approach represents each cell as a probabilistic mixture of transcriptional programs (topics) rather than assigning cells to discrete clusters, naturally capturing the continuous nature of developmental transitions. We combined LDA with Gene Set Variation Analysis (GSVA) to enable topic transfer across datasets, allowing topics learned from one context to be scored in independent datasets. The framework was applied across three biological contexts: (1) mouse embryogenesis to establish validity against comprehensive ground truth, (2) human cerebellar development and

medulloblastoma to demonstrate developmental-cancer correspondence, and (3) transcript-level analysis of TARGET pediatric cancers to reveal isoform-specific disease associations.

Results. In mouse organogenesis, topic modeling successfully recovered 50 transcriptional programs organized into 11 lineage-specific groups that recapitulated known developmental hierarchies while revealing transitional states invisible to conventional clustering. Topics subdivided major cell lineages (hematopoietic, neural, epithelial) and captured developmental gradients from progenitor to mature cell identities.

In human cerebellar development and medulloblastoma, developmental topics transferred successfully across timepoints and sequencing technologies. Specific developmental programs showed enrichment in molecularly defined medulloblastoma subgroups, with developmental topic enrichment correlating with clinical outcomes. This established topic transfer as a viable approach for linking normal development to cancer biology.

Extension to transcript-level resolution revealed that alternative isoforms organize according to developmental context rather than protein structure or function. Analysis of the *NTRK* receptor tyrosine kinase family identified three focus topics with distinct disease associations: an *NTRK3*-containing topic capturing sympathoadrenal differentiation signatures enriched in neuroblastoma, and two *NTRK2*-containing topics representing distinct renal developmental programs differentially enriched in Wilms tumor, clear cell sarcoma of the kidney, and rhabdoid tumor. Critically, full-length and truncated isoforms of the same gene showed distinct topic associations and disease enrichment patterns, demonstrating that isoform-level analysis reveals regulatory complexity invisible to gene-level approaches.

Conclusions. Topic modeling provides a generalizable computational framework for identifying conserved transcriptional programs across developmental and disease contexts. The approach successfully scales from mouse to human systems, from single-cell to bulk sequencing, and from gene-level to isoform-level resolution. Key contributions include: (1) demonstration that topics capture transitional developmental states and hierarchical organization within lineages, (2) validation that developmental programs persist in pediatric cancers and correlate with clinical outcomes, (3) revelation that alternative splicing creates context-dependent expression programs organized by developmental lineage, and (4) establishment of a publicly available, technology-agnostic framework for cross-dataset comparison.

The findings support developmental context as a fundamental determinant of both normal and pathological gene expression programs. Future directions include functional validation of identified programs, spatial transcriptomics integration, long-read sequencing validation of isoform predictions, and extension to additional pediatric cancer types and adult cancers with developmental origins.

Significance. This work establishes topic modeling as a powerful analytical framework for understanding how developmental programs organize at multiple scales of resolution and how these programs contribute to disease. The transferable nature of the approach enables systematic comparison across the growing landscape of single-cell atlases without requiring computationally intensive integration methods. The framework is publicly available and broadly applicable to diverse biological contexts where identification of conserved expression programs is relevant, from comparative developmental biology to disease modeling and therapeutic target identification.

For little Ashmi. This one is for you.

Acknowledgements

This research story, like any story worth telling, has been shaped by many hands and voices--a symphony of mentors, friends, and chosen family who transformed ambition into form and chaos into constellations.

Dr. Siobhan Pattwell accepted my chaotic energies and animated spirit, let me be authentic and candid and make my own mistakes to find my way. You are a rigorous scientist, therapist, mentor, cheerleader all at once. You heal work trauma. You have created something so rare that people will not want to leave, and I count myself blessed beyond measure to have been chosen by you and to be part of the Pattwell science family. This gratitude will thread through everything I do in my career. I hope we will reach out, collaborate, and send mentees to one another for years to come.

Every Friday morning, Dr. John Gennari revamps my excitement for science. Our conversations and morning coffee are a gift. You make it easy to speak of hurdles and setbacks, victories both scientific and personal. I hope you will keep making time for my silly tales, and I promise to send you a copy of my book, even if it's a vanity publish.

My committee--Dr. Jay Shendure, Dr. Ali Shojaie, Dr. Eleanor Chen--has provided diverse guidance that proved immeasurably helpful. Jay, your openness and support have been immensely encouraging. I appreciate how you have been extraordinarily kind and available. Ali, thank you for urging me toward focus, for helping me find concrete ground beneath broad ideas. Eleanor, that first piece of feedback you gave me, to explore cancer persistence versus initiation, sparked collaborations across Seattle Children's that continue to bloom. I am so grateful.

I could not have gotten here without my unofficial committee whose continued support, time, and patience with my rants sustained me through this journey. Dr. Kathleen Millen and Dr. Parthiv Haldipur, thank you for showing me the beauty in mosaics of cellular differentiation and migration in cerebellar development and also being cheerleaders. I couldn't have gotten through the everyday without the constant support of Dr. Leyre Merino Galan and Dr. David Johnson.

In the Pattwell lab, I found best friends, my people and science all in one rare and precious space. My gratitude is undying.

The BIME community, especially my cohort, carried me through the first two years--dimly lit game nights, karaoke, dinners and bars are a majority of the reason I survived our classes and qualification exams. I know you will be longtime friends of mine. Ehsan and Atefe have become part of my chosen family and home in Philly (for now) and Bhargav a definite lifelong friend.

Craft night crew--my close friends, Kiran, Emily, Emma, Sherina, Hawa--welcomed me when I showed up tired, listened to my rants, forgave my periodic disappearances for deadlines. We can finally cross off that bingo square.

My hometown friends scattered across states and Canada--Atmu, Pranz, Sam, Ash, Nyk, GK, Monisha, Kish, Morgan, Swati, Samir, and Shar--and my kudumbam, Mathuri and Praa, have kept me humble and given me homes everywhere I go.

My spouse, beautiful and wonderful and phenomenal beyond words--it is impossible to speak of you coherently. Saying I love you, that you are supportive, feels stupid and insufficient. This dissertation is ours as much as it is mine--every word, struggle, triumph was shared (literally) and I'm not sure what universal being I satisfied to get to be your wife.

It is impossible to think of this achievement without thinking of the privilege that carried me here. I will be the first doctor on either side of my family, and this represents the immense work and dedication of my parents who came from poverty in India--especially my father, one of the first to finish high school, who put himself through college while supporting his family. Our relationship may not be movie-perfect, but it is perfect for me to be who I am now, proud of who I am. I owe them the world and more.

I must also acknowledge that the privilege of caste, though some say it does not exist in the US, is a major reason I am where I am. Acknowledgment alone is insufficient. I promise to continue speaking about caste and its effects, to amplify the voices like Thenmozhi Soundararajan and Sujatha Gidla and so many others.

It is impossible to think of this PhD defense without also thinking of the countless people in Sudan, Congo, Palestine who have been robbed of the ability to even dream of higher education. As I write this, Palestine is recovering from ashes caused by genocide perpetrated by Israel and their close partner and conspirator, the US. Genocides continue in Sudan and the Congo. Iran has been in internet darkness for over a month. Minnesota is occupied and occupation spreads across the US in ways that mimic what native folks have already experienced and continue to see. Even without seeking the daily news, these global affairs seep into daily life--grants unfunded, colleagues let go, friends and family unable to stay in the US despite a decade of waiting, neighbors kidnapped, the rights of myself and my partner and my family and nearly all my friends actively targeted.

I am grateful everyday for the privilege of resistance--of pursuing research, finding joy in small moments, relying on community, having resources to activate change. However, the nervous system was not built for continuous onslaught, for this relentless violence. This dissertation was written through days of numbness, grief, and fear, but also through days of intense community-building, gratefulness, and hope.

Frequently I found myself wondering if this career was one I could live with, knowing other skills might be necessary, my abilities better used elsewhere. Through this process I learned that showing up every day is itself resistance. The skills I have gained are not just technical--thinking in code, dreaming in terminal, deciphering one astrocyte from another. They are critical thinking, the ability to separate myself from initial reactions to understand the situation at hand, the ability to work under pressure until my eyelids can't stay open a second longer because I am passionate and capable. I have learned to see overlaps--the abstraction and statistics of the lab connect to learning the best native smart mulching techniques for the farm.

I vow to use the privilege of these skills and this title to traverse systems and rooms, to make change for the better--to show that diversity, queerness, all abilities belong and deserve equity and inclusion. I vow to ensure that the institutions I am part of will hear from me relentlessly, that they must do better, that their money and influence must better our society rather than contribute to exclusive rings of corrupt and unlawful syndicates.

This is evidence of my resistance.

Table of Contents

1. CHAPTER 1. INTRODUCTION	10
1.1. Research Objectives and Dissertation Overview	10
1.2. Contributions to the Field	10
2. CHAPTER 2: BACKGROUND	12
2.1. The Single-Cell Genomics Revolution	12
2.2. Topic Modeling Framework	12
2.3. Neurodevelopmental Origins of Pediatric Cancers	16
2.4. The Framework and Structure of this dissertation	17
3. CHAPTER 3. Single Cell Topic Modeling Identifies Transitional Cell Types in Mouse Fetal Development	18
3.1. Background in Mouse Organogenesis and Developmental Cell Types	18
3.2. Background in Current Single Cell Analysis Approaches	18
3.3. Methods	21
3.4. Results	23
3.5. Discussion	31
3.6. Conclusion	32
4. CHAPTER 4: Transferable Topic Modeling Reveals Developmental Cellular Lineage Persistence in Medulloblastoma	33
4.1. Background in Cross Dataset Developmental and Pediatric Cancer Studies	34
4.2. Materials and Methods	36
4.3. Results	37
4.4. DISCUSSION	49
5. CHAPTER 5. NTRK Family Isoform Programs Distinguishes Pediatric Cancers Revealed by Transcript-Level Topic Modeling	52
5.1. Background	52
5.2. Methods	54
5.3. Results	55
5.4. Discussion	61
5.5. Limitations	62
5.6. Future Directions	63
5.7. Conclusion	63
6. CHAPTER 6. Discussion, Conclusion, and Future Directions	64
6.1. Technical Challenges: Genome Versions, Transcript Annotations, and Reproducibility	64
6.2. Ethical Considerations of Embryonic Tissue Research	66
6.3. Future Directions	68
6.4. Contributions	69
6.5. Limitations	69
6.6. Conclusion	70
7. APPENDIX	70
7.1. Chapter 3 Supplemental Topic Groups	70

List of Figures and Tables

Figure 3.1 Elbow plot of the rate of perplexity change (RPC, y-axis) between iterations of topic modeling with changing numbers of topics (K, x-axis).	22
Figure 3.2. Topic-trajectory mean association scores reveal lineage-specific transcriptional programs during mouse organogenesis.	24
Figure. 3.3. Dot plot of topic expression across the 26 major developmental trajectories.	25
Table 3.1. Mouse organogenesis cell atlas derived topic groupings based on hierarchical clustering and gene composition similarities.	25
Figure 3.4. Topic-trajectory Pearson correlation heatmap reveals lineage-specific transcriptional programs.	27
Figure 3.5. Topic-topic correlation matrix.	28
Figure 3.6. CNS topics capture neural tube development and neuronal/glial diversification.	30
Figure 4.1. Computational workflow for topic modeling and transfer from Source dataset to two Applied datasets.	36
Figure 4.2. Spatial organization of neurogenic and gliogenic cell populations across the external granule layer during cerebellar development.	38
Figure 4.3. Topic modeling reveals distinct transcriptional programs defining rhombic lip progenitor lineage trajectories in human cerebellar development.	41
Figure 4.4. Cross-dataset validation demonstrates robust conservation of rhombic lip lineage topics from source dataset in an independent human developmental cerebellar dataset (applied data).	43
Figure 4.5. Spatial validation of EGL-associated topic marker genes in developing mouse cerebellum at postnatal day 4.	46
Figure 4.6. Atlas of medulloblastoma bulk RNA sequencing (Applied dataset 2).	47
Figure 4.7. SHH subtypes express different developmental source topics and correspond to specific stages of EGL transitions.	48
Table 5.1. Select NTRK Family Transcripts Detected in the TARGET Pediatric Cancer Topic Models	54
Figure 5.1. TARGET cohort overview.	56
Figure 5.2. Disease-specific expression of individual NTRK family transcripts.	58
Figure 5.3. Heatmap of NTRK transcript β weights across topics.	58
Table 5.2. Focus topics selected for detailed characterization based on NTRK transcript rank.	58
Figure 5.4. Disease-specific enrichment of the three focus NTRK topics.	59
Figure 5.5. Transcript composition of the three focus NTRK topics.	60
Figure S1. Group 1: Hematopoietic lineage topics capture the full hierarchy of blood cell development.	72
Figure S2. Group 2: Tissue myeloid topics represent functional states of tissue-resident macrophages.	73
Figure S3. Group 3: Myogenic topics capture skeletal and cardiac muscle differentiation from mesoderm.	75
Figure S4. Group 4: Mesenchymal/skeletal topics delineate mesenchymal stem cell differentiation into connective tissue lineages.	77
Figure S5. Group 5: Epithelial topics capture epithelial programs across distinct germ layer origins.	78
Figure S6. Group 6: Sensory/PNS topics mark peripheral nervous system and sensory neuron development from neural crest and cranial placodes.	79
Figure S7. Group 8: Endodermal organ topics mark hepatocyte and intestinal epithelium differentiation from definitive endoderm.	80
Figure S8. Group 9: Adipocyte topics distinguish metabolic and regulatory programs in adipose tissue.	81
Figure S9. Group 10: Vascular topic marks endothelial cell identity.	82
Figure S10. Group 11: Ciliated topic captures shared motile cilia program across developmentally distinct cell types.	83
Figure S11. Group 12: Cell cycle topic represents a universal proliferation signature across lineages.	83
Figure S12. Ungrouped topics display diffuse expression patterns or ambiguous biological signatures.	86
Table S1. Summary of Topic Identities, Key Marker Genes, and Expression Specificity.	86

Preamble

The Inbetweens

A book can be a fortress, a portal to where one is hurdled through places of wild imaginations, traversing deserts with dragons, crossing rivers with no bridges, finding gardens with magic fruit, and willows with leaves intertwined with drops of jewels. The power wielded on paper can create fires that rise up, singe the hair, and make the eyelids sting. They create the fair and fatal roads of which even dust and stone can be ornamented in a way to be held dear to the heart. In the past years, with the sanguineness of a knight-errant, I embraced the diligence of the schoolroom, the endless battles with numerical magics and statistical wizardry, crawled through space yet to be charted while allowing the jagged stones and unfettered growth of the path to rip at my shoulders and shins... all to grace in the privilege of creating and unearthing and sharing an adventure.

Though the findings and results do push forth our worldly understanding at a miniscule scale--I cherish from this adventure new perspectives---I have been changed in a way that breaks my previous molds of life, broadens my understandings of purpose and belonging and compassion, and teaches me new ways of thinking, speaking, and sorcery.

My community's patient endurance of my rambling and moodiness while I traversed the labyrinthine of the growth pains of the mid-late 20s is why I reach this final curtain of flame and am ready to embrace the uncertainty and chaos with new found confidence. Though the final story written here remains brief (ish) and incomplete---a sweet fleeting glance, a soft feather in the wind---I have learned that the greatest part of an adventure is during, in the in-between.

I don't suspect the story here to linger past the pages of this book (or perhaps an additional manuscript or two), however, I hope to showcase the epic and exhilarating adventure of the in-between.



1. CHAPTER 1. INTRODUCTION

The developing human brain represents one of the most complex and precisely orchestrated biological systems. Understanding cellular heterogeneity during brain development has profound implications for both basic neuroscience and clinical medicine, particularly in pediatric brain cancers where developmental programs can persist in disease states. Despite remarkable advances in single-cell genomics that have transformed our ability to characterize cellular diversity, fundamental challenges still remain in being able to understand underlying cell types and differentiation across datasets..

1.1. Research Objectives and Dissertation Overview

Brain cancers are the leading cause of cancer related death in children and adolescents.¹ Many of these childhood central nervous system (CNS) cancers have prenatal origins. Discerning the developmental links that promote cancer progression and development has been challenged by limited availability in individual datasets that can be integrated and methods for decoding these granular shared patterns. Here, we integrate and expand upon several of the largest single cell atlases of the human and mouse prenatal brain and across several pediatric brain tumor types to identify developmental transcriptional programs and cellular origins of tumor initiation. The goal of this dissertation is to demonstrate an additional -omics approach to further improve our understanding of cell types and transitions in the context of neurodevelopment and pediatric brain cancers. I accomplish this through the use of topic modeling which is a probabilistic framework borrowed from natural language processing that treats cells as "documents" and genes as "words" to identify conserved transcriptional programs. The central hypotheses of this dissertation is that *(1) there are transcriptional signatures in neurodevelopment that persist in pediatric brain cancers and (2) which can be uncovered with topic modeling while enabling systematic comparison of developmental and disease states across diverse datasets and technologies.*

This dissertation demonstrates the development, validation, and application of a transferable topic modeling framework across three biological contexts (elaborated in chapters 2, 3, and 4) of increasing complexity and translational relevance.

Chapter 2 provides background on single-cell RNA sequencing technology, the major computational challenges in the field, and how topic modeling addresses these limitations. This chapter focuses on the history and evolution of single-cell genomics, big-picture problems in single-cell and bulk sequencing approaches, and comprehensive background on topic modeling methodologies including Latent Dirichlet Allocation (LDA) and its applications in genomics.

Chapter 3 demonstrates the use of topic modeling in mouse embryogenesis. This is a biological context where data is comprehensive and extensively characterized which allows for robust validation through the existing literature. Using the mouse organogenesis cell atlas, I show that topics can subdivide all primary cell type trajectories initially identified through conventional clustering approaches. I focus primarily on central nervous system (CNS)-related cell types to demonstrate the additional specificity and sensitivity of cell types identifiable with topics, while comprehensive analysis of all 13 major developmental groups is presented in supplementary materials.

Chapter 4 demonstrates how the same topic modeling approach can be applied to human neurodevelopment, where data is more scarce and biological knowledge is less comprehensive. I develop topics from human cerebellar development and show these topics transfer successfully across independent developmental datasets. I then apply these neurodevelopmental topics to medulloblastoma to showcase whether our transfer approach works across the developmental-cancer boundary. I highlight specific neurodevelopmental signatures (topics) that relate to pediatric brain cancer subtypes and clinical outcomes.

Chapter 5 explores a further use case of topic modeling by dissecting genes into their constituent transcripts to create more specific transcript/isoform-related signatures for pediatric brain cancers. This chapter extends topic modeling to isoform-level resolution across multiple pediatric and adult brain tumor types from the TARGET cohort, demonstrating that alternative splicing creates context-dependent isoform programs invisible to gene-level analysis. I demonstrate this principle through detailed examination of *NTRK2*/TrkB receptor isoforms, showing that developmental context rather than functional classification determines which isoforms are co-expressed.

1.2. Contributions to the Field

This dissertation makes several distinct contributions. From a bioinformatics and biomedical informatics standpoint, this transferable topic modeling framework enables cross-dataset comparison applicable across sequencing technologies (single-cell to bulk), species (mouse to human), and biological contexts (development to disease). From a pediatric neuro-oncology perspective, I demonstrate that medulloblastoma subtypes retain specific developmental signatures that associate with age of onset and clinical outcomes. Additionally, with the extension of topic modeling to isoform-level analysis, I highlight how splicing programs and alternative isoforms of the same genes can have distinct disease associations. From a developmental neuroscience approach, I identify cell type and developmental transition signatures of neurodevelopmental processes, including transitional states that enhance conventional clustering approach results.

This work not only establishes a generalizable and accessible computational framework appropriate across sequencing technologies and biological context, we also begin identifying and creating genetic signatures from these topics that can then be validated in vivo in future studies. Additionally, this framework is publicly available and can be explored in different bioinformatics inquiries.



2. CHAPTER 2: BACKGROUND

This chapter provides background on the foundational concepts underlying this dissertation. Rather than cataloging technical details better suited to individual chapters, this background focuses on the big-picture themes that unite the work throughout. These themes include a brief background and limitations of single-cell genomics, the fundamental problems these limitations create for understanding biology, how topic modeling may be a suitable approach to address these problems, and why the developmental origins of cancer provide the biological context where these methods prove valuable.

2.1. The Single-Cell Genomics Revolution

For much of molecular biology's history, gene expression studies measured average signals across millions of cells. Bulk RNA sequencing shares what genes are expressed in a tissue sample. Though this technique continues to be powerful for overall gene or transcript expression, using the population averages obscures the cellular heterogeneity. A tissue biopsy contains dozens of cell types in varying proportions, each with distinct expression patterns. Developmental transitions involve rare intermediate states. Tumors harbor subpopulations with different proliferative capacities and drug sensitivities. All of this variation is invisible when the tissue is homogenized and when methods measure aggregate expression. Single-cell RNA sequencing fundamentally changed this paradigm by enabling transcriptome-wide profiling of individual cells. Technologies evolved from early plate based methods analyzing tens to hundreds of cells, to droplet-based approaches which can capture tens of thousands, to combinatorial indexing methods^{2,3} now profiling millions of cells in single experiments. These advances have generated comprehensive atlases of mouse⁴⁻⁸ and human development, adult tissues across organs^{9,10}, and disease states including cancer^{11,12}.

One major finding from being able to assess each individual cell is that "cell types" defined by classical markers actually represent continua of states. Developmental lineages branch through intermediate populations expressing mixed signatures of progenitor and differentiated identities. Tumors contain hierarchies of cells spanning stem-like to differentiated states. The resolution of single-cell analysis has revealed complexity invisible to bulk approaches, reshaping our understanding of development, homeostasis, and disease. Despite this revolution, fundamental limitations have emerged that constrain our ability to synthesize findings across studies and translate discoveries to clinical contexts. Here, I describe three limitations that reflect deep conceptual challenges in how we analyze and compare single-cell data.

(1) The discrete classification problem: Standard single-cell analysis workflows rely heavily on clustering algorithms¹³ that assign each cell to a single discrete cluster. This approach assumes cells belong to distinct, separable groups. Though this is powerful for identifying the major cellular or subcellular trajectories, it limits the ability to identify continua or transitory cells. This is particularly important during development and in disease as these cases involve cells existing in transitional states expressing multiple transcriptional programs simultaneously. A cell differentiating from progenitor to neuron expresses both proliferation genes and early neuronal markers. A tumor cell may express stem-like programs alongside differentiation signatures. Forcing such cells into discrete clusters either creates arbitrary boundaries along continuous trajectories or lumps together cells at different transition stages, losing critical biological information.

(2) The cross-dataset comparison problem: A fundamental methodology of science is to synthesize findings across studies to build generalizable knowledge. In single-cell genomics, this synthesis faces severe challenges. Technical variation between experiments can be as large as biological variation of interest. Different sequencing platforms, dissociation protocols, laboratories, and time points introduce systematic differences that confound biological interpretation. Existing batch correction methods¹⁴⁻¹⁶ can align datasets but risk removing genuine biological signals when technical effects are confounded with biology of interest. More critically, these methods require datasets to share at least some cell populations, fail when comparing radically different biological contexts like development versus adult disease, and cannot bridge the gap between single-cell and bulk RNA sequencing that dominates clinical datasets.

(3) The annotation inconsistency problem: Different studies apply different cell type nomenclature, resolution criteria, and annotation standards. The same population may be called "neural progenitors" in one study, "radial glia" in another, and subdivided into multiple subtypes in a third. This inconsistency undermines systematic comparison and meta-analysis. A method that could identify conserved gene expression programs beneath superficial annotation differences would significantly advance the field.

These three problems create barriers to understanding developmental trajectories and their persistence in disease and motivate the need for alternative analytical frameworks. This is particularly important for a breadth of fundamental research questions that integrate different biological contexts such as those involving developmental origins of disease. Here, I propose topic modeling as an additive methodology to the single cell analytical pipelines to address the challenges presented above.

2.2. Topic Modeling Framework

Topic modeling originated in natural language processing as a method for discovering latent themes in document collections¹⁷. The intuition of this methodology is as follows: documents are mixtures of topics, and topics are probability distributions over words. A news article might be 60% about politics, 30% about economics, and 10% about international relations. Each of these topics is defined by words that tend to appear together i.e. "politics" topics contain words like "election," "senator," and "legislation," while

"economics" topics contain words like "inflation," "market," and "recession." Topic modeling does not require knowing topics in advance. Given a collection of documents, the algorithm infers both what topics exist and how much each document expresses each topic. This unsupervised discovery approach allows researchers to identify themes even in unfamiliar corpora.

The mathematical framework most commonly used is Latent Dirichlet Allocation (LDA)¹⁷. LDA assumes a generative process as follows: (1) with a document from collection of documents, sample a distribution over topics from a prior distribution, then (2) for each word position, sample a topic according to the document's topic distribution, and (3) finally sampling a word according to that topic's word distribution. Inference reverses this process, learning the topic-word distributions and document-topic distributions that best explain the observed words.

2.2.1. Adapting Topic Modeling to Genomics

The adaptation of topic modeling to single cell sequencing follows a similar intuition with the analogy of cells are akin to documents and genes akin to words. In this case, gene expression counts are word frequencies and topics become co-expressed gene programs. For example, a cell (for single cell sequencing) or sample (for bulk RNA sequencing) might be 40% proliferation program, 30% neural differentiation program, 20% response to signaling, and 10% metabolic regulation. Each program is defined by genes that tend to be expressed together. This type of framing addresses the discrete classification problem directly. Unlike clustering, which assigns each cell to a single cluster, topic modeling represents each cell as a mixture of programs with varying intensities. A cell transitioning from progenitor to differentiated state naturally expresses both programs at intermediate levels. There are no forced boundaries and the transition is captured as continuous variation in topic proportions.

The approach also differs fundamentally from other dimensionality reduction methods commonly used in single-cell analysis approaches. For instance, principal component analysis (PCA)¹⁸ finds orthogonal axes of maximum variance, which is useful for capturing major patterns of variation and reducing noise, but produces components that mix positive and negative gene weights, making biological interpretation difficult. Non-negative matrix factorization (NMF)¹⁹ provides similar decomposition to topic modeling but lacks the probabilistic interpretation in that NMF is purely a matrix factorization technique without an underlying generative model, whereas LDA provides a full probabilistic framework with interpretable priors and posterior distributions that enable uncertainty quantification and principled statistical inference. Independent component analysis (ICA)²⁰ assumes statistical independence between components, meaning the components cannot co-occur, which is inappropriate for biological programs that often co-occur simultaneously within cells.

Topic modeling's probabilistic framework provides several advantages. Each cell has a distribution over topics representing its current state. Each topic has a distribution over genes representing a program or functional signature. These distributions can be examined directly for biological interpretation as the level of topics or the genes it encompasses. This allows for soft topic assignments rather than hard which then handles uncertainty and mixed states. Additionally, as we will further examine, topics defined by their gene distributions can be transferred to new datasets without requiring whole data integration and transfer of just relevant information.

In spite of these conceptual advantages, topic modeling has seen limited adoption in single-cell genomics compared to clustering. There are several contributing factors. First, topic modeling requires choosing the number of topics, analogous to choosing cluster number but with less developed heuristics for the choice. Second, interpretation requires examining many topics rather than a smaller number of clusters, increasing cognitive load. Third, software for topic modeling was initially less mature than clustering tools especially in -omics contexts, though this gap has narrowed recently. Fourth, and perhaps most importantly, the field has developed extensive infrastructure around discrete cell type definitions such as ontologies, databases, and reference atlases that assumes cells belong to types rather than existing in mixed states.

Despite these barriers, topic modeling offers unique capabilities for the problems outlined in section 2.1 and these capabilities prove essential for understanding early developmental cellular transitions and their persistence in disease.

2.2.2. Early Applications to Bulk RNA Sequencing

The conceptual parallel between documents containing words and biological samples containing gene expression measurements motivated early applications of topic modeling to transcriptomic data. In this framework, each sample (whether bulk tissue, cell line, or tumor) is treated as a document, genes are treated as words, and the learned topics represent transcriptional signatures active across samples.

Early applications of topic modeling to bulk expression data²¹ focused on cancer genomics, where tumor heterogeneity and the coexistence of multiple transcriptional programs make mixture models particularly appropriate. Tumors are not homogeneous entities but rather complex ecosystems containing cancer cells in different states, stromal cells, immune infiltrates, and vasculature. Topic modeling can decompose bulk tumor expression profiles into contributing programs without requiring prior knowledge of what those programs represent.

Studies applying LDA to The Cancer Genome Atlas (TCGA) data^{21,22} demonstrated that topics learned from tumor expression data often corresponded to known biological processes including cell cycle, immune response, metabolism, and tissue-specific

differentiation programs. Importantly, topic modeling could identify programs active in subsets of samples that might be missed by differential expression analysis between predefined groups, enabling discovery-driven rather than hypothesis-driven analysis.

The application of topic modeling to bulk RNA sequencing established several principles that informed later single-cell applications²¹: (1) the number of topics requires careful optimization for each dataset, (2) biological interpretation of topics requires examination of top-weighted genes and enrichment analysis, (3) topic scores can be used as features for downstream analysis including survival modeling and subtype classification, and (4) topics learned from one dataset can potentially be applied to independent datasets to assess program conservation.

2.2.3. Applications to Single-Cell RNA Sequencing

The advent of single-cell RNA sequencing created new opportunities and challenges for topic modeling approaches. Single-cell data provides cellular resolution that bulk analysis cannot achieve, but also introduces substantial technical noise from dropout events, low capture efficiency, and amplification bias. The high dimensionality (thousands of genes) and large sample sizes (thousands to millions of cells) of single-cell datasets require computationally efficient implementations. Several groups have developed topic modeling frameworks specifically designed for single-cell transcriptomic analysis. These methods share the core LDA framework but differ in their inference algorithms, handling of zero-inflation, integration with other single-cell tools, and downstream analysis capabilities.

Celda (Cellular Latent Dirichlet Allocation)²³: Wang and colleagues developed Celda as a Bayesian model that simultaneously performs co-clustering of genes into modules and cells into subpopulations using single-cell RNA-seq data. Celda extends the basic LDA framework by incorporating a cell-level clustering component, enabling joint inference of gene modules (analogous to topics) and cell populations. This joint modeling can improve both gene module discovery and cell clustering by sharing information between the two tasks. Celda is implemented as part of the singleCellTK package and integrates with standard single-cell analysis workflows in R. Though this is a powerful framework, this method is not used here as it limits the parameter adjustment capabilities and is limited to preprocessed single cell RNA sequencing.

fastTopics^{24,25}: Developed by Carbonetto and colleagues, fastTopics provides a computationally efficient implementation of topic modeling using NMF for large single-cell datasets. The method uses a variational inference approach that scales to millions of cells, addressing a key limitation of Gibbs sampling-based methods for very large datasets. fastTopics also provides tools for differential topic analysis, enabling identification of topics that distinguish cell populations or conditions. In general, fastTopics finds maximum a posteriori/MLE versus Gibbs sampling samples from the full posterior distribution. Though this reduces run time (from hours to minutes for datasets of a million cells), Gibbs sampling provides for a better understanding of the uncertainty in the topic assignments. In the context of identifying known cell types, fastTopics may be the better option. In this dissertation, we are exploring rare and transitional cell types where the Gibbs approach is more appropriate.

scTM²⁶ and related methods: Several additional methods have applied topic models or related matrix factorization approaches to single-cell data. These include methods based on non-negative matrix factorization (NMF), which shares mathematical connections with topic modeling but uses different inference procedures. NMF-based methods such as cNMF (consensus NMF) identify gene expression programs and have been widely applied to identify activity programs in cancer single-cell datasets. A landmark application of topic modeling to single-cell cancer data came from the Tirosh laboratory, which used NMF to identify recurrent transcriptional programs in metastatic melanoma²⁶. This study demonstrated that single-cell topic modeling could identify programs including a resistance program associated with therapeutic response, immune evasion signatures, and cell cycle programs. The ability to score individual cells for multiple programs simultaneously revealed intra-tumoral heterogeneity that bulk analysis could not resolve. This is also a powerful approach, however,

Topic modeling offers several advantages over standard clustering for single-cell analysis. First, it provides soft assignments where cells can express multiple topics rather than belonging to a single cluster. This better reflects biological reality, as cells often simultaneously express programs related to cell type identity, cell cycle state, stress response, and other biological processes. Second, topics are defined by genes rather than by cell groupings, facilitating biological interpretation through examination of top genes and pathway enrichment. Third, the topic-gene matrix learned from one dataset can be applied to new datasets, enabling cross-dataset comparison without batch correction.

2.2.4. Applications to Single-Cell ATAC Sequencing

The success of topic modeling for transcriptomic data motivated applications to other genomic modalities, most notably single-cell ATAC sequencing (scATAC-seq), which measures chromatin accessibility at single-cell resolution. scATAC-seq data presents even greater sparsity than scRNA-seq, as each cell contains only two copies of each genomic locus that can either be accessible or inaccessible.

cisTopic²⁷: Bravo González-Blas and colleagues developed cisTopic specifically for topic modeling of single-cell ATAC-seq data. In cisTopic, cells are treated as documents and accessible regulatory regions are treated as words. The learned topics represent sets of co-accessible regulatory regions that may correspond to cell type-specific or condition-specific regulatory programs. cisTopic has proven particularly valuable for identifying cell type-specific enhancer landscapes and for linking regulatory region accessibility to

gene expression. The method includes functionality for topic annotation through motif enrichment analysis, identifying transcription factors whose binding sites are enriched in topic-defining regions. This enables inference of regulatory mechanisms underlying each topic. A major learning from cisTopic is that chromatin accessibility topics often correspond to developmental lineages and differentiation states, consistent with the role of chromatin remodeling in establishing cell identity. Topics can capture both broadly active regulatory programs and cell type-specific enhancer landscapes, providing complementary information to transcriptomic analysis. The cisTopic framework has been extended to integrate scATAC-seq with scRNA-seq data, enabling joint analysis of chromatin accessibility and gene expression. This multimodal integration can reveal regulatory relationships between accessible chromatin regions and gene expression, supporting inference of gene regulatory networks.

2.2.5. Parameter Selection and Model Evaluation

A challenge in topic modeling is selecting the number of topics and other hyperparameters. Unlike clustering, where silhouette scores and other metrics provide guidance on cluster number, topic model selection can be less straightforward. Several techniques can be used to assist with topic number selection. (1) Perplexity²⁸ measures how well a model predicts held-out data, with lower values indicating better fit. While commonly used for topic model selection, perplexity does not always correlate with human interpretability of topics. Models with lower perplexity may have more topics that are less interpretable. (2) Coherence metrics measure the semantic consistency of top words within topics, with more coherent topics expected to be more interpretable. Various coherence measures have been proposed, including metrics based on word co-occurrence statistics^{29,30} and normalized pointwise mutual information³¹. For genomic applications, coherence could be adapted to measure gene co-expression or pathway co-membership. (3) Standard model selection approaches including held-out likelihood and cross-validation^{32,33} can be applied to topic models. Cells or samples are divided into training and test sets, models are trained on training data, and performance is evaluated on held-out test data. (4) Topics should be robust to perturbations including random initialization, subsampling of cells, and subsampling of genes. Stability analysis runs the model multiple times and assesses concordance of topics across runs³⁴. Highly stable topics are more likely to represent true biological signals rather than noise. (5) Ultimately, the value of topic models for genomic analysis depends on biological interpretability and utility. Topics should correspond to known biological processes (validated through enrichment analysis), should be reproducible across datasets, and should provide insights not available from standard clustering approaches. The optimal number of topics may depend on the biological question: fewer topics for broad cell type identification, more topics for fine-grained cell state analysis.

2.2.6. Topic Transfer Framework for Cross Dataset Comparisons

Topics are defined by their top genes. A "neuronal differentiation" topic might be characterized by genes including quintessential differentiation genes such as *Neurog2*, *Neurod1*, *Elavl3*, and *Tubb3* in a ranked order but may also include genes that overlap with a "cell cycle" topic that may be defined by *Mki67*, *Top2a*, and *Ccnb1*. These gene sets represent the transcriptional program. We can then take these gene sets that we have learned from one dataset and understand their correspondence in another. By enabling cross-dataset comparison we can score these gene sets in any dataset where those genes are measured. Gene Set Variation Analysis (GSVA)³⁵ provides a non-parametric method for computing enrichment scores at the single-sample or single-cell level. Given a gene set and an expression profile, GSVA calculates how strongly that sample expresses the coordinated program defined by the gene set.

This creates a workflow for topic transfer. First, learn topics from a reference dataset by identifying co-expressed gene programs. Second, extract the top genes defining each topic as gene sets. Third, score these gene sets in target datasets using GSVA, quantifying how strongly each cell or sample expresses each topic's program. This approach enables comparison across datasets without requiring shared technology. In this dissertation, I apply this approach to single and bulk RNA sequencing, demonstrating this transfer capability cross technologies.

The topic transfer framework proves particularly powerful for connecting developmental and disease datasets. The challenge of mapping tumors to "normal" developmental stages includes high-resolution genomics atlases of brain development profiled with one technology, and genomics of hundreds of tumors profiled with the same or different technologies where all of the datasets are from different labs and studies. We learn developmental topics from the developmental atlas, identifying topics related to "proliferating granule neuron progenitors" and "differentiating granule neurons" for instance. We then score these developmental programs in bulk or single cell tumor samples. Tumors expressing high levels of the proliferative progenitor program may represent arrested development at that stage. This comparison preserves the original data, and naturally handles the technology difference.

The same framework enables comparing datasets across laboratories, timepoints, and species. Topics learned from one developmental stage can be scored at earlier or later stages to track program dynamics. The main requirement is overlapping gene coverage--if both datasets measure the same genes, topics can be transferred.

Topics remain biologically interpretable throughout this process. Unlike batch correction methods that transform expression values in opaque ways, topic transfer keeps the gene sets visible. We can examine which genes define each topic, perform pathway enrichment, compare to known markers from literature, and validate through orthogonal data like spatial expression patterns. This interpretability proves essential for biological discovery. When a topic transfers successfully across datasets, we can explore what biological program these co-expressed genes represent. When a topic shows strong scores in tumors, we can identify specific genes driving the signature and whether they suggest therapeutic targets. The framework enables both discovery and mechanistic follow-up. As mentioned in the

above sections, this is particularly useful for biological questions that require the combination of contexts. Here, we explore the neurodevelopmental contexts that contribute to pediatric brain cancers.

2.3. Neurodevelopmental Origins of Pediatric Cancers

Embryonic development involves continuous differentiation trajectories where progenitor cells progressively acquire different levels of maturation of cell identities through tightly regulated developmental programs^{36,37}. Understanding these developmental mechanisms is fundamental to understanding where aberrations can lead to dysregulation and downstream diseases, including pediatric cancers. Unlike malignancies arising later in life that accumulate somatic mutations over years, pediatric cancers typically carry significantly fewer somatic mutations but instead exhibit profound transcriptomic dysregulation³⁸. This observation has led to the theory of developmental origins for cancers where many pediatric cancers arise when normal development becomes dysregulated and traps cells in aberrant progenitor states or blocking normal differentiation trajectories.

2.3.1. Developmental Processes in Humans and Mice

Mammalian development proceeds through highly conserved but temporally distinct programs in humans and mice. While the fundamental molecular pathways and cell type specifications are preserved across species, the timescales do differ dramatically. Human brain development, for instance, is highly protracted compared to mice, with a sixfold longer maturation period³⁹. Cerebellar development in humans extends from approximately post-conception week 9 through the first two postnatal years, whereas in mice this process spans from embryonic day 12 through postnatal day 21. Despite these temporal differences, the core developmental logic remains conserved where progenitor populations arise from specific germinal zones, undergo proliferative expansion driven by developmental signaling pathways, and differentiate along defined lineages marked by sequential activation of lineage-specific transcription factors. These developmental processes involve continuous state transitions rather than discrete cell type switches^{37,40}. Progenitor cells traverse intermediate states characterized by mixed transcriptional signatures while differentiating. A differentiating neural progenitor, for example, progressively downregulates stem cell maintenance programs while upregulating neuronal differentiation markers, passing through transitional states that express both programs simultaneously. This continuous nature of developmental transitions has important implications for understanding how dysregulation at specific developmental stages can produce tumors with distinct molecular and clinical characteristics.

2.3.2. The Developmental Origins of Pediatric Tumors

This developmental origins hypothesis posits that pediatric tumors arise from specific developmental progenitor populations whose normal regulatory mechanisms have been subverted. This paradigm emerged from converging evidence where (1) pediatric tumors frequently occur at developmental timepoints and anatomical locations where active progenitor proliferation is ongoing, (2) tumor cells retain transcriptional programs characteristic of specific developmental progenitors, and (3) functional studies have demonstrated that introducing oncogenic mutations in defined progenitor populations recapitulates human tumors, while the same mutations in other cell types do not.

The mechanisms underlying this developmental dysregulation operate at multiple levels. Genetically, driver mutations in pediatric cancers commonly target developmental signaling pathways that normally regulate progenitor self-renewal and differentiation, such as Sonic Hedgehog (SHH), WNT, and Notch signaling^{38,41}. In general, these mutations promote proliferation in healthy tissues and cells, and they also disrupt the balance between self-renewal and differentiation that defines progenitor behavior. Epigenetically, pediatric tumors show DNA methylation landscapes and enhancer activity patterns that correspond to their proposed cells of origin, suggesting that the epigenetic state established during normal development is maintained in transformed cells. This epigenetic memory may both reflect the developmental origin and actively maintain tumor identity^{41,42}.

The concept of impaired differentiation is central to this paradigm. Normal progenitors are defined by their developmental potential i.e. their capacity to both self-renew and differentiate. Transformation requires disrupting this balance, creating cells that retain proliferative capacity and fail to complete normal differentiation programs. Different modes of developmental arrest may produce tumors with distinct molecular and clinical characteristics even when arising from the same progenitor lineage.

2.3.3. Implications for Understanding Cancer Biology

The developmental origins framework has reframed some of our understanding of pediatric cancer biology. If tumors represent frozen or disrupted developmental states, then understanding normal development becomes essential for understanding cancer. The transcriptional signatures, signaling dependencies, and splicing patterns of these tumors may be inherited from their cells of origin rather than being independently evolved. Understanding which developmental stage a tumor resembles may also inform prognosis and treatment strategies. Tumors arrested at more primitive progenitor states may behave more aggressively than those retaining more differentiated programs. The heterogeneity within tumor types may reflect the hierarchical structure of developmental lineages, with different tumor subtypes corresponding to transformation at different developmental stages. This dissertation applies computational approaches to systematically identify and transfer developmental programs across datasets, enabling mapping of which developmental states persist in different tumor types.

2.3.4. Data Access and Acknowledgement

This dissertation relies on data derived from human embryonic and fetal tissue from elective abortions across multiple studies. Chapter 4 depends critically on datasets generated from human developmental tissue and pediatric tumors. The data sources include human cerebellar development single-cell atlases derived from tissue obtained from elective pregnancy terminations and surgical resections across different stages of prenatal development. The use of this tissue is governed by institutional review boards, informed consent procedures, and data protection policies. The original data generators obtained appropriate ethical approvals, and publicly shared datasets are de-identified to protect patient privacy. This dissertation analyzed only previously published, de-identified datasets and did not involve direct tissue collection or patient interaction. This scientific dissertation takes no direct stance on political, religious, or ethical perspectives on the concept of personhood, however, I acknowledge the ethical boundaries for accessing these types of tissues and the political and cultural frameworks that affect this type of research globally. Further exploration of this topic is explored in the Discussion (Chapter 6).

2.4. The Framework and Structure of this dissertation

The framework developed in this dissertation adapts topic modeling to identify transferable developmental genetic and isoform signatures, transfer and score them across datasets and disease contexts, and addresses the fundamental challenges outlined above. It handles continuous developmental transitions through probabilistic topic mixtures rather than discrete clusters. It enables cross-dataset comparison through gene set scoring rather than requiring integration. It identifies conserved programs beneath annotation inconsistencies. Additionally, it provides a systematic method for mapping developmental atlases to disease states. In the subsequent chapters, we will apply these ideas to the three domains mentioned in chapter 1.



3. CHAPTER 3. Single Cell Topic Modeling Identifies Transitional Cell Types in Mouse Fetal Development

Single-cell RNA sequencing has transformed our understanding of cellular heterogeneity during development yet conventional clustering approaches face significant challenges when applied to complex datasets especially with rare and transient cell populations^{43,44}. Standard methods such as Louvain or Leiden¹³ clustering assign each cell to a single discrete cluster, which, although provides powerful information about genetic similarities between cells, also fundamentally conflicts with the biological reality that developing cells often exist in transitional states expressing multiple transcriptional signatures simultaneously.

Topic modeling, specifically Latent Dirichlet Allocation (LDA)⁴⁵, offers a probabilistic alternative that addresses these limitations. Originally developed for natural language processing to identify latent themes in document collections, LDA can be adapted to transcriptomics by treating cells as “documents”, genes as “words” and gene expression values as “word frequencies”. This framework generates probabilistic distributions over genes otherwise known as topics that can be combined in varying proportions within individual cells. This approach has the advantage of characterizing cells by their mixture of multiple transcriptional signatures (topics) instead of forced into discrete categories. This in turn more accurately reflects the continuous nature of differentiation and developmental processes.

The advantages of topic modeling are especially beneficial for biological contexts where nuances in cell trajectories and cell type differences are challenging to capture with clustering alone. In this case, we focus on prenatal brain development where there is constant transience of cell types across differentiation states and ephemerality with cellular structures and anatomical features. Before applying topic modeling to human developmental datasets where more uncertainty awaits, we validate the model structure on a more well characterized system. We use the mouse model as the availability of comprehensive single cell atlases spanning embryonic development, established knowledge of developmental trajectories, and the ability to compare topic results against decades of developmental biology research is plentiful^{2,4–8,46–52}. If topic modeling can successfully recapitulate known developmental programs and cell type identities in the mouse, then this provides additional confidence for its application to human datasets where ground truth is less established.

This chapter presents preliminary analysis of topic modeling applied to the Qiu et al 2022 mouse organogenesis cell atlas^{2,5}. This is a comprehensive single cell dataset spanning embryonic days 8.5 (E8.5) through postnatal day 0 (P0). The objectives of these analyses are as follows: (1) determine whether LDA can identify biologically meaningful transcriptional signatures in a whole-embryo developmental context, (2) assess whether topics correspond to known cell types and developmental processes, (3) establish pipeline parameters that can subsequently be applied to human developmental studies.

3.1. Background in Mouse Organogenesis and Developmental Cell Types

Mouse embryonic development from gastrulation through birth represents a thoroughly characterized prenatal biological processes due to availability of tissue and data. Decades of classical embryology, genetic studies, and more recently, single-cell atlases⁴ have established detailed understanding of major developmental lineages (mesoderm, endoderm, ectoderm derivatives), timing and molecular control of cell fate decisions, marker genes defining progenitor and differentiated populations, spatial organization and migration patterns. Within mouse embryogenesis, there are several known developmental trajectories that are also represented in the atlas we will explore in this chapter. These trajectories include hematopoietic, myogenic, skeletal/mesenchymal, neural, epithelial, and vascular. Here, we will focus on neural lineages where the nervous system develops from neural plate through neural tube into central nervous system (brain and spinal cord) and neural crest-derived peripheral nervous system.

3.2. Background in Current Single Cell Analysis Approaches

Understanding standard single-cell analysis workflows is essential for contextualizing how topic modeling differs from and complements conventional approaches. Before any analysis, including topic modeling, single-cell RNA sequencing data must undergo several preprocessing steps that fundamentally shape downstream results. These steps address technical artifacts and reduce noise while preserving biological signal, making them critical prerequisites regardless of the analytical framework applied.

3.2.1. Quality Control and Preprocessing

Quality control generally is the first step of processing single cell data as it contains multiple sources of technical artifacts that are suggested to be removed before analysis. Standard quality metrics include the number of detected genes per cell, the total number of unique molecular identifiers (UMIs) per cell, and the percentage of reads mapping to mitochondrial genes. These metrics identify problematic cells: cells with very few detected genes likely represent empty droplets or lysed cells, while cells with unusually high gene counts may represent doublets where two cells were captured in a single droplet. Elevated mitochondrial gene expression often indicates cellular stress or damage, as cytoplasmic mRNA is lost while mitochondrial transcripts are retained in the mitochondrial double membrane. Thresholds for these metrics are typically dataset-specific and determined through examination of quality metric distributions. Adaptive thresholding approaches that account for dataset-specific variation have been developed to reduce subjectivity in filtering decisions. Additionally, computational methods such as⁵³ and Scrublet⁵⁴ can identify likely doublets based on simulated artificial doublets from the data. Gene filtering typically removes genes detected in very few cells, as these provide little information

for downstream analysis and may increase noise. The specific approach to quality control affects which cells and genes enter the analysis, directly impacting the ability to detect rare cell types or transitional states that topic modeling aims to capture.

3.2.2. Normalization Methods

Normalization in transcriptomics aims to remove technical variation while preserving biological signals and enabling meaningful comparison of gene expression across cells. Several normalization strategies have been developed for single-cell data, each with distinct assumptions and tradeoffs.

Library size normalization is one of the more common and simpler approaches which divides each cell's counts by its total count (library size) and scales to a common factor such as counts per million (CPM) or transcripts per million (TPM)⁵⁵. This assumes that total RNA content is similar across cells and that the majority of genes are not differentially expressed. While computationally simple, this approach can be biased when cells differ substantially in total RNA content or when a small number of highly expressed genes dominate the library.

Log-normalization⁵⁶ follows library size normalization, a log transformation is applied to reduce the influence of highly expressed genes and make the data more normally distributed. This is the default approach in tools like Seurat and is effective for many downstream analyses. The log transformation compresses the dynamic range of expression values, reducing the impact of outliers and technical noise in highly expressed genes.

Slightly more complicated and sophisticated approaches address specific limitations from the normalizations above. This includes Scraper pooling-based normalization⁵⁷ which addresses the challenge that low-coverage cells have imprecise size factor estimates. Cells are pooled together, pool-based size factors are computed, and then deconvolved to obtain cell-specific size factors. This approach is more robust for heterogeneous datasets with varying cell sizes and RNA content. Another tool is *SCTransform*⁵⁸ which was developed for Seurat v3 and beyond, the *SCTransform* function uses regularized negative binomial regression to model the relationship between gene expression and sequencing depth. This approach directly models count data, accounts for the mean-variance relationship in UMI counts, and identifies variable genes in a principled manner. *SCTransform* has shown improved performance for identifying rare cell populations and reducing the influence of sequencing depth on downstream analysis.

The choice of normalization method can substantially impact downstream results. Most single cell analysis that follow the Seurat tutorial will use *SCTransform* for ease of use and ample documentation, however, it is imperative that the type of normalizations used are chosen with consideration of the data characteristics and type. In this work, different types of normalization methods are used depending on the type of data and the dataset which will be mentioned in the methods in each chapter.

3.2.3. Feature Selection

Single-cell datasets contain expression measurements for tens of thousands of genes, but not all genes are informative for distinguishing cell types or states. Feature selection identifies highly variable genes that show greater variation than expected from technical noise, focusing downstream analysis on biologically meaningful signal. The following methods are the most commonly used ones for feature selection. Variance-based selection identifies genes with highest variance across cells. However, mean expression and variance are correlated in count data, so genes are typically binned by mean expression and the most variable genes within each bin are selected, or variance is modeled as a function of mean expression. Methods such as Seurat's *FindVariableFeatures*⁵⁹ function fit a loess curve to the mean-variance relationship and identify genes with variance exceeding the expected value. *SCTransform*⁶⁰ incorporates variable gene identification into the normalization procedure by identifying genes with high residual variance after accounting for sequencing depth. Some methods^{61,62} identify variable genes based on the relationship between mean expression and dropout rate (fraction of zero counts). Genes with more zeros than expected given their mean expression may represent genes with bimodal expression patterns across cell populations.

Typically, 2,000-5,000 highly variable genes are selected for downstream analysis. This number represents a tradeoff between retaining biological signals and reducing noise. Including too few genes may miss important biological variation, while including too many may introduce noise that obscures cell type differences. However, these values are modified based on the biological question at hand and the data attributes.

3.2.4. Dimensionality Reduction

Understanding conventional dimensionality reduction and clustering approaches provides essential context for how topic modeling differs in its treatment of cellular states and transitions. Dimensionality reduction transforms the high-dimensional gene expression space into a lower-dimensional representation that captures the major axes of variation while reducing noise. These reduced representations are used for visualization, clustering, and trajectory analysis.

Principal Component Analysis (PCA)¹⁸ identifies linear combinations of genes (principal components) that capture the maximum variance in the data. The first principal component captures the most variance, the second captures the most remaining variance orthogonal to the first, and so on. PCA is typically applied to log-normalized, scaled expression values for highly variable genes. The resulting principal components provide a reduced representation that serves as input for clustering and visualization methods. The

number of principal components to retain is a critical parameter. Too few components may lose biological signal, while too many may retain noise. Methods for selecting the number of components include examining the "elbow" in a scree plot of variance explained, jackstraw analysis to identify statistically significant components, or simply using a standard number (i.e. 30-50 PCs) that works well empirically.

Uniform Manifold Approximation and Projection (UMAP)⁶³ is a nonlinear dimensionality reduction method that aims to preserve both local and global structure in the data. It constructs a high-dimensional graph representation of the data based on nearest neighbor distances, then optimizes a low-dimensional embedding that preserves this graph structure. UMAP is the most widely used visualization method for single-cell data due to its ability to reveal cluster structure and continuous trajectories while being computationally efficient. UMAP parameters include the number of neighbors (controlling the balance between local and global structure), minimum distance (controlling how tightly points are packed in the embedding), and the distance metric. UMAP is typically applied to PCA-reduced data rather than the original expression matrix.

t-distributed Stochastic Neighbor Embedding (t-SNE)⁶⁴ is another nonlinear dimensionality reduction method that preserves local structure. t-SNE models pairwise similarities as probability distributions and minimizes the Kullback-Leibler divergence between high-dimensional and low-dimensional representations. While t-SNE produces visually appealing plots with well-separated clusters, it distorts global structure and distances between clusters are not meaningful. t-SNE is also more computationally intensive than UMAP, limiting its application to very large datasets.

3.2.5. Clustering Methods

Clustering methods partition cells into groups with similar expression profiles, representing the dominant paradigm in single-cell analysis that topic modeling offers an alternative to. Graph-based clustering (Louvain⁶⁵ and Leiden¹³) are the most widely used approaches. These methods construct a k-nearest neighbor (KNN) graph in PCA space, where cells are nodes and edges connect cells to their nearest neighbors. The Louvain or Leiden algorithm then identifies communities in this graph by optimizing a modularity function that measures the density of connections within communities versus between communities. The Louvain algorithm iteratively moves nodes between communities to maximize modularity, then aggregates communities and repeats. The Leiden algorithm improves upon Louvain by guaranteeing that communities are well-connected and avoiding the issue of poorly connected communities that can arise with Louvain. An important hyperparameter is resolution, which controls the granularity of clustering. Higher resolution produces more clusters, while lower resolution produces fewer, larger clusters. There is no single "correct" resolution, and different biological questions may require different granularities. Clustering at multiple resolutions and examining marker genes at each level can help identify biologically meaningful groupings.

Alternative clustering approaches exist but are seen less frequently. This includes traditional hierarchical clustering builds a tree (dendrogram) of nested clusters by iteratively merging the most similar clusters (agglomerative) or splitting the most heterogeneous clusters (divisive). While computationally expensive for large datasets, hierarchical clustering provides a natural representation of cell type relationships and can be useful for defining cell type hierarchies. Another method is K-means which partitions cells into k clusters by minimizing within-cluster variance. The algorithm iteratively assigns cells to the nearest centroid and updates centroids until convergence. K-means is fast but requires specifying the number of clusters a priori and assumes spherical clusters of similar size, which may not match biological reality.

Density-based clustering (DBSCAN^{66,67}, HDBSCAN⁶⁸) identifies clusters as regions of high density separated by regions of low density. Unlike k-means, density-based methods do not require specifying the number of clusters and can identify clusters of arbitrary shape. HDBSCAN (Hierarchical DBSCAN) extends DBSCAN to handle varying density and produces a hierarchy of clusters. Model-based clustering methods such as Gaussian mixture models assume that data arise from a mixture of probability distributions and use expectation-maximization to infer cluster assignments and distribution parameters. These methods provide probabilistic cluster assignments and can model clusters of different shapes and sizes.

Understanding these standard approaches is essential because they establish the baseline against which topic modeling's advantages become clear. All of these clustering methods share a fundamental limitation: they assign each cell to a single discrete cluster, assuming cells belong to distinct, separable groups. This assumption conflicts with developmental biology, where cells exist in transitional states expressing multiple programs simultaneously. Topic modeling addresses this limitation by representing each cell as a probabilistic mixture of topics, naturally capturing the continuous nature of developmental transitions that discrete clustering obscures.

3.3. Methods

3.3.1. Dataset (Mouse Organogenesis Cell Atlas)

We analyzed the mouse organogenesis cell atlas from Qiu et al. (2024), which provides a comprehensive single-cell time-lapse of mouse prenatal development from gastrulation to birth (E9.5 to P0). This dataset employs sci-RNA-seq3, a combinatorial indexing approach enabling profiling of millions of nuclei (>12 million nuclei) without requiring physical cell isolation. This dataset's comprehensive coverage across developmental time, large cell numbers, and whole-organism scope enable validation of whether

topics can recover known major developmental lineages, distinguish different states of cell type lineages from progenitors to fully differentiated, identify shared biological programs or signatures (i.e. cell cycle) across lineages, and capture tissue specific signatures. The published dataset includes quality-filtered nuclei that underwent library size normalization, with cell type annotations from the original authors based on Leiden clustering of the top 15 principal components and marker gene expression. For this analysis, we obtained the published quality-filtered dataset from GEO accession GSE228590 and randomly downsampled to approximately 100,000 cells for computational efficiency while maintaining representation across all major cell populations. Cell type annotations were taken directly from the original publication. We used these annotations for validation but note that topic modeling was performed in an unsupervised manner without using annotation information. This independence is important because if topic modeling recovers the same biological structure identified by clustering and expert annotation, but does so without supervision, this demonstrates that topics capture genuine biological signal rather than simply recapitulating provided labels.

3.3.2. Topic Modeling Implementation

LDA was implemented using the collapsed Gibbs sampling algorithm⁶⁹ in the R *lda* package (version 1.5.2)⁷⁰. The sparse expression matrix was extracted from the RNA assay, multiplied by 10, and rounded to convert continuous expression values to discrete count data suitable for the Gibbs sampler. This preprocessing approach follows established protocols for adapting continuous single-cell expression data to the discrete count framework required by LDA. Models were trained with alpha (document-topic concentration) set to 50 and beta/eta (topic-word concentration) set to 0.1, with 100 iterations following a 300-iteration burn-in period. These hyperparameters were selected based on preliminary optimization and are consistent with recommendations for topic modeling of genomic data^{23,27,71,72}.

Optimal Topic Number Selection.

Determining the optimal number of topics (K) is critical for balancing model complexity with interpretability. Previous studies have demonstrated that perplexity-based selection can show variability across iterations, motivating the use of rate of perplexity change as a more stable criterion^{27,28,73}. Models were trained iteratively across a range of topic numbers from K=20 to K=150, incrementing by 5. For each model, perplexity was calculated on held-out data using the *text2vec* package²⁹. The optimal K was identified by examining both absolute perplexity values and the rate of perplexity change across K values (Figure 3.1). The elbow point, where additional topics provide diminishing returns in model fit, occurred at K=50. This value was further validated by biological assessment: 50 topics provided sufficient resolution to capture all major cell types and developmental programs without excessive redundancy between topics. Lower K values merged biologically distinct programs, while higher K values generated redundant topics with highly overlapping gene signatures.

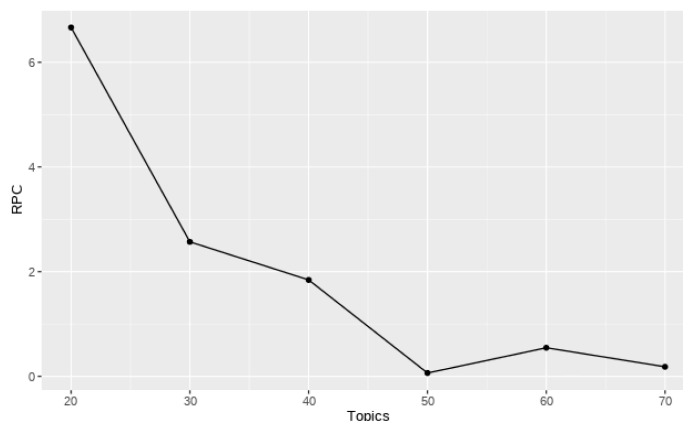


Figure 3.1 Elbow plot of the rate of perplexity change (RPC, y-axis) between iterations of topic modeling with changing numbers of topics (K, x-axis).

Latent Dirichlet Allocation (LDA) models were trained iteratively across a range of topic numbers (K = 20 to 100, incrementing by 5). For each model, perplexity was calculated on held-out data, and the rate of perplexity change (RPC) was computed as the absolute difference in perplexity between consecutive models divided by the change in K. The optimal number of topics is identified at the elbow point, where additional topics yield diminishing improvements in model fit. This statistical criterion was combined with biological assessment to ensure the selected K provided sufficient resolution to capture major cell types and developmental programs without generating redundant topics.

Topic Gene Extraction and Annotation

For each of the 50 topics, the top 50 genes ranked by probability score were extracted using the *lda::top.topic.words* function. Gene identifiers were converted from Ensembl IDs to gene symbols using the GSE228590 gene annotation file containing 49,585 genes. Topic annotation was performed through three complementary approaches: alignment with existing cell type labels to identify which

topics correspond to known populations, manual gene inspection comparing top-ranked genes against established cell type markers from CellMarker 2.0⁷⁴, PangaloDB⁴⁸, and examination of topic spatial distribution on UMAP embeddings. Gene ontology analysis was performed using annotations from the Gene Ontology Consortium⁷⁵ (2023) and KEGG pathway database⁷⁶ to validate biological interpretations.

Topic Grouping Methodology

Following model training and topic extraction, topics were organized into biologically coherent groups through a systematic multi-criteria framework designed to capture developmental relationships while minimizing arbitrary assignments.

Topic grouping employed complementary analytical approaches.

Spatial Co-localization. UMAP feature plots displaying topic weights across the single-cell embedding revealed spatial relationships between topics. Topics showing overlapping or adjacent expression domains on the UMAP were prioritized for co-grouping, as spatial proximity in the embedding typically reflects transcriptional similarity. Sequential expression patterns along apparent differentiation trajectories in the UMAP (ie. a topic expressed in progenitors transitioning to a topic expressed in mature cells within the same spatial region) informed hierarchical relationships within groups.

Cell Type and Trajectory Enrichment. Mean topic scores across annotated trajectories identified primary associations for each topic. Topics with peak expression (maximum mean score) in the same or developmentally related trajectories were grouped together. A specificity threshold was applied: topics with maximum expression scores >5 in a single trajectory were considered highly specific and prioritized for assignment to the corresponding lineage group. Topics showing elevated expression across multiple related trajectories (ie. mesoderm-derived lineages) were assigned based on the trajectory with highest enrichment and known developmental relationships.

Inter-Topic Correlation Analysis. Pairwise Pearson correlation coefficients were computed between all topic pairs using their cell-level activity scores across the full dataset. High positive correlations ($r > 0.5$) indicate that two topics tend to be co-expressed in the same cells, suggesting shared developmental contexts or complementary functions within a lineage. Topics with strong positive correlations were considered for co-grouping, while anti-correlated topics ($r < -0.3$) were assigned to separate groups reflecting mutually exclusive developmental programs.

Several other grouping methodologies were tested as well to determine topic groups. This included primary assignments via peak enrichment where each topic was initially assigned to the group corresponding to its highest-scoring trajectory. For example, a topic with peak expression in the hepatocyte trajectory was assigned to the endodermal organs group. Next was a refinement of the groupings based on relevant literature validations. Topics were reassigned when existing biological knowledge indicated a more appropriate grouping (i.e. kidney tubular epithelium was grouped with other epithelia despite its mesodermal association as it has a stronger shared epithelial program). Within groups as well, topics were ordered to reflect differentiation sequences where supported by (a) temporal expression patterns across developmental timepoints, (b) spatial gradients in the UMAP from progenitor to mature cell regions, and (c) known differentiation hierarchies from the literature. Finally, Topics with low maximum expression scores (<0.5 across all trajectories), diffuse expression patterns spanning multiple unrelated lineages, or ambiguous biological signatures were classified as ungrouped pending additional characterization.

3.4. Results

The 50 topics were organized into 11 lineage-specific groups plus one universal modifier signature. This organization reflects the major differentiation trajectories of mouse embryonic development, from pluripotent progenitors through germ layer specification to terminal cell type identities.

3.4.1. Topics Group based on Biological Processes and Major Cell Type Trajectories

To characterize the transcriptional programs underlying mouse organogenesis, we computed mean topic association scores for each of the 50 topics across 26 major developmental trajectories (Figure 3.2, Figure 3.3) as defined by the original study. Hierarchical clustering of trajectories revealed biologically coherent groupings that recapitulate known developmental relationships (Figure 3.2) hematopoietic lineages (B cells, T cells, erythroid, myeloid) clustered together, as did neural lineages (CNS neurons, intermediate progenitors, oligodendrocytes, brain/spinal cord), and mesodermal derivatives (mesoderm, muscle, cardiomyocytes). There were two major classes of topics (Figure 3.2) lineage-specific topics showing strong association with one or few trajectories, and shared topics exhibiting broader expression across related cell types. Highly lineage-specific topics included Topic 22 (hepatocytes, score = 6.69), Topic 43 (endothelium, score = 6.06), Topic 23 (intestinal epithelium, score = 6.44), Topic 49 (B and T lymphocytes, scores = 6.43 and 6.41), and Topic 30 (adipocytes, score = 5.83). In contrast, several topics were shared between developmentally related trajectories, reflecting common progenitor states or conserved differentiation programs. For example, Topic 42 (maturing erythroid) showed high scores in both primitive erythroid (5.79) and definitive erythroid (4.92) lineages, while Topic 27 (mature muscle) was shared between cardiomyocytes (4.42) and skeletal muscle cells (3.88). Based on the primary trajectory associations and clustering patterns, we organized the 50 topics into 12 developmental groups representing major branches of organogenesis (Figure 3.3, Table

3.1). We determined topic group names based on prominent marker genes across all of the included topics within that group as well ranking the major primary trajectories (as determined by the original study from clustering) for the cell types highly associated with the topics (Figure 3.3, Table 3.1)

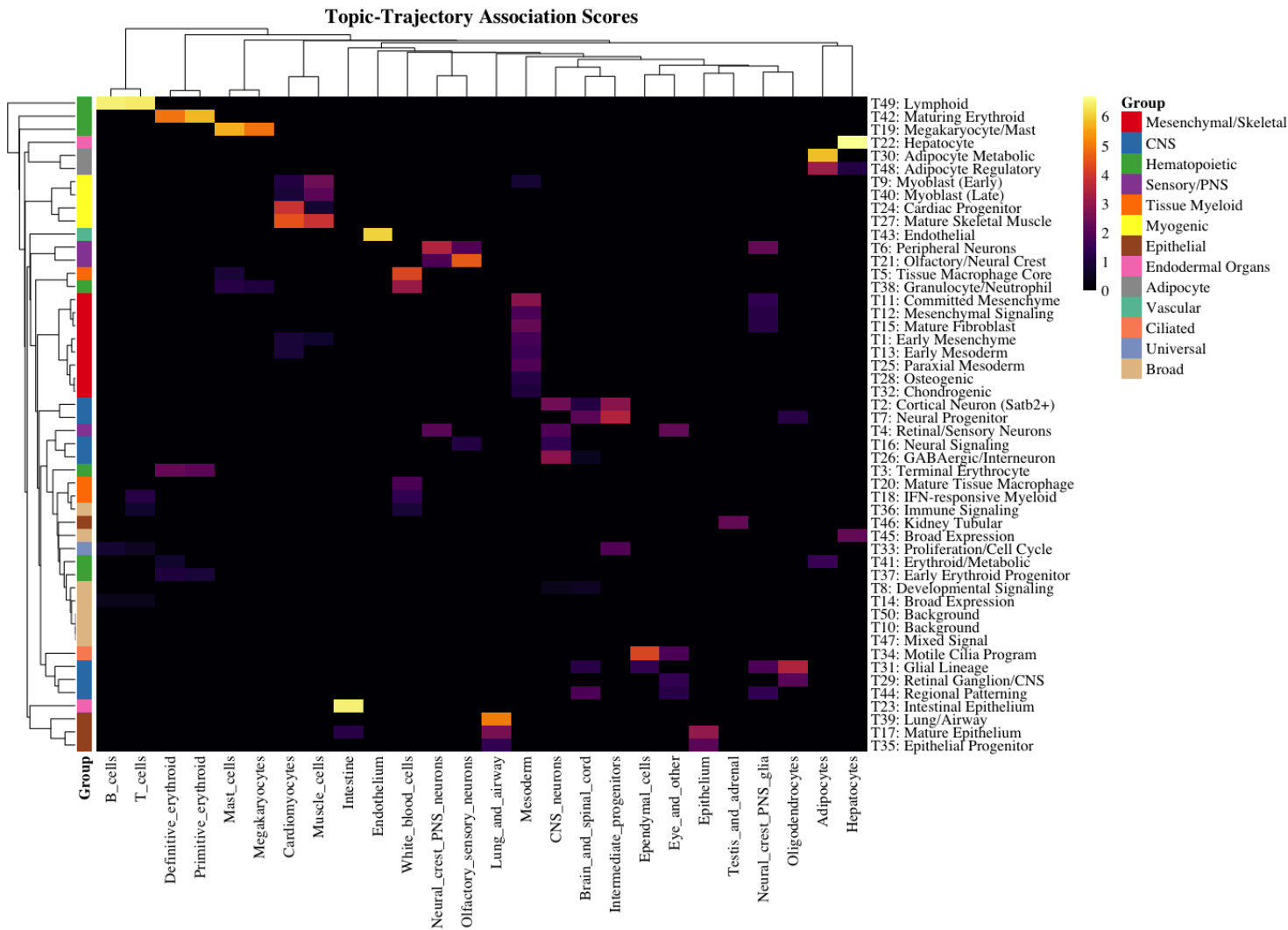


Figure 3.2. Topic-trajectory mean association scores reveal lineage-specific transcriptional programs during mouse organogenesis.

Heatmap displaying mean topic scores for each of the 50 transcriptional topics (rows) across 26 major developmental trajectories (columns). Color intensity represents the mean topic weight for cells within each trajectory, with darker blue indicating stronger association. Row annotations indicate the developmental group assignment for each topic based on primary trajectory associations. Hierarchical clustering (Euclidean distance, Ward's method) was applied to both rows and columns to group topics and trajectories with similar expression patterns.

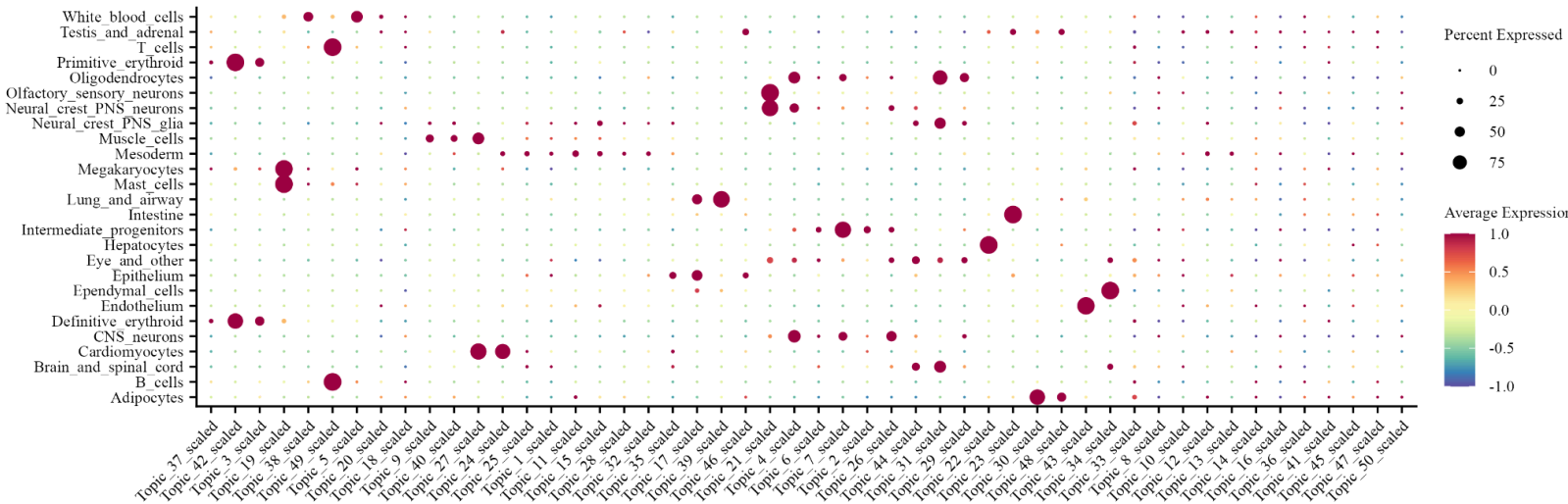


Figure. 3.3. Dot plot of topic expression across the 26 major developmental trajectories. Dot size indicates percent of cells expressing each topic and color indicates mean expression level. Topics are organized by assigned groupings (columns), and major cellular trajectories (predetermined from the original publication) are the rows.

Table 3.1. Mouse organogenesis cell atlas derived topic groupings based on hierarchical clustering and gene composition similarities.

The topic groups were given a group name based on defining marker genes and the primary trajectories involved as determined by the cell types with the highest topic-cell associations.

Group	Group Name	Topics	Primary Trajectories	Defining Marker Genes
1	Hematopoietic	37, 42, 3, 19, 38, 49	Definitive/primitive erythroid, megakaryocytes, mast cells, B cells, T cells	<i>Hemgn, Sptal, S100a8/a9, Ptprc</i>
2	Tissue Myeloid	5, 20, 18	White blood cells	<i>Dock10, Mertk, Stat1</i>
3	Myogenic	9, 40, 27, 24	Muscle cells, cardiomyocytes	<i>Nes, Des, Ttn, Ryr2</i>
4	Mesenchymal/Skeletal	25, 1, 11, 15, 28, 32	Mesoderm	<i>Prrx1, Col3a1, Runx2, Col9a1</i>
5	Epithelial	35, 17, 39, 46	Epithelium, lung/airway	<i>Igfbp2, Esrp1, Sftpc, Pkhd1</i>
6	Sensory/PNS	21, 4, 6	Olfactory sensory neurons, neural crest PNS	<i>Scn9a, Rims1, Map2</i>
7	CNS	7, 2, 26, 44, 31, 29	CNS neurons, brain/spinal cord, oligodendrocytes	<i>Dab1, Satb2, Erbb4, Slc1a2</i>
8	Endodermal Organs	22, 23	Hepatocytes, intestine	<i>Cps1, Alb, Cubn, Hnf4g</i>
9	Adipocyte	30, 48	Adipocytes	<i>Acadl, Glul, Foxo1</i>
10	Vascular	43	Endothelium	<i>Pecam1, Kdr, Tek</i>
11	Ciliated	34	Ependymal cells, eye/other	<i>Dnah7b, Ccdc39, Ccdc148</i>
12	Universal (Cell Cycle)	33	Pan-lineage (progenitors)	<i>Mki67, Top2a, Aspm</i>
13	Ungrouped	8, 10, 12, 13, 14, 16, 36, 41, 45, 47, 50	Various/unclear	---

Validation of Groupings Through Correlation Structure

Pairwise Pearson correlations between topics (Figure 3.4) within assigned groups confirmed the biological coherence of group assignments. Topics within the hematopoietic group showed positive inter-correlations (mean $r = 0.42$ among erythroid topics 37, 42, and 3), reflecting their shared developmental context within the hematopoietic hierarchy. The CNS group topics similarly showed positive correlations (mean $r = 0.38$), consistent with the overlapping nature of neural progenitor, neuronal, and glial programs during brain development.

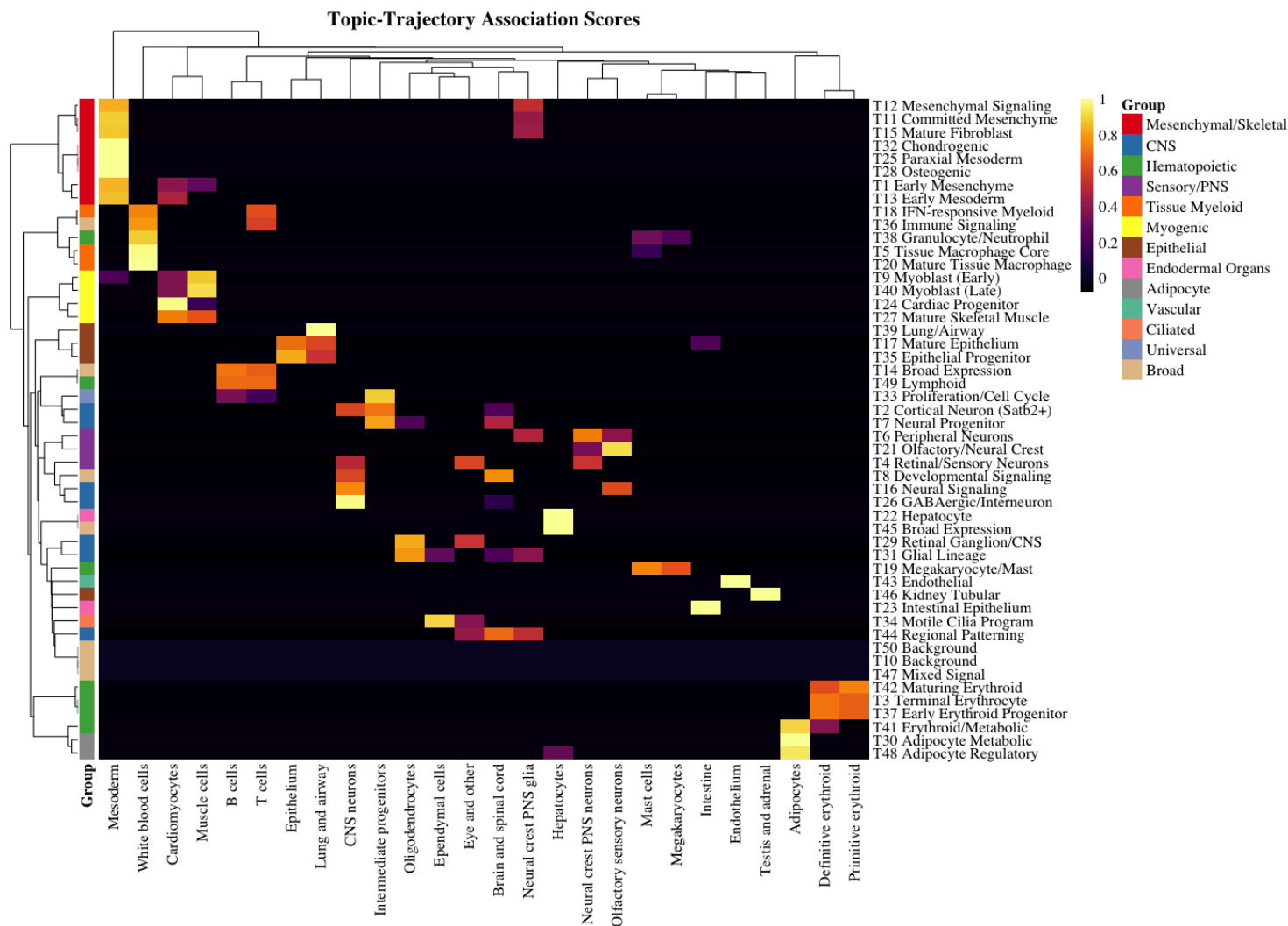


Figure 3.4. Topic-trajectory Pearson correlation heatmap reveals lineage-specific transcriptional programs.

Heatmap displaying Pearson correlation coefficients between 50 transcriptional topics (rows) and 26 major developmental trajectories (columns). Color intensity reflects correlation strength, with yellow indicating high positive correlation ($r \approx 1$, highly specific association) and dark purple indicating no correlation ($r \approx 0$). Hierarchical clustering (Euclidean distance, Ward's method) was applied to both rows and columns, revealing coherent developmental modules. Row annotations indicate topic group assignments.

Conversely, topics from developmentally distinct groups showed low or negative correlations. Hematopoietic topics were anti-correlated with mesenchymal topics (mean $r = -0.28$), reflecting the mutual exclusivity of blood versus connective tissue fates. The universal cell cycle topic (Topic 33) showed positive correlations with progenitor-associated topics across all lineages and negative correlations with mature cell type topics, validating its classification as a proliferation signature independent of lineage identity.

To assess the independence of learned transcriptional programs, we computed pairwise correlations between all 50 topics (Figure 3.5). The correlation matrix revealed predominantly low correlations ($|r| < 0.3$) between most topic pairs, demonstrating that LDA successfully identified independent transcriptional programs. Some topics within the same developmental group showed moderate positive correlations ($r = 0.3-0.5$), consistent with sequential or complementary programs within a lineage. However, the overall independence of topics, even within the same group, reflects the modular nature of developmental programs where distinct transcriptional states can co-exist within a lineage without strong co-expression.

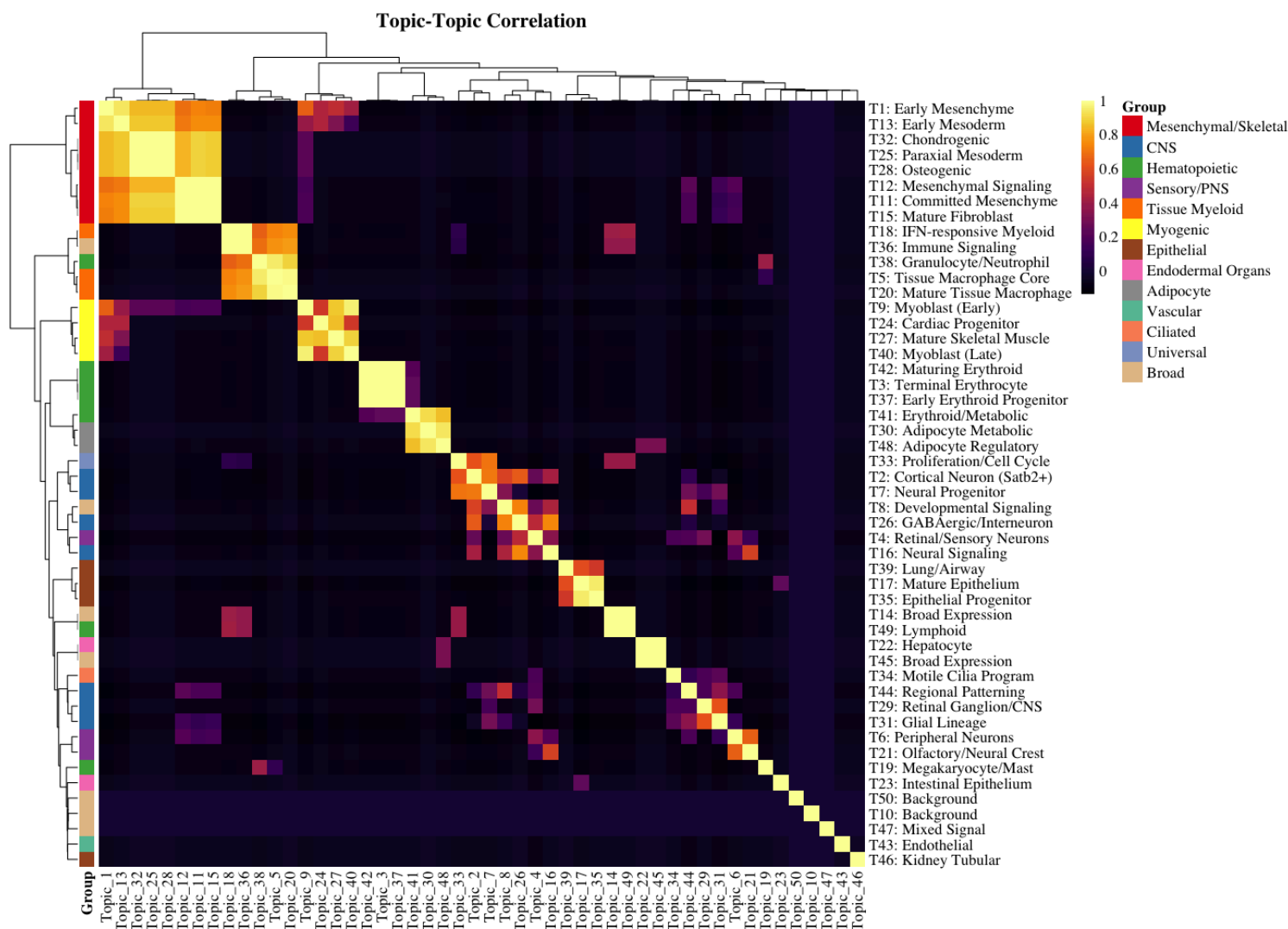


Figure 3.5. Topic-topic correlation matrix.

Pearson correlations computed between all 50 topics using topic weights across cells. Topics ordered by developmental group (left sidebar).

3.4.2. Central Nervous System related Topic Group identifies Progenitor to Neuron and Glia related Cell Types

The resulting 50 topics were organized into 11 lineage specific groups, one universal modifier group, and a set of ungrouped topics (13 total groups) (Table 3.1). This organization reflects different differentiation trajectories of mouse embryonic development that includes pluripotent progenitors through germ layer specification to terminal cell type identities. Here, we do a deep dive into one group specifically that represents CNS related cell types and activity. This group (titled CNS topic group) encompasses six topics (7,2,26,44,31, and 29) which collectively capture parts of the developmental spectrum from neural progenitors through neuronal and glial differentiation with distinctions that enhance and go beyond the original clustering approaches.

The central nervous system (brain and spinal cord) derives from the neural tube, which forms by folding of the neural plate ectoderm. Neural tube neuroepithelial cells transform into radial glia which are the primary neural stem cells that generate both neurons (neurogenesis) and glia (gliogenesis). Radial glia divide asymmetrically where one daughter remains a radial glial cell while the other becomes either a neuron directly or an intermediate progenitor that amplifies neuronal output. This process is precisely orchestrated in time and space to produce the diverse neuronal subtypes and layered architecture of the brain³⁷.

Branching Structure

Neural Progenitors (Topic 7) Topic 7 marks radial glia and neural progenitors expressing *Xkr4*, *Adgrb3*, and *Dab1* (disabled-1, an adaptor protein in the Reelin signaling pathway)^{37,77}. Reelin signaling is essential for neuronal migration and cortical lamination³⁷.

Excitatory Neurons (*Topic 2*) Glutamatergic (excitatory) projection neurons constitute the majority of cortical neurons and are generated locally within the dorsal telencephalon. Topic 2 represents maturing cortical neurons with high expression of *Satb2* (special AT-rich binding protein 2), a chromatin remodeling factor that specifies upper-layer cortical identity and callosal projection neuron fate³⁷. SATB2 functions by repressing *Ctip2*, thereby preventing subcortical projection identity and promoting interhemispheric connections through the corpus callosum³⁷.

Inhibitory Neurons (*Topic 26*) GABAergic interneurons provide inhibitory control over cortical circuits and originate from ventral telencephalon (ganglionic eminences), migrating tangentially into the cortex³⁷. Topic 26 captures interneuron identity through expression of *ErbB4* (the neuregulin receptor NRG1-ERBB4 signaling regulates interneuron migration and survival); *Astn1* (astrotactin 1, mediating neuron-glia interactions during migration); and *Nkain3*³⁷.

Regional Identity and Glia (*Topics 44, 31*) Topic 44 represents regionally patterned neuronal populations expressing *Meis2*, *Nnat* (neuronatin), and *Pbx3* (transcription factors establishing positional identity along the neuraxis). Topic 31 captures glial lineages with expression of *Slc1a2* (EAAT2/GLT-1, an astrocytic glutamate transporter); *Ptn* (pleiotrophin); and *Ntrk2* (✨TrkB✨, the BDNF receptor)³⁷.

Retinal Ganglion Cells (*Topic 29*) Although the retina is anatomically peripheral, it is developmentally CNS-derived (from the optic vesicle, an outpouching of the diencephalon). Retinal ganglion cells are the output neurons of the retina, whose axons form the optic nerve and transmit visual information to the brain. Topic 29 expresses *Kcnb2*, *Adgrb3*, and *Lrp1b*, bridging CNS and sensory system groups.

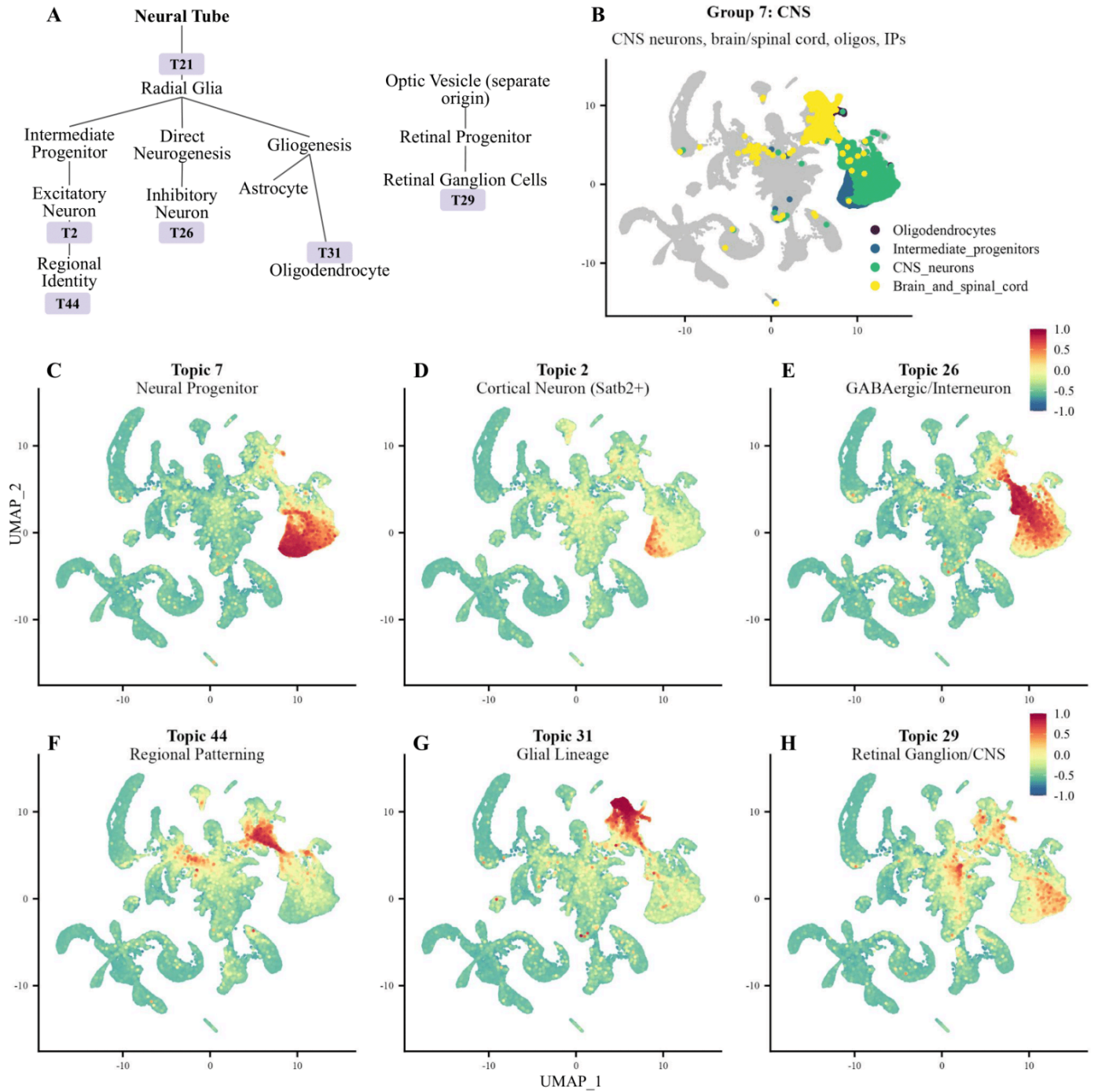


Figure 3.6. CNS topics capture neural tube development and neuronal/glial diversification.

(A) Schematic of neural tube differentiation from radial glia through intermediate progenitors to excitatory neurons (Topic 2), inhibitory neurons, astrocytes, and oligodendrocytes. Regional patterning and retinal ganglion cells are also represented. (B) UMAP visualization highlighting oligodendrocytes, intermediate progenitors, CNS neurons, and brain/spinal cord trajectories. (C-H) UMAPs displaying topic weights for Topics 7 (Neural Progenitor), 2 (Cortical Neuron/Satb2+), 26 (GABAergic/Interneuron), 44 (Regional Patterning), 31 (Glial Lineage), and 29 (Retinal Ganglion/CNS). This group represents the largest and most complex developmental hierarchy in the dataset.

3.4.3. Summary of Other Major Topic Groupings

While the CNS example demonstrates how topics subdivide known lineages with unprecedented resolution from an unsupervised approach, similar patterns emerged across all 12 major developmental topic groups, revealing general principles about developmental organization.

Hematopoiesis exemplifies hierarchical topic organization. Six topics (Group 1, Figure S1) captured the complete blood cell hierarchy, with Topic 37 marking hematopoietic stem cells, Topics 42 and 3 distinguishing primitive versus definitive erythropoiesis through temporal hemoglobin switching (embryonic Hbb-bh1 versus adult Hbb-b1), and Topics 38 and 49 separating myeloid and lymphoid maturation. This hierarchical structure mirrors the well-established hematopoietic developmental tree, demonstrating that topics recover known branching relationships without supervision.

Mesenchymal lineages revealed unexpected program sharing. Group 4 (Figure S3) had six topics spanned skeletal development, but rather than segregating strictly by tissue type, topics reflected developmental stages shared across tissues. Topic 25 captured early mesodermal commitment (*Meox1*, *Tbx6*) expressed across presumptive skeletal, connective, and smooth muscle precursors. Only later topics showed tissue restriction with Topic 32 for mature cartilage (*Col9a1*, *Sox9*) and Topic 28 for osteoblasts (*Bglap*, *Spp1*). This progression from shared progenitor programs to tissue-specific differentiation programs illustrates how development proceeds through common molecular states before lineage restriction, a pattern obscured when clustering assigns cells to tissue labels from the outset.

Myogenic topics captured parallel but distinct programs. Group 3 (Figure S3) had four topics revealed that cardiac and skeletal muscle, despite sharing myogenic transcription factors, activate these factors in different contexts. Topic 9 marked early mesodermal progenitors (*Meox2*, *Prrx1*) common to both lineages. Topic 40 captured skeletal muscle determination (*Myod1*, *Myf5*), Topic 27 skeletal muscle differentiation (Myog, *Tnnt3*), and Topic 24 cardiac specification (*Nkx2-5*, *Gata4*). The segregation of these programs despite shared fundamental myogenic machinery (MyoD family transcription factors) demonstrates that topics distinguish context-dependent deployment of the same transcriptional regulators.

Epithelial topics unified diverse organs through shared programs. Group 5 (Figure S5) had four topics marked lung (*Sftpc*, *Lamp3*), intestine (*Vill*, *Cdx2*), kidney tubule (*Pkhd1*, *Aqp2*), and epidermal (*Krt14*, *Tp63*) epithelium. While these tissues arise from different germ layers and serve distinct functions, topics captured their shared epithelial identity through E-cadherin, tight junction components, and apicobasal polarity genes, overlaid with organ-specific transcriptional programs. This demonstrates topics can simultaneously represent general biological properties (epithelial architecture) and context-specific functions.

Single-topic groups revealed highly specialized programs. Groups 8-12 (Figure S7-11) comprised one or two topics each, representing tightly coordinated programs where further subdivision provides no additional resolution. Topic 43 (vascular endothelium) captured a relatively coherent signature with *Pecam1*, *Kdr*, and *Tek* defined endothelial identity across all vessels regardless of organ context, reflecting the conserved nature of vascular specification. Similarly, Topic 33 (universal cell cycle) showed that proliferation represents a modular program deployable across all lineages, with *Mki67*, *Top2a*, and *Aspm* enriched in dividing cells from neural, hematopoietic, epithelial, and mesenchymal compartments.

Ungrouped topics (Figure S11, Table S1) had different diffuse patterns with possible associations with different major cellular trajectories though further analysis is required to confirm.

These patterns reveal general principles of developmental gene expression organization. Development proceeds through hierarchical programs where early topics mark broad fate decisions (mesoderm versus ectoderm) and later topics mark terminal fates (cardiomyocyte versus skeletal muscle). Some programs are modular and lineage-independent (cell cycle, stress response), while others are highly lineage-restricted. Most developmental transitions involve gradual shifts in topic proportions rather than abrupt switches, reflecting the continuous nature of differentiation that discrete clustering cannot capture.

Detailed characterization of all groups including top marker genes for each topic, cell type enrichment patterns, validation against literature, spatial expression patterns, temporal dynamics, and comparison to clustering results is provided in Supplementary Materials.

3.5. Discussion

This chapter demonstrates that topic modeling successfully identifies biologically meaningful transcriptional programs in mouse organogenesis and provides validation for applying this framework to less-characterized systems. The 50 topics organized into 12 developmental groups recapitulate known lineage relationships while revealing hierarchical organization within lineages that discrete clustering approaches obscure.

The central nervous system analysis exemplifies how topics capture developmental progressions. Rather than assigning cells to discrete bins labeled "neural progenitor" or "neuron," topics quantify the continuous spectrum from radial glia through neurogenesis to glial differentiation. Topic 2 marks proliferative radial glia, Topic 7 captures active neurogenesis, and Topic 26 identifies oligodendrocyte specification as defined by marker genes encompassed in the respective topics and the cell-topic association scores. Cells in transition express multiple topics simultaneously with intensities reflecting their position along the differentiation trajectory. This probabilistic representation more accurately reflects the biological reality of development than forced assignment to mutually exclusive clusters.

The grouping structure validates biological coherence. Topics within lineages show positive correlations through trajectory enrichment, while topics from distinct developmental fates show minimal overlap. The universal cell cycle topic (Topic 33)

demonstrates that LDA can simultaneously capture lineage-independent processes and lineage-specific differentiation programs with canonical cell cycle and proliferation genes. The topic was enriched across proliferative cells regardless of fate.

Several limitations warrant discussion. First, the choice of $K=50$ topics, while validated through perplexity analysis and biological assessment, represents one resolution among many possible decompositions. Different K values capture different granularities with lower K merges related programs while higher K subdivides programs into increasingly specific signatures. Second, interpretation requires biological expertise to map topics to known developmental processes. Unlike clustering methods that generate discrete, countable populations, topic modeling produces probability distributions requiring more careful analysis. Third, the downsampling to 100,000 cells, while maintaining representation of major lineages, may undersample rare cell populations and reduce power to detect transient intermediate states.

Despite these limitations, the successful recovery of known developmental programs in mouse organogenesis provides confidence for applying topic modeling to human datasets where ground truth is less established. The framework's ability to handle continuous transitions, identify rare and transient states, and generate transferable gene signatures addresses key limitations in current single-cell analysis approaches. **Chapter 4 applies these validated methods to human cerebellar development and demonstrates topic transfer to medulloblastoma, testing whether developmental signatures persist in disease. Chapter 5 extends the framework to transcript-level resolution, revealing that alternative isoforms co-express according to developmental context rather than protein function, a layer of regulation invisible to gene-level analysis.**

3.6. Conclusion

This chapter demonstrated that topic modeling successfully identifies biologically meaningful transcriptional programs in a well-characterized developmental dataset. The approach recovered all major developmental lineages, revealed hierarchical organization within lineages, and provided advantages over discrete clustering for capturing transitional states and shared programs. The detailed examination of CNS topics illustrated how topic modeling can dissect developmental trajectories with sensitivity beyond conventional approaches.

The validation achieved here justifies proceeding to less-characterized biological contexts in subsequent chapters. Having established that topic modeling works when tested against known biology, we can now apply it to human neurodevelopment and disease where ground truth is less complete but biological questions are equally if not more urgent.



4. CHAPTER 4: Transferable Topic Modeling Reveals Developmental Cellular Lineage Persistence in Medulloblastoma

The contents of this chapter are part of a manuscript currently under review--specifically the methods and results. The introduction and discussion have been modified for this dissertation.

Single-cell RNA-seq has revolutionized the study of cellular heterogeneity in brain development and disease^{3,78-81}. However, a critical challenge of this method that persists is inconsistent cell type annotations across datasets, which can undermine comparative analysis and translational applications. This gap continues to grow as datasets continuously expand across laboratories, technologies, and biological contexts but our ability to systematically compare and integrate findings has not kept pace. Though powerful statistical and machine learning approaches have been proposed, a recent perspectives of causal machine learning for single cell genomics highlighted several open problems in the field⁸². First, there is a lack of generalization of models to novel experimental conditions; second, there still exists the complexity of interpreting learned models (the "black box"); third, there still is a difficulty of learning cell dynamics instead of one consistent state. Additionally, with the expansion of data, batch correction and whole data merging methods, while valuable, require careful parameter tuning, can remove genuine biological signals, and become computationally prohibitive⁸²⁻⁸⁵.

The limitations listed above are consequential for understanding biological problems that require integration of different contexts. For instance, for pediatric brain cancers where origins are strongly tied to developmental cellular dynamics and hierarchies. This precise mapping of tumor-developmental relationships can have downstream effects on therapeutic strategies. The cerebellum is an ideal system for developing such methods as it is the known origin for several pediatric brain tumors and has defined developmental stages. The cerebellar rhombic lip (RL) orchestrates neurogenesis through precisely timed proliferation and differentiation waves. This process over half of all neurons in the adult human brain^{37,86}. Beginning around post-conception week (PCW) 8 in humans, specific progenitors migrate from the RL to form the external granule layer (EGL). The EGL is a bilayered germinal zone where granule neuron progenitors (GNPs) undergo massive Sonic hedgehog (SHH)-driven expansion in the outer EGL before transitioning to the inner EGL for differentiation and eventual migration to the internal granule layer^{37,87,88}. This developmental story, though meticulously coordinated, is transient. Its ephemeral existence proves challenging to capture with conventional clustering approaches. These methods are powerful for more prominent cell types but overlook small populations undergoing rapid state transitions that get buried in umbrella clusters. Dysregulation of these EGL developmental processes, particularly aberrant EGL proliferation, can directly contribute to post-natal pathogenesis, specifically medulloblastoma⁸⁹⁻⁹².

Medulloblastoma is the most common malignant pediatric brain tumor that arises from these cerebellar progenitor populations. This cancer has four major molecularly defined subgroups: WNT, SHH, Group 3, and Group 4⁸⁹. Seminal work has established that impaired differentiation of specific neural progenitors underlies pediatric brain cancers⁹³. By projecting bulk tumor transcriptomes onto developmental single cell atlases, Vladoiu et al. demonstrated that medulloblastoma subgroups match to specific developmental cellular lineages. Here, they revealed highly defined WNT medulloblastomas map to rhombic lip derived mossy fiber neurons. Similarly, from another pivotal integrative multi-omic work from Northcott et al⁹², Group 3 and Group 4 medulloblastoma tumors mapped to photoreceptor and unipolar brush cell expansions from the RL. Another complementary work from Hendrikse et al⁹⁴ employed unique approaches to map tumor cells of origin through bulk transcriptomics projected onto single cell developmental atlases, DNA methylation, and enhancer landscape analysis and further validated medulloblastoma tumor subgroup origins.

While these approaches established the developmental origins of medulloblastoma subgroups, there are several unaddressed challenges. First, projection-based methods rely on pre-defined developmental signatures which miss novel information or skew our understanding of existing cell types. Second, these require integration across disparate datasets with different sequencing technologies, processing pipelines, and annotation schemes. These integration and data harmonization methods invariably lose data and can oversimplify the existing biology. Finally, the developmental populations that are the most relevant to tumor origins are small, transient, and poorly represented in most of these atlases which prevents more robust and detailed mapping. These methodological constraints highlight the need for analytical frameworks that can extract transferable biological information.

We demonstrate here that topic modeling⁴⁵, a natural language processing approach, can treat cells as "documents" and genes as "words"²⁷ in order to identify conserved transcriptional programs across diverse datasets (analogy demonstrated in Figure S1 and workflow demonstrated in Figure S2). Topics from one dataset can be directly applied to others, enabling connections between developmental programs and cancer. While different studies might use varying labels for similar populations, topic modeling reveals shared or distinct transcriptional programs beneath annotation disagreements. Ultimately, we show here that this topic modeling framework identifies biologically meaningful gene expression patterns that persist across technologies, timepoints, and between healthy and malignant tissues.

As proof of principle, we analyzed midgestational fetal cerebellar landmark datasets. The Aldinger et al. dataset⁷⁸ (subset to 15,556 nuclei from post-conception week 17) serves as our source for establishing cerebellar-specific transcriptional programs, benefiting from expert annotation. Post-conception week 17 represents an ideal developmental timepoint for identifying EGL-associated transcriptional programs, as the EGL is well-established and actively expanding at this stage, with robust populations of both proliferating oEGL and differentiating iEGL granule neurons. The Cao et al. dataset³ (subset to 119,954 nuclei from weeks 12.5) serves as our validation set for testing topic conservation across different experimental approaches. Week 12.5 was selected as it contained the largest number of nuclei in the Cao dataset and represents an earlier developmental stage where we would expect

distinct distributions of developmental topics. This distribution would be focused specifically with greater enrichment of RL-associated programs and proportionally less EGL activity compared to the later week 17 timepoint. This temporal comparison allows us to capture the developmental transition from RL-dominant to EGL-dominant neurogenesis. For disease context, we use the Arora et al. dataset which encompasses medulloblastoma bulk RNA sequencing⁹⁵ across multiple subgroups, examining how developmental programs persist or alter in malignancy.

To further validate topic conservation, we applied our topics to a mouse developmental atlas from Qiu et al.⁵² and examined expression patterns of top topic genes using mouse in situ hybridization. These analyses confirmed EGL localization and validated the genetic signatures proposed here as capturing transitional cell states within the EGL. This also demonstrates the ability of our pipeline to combine and analyze data across three labs, different genomic data, and collection technologies.

Our analysis shows several advances. First, topic modeling identifies transitional states and rare cell types, particularly developing human EGL progressions which includes oEGL to iEGL transitions and migration of differentiating granule neurons. These specific transitions are missed by conventional clustering approaches. This captures the spectrum of GNP proliferation, cell cycle exit, and early differentiation within the EGL structure. Second, topic transfer using gene set variation analysis (GSVA) successfully identifies conserved cell types and developmental processes across independent developmental datasets from different labs and technologies. The transferability demonstrates robust detection of EGL associated genetic signatures across developmental timepoints. Third, applying cerebellar developmental topics to medulloblastoma bulk RNA sequencing confirms the GNP and EGL origins of SHH tumors and reveals age-specific molecular signatures within SHH subtypes. This disease application demonstrates the transferability of topics across data types and biological contexts. Notably, the EGL-derived topics allow us to distinguish which stage of granule neuron development, such as proliferative oEGL versus differentiating iEGL, is most represented in different SHH medulloblastoma subtypes. A major contribution to the field here is the generalizability of this pipeline, which enables systematic mapping of EGL developmental programs onto disease states. These findings improve our understanding of cerebellar neurodevelopment, particularly the molecular architecture of the EGL, and guide downstream characterization of tumor substructure that can inform and alter treatment strategies.

This chapter demonstrates that topic modeling overcomes barriers to cross-dataset comparison by learning transferable transcriptional signatures that generalize across datasets, technologies, and biological contexts. We apply this framework to over one million fetal cerebellar nuclei to identify topics capturing the developmental spectrum from rhombic lip progenitors through EGL differentiation, including transitional states missed by conventional clustering. We then demonstrate that these topics transfer successfully across sequencing technologies (SPLiT-seq to sci-RNA-seq3), developmental timepoints (PCW 17 to 12.5), species (human to mouse), and from single-cell to bulk RNA sequencing of medulloblastoma tumors.

The chapter is organized as follows. Section 4.2 reviews related work on developmental origins of medulloblastoma, cross-dataset analysis methods, and topic modeling applications in genomics. Section 4.3 describes the datasets, computational methods, and validation approaches. Section 4.4 presents results including topic discovery, cross-dataset validation, spatial validation, and application to medulloblastoma. Section 4.5 discusses the implications of these findings for understanding EGL development, SHH medulloblastoma biology, and the broader utility of topic modeling for comparative genomics.

4.1. Background in Cross Dataset Developmental and Pediatric Cancer Studies

Developmental Origins of Medulloblastoma

The concept that pediatric brain tumors arise from specific developmental progenitor populations has gained substantial support over the past decade. Seminal work established that impaired differentiation of specific neural progenitors underlies pediatric brain cancers⁹³. By projecting bulk tumor transcriptomes onto developmental single-cell atlases, Vladoiu and colleagues demonstrated that medulloblastoma subgroups match to specific developmental cellular lineages, revealing that WNT medulloblastomas map to rhombic lip-derived mossy fiber neurons⁹⁶.

Subsequent integrative multi-omic studies refined these cellular origin assignments. Northcott, Smith, and colleagues demonstrated that Group 3 and Group 4 medulloblastoma tumors mapped to photoreceptor and unipolar brush cell expansions from the rhombic lip⁹². Hendrikse and colleagues employed complementary approaches including bulk transcriptomics projected onto single-cell developmental atlases, DNA methylation, and enhancer landscape analysis to further validate medulloblastoma tumor subgroup origins, concluding that failure of human rhombic lip differentiation underlies medulloblastoma formation⁹⁴.

For SHH medulloblastoma specifically, the granule neuron precursor origin is well established. Mouse models in which *Ptch1* is conditionally deleted in GNPs faithfully recapitulate human SHH medulloblastoma⁴². However, the heterogeneity within SHH medulloblastoma with four subtypes (SHH α , SHH β , SHH γ , SHH δ) showing distinct age distributions and clinical outcomes suggests that the specific developmental stage or state of the cell of origin matters⁹⁷. The molecular basis for this subtype heterogeneity and its relationship to EGL developmental stages has not been fully characterized yet.

These prior approaches established developmental origins of medulloblastoma subgroups in pivotal studies. However, several challenges remained unaddressed. First, projection-based methods relied on pre-defined developmental signatures, potentially missing

novel information or skewing understanding of existing cell types. Second, these methods required integration across disparate datasets with different sequencing technologies, processing pipelines, and annotation schemes where the integration invariably loses data and can oversimplify existing biology. Third, the developmental populations most relevant to tumor origins are small, transient, and poorly represented in most atlases, preventing robust and detailed mapping.

Cross-Dataset Analysis Methods in Single-Cell Genomics

The challenge of comparing single-cell datasets across studies has motivated development of numerous computational methods. Batch correction approaches modify expression values to remove technical variation while preserving biological signal. Mutual nearest neighbors (MNN) identifies pairs of cells from different batches that represent the same cell type and computes correction vectors to align them¹⁵. Harmony iteratively adjusts PCA embeddings to remove batch effects using soft clustering to identify cell types¹⁶. Deep learning approaches including scVI use variational autoencoders to learn latent representations accounting for batch effects⁹⁸. These methods have proven valuable for integrating datasets with similar cell type compositions, but they have important limitations for the developmental-cancer comparison problem. First, they require datasets to share at least some cell populations to identify corresponding cells across batches which is a requirement not met when comparing fetal development to adult tumors. Second, batch effects may be confounded with biological differences of interest, making their removal problematic. Third, integration methods cannot bridge the gap between single-cell and bulk RNA sequencing, precluding analysis of clinical cohorts available only as bulk data.

Reference mapping approaches offer an alternative strategy. Methods such as Seurat's reference mapping and scArches project query data onto pre-computed reference atlases and transfer cell type labels based on proximity in the reference embedding^{85,99}. These approaches require that query cells have counterparts in the reference which also means that novel cell types or disease states not represented cannot be appropriately classified. Furthermore, reference mapping operates at the level of discrete cell type labels rather than continuous programs, potentially losing information about graded biological variation.

Signature scoring methods including Seurat's AddModuleScore compute enrichment of predefined gene sets in single-cell data. These approaches are conceptually related to the topic transfer framework developed here but typically rely on manually curated gene sets from literature or databases. Such predefined signatures may not capture the specific transcriptional programs present in the datasets under study, may be biased toward well-studied cell types and pathways, and may miss novel or dataset-specific programs. Here we

Topic Modeling Applications in Genomics

Topic modeling has been increasingly applied to genomic data analysis, as reviewed in Chapter 2. Early applications to bulk expression data focused on cancer genomics, demonstrating that topics learned from tumor expression data often corresponded to known biological processes including cell cycle, immune response, and metabolism. The application of topic modeling to single-cell transcriptomics was pioneered by several groups.

Celda (Cellular Latent Dirichlet Allocation) developed by Wang and colleagues simultaneously performs co-clustering of genes into modules and cells into subpopulations using single-cell RNA-seq data²³. fastTopics provides a computationally efficient implementation using variational inference that scales to millions of cells²⁴. The cisTopic framework applies topic modeling to single-cell ATAC-seq data, treating cells as documents and accessible regulatory regions as words to identify co-accessible regulatory programs²⁷. A landmark application of topic modeling to single-cell cancer data came from the Tirosh laboratory, which used non-negative matrix factorization (NMF), mathematically related to topic modeling, to identify recurrent transcriptional programs in metastatic melanoma²⁶. This study demonstrated that single-cell topic modeling could identify programs including resistance programs associated with therapeutic response, revealing intra-tumoral heterogeneity that bulk analysis could not resolve. Despite these advances, the systematic application of topic modeling to compare developmental and disease datasets across technologies has not been extensively explored. The present work addresses this gap by demonstrating that topics learned from developmental single-cell data can be effectively transferred to independent developmental datasets and to disease bulk RNA sequencing data using GSVA-based scoring.

Gene Set Variation Analysis for Topic Transfer

Gene Set Variation Analysis (GSVA) computes gene set enrichment scores for individual samples or cells without requiring explicit sample groupings or phenotype labels¹⁰⁰. Unlike Gene Set Enrichment Analysis (GSEA), which compares two conditions, GSVA transforms a gene expression matrix into a gene set enrichment matrix, enabling sample-level quantification of pathway activity.

The GSVA algorithm proceeds through several steps: gene expression values are transformed to ranks within each sample, a kernel estimation creates a cumulative density function for each gene, genes are ranked by their sample-specific statistics, and enrichment scores are calculated as the maximum deviation of gene set genes from the overall distribution. GSVA scores represent relative enrichment of each gene set in individual cells or samples, with positive scores indicating coordinate upregulation relative to genes outside the set.

The combination of topic modeling and GSVA provides a principled framework for cross-dataset analysis. Topics are learned from a reference dataset by identifying co-expressed gene programs. The top genes defining each topic are extracted as gene sets. These gene

sets are then scored in target datasets using GSVA, quantifying how strongly each sample or cell expresses each topic's transcriptional program.

4.2. Materials and Methods

Three primary datasets were analyzed in this study, each selected for specific biological and technical characteristics that enable rigorous validation of the topic modeling framework (Figure 4.1).

Source Dataset (Aldinger et al.)⁷⁸: The source dataset comprised 15,556 nuclei from human fetal cerebellum at post-conception week (PCW) 17, processed using SPLiT-seq technology (Aldinger et al., 2021). This dataset was selected as the source for topic generation for several reasons. First, PCW 17 represents an ideal developmental timepoint for identifying EGL-associated transcriptional programs, as the EGL is well-established and actively expanding at this stage, with robust populations of both proliferating oEGL and differentiating iEGL granule neurons. Second, the dataset benefits from expert annotation by developmental neurobiologists with deep expertise in cerebellar development, providing high-confidence cell type labels against which topic localization can be validated. Third, SPLiT-seq technology provides sufficient gene detection for robust topic modeling while representing a distinct technical platform from the validation dataset.

Applied Dataset 1 (Cao et al.)³: The first applied dataset comprised 119,954 nuclei from human fetal cerebellum at PCW 12.5, processed using sci-RNA-seq3 technology as part of a comprehensive human cell atlas of fetal gene expression (Cao et al., 2020). PCW 12.5 was selected because it contained the largest number of nuclei in the Cao dataset and represents an earlier developmental stage where distinct distributions of developmental topics would be expected, specifically, greater enrichment of rhombic lip-associated programs and proportionally less EGL activity compared to the later PCW 17 timepoint. The use of sci-RNA-seq3 technology provides validation that topics transfer across fundamentally different single-cell platforms. This temporal comparison allows capture of the developmental transition from rhombic lip-dominant to EGL-dominant neurogenesis.

Applied Dataset 2 (Arora et al.)⁹⁵: The disease application dataset comprised bulk RNA sequencing from 876 medulloblastoma tumors (274 labeled as SHH) from a comprehensive medulloblastoma reference landscape (Arora et al., 2025). This dataset enables examination of how developmental programs persist or alter in malignancy and tests whether topics transfer successfully from single-cell to bulk RNA sequencing data.

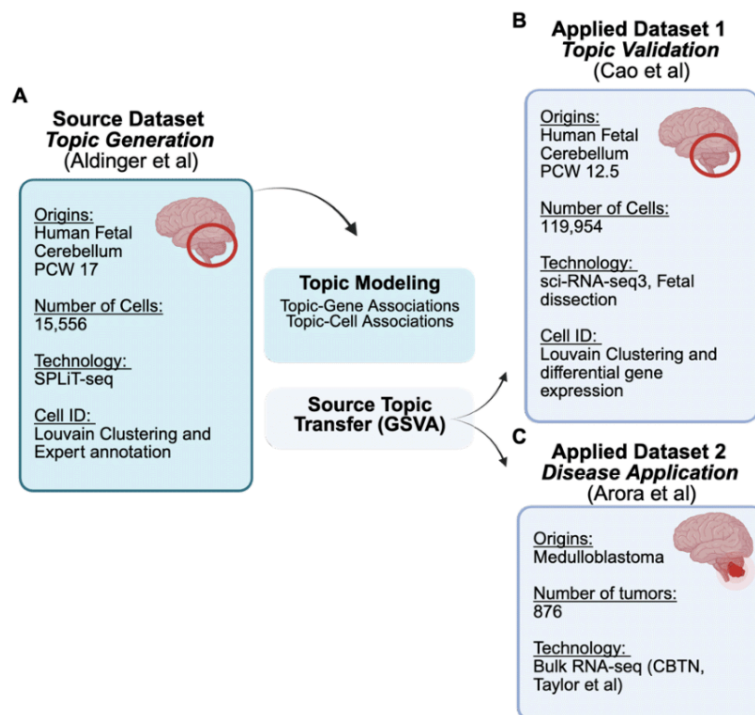


Figure 4.1. Computational workflow for topic modeling and transfer from Source dataset to two Applied datasets. (A) Source Dataset (Aldinger et al., subset of 15,556 nuclei from PCW 17 of human cerebellar development) (B) Applied Dataset 1 (Cao et al., subset of 119,954 cells from PCW 12.5 of human cerebellar development) (C) Applied Dataset 2 (medulloblastoma bulk RNA sequencing).

Topic modeling identifies conserved transcriptional programs across cerebellar development datasets

Our computational framework transfers developmental transcriptional signatures (topics) across datasets and biological contexts. Figure 1 illustrates an overview of the workflow: topics are derived from the source dataset (Aldinger et al., PCW 17), validated in an independent developmental dataset (Cao et al., PCW 12.5), and applied to medulloblastoma bulk RNA-seq to connect developmental programs with disease states.

Analysis of the Aldinger et al. dataset (subset to 15,556 nuclei from post-conception week 17) (Figure 1A) using topic modeling identified 65 topics. Our method included the use of iterative modeling to select the optimal number of topics based on perplexity (Table S1 and see Methods). Perplexity is a statistical measurement that quantifies how well a model predicts on held-out data. By selecting a minimum plateau in perplexity across iterations allows for selecting a model with less redundancy while maintaining complexity. Initially, these topics are given arbitrary numeric identifiers that we then renamed based on biological function.

Then, we transferred the source topics to the independent Cao et al. dataset (subset to 119,954 nuclei from week 12.5) using GSVA (Figure 1B) of the top 50 genes per topic from the source dataset (Table S3), these topics showed conservation despite differences in technology (SPLiT-seq vs sci-RNA-seq3) and developmental timepoints, validating their biological relevance. These topics were then tested on Applied dataset 2 in a medulloblastoma application to verify that the same topics related to GNPs and the EGL are also identifiable in medulloblastoma bulk RNA sequencing (Figure 1C). This dataset is a reference landscape for human brain disease of which we subset for medulloblastoma (n=876 tumors).

To specifically identify topics that may represent the EGL, we (1) identified topics with the presence of known EGL marker genes and pathways among their top-ranked gene-topic associations, (2) compared topic-gene associations with the top differentially expressed genes (DEGs) from the EGL through hierarchical clustering analysis, and (3) examined GSVA-based UMAP localization to granule neuron, granule cell progenitor, and rhombic lip cell type clusters in the source dataset with existing expert annotations as well as follow up expert validation. The selected topics were then validated across independent datasets using GSVA-based topic transfer and through spatial validation via in situ hybridization of top marker genes from each topic in mouse cerebellar tissue.

4.3. Results

4.3.1. Topics Identify Transition Cell States in the External Granule Layer (EGL)

The resulting 65 topics derived from the source dataset included a subset that captured the EGL which we identified through combinations of EGL-based genetic markers and RL-derived lineages within the EGL (Figure 4.2), identified through the presence of known RL marker genes and EGL genes. The cerebellar EGL is organized into distinct zones from the pial surface to the internal granule layer (IGL). The outer EGL contains mitotic and proliferative cells, including early transient Rhombic Lip-derived (RL) and locally proliferating progenitors (Topic 22) and intermediate progenitors (Topic 2). The transition zone harbors differentiating neuronal cells. The inner EGL consists of post-mitotic and pre-migratory neurons, including differentiating cells (Topic 49), early differentiated granule neurons (GNs) (Topic 46), and post-mitotic cells (Topic 61). Mature neurons migrate from the inner EGL toward the IGL, with transitioning mature GNs represented by Topic 25. Glial progenitors (Topic 27) span multiple zones throughout this developmental trajectory.

These topics were selected from the pool of 65 topics through three complementary approaches. First, these topics were identified through the presence of known EGL marker genes and pathways among their top-ranked gene-topic associations, including proliferation markers *MKI67* and *CENPF*, SHH pathway component *GLI2*, and the human-specific *ARHGAP11B* associated with progenitor expansion, alongside EGL differentiation markers for the neuronal lineage topics. Second, hierarchical clustering analysis comparing topic-gene associations with the top differentially expressed genes (DEGs) from the EGL identified in the Aldinger et al. study showed that Topics 49, 22, 2, 27, and 46 showed the strongest selective enrichment for these EGL-specific markers (Figure S4.3). Third, we examined GSVA-based UMAP localization to granule neuron, granule cell progenitor, and rhombic lip cell type clusters (Figure 4.3A) in the source dataset. The resulting topics begin to reveal previously unrecognizable substructure within the EGL developmental continuum. They capture cellular heterogeneity that conventional clustering methods struggle to resolve without extensive expert knowledge and manual curation.

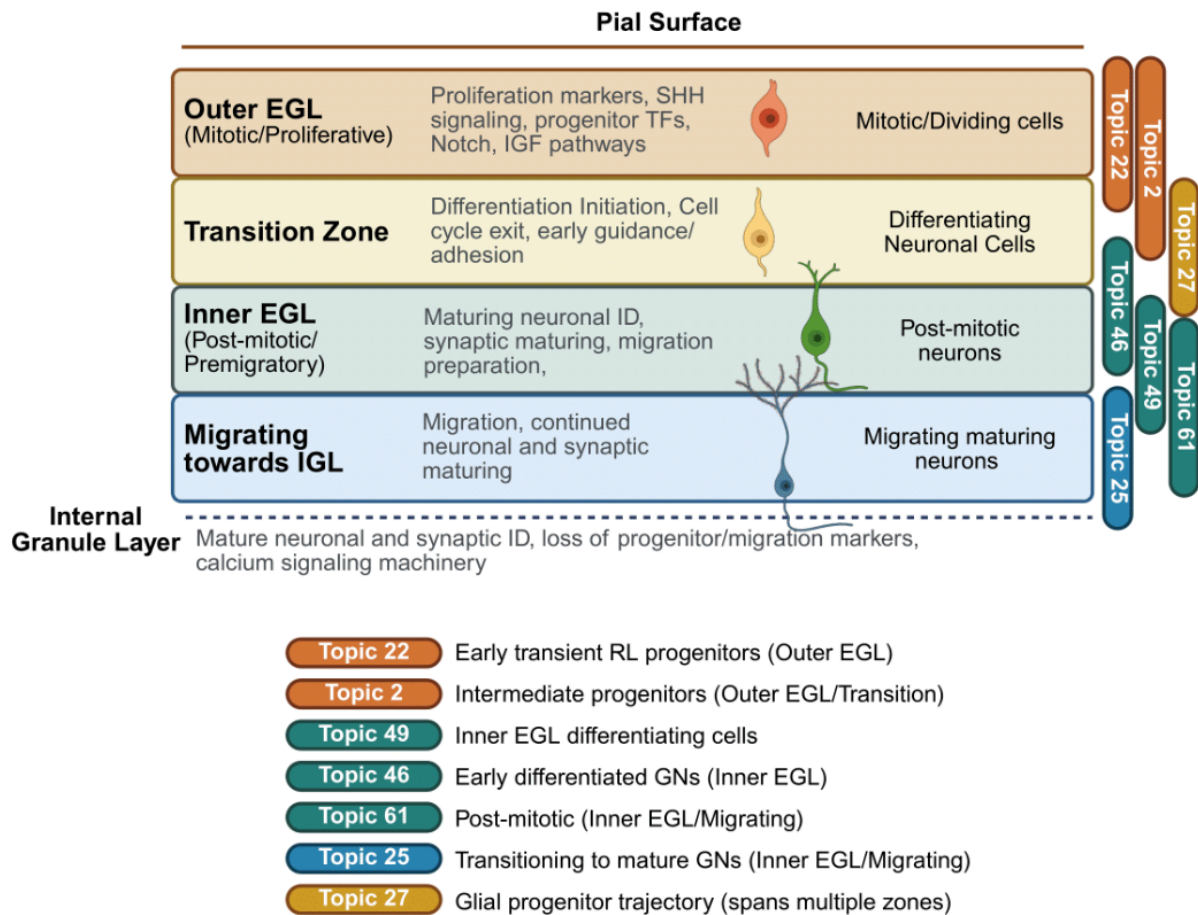


Figure 4.2. Spatial organization of neurogenic and gliogenic cell populations across the external granule layer during cerebellar development.

Schematic representation of cerebellar EGL structure showing the distribution of seven EGL-associated topics (Topics 2, 22, 25, 27, 46, 49, 61) across developmental zones from outer EGL to internal granule layer. Topics are color-coded and labeled according to their spatial localization patterns.

Early Proliferative Rhombic Lip Progenitors in the Outer EGL (Topic 22)

Topic 22 represents an early, transient rhombic lip-derived early proliferating progenitor population genetic signature marked by proliferation genes *MKI67*, *CENPF*, and *PARD3* (centrosome marker) indicating active cell division. This topic also contains *GLI2* (essential for SHH signaling). The presence of *PARD3*, a subpolarity gene, is consistent with cells undergoing active spatial reorganization and is critical for germinal zone exit³⁷. Additional proliferation and cell division genes include *CDK6* (cell cycle regulator), *EZH2* (chromatin remodeling), *MPPED2*, *DIAPH3* (cytoskeleton regulation), *STARD13*, *KNTC1* (kinetochore function), and *CIT* (cytokinesis). This population may be associated with cerebellar foliation processes, as evidenced by the presence of *ARHGAP11B*, a human-specific gene that has been linked to increased cerebral gyration and cerebellar foliation^{101–103}. In addition, Topic 22 has its strongest expression in the rhombic lip (RL) with very minimal expression across other cell types (Figure 4.3B), consistent with its identity as an early progenitor signature. The topic-cell association values projected using UMAP (Figure 4.3E) shows Topic 22 is sharply localized to a small discrete population within the RL cluster (upper left region), with virtually no expression in the large GN/GNP cluster. This type of pattern reflects the transient and spatially restricted nature of this progenitor state at this developmental timepoint. The exact temporal relationship between Topic 22 and Topic 2 cannot be definitively ordered, as both represent early transient states.

Intermediate Progenitor State within EGL Transition (Topic 2)

Topic 2 represents an intermediate progenitor state starting to transition from the outer EGL to differentiated populations, characterized by expression of *SOX2*, *MEIS1*, *PAX3*, and *NFLA*. *MEIS1* is a transcription factor that promotes BMP signaling and *ATOHI* degradation, driving cell cycle exit and differentiation^{37,104}. This topic contains SHH pathway components (*PTCH1*, *GLI2*) essential for glia survival^{37,105,106}. The co-expression of *GRIA4* (astrocyte-associated), *SOX2-OT* (glial marker), and *NFLA* (downstream

of *PAX3* in oligodendrocyte precursor specification) suggests this captures progenitors at a critical lineage decision point. The proliferation signature (*MKI67*, *CENPF*) combined with *GLI2* expression without *LMX1A/ATOH1* confirms its differentiating EGL progenitor identity. Like Topic 22, this population contains *PARD3* expression consistent with its transitional nature and may be present during active cerebellar foliation associated with *ARHGAP11B* expression. Additionally, Topic 2 has the highest expression in Bergmann glia (BG) and astrocytes (Ast/Ependymal, Ast), with moderate expression in glia and GNP, suggesting a glial-progenitor intermediate state (Figure 4.3B). Through scaled topic-cell expression patterns (Figure 3D), Topic 2 is enriched in both the glial cluster and portions of the GN/GNP cluster, with particularly strong expression at the boundary between these populations, consistent with cells at a progenitor lineage decision point bridging neuronal and glial fates.

Glial Progenitors diverging from Neuronal EGL Lineage (Topic 27)

Topic 27 represents an early glial progenitor trajectory located in the outer rhombic lip, enriched in radial and Bergmann glia annotated and related progenitor populations. This topic is identifiable with markers *PTN*, *SMOC1*, *TNC*, and *ETV5* (apical, basal radial and Bergmann glial markers) alongside *SLC1A3* (definitive glial transporter) and *GDF10* (glial marker). The presence of *PAX3* and downstream effector *NFIA* links this population to oligodendrocyte precursor specification⁷⁸. Additionally when compared to existing cell type annotations (Figure 4.3A-B), Topic 27 showed highest expression in Bergmann glia (BG), astrocytes (Ast, Ast/Ependymal), and radial glia, with some expression in RL, suggesting a glial progenitor identity. This is reflected in the scaled GSVA values (Figure 4.3C) where Topic 27 is strongly enriched in the large glial/astrocyte cluster, with minimal expression in the GN/GCP cluster, demonstrating clear spatial segregation consistent with commitment to the glial rather than neuronal fate. This topic likely represents early glial progenitors distinct from the neuronal lineage trajectories captured by the other topics.

Early Post-Mitotic Cells in Inner EGL initiating Differentiation (Topic 49)

Topic 49 contains a geneset that represents a population of recently differentiated cells within the inner EGL. These cells are characterized by expression of NEUN (*RBFOX3*), *RELN*, *PAX6*, and *ROR*, consistent with their post-mitotic identity. Notably, *RELN* expression shows enrichment in the outer relative to the inner EGL, a pattern confirmed in in situ hybridization of mouse brain at P4 from the Allen Cell Atlas (Figure S6C). The very low gene-topic association of *ATOH1* suggests these are transitioning cells that have moved beyond the early progenitor stage¹⁰⁷. The presence of *NEUROD1*, which is initially expressed during neuronal differentiation, indicates that Topic 49 likely represents cells upstream of *NEUROD1* expression in the differentiation trajectory. This population also expresses synaptic adhesion molecules (*LRRTM4*, *NLGN1*, *NLGN4X*) and guidance cues (*ROBO2*, *DSCAM*), suggesting these are migrating granule neuron progenitors beginning synaptic organization⁷⁸. The scaled GSVA values (Figure 4.3F) reveals that topic 49 is localized to specific regions within the granule neuron cluster from the cluster annotations. This is consistent with cells that would be at an early post-mitotic stage of differentiation.

Inner EGL Post-Mitotic Neurons with developing Synaptic Connectivity (Topic 61)

Topic 61 represents a distinct inner EGL genetic signature characterized by *RBFOX3* (NEUN) expression, which marks post-mitotic neurons and definitively excludes these cells from the outer EGL compartment (Lee et al., 2020). This population expresses glutamate receptors (*GRIK2*, *GRIA4*), synaptic proteins (*RELN*), and calcium signaling components (*RYR3*, *PRKCB*), indicating these cells are establishing functional connectivity while still undergoing maturation. This population appears to represent a slightly later stage in the differentiation process compared to Topic 49, representing another transition state within the inner EGL as granule neuron progenitors commit to their neuronal fate. These populations of cells that have high expression of this topic may represent novel "states" of EGL cells that have not been previously well-characterized. Topic 61 is also enriched in intermediate regions of the granule neuron cluster (Figure 4.3G), that is consistent with cells that have progressed beyond early differentiation but have not completed maturation.

Maturing Neurons Transitioning from inner EGL towards Internal Granule Layer (Topic 25)

Topic 25 represents a more differentiated neuronal state among these populations, characterized by *ROBO1* (definitive EGL marker), mature synaptic proteins (*UNC13C*), and neuronal specification factors (*RBFOX1*, *MEG3*). This population also expresses *DSCAM* (granule neuron marker), *RELN* (EGL attractant), and *GRIK2* (Lee et al., 2020). When compared to the existing annotations of the cells (Figure 3B), Topic 25 showed strong association with granule neurons (GN) and moderate expression in eCN/UBC, suggesting a mature neuronal signature. From UMAP visualization of the scaled GSVA values (Figure 3H), Topic 25 is enriched throughout much of the GN cluster with a relatively uniform distribution, extending into some adjacent populations, consistent with cells that have completed migration and are integrating into cerebellar circuits. These cells may represent an intermediate state between the inner EGL and fully differentiated granule neurons that will eventually migrate to form the internal granule layer⁷⁸.

Progenitor State with sustained plasticity in EGL (Topic 46)

Topic 46 represents a slightly distinct population marked by *MEIS1* (which promotes cell cycle exit and differentiation¹⁰⁴), *PDE1A* (expressed throughout granule cell development), and *DACH1* which are all markers of EGL identity. However, this population also expresses neural developmental transcription factors including *ERBB4*, *NFIA*, *NFIB*, *TCF4*, and *CHD7* (*CHD7* promotes clonal expansion of granule neuron progenitors^{37,108}). *ERBB4* is found in the inner EGL and in regions distinct from Purkinje cells, suggesting these may be among the earliest born granule neurons. The expression of *ZEB1* is particularly notable, as this factor controls neuron

differentiation and germinal zone exit through regulation of polarity genes including *PARD3* and *PARD6*¹⁰⁹. *ZEB1* is downregulated as GNPs exit the EGL, and its persistent expression delays morphological maturation and migration to the IGL. Importantly, *ZEB1* expression is elevated in SHH medulloblastoma and has been identified as a subgroup-specific marker of prognosis and potential therapeutic target¹¹⁰. This topic also contains synaptic adhesion molecules (*LRRTM4*, *NLGN1*, *NLGN4X*) and the CSMD family of cell surface glycoproteins (*CSMD1*, *CSMD2*, *CSMD3*), along with additional neural development markers including *LHFPL3*, *NTNG1* (axon guidance), *AGBL4*, and *STK32B*.

When compared to preexisting cell type annotations, Topic 46 had strongest expression in eCN/UBC, GCP, and BG, with notable expression also in RL and glia, indicating a broad expression pattern across multiple cell types (Figure 3B) though strongly in the RL. Additionally, Topic 46 (Figure 4.3I) has diffuse enrichment spanning the RL region, portions of the GN/GCP cluster, and extending into glial populations, confirming its expression across multiple cell clusters, consistent with a rare or multipotent progenitor state. These findings represent some of the first clear distinctions of granule neuron proliferation and differentiation states, particularly the transition between outer and inner EGL compartments.

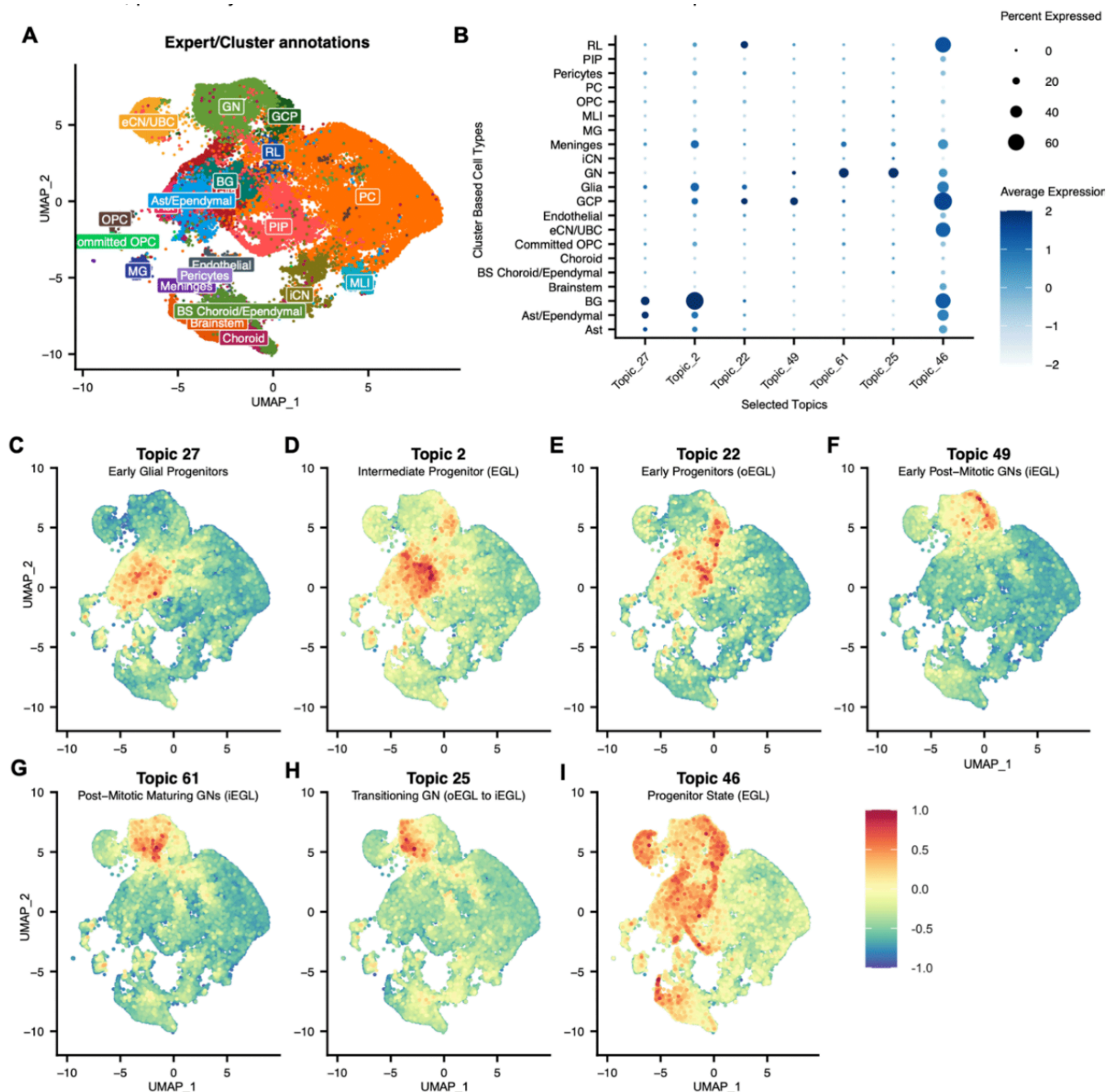


Figure 4.3. Topic modeling reveals distinct transcriptional programs defining rhombic lip progenitor lineage trajectories in human cerebellar development.

(A) UMAP visualization of 15,556 nuclei from post-conception week 17 (Aldinger et al. source dataset)⁷⁸ colored by original cell type annotations, including rhombic lip (RL), granule cell precursors (GCP), granule neurons (GN), Purkinje cells (PC), molecular layer interneurons (MLI), cerebellar nuclei neurons (iCN/eCN), oligodendrocyte precursors (OPC), committed OPCs, astrocytes (Ast), Bergmann glia (BG), microglia (MG), and vascular/meningeal populations.

(B) Dot plot showing average expression (color intensity) and percentage of cells expressing (dot size) six key topics across major preannotated cerebellar cell types. Topics capture the rhombic lip-to-glial (Topics 27, 2) and rhombic lip-to-neuronal (Topics 22, 49, 61, 25) developmental trajectories.

(C-H) UMAP projections of individual topic expression scores from GSVA analysis. Color scale represents scaled GSVA scores.

(C) Topic 27 (Early Glial Progenitors)

(D) Topic 2 (Intermediate Progenitor (EGL))

(E) Topic 22 (Early Progenitors (oEGL))

(F) Topic 49 (Early Post-Mitotic GNs (iEGL))

(G) Topic 61 (Post-Mitotic Maturing GNs (iEGL))

(H) Topic 25 (Transitioning GN (oEGL to iEGL))

Cross-dataset validation confirms robust conservation of developmental programs

To assess whether topics generated using the Source data (Figure 4.1A) can be used to identify similar developmental programs in a different dataset (Figure 4.1B), regardless of how that dataset was processed, we applied source topics to applied dataset 1 (Cao et al.)³. Applied dataset 1 is another human developmental cerebellar dataset using sci-RNA-seq3 where we used subset for PCW 12.5 (Figure 4.4A). Here, we focus on the topics above that overlap with the EGL lineages: Topics 27, 2, 22, 49, 61, and 25. These topics showed consistent enrichment patterns (Figure 4.4B) despite differences in sequencing technology (SPLiT-seq vs sci-RNA-seq3) and specimen ages.

Topic 27 (Figure 4.4C) and Topic 2 (Figure 4.4D) showed enrichment patterns consistent with glial progenitors and intermediate EGL states, respectively. On average, they both congregated in the cluster previously identified broadly as ‘astrocytes’ (Figure 4.4B). However, they both identify substructure within this cluster and show overlap with part of the granule neuron cluster at different intensities (Figure 4.4C-D).

Topic 22, which we identified as marking early progenitors in the oEGL, identified a distinct population in the Cao dataset that had not been annotated as such in the original study. This population is located between the clusters previously identified as astrocytes and granule neurons which is the expected region where rhombic lip derived early progenitors would exist (Figure 4.4E).

Topics 49 (Figure 4.4F), 61 (Figure 4.4G), and 25 (Figure 4.4H), which capture the progression from early neuronal differentiation into maturation, reveals previously unrecognized substructure within the granule neuron cluster in the Cao dataset. Topic 49 consistently marked cells at the periphery of the granule neuron cluster, suggesting these cells represent newly differentiating cells beginning their migration from the EGL. Topic 61 is enriched in the intermediate regions, while Topic 25 is concentrated in the cluster core, indicating the most mature granule neurons. This spatial organization within the UMAP recapitulates the known developmental trajectory, providing orthogonal validation of these topic identities.

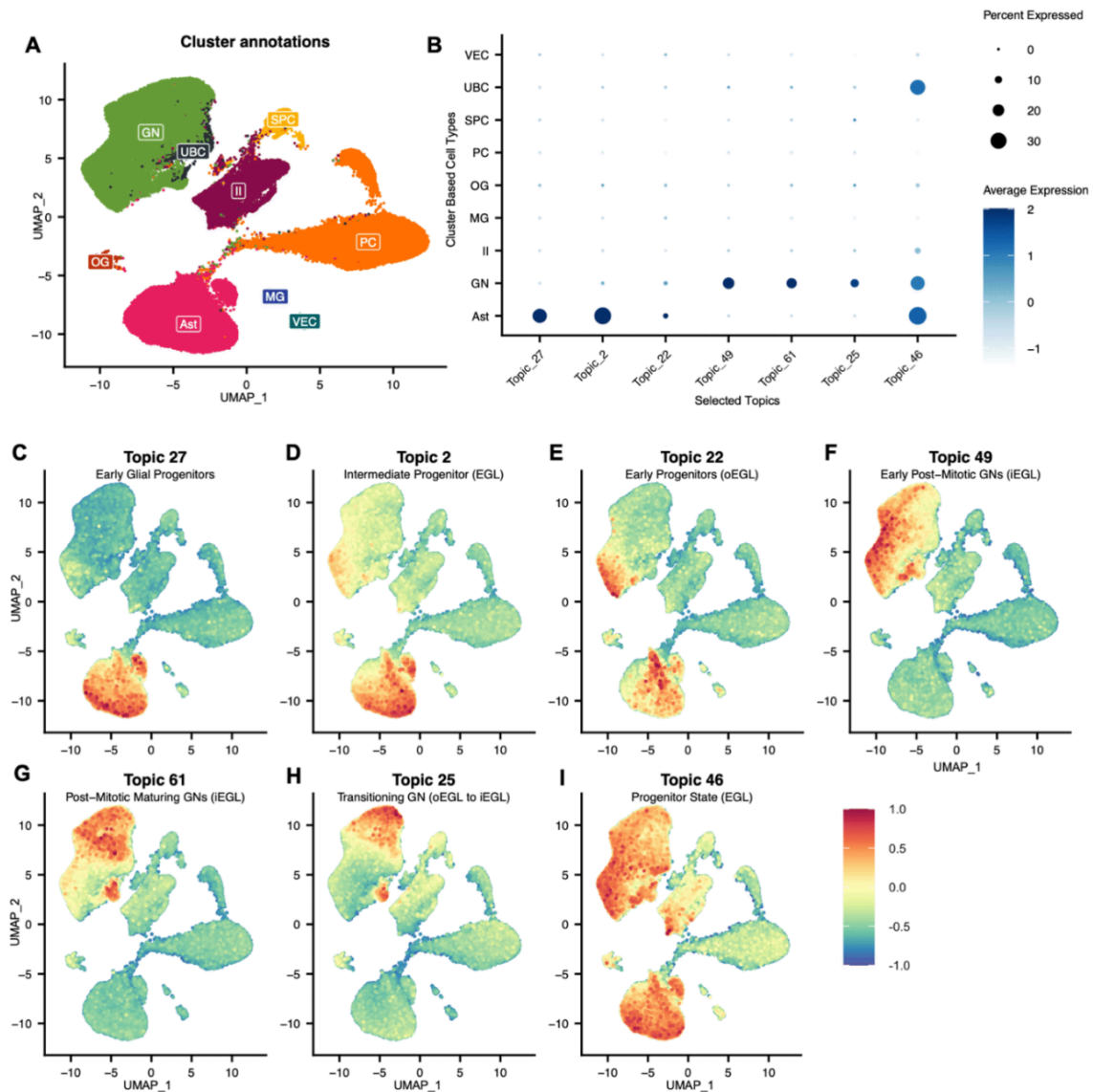


Figure 4.4. Cross-dataset validation demonstrates robust conservation of rhombic lip lineage topics from source dataset in an independent human developmental cerebellar dataset (applied data).

(A) UMAP visualization of 15,556 nuclei cells from post-conception week 12.5 (Cao et al. applied dataset)³ colored by original cell type annotations, including granule neurons (GN), unipolar brush cells (UBC), astrocytes (Ast), Purkinje cells (PC), oligodendrocytes (OG), microglia (MG), molecular layer interneurons (II), vascular endothelial cells (VEC), and SLC24A4+/PEX5L+ cells (SPC).

(B) Dot plot showing average topic expression (color intensity) and percentage of cells expressing (dot size) for the same six rhombic lip lineage topics identified in the source dataset, now applied to this independent dataset via GSVA transfer.

(C-H) UMAP projections of individual topic expression scores following GSVA-based topic transfer. Color scale represents scaled GSVA scores.

(C) Topic 27 (Early Glial Progenitors)

(D) Topic 2 (Intermediate Progenitor (EGL))

(E) Topic 22 (Early Progenitors (oEGL))

(F) Topic 49 (Early Post-Mitotic GNs (iEGL))

(G) Topic 61 (Post-Mitotic Maturing GNs (iEGL))

(H) Topic 25 (Transitioning GN (oEGL to iEGL))

Spatial validation confirms EGL localization of topic marker genes

To validate that our computationally-derived topics capture genuine developmental programs with anatomical correlates in the EGL, we examined the spatial expression patterns of key topic marker genes using in situ hybridization from the Allen Developing Mouse

Brain Atlas^{8,49} at postnatal day 4. We selected P4 as it represents a developmental stage when the mouse EGL is well-established and actively producing granule neurons, analogous to the mid-gestation human cerebellar development captured in our source dataset. As a secondary validation, we confirmed that human topics successfully transferred to mouse cerebellar single-cell data (Supplementary Figure S4.4, using gene conversion Table S4.4), where they localized to predicted cell types in the developmental mouse atlas (Figure S4.4A-B), demonstrating cross-species conservation of these transcriptional programs. Each of the topics was projected onto the mouse developmental dataset and scaled GSVA scores begin to find substructure in the broader cell type categories in the CNS (Figure S4.4C-I).

For spatial validation, we systematically examined nine genes representing core markers across our EGL-associated topics available through the atlas, focusing on their localization to EGL versus non-EGL compartments. These nine genes showed concordance to their predicted anatomical regions as defined below.

Spatial validation of Pan-EGL and Transitional Markers

PAX6 is a canonical marker of RL derivatives expressed throughout granule neuron development. *PAX6* showed robust, specific expression throughout the EGL with particularly strong signal in both outer and inner zones (Figure 4.5A), confirming its role as a pan-EGL marker. Computationally, *PAX6* showed strong association with Topic 49 (association value: 14,193), which represents early post-mitotic granule neurons in the inner EGL. The strong association to *PAX6* validates our topic assignment to the EGL and in conjunction with the other top markers of Topic 49 we can distinguish oEGL proliferative states from iEGL differentiating states.

BOC (brother of *CDO*), a component of the Hedgehog signaling pathway, displayed expression in the inner EGL with some extension into the outer EGL (Figure 4.5B), consistent with its role in maintaining *SHH* responsiveness in GNPs. Computationally, *BOC* showed strongest associations with Topic 2 (17,698) representing intermediate progenitors in the EGL, Topic 27 (15,636) marking glial progenitors, and Topic 22 (12,736) representing early progenitors in the outer EGL, supporting its expression across multiple EGL compartments and progenitor lineages.

RELN (Reelin), strongly associated with Topics 61 (50,282), 49 (24,603), and 25 (36,146) representing the progression from inner EGL post-mitotic neurons through mature granule neurons, revealed critical spatial organization within the EGL. *RELN* showed robust expression throughout the EGL with enrichment in the outer EGL relative to the inner EGL (Figure 4.5C). This outer-to-inner gradient validates our topic assignments that distinguish oEGL proliferative states from iEGL differentiating states, demonstrating that our topics capture spatial organization within the EGL structure.

MEIS1, strongly associated with Topic 2 (33,190) representing intermediate progenitors and Topic 46 (24,443) marking progenitors with sustained developmental plasticity, exhibited expression throughout the EGL with particular enrichment in differentiating zones (Figure 4.6D). This spatial pattern is consistent with *MEIS1*'s known role in promoting cell cycle exit and differentiation, validating Topic 2's identity as an intermediate progenitor at the transition from proliferation to differentiation.

Spatial Validation of Inner EGL and Transitioning Populations

GRIK2 strongly associated with Topics 61 (53,729) and 25 (37,713) represents inner EGL post-mitotic maturing granule neurons and transitioning granule neurons, showed expression in the inner EGL and differentiating granule neurons with signals extending from the iEGL into regions preparing for migration (Figure 4.5F). The spatial distribution of this synaptic marker, concentrated in the iEGL, confirms that Topics 61 and 25 capture sequential stages of functional maturation as granule neurons exit the germinal zone and establish synaptic connectivity.

EPHA6 associated with Topic 25 (36,018) represents transitioning and maturing granule neurons, displayed expression in the IGL (Figure 4.5E), consistent with more mature, post-migratory granule neurons that have completed their transition from the EGL. This localization validates Topic 25's identity as capturing the late stages of granule neuron maturation and migration.

ERBB4, showing strong computational associations with Topics 49 (46,196), 61 (40,835), and 25 (30,986) spanning the inner EGL through transitioning populations, as well as Topic 46 (15,232) marking progenitors with sustained plasticity, displayed robust expression in the inner EGL (Figure 4.5I). This spatial pattern supports its role in marking granule neurons undergoing differentiation and early maturation within the iEGL compartment, consistent with its presence in multiple topics capturing the iEGL-to-migration transition.

Spatial validation of Progenitor and Early Differentiation Markers

NFIA is a transcription factor that has a high topic-gene association with Topic 46 (43,944), which represents progenitors with sustained plasticity, Topic 49 (19,015) marking early post-mitotic neurons, and Topic 2 (13,044) representing intermediate progenitors, displayed strong expression throughout the EGL in progenitor zones (Figure 4.5G). This spatial pattern validates *NFIA*'s role in maintaining developmental flexibility across multiple progenitor and early differentiation states captured by our topics.

TCF4, a neural developmental transcription factor showing strongest computational association with Topic 46 (42,576) representing progenitors with sustained developmental plasticity, showed expression in progenitor populations within and adjacent to the EGL (Figure 4.5H). This spatial pattern supports Topic 46's identity as marking cells that retain developmental flexibility even as they progress through early differentiation stages.

Across all nine genes examined, spatial expression patterns were consistent with predicted topic associations and developmental functions. EGL-enriched genes (*PAX6*, *BOC*, *RELN*, *MEIS1*, *NFIA*, *TCF4*) showed robust localization to the germinal zone, with several displaying layer-specific patterns (outer vs inner EGL) that correspond to the oEGL-to-iEGL progression captured by our topics, specifically Topics 2, 22, 27, and 46). Genes associated with neuronal maturation (*GRIK2*, *EPHA6*, *ERBB4*) showed expression patterns consistent with the inner EGL through post-migratory populations, validating the sequential developmental stages captured by early (Topic 2, 22, 27, 46) through later transitions (Topics 49, 61, and 25).

This spatial validation begins to demonstrate that our topic modeling framework successfully identifies biologically meaningful and anatomically organized developmental programs within the EGL. The concordance between computational topic assignments (based on gene-topic association values from Table S4.2) and physical tissue localization (from in situ hybridization) provides orthogonal confirmation that extends beyond computational predictions.

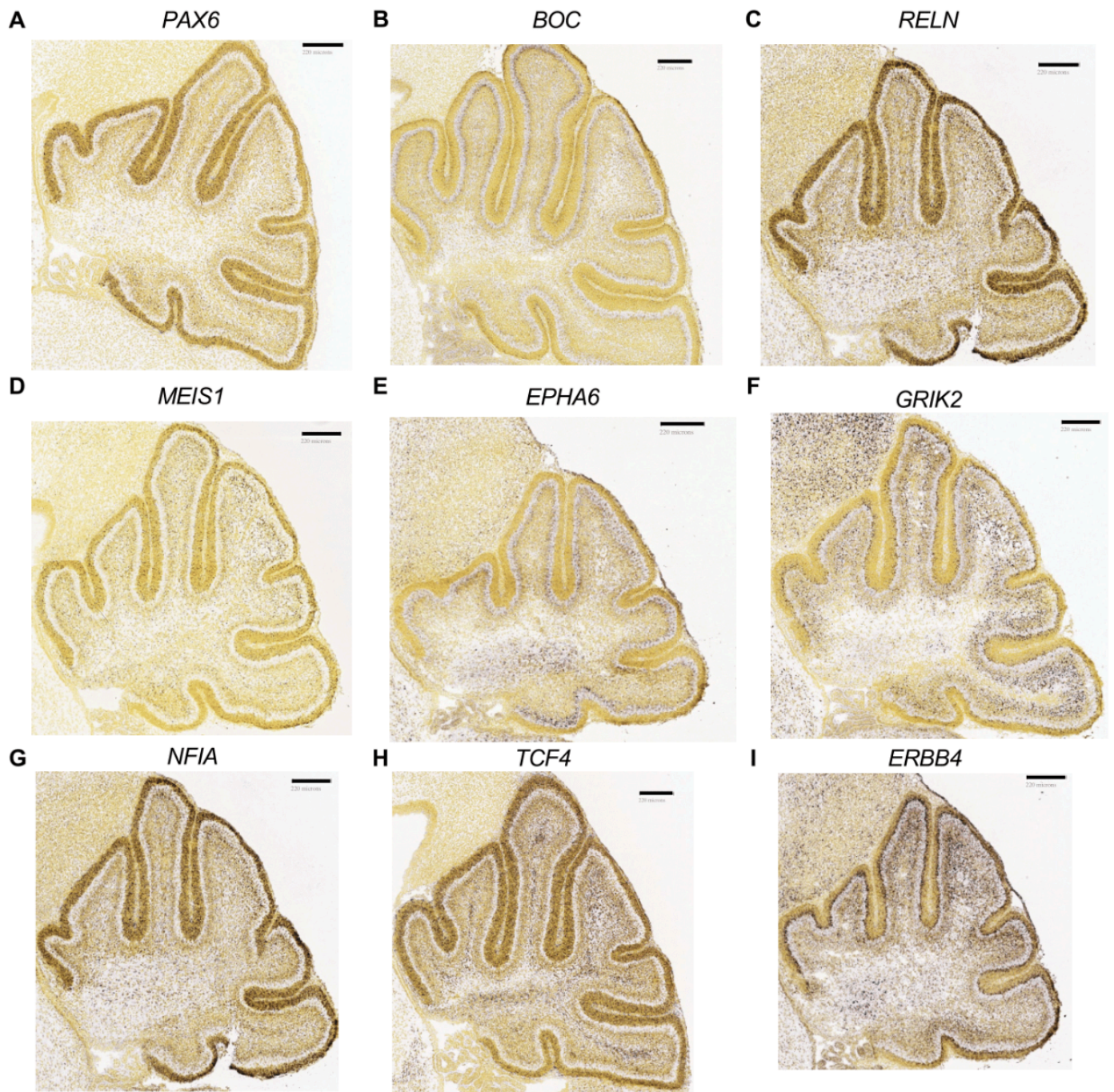


Figure 4.5. Spatial validation of EGL-associated topic marker genes in developing mouse cerebellum at postnatal day 4. In situ hybridization images from the Allen Developing Mouse Brain Atlas showing spatial localization of genes computationally identified as markers of EGL-associated topics. Gene-topic associations are based on association values from Table S2. Scale bars = 220 μm with darker color indicating positive cells.

- (A) *PAX6*
- (B) *BOC*
- (C) *RELN*
- (D) *MEIS1*
- (E) *EPHA6*
- (F) *GRIK2*
- (G) *NFIA*
- (H) *TCF4*
- (I) *ERBB4*

SHH medulloblastoma retains human EGL transition states

To validate our framework's ability to bridge developmental and cancer contexts, we applied cerebellar developmental topics from the source dataset to medulloblastoma samples from a published bulk RNA sequencing atlas⁹⁵ which includes 876 medulloblastoma tumors (274 labeled as SHH). The UMAP visualization reveals clear separation of the four molecular subgroups, WNT, SHH, Group 3, and Group 4, with SHH forming a distinct cluster (Figure 6A). Within the 274 SHH tumors, the dataset is further divided into four subtypes (SHH α , SHH β , SHH γ , and SHH δ) (Figure 6E) that generally follow age-specific patterns (Figure 6B). These SHH tumors show a clear age distribution with infant cases (<3 years) predominantly in SHH β / γ , childhood cases (3-16 years) in SHH α , and adult cases (>17 years) in SHH δ (Figure 6C), with relatively balanced gender distribution across SHH subtypes (Figure 6D). Each subtype has different 5 year survival rates with SHH δ having the highest rates at around 80% followed by SHH β , SHH γ , then SHH α with the worst at about 40%^{42,89,111} (Figure 6E).

Application of source topics showed several with high hierarchical association with SHH subgroup, specifically topics 2, 49, 22, and 46, with variable expression across other subgroups (Figure 4.7A). This gradient across subgroups suggests the granule cell precursor relationship with SHH medulloblastoma and varying degrees of developmental program retention across subgroups.

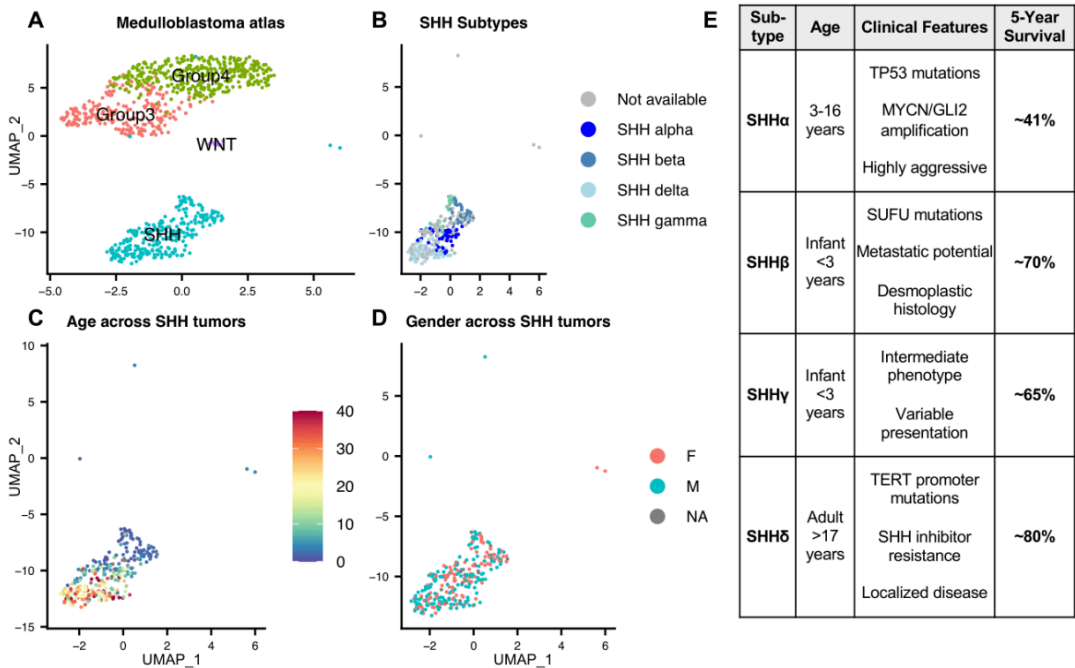


Figure 4.6. Atlas of medulloblastoma bulk RNA sequencing (Applied dataset 2).
(A) UMAP visualization of 876 medulloblastoma tumors colored by molecular subgroup: WNT (n=70), SHH (n=274), Group 3 (n=144), and Group 4 (n=388).
(B) UMAP of SHH subgroup subset showing four molecular subtypes (SHH α , SHH β , SHH γ , SHH δ).
(C) UMAP of SHH subgroup subset colored by age distribution across tumors.
(D) UMAP of SHH subgroup subset colored by gender distribution across tumors.
(E) Clinical characteristics table summarizing age ranges, molecular features, clinical behavior, and 5-year survival rates for each SHH subtype.

SHH medulloblastoma subtypes associate with distinct developmental programs

The four SHH medulloblastoma subtypes showed distinct topic associations that corresponded with their age of onset. Each subtype appears to retain transcriptional signatures from a specific stage of EGL development (associated with the topics), suggesting that the developmental timing at tumor initiation fundamentally shapes tumor biology.

SHH δ tumors occur predominantly in adults with TERT promoter mutations and showed preferential enrichment for Topic 2 (Figure 6B and E)¹¹². This topic contains SHH pathway components *PTCH1* and *GLI2*, transcription factors *MEIS1*, *DACH1*, and *NFIA*, alongside additional markers whose functions are less characterized (*EGFEM1P*, *NAALADL2*, *XYLT1*) and are distinct for the adult tumors versus pediatric.

SHH α tumors arise in children aged 3-16 years with frequent TP53 mutations and MYCN/GLI2 amplifications⁹⁷. These tumors showed almost exclusive expression of Topic 22 (Figure 4.7C and E). This topic contained proliferation-associated genes *MKI67* (cell division marker), *CDK6* (cell cycle regulator), *EZH2* (chromatin remodeling), and *GLI2* (SHH signaling), alongside additional proliferation and cell division genes including *MPPED2*, *DIAPH3* (cytoskeleton regulation), *STARD13*, *KNTC1* (kinetochore function), and *CIT* (cytokinesis)⁷⁷. This represents the most proliferative, outer EGL-like signature among SHH subtypes which is consistent with this subtype's aggressive clinical behavior and poor prognosis.

SHH β tumors, arising in infants under 3 years with frequent SUFU mutations⁹⁷, showed co-enrichment of Topic 46 (Figure 7E), containing neural developmental transcription factors *NFIA*, *NFIB*, *TCF4*, *CHD7*, and *MEIS1*. The presence of *ZEB1*, which regulates germinal zone exit, alongside markers of developmental plasticity suggests that these tumors retain characteristics of a progenitor state with sustained differentiation potential. These same SHH β tumors also demonstrated predominant enrichment for Topic 49 (Figure 4.7C and E), which contained synaptic adhesion molecules (*LRRTM4*, *NLGN1*, *NLGN4X*), the *CSMD* family of cell surface glycoproteins (*CSMD1*, *CSMD2*, *CSMD3*), and additional neural development markers including *LHFPL3*, *NTNG1* (axon guidance), *AGBL4*, and *STK32B*. This co-enrichment pattern suggests SHH β tumors arise from or retain characteristics of inner EGL cells initiating differentiation and synaptic development. SHH γ tumors, also occurring in infants, showed intermediate expression patterns with moderate enrichment for Topics 2 and 49, but lacking the strong Topic 46 co-enrichment observed in SHH β (Figure 4.6E-F).

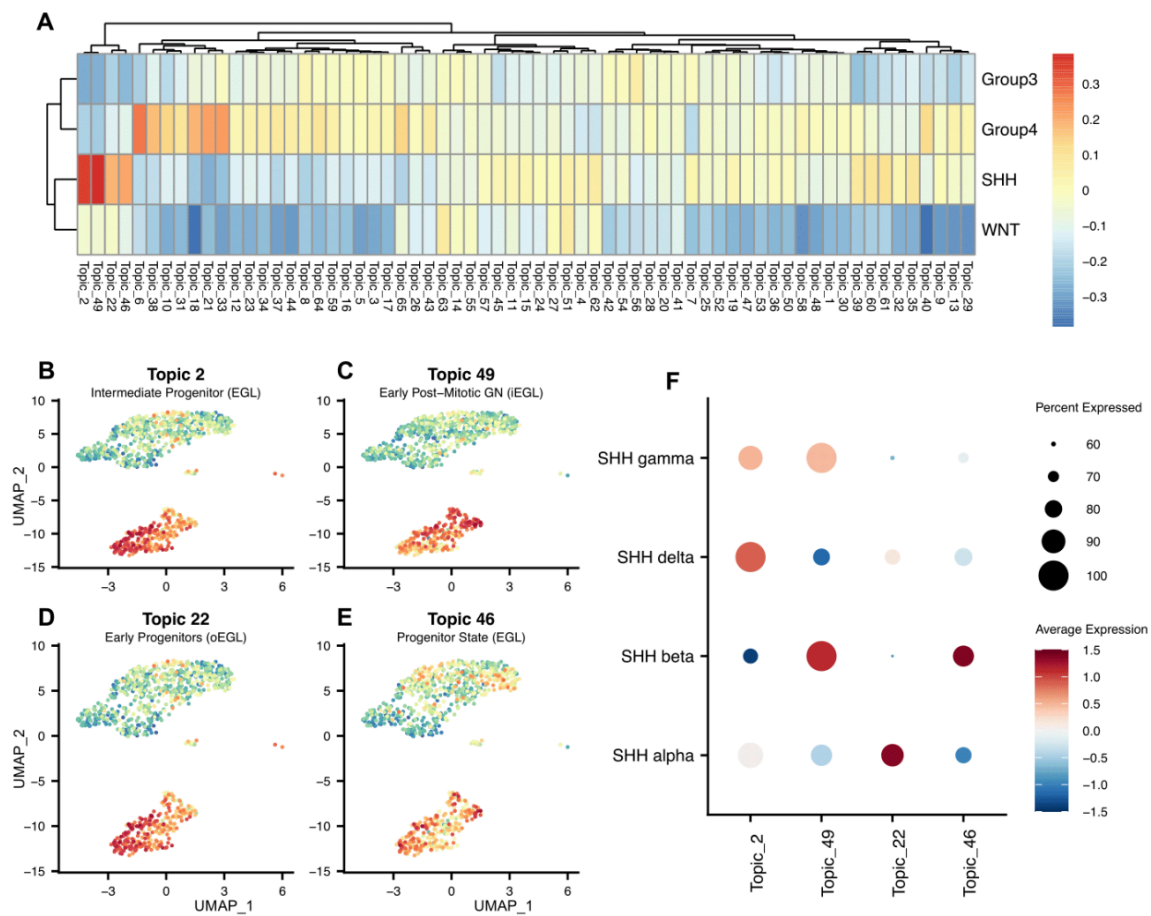


Figure 4.7. SHH subtypes express different developmental source topics and correspond to specific stages of EGL transitions. (A) Heatmap showing hierarchical clustering of 65 cerebellar developmental topics (columns) across medulloblastoma subgroups (rows). Color intensity represents scaled GSVA enrichment scores, with red indicating high expression and blue indicating low expression. (B-D) UMAP projections of topic expression scores across all 876 medulloblastoma samples, with color intensity representing GSVA enrichment scores. (B) Topic 2 (Intermediate Progenitor/EGL) (C) Topic 49 (Early Post-Mitotic GNs/iEGL) (D) Topic 22 (Early Progenitors/oEGL) (E) Topic 46 (Progenitor State/ EGL) (F) Dot plot quantifying topic association across SHH subtypes with dot size indicating percent of tumors under SHH subtype categorization with non-zero GSVA scaled score of the topics and color indicating average scaled GSVA score.

4.4. DISCUSSION

Topic modeling addresses fundamental challenges in comparative genomics

In the era of big data in genomics, several critical challenges continue to perplex the field especially in the context of statistical and machine learning: the lack of generalization of models to novel experimental conditions, the complexity of interpreting learned models (“black box”), and the application to complex cell dynamics. Here, we show how topic modeling can directly address these challenges and validate in the context of neurodevelopment and pediatric brain tumor origins.

The conservation of topics across datasets with different technologies, processing methods, and biological contexts as we show here with sci-RNA-seq3, SPLiT-seq, and bulk-RNA-seq, demonstrates that topic modeling captures fundamental biological information rather than dataset-specific artifacts. This generalization reflects the principle that cells coordinate gene expression in modules that drive specific biological functions^{113,114}. This transferable topic modeling framework learns these modules from genetic signatures from source data and projects them onto target datasets through GSVA. This enables cross study comparison without computationally intensive integrations and dramatically reduces data loss as all genes are retained in the analysis.

In addition, the interpretability of topic modeling derives from its probabilistic, modular structure. Each topic consists of genes with quantified association strengths, enabling direct biological interpretation through pathway analysis, literature review, and spatial validation. Our EGL-associated topics map onto known developmental processes (proliferation (Topics 22, 2), glial specification (Topic 27), and neuronal differentiation (Topics 49, 61, 25, 46)) while revealing previously unrecognized transitional states within these broad categories. The topic-gene association values provide quantitative rankings that can be validated through orthogonal approaches such as in situ hybridization, as we demonstrated for nine marker genes in developing mouse cerebellum. This type of interpretability aligns with the goal of deriving biological insights from computational models rather than treating them as black boxes.

The ability to capture cell dynamics through transitional states represents perhaps the most significant advance that builds upon conventional clustering. Standard clustering fragments or entirely misses populations undergoing rapid state transitions due to their small size and intermediate expression profiles³². Topic modeling's probabilistic assignments allow cells to express multiple topics simultaneously, naturally capturing the gradual nature of differentiation. Our identification of Topic 22 as a rhombic lip progenitor signature exemplifies this capability. These cells were present in multiple independent datasets but missed by conventional analysis due to their small numbers and transitional nature. The topic emerged not because we searched for rhombic lip markers, but because these genes consistently co-occurred in a subset of cells, revealing an underlying biological program invisible to discrete clustering approaches.

Biological Insights into EGL Lineage Transitions and Specification

Our analysis reveals previously underappreciated complexity in the EGL while building upon conventional single cell and bulk RNA sequencing analysis. The identification of distinct transcriptional programs for the more proliferative progenitors (Topics 27 and 2) versus the oEGL-to-iEGL neuronal progression (Topics 22, 49, 61, 25) shows that these lineage decisions involve fundamentally different molecular mechanisms. The progenitor signatures appear more direct, with Topic 27 capturing multipotent glial progenitors. The presence of both TNC and ETV5 (glial markers) alongside SLC1A3 and GDF10 (broader glial markers) in Topic 27 suggests these progenitors retain flexibility. The neuronal lineage shows complexity above conventional clustering with our topics capturing a developmental spectrum from proliferative oEGL through differentiating iEGL to migrating granule neurons. This progression involves at least five distinct transcriptional states (Topics 22, 49, 46, 61, 25). This reflects the nuances of granule neuron development, which requires massive clonal expansion in the oEGL, cell cycle exit and initial differentiation in the transition zone, further differentiation and synaptic specification in the iEGL, radial migration through multiple cerebellar layers, and finally synaptic integration with Purkinje cells and other neuronal populations. Each stage requires distinct transcriptional programs, which our topics successfully delineate.

Developmental Timing and Tumor Biology

The subtype-specificity of topic expression within SHH medulloblastomas provides compelling evidence that developmental timing fundamentally shapes tumor transcriptional architecture. Rather than all SHH tumors arising from identical cellular origins with subsequent divergence, our results suggest that the developmental stage at tumor initiation, influenced by patient age, determines which transcriptional programs persist in the tumor itself.

Adult SHH δ tumors' enrichment for Topic 2 suggests retention of an intermediate EGL progenitor state core SHH signaling components. These tumors arise in a cerebellum that has completed development, where any tumorigenic transformation must occur in rare residual or reactivated progenitors that existed during development. The combination of SHH pathway activity with mature transcription factor networks may provide alternative survival mechanisms when the primary SHH pathway is blocked, potentially explaining the variable response to therapeutic inhibitors observed clinically in adults^{111,112}. In addition, the relatively uncharacterized markers *EGFEMIP*, *NAALADL2*, and *XYLT1* identified in this signature warrant further investigation as potential age-specific markers distinguishing adult from pediatric SHH tumors.

Infant SHH β tumors' enrichment for Topic 49 reflects their origin during active cerebellar development when the inner EGL is actively generating differentiating granule neurons engaged in synaptogenesis and migration. The abundance of synaptic adhesion molecules (*LRRTM4*, *NLGN1*, *NLGN4X*) and the complete *CSMD* family suggests these tumors arise from or retain characteristics of inner EGL cells that are locked in a specific developmental window focused on establishing neural connections. The co-enrichment of Topic 46's neural transcription factors (*NFIA*, *NFIB*, *TCF4*, *CHD7*, *MEIS1*) indicates these tumors retain broader developmental plasticity, potentially contributing to their metastatic potential and these cells may maintain the migratory and adaptive programs of normal EGL development. The unique markers *LHFPL3*, *NTNG1*, *AGBL4*, and *STK32B* could serve as infant-specific biomarkers for diagnostic stratification.

Childhood SHH α tumors' exclusive expression of Topic 22 represents a particularly aggressive phenotype where outer EGL-like proliferation programs dominate without accompanying differentiation signals. The concentration of cell cycle genes (*MKI67*, *CDK6*, *EZH2*) alongside factors associated with genomic instability (*KNTC1* overexpression linked to chromosomal instability, *DIAPH3* regulating cytoskeleton dynamics, *CIT* essential for cytokinesis) creates a molecular signature consistent with their dismal prognosis (5-year survival ~41%)^{89,115}. This hyperproliferative outer EGL-like state, combined with TP53 loss, generates uncontrolled proliferation and genomic instability that may explain why these tumors have the worst outcomes among SHH subtypes despite occurring in children who would otherwise have better regenerative capacity than adults. The dominance of proliferation markers suggests potential vulnerability to cell cycle inhibitors, particularly CDK4/6 inhibitors targeting *CDK6* or EZH2 inhibitors targeting the chromatin remodeling machinery^{42,89,111,112}, though further in vitro validation followed by in vivo and clinical validation would be required.

SHH γ tumors' intermediate expression pattern, with moderate enrichment for Topics 2 and 49 but lacking strong Topic 46 co-enrichment, may reflect their transitional position between infant and childhood tumor biology. This pattern suggests these tumors retain some developmental flexibility without being locked into either a more proliferative outer EGL-like state of SHH α or the inner EGL synaptic differentiation program of SHH β . Further investigation with larger sample sizes and in vitro and in vivo validation is needed to fully characterize this subtype's developmental relationships and clinical implications.

Methodological Advances and Broader Applications

Beyond the specific biological insights, this work showcases topic modeling as a powerful tool for comparative and big-data genomics. The ability to transfer topics across datasets without batch correction or shared processing pipelines solves challenges as the field generates ever-larger atlases with evolving technologies. This transferability corresponds to topics capturing genuine biological programs that are beyond the technical variations.

The framework's ability to work across data modalities, as shown here in single cells to bulk tumors, is particularly valuable for translational research. Many clinical samples are only available as bulk RNA-seq due to historical technologies, sample quality, quantity, or cost constraints. Additionally, the current gold standard analysis for integrating single cell datasets is to reanalyze the original datasets with a consistent, current framework across the datasets. Original fastq files are not always available for multiple reasons and therefore not all prior datasets can be reanalyzed in this way. The topic analysis overcomes this limitation and provides a method for comparing and expanding the original analyses.

Our demonstration that developmental topics remain detectable and informative in bulk tumor samples opens new avenues for analyzing clinical cohorts and understanding how developmental programs contribute to diverse pathologies.

Future Directions and Implications

The topics and statistical framework presented here have immediate applications beyond medulloblastoma and cerebellar development. Other pediatric brain tumors suspected to have developmental origins, including other embryonal tumors, pediatric gliomas, and ependymomas, could be analyzed using the same cerebellar developmental topics or topics derived from appropriate developmental atlases. Adult brain tumors might also retain or reactivate developmental programs detectable through this approach.

More broadly, this framework could transform how we approach the relationship between development and disease across all organ systems. As developmental atlases become available for heart, lung, kidney, and other organs, the same methodology could identify which developmental programs persist in congenital anomalies, contribute to adult disease, or become reactivated in cancer. The key insight is that development provides a template for understanding disease and by comprehensively mapping normal developmental programs we continue to create a reference frame for understanding pathology.

Additional future directions includes further in vitro validation of the genetic signatures within the topics explored here to further strengthen the applicability of these programs. In this case, exploration of the top 5-10 genes from the topics associated with the RL and bifurcating trajectories, particularly those distinguishing oEGL from iEGL states, can be explored in new human developmental tissue, mouse models, and medulloblastoma samples. Additionally, validation of these topics in different cancer treatment settings including chemotherapy induced cytodifferentiation, can provide insights into altered differentiation trajectory in tumors as well. Patient-derived organoid models could be particularly valuable for testing whether manipulation of stage-specific EGL programs alters tumor behavior or drug sensitivity.

As single-cell technologies generate ever-larger datasets, approaches that extract meaningful patterns without computationally intensive integrations become necessary. Topic modeling meets this need by providing a scalable framework for building cumulative knowledge from the exponentially growing corpus of transcriptomic data. The topics and methods presented here are immediately applicable to other biological systems and malignancies, offering a path forward for integrating datasets across different biological contexts and technologies. By using this framework in the context of neurodevelopment and early cancer pathogenesis, we showcase how hard to identify transitional cellular populations can be characterized efficiently. This characterization was transferred to tumor data and we showed persisting developmental features from specific stages of EGL development that may drive medulloblastoma subtype diversity and could inform personalized treatment strategies.

Having validated topic modeling in mouse development (Chapter 3) and demonstrated successful transfer across human developmental datasets and into medulloblastoma, we have established that this framework captures biologically meaningful and clinically relevant transcriptional programs. Chapter 5 extends this approach to transcript-level resolution, revealing that developmental context shapes not only which genes are expressed but also how they are spliced. This adds a layer of regulatory complexity with important implications for understanding pediatric brain tumor biology.

Limitations of the study

Limitations of this study include that topic number selection requires balancing granularity with interpretability. Different values might reveal additional insights. The uncharacterized markers and marker combinations found with these topics require further functional validation to confirm their roles in development and disease in human. Our analysis focused on a limited pre-natal development time whereas earlier and later stages would complete the developmental picture. Despite these limitations, the framework showcased valuable insights including conserved programs across independent datasets, cell populations hidden from commonly used clustering approaches, and connected developmental programs to cancer subtypes



5. CHAPTER 5. NTRK Family Isoforms and Signatures Associates with Pediatric Cancers Revealed by Transcript-Level Topic Modeling

The preceding chapters established probabilistic topic modeling as a powerful framework for identifying coordinated gene expression programs across developmental and disease contexts. Chapter 3 demonstrated that topic modeling captures biologically meaningful transcriptional programs in normal mouse development, while Chapter 4 revealed that pediatric brain tumors retain developmental signatures that associate with clinical outcomes. These findings established gene-level topic modeling as a valuable tool for identifying cells in transition within developmental datasets and their relationships to cancer. However, gene-level quantification collapses expression across all transcript isoforms produced from a single locus, masking regulatory complexity that may be central to both developmental biology and disease pathogenesis.

Most human genes undergo alternative splicing, which generates multiple transcripts through a variety of mechanisms including differential exon inclusion, alternative splice site usage, and alternative promoter or polyadenylation site selection. These isoforms can encode proteins with distinct or opposing functions, different subcellular localizations, and altered regulatory properties. During development and in disease, specific isoform expression combinations are coordinated through splicing factor networks that respond to developmental signals, creating context-dependent splicing landscapes that gene-level analysis cannot resolve. Aberrant alternative splicing is increasingly recognized as a hallmark of cancer biology^{116,117}, and in pediatric cancers specifically, splicing dysregulation during developmental processing is linked to tumor formation^{38,42,116–118}. Neuroblastoma cells reactivate fetal splicing patterns through aberrant expression of RNA-binding proteins^{118–120}. Wilms tumor shows distinctive splicing of the WT1 gene, with the ratio of different isoforms affecting transcriptional versus splicing functions^{121,122}.

The NTRK family of neurotrophin receptors (NTRK1 (TrkA), NTRK2 (TrkB), and NTRK3 (TrkC)) demonstrates a case for the importance of isoform-level analysis. Each NTRK locus produces multiple transcripts through alternative splicing, generating both full-length kinase-active receptors and truncated kinase-inactive variants with distinct or opposing biological functions. Yet clinical assays and most research studies measure total NTRK expression without isoform resolution. NTRK1/TrkA expression has long been associated with favorable prognosis in neuroblastoma, where high TrkA levels predict spontaneous tumor regression^{123,124}. NTRK2/TrkB plays dual roles where its full-length kinase-active form mediates BDNF-dependent survival signaling in aggressive neuroblastoma^{125,126}, while the truncated TrkB.T1 variant is an oncogenic driver on its own. Pattwell and colleagues demonstrated that TrkB.T1 overexpression is sufficient to drive tumorigenesis in mice through Wnt, TGF- β , and Sonic Hedgehog pathways independently of kinase activity^{127–129}. Furthermore, TrkB.T1 was shown to be the predominant NTRK2 isoform across embryonic organogenesis and highly expressed in a wide range of adult and pediatric tumors, including those harboring various NTRK fusions¹³⁰. NTRK3/TrkC has been linked to differentiation-competent neuroblastoma subtypes and plays essential roles in sympathoadrenal development^{124,129}. As detailed in a recent comprehensive review from the Pattwell lab, the NTRK family has gained substantial prominence in precision oncology following FDA approval of TRK inhibitors, yet our understanding of individual isoform contributions remains incomplete¹²⁹.

This chapter begins to explore the hypothesis that NTRK family isoforms co-express distinctly from each other either according to developmental context (i.e. disease lineage) or structural or functional classifications (i.e. kinase domain status) in pediatric cancer. Here we use transcript level topic modeling of the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) pediatric cancer dataset. The specific objectives of this chapter are as follows. First, we adapt transcript-level topic modeling to the TARGET pediatric cancer dataset to discover coordinated isoform programs across multiple model resolutions ($K=20$ through $K=95$). Second, we identify all NTRK family transcripts within the topic landscape and characterize their disease-specific expression patterns. Third, we employ Spearman correlation analysis within NTRK-containing topics to define the co-expression neighborhoods of each NTRK isoform, testing whether these neighborhoods reflect developmental programs or kinase-based groupings. The chapter is organized as follows: Section 5.2 provides background on NTRK family biology, precision oncology considerations, and the TARGET dataset. Section 5.3 describes the computational methods for transcript-level topic modeling and NTRK isoform analysis. Section 5.4 presents results demonstrating that sixteen NTRK-containing topics partition into renal and neural crest programs, that individual NTRK transcripts show disease-specific expression independent of gene family membership, and that Spearman correlation reveals distinct co-expression networks organized by developmental lineage rather than receptor identity. Section 5.5 discusses the implications for NTRK biology and TRK inhibitor therapy, and Section 5.6 provides conclusions.

5.1. Background

5.1.1. The NTRK Family of Neurotrophin Receptors

The NTRK (neurotrophic receptor tyrosine kinase) family comprises three members (NTRK1, NTRK2, and NTRK3) that encode the TrkA, TrkB, and TrkC receptors, respectively. These receptors mediate cellular responses to the neurotrophin family of growth factors: nerve growth factor (NGF) binds preferentially to TrkA, brain-derived neurotrophic factor (BDNF) to TrkB, and neurotrophin-3 (NT-3) to TrkC, with NT-3 also showing lower-affinity binding to TrkA and TrkB¹²⁹ (Jagana et al., 2026). Upon neurotrophin binding, full-length Trk receptors dimerize and autophosphorylate, activating canonical downstream signaling cascades including the RAS-MAPK, PI3K-AKT, and PLC γ pathways, which promote neuronal survival, differentiation, and synaptic plasticity during nervous system development.

Each NTRK locus generates multiple transcript isoforms through alternative splicing, producing both full-length kinase-active receptors and truncated kinase-inactive variants that lack the intracellular tyrosine kinase domain. Truncated isoforms were historically viewed as dominant-negative regulators that sequester ligand without activating downstream signaling. However, this view has been fundamentally revised. Truncated TrkB isoforms, particularly TrkB.T1, possess kinase-independent signaling capabilities, activating RhoA, mobilizing intracellular calcium, and modulating cellular morphology through distinct mechanisms¹²⁹ (Jagana et al., 2026). The Pattwell laboratory demonstrated that TrkB.T1 overexpression is sufficient to drive tumorigenesis in mice, with the truncated variant engaging developmental signaling pathways such as Wnt, TGF- β , and Sonic Hedgehog, independently of kinase activity^{127–129}. Furthermore, TrkB.T1 was identified as the predominant NTRK2 isoform across embryonic organogenesis and was found to be highly expressed in a broad range of adult and pediatric tumors, including those harboring NTRK fusions¹³⁰. These findings establish that truncated isoforms are active signaling molecules with context-dependent oncogenic functions.

5.1.2. NTRK in Pediatric Cancer and Precision Oncology

The NTRK family has become a focal point in precision oncology following FDA approval of TRK inhibitors (larotrectinib and entrectinib) for NTRK fusion-positive cancers. These agents target the intracellular kinase domain of full-length Trk receptors and have demonstrated remarkable efficacy in tumors driven by NTRK gene fusions regardless of histological type (Jagana et al., 2026). However, many current clinical assays typically measure total NTRK expression without isoform resolution, creating a critical knowledge gap. Tumors predominantly expressing truncated isoforms would test NTRK-positive yet may derive limited benefit from kinase inhibitors targeting the absent catalytic domain. Conversely, if truncated isoforms serve as oncogenic drivers through kinase-independent mechanisms such as demonstrated for TrkB.T1, then therapeutic strategies may need to target ligand binding or extracellular domains rather than kinase activity. Therefore, isoform resolution is essential for predicting therapeutic response and developing rational treatment strategies. For instance, the importance of NTRK family members in neuroblastoma biology has been recognized for decades. High TrkA expression is associated with favorable biology, younger age at diagnosis, and spontaneous regression in neuroblastoma^{123,124}. In contrast, TrkB/BDNF co-expression has been associated with aggressive, unfavorable neuroblastoma, promoting survival signaling^{125,131}. As reviewed by Pattwell, Arber, and Brodeur, the neurotrophin receptor family plays multifaceted roles throughout neuroblastoma development and progression¹³².

5.1.3. The TARGET Initiative and Developmental Lineage Diversity

The Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative provides RNA sequencing data across seven pediatric malignancy types representing three distinct developmental lineages. The neural crest lineage is represented by neuroblastoma (NBL, n=162). The renal lineage encompasses Wilms tumor (WT, n=132), clear cell sarcoma of the kidney (CCSK, n=13), and rhabdoid tumor (RT, n=5). The hematopoietic lineage includes acute lymphoblastic leukemia (ALL, n=194), acute myeloid leukemia (AML, n=196), and acute myeloid leukemia with induction failure (AML-IF, n=32). This developmental diversity provides an ideal system for testing whether NTRK isoform expression patterns reflect developmental origin versus functional classification. If isoforms from the same gene but with different kinase domain status consistently co-express, this would support a structural/functional model. If, instead, isoforms from different genes co-express when they share developmental lineage, this could support a developmental context model.

5.1.4. Technical Considerations for Transcript-Level Analysis

Accurate transcript-level analysis from short-read RNA-seq faces substantial technical challenges, particularly for complex loci like those encoding NTRK receptors. Short reads of 50 to 150 base pairs are computationally assigned to transcripts spanning tens of thousands of bases, and reads mapping to exons shared between isoforms are inherently ambiguous. The NTRK2 locus is particularly challenging because multiple isoforms differ only in their 3' terminal exons. Quantification tools such as Salmon employ expectation-maximization algorithms to probabilistically assign ambiguous reads, but accuracy depends critically on reference transcript annotations which may have differing transcript nomenclature and definitions across versions. The TARGET data analyzed here use Ensembl v104 annotations and Salmon v1.4.0 for transcript quantification. It is also imperative to note that several fundamental genes that are known to be associated with different cancers are missing in this dataset (i.e. P75 with aliases CD271, TNFRSF16, NGFR, P75NTR). Though this is a major limitation of the TARGET dataset---there are still meaningful findings in this chapter and to be found in this data.

Despite these limitations, transcript-level analysis captures biologically meaningful variation that gene-level aggregation obliterates in general. The multi-resolution topic modeling approach employed here, provides additional robustness by identifying patterns that are stable across model granularity.

5.2. Methods

5.2.1. Data Acquisition and Preprocessing

RNA sequencing data and clinical annotations were obtained from the TARGET initiative via dbGaP (accession: phs000218). Transcript-level quantification was performed using Salmon v1.4.0¹³³ with the Ensembl v104 transcriptome reference (GRCh38). Salmon employs a lightweight alignment algorithm combined with expectation-maximization to probabilistically assign reads to

transcripts, providing transcript-per-million (TPM) normalized expression values that account for both library size and transcript length. After quality control filters removed samples with fewer than 10 million aligned reads or mapping rates below 70%, resulting in a final cohort of 734 samples across seven disease types.

Latent Dirichlet Allocation (LDA) was applied to the transcript-level expression matrix using the methodology established in Chapters 3 and 4. Topic modeling was performed at multiple resolutions (K=20 through K=95 at intervals of 5), yielding 920 total topics across 16 K values. For each model, the top 1,000 transcripts per topic ranked by their topic-transcript association scores (β weights) were extracted. To enable quantitative analysis of topic expression across samples and diseases, Gene Set Variation Analysis (GSVA) was applied to convert topic-transcript associations into per-sample topic scores. For each topic, the top 100 most highly associated transcripts were extracted as a gene set, and GSVA was applied to the continuous transcript expression matrix ($\log_2(\text{TPM}+1)$ transformed). GSVA scores were scaled for visualization, with a range of -1 to +1.

5.2.2. NTRK Family Topic Identification and Profiling

All NTRK family transcripts were defined using Ensembl annotations. One hundred and thirty four NTRK transcripts were queried in the dataset, though 35 were present in the dataset. Topics were screened for the presence of any NTRK transcript within the top 100 ranked transcripts by topic-transcript association value. This identified 154 NTRK-containing topics spanning all model resolutions. For each NTRK-containing topic, the topic-transcript probability matrix was extracted. All unique transcripts from the top 100 positions across the 35 NTRK topics were assembled into a combined matrix (topics $\times$ transcripts). Individual NTRK transcript expression was visualized across disease types using violin plots of raw transcript counts grouped by primary disease code. To identify transcripts co-regulated with NTRK isoforms, Spearman rank correlation was computed across the combined topic-transcript probability matrix. For each NTRK isoform, the top 20 most positively correlated transcripts were extracted. Topics were clustered using Euclidean distance with complete linkage on the full transcript probability matrix. UMAP feature plots, violin plots, heatmaps, and dendrograms were generated using Seurat, ggplot2, pheatmap, and gg dendro in R.

A Note on NTRK Transcript Nomenclature

Ensembl transcript naming conventions (e.g., NTRK2-204, NTRK2-205) vary across Ensembl versions, databases, and publications. The same biological transcript may be designated differently depending on the annotation source. For unambiguous identification, all NTRK transcripts in this chapter are referenced by their stable Ensembl transcript IDs (ENST identifiers with version numbers), which remain consistent across contexts. Table 5.1 provides the correspondence between Ensembl IDs and transcript names used in this chapter based on Ensembl v104 annotations. Readers should note that in other sources, particularly earlier biological literature, ENST00000359847 (the major truncated TrkB.T1 isoform) is frequently referred to as "NTRK2-204", whereas in Ensembl v104 it is designated "NTRK2-205". We follow the Ensembl v104 nomenclature throughout but emphasize that the ENST identifiers are the authoritative references.

Table 5.1. Select NTRK Family Transcripts Detected in the TARGET Pediatric Cancer Topic Models

Ensembl Transcript ID	Gene	Transcript Name	Length (bp)	Biotype	Kinase Status
ENST00000358660	NTRK1	NTRK1-201	2,492	Protein coding	Intact kinase
ENST00000368196	NTRK1	NTRK1-202	2,701	Protein coding	Intact kinase
ENST00000392302	NTRK1	NTRK1-203	2,665	Protein coding	Intact kinase
ENST00000489021	NTRK1	NTRK1-204	570	Protein coding	Truncated
ENST00000497019	NTRK1	NTRK1-205	2,508	Nonsense-mediated decay	N/A
ENST00000524377	NTRK1	NTRK1-206	2,669	Protein coding	Intact kinase
ENST00000530298	NTRK1	NTRK1-207	3,052	Protein coding	Intact kinase
ENST00000531606	NTRK1	NTRK1-208	658	Protein coding	Truncated
ENST00000533630	NTRK1	NTRK1-209	645	Protein coding	Truncated
ENST00000534682	NTRK1	NTRK1-210	844	Protein coding	Truncated
ENST00000277120	NTRK2	NTRK2-201 (canonical)	8,666	Protein coding	Intact kinase
ENST00000304053	NTRK2	NTRK2-202	8,108	Protein coding	Intact kinase
ENST00000323115	NTRK2	NTRK2-203	8,723	Protein coding	Intact kinase
ENST00000359847	NTRK2	NTRK2-204 (TrkB.T1)	7,286	Protein coding	Truncated
ENST00000376208	NTRK2	NTRK2-205	8,633	Protein coding	Intact kinase
ENST00000376213	NTRK2	NTRK2-206	8,693	Protein coding	Intact kinase
ENST00000395882	NTRK2	NTRK2-207	7,499	Protein coding	Intact kinase
ENST00000317501	NTRK3	NTRK3-201	4,134	Protein coding	Intact kinase
ENST00000357724	NTRK3	NTRK3-202	2,980	Protein coding	Truncated
ENST00000394480	NTRK3	NTRK3-203 (825aa)	19,984	Protein coding	Intact kinase

ENST00000542733	NTRK3	NTRK3-204	2,287	Protein coding	Truncated
ENST00000557856	NTRK3	NTRK3-205	2,622	Protein coding	Truncated
ENST00000557897	NTRK3	NTRK3-206	398	Protein coding	Truncated
ENST00000558306	NTRK3	NTRK3-207	629	Protein coding	Truncated
ENST00000558576	NTRK3	NTRK3-208	717	Protein coding	Truncated
ENST00000558676	NTRK3	NTRK3-209	4,670	Protein coding	Intact kinase
ENST00000559188	NTRK3	NTRK3-210	535	Protein coding	Truncated
ENST00000559680	NTRK3	NTRK3-211	441	Protein coding	Truncated
ENST00000559764	NTRK3	NTRK3-212	538	Protein coding	Truncated
ENST00000560017	NTRK3	NTRK3-213	257	Protein coding	Truncated
ENST00000560201	NTRK3	NTRK3-214	424	Protein coding	Truncated
ENST00000560739	NTRK3	NTRK3-215	572	Protein coding	Truncated
ENST00000626019	NTRK3	NTRK3-216	2,741	Protein coding	Truncated
ENST00000629765	NTRK3	NTRK3-217	20,018	Protein coding	Intact kinase
ENST00000569588	NTRK3-AS1	NTRK3-AS1-201	2,663	lncRNA (antisense)	

5.3. Results

5.3.1. The TARGET Pediatric Cancer Cohort Resolves into Developmentally Distinct Transcriptional Clusters

The TARGET cohort consists of 734 samples across seven pediatric malignancy types representing three developmental lineages: hematopoietic (Acute Lymphoblastic Leukemia [ALL], n=194; Acute Myeloid Leukemia [AML], n=196; Acute Myeloid Leukemia with Induction Failure [AML-IF], n=32), neural crest (neuroblastoma [NBL], n=162), and renal (Wilms tumor [WT], n=132; Clear Cell Sarcoma of the Kidney [CCSK], n=13; rhabdoid tumor [RT], n=5). Transcript-level expression profiles separate by developmental origin, with hematopoietic malignancies forming a transcriptionally distinct group from solid tumors, and neural crest-derived neuroblastoma separating cleanly from renal lineage tumors (Figure 5.1A).

Analysis by sex reveals no major confounding effects on transcriptional clustering (Figure 5.1B). Age at diagnosis shows disease-specific patterns consistent with known epidemiology where neuroblastoma and Wilms tumor occur predominantly in early childhood (approximately 2-5 years), while ALL and AML show broader age distributions extending into adolescence (Figure 5.1C). Within the neuroblastoma cluster, younger patients occupy distinct transcriptional space from older children, consistent with the established inverse relationship between age at diagnosis and tumor differentiation status in this malignancy.

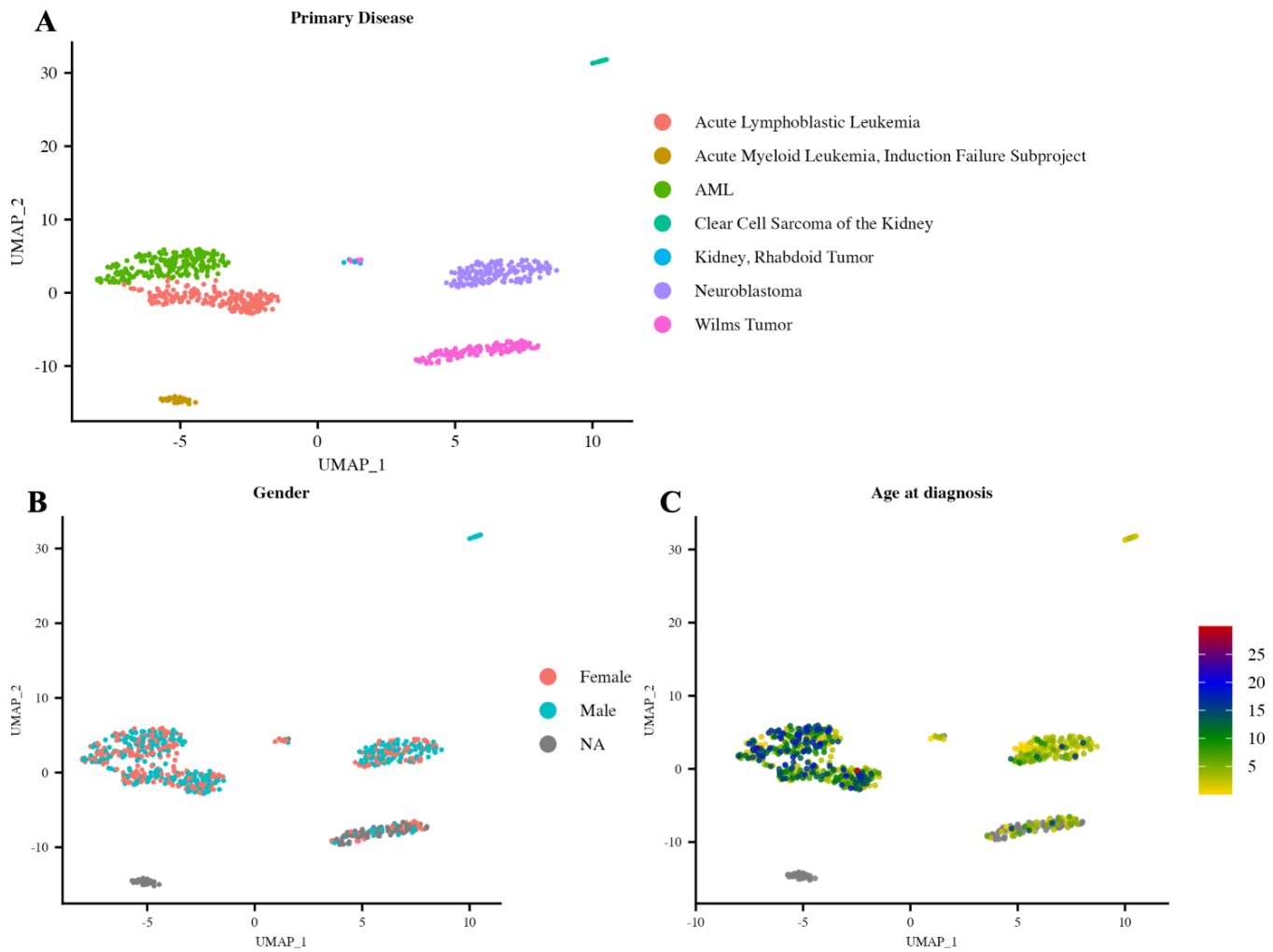


Figure 5.1. TARGET cohort overview.

(A) UMAP visualization colored by primary disease.

(B) UMAP colored by sex.

(C) UMAP colored by age at diagnosis (years).

5.3.2. Individual NTRK Transcripts Show Variable Expression Across Solid Tumors

NTRK family transcripts were detected across the TARGET dataset, spanning all three NTRK genes across all diseases. A consistent feature is that hematopoietic malignancies (ALL, AML, AML-IF) rarely express NTRK transcripts, while solid tumors show variable expression (Figure 5.1). Within the solid tumors, however, most isoforms show broadly overlapping distributions that do not cleanly delineate a single disease type.

NTRK2 transcript ENST00000395882 (NTRK2-207) shows the highest expression in WT, with positive median values, while also being detected at elevated levels in CCSK and RT (Figure 5.2A-B). Neuroblastoma shows intermediate expression with a distribution centered around zero. Hematopoietic malignancies demonstrate uniformly low expression of this transcript. The UMAP projection reveals that high-expressing samples localize specifically to the WT and renal tumor clusters, with expression intensity forming a gradient across the WT population. NTRK2 transcript ENST00000359847 (NTRK2-204/TrkB.T1, the truncated kinase-inactive isoform) demonstrates highest expression in WT and RT, with CCSK showing a broad distribution and neuroblastoma expression comparatively lower but still detectable above hematopoietic malignancies (Figure 5.2C-D). However, TrkB.T1 expression extends into portions of the NBL cluster adjacent to the renal tumors. NTRK3 transcript ENST00000317501 (NTRK3-201) is predominantly expressed in neuroblastoma and WT, with CCSK showing elevated but variable expression (Figure 5.2E-F). This transcript has heterogeneous expression within the NBL cluster, with highest-expressing samples concentrated in a subregion rather than uniformly distributed. NTRK1 transcript ENST00000524377 (NTRK1-206) shows specific expression in neuroblastoma alone, with all other disease types including other solid tumors demonstrating minimal to no expression (Figure 5.2H) although expression intensity varies within NBL, suggesting molecular heterogeneity across tumors (Figure 5.2G). NTRK3 transcript ENST00000394480 (NTRK3-203) is

expressed primarily in neuroblastoma with a broad distribution, while WT shows intermediate expression and CCSK shows variable expression (Figure 5.2I-J).

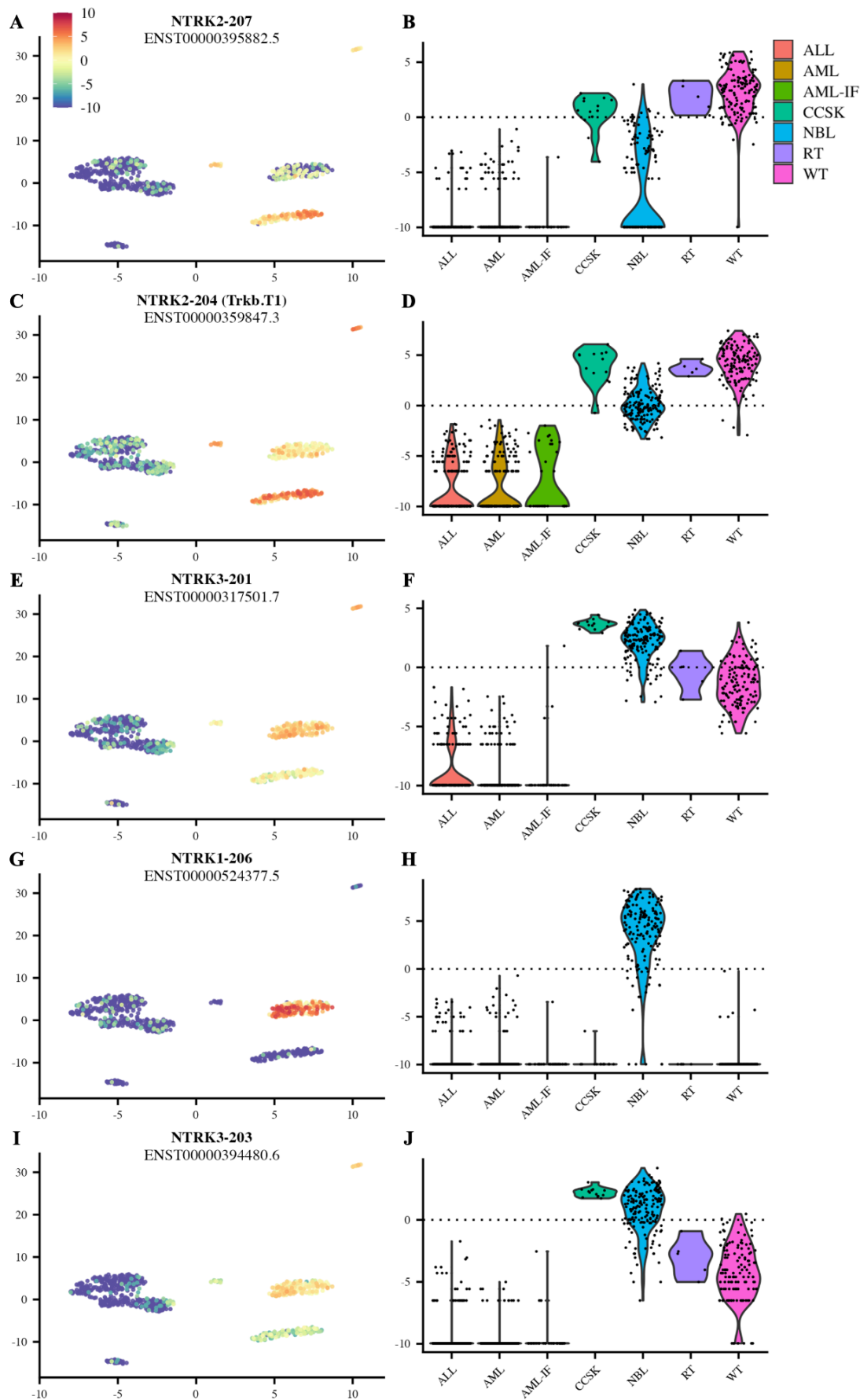


Figure 5.2. Disease-specific expression of individual NTRK family transcripts.
 UMAP projections (A, C, E, G, I) and violin plots (B, D, F, H, J) of transcript-level expression across the TARGET cohort.
 (A-B) NTRK2-207/ENST00000395882.
 (C-D) NTRK2-204/ENST00000359847.
 (E-F) NTRK3-201/ENST00000317501.
 (G-H) NTRK1-206/ENST00000524377.
 (I-J) NTRK3-203/ENST00000394480.

5.3.3. NTRK Isoforms Populate Distinct Topics Across the Model Landscape

Since individual NTRK transcripts may not delineate disease types on their own, they may be embedded within broader co-expression transcript signatures that achieve greater specificity. Across the 920 topics generated at K=20 through K=95, 35 NTRK transcripts appear in the top-ranked positions of at least one topic. Hierarchical clustering of the transcript-topic association weight matrix reveals that NTRK transcripts cluster by co-expression context rather than by gene of origin (Figure 5.3). NTRK3-201, NTRK2-204 (TrkB.T1), and NTRK1-207 form a distinct cluster with the highest association, while full-length kinase-active isoforms NTRK2-201 and NTRK2-207 cluster separately at lower weights. The majority of truncated NTRK3 isoforms (NTRK3-202 through NTRK3-216) cluster together with uniformly low topic involvement. This organization where isoforms from different genes cluster together while isoforms from the same gene segregate supports the hypothesis that developmental context rather than receptor identity determines NTRK co-expression. To characterize the biological programs associated with NTRK expression, we selected three focus topics with the highest-ranking NTRK placements (Table 5.2) Topic 80 from K=80 (NTRK3-201/ENST00000317501 at rank 7), Topic 16 from K = 80 (NTRK2-207/ENST00000395882 at rank 2), and Topic 62 from K = 70 (TrkB.T1/NTRK2-204/ENST00000359847 at rank 6).



Figure 5.3. Heatmap of NTRK transcript-topic weights across topics.
 Hierarchical clustering of 35 NTRK transcripts (rows) across NTRK-containing topics (columns). Color intensity represents topic-transcript probability weights. Clustering performed using Euclidean distance and complete linkage.

Table 5.2. Focus topics selected for detailed characterization based on NTRK transcript rank.
 Topics were selected from the 920 total topics (K=20-95) based on having an NTRK transcript in the top 10 ranked positions by transcript-topic association weight.

Topic	Transcript	Transcript-topic rank	Topic name
K = 80, Topic 80	NTRK3-201 ENST00000317501	7	NTRK3-201 Topic
K = 80, Topic 16	NTRK2-207 ENST00000395882	2	NTRK2-207 Topic
K = 70, Topic = 62	TrkB.T1 NTRK2-204 ENST00000359847	6	NTRK2-204 Topic

5.3.4. Focus Topics Achieve Disease Specificity Far Exceeding Individual Transcripts

Each of the three focus topics shows specific disease enrichment (Figure 5.4). The NTRK3-201 Topic demonstrates strong enrichment in neuroblastoma (Figure 5.4A-B). CCSK shows intermediate scores near zero, distinct from the uniformly negative scores in other non-NBL diseases. The UMAP projection confirms that high topic scores localize specifically to the NBL cluster with a specific

favoring towards one subgroup of neuroblastoma tumors. The NTRK2-207 Topic and NTRK2-204 Topic show strong enrichment in both Wilms tumor and rhabdoid tumor (Figure 5.4C-F). Both also display a similar patterns of a subset of WT tumors having a stronger enrichment of these topics (Figure 5.4C & E)

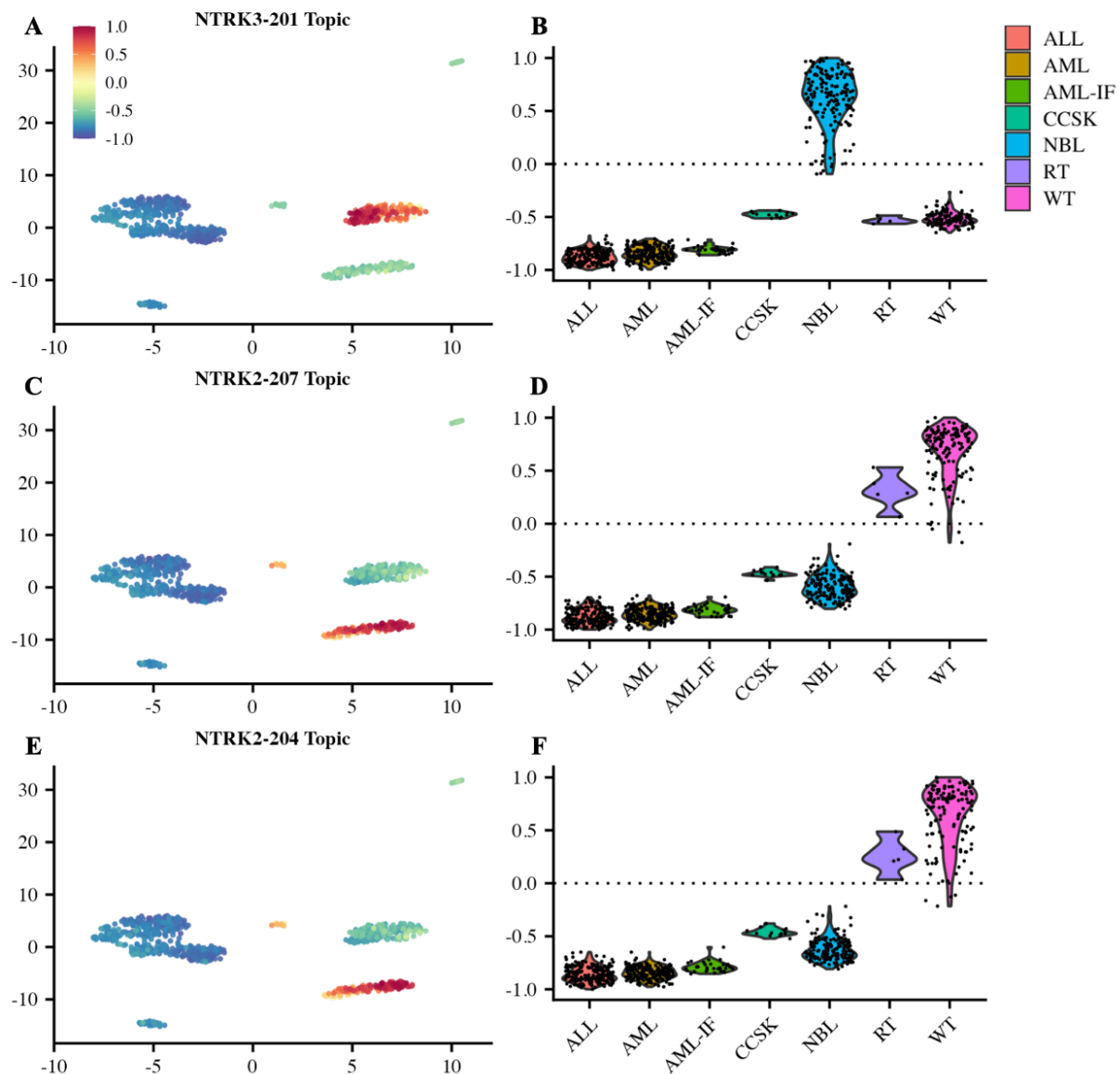


Figure 5.4. Disease-specific enrichment of the three focus NTRK topics.

UMAP projections (A, C, E) and violin plots (B, D, F) of scaled GSVA scores for each focus topic.

(A-B) NTRK3-201 Topic

(C-D) NTRK2-207 Topic

(E-F) NTRK2-204/TrkB.T1 Topic

5.3.5. The Three Focus Topics with NTRK Transcript Associations Contain Distinct Transcript Signatures

The three focus topics contribute largely non-overlapping gene sets (Figure 5.5). The NTRK3-201 Topic assigns high transcript-topic association weights to neural/neuroendocrine transcripts, while the NTRK2-207 and NTRK2-204 Topics assign weights to renal/developmental transcripts with distinct compositions. Transcripts highly ranked in one topic carry near-zero weight in the others. The NTRK3-201 Topic shares no transcripts with either renal topic at top-30 depth, confirming that neural crest and renal programs represent biologically independent transcriptional repertoires.

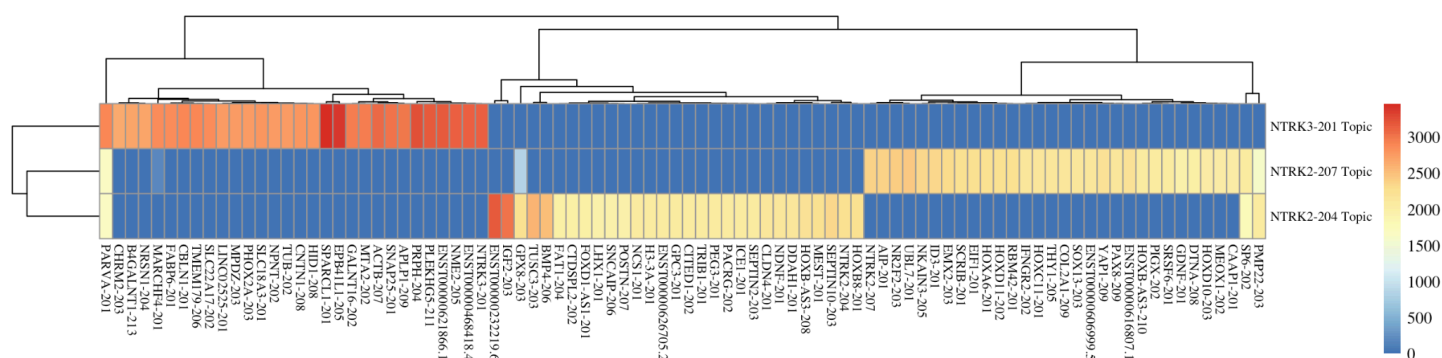


Figure 5.5. Transcript composition of the three focus NTRK topics.

Heatmap of thirty top-ranked transcripts (rows) across the three focus topics (columns). Color intensity represents topic-transcript association weights.

NTRK3-201 Topic---Sympathoadrenal Differentiation Signature

The NTRK3-201 Topic captures a transcriptional signature that may define sympathoadrenal identity, which is the developmental lineage from which neuroblastoma originates. At the core of this topic are *HAND2* and *PHOX2A*, which together with *PHOX2B* form the core regulatory circuit (CRC) driving sympathoadrenal specification¹³⁴. *PHOX2B* activates *HAND2* and *GATA* transcription factors, which in turn induce the noradrenergic gene battery encoding catecholamine biosynthetic enzymes¹³⁵. Germline *PHOX2B* mutations represent the first identified genetic predisposition for neuroblastoma, with haploinsufficiency promoting proliferation while blocking differentiation of sympathoadrenal progenitors^{136,137}. The topic also contains peripherin (PRPH), a type III intermediate filament protein that serves as a definitive marker of peripheral neuronal identity and correlates with differentiation status in neuroblastoma¹³⁸, alongside synaptic proteins including *SNAP25* that indicate progression toward functional neuronal maturity. NTRK3-201 (ENST00000317501) is within the top ten associated transcripts within this topic as well. NTRK3-201 is also embedded within the sympathoadrenal differentiation network rather than appearing in isolation. TrkC/NTRK3 expression has long been associated with favorable neuroblastoma prognosis and differentiation-competent tumor subtypes. In a comprehensive analysis of 814 primary neuroblastomas, TrkC expression correlated with favorable features including younger age and absence of MYCN amplification, while TrkA and TrkC overexpression are considered favorable prognostic biomarkers in contrast to TrkB, which marks aggressive MYCN-amplified tumors^{139,140}. The positioning of NTRK3-201 alongside *HAND2*, *PHOX2A*, and peripherin may suggest that TrkC contributes to the differentiation program itself.

The strong enrichment of this topic in neuroblastoma samples, with heterogeneous distribution across the NBL cluster (Figure 5.4A-B), likely reflects developmental heterogeneity within this tumor type. Neuroblastomas retaining robust expression of this differentiation signature may represent tumors arrested at a later developmental stage or those preserving capacity for terminal differentiation with both associated with favorable clinical behavior. This aligns with the paradigm that neuroblastoma is fundamentally a disease of arrested differentiation, where the specific point of developmental blockade determines tumor biology¹³⁴. Tumors expressing the NTRK3-201 Topic signature may therefore be more amenable to differentiation-inducing therapies such as retinoic acid, which drives neuroblastoma cells through the same developmental cascade captured by this topic⁴¹.

NTRK2-207 Topic---Nephron Progenitor Signature

The NTRK2-207 Topic reconstructs a renal developmental program anchored by *NR2F2/COUP-TFII* as it's top associated transcript. , *NR2F2/COUP-TFII* is a nuclear receptor essential for kidney development that represses neural differentiation while promoting mesenchymal fates^{141,142}. NTRK2-207 (ENST00000395882) occupies rank 2, positioned above canonical renal markers including *PAX8*, *YAP1*, *GDNF*, and extensive *HOX* gene representation (*HOXD11*, *HOXA6*, *HOXC11*, *HOXD10*, *HOXB8*). This *HOX* signature reflects anterior-posterior patterning of intermediate mesoderm during nephrogenesis, as posterior *HOX* genes (particularly *HOX11* paralogs) are essential for metanephric kidney specification^{143,144}. *PAX8* is a definitive renal lineage marker required for nephron progenitor survival¹⁴⁵ and appears alongside *YAP1*. *YAP1* is a Hippo pathway effector implicated in nephron progenitor self-renewal and Wilms tumor pathogenesis¹⁴⁶. *GDNF*, which drives ureteric bud branching morphogenesis through RET signaling¹⁴⁷, completes this nephrogenic signature.

The positioning of a full-length kinase-active NTRK2 isoform as the second associated transcript in this topic above all *HOX* genes and alongside *NR2F2* may indicate that NTRK2-207 could be a component of the renal developmental program. While NTRK2-207 is individually expressed across multiple solid tumor types (Figure 5.2A-B), the NTRK2-207 Topic enriches specifically in Wilms tumor and rhabdoid tumor (Figure 5.4C-D). This dissociation between individual transcript expression and topic enrichment demonstrates that developmental co-expression context drives disease specificity, and underscores the value of topic-level analysis over single-gene approaches.

NTRK2-204 Topic---Imprinted Renal Progenitor Signature

The NTRK2-204 Topic captures an early renal progenitor signature distinguished by enrichment for imprinted genes and cap mesenchyme markers. *IGF2*, a paternally expressed growth factor whose loss of imprinting (LOI) is among the most common molecular lesions in Wilms tumor and drives tumor cell proliferation¹⁴⁸⁻¹⁵⁰, appears prominently alongside additional imprinted loci including *MEST/PEG1* and *PEG3*. The cap mesenchyme is a self-renewing nephron progenitor population and is marked by *CITED1*¹⁵¹, which appears in this topic alongside *BMP4*, a regulator of metanephric mesenchyme induction. Glypican-3 (GPC3), which modulates Wnt and Hedgehog signaling, is associated with Wilms tumor predisposition in Simpson-Golabi-Behmel syndrome and is overexpressed in a subset of Wilms tumors¹⁵².

TrkB.T1/NTRK2-204 (ENST00000359847) is also within the top ten transcripts of this topic. Unlike full-length TrkB, this truncated isoform lacks the intracellular kinase domain yet retains signaling capacity through kinase-independent mechanisms including RhoA activation and engagement of developmental pathways such as Wnt, TGF- β , and Sonic Hedgehog^{127,132}. The co-expression of TrkB.T1 with imprinted genes (*IGF2*, *MEST*, *PEG3*), cap mesenchyme markers (*CITED1*), and Wnt modulators (*GPC3*) embeds this truncated receptor within an early renal progenitor state signature. TrkB.T1's capacity to activate Wnt signaling independently of kinase activity is biologically coherent with the Wnt-dependent nature of nephrogenesis, where canonical Wnt/ β -catenin signaling is essential for nephron progenitor self-renewal^{153,154}. This may suggest that TrkB.T1 may contribute to the developmental programs retained in Wilms tumor through mechanisms distinct from those targetable by TRK kinase inhibitors.

5.4. Discussion

This chapter demonstrates that transcript-level topic modeling reveals organizational principles of NTRK family isoform expression in pediatric cancers that are invisible to gene-level analysis. One of the central findings is that NTRK isoforms co-express according to developmental lineage rather than kinase domain status. This finding has implications for understanding neurotrophin receptor biology and for therapeutic targeting in precision oncology.

5.4.1. Transcriptional Developmental Signatures Organize NTRK Isoform Co-expression

The results challenge the conventional framework for understanding neurotrophin receptor biology, which classifies isoforms primarily by kinase domain status (active versus truncated). Under a structural/functional model, one would predict that kinase-active isoforms across the NTRK family would co-express due to shared signaling capabilities, while truncated isoforms would form a separate co-expression group.

The hierarchical clustering of NTRK transcripts across topic-transcript association matrices shows that isoforms from different genes cluster together when they share developmental lineage, while isoforms from the same gene segregate when they are embedded in different developmental programs. NTRK3-201, NTRK2-204 (TrkB.T1), and NTRK1-207 form a co-expression cluster despite encoding receptors for different neurotrophins and having different kinase domain configurations. Conversely, full-length kinase-active NTRK2-201 and truncated NTRK2-204 participate in distinct transcriptional programs, the former showing broader distribution while the latter is specifically embedded within a renal developmental signature enriched for HOX transcription factors and PAX8.

This organization parallels the Pattwell laboratory's demonstration that TrkB.T1 engages developmental signaling pathways (Wnt, TGF- β , Sonic Hedgehog) independently of kinase activity. Our finding that TrkB.T1 co-expresses with imprinted genes (*IGF2*, *MEST*, *PEG3*), cap mesenchyme markers (*CITED1*), and Wnt modulators (*GPC3*) in the NTRK2-204 Topic provides transcriptomic evidence supporting this kinase-independent functional framework. The developmental context in which an isoform is expressed may be more functionally relevant than its domain architecture, suggesting that the signaling pathways engaged by a given NTRK isoform depend on the transcriptional milieu in which it operates rather than solely on its structural properties.

5.4.2. Three Focus Topics Define Distinct Developmental Signatures

The three focus topics--NTRK3-201 Topic, NTRK2-207 Topic, and NTRK2-204 Topic--demonstrate how transcript-level topic modeling achieves disease specificity.

The NTRK3-201 Topic defines a sympathoadrenal differentiation signature. This topic comprises transcripts marking neural crest-derived sympathoadrenal identity. That NTRK3-201 ranks alongside these canonical sympathoadrenal specifiers at position 7 indicates that TrkC may contribute to this sympathoadrenal developmental program. This finding is consistent with decades of neuroblastoma biology demonstrating that TrkC expression associates with differentiation-competent tumor subtypes.

The NTRK2-207 Topic reconstructs a nephron progenitor signature. *NR2F2/COUP-TFII* at rank 1, NTRK2-207 at rank 2, followed by *PAX8*, *YAPI*, *GDNF*, and extensive HOX representation (*HOXD11*, *HOXA6*, *HOXC11*, *HOXB-AS3*, *HOXD10*, *HOXB8*). *NR2F2* is essential for kidney development and actively represses neural differentiation while promoting mesenchymal fates. The co-expression of *YAPI* is notable given *YAPI*'s established role in nephron progenitor maintenance and Wilms tumor pathogenesis. This places a kinase-active NTRK2 isoform within a transcription factor network governing renal development, which is an unexpected finding given TrkB's canonical association with nervous system biology.

The NTRK2-204 Topic captures an imprinted renal progenitor signature. This topic is characterized by *IGF2*, whose loss of imprinting is among the most common molecular lesions in Wilms tumor, alongside additional imprinted genes (*MEST/PEG1*, *PEG3*). *TrkB.T1* at rank 6, with *BMP4*, *CITED1* (a definitive cap mesenchyme marker), and *GPC3*. This creates an early renal progenitor program distinct from the NTRK2-207 Topic though they are expressed in the same tumors. The enrichment for imprinted loci suggests this topic captures a developmentally earlier renal program, consistent with the epigenetic dysregulation that characterizes Wilms tumor pathogenesis.

The disease-specificity achieved by these topics demonstrates that coordinated transcript signatures defined by topic modeling can define the molecular boundaries between pediatric cancer lineages. NTRK expression alone cannot distinguish neuroblastoma from Wilms tumor, however the developmental transcription factor networks in which NTRK isoforms are embedded provide this discrimination.

5.4.3. Methodological Advances: Transcript-Level Topic Modeling

This chapter extends the topic modeling framework established in Chapters 3 and 4 to transcript-level resolution. A significant methodological insight is that transcript-level quantification enables detection of regulatory complexity that gene-level analysis fundamentally cannot capture. Two isoforms from the same gene can participate in completely different developmental programs and associate with distinct disease types. This information is lost when isoform expression is collapsed to gene-level values. The multi-resolution approach (K=20 through K=95) provides robustness by identifying patterns stable across model granularity. The detection of 35 NTRK transcripts across 920 topics, with 154 NTRK-containing topics spanning all model resolutions, demonstrates that NTRK-associated transcriptional programs are consistent features of pediatric cancer biology rather than artifacts of specific model parameterization. The application of GSVA to convert topic-transcript associations into per-sample topic scores enables quantitative comparison of topic enrichment across diseases and samples. This approach leverages the continuous nature of topic modeling where samples are not assigned to discrete categories but rather scored along continuous axes representing developmental programs. The GSVA framework also enables integration with clinical metadata, although such analyses are beyond the scope of the current chapter.

5.5. Limitations

Several limitations warrant consideration when interpreting these results.

First, short-read RNA-seq faces inherent challenges for transcript-level quantification. Reads of 50-150 base pairs must be computationally assigned to transcripts spanning tens of thousands of bases, and reads mapping to exons shared between isoforms are ambiguous. The NTRK2 locus is particularly challenging because multiple isoforms differ only in 3' terminal exons. Salmon's expectation-maximization algorithm provides probabilistic assignments, but accuracy depends on reference annotations and may be imperfect for complex loci. Long-read sequencing validation would strengthen confidence in isoform quantification, particularly for distinguishing closely related variants.

Second, the TARGET cohort has unequal sample sizes across disease types. Neuroblastoma (n=162) and Wilms tumor (n=132) provide adequate power for robust pattern detection, but clear cell sarcoma of the kidney (n=13) and rhabdoid tumor (n=5) are represented by small numbers that limit statistical confidence. The consistent enrichment of renal topics across WT, CCSK, and RT despite these sample size differences provides some reassurance, but replication in larger renal tumor cohorts would be valuable.

Third, topic modeling identifies co-expression patterns but cannot establish causality. The observation that *TrkB.T1* co-expresses with Wnt pathway components does not demonstrate that *TrkB.T1* activates Wnt signaling in these tumors. Functional validation through isoform-specific perturbation experiments such as overexpression, knockdown, CRISPR editing of specific splice sites would be required to test whether developmental context determines isoform function or whether the associations are correlational.

Fourth, the correlation analysis is restricted to transcripts within NTRK-containing topics and may miss relevant correlations in the broader transcriptome. Transcripts that co-express with NTRK isoforms but do not rank highly in any NTRK-containing topic would not be captured by this approach. A transcriptome-wide correlation analysis might reveal additional co-expression partners, though at the cost of increased multiple testing burden.

Fifth, several biologically important genes are missing from the TARGET dataset. Most notably, *P75/NGFR* (also known as *CD271*, *TNFRSF16*), is a pan-neurotrophin receptor that modulates Trk signaling and has important roles in neuroblastoma biology. This gene is absent from the original TARGET data possibly due to processing or sequencing technicalities. This limits the ability to assess the complete neurotrophin receptor landscape and its organization across pediatric cancers as there is uncertainty on the quality of sequencing.

Sixth, Ensembl transcript nomenclature varies across versions. The same biological transcript may be designated differently depending on annotation source, creating potential for confusion when comparing results to existing literature. We have provided ENST identifiers throughout to enable unambiguous identification, but one should exercise caution when relating findings to studies using different annotation sources.

5.6. Future Directions

These findings suggest several directions for future investigation.

Long-read sequencing validation of selected TARGET samples would address uncertainties in short-read isoform quantification. Technologies such as Oxford Nanopore or PacBio can sequence full-length transcripts, enabling direct observation of isoform structure rather than computational inference. A focused study of NTRK2 transcripts in Wilms tumor samples would be particularly valuable given the complexity of this locus.

Isoform-specific perturbation experiments in Wilms tumor and neuroblastoma cell lines would test whether developmental context determines isoform function. CRISPR-based editing of specific splice sites could selectively eliminate individual isoforms while preserving others, enabling dissection of isoform-specific contributions to cell behavior. Alternatively, ectopic expression of specific isoforms in cell lines from different developmental contexts could test whether the same isoform has different effects depending on the transcriptional milieu.

Clinical outcome analysis within the neuroblastoma cohort could test whether topic scores provide prognostic information beyond what individual NTRK expression levels or existing risk stratification schemes provide. The NTRK3-201 Topic, with its sympathoadrenal differentiation signature, might identify patients with favorable biology who could benefit from treatment de-escalation.

5.7. Conclusion

This chapter completes the progression from gene-level topic modeling in mouse development (Chapter 3), to human cerebellar development and medulloblastoma (Chapter 4), to transcript-level resolution in pediatric cancers. One of the central findings here being that developmental context organizes isoform co-expression and adds a layer of regulatory complexity to the developmental framework established in earlier chapters.

In Chapter 3, topic modeling revealed that transcriptional programs in mouse organogenesis are organized hierarchically, with early topics marking broad fate decisions and later topics marking terminal differentiation. Chapter 4 demonstrated that developmental programs persist in pediatric brain tumors, with medulloblastoma subtypes retaining signatures of specific developmental stages. Chapter 5 extends this framework by showing that alternative splicing produces multiple isoforms from single genes, and these isoforms are incorporated into developmental programs according to the lineage context rather than the gene of origin. This progression reveals a principle that developmental transcription factor networks coordinate expression not only of genes and specific isoforms, creating context-dependent transcriptomes that gene-level analysis cannot resolve.

The methodological continuity across chapters demonstrates that this computational framework (transcriptomic or genomic topic modeling with GSVA transfer) scales from mouse to human, from single-cell to bulk RNA-seq, and from gene-level to transcript-level resolution. This scalability supports the broader goal of using topic modeling as a unifying framework for integrating data across developmental and disease contexts.

Transcript-level topic modeling of the TARGET pediatric cancer dataset reveals that NTRK family isoforms are organized into co-expression programs that follow developmental lineage rather than protein structural classification. The three focus topics--NTRK3-201 Topic (sympathoadrenal differentiation), NTRK2-207 Topic (nephron progenitor), and NTRK2-204 Topic (imprinted renal progenitor)--achieve disease specificity in a way that individual NTRK transcript expression does not provide. These topics highlight how NTRK1 and NTRK3 isoforms embedded in neural crest signatures and NTRK2 isoforms embedded in renal developmental signatures.

The implications extend to precision oncology where TRK inhibitor efficacy may depend not only on whether a tumor expresses NTRK genes but on which isoforms predominate and what developmental programs they participate in. Isoform-resolved expression profiling and attention to developmental transcriptional context may improve patient stratification for neurotrophin receptor-targeted therapies.

More broadly, this chapter demonstrates that alternative splicing creates a layer of regulatory complexity invisible to gene-level analysis, which is a finding with relevance beyond the NTRK family to the broader landscape of cancer biology and therapeutic development.



6. CHAPTER 6. Discussion, Conclusion, and Future Directions

This dissertation demonstrates that topic modeling provides a transferable computational framework for identifying conserved transcriptional signatures and relevant biological programs across developmental and disease contexts. The approach was validated across three biological systems of increasing complexity: mouse embryonic development with a focus on neural pathways (Chapter 3), human cerebellar development and its relation to medulloblastoma (Chapter 4), and transcript-level analysis of pediatric cancers specifically focusing on NTRK isoforms (Chapter 5).

From each chapter, there are several major findings. From chapter 3, topics derived from mouse organogenesis successfully subdivided major cell lineages identified through conventional clustering approaches while adding transitional states and partial differentiation programs invisible to discrete classification methods. Here, we not only constructed and validated the topic modeling and GSVA pipeline, but also demonstrated how additional layers of cell type identification can be uncovered across lineages. From chapter 4, topics learned from human cerebellar development transferred across independent datasets using different sequencing technologies, and these developmental topics enriched in specific medulloblastoma subtypes with associations to clinical outcomes. We demonstrated how the established topic modeling and GSVA pipeline can be applied in a multiomics approach across distinct biological contexts successfully and identify genetic signatures and biological processes that furthers our conventional pan-atlas analysis approaches. Finally, from chapter 5, transcript-level topic modeling of NTRK family genes revealed that isoforms organize according to developmental lineage rather than protein domain structure, with implications for understanding neurotrophin receptor biology and therapeutic targeting in pediatric cancers. Not only did we demonstrate that transcripts provide more in-depth information, we also showed that the correlated networks of transcripts underscore relevant biology that transcripts alone or any gene analysis miss. These results support the hypothesis that topic modeling can identify conserved transcriptional signatures that persist across biological contexts as demonstrated through the developmental-cancer boundary, enabling systematic comparison. The framework scales across technical platforms including single-cell to bulk RNA-seq, across species from mouse to human, and across levels of resolution from gene to isoform-level analysis.

This methodology extends beyond the bioinformatics and biomedical context provided here, as well. The approach can be applied to other sequencing-based studies within a bioinformatics project and can be adapted to contexts outside of sequencing. Current applications include analysis of electronic health records, insurance claims data, and other medical data types including mixed-modality datasets. The framework's broad applicability and generalizability position it as a versatile tool across multiple domains of computational biology and data science.

6.1. Challenges with Genome Versions and Annotations, and Reproducibility with Transcriptomic Analyses

Through this work we have explored the process of creating and demonstrating this generalizable analytical approach. However, two significant challenges emerged that warrant detailed discussion: the technical challenges posed by inconsistent genome annotations and transcript identifiers, and the ethical considerations surrounding the use of embryonic and fetal tissue in research. These issues have substantial implications for reproducibility, future research directions, and the sustainability of developmental biology as a field.

6.1.1. The Problem of Annotation Instability

Genomic research depends fundamentally on reference annotations such as genome builds that define chromosomal coordinates, gene models that specify exon-intron structure, and transcript databases that catalog splice variants. These annotations are treated as stable references in published literature, but in practice they are revised regularly as sequencing technologies improve and computational methods advance. This creates substantial challenges for reproducibility and cross-study comparison. A transcript identified in a 2015 publication using Ensembl release 75 may have a different identifier in 2026 publications using Ensembl release 110. The same biological isoform may be annotated differently, or may not be annotated at all in earlier releases. Gene coordinates shift between genome builds such as from GRCh37/hg19 to GRCh38/hg38, affecting variant calling, splice site prediction, and any position-dependent analysis.

6.1.2. Sources of Annotation Variation

Several factors contribute to annotation instability. Genome build updates represent one major source of variation. The transition from GRCh37/hg19 to GRCh38/hg38 involved not just coordinate updates but also corrections to assembly errors, addition of alternative haplotypes, and revisions to gene models. Genes on chromosome scaffolds that were repositioned may have entirely different coordinates. Some genes that were erroneously duplicated in GRCh37 were corrected in GRCh38. Improved transcript evidence also drives annotation changes. As RNA sequencing data accumulates, evidence for new splice variants emerges. Ensembl and RefSeq periodically update their transcript databases to incorporate new isoforms which might overlap with previous versions or split up isoforms. Conversely, some predicted transcripts supported only by computational evidence may be removed if RNA-seq data fails to confirm their expression. Annotation algorithm changes contribute as well, since the computational methods used to predict gene structures and identify transcripts are continually refined. Updates to these algorithms can result in modified gene models even when the underlying sequence data has not changed. Database-specific policies create additional variation, as Ensembl, RefSeq, and GENCODE use different criteria for transcript annotation. Ensembl tends to be more comprehensive, including computationally

predicted transcripts with limited expression evidence. RefSeq is more conservative, requiring stronger experimental support. This means a transcript present in Ensembl may be absent from RefSeq.

These annotation issues have particular relevance for transcript-level quantification from short-read RNA-seq data, as performed in Chapter 5. Tools like Salmon¹³³ and kallisto¹⁵⁵ quantify transcript expression by probabilistically assigning reads to transcripts in the reference annotation. This process is inherently limited by annotation completeness and accuracy. If a biologically important splice variant is not included in the annotation, it cannot be quantified. Reads originating from that isoform will be assigned to other annotated transcripts sharing the same exons, creating systematically biased quantification. As annotations improve and new isoforms are added, previously invisible transcripts become detectable, potentially changing biological interpretations.

6.1.3. Case Study with the NTRK2/TrkB Isoform Nomenclature

For the NTRK2 locus specifically, the complexity of alternative splicing creates challenges. Multiple isoforms differ only in 3' terminal exons or in retention of specific internal exons. Short reads mapping to shared exons cannot definitively identify which isoform they originated from. Salmon uses an expectation-maximization algorithm to probabilistically distribute such ambiguous reads, but the accuracy depends on the annotation being comprehensive and correct.

The challenges of annotation instability are well illustrated by the NTRK2 locus, which was central to the analysis in Chapter 5. The truncated TrkB.T1 isoform has been studied in neuroscience and cancer biology for over two decades. However, identifying this isoform consistently across annotation systems and genome builds is non-trivial.

In current Ensembl annotations using release 110 and GRCh38, TrkB.T1 corresponds to NTRK2-204 with the identifier ENST00000359847 and also NTRK2-207 with identifier ENST00000359847. From other releases, TrkB.T1 may refer to ENST00000690281, ENST00000689685, or ENST00000692506. From UCSC, this variant is notated as uc004aob.1 and uc011ltb.1. In earlier genome builds or annotation releases, the same biological transcript may have been assigned different ENST identifiers or may not have been annotated as a distinct isoform. Literature published over the past 20 years may use any of these identifiers, or may simply refer to "TrkB.T1" without providing database identifiers. Definitive identification requires direct sequence comparison, exon structure, protein length, etc rather than relying on isoform numbers alone. This challenge is compounded when incorporating cross-species data, as determining transcript correspondence between human and mouse NTRK2 isoforms requires sequence alignment and cannot be inferred from annotation schemes, multiplying the ambiguity inherent in identifier-based approaches.

This nomenclature drift creates practical problems. Literature synthesis becomes difficult as meta-analyses or systematic reviews must reconcile not just different naming schemes but also different versions and require sequence comparisons to determine whether studies are examining the same isoform. Method replication is challenging because a researcher attempting to validate findings from a 2015 paper may be unable to identify which transcript in current annotations corresponds to the isoform studied in that paper. Data integration requires manual curation since combining datasets processed with different annotation versions requires mapping between identifier schemes, which is not always one-to-one. Computational pipelines will have trouble when annotations get updated, because scripts that may hard-code specific ENST identifiers will fail when run with newer annotation releases if those identifiers have been deprecated or renamed.

6.1.4. Solutions Employed in this Dissertation and Recommendations

To address these challenges, this dissertation implemented several practices. All analyses explicitly document versions by specifying the genome build such as GRCh38/hg38 for TARGET data in Chapter 5 and annotation source. Where applicable, Ensembl release numbers are provided. All isoform-level findings report complete ENST identifiers with version numbers such as ENST00000359847.3 rather than truncated identifiers. Transcript identifiers are accompanied by gene symbols such as NTRK2-204/TrkB.T1 to facilitate comparison with literature using different identifier schemes. The exact FASTA files used for transcript quantification are maintained and can be reviewed to enable exact replication. Findings are validated by biological coherence such as co-expression with functionally related genes rather than relying solely on annotation-based claims.

Based on these challenges, several improvements to genomic research infrastructure would enhance reproducibility. Persistent identifiers with comprehensive version tracking would help, where transcript identifiers should include version numbers as standard practice, and databases should maintain comprehensive archives of deprecated identifiers with clear mappings to current annotations. Cross-database identifier mapping through a centralized resource providing bidirectional mappings between Ensembl, RefSeq, UCSC, and GENCODE identifiers would facilitate data integration and literature synthesis. Understandably, these solutions require extensive monetary resources and people effort and time that are some of the major hurdles at the moment, though extremely worth it. While partial solutions exist such as biomaRt and MyGene.info, they are not comprehensive and not always maintained. Mandatory reporting standards should be implemented where journals require that genomic studies report genome build, annotation database and version, and quantification tool versions with the same rigor applied to sample sizes and statistical methods, with supplementary materials including or linking to the exact reference files used. Long-read validation of short-read findings should become standard, as long-read sequencing technologies like Oxford Nanopore and PacBio directly observe full-length transcripts rather than inferring them from short reads. As costs decrease, long-read validation should become standard for transcript-level claims, particularly for complex loci with extensive alternative splicing. Annotation-agnostic analysis frameworks would also help, as computational methods that focus on

co-expression patterns or other relational properties rather than individual gene identities are more robust to annotation updates, with topic modeling as employed in this dissertation exemplifying this approach since topics are defined by expression relationships, and the specific genes contributing to a topic can be updated as annotations improve without invalidating the topic structure.

6.1.5. Long-term Perspective

Annotation instability is not a solvable problem. It is an inherent feature of biological research. The dynamic nature of our changing annotation systems make understanding our complex biological systems challenging. As technology advances, our ability to detect rare transcripts, cell-type-specific isoforms, and developmental stage-specific variants improves. Annotations will continue to be refined. Beyond the practical challenges of changing identifiers, there is a deeper issue. We are making foundational assumptions that the sequence from one human reference genome will be the same in the next human and the next, that transcripts will be consistently structured across individuals, that the gene boundaries we define based on averaged or consensus sequences that may have vastly different expressions accurately represent biological reality. Each layer of annotation builds on these assumptions, and the averaged transcripts we quantify are themselves abstractions of variable biological entities. This realization leads to existentialistic thinking about the field of genomic science and heavily questions whether we as a field are measuring something real or are we measuring our own constructed categories. The appropriate response is not to lament this instability, which I and Dr. Pattwell have done throughout this process, but to design research practices that accommodate it. This includes prioritizing biological validation over annotation-based claims, maintaining detailed and standardized version documentation, focusing on conserved patterns rather than individual identifiers, and contributing to community efforts to improve annotation databases and cross-reference tools.

The dissertation demonstrates that robust biological findings can be obtained despite annotation challenges. The developmental topics identified in Chapter 4 transfer across datasets even though they were generated using different genome builds. The isoform-level organization principles in Chapter 5 remain valid even if specific transcript identifiers change in future releases. This robustness derives from focusing on co-expression relationships and biological coherence rather than treating annotations as immutable truth.

6.2. Ethical Considerations of Embryonic Tissue Research

This dissertation used only previously published, de-identified datasets and did not involve direct tissue collection or patient interaction. However, regulatory compliance does not exhaust ethical responsibilities.

Outside of the scope of this dissertation, I was taught that scientific research ethics be grounded in principles of beneficence, respect for persons, and justice. While values necessarily inform ethical frameworks, conflating political positions with ethical requirements undermines both scientific integrity and ethical reasoning. I find it important to advocate for evidence-based policy while maintaining scientific objectivity. Restrictions on embryonic and fetal tissue research have consequences for public health and our understanding of developmental disorders, cancer biology, and infectious disease may be impeded. Therapeutic development timelines may be extended. In these processes, I have learned that the burden of these consequences falls disproportionately on patients with diseases that require developmental tissue for research progress. Divergent policies create challenges for collaboration in general. Researchers in restrictive jurisdictions may be unable to participate in studies using fetal tissue, limiting scientific exchange and potentially concentrating expertise in more permissive regions. Therefore, I do find it necessary to explore these ethics briefly.

6.2.1. Scientific Necessity of Developmental Tissue

Human embryonic and fetal tissue represents an irreplaceable resource for biomedical research, providing unique insights into human development that currently cannot be obtained through alternative means. The scientific necessity of this tissue extends across multiple domains of inquiry, from fundamental developmental biology to translational medicine and therapeutic development.

The research presented in this dissertation relied on human fetal cerebellar tissue for topic modeling and developmental cell type genetic signature identification. Datasets from Aldinger et al. (2021) and Cao et al. (2020) were essential for identifying conserved biological programs and genetic signatures that are unique to humans and also have connections to early pediatric brain cancers. Without access to such tissue, characterization of these human-specific findings, including those that are taken advantage of in cancer, would be impossible.

Mouse models, while essential for mechanistic studies, have several limitations for human research. In relation to this dissertation alone, the mouse cerebellum develops over approximately three weeks compared to more than two years in humans. The rhombic lip is in fact a human specific anatomical feature that doesn't exist in the mouse. Many genes and cell types are human-specific or show different expression patterns between species. Disease models (such as xenografts or animal models) often fail to recapitulate human pathology, as demonstrated by challenges in modeling Group 3 and Group 4 medulloblastoma in mice. Organoid and induced pluripotent stem cell (iPSC) models offer promising alternatives but remain limited in their ability to fully recapitulate in vivo developmental programs, particularly for later stages of cerebellar development including external granule layer formation and granule neuron migration. Understanding pediatric brain cancer developmental origins requires direct analysis of human developmental tissues. The identification of tumor-specific developmental signatures and potential therapeutic targets depends on accurate characterization of normal human development. Treatments developed without understanding human-specific biology may fail in clinical translation.

6.2.2. Global Disparities in Tissue Availability

Sex-selective abortion practices in certain regions create disparities in the sex distribution of available fetal tissues. In some contexts, these practices result in overrepresentation of female fetal tissues available for research, while in other contexts male samples may predominate. Though these datasets are mostly confined to countries outside of the US, their results and publications are still in the scientific field. This creates potential biases in developmental research, as male and female brains show subtle but significant developmental differences^{156,157}, and sex as a biological variable must be considered in study design and interpretation. Sex-selective practices vary widely across regions, influenced by cultural, economic, and policy factors. These disparities affect not only tissue availability but also the generalizability of research findings. Biased sample availability may influence research conclusions if sex-specific developmental programs are differentially represented in available tissues. Researchers must report sample sex ratios and consider potential biases in interpretation, and efforts to balance sample collection across sexes should be prioritized.

Religious and cultural beliefs significantly impact willingness to donate fetal tissue for research. Some traditions prohibit any use of fetal remains, while others permit research use under specific conditions. These differences create geographic disparities in tissue availability and influence the composition of biobanks and research cohorts. Cross-border sharing of fetal tissue faces legal, ethical, and logistical challenges as well. Consent requirements, data protection regulations, and material transfer agreements vary by jurisdiction. Researchers must navigate complex regulatory landscapes to enable international collaboration while respecting local ethical frameworks. This variation complicates multi-site studies and may influence where research is conducted.

Restrictions on abortion procedures may reduce tissue availability, alter the gestational age distribution of available samples, or impose conditions that complicate research use. Conversely, policies supporting reproductive autonomy may facilitate tissue procurement within ethical guidelines. Reduced tissue availability impacts research in developmental biology, disease modeling, vaccine development, and regenerative medicine. The effects extend beyond immediate research communities to influence public health outcomes and therapeutic development pipelines.

6.2.3. Acknowledgment and Responsibility

The data used in this dissertation exist because individuals and families chose to donate tissue for research, clinicians and researchers invested effort in tissue collection and processing, and funding agencies supported biobanking infrastructure. None of this was inevitable. The tissue samples represent tumors removed from children with cancer, developmental tissue from pregnancies that did not continue, and cells that once were part of living individuals. This origin imposes responsibilities on researchers using the resulting data. First, analytical rigor is essential, as sloppy analysis or poor statistical methods dishonor the contribution of tissue donors, and the data should be analyzed with appropriate care and expertise. Honest interpretation requires that results should be reported accurately, with limitations acknowledged, since overstated conclusions or misleading presentations betray the trust of those who donated tissue in hope of advancing knowledge. Translational intent is important because while not every study directly impacts clinical care, research using human tissue should have plausible pathways to improving understanding of disease or informing future therapeutic approaches. Open science practices should be implemented where possible, as analytical methods and code should be shared to maximize the value extracted from limited tissue resources, and this dissertation provides code repositories and detailed methods to facilitate replication and extension of the work. These implementation steps can be applied to all animal and human based research with proper compensation, acknowledgement, and respect.

This dissertation has attempted to honor these responsibilities through rigorous methods, transparent reporting of limitations, focus on questions with clinical relevance, and contribution to open computational frameworks.

6.2.4. Current Political and Regulatory Context

Access to embryonic and fetal tissue for research is subject to political and regulatory variation across jurisdictions and over time. In the United States, the regulatory environment has been unstable, with periods of expanded federal funding for fetal tissue research alternating with periods of restriction. Recent political developments have created additional uncertainty. The 2022 Dobbs v. Jackson decision overturned federal protection for abortion access, and subsequent state-level legislation has varied widely. Some states have enacted restrictions not only on abortion but also on the use of tissue from elective terminations in research. Federal funding policies have shifted with changes in presidential administrations. Internationally, regulations vary substantially. Some countries maintain permissive frameworks for embryonic and fetal tissue research under strict ethical oversight. Others have restrictive policies based on religious, cultural, or philosophical positions regarding the moral status of embryos and fetuses. This heterogeneity creates practical challenges for research continuity. Datasets that are currently accessible may not remain so indefinitely. Tissue procurement pipelines may be disrupted by regulatory changes. Collaborative research across jurisdictions must navigate differing legal frameworks.

6.2.5. Implications for Future Research in Embryonic Tissue Research

The political and regulatory instability surrounding embryonic tissue research has several implications. Current datasets become more valuable because comprehensive atlases of human development should be priorities for the field, as future tissue access may be restricted, and thorough characterization and open sharing of existing datasets maximizes their utility. Alternative approaches warrant investment, and development of stem cell organoid systems, computational prediction methods, and other approaches that reduce

dependence on primary tissue should be pursued, while recognizing their limitations. Public engagement is necessary because the scientific community has a responsibility to communicate clearly about why embryonic tissue research is important, what protections exist, and what would be lost if access were curtailed, though this communication must respect diverse views on the moral status of embryonic life rather than dismissing ethical concerns as anti-scientific. International collaboration becomes crucial since if access to tissue becomes restricted in some jurisdictions, collaborative frameworks that enable researchers in restrictive locations to access datasets generated elsewhere become important.

6.2.6. Ethical Nuance and Uncertainty in Embryonic Tissue Research

The ethical issues surrounding embryonic tissue research do not have simple resolutions since conflicts continue about the moral status of embryos and fetuses, about when personhood begins, and about what uses of tissue are ethically permissible. These disagreements are grounded in differing philosophical frameworks, religious beliefs, and cultural values. Science can inform these discussions by clarifying questions such as when different developmental processes begin and what organoid models can replicate and what they miss, but cannot resolve fundamental value disagreements. The claim that embryonic tissue research is essential for advancing human health is factually supportable but does not automatically override concerns about the moral status of embryonic life.

This scientific dissertation takes no position on when personhood begins or what moral status should be accorded to embryos at different developmental stages. These are questions that science is not equipped to answer. Though, personally, it begs larger questions of how we can consider an embryo a person when people post natally are not given personhood. Perhaps an alternative would be to focus on improving the quality of living for people we have now (perhaps just ensuring they have a chance to live) before we begin making claims about the unborn. In terms of science, what can be stated is that understanding human development requires human developmental tissue for certain questions that cannot be addressed with animal models or alternative approaches. Developmental neuroscience has contributed to understanding of congenital disorders, pediatric cancers, and normal brain development in ways that have benefited human health. Research using embryonic tissue is conducted under ethical oversight with informed consent and privacy protections. Alternative approaches should be developed but currently cannot fully replace primary tissue for many research questions. Access to tissue is not guaranteed in perpetuity and may continue to become more restricted.

6.2.7. Moving Forward

This dissertation used embryonic tissue-derived data while working within established ethical frameworks and regulatory structures. Future work in this area should maximize the value of existing datasets through rigorous analysis and open sharing and also develop complementary approaches such as organoids and computational models where feasible. Additionally, it would be important to engage transparently with ethical oversight processes, and communicate clearly about the scientific importance of developmental research and the protections in place. In addition, the field should actively prepare for possible continued restrictions on tissue access, and respect the diversity of views on embryonic tissue research while advocating for policies that enable important science under appropriate ethical oversight. The ethical questions are not resolved by this work, nor can they be. However, we carried out this work with awareness of the ethical considerations, with gratitude to tissue donors, and with commitment to honoring their contribution through rigorous science.

6.3. Future Directions

The findings and challenges identified in this dissertation suggest several directions for future research. Here we outline three broad categories of future work that tie into previous discussions from Chapter 3, 4, and 5.

6.3.1. Methodological Advances

Long-read sequencing validation would address assumptions and uncertainties especially with isoform analyses presented in this dissertation. The transcript-level findings in Chapter 5 relied on short-read RNA-seq and probabilistic isoform quantification. Long-read sequencing of selected TARGET samples would provide direct observation of full-length transcripts, validating isoform assignments and potentially identifying additional splice variants not captured in current annotations. Similarly, using transcript based long read sequencing for the mouse and human embryogenesis and cancer analyses presented in chapters 3 and 4 would improve the specificity and detail of the analysis. This would allow us to target specific transcript networks that associate with cell types that may relate to cancer persistence---creating more specific avenues for therapeutic explorations.

Integration with spatial transcriptomics represents another important direction. Combining topic modeling with spatial transcriptomics would reveal how transcriptional programs organize in three-dimensional tissue architecture. This could identify whether topic states correspond to spatial niches and how cellular neighborhoods influence topic expression. In chapter 3, we briefly explored this avenue through the use of the Allen Institute's in situ hybridization data. Further explorations with human embryogenesis data is warranted. This is an approach we will be exploring specifically for topics mentioned in Chapter 4 with collaborators at the gene level.

Single-cell isoform quantification would extend the approach further. Current single-cell RNA-seq protocols provide limited information about alternative splicing due to 3' bias in many platforms. Emerging full-length single-cell methods would enable topic modeling at isoform resolution in single cells, potentially revealing cell-type-specific splicing programs.

6.3.2. Informatics Advances

Several key advances would benefit the omics topic modeling and GSVA pipeline. First, incorporation of prior knowledge could assist in identifying or guiding topics towards specific cell types or biological processes of interest (semi-supervised topic modeling). Andrzejewski, Zhu, and Craven¹⁵⁸ developed methods for incorporating domain knowledge into topic modeling through Dirichlet Forest priors. These methods allow specification of must-link constraints (genes that should be grouped together) and cannot-link constraints (genes that should be separated) based on prior biological knowledge. For genomic applications, such constraints could encode known pathway memberships or regulatory relationships.

Additional extensions to the pipeline could incorporate hierarchical¹⁵⁹ and dynamic topic^{160–162} models. Hierarchical topic models and dynamic topic models have been developed for text analysis and could be adapted for genomic data. Hierarchical models learn topics at multiple levels of granularity, potentially corresponding to broad cell type programs and more specific cell state programs. Dynamic topic models allow topics to evolve over time, which could model developmental trajectories or disease progression.

Finally, additional methods could include improved methods for specifying the number of topics and other parameters based on an appropriate level of complexity from the data rather than a priori. Existing methods such as

6.3.3. Biological Validation

Functional perturbation experiments would test causality of the identified patterns. The topics identified represent co-expression patterns, but causality has not been established. Perturbing key transcription factors enriched in specific topics such as COUP-TFII in the NTRK2-207 renal topic could test whether topic states are functionally coordinated or merely correlational. Isoform-specific manipulation would address questions about splicing function. CRISPR-based editing of splice sites could selectively eliminate specific NTRK isoforms while preserving others, enabling dissection of isoform-specific functions. This is particularly relevant for testing whether TrkB.T1 function depends on the developmental context in which it is expressed. Developmental trajectory mapping would reveal temporal dynamics of topic expression. Time-series single-cell analysis of cerebellar development could map how topic expression changes along developmental trajectories, potentially identifying the transcription factors that drive topic transitions. Further in situ or other related assays could be applied to validate the top genes of topics found in chapter 3 and 4.

6.3.4. Broader Applications

Extension to other cancer types could test the generalizability of the framework. The approach could be applied to other cancers with developmental origins, such as additional neuroblastoma subtypes, rhabdomyosarcoma, or leukemias, to test whether developmental topic enrichment represents a general principle in pediatric oncology. Adult cancer applications represent another potential direction, as some adult cancers show evidence of dedifferentiation or reactivation of developmental programs. Topic modeling could identify which developmental signatures are reactivated and whether they correlate with therapeutic vulnerability. Cross-species developmental biology could benefit from topic transferability, as topic models trained on well-characterized mouse development could transfer to less-studied species, potentially enabling comparative developmental biology at scale. Beyond developmental biology and oncology, this approach has broad applicability to diverse sequencing-based studies, including plant genome biology, metagenomics of soil or bacterial communities, and other biological systems where identification of conserved expression programs is relevant.

6.4. Contributions

This dissertation makes several contributions to the field. From a methodological perspective, it demonstrates that topic modeling with GSVA-based transfer enables cross-dataset comparison applicable across technologies, species, and biological contexts. For developmental neuroscience, it identifies cell type and transitional signatures in cerebellar development, showing that topics supplement conventional clustering by revealing partial programs and transitional states. In pediatric oncology, it shows that medulloblastoma subtypes retain specific developmental signatures that associate with clinical outcomes, supporting the developmental origins of cancer framework. For molecular oncology, it reveals that NTRK isoforms organize according to developmental lineage rather than structural features, with implications for understanding receptor function and therapeutic targeting. Regarding reproducibility, it provides detailed documentation of analytical approaches, code repositories, and discussion of annotation challenges to facilitate replication and extension.

6.5. Limitations

Several limitations apply to the dissertation as a whole and warrant consideration. First, the use of topic modeling is fundamentally correlational rather than causal. Topic modeling, like other unsupervised matrix-decomposition approaches, identifies coordinated patterns of expression but cannot establish whether the genes or transcripts that co-occur within a topic are mechanistically linked, share an upstream regulator, or simply respond in parallel to a common cellular state. This caveat applies equally to the developmental

trajectories described in Chapter 3, the medulloblastoma-cerebellar correspondences in Chapter 4, and the isoform co-expression structures in Chapter 5. Causal claims about any of these programs would require experimental perturbation such as genetic, pharmacological, or otherwise that lies outside the scope of a purely computational thesis.

Second, every analysis presented here relies on retrospective, publicly archived data generated by other groups under conditions that cannot be revisited. Sample selection, dissociation protocols, library preparation, sequencing depth, and upstream preprocessing decisions were all fixed before this work began, and any biases embedded in those choices propagate into the topics learned downstream. Cross-dataset transfer mitigates but does not eliminate this concern, because shared biases (for example, the over-representation of certain anatomical regions or developmental time points in published atlases) can be reinforced rather than revealed when datasets are compared. Prospective validation in independently collected cohorts remains necessary before any of these findings can be considered actionable in a translational setting.

Third, the dissertation is dependent on the annotation ecosystem in which it was produced. Genome builds, transcript catalogues, and the cell-type labels inherited from the original atlas publications all function as fixed reference frames in this work, yet each is itself a moving target, as discussed at length in Section 6.2. Topics defined over Ensembl transcripts will shift as those transcripts are added, deprecated, or re-numbered. Topic-trajectory associations rest on cell identities assigned by the original authors using their own clustering choices. Any cross-dataset comparison inherits the union of these upstream decisions. Related to this is a broader methodological caveat in that topic modeling itself involves user-selected hyperparameters, such as the number of topics. Although the selection procedures used here are principled, alternative choices would yield related but not identical decompositions. Together, these limitations mean that the specific topics reported in this dissertation should be read as one coherent and reproducible view of the underlying biology rather than as a unique or final partition of it.

6.6. Conclusion

This dissertation demonstrates that topic modeling provides a transferable computational framework for identifying conserved transcriptional programs across developmental and disease contexts. The approach successfully scaled from mouse to human, from single-cell to bulk sequencing, and from gene-level to isoform-level resolution.

Here, we presented several major contributions. Topics subdivide major developmental lineages and identify transitional states that are invisible to discrete clustering approaches. Developmental topics transfer across datasets and persist in pediatric brain cancers, enabling cross-context comparison. Isoform organization follows developmental lineage rather than structural classification, revealing that alternative splicing creates context-dependent expression programs. Developmental topic enrichment correlates with clinical outcomes in medulloblastoma, suggesting potential prognostic applications. Significant challenges were identified, particularly regarding genome annotation instability and the political and regulatory uncertainty surrounding embryonic tissue research. These challenges have implications for reproducibility and future research sustainability. Future directions include methodological improvements such as long-read validation and spatial integration, biological validation through functional perturbation and isoform manipulation, clinical translation via prospective validation and biomarker development, and broader applications to other cancer types.

The practical implications extend to multiple domains. For developmental neuroscience, topic modeling supplements clustering by identifying transitional and differentiation states across cells during development. For pediatric neuro-oncology, developmental topic enrichment may enable more precise patient stratification than histology or genomics alone. For precision oncology, isoform-level resolution reveals context-dependent expression patterns relevant to therapeutic targeting. From an informatics perspective, this topic modeling pipeline addresses several major challenges in the sequencing and big data fields. The first is the *discrete classification* problem, where in large datasets, samples, cells, patients, or other entities are assigned hard and fast identities, whereas topic modeling enables definition of identities with more granularity without resorting to black box approaches. The second is the *cross-dataset comparison* problem, where integration of multiple datasets can incorporate extensive computational hardware and support that is often inaccessible and, moreover, results in data loss. The third is the *annotation inconsistency* problem, where cells, tumors, or patients are assigned labels that vary in definition across datasets.

The work contributes to understanding how developmental programs organize at multiple scales of resolution and how these programs persist in disease states. The findings support continued development of topic modeling as an analytical framework for integrating data across biological contexts and suggest that developmental context is a fundamental determinant of both normal and pathological gene expression programs.

Most importantly, this dissertation has provided the foundation for the development of a well-rounded, critical thinking, and passionate scientist.



7. APPENDIX

7.1. Chapter 3 Supplemental Topic Groups

In this supplementary section, we explore the mouse topics derived from Chapter 3 exploring all other topic groups based on all defined cellular lineages.

Group 1: Hematopoietic Lineages (Topics 37, 42, 3, 19, 38, 49)

Hematopoiesis generates all blood cell types from hematopoietic stem cells (HSCs) through a hierarchical branching process^{163,164}. HSCs reside in the bone marrow and possess both self-renewal capacity and multipotency. They give rise to multipotent progenitors (MPPs), which lose self-renewal but retain multilineage potential. MPPs bifurcate into two major differentiation branches including the common myeloid progenitor (CMP) and the common lymphoid progenitor (CLP). CMPs generate red blood cells, platelets, and innate immune cells. CLPs produce B and T lymphocytes of the adaptive immune system^{163,164}.

Branching Structure.

The CMP further divides into the megakaryocyte-erythroid progenitor (MEP) and granulocyte-monocyte progenitor (GMP). The MEP gives rise to both erythrocytes (red blood cells that carry oxygen via hemoglobin) and megakaryocytes (large polyploid cells that fragment to produce platelets for blood clotting). These two lineages are siblings, sharing a recent common progenitor. The GMP produces granulocytes (neutrophils, eosinophils, basophils) and monocytes (which differentiate into macrophages in tissues). Mast cells, which mediate allergic responses and contain histamine-rich granules, arise from a myeloid progenitor but complete their maturation in peripheral tissues^{163,164}.

Erythroid Branch. (Topics 37 → 42 → 3). This branch represents a differentiation sequence within the MEP-derived erythroid lineage. Topic 37 marks early erythroid progenitors at the colony forming unit erythroid (CFU-E) stage, expression *Trak2*, *Wdr26*, and *Gapvd1*. Topic 42 captures maturing erythroblasts undergoing hemoglobin accumulation and enucleation preparation, with expression of *Hemgn* (hemogen), *Rhd* (Rh blood group antigen), and *Slc4a1* (band 3 anion exchanger), which mediates CO₂/bicarbonate exchange in mature red cells. Topic 3 represents terminal erythrocytes expressing *Spta1* and *Sptb* (alpha and beta spectrin), structural proteins forming the membrane cytoskeleton that gives red blood cells their characteristic biconcave shape and deformability^{77,165,166}.

Megakaryocyte/Mast Branch (Topic 19). Topic 19 captures the MEP-derived megakaryocyte lineage along with mast cells, which share dependence on *Kit* (stem cell factor receptor) signaling. Megakaryocytes are unique cells that undergo endomitosis (DNA replication without cell division) to become highly polyploid, then extend cytoplasmic protrusions called proplatelets that fragment into thousands of platelets¹⁶⁷. This topic shows enrichment in megakaryocyte progenitors and mast cells.

Granulocyte Branch (Topic 38). Arising from the GMP (a sibling lineage to MEP), Topic 38 defines neutrophils and other granulocytes through expression of *S100a8* and *S100a9*¹⁶⁸. Neutrophils are the most abundant white blood cells and serve as first responders to bacterial infection through phagocytosis and release of antimicrobial granule contents^{168,169}.

Lymphoid Branch (Topic 49). Representing the earliest branch point from MPPs, the CLP gives rise to both B lymphocytes (which produce antibodies) and T lymphocytes (which coordinate adaptive immunity and kill infected cells). Topic 49 captures shared lymphoid identity through expression of *Ptpnc* (CD45, the common leukocyte antigen and a tyrosine phosphatase essential for antigen receptor signaling), *Bach2* (a transcription factor required for lymphocyte development and memory formation), *Lef1* (lymphoid enhancer factor 1, essential for T cell development), and *St8sia4* (a sialyltransferase involved in polysialic acid synthesis important for lymphocyte trafficking)⁷⁷. This topic shows equal enrichment in B cells and T cells, consistent with its capture of shared lymphoid identity upstream of B and T cells lineage commitment¹⁷⁰.

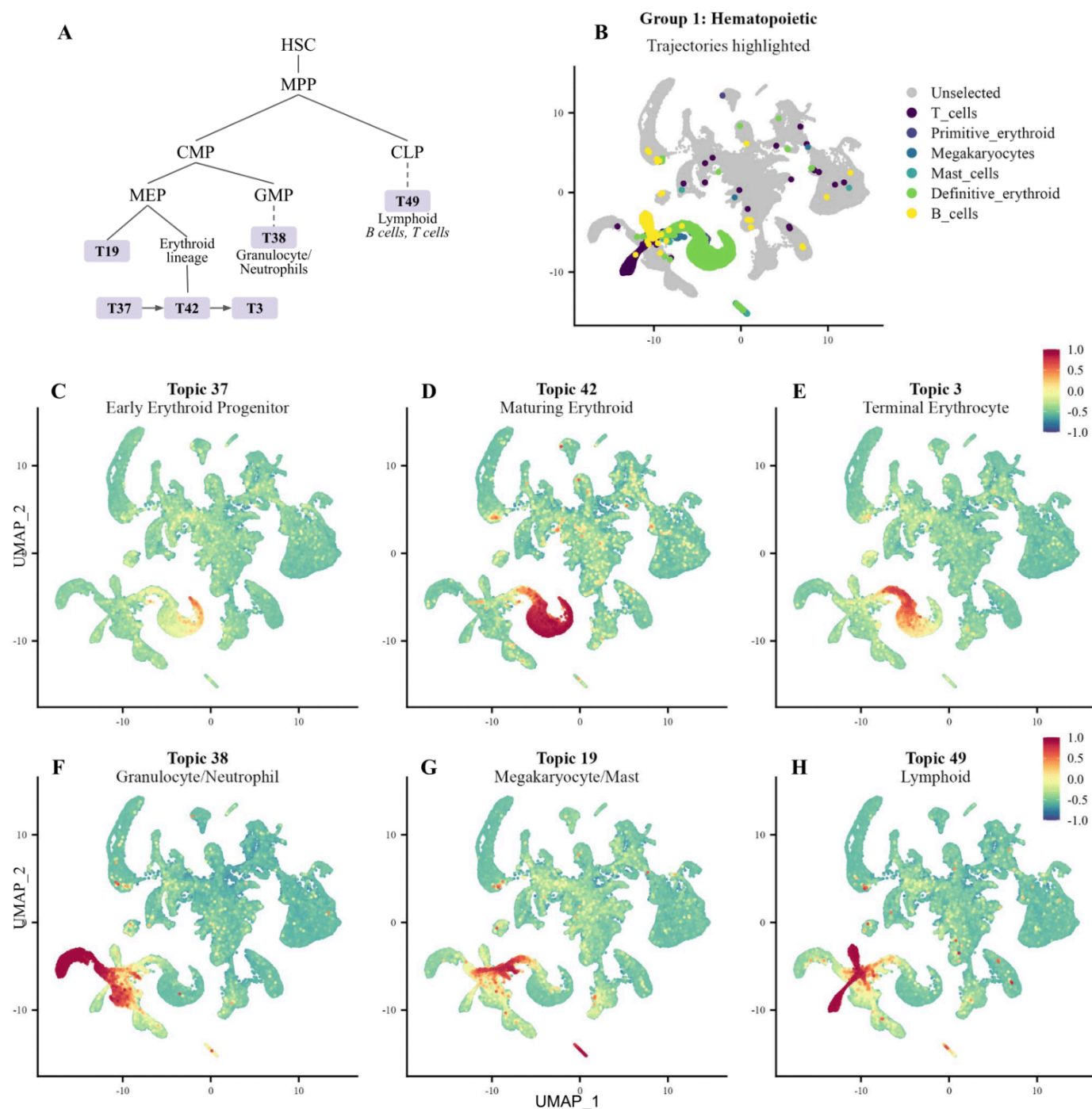


Figure S1. Group 1: Hematopoietic lineage topics capture the full hierarchy of blood cell development.

(A) Schematic of hematopoietic differentiation from hematopoietic stem cells (HSC) through multipotent progenitors (MPP) to myeloid (CMP) and lymphoid (CLP) branches, with topic assignments indicated.

(B) UMAP visualization highlighting hematopoietic trajectories including T cells, primitive erythroid, megakaryocytes, mast cells, definitive erythroid, and B cells.

(C-H) UMAPs displaying topic weights across the single-cell embedding for Topics 37 (Early Erythroid Progenitor), 42 (Maturing Erythroid), 3 (Terminal Erythrocyte), 38 (Granulocyte/Neutrophil), 19 (Megakaryocyte/Mast), and 49 (Lymphoid). Color scale indicates scaled topic weight from -1.0 (blue) to 1.0 (red).

Group 2: Tissue Myeloid Cells (Topics 5, 20, 18)

Tissue-resident macrophages are phagocytic immune cells that maintain tissue homeostasis by clearing dead cells (efferocytosis), pathogens, and debris. Unlike circulating monocytes, many tissue macrophages derive from embryonic progenitors (yolk sac and fetal liver) that seed tissues before birth and self-maintain through local proliferation. Different tissues harbor specialized macrophage

populations such as microglia in the brain, Kupffer cells in the liver, alveolar macrophages in the lung, and osteoclasts in bone. These cells exhibit plasticity, adopting different activation states in response to environmental signals¹⁷¹.

Functional States

Unlike the hematopoietic group, the myeloid topics represent functional states rather than sequential differentiation stages. Topic 5 captures the core tissue macrophage identity with expression of *Dock10*, *Inpp5d* (SHIP1, a phosphatase that negatively regulates immune signaling), and *Ptpnrc*. Topic 20 represents homeostatic macrophages engaged in tissue maintenance, expressing *Mertk* (a receptor tyrosine kinase essential for efferocytosis which is the recognition and engulfment of apoptotic cells), *Abca1* (ATP-binding cassette transporter A1, which mediates cholesterol efflux to prevent foam cell formation), and *Zeb2* (a transcription factor that maintains tissue macrophage identity). Topic 18 captures interferon-activated myeloid cells expressing *Stat1* and *Samhd1*, genes induced by type I interferons during viral infection or inflammatory responses^{172,173}.

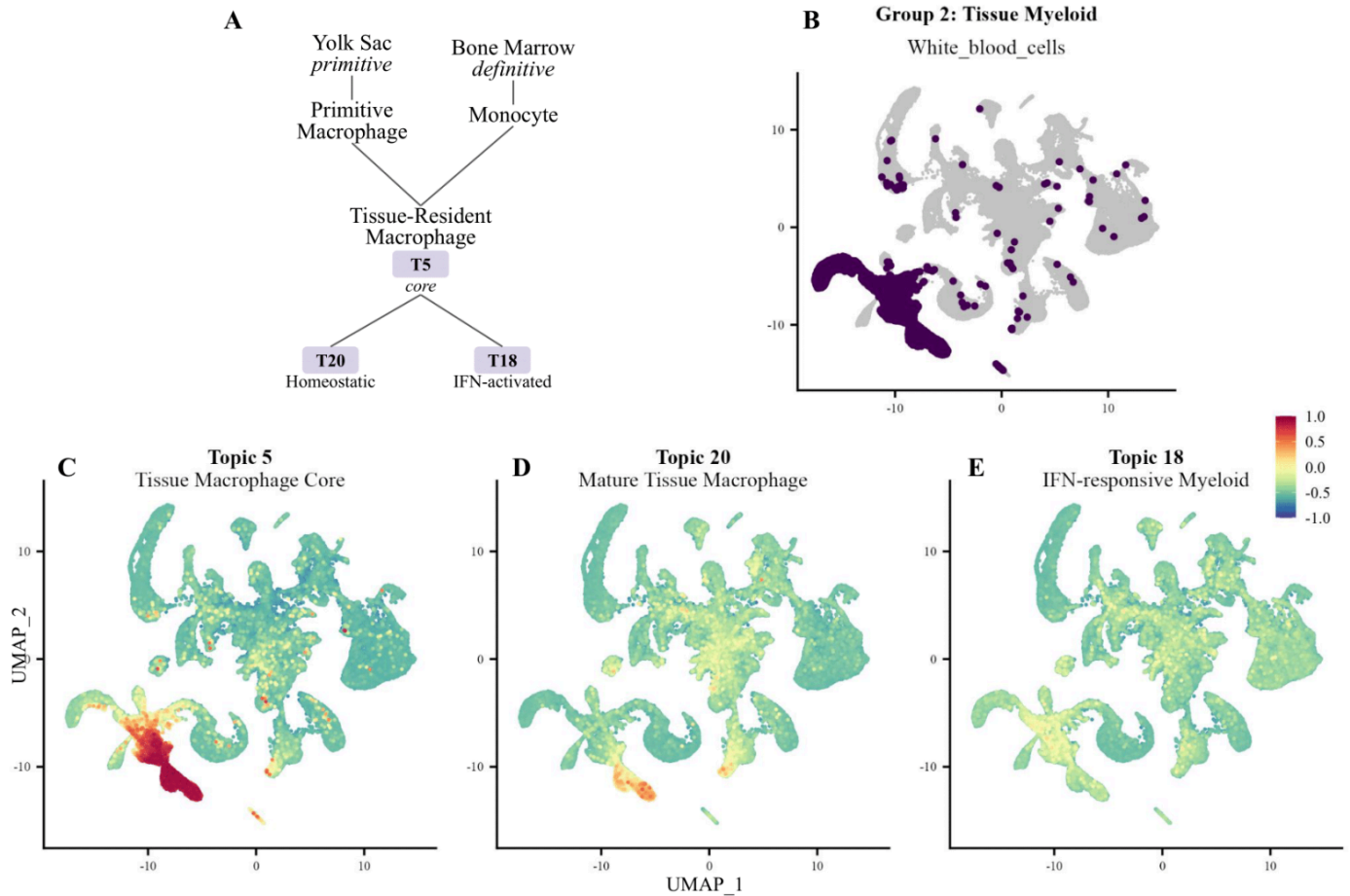


Figure S2. Group 2: Tissue myeloid topics represent functional states of tissue-resident macrophages.

(A) Schematic illustrating dual developmental origins of tissue-resident macrophages from yolk sac (primitive) and bone marrow (definitive/monocyte-derived) precursors, converging on shared tissue macrophage identity.

(B) UMAP visualization highlighting white blood cell trajectories.

(C-E) UMAPs displaying topic weights for Topics 5 (Tissue Macrophage Core), 20 (Mature Tissue Macrophage), and 18 (IFN-responsive Myeloid).

Group 3: Myogenic Lineage (Topics 9, 40, 27, 24)

Muscle development produces three distinct muscle types from mesodermal precursors: skeletal muscle (voluntary movement), cardiac muscle (heart contraction), and smooth muscle (involuntary visceral movement). Skeletal and cardiac muscle share the property of striation (organized sarcomeres containing myosin and actin filaments), but arise from different mesodermal populations and utilize distinct transcriptional programs¹⁷⁴.

Branching Structure.

Skeletal Muscle Branch (*Topics 9 → 40 → 27*) Skeletal muscle derives from paraxial mesoderm, which segments into somites flanking the neural tube. The dorsolateral somite (dermomyotome) gives rise to both dermis and skeletal muscle. Myogenic progenitors express *Pax3* and *Pax7*, then activate myogenic regulatory factors (*MyoD*, *Myf5*, *Myogenin*) to become committed myoblasts. Myoblasts are mononucleated proliferative cells that subsequently fuse to form multinucleated myotubes, which mature into contractile myofibers^{175,176}.

Topics 9 and 40 represent overlapping myoblast populations at early and late stages, respectively, expressing *Nes* (nestin, an intermediate filament marking muscle progenitors), *Pdgfc*, and *Plxna2*. Topic 27 defines mature skeletal muscle fibers through sarcomeric protein expression: *Myl1* (myosin light chain 1, specific to fast-twitch fibers), *Des* (desmin, the muscle-specific intermediate filament that links sarcomeres to the sarcolemma), *Tin* (titin, the giant protein spanning half the sarcomere), and *Actn2* (alpha-actinin 2)^{77,175}.

Cardiac Muscle Branch (*Topic 24*) Cardiac muscle derives from lateral plate mesoderm through the first and second heart fields which are distinct progenitor populations that contribute to different heart regions¹⁷⁷. Topic 24 captures cardiac progenitors and cardiomyocytes through expression of *Gucyl1a3* and *Gucyl1b3* (soluble guanylyl cyclase subunits mediating nitric oxide signaling), *Cacnb2* and *Cacna1c* (L-type calcium channel subunits essential for cardiac action potentials and excitation-contraction coupling), and *Ryr2* (ryanodine receptor 2, the major sarcoplasmic reticulum calcium release channel in cardiomyocytes)¹⁷⁷. This topic bridges differentiation from cardiac progenitors to mature cardiomyocytes.

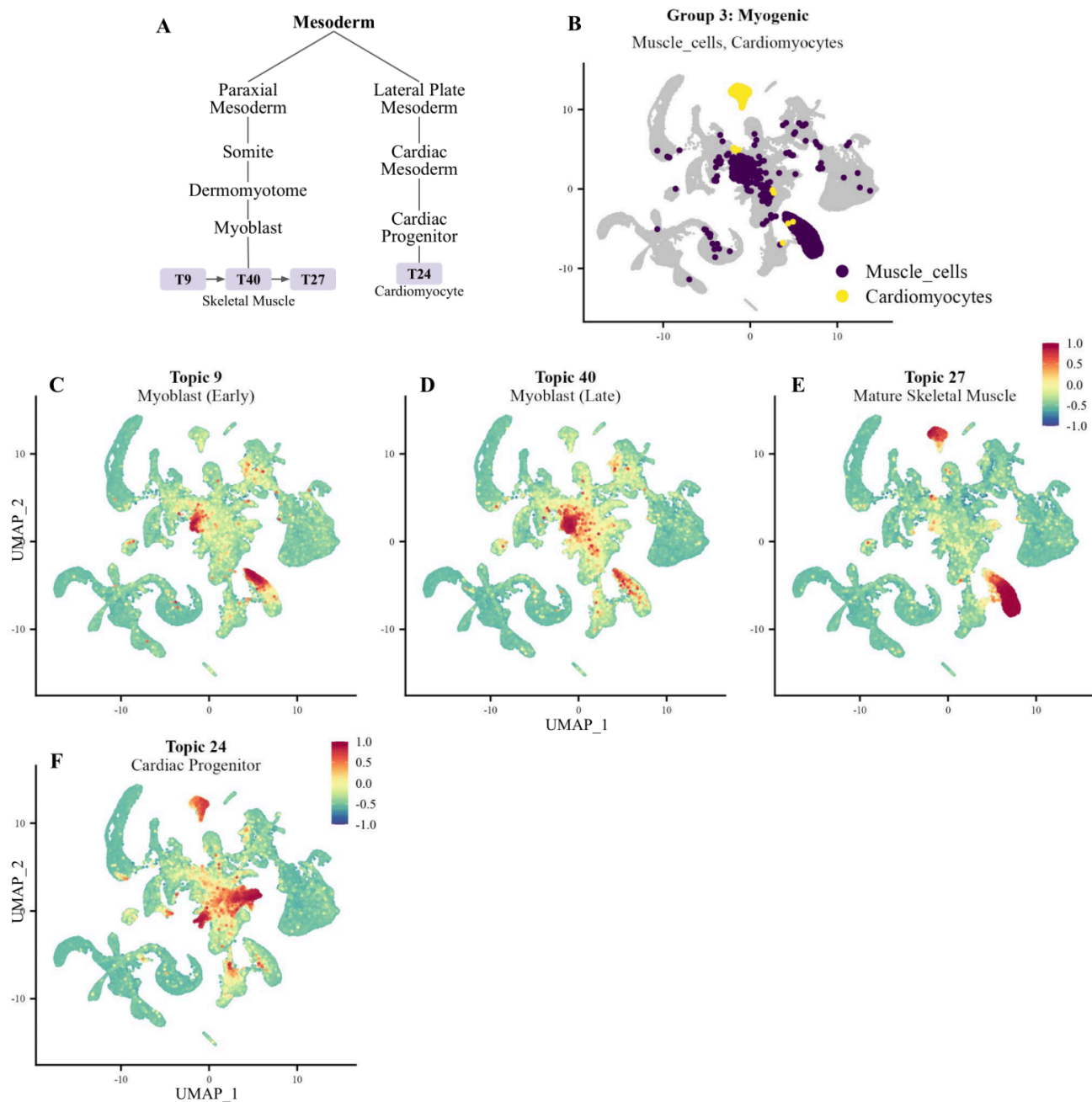


Figure S3. Group 3: Myogenic topics capture skeletal and cardiac muscle differentiation from mesoderm.

(A) Schematic of mesoderm-derived muscle development showing paraxial mesoderm giving rise to somite, dermomyotome, and myoblast lineages, and lateral plate mesoderm generating cardiac mesoderm, cardiac progenitors, and cardiomyocytes.

(B) UMAP visualization highlighting muscle cells and cardiomyocytes.

(C-F) UMAPs displaying topic weights for Topics 9 (Myoblast Early), 40 (Myoblast Late), 27 (Mature Skeletal Muscle), and 24 (Cardiac Progenitor).

Group 4: Mesenchymal/Skeletal Lineage (Topics 25, 1, 11, 15, 28, 32)

Mesenchymal cells are multipotent stromal cells that give rise to connective tissues including bone, cartilage, fat, and fibrous tissue. During development, mesenchymal progenitors derive primarily from paraxial mesoderm (sclerotome portion of somites) and neural crest (for craniofacial structures). These cells are characterized by their spindle-shaped morphology, ability to adhere to plastic in culture, and capacity for tri-lineage differentiation into osteoblasts, chondrocytes, and adipocytes¹⁷⁸.

Branching Structure

Differentiation Hierarchy: (*Topic 25, 1, 11*) The mesenchymal lineage follows a branching pattern from multipotent progenitors to specialized cell types. Topic 25 represents the earliest stage (paraxial mesoderm and sclerotome) expressing *Prrx1* (paired-related homeobox 1, marking limb and craniofacial mesenchyme) and *Mdk* (midkine). Topics 1 and 11 capture progressive mesenchymal commitment with expression of fibrillar collagens *Col3a1* and *Col5a2*, plus *Fap* (fibroblast activation protein)¹⁷⁸.

Terminal Differentiation Branches: (*Topic 15, 28, 32*) Topic 15 defines mature fibroblasts which are the principal cells of connective tissue that synthesize and maintain extracellular matrix while expressing *Vim* (vimentin), *Colla2*, and *Postn* (periostin, secreted during tissue remodeling). Topic 28 represents the osteogenic branch leading to osteoblasts (bone-forming cells) with expression of *Ibsp* (bone sialoprotein, a mineralization nucleator), *Satb2* (a transcription factor for osteoblast differentiation), and *Runx2* (the master regulator of osteogenesis). Topic 32 captures the chondrogenic branch leading to chondrocytes (cartilage cells) through cartilage-specific collagens *Col9a1*, *Col9a2*, *Col9a3*, *Col11a1*, and *Col27a1*⁷⁷. Chondrocytes are embedded in an avascular matrix rich in type II collagen and proteoglycans, providing cushioning at joints and structural support in the developing skeleton before ossification^{77,179,180}.

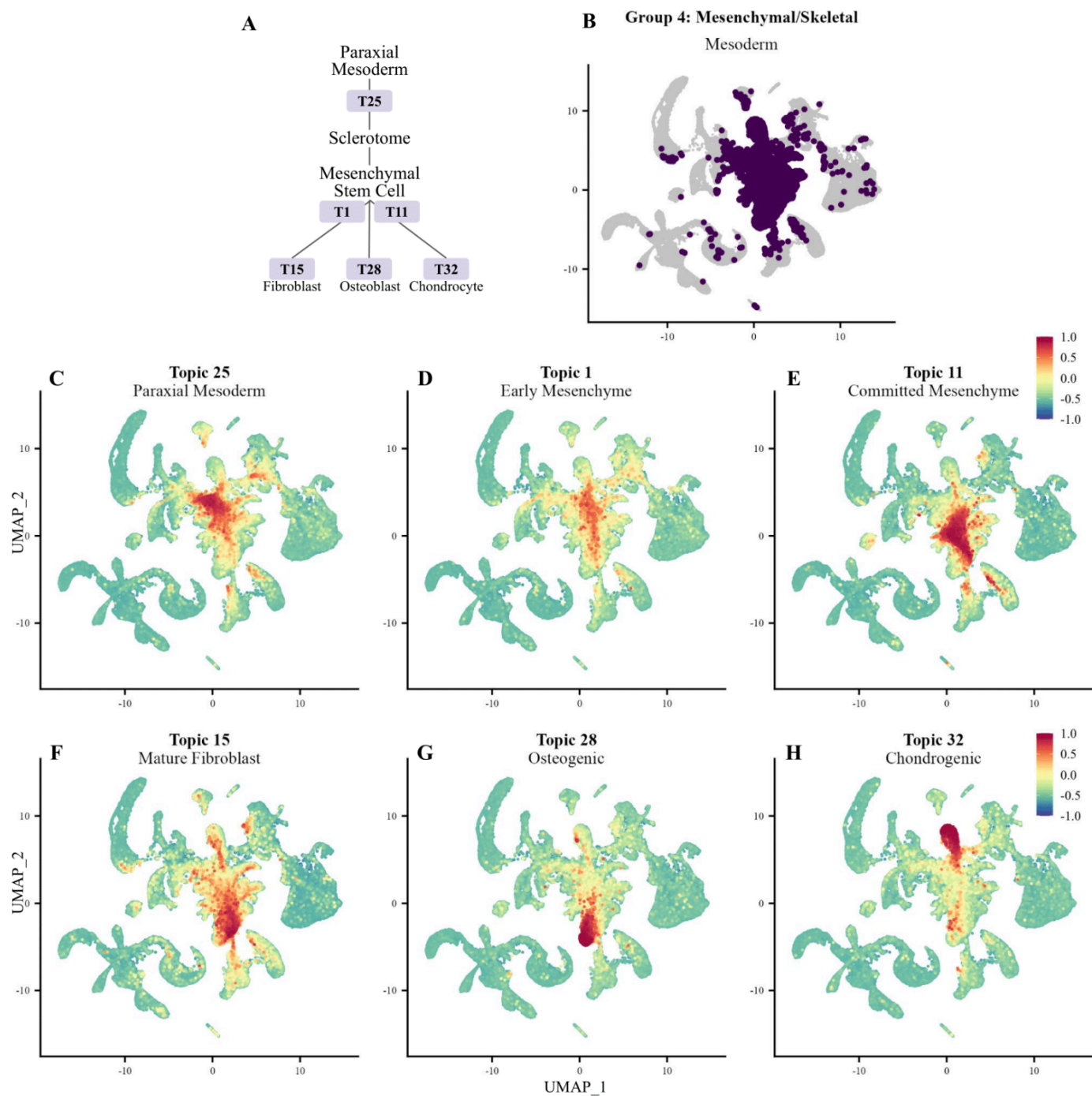


Figure S4. Group 4: Mesenchymal/skeletal topics delineate mesenchymal stem cell differentiation into connective tissue lineages.

(A) Schematic showing paraxial mesoderm giving rise to sclerotome and subsequently mesenchymal stem cells, which differentiate into fibroblasts, osteoblasts, and chondrocytes.

(B) UMAP visualization highlighting mesoderm trajectories.

(C-H) UMAPs displaying topic weights for Topics 25 (Paraxial Mesoderm), 1 (Early Mesenchyme), 11 (Committed Mesenchyme), 15 (Mature Fibroblast), 28 (Osteogenic), and 32 (Chondrogenic).

Group 5: Epithelial Lineage (Topics 35, 17, 39, 46)

Epithelial cells form sheets that line body surfaces and cavities, providing barrier function, selective transport, and secretion. Epithelia derive from all three germ layers: surface ectoderm (skin, cornea), endoderm (gut, lung, liver), and mesoderm. Despite diverse origins,

epithelial cells share key features such as apical-basal polarity, tight junctions connecting adjacent cells, and attachment to a basement membrane¹⁸¹.

Branching Structure.

Surface Epithelium Branch (*Topics 35 → 17*) Topic 35 represents epithelial progenitors expressing *Igfbp2* and *Mpped2*¹⁸¹, with broad expression across epidermal, corneal, and airway progenitor populations. Topic 17 marks mature surface epithelium with expression of *Lama5* (laminin alpha 5, a basement membrane component), *Esrp1* (epithelial splicing regulatory protein 1, which controls epithelial-specific alternative splicing of genes like *Fgfr2* and *Cd44*), and *Grhl2* (grainyhead-like 2, a transcription factor maintaining epithelial identity and suppressing epithelial-to-mesenchymal transition)^{77,181}.

Lung Epithelium (*Topic 39*) Derived from foregut endoderm, the lung epithelium differentiates into diverse cell types including alveolar type I cells (gas exchange), alveolar type II cells (surfactant production), and airway epithelial cells (ciliated and secretory). Topic 39 captures alveolar type II cell identity through expression of *Sftpb* and *Sftp*c (surfactant proteins B and C)^{182,183}.

Kidney Tubular Epithelium (*Topic 46*) Uniquely, kidney tubular epithelium derives from intermediate mesoderm (not endoderm) through mesenchymal-to-epithelial transition. Topic 46 expresses *Pkhd1* (polycystic kidney and hepatic disease 1, mutated in autosomal recessive PKD); *Slc4a4* (sodium bicarbonate cotransporter, essential for proximal tubule acid-base handling); and *Esrrg* (estrogen-related receptor gamma, which regulates metabolic genes in this highly metabolically active tissue)^{77,184}.

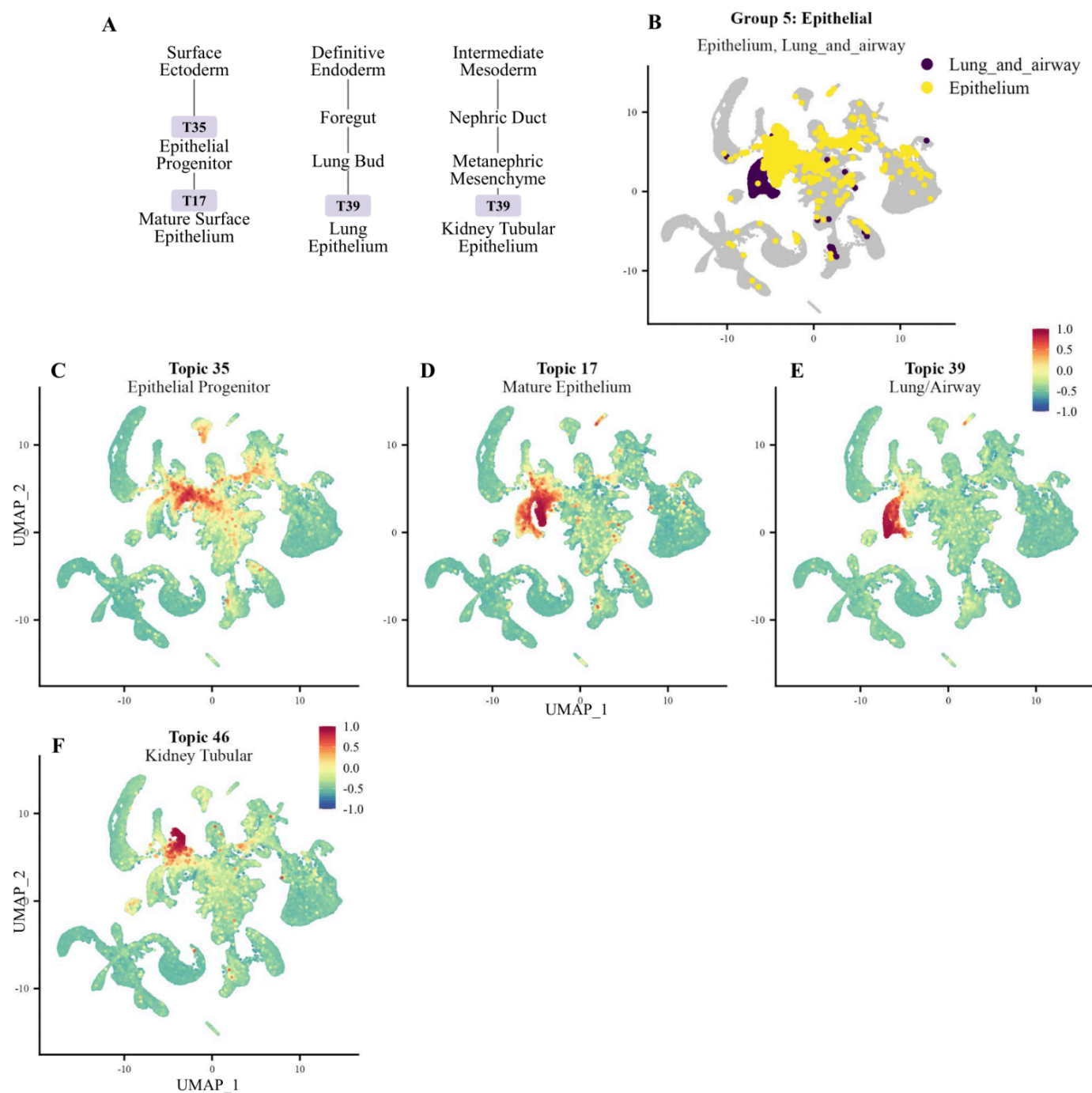


Figure S5. Group 5: Epithelial topics capture epithelial programs across distinct germ layer origins.

(A) Schematic showing three developmental sources of epithelium: surface ectoderm, definitive endoderm through foregut and lung bud, and intermediate mesoderm through nephric duct and metanephric mesenchyme.

(B) UMAP visualization highlighting lung/airway and epithelium trajectories.

(C-F) UMAPs displaying topic weights for Topics 35 (Epithelial Progenitor), 17 (Mature Epithelium), 39 (Lung/Airway), and 46 (Kidney Tubular).

Group 6: Sensory and Peripheral Nervous System (Topics 21, 4, 6)

The peripheral nervous system (PNS) derives from two embryonic sources including the neural crest cells and cranial placodes. Neural crest cells delaminate from the dorsal neural tube and migrate extensively to form sensory neurons of dorsal root ganglia (DRG), autonomic neurons (sympathetic and parasympathetic), and peripheral glia (Schwann cells). Cranial placodes are ectodermal

thickenings that give rise to specialized sensory structures including the olfactory placode (olfactory sensory neurons), otic placode (inner ear hair cells and spiral ganglion neurons), and lens placode (lens of the eye)^{37,185,186}.

Branching Structure

Olfactory Sensory Neurons (Topic 21) Derived from the olfactory placode, olfactory sensory neurons are bipolar cells with dendrites extending into the nasal epithelium (bearing odorant receptors) and axons projecting to the olfactory bulb. These neurons are unique in undergoing continuous neurogenesis throughout life^{37,187}. Topic 21 expresses *Scn9a* (Nav1.7 voltage-gated sodium channel, essential for action potential propagation in sensory neurons)¹⁸⁵; *Kcnh7* (potassium channel); and *Stmn2* (stathmin-2, a neuronal marker)^{77,187}.

Neural Crest-Derived Neurons (Topics 4 and 6) Topics 4 and 6 represent overlapping populations of neural crest-derived sensory and autonomic neurons. Topic 4 expresses *Sntg1*, *Rims1* (regulating synaptic membrane exocytosis 1, essential for neurotransmitter release), and *Cntnap5a* (a neurexin family member involved in cell adhesion)^{77,185}. Topic 6 shows expression of *Map2* (microtubule-associated protein 2, enriched in dendrites); *Frmd4a*; and *Dclk1* (doublecortin-like kinase 1, involved in neuronal migration). These topics also show enrichment in retinal neurons, reflecting shared neural crest contributions to eye development.

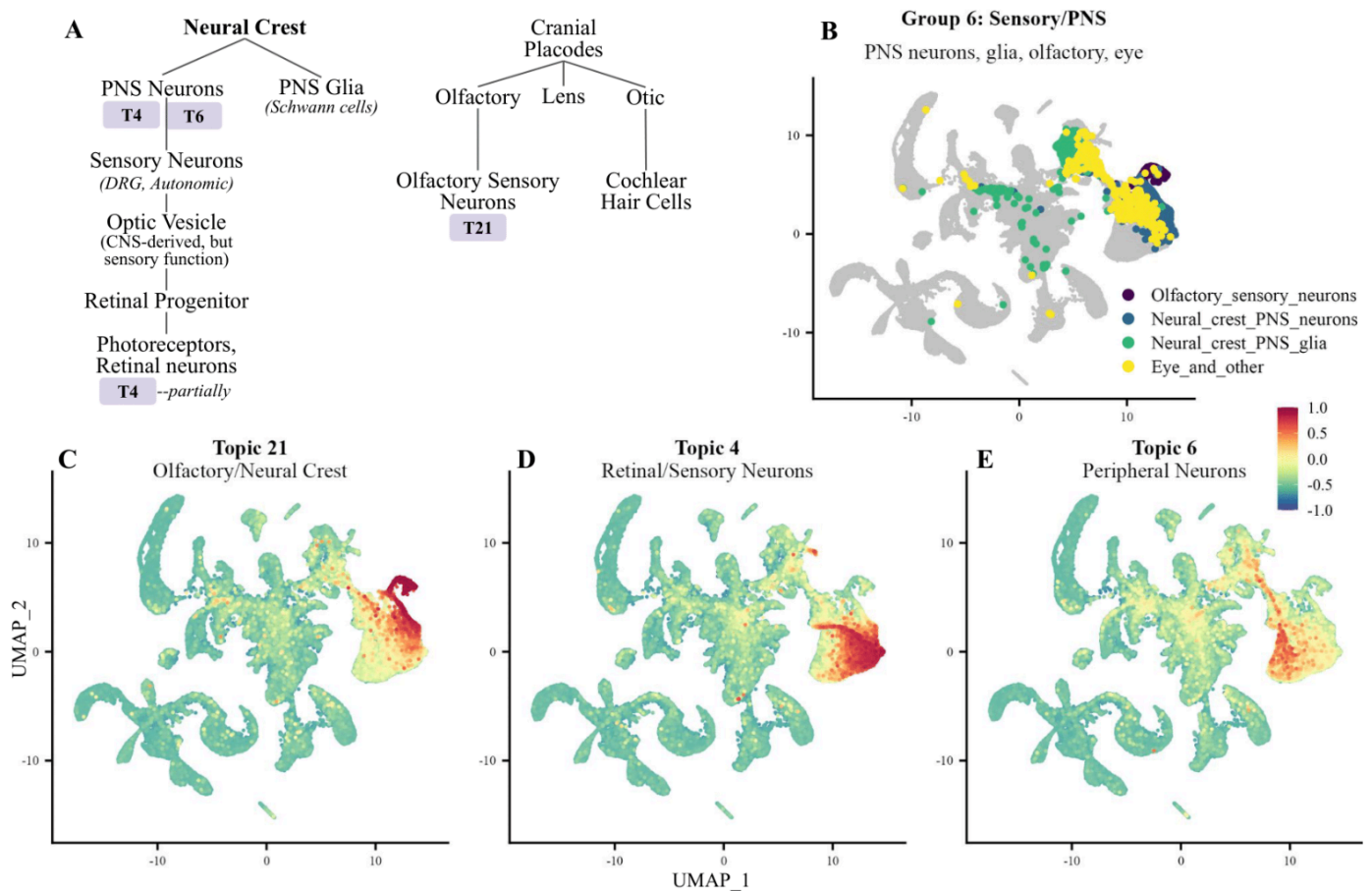


Figure S6. Group 6: Sensory/PNS topics mark peripheral nervous system and sensory neuron development from neural crest and cranial placodes.

(A) Schematic depicting neural crest derivatives including PNS neurons and PNS glia (Schwann cells), alongside cranial placode contributions to olfactory, lens, and otic structures. Sensory neurons derive from both dorsal root ganglia (DRG) and autonomic lineages; retinal neurons are CNS-derived but functionally sensory.

(B) UMAP visualization highlighting olfactory sensory neurons, neural crest PNS neurons, neural crest PNS glia, and eye/other trajectories.

(C-E) UMAPs displaying topic weights for Topics 21 (Olfactory/Neural Crest), 4 (Retinal/Sensory Neurons), and 6 (Peripheral Neurons).

Group 8: Endodermal Organs (Topics 22, 23)

The definitive endoderm gives rise to the epithelial lining of the gastrointestinal tract and associated organs. It is regionalized along the anterior-posterior axis into foregut (pharynx, esophagus, stomach, liver, pancreas, lung), midgut (small intestine), and hindgut

(colon, rectum). Organogenesis involves reciprocal signaling between endoderm and surrounding mesenchyme, leading to budding and branching morphogenesis of organs like liver, pancreas, and lung¹⁸⁸.

Branching structure

Hepatocytes (Topic 22) The liver bud emerges from the foregut endoderm adjacent to the developing heart, induced by FGF and BMP signals from cardiac mesoderm and septum transversum mesenchyme. Hepatoblasts differentiate into hepatocytes (parenchymal cells performing metabolic functions) and cholangiocytes (biliary epithelium)^{189,190}. Topic 22 captures hepatocyte identity with expression of *Cps1* (carbamoyl phosphate synthetase 1, the rate-limiting enzyme of the urea cycle that detoxifies ammonia); *Alb* (albumin, the most abundant plasma protein, synthesized exclusively by hepatocytes); *Afp* (alpha-fetoprotein, the fetal equivalent of albumin); and *Fga/Fgb/Fgg* (fibrinogen chains, coagulation factors)^{77,190}.

Intestinal Epithelium (Topic 23) The intestinal epithelium is a rapidly self-renewing tissue organized into crypts (containing stem cells) and villi (differentiated absorptive surface). Stem cells at the crypt base generate transit-amplifying cells that differentiate into absorptive enterocytes, mucus-secreting goblet cells, hormone-producing enteroendocrine cells, and antimicrobial Paneth cells^{191–193}. Topic 23 expresses *Cubn* (cubilin, a multiligand receptor in enterocytes for vitamin B12-intrinsic factor complex and other molecules); *Mttp* (microsomal triglyceride transfer protein, essential for chylomicron assembly and dietary lipid absorption); *Hnf4g* (hepatocyte nuclear factor 4 gamma, regulating intestinal gene expression); and *Mgat4a*^{77,191}. This topic is enriched in midgut/hindgut epithelium and goblet cells.

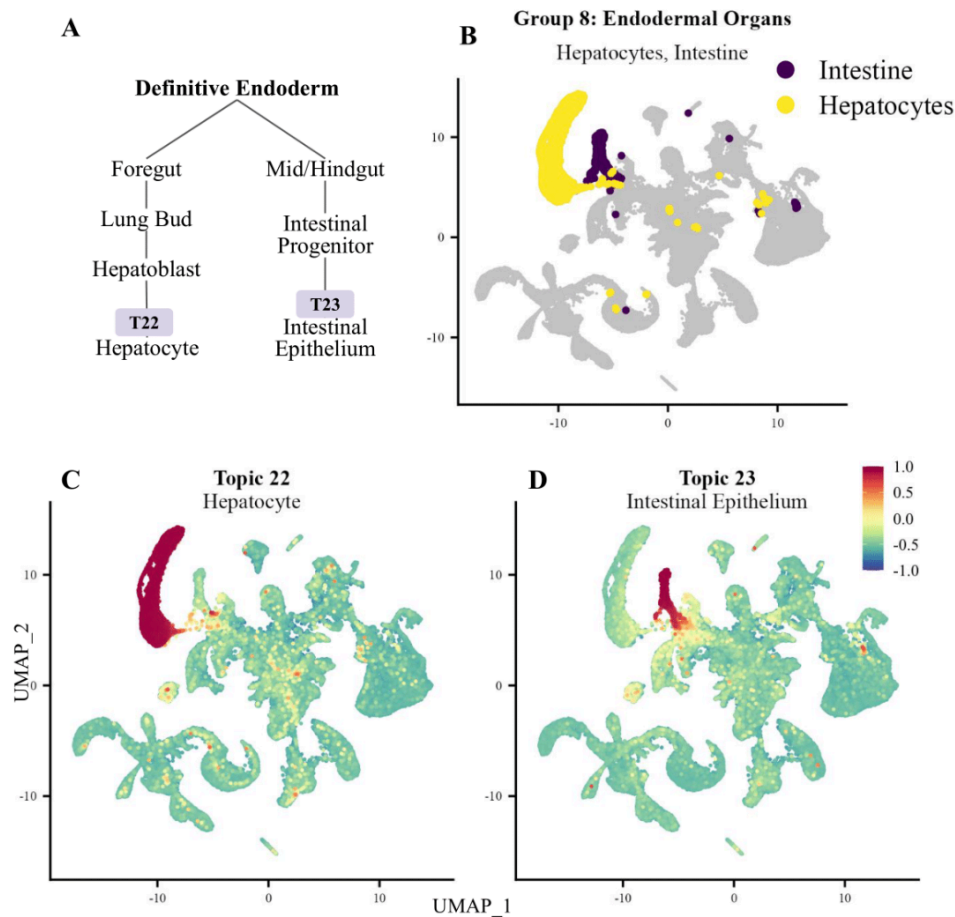


Figure S7. Group 8: Endodermal organ topics mark hepatocyte and intestinal epithelium differentiation from definitive endoderm.

(A) Schematic of definitive endoderm bifurcation into foregut (leading to lung bud and hepatoblast/hepatocyte) and mid/hindgut (leading to intestinal progenitor and intestinal epithelium).

(B) UMAP visualization highlighting intestine and hepatocyte trajectories.

(C-D) UMAPs displaying topic weights for Topics 22 (Hepatocyte) and 23 (Intestinal Epithelium).

Group 9: Adipocyte Program (Topics 30, 48)

Adipocytes (fat cells) differentiate from mesenchymal stem cells through a well-characterized transcriptional cascade involving C/EBP and PPAR family transcription factors. Two major types exist including white adipocytes (energy storage as triglycerides, single large lipid droplet) and brown adipocytes (thermogenesis via uncoupling protein 1, multiple small lipid droplets, abundant mitochondria). Adipose tissue is now recognized as a major endocrine organ secreting hormones (adipokines) like leptin and adiponectin that regulate systemic metabolism¹⁹⁴.

Adipocyte Identities (Topics 30 and 48). Unlike sequential differentiation, Topics 30 and 48 capture different functional aspects of adipocyte identity. Topic 30 represents the metabolic program with expression of *Acadl* (acyl-CoA dehydrogenase, long chain, catalyzing the first step of mitochondrial fatty acid beta-oxidation); *Glul* (glutamine synthetase); *Cat* (catalase, detoxifying hydrogen peroxide generated during lipid metabolism); and *Car3* (carbonic anhydrase 3, highly expressed in adipocytes). Topic 48 captures regulatory programs with expression of *Foxo1* (forkhead box O1, a transcription factor regulated by insulin signaling that controls adipocyte differentiation and metabolic gene expression); *Txnip* (thioredoxin-interacting protein, linking glucose sensing to redox state); and *Rapgef4*^{77,194}. Both topics show enrichment in brown and white adipocytes.

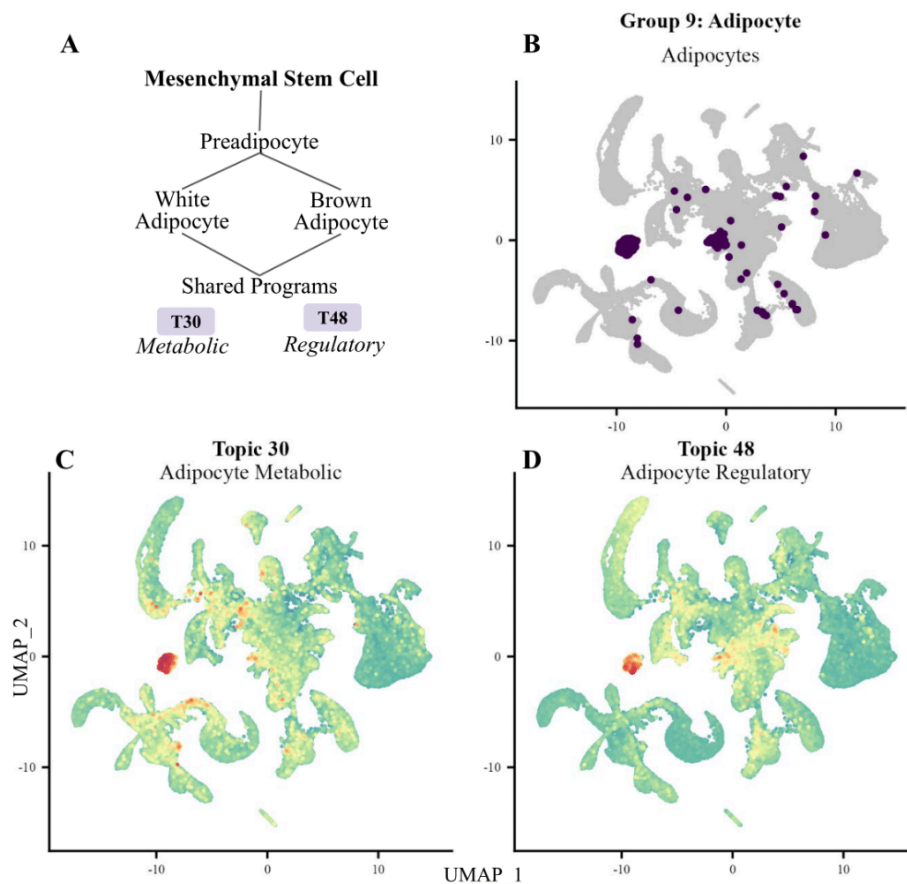


Figure S8. Group 9: Adipocyte topics distinguish metabolic and regulatory programs in adipose tissue.

(A) Schematic showing mesenchymal stem cell differentiation through preadipocytes to white and brown adipocytes, with shared signatures related to metabolic and regulatory biological mechanisms.

(B) UMAP visualization highlighting adipocyte trajectories.

(C-D) UMAPs displaying topic weights for Topics 30 (Adipocyte Metabolic) and 48 (Adipocyte Regulatory).

Group 10: Vascular Endothelium (Topic 43)

Endothelial cells line blood and lymphatic vessels, forming a selective barrier between blood and tissues. They derive from lateral plate mesoderm through hemangioblasts, which are bipotent progenitors that generate both hematopoietic and endothelial lineages. Vasculogenesis (de novo vessel formation from angioblasts) establishes the primary vascular plexus, which is then remodeled by angiogenesis (sprouting and branching from existing vessels). Endothelial cells differentiate into arterial, venous, and lymphatic subtypes with distinct molecular signatures¹⁹⁵.

Endothelial Identity (Topic 43) Topic 43 expresses canonical endothelial markers including *Pecam1* (CD31/platelet-endothelial cell adhesion molecule, concentrated at endothelial junctions where it mediates cell-cell adhesion and leukocyte transmigration and is widely used as an immunohistochemical marker); *Kdr* (VEGFR2/Flk1, the major VEGF receptor driving angiogenesis; VEGF signaling is essential for vascular development and is targeted in anti-cancer therapy); *Tek* (Tie2, the angiopoietin receptor regulating vessel stability and permeability); *Flt1* (VEGFR1); *Egfl7* (EGF-like domain 7, secreted specifically by endothelial cells during vascular development); and *Emcn* (endomucin)^{2,6,195}. This topic shows high enrichment in venous, capillary, and microvascular endothelium across multiple organs.

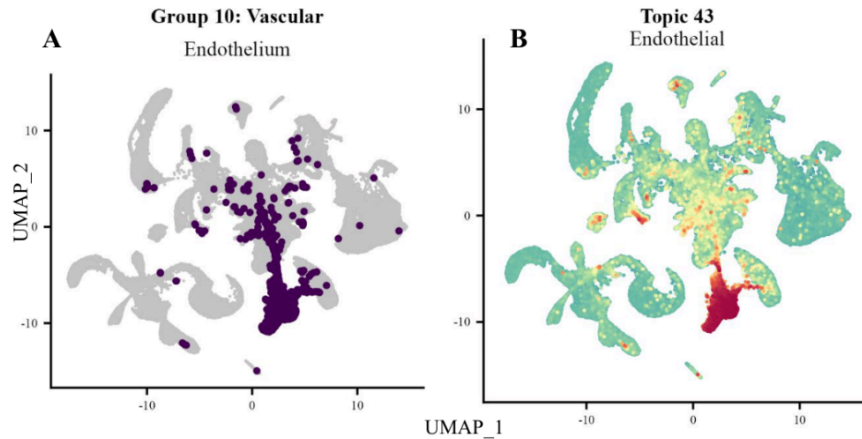


Figure S9. Group 10: Vascular topic marks endothelial cell identity.

(A) UMAP visualization highlighting endothelium trajectory.
(B) UMAP displaying topic weights for Topic 43 (Endothelial).

Group 11: Ciliated Cell types (Topic 34)

Cilia are microtubule-based organelles projecting from the cell surface. There are two major types of cilia: motile cilia (containing dynein motors, generating fluid flow) and primary cilia (sensory organelles, one per cell). Motile cilia are found on ependymal cells (lining brain ventricles, moving cerebrospinal fluid), airway epithelium (mucociliary clearance), fallopian tube epithelium (ovum transport), and the embryonic node (establishing left-right asymmetry)^{5,37,106}.

Ciliary Program (Topic 34) Topic 34 captures the shared transcriptional program of ciliated cells across diverse tissues. It expresses genes encoding axonemal and dynein regulatory complex components: *Dnah7b* (dynein axonemal heavy chain 7B); *Ccdc39*, *Ccdc148*, *Ccdc171* (coiled-coil domain proteins required for dynein arm assembly); and *Wdr78*. This topic shows highest expression in choroid plexus epithelium (produces CSF), ependymal cells (moves CSF), cochlear hair cells (mechanosensation), and nodal cilia (left-right patterning)^{37,77}.

Especially notable is that Topic 34 shows spatial overlap with the “Eye and other” major cellular trajectory from the original annotations, reflecting a fundamental biological relationship. Photoreceptor outer segments are highly specialized primary cilia, and the connecting cilium between inner and outer segments is essential for transporting phototransduction proteins. Mutations in ciliary genes, including dynein components like those in Topic 34, cause retinal degenerations such as Leber congenital amaurosis and retinitis pigmentosa. The “eye and other” cellular trajectory also encompasses inner ear structures (cochlear hair cells), which share placodal developmental origins with the eye and depend on ciliary gene programs. The overlap between Topic 34 and sensory structures thus provides biological validation of the topic's capture of a genuine ciliary program utilized across diverse cell types.

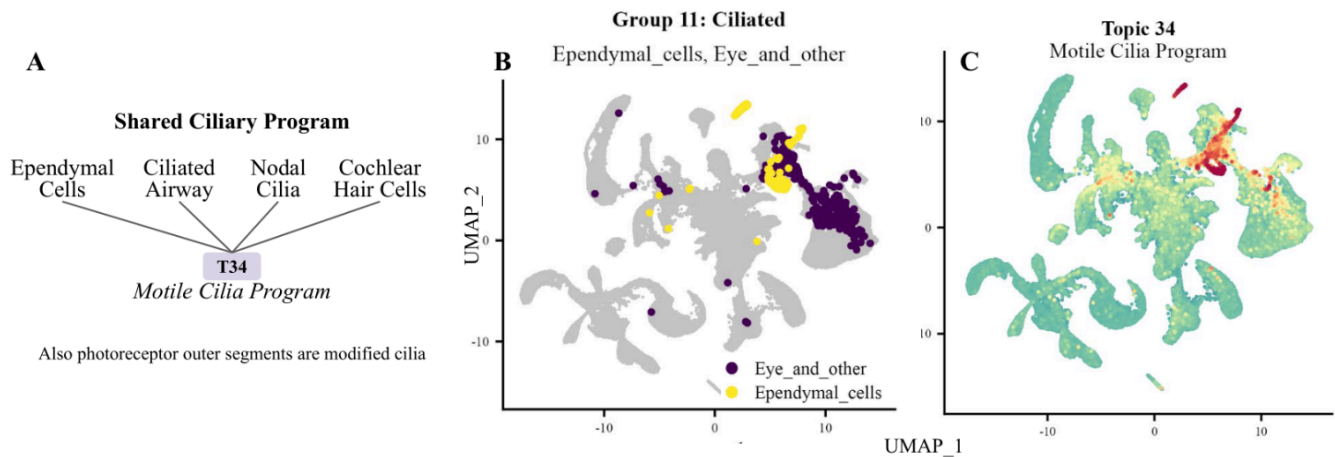


Figure S10. Group 11: Ciliated topic captures shared motile cilia program across developmentally distinct cell types.

(A) Schematic showing convergent expression of the motile cilia program in ependymal cells, ciliated airway epithelium, nodal cilia, and cochlear hair cells which are all cell types with different developmental origins that share ciliary machinery. Photoreceptor outer segments, which are modified cilia, also partially express this program.

(B) UMAP visualization highlighting ependymal cells and eye/other trajectories.

(C) UMAP displaying topic weights for Topic 34 (Motile Cilia Program).

Group 12: Cell Cycle related (Topic 33)

Cell Cycle Program. Unlike the lineage-specific groups, Topic 33 represents a cellular state of active proliferation rather than a differentiation lineage. The cell cycle consists of G1 (growth), S (DNA synthesis), G2 (preparation for mitosis), and M (mitosis) phases; most differentiated cells exit to G0 (quiescence). Topic 33 captures cells in S/G2/M phases through expression of canonical cell cycle genes such as *Mki67* (Ki-67, a widely used proliferation marker that associates with ribosomal RNA transcription); *Top2a* (topoisomerase II alpha, essential for chromosome condensation and segregation); *Aspm* (assembly factor for spindle microtubules, mutations cause microcephaly); *Cenpf* (centromere protein F); *Bard1* (BRCA1-associated RING domain 1, involved in DNA damage response); and *Kif14* (a mitotic kinesin)^{76,77}.

Topic 33 is expressed in dividing progenitors across many lineages: hematopoietic stem cells, neural progenitors, myoblasts, transit-amplifying intestinal cells, and any other actively cycling population. It shows positive correlation with topics capturing progenitor states ($r=0.35$ with erythroid progenitors in Topic 3) and anti-correlation with mature cell types ($r=-0.21$ with differentiated neurons in Topic 7). High expression in progenitor populations combined with low expression in post-mitotic cells (neurons, hepatocytes, mature muscle) confirms that Topic 33 marks the universal proliferative state independent of lineage identity. This topic serves as a useful modifier when interpreting other topics, as high Topic 33 co-expression indicates a progenitor rather than terminally differentiated state.

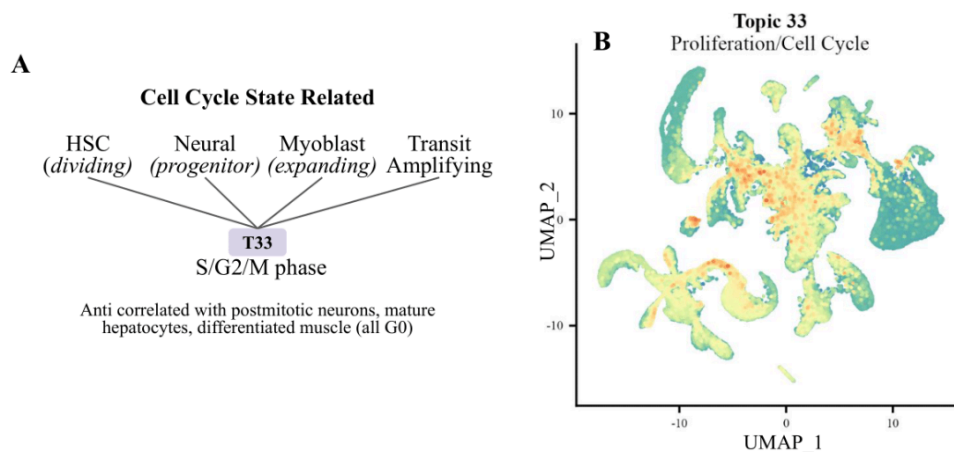


Figure S11. Group 12: Cell cycle topic represents a universal proliferation signature across lineages.

(A) Schematic illustrating that Topic 33 (S/G2/M phase) is expressed in actively dividing cells across all lineages, including HSCs, neural progenitors, expanding myoblasts, and transit-amplifying populations. This topic is anti-correlated with postmitotic neurons, mature hepatocytes, and differentiated muscle (all G0).

(B) UMAP displaying topic weights for Topic 33 (Proliferation/Cell Cycle).

Ungrouped Topics

Ten topics (8, 10, 12, 14, 16, 36, 41, 45, 47, and 50) showed low maximum expression scores (<0.5) or diffuse expression patterns across multiple unrelated cell types and were not assigned to lineage-specific groups. These topics may represent technical variation, subtle ubiquitous programs (housekeeping functions, stress responses), rare cell populations inadequately sampled, or biological signals requiring additional characterization. While these topics remain outside our established grouping framework, they may be instrumental in revealing the elusive biology of transient and rare cell types that have outwit our characterizations until now. In future works, these topics will be further explored.

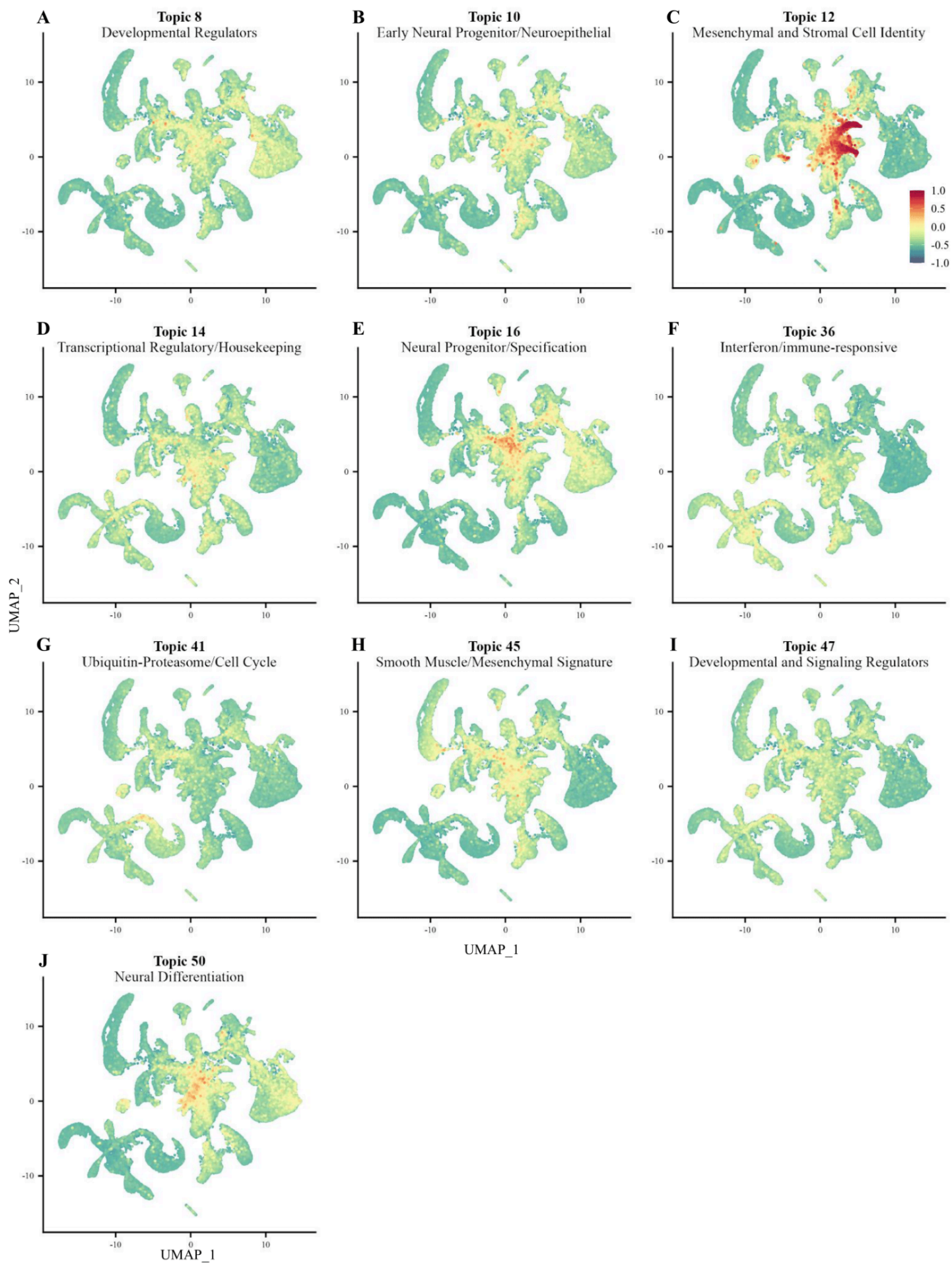


Figure S11. Ungrouped topics display diffuse expression patterns or ambiguous biological signatures.

UMAPs display topic weights for topics showing low maximum expression scores (<0.5 across all trajectories), expression patterns spanning multiple unrelated lineages, or ambiguous biological signatures, precluding confident assignment to lineage-specific groups pending additional characterization.

- (A) Topics 8 (Developmental Regulators),
- (B) 10 (Early Neural Progenitor/Neuroepithelial),
- (C) 12 (Mesenchymal and Stromal Cell Identity),
- (D) 14 (Transcriptional Regulatory/Housekeeping),
- (E) 16 (Neural Progenitor/Specification), 36 (Interferon/Immune-responsive),
- (F) 41 (Ubiquitin-Proteasome/Cell Cycle),
- (H) 45 (Smooth Muscle/Mesenchymal Signature),
- (I) 47 (Developmental and Signaling Regulators), and
- (J) 50 (Neural Differentiation).

Topic	Primary Identity	Defining Marker Genes	Lineage Specificity
8	Developmental regulators	Pbx1, Meis2, Nfib, Celf2, Ptprd	Low (diffuse)
10	Neural/mesenchymal progenitor	Pbx1, Nfia, Reln, Tshz2, Nbea	Low (diffuse)
12	Mesenchymal/stromal	Fn1, Col25a1, Lama2, Lama4, Tgfb3, Adam12	Moderate
14	General housekeeping	Actb, Foxp1, Iqgap1, Gnas, Macf1	Very low (ubiquitous)
16	Neural progenitor	Aff3, Pbx1, Pbx3, Unc5c, Epha7, Nnat	Low-moderate
36	Interferon/immune-responsive	Stat1, Stat2, Stat3, Samhd1	Moderate
41	Signaling/ubiquitin-proteasome	Herc1, Maml3, March8, Cep152	Very low (diffuse)
45	Smooth muscle/mesenchymal	Myocd, Cald1, Mylk, Fn1, Utn	Moderate
47	General developmental	Nfia, Fgfr2, Mbnl1, Bach2, Foxp1	Very low (diffuse)
50	Neural differentiation	Ebf1, Ebf2, Ebf3, Ntrk3, Nrp1, Prickle2	Low-moderate

Table S1. Summary of Topic Identities, Key Marker Genes, and Expression Specificity.

Table summarizing the primary biological identity, key marker genes, and degree of lineage specificity for each of the ten analyzed topics. Topic identities were inferred from the top-weighted genes and specificity was assessed by visual inspection of the corresponding cell-topic associations relative to the major trajectory annotations. Topics with "low" or "very low" specificity exhibit diffuse expression across multiple cell types and likely represent shared developmental or housekeeping programs. Topics with "moderate" specificity show enrichment in specific regions of the UMAP corresponding to defined cell lineages.



References

1. Subramanian, S., and Ahmad, T. (2023). Childhood Brain Tumors. In StatPearls (StatPearls Publishing).
2. Qiu, C., Cao, J., Martin, B.K., Li, T., Welsh, I.C., Srivatsan, S., Huang, X., Calderon, D., Noble, W.S., Disteché, C.M., et al. (2022). Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nat. Genet.* *54*, 328–341. <https://doi.org/10.1038/s41588-022-01018-x>.
3. Cao, J., O'Day, D.R., Pliner, H.A., Kingsley, P.D., Deng, M., Daza, R.M., Zager, M.A., Aldinger, K.A., Blecher, R., Zhang, F., et al. (2020). A human cell atlas of fetal gene expression. *Science* *370*, eaba7721. <https://doi.org/10.1126/science.aba7721>.
4. Pijuan-Sala, B., Griffiths, J.A., Guibentif, C., Hiscock, T.W., Jawaaid, W., Calero-Nieto, F.J., Mulas, C., Ibarra-Soria, X., Tyser, R.C.V., Ho, D.L.L., et al. (2019). A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature* *566*, 490–495. <https://doi.org/10.1038/s41586-019-0933-9>.
5. Qiu, C., Martin, B.K., Welsh, I.C., Daza, R.M., Le, T.-M., Huang, X., Nichols, E.K., Taylor, M.L., Fulton, O., O'Day, D.R., et al. (2023). A single-cell transcriptional timelapse of mouse embryonic development, from gastrula to pup. Preprint at bioRxiv, <https://doi.org/10.1101/2023.04.05.535726> <https://doi.org/10.1101/2023.04.05.535726>.
6. Khouri-Farah, N., Guo, Q., Morgan, K., Shin, J., and Li, J.Y.H. (2022). Integrated single-cell transcriptomic and epigenetic study of cell state transition and lineage commitment in embryonic mouse cerebellum. *Sci. Adv.* *8*, eabl9156. <https://doi.org/10.1126/sciadv.abl9156>.
7. Yao, Z., van Velthoven, C.T.J., Kunst, M., Zhang, M., McMillen, D., Lee, C., Jung, W., Goldy, J., Abdelhak, A., Aitken, M., et al. (2023). A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* *624*, 317–332. <https://doi.org/10.1038/s41586-023-06812-z>.
8. Allen Mouse Brain Atlas. <https://developingmouse.brain-map.org/>.
9. Siletti, K., Hodge, R., Albiach, A.M., Hu, L., Lee, K.W., Lönnerberg, P., Bakken, T., Ding, S.-L., Clark, M., Casper, T., et al. (2022). Transcriptomic diversity of cell types across the adult human brain. Preprint at bioRxiv, <https://doi.org/10.1101/2022.10.12.511898> <https://doi.org/10.1101/2022.10.12.511898>.
10. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* *45*, 580–585. <https://doi.org/10.1038/ng.2653>.
11. TARGET Project Team (2013). The Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative: overview and progress. *Pediatr. Blood Cancer* *60*, S1–S10. <https://doi.org/10.1002/pbc.24650>.
12. Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* *45*, 1113–1120. <https://doi.org/10.1038/ng.2764>.
13. Traag, V.A., Waltman, L., and van Eck, N.J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* *9*, 5233. <https://doi.org/10.1038/s41598-019-41695-z>.
14. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. *Cell* *177*, 1888–1902.e21. <https://doi.org/10.1016/j.cell.2019.05.031>.
15. Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* *36*, 421–427. <https://doi.org/10.1038/nbt.4091>.
16. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* *16*, 1289–1296. <https://doi.org/10.1038/s41592-019-0619-0>.
17. Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* *3*, 993–1022.
18. Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemom. Intell. Lab. Syst.* *2*, 37–52. [https://doi.org/10.1016/0169-7439\(87\)80084-9](https://doi.org/10.1016/0169-7439(87)80084-9).

19. Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 788–791. <https://doi.org/10.1038/44565>.
20. De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). An introduction to independent component analysis. *J. Chemom.* 14, 123–149. [https://doi.org/10.1002/1099-128X\(200005/06\)14:3%3C123::AID-CEM589%3E3.0.CO;2-1](https://doi.org/10.1002/1099-128X(200005/06)14:3%3C123::AID-CEM589%3E3.0.CO;2-1).
21. Valle, F., Caselle, M., and Osella, M. (2025). Exploring the latent space of transcriptomic data with topic modeling. *NAR Genomics Bioinforma.* 7, lqaf049. <https://doi.org/10.1093/nargab/lqaf049>.
22. Valle, F., Osella, M., and Caselle, M. (2020). A Topic Modeling Analysis of TCGA Breast and Lung Cancer Transcriptomic Data. *Cancers* 12, 3799. <https://doi.org/10.3390/cancers12123799>.
23. Wang, Z., Yang, S., Koga, Y., Corbett, S.E., Johnson, W.E., Yajima, M., and Campbell, J.D. (2021). Celda: A Bayesian model to perform co-clustering of genes into modules and cells into subpopulations using single-cell RNA-seq data. Preprint at bioRxiv, <https://doi.org/10.1101/2020.11.16.373274> <https://doi.org/10.1101/2020.11.16.373274>.
24. Carbonetto, P., Sarkar, A., Wang, Z., and Stephens, M. (2025). Non-negative matrix factorization algorithms generally improve topic model fits. Preprint at arXiv, <https://doi.org/10.48550/arXiv.2105.13440> <https://doi.org/10.48550/arXiv.2105.13440>.
25. Yao, L., Mimno, D., and McCallum, A. (2009). Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (ACM)*, pp. 937–946. <https://doi.org/10.1145/1557019.1557121>.
26. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., et al. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352, 189–196. <https://doi.org/10.1126/science.aad0501>.
27. Bravo González-Blas, C., Minnoye, L., Papasokrati, D., Aibar, S., Hulselmans, G., Christiaens, V., Davie, K., Wouters, J., and Aerts, S. (2019). cisTopic: cis-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* 16, 397–400. <https://doi.org/10.1038/s41592-019-0367-1>.
28. Zhao, W., Chen, J.J., Perkins, R., Liu, Z., Ge, W., Ding, Y., and Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics* 16, S8. <https://doi.org/10.1186/1471-2105-16-S13-S8>.
29. Selivanov, D., models), M.B. (Coherence measures for topic, and code), Q.W. (Author of the W.C. (2025). text2vec: Modern Text Mining Framework for R. Version 0.6.6.
30. Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, R. Barzilay and M. Johnson, eds. (Association for Computational Linguistics), pp. 262–272.
31. Rijcken, E., Zervanou, K., Spruit, M., Scheepers, F., and Kaymak, U. (2023). Effect of Calculating Pointwise Mutual Information using a Fuzzy Sliding Window in Topic Modeling. In *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, pp. 1–6. <https://doi.org/10.1109/FUZZ52849.2023.10309675>.
32. Fong, E., and Holmes, C.C. (2020). On the marginal likelihood and cross-validation. *Biometrika* 107, 489–496. <https://doi.org/10.1093/biomet/asz077>.
33. Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* 27, 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>.
34. Hosseiny Marani, A., and Baumer, E.P.S. (2023). A Review of Stability in Topic Modeling: Metrics for Assessing and Techniques for Improving Stability. *ACM Comput Surv* 56, 108:1-108:32. <https://doi.org/10.1145/3623269>.
35. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14, 7. <https://doi.org/10.1186/1471-2105-14-7>.
36. HALDIPUR, P., DANG, D., and MILLEN, K.J. (2018). Embryology. *Handb. Clin. Neurol.* 154, 29–44. <https://doi.org/10.1016/B978-0-444-63956-1.00002-3>.

37. Rubenstein, J.L.R., Rakic, P., Chen, B., and Kwan, K.Y. eds. (2020). *Patterning and Cell Type Specification in the Developing CNS and PNS* 2nd ed. (Academic Press).
38. Filbin, M., and Monje, M. (2019). Developmental origins and emerging therapeutic opportunities for childhood cancer. *Nat. Med.* 25, 367–376. <https://doi.org/10.1038/s41591-019-0383-9>.
39. Semple, B.D., Blomgren, K., Gimlin, K., Ferriero, D.M., and Noble-Haeusslein, L.J. (2013). Brain development in rodents and humans: Identifying benchmarks of maturation and vulnerability to injury across species. *Prog. Neurobiol.* 0, 1–16. <https://doi.org/10.1016/j.pneurobio.2013.04.001>.
40. Cerebellar patterning (2020). In *Patterning and Cell Type Specification in the Developing CNS and PNS* (Academic Press), pp. 107–135. <https://doi.org/10.1016/B978-0-12-814405-3.00006-0>.
41. Chen, X., Yang, W., Roberts, C.W.M., and Zhang, J. (2024). Developmental origins shape the pediatric cancer genome. *Nat. Rev. Cancer* 24, 382–398. <https://doi.org/10.1038/s41568-024-00684-9>.
42. Gibson, P., Tong, Y., Robinson, G., Thompson, M.C., Curre, D.S., Eden, C., Kranenburg, T.A., Hogg, T., Poppleton, H., Martin, J., et al. (2010). Subtypes of medulloblastoma have distinct developmental origins. *Nature* 468, 1095–1099. <https://doi.org/10.1038/nature09587>.
43. Camara, P.G. (2018). Methods and challenges in the analysis of single-cell RNA-sequencing data. *Curr. Opin. Syst. Biol.* 7, 47–53. <https://doi.org/10.1016/j.coisb.2017.12.007>.
44. Heumos, L., Schaar, A.C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M.D., Strobl, D.C., Henao, J., Curion, F., et al. (2023). Best practices for single-cell analysis across modalities. *Nat. Rev. Genet.*, 1–23. <https://doi.org/10.1038/s41576-023-00586-w>.
45. Blei, D.M. Latent Dirichlet Allocation.
46. Haldivur, P., Dang, D., Aldinger, K.A., Janber, S., and others (2018). Phenotypic outcomes in mouse and human *Foxc1* dependent Dandy-Walker cerebellar malformation suggest shared mechanisms. *eLife* 6, e20898. <https://doi.org/10.7554/eLife.20898>.
47. La Manno, G., Gyllborg, D., Codeluppi, S., Nishimura, K., Salto, C., Zeisel, A., Borm, L.E., Stott, S.R.W., Toledo, E.M., Villaescusa, J.C., et al. (2016). Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell* 167, 566–580.e19. <https://doi.org/10.1016/j.cell.2016.09.027>.
48. Franzén, O., Gan, L.-M., and Björkegren, J.L.M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database* 2019, baz046. <https://doi.org/10.1093/database/baz046>.
49. Lein, E.S., Hawrylycz, M.J., Ao, N., Ayres, M., Bensinger, A., Bernard, A., Boe, A.F., Boguski, M.S., Brockway, K.S., Byrnes, E.J., et al. (2007). Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445, 168–176. <https://doi.org/10.1038/nature05453>.
50. Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* 347, 1138–1142. <https://doi.org/10.1126/science.aaa1934>.
51. Cerebellum (2012). In *The Mouse Nervous System* (Academic Press), pp. 360–397. <https://doi.org/10.1016/B978-0-12-369497-3.10011-1>.
52. Qiu, C., Martin, B.K., Welsh, I.C., Daza, R.M., Le, T.-M., Huang, X., Nichols, E.K., Taylor, M.L., Fulton, O., O’Day, D.R., et al. (2024). A single-cell time-lapse of mouse prenatal development from gastrula to birth. *Nature* 626, 1084–1093. <https://doi.org/10.1038/s41586-024-07069-w>.
53. 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. <https://doi.org/10.1016/j.cels.2019.03.003>.
54. Wolock, S.L., Lopez, R., and Klein, A.M. (2019). Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst.* 8, 281–291.e9. <https://doi.org/10.1016/j.cels.2018.11.005>.

55. Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12, 323. <https://doi.org/10.1186/1471-2105-12-323>.
56. Boeshaghi, A.S., Hallgrímsdóttir, I.B., Gálvez-Merchán, Á., and Pachter, L. (2022). Depth normalization for single-cell genomics count data. Preprint at bioRxiv, <https://doi.org/10.1101/2022.05.06.490859>.
57. L. Lun, A.T., Bach, K., and Marioni, J.C. (2016). Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* 17, 75. <https://doi.org/10.1186/s13059-016-0947-7>.
58. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* 20, 296. <https://doi.org/10.1186/s13059-019-1874-1>.
59. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. *Cell* 177, 1888-1902.e21. <https://doi.org/10.1016/j.cell.2019.05.031>.
60. Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Preprint at bioRxiv, <https://doi.org/10.1101/576827> <https://doi.org/10.1101/576827>.
61. Andrews, T.S., and Hemberg, M. (2018). Dropout-based feature selection for scRNASeq. Preprint at bioRxiv, <https://doi.org/10.1101/065094> <https://doi.org/10.1101/065094>.
62. Andrews, T.S., and Hemberg, M. (2019). M3Drop: dropout-based feature selection for scRNASeq. *Bioinformatics* 35, 2865–2867. <https://doi.org/10.1093/bioinformatics/bty1044>.
63. McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. Preprint at arXiv, <https://doi.org/10.48550/arXiv.1802.03426> <https://doi.org/10.48550/arXiv.1802.03426>.
64. Maaten, L. van der, and Hinton, G. (2008). Visualizing Data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.
65. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* 2008, P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>.
66. Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining KDD'96*. (AAAI Press), pp. 226–231.
67. Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A. (2011). Density-based clustering. *WIREs Data Min. Knowl. Discov.* 1, 231–240. <https://doi.org/10.1002/widm.30>.
68. Campello, R.J.G.B., Moulavi, D., and Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*, J. Pei, V. S. Tseng, L. Cao, H. Motoda, and G. Xu, eds. (Springer), pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14.
69. Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining KDD '08*. (Association for Computing Machinery), pp. 569–577. <https://doi.org/10.1145/1401890.1401960>.
70. Jonathan Chang (2009). lda: Collapsed Gibbs Sampling Methods for Topic Models. <https://doi.org/10.32614/CRAN.package.lda> <https://doi.org/10.32614/CRAN.package.lda>.
71. Swapna, L.S., Huang, M., and Li, Y. (2023). GTM-decon: guided-topic modeling of single-cell transcriptomes enables sub-cell-type and disease-subtype deconvolution of bulk transcriptomes. *Genome Biol.* 24, 1–32. <https://doi.org/10.1186/s13059-023-03034-4>.
72. Pancheva, A., Wheadon, H., Rogers, S., and Otto, T.D. (2022). Using topic modeling to detect cellular crosstalk in scRNA-seq. *PLoS Comput. Biol.* 18, e1009975. <https://doi.org/10.1371/journal.pcbi.1009975>.
73. Neishabouri, A., and Desmarais, M.C. (2012). Reliability of Perplexity to Find Number of Latent Topics. In *Late-Breaking Results, Project Papers and Workshop Proceedings of the 21st Conference on User Modeling, Adaptation, and Personalization*.

74. Hu, C., Li, T., Xu, Y., Zhang, X., Li, F., Bai, J., Chen, J., Jiang, W., Yang, K., Ou, Q., et al. (2023). CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res.* *51*, D870–D876. <https://doi.org/10.1093/nar/gkac947>.
75. The Gene Ontology Consortium, Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., et al. (2023). The Gene Ontology knowledgebase in 2023. *Genetics* *224*, iyad031. <https://doi.org/10.1093/genetics/iyad031>.
76. Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* *28*, 27–30. <https://doi.org/10.1093/nar/28.1.27>.
77. Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., Stein, T.I., Nudel, R., Lieder, I., Mazon, Y., et al. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinforma.* *54*, 1.30.1–1.30.33. <https://doi.org/10.1002/cpbi.5>.
78. Aldinger, K.A., Thomson, Z., Phelps, I.G., Haldipur, P., Deng, M., Timms, A.E., Hirano, M., Santpere, G., Roco, C., Rosenberg, A.B., et al. (2021). Spatial and cell type transcriptional landscape of human cerebellar development. *Nat. Neurosci.* *24*, 1163–1175. <https://doi.org/10.1038/s41593-021-00872-y>.
79. Braun, E., Danan-Gothold, M., Borm, L.E., Vinsland, E., Lee, K.W., Lönnerberg, P., Hu, L., Li, X., He, X., Andrusivová, Ž., et al. (2022). Comprehensive cell atlas of the first-trimester developing human brain (Developmental Biology) <https://doi.org/10.1101/2022.10.24.513487>.
80. Ramos, S.I., Mussa, Z.M., Falk, E.N., Pai, B., Giotti, B., Allette, K., Cai, P., Dekio, F., Sebra, R., Beaumont, K.G., et al. (2022). An atlas of late prenatal human neurodevelopment resolved by single-nucleus transcriptomics. *Nat. Commun.* *13*, 7671. <https://doi.org/10.1038/s41467-022-34975-2>.
81. Fan, X., Dong, J., Zhong, S., Wei, Y., Wu, Q., Yan, L., Yong, J., Sun, L., Wang, X., Zhao, Y., et al. (2018). Spatial transcriptomic survey of human embryonic cerebral cortex by single-cell RNA-seq analysis. *Cell Res.* *28*, 730–745. <https://doi.org/10.1038/s41422-018-0053-3>.
82. Tejada-Lapuerta, A., Bertin, P., Bauer, S., Aliee, H., Bengio, Y., and Theis, F.J. (2025). Causal machine learning for single-cell genomics. *Nat. Genet.* *57*, 797–808. <https://doi.org/10.1038/s41588-025-02124-2>.
83. Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M., et al. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* *19*, 41–50. <https://doi.org/10.1038/s41592-021-01336-8>.
84. Tran, H.T.N., Ang, K.S., Chevrier, M., Zhang, X., Lee, N.Y.S., Goh, M., and Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* *21*, 12. <https://doi.org/10.1186/s13059-019-1850-9>.
85. Lotfollahi, M., Wolf, F.A., and Theis, F.J. (2019). scGen predicts single-cell perturbation responses. *Nat. Methods* *16*, 715–721. <https://doi.org/10.1038/s41592-019-0494-8>.
86. Leto, K., Arancillo, M., Becker, E.B.E., Buffo, A., Chiang, C., Ding, B., Dobyns, W.B., Dusart, I., Haldipur, P., Hatten, M.E., et al. (2016). Consensus Paper: Cerebellar Development. *Cerebellum Lond. Engl.* *15*, 789–828. <https://doi.org/10.1007/s12311-015-0724-2>.
87. Butts, T., Green, M.J., and Wingate, R.J.T. (2014). Development of the cerebellum: simple steps to make a “little brain.” *Dev. Camb. Engl.* *141*, 4031–4041. <https://doi.org/10.1242/dev.106559>.
88. Consalez, G.G., Goldowitz, D., Casoni, F., and Hawkes, R. (2021). Origins, Development, and Compartmentation of the Granule Cells of the Cerebellum. *Front. Neural Circuits* *14*. <https://doi.org/10.3389/fncir.2020.611841>.
89. Hovestadt, V., Smith, K.S., Bihannic, L., Filbin, M.G., Shaw, M.L., Baumgartner, A., DeWitt, J.C., Groves, A., Mayr, L., Weisman, H.R., et al. (2019). Resolving medulloblastoma cellular architecture by single-cell genomics. *Nature* *572*, 74–79. <https://doi.org/10.1038/s41586-019-1434-6>.

90. Luo, W., Lin, G.N., Song, W., Zhang, Y., Lai, H., Zhang, M., Miao, J., Cheng, X., Wang, Y., Li, W., et al. (2021). Single-cell spatial transcriptomic analysis reveals common and divergent features of developing postnatal granule cerebellar cells and medulloblastoma. *BMC Biol.* *19*, 135. <https://doi.org/10.1186/s12915-021-01071-8>.
91. Hendrikse, L.D., Haldipur, P., Saulnier, O., Millman, J., Sjoboen, A.H., Erickson, A.W., Ong, W., Gordon, V., Coudière-Morrison, L., Mercier, A.L., et al. (2022). Failure of human rhombic lip differentiation underlies medulloblastoma formation. *Nature* *609*, 1021–1028. <https://doi.org/10.1038/s41586-022-05215-w>.
92. Smith, K.S., Bihannic, L., Gudenas, B.L., Haldipur, P., Tao, R., Gao, Q., Li, Y., Aldinger, K.A., Iskusnykh, I.Y., Chizhikov, V.V., et al. (2022). Unified rhombic lip origins of group 3 and group 4 medulloblastoma. *Nature* *609*, 1012–1020. <https://doi.org/10.1038/s41586-022-05208-9>.
93. Jessa, S., Blanchet-Cohen, A., Krug, B., Vladoiu, M., Coutelier, M., Faury, D., Poreau, B., Jay, N.D., Hébert, S., Monlong, J., et al. (2019). Stalled developmental programs at the root of pediatric brain tumors. *Nat. Genet.* *51*, 1702. <https://doi.org/10.1038/s41588-019-0531-7>.
94. Hendrikse, L.D., Haldipur, P., Saulnier, O., Millman, J., Sjoboen, A.H., Erickson, A.W., Ong, W., Gordon, V., Coudière-Morrison, L., Mercier, A.L., et al. (2022). Failure of human rhombic lip differentiation underlies medulloblastoma formation. *Nature* *609*, 1021–1028. <https://doi.org/10.1038/s41586-022-05215-w>.
95. Arora, S., Nuechterlein, N., Jensen, M., Glatzer, G., Sievers, P., Varadharajan, S., Korshunov, A., Sahm, F., Mack, S.C., Taylor, M.D., et al. (2025). Integrated transcriptomic landscape of medulloblastoma and ependymoma reveals novel tumor subtype-specific biology. Preprint at bioRxiv, <https://doi.org/10.1101/2024.10.21.619495>.
96. Vladoiu, M.C., El-Hamamy, I., Donovan, L.K., Farooq, H., Holgado, B.L., Sundaravadanam, Y., Ramaswamy, V., Hendrikse, L.D., Kumar, S., Mack, S.C., et al. (2019). Childhood cerebellar tumours mirror conserved fetal transcriptional programs. *Nature* *572*, 67–73. <https://doi.org/10.1038/s41586-019-1158-7>.
97. Cavalli, F.M.G., Remke, M., Rampasek, L., Peacock, J., Shih, D.J.H., Luu, B., Garzia, L., Torchia, J., Nor, C., Morrissy, A.S., et al. (2017). Intertumoral Heterogeneity within Medulloblastoma Subgroups. *Cancer Cell* *31*, 737–754.e6. <https://doi.org/10.1016/j.ccell.2017.05.005>.
98. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., and Yosef, N. (2018). Deep Generative Modeling for Single-cell Transcriptomics. *Nat. Methods* *15*, 1053–1058. <https://doi.org/10.1038/s41592-018-0229-2>.
99. Hao, Y., Stuart, T., Kowalski, M., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C., et al. (2023). Dictionary learning for integrative, multimodal, and massively scalable single-cell analysis. *Nat. Biotechnol.* *42*, 293. <https://doi.org/10.1038/s41587-023-01767-y>.
100. Hänzelmann, S., Castelo, R., and Guinney, J. (2013). GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* *14*, 7. <https://doi.org/10.1186/1471-2105-14-7>.
101. Zhong, S., Wang, M., Huang, L., Chen, Y., Ge, Y., Zhang, J., Shi, Y., Dong, H., Zhou, X., Wang, B., et al. (2023). Single-cell epigenomics and spatiotemporal transcriptomics reveal human cerebellar development. *Nat. Commun.* *14*, 7613. <https://doi.org/10.1038/s41467-023-43568-6>.
102. Heide, M., Haffner, C., Murayama, A., Kurotaki, Y., Shinohara, H., Okano, H., Sasaki, E., and Huttner, W.B. (2020). Human-specific ARHGAP11B increases size and folding of primate neocortex in the fetal marmoset. *Science* *369*, 546–550. <https://doi.org/10.1126/science.abb2401>.
103. Kalebic, N., Gilardi, C., Albert, M., Namba, T., Long, K.R., Kostic, M., Langen, B., and Huttner, W.B. (2018). Human-specific ARHGAP11B induces hallmarks of neocortical expansion in developing ferret neocortex. *eLife* *7*, e41241. <https://doi.org/10.7554/eLife.41241>.
104. Owa, T., Taya, S., Miyashita, S., Yamashita, M., Adachi, T., Yamada, K., Yokoyama, M., Aida, S., Nishioka, T., Inoue, Y.U., et al. (2018). Meis1 Coordinates Cerebellar Granule Cell Development by Regulating Pax6 Transcription, BMP Signaling and Atoh1 Degradation. *J. Neurosci.* *38*, 1277–1294. <https://doi.org/10.1523/JNEUROSCI.1545-17.2017>.

105. Corrales, J.D., Rocco, G.L., Blaess, S., Guo, Q., and Joyner, A.L. (2004). Spatial pattern of sonic hedgehog signaling through Gli genes during cerebellum development. *Development* 131, 5581–5590. <https://doi.org/10.1242/dev.01438>.
106. Chizhikov, V.V., Davenport, J., Zhang, Q., Shih, E.K., Cabello, O.A., Fuchs, J.L., Yoder, B.K., and Millen, K.J. (2007). Cilia Proteins Control Cerebellar Morphogenesis by Promoting Expansion of the Granule Progenitor Pool. *J. Neurosci.* 27, 9780–9789. <https://doi.org/10.1523/JNEUROSCI.5586-06.2007>.
107. Flora, A., Garcia, J.J., Thaller, C., and Zoghbi, H.Y. (2007). The E-protein Tcf4 interacts with Math1 to regulate differentiation of a specific subset of neuronal progenitors. *Proc. Natl. Acad. Sci. U. S. A.* 104, 15382–15387. <https://doi.org/10.1073/pnas.0707456104>.
108. Whittaker, D.E., Riegman, K.L.H., Kasah, S., Mohan, C., Yu, T., Sala, B.P., Hebaishi, H., Caruso, A., Marques, A.C., Michetti, C., et al. The chromatin remodeling factor CHD7 controls cerebellar development by regulating reelin expression. *J. Clin. Invest.* 127, 874–887. <https://doi.org/10.1172/JCI83408>.
109. Singh, S., Howell, D., Trivedi, N., Kessler, K., Ong, T., Rosmaninho, P., Raposo, A.A., Robinson, G., Roussel, M.F., Castro, D.S., et al. Zeb1 controls neuron differentiation and germinal zone exit by a mesenchymal-epithelial-like transition. *eLife* 5, e12717. <https://doi.org/10.7554/eLife.12717>.
110. Fratini, L., Dalmolin, M.G.S., Sinigaglia, M., da Silveira Perla, A., de Farias, C.B., Brunetto, A.L., Brunetto, A.T., da Cunha Jaeger, M., and Roesler, R. (2023). ZEB1 is a Subgroup-Specific Marker of Prognosis and Potential Drug Target in Medulloblastoma. *Neuromolecular Med.* 25, 64–74. <https://doi.org/10.1007/s12017-022-08716-z>.
111. Krausert, S., Brabetz, S., Mack, N.L., Schmitt-Hoffner, F., Schwalm, B., Peterziel, H., Mangang, A., Holland-Letz, T., Sieber, L., Korshunov, A., et al. (2022). Predictive modeling of resistance to SMO inhibition in a patient-derived orthotopic xenograft model of SHH medulloblastoma. *Neuro-Oncol. Adv.* 4, vdac026. <https://doi.org/10.1093/oaajnl/vdac026>.
112. Kool, M., Jones, D.T.W., Jäger, N., Northcott, P.A., Pugh, T.J., Hovestadt, V., Piro, R.M., Esparza, L.A., Markant, S.L., Remke, M., et al. (2014). Genome sequencing of SHH medulloblastoma predicts genotype-related response to smoothened inhibition. *Cancer Cell* 25, 393–405. <https://doi.org/10.1016/j.ccr.2014.02.004>.
113. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34, 166–176. <https://doi.org/10.1038/ng1165>.
114. Badia-i-Mompel, P., Wessels, L., Müller-Dott, S., Trimbou, R., Ramirez Flores, R.O., Argelaguet, R., and Saez-Rodriguez, J. (2023). Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* 24, 739–754. <https://doi.org/10.1038/s41576-023-00618-5>.
115. Ray, S., Chaturvedi, N.K., Bhakat, K.K., Rizzino, A., and Mahapatra, S. (2021). Subgroup-Specific Diagnostic, Prognostic, and Predictive Markers Influencing Pediatric Medulloblastoma Treatment. *Diagnostics* 12, 61. <https://doi.org/10.3390/diagnostics12010061>.
116. Baralle, F.E., and Giudice, J. (2017). Alternative splicing as a regulator of development and tissue identity. *Nat. Rev. Mol. Cell Biol.* 18, 437–451. <https://doi.org/10.1038/nrm.2017.27>.
117. Zhang, Y., Qian, J., Gu, C., and Yang, Y. (2021). Alternative splicing and cancer: a systematic review. *Signal Transduct. Target. Ther.* 6, 78. <https://doi.org/10.1038/s41392-021-00486-7>.
118. Chen, J., Hackett, C.S., Zhang, S., Song, Y.K., Bell, R.J.A., Molinaro, A.M., Quigley, D.A., Balmain, A., Song, J.S., Costello, J.F., et al. (2015). The genetics of splicing in neuroblastoma. *Cancer Discov.* 5, 380–395. <https://doi.org/10.1158/2159-8290.CD-14-0892>.
119. Shi, Y., Raklli, V., Maxymovitz, E., Li, S., Westerlund, I., Bedoya-Reina, O.C., Yuan, J., Bullova, P., Juhlin, C.C., Stenman, A., et al. (2020). Alternative splicing in neuroblastoma generates RNA-fusion transcripts and is associated with vulnerability to spliceosome inhibitors. Preprint at bioRxiv, <https://doi.org/10.1101/851238> <https://doi.org/10.1101/851238>.
120. Westermann, F., Muth, D., Benner, A., Bauer, T., Henrich, K.-O., Oberthuer, A., Brors, B., Beissbarth, T., Vandesompele, J., Pattyn, F., et al. (2008). Distinct transcriptional MYCN/c-MYC activities are associated with spontaneous regression or malignant progression in neuroblastomas. *Genome Biol.* 9, R150. <https://doi.org/10.1186/gb-2008-9-10-r150>.

121. Hastie, N.D. (2017). Wilms' tumour 1 (WT1) in development, homeostasis and disease. *Development* 144, 2862–2872. <https://doi.org/10.1242/dev.153163>.
122. Anaka, M.R., Morison, I.M., and Reeve, A.E. (2010). The developmental programme for genesis of the entire kidney is recapitulated in Wilms tumour. *PLoS One* 5, e9488. <https://doi.org/10.1371/journal.pone.0009488>.
123. Nakagawara, A. (2001). Trk receptor tyrosine kinases: a bridge between cancer and neural development. *Cancer Lett.* 169, 107–114. [https://doi.org/10.1016/s0304-3835\(01\)00530-4](https://doi.org/10.1016/s0304-3835(01)00530-4).
124. Brodeur, G.M. (2003). Neuroblastoma: biological insights into a clinical enigma. *Nat. Rev. Cancer* 3, 203–216. <https://doi.org/10.1038/nrc1014>.
125. Nakagawara, A., Arima-Nakagawara, M., Scavarda, N.J., Azar, C.G., Cantor, A.B., and Brodeur, G.M. (1993). Association between high levels of expression of the TRK gene and favorable outcome in human neuroblastoma. *N. Engl. J. Med.* 328, 847–854. <https://doi.org/10.1056/NEJM199303253281205>.
126. Pattwell, S.S., Arber, C.E., and Brodeur, G.M. (2022). The role of neurotrophins in neuroblastoma: from early development to clinical intervention. *Semin. Cancer Biol.* 82, 76–91. <https://doi.org/10.1016/j.semcancer.2021.03.027>.
127. Pattwell, S.S., Arora, S., Cimino, P.J., Ozawa, T., Szulzewsky, F., Hoellerbauer, P., Bonifert, T., Hoffstrom, B.G., Boiani, N.E., Bolouri, H., et al. (2020). A kinase-deficient NTRK2 splice variant predominates in glioma and amplifies several oncogenic signaling pathways. *Nat. Commun.* 11, 2977. <https://doi.org/10.1038/s41467-020-16786-5>.
128. Pattwell, S.S., Arora, S., Nuechterlein, N., Zager, M., Loeb, K.R., Cimino, P.J., Holland, N.C., Reche-Ley, N., Bolouri, H., Bonnin, D.A.A., et al. (2022). Oncogenic role of a developmentally regulated NTRK2 splice variant. *Sci. Adv.* 8, eabo6789. <https://doi.org/10.1126/sciadv.abo6789>.
129. Jagana, H.L., Shabar, M.M., Jackson, T.S., Johnson, D.E., Hemenway, J.M., Mukkamala, V., Lukasik, A., Hathaway, M.R., and Pattwell, S.S. (2026). Oncogenic influences of neurotrophin receptors: Shedding light on Trk biology. *Cell Rep.* 45, 116928. <https://doi.org/10.1016/j.celrep.2026.116928>.
130. Luberg, K., Wong, J., Bhattarai, P., Bhattacharya, M., and others (2022). TrkB-T1 is expressed embryonically and is highly expressed in adult tumors. *Cell Rep.* 38, 110173. <https://doi.org/10.1016/j.celrep.2021.110173>.
131. Nakagawara, A., Azar, C.G., Scavarda, N.J., and Brodeur, G.M. (1994). Expression and function of TRK-B and BDNF in human neuroblastomas. *Mol. Cell. Biol.* 14, 759–767. <https://doi.org/10.1128/mcb.14.1.759-767.1994>.
132. Pattwell, S.S., Arber, C.E., Bhattarai, P., and Brodeur, G.M. (2022). The role of neurotrophins in neuroblastoma: from early development to targeted therapy. *Nat. Rev. Cancer*.
133. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat. Methods* 14, 417–419. <https://doi.org/10.1038/nmeth.4197>.
134. Zeineldin, M., Patel, A.G., and Dyer, M.A. (2022). Neuroblastoma: When differentiation goes awry. *Neuron* 110, 2916–2928. <https://doi.org/10.1016/j.neuron.2022.07.012>.
135. Tomolonis, J.A., Agarwal, S., and Shohet, J.M. (2018). Neuroblastoma pathogenesis: deregulation of embryonic neural crest development. *Cell Tissue Res.* 372, 245–262. <https://doi.org/10.1007/s00441-017-2747-0>.
136. Reiff, T., Tsarovina, K., Majdazari, A., Schmidt, M., del Pino, I., and Rohrer, H. (2010). Neuroblastoma Phox2b variants stimulate proliferation and dedifferentiation of immature sympathetic neurons. *J. Neurosci.* 30, 905–915. <https://doi.org/10.1523/JNEUROSCI.5368-09.2010>.
137. Mossé, Y.P., Laudenslager, M., Khazi, D., Carlisle, A.J., Winter, C.L., Rappaport, E., and Maris, J.M. (2004). Germline PHOX2B mutation in hereditary neuroblastoma. *Am. J. Hum. Genet.* 75, 727–730. <https://doi.org/10.1086/424530>.
138. Yuan, A., Sasaki, T., Kumar, A., Peterhoff, C.M., Rao, M.V., Liem, R.K., Julien, J.-P., and Nixon, R.A. (2012). Peripherin is a subunit of peripheral nerve neurofilaments: implications for differential vulnerability of CNS and peripheral nervous system axons. *J. Neurosci. Off. J. Soc. Neurosci.* 32, 8501–8508. <https://doi.org/10.1523/JNEUROSCI.1081-12.2012>.

139. Light, J.E., Koyama, H., Minturn, J.E., Ho, R., Simpson, A.M., Iyer, R., Mangino, J.L., Kolla, V., London, W.B., and Brodeur, G.M. (2012). Clinical significance of NTRK family gene expression in neuroblastomas. *Pediatr. Blood Cancer* 59, 226–232. <https://doi.org/10.1002/pbc.23343>.
140. Gonzalez Castro, L.N., Tirosh, I., and Suvà, M.L. (2021). Decoding Cancer Biology One Cell at a Time. *Cancer Discov.* 11, 960–970. <https://doi.org/10.1158/2159-8290.CD-20-1376>.
141. Pereira, F.A., Qiu, Y., Zhou, G., Tsai, M.-J., and Tsai, S.Y. (1999). The orphan nuclear receptor COUP-TFII is required for angiogenesis and heart development. *Genes Dev.* 13, 1037–1049. <https://doi.org/10.1101/gad.13.8.1037>.
142. Li, L., Galichon, P., Bhargava, V., D’Agati, V., Bhargava, V., and others (2015). The nuclear receptor COUP-TFII regulates kidney development and disease. *Nat. Commun.* 6, 6771. <https://doi.org/10.1038/ncomms7771>.
143. Wellik, D.M., Hawkes, P.J., and Capecchi, M.R. (2002). Hox11 paralogous genes are essential for metanephric kidney induction. *Genes Dev.* 16, 1423–1432. <https://doi.org/10.1101/gad.993302>.
144. Patterson, L.T., and Potter, S.S. (2004). Hox genes and kidney patterning. *Curr. Opin. Nephrol. Hypertens.* 13, 1–7. <https://doi.org/10.1097/00041552-200401000-00001>.
145. Bouchard, M., Souabni, A., Mandler, M., Neubüser, A., and Busslinger, M. (2002). Nephric lineage specification by Pax2 and Pax8. *Genes Dev.* 16, 2958–2970. <https://doi.org/10.1101/gad.240102>.
146. Reginensi, A., Scott, R.P., Gregorieff, A., Bagherie-Lachidan, M., Chung, C., Lim, D.-S., Pawson, T., Wrana, J., and McNeill, H. (2013). Yap- and Cdc42-dependent nephrogenesis and morphogenesis during mouse kidney development. *PLOS Genet.* 9, e1003380. <https://doi.org/10.1371/journal.pgen.1003380>.
147. Costantini, F., and Shakya, R. (2006). GDNF/Ret signaling and the development of the kidney. *BioEssays* 28, 117–127. <https://doi.org/10.1002/bies.20357>.
148. Ogawa, O., Eccles, M.R., Szeto, J., McNoe, L.A., Yun, K., Maw, M.A., Smith, P.J., and Reeve, A.E. (1993). Relaxation of insulin-like growth factor II gene imprinting implicated in Wilms’ tumour. *Nature* 362, 749–751. <https://doi.org/10.1038/362749a0>.
149. Ravenel, J.D., Broman, K.W., Perlman, E.J., Niemitz, E.L., Jayawardena, T.M., Bell, D.W., Haber, D.A., Uejima, H., and Feinberg, A.P. (2001). Loss of imprinting of insulin-like growth factor-II (IGF2) gene in distinguishing specific biologic subtypes of Wilms tumor. *J. Natl. Cancer Inst.* 93, 1698–1703. <https://doi.org/10.1093/jnci/93.22.1698>.
150. Brouwer-Visser, J., and Huang, G.S. (2014). IGF2 signaling and regulation in cancer. *Cytokine Growth Factor Rev.* 25, 667–675. <https://doi.org/10.1016/j.cytogfr.2014.09.001>.
151. Boyle, S., Misfeldt, A., Chandler, K.J., Deal, K.K., Southard-Smith, E.M., Mortlock, D.P., Baldwin, H.S., and de Caestecker, M. (2008). Fate mapping using Cited1-CreERT2 mice demonstrates that the cap mesenchyme contains self-renewing progenitor cells and gives rise exclusively to nephronic epithelia. *Dev. Biol.* 313, 234–245. <https://doi.org/10.1016/j.ydbio.2007.10.014>.
152. Tretiakova, M., Turkyilmaz, M., Engelman, D., Lin, O., Rajan, N., and others (2015). Glypican-3 overexpression in primary and metastatic Wilms tumors. *Virchows Arch.* 466, 67–76. <https://doi.org/10.1007/s00428-014-1669-4>.
153. Karner, C.M., Das, A., Ma, Z., Self, M., Chen, C., Lum, L., Oliver, G., and Carroll, T.J. (2011). Canonical Wnt9b signaling balances progenitor cell expansion and differentiation during kidney development. *Development* 138, 1247–1257. <https://doi.org/10.1242/dev.057646>.
154. Park, J.-S., Ma, W., O’Brien, L.L., Chung, E., Guo, J.-J., Cheng, J.-G., Valerius, M.T., McMahon, J.A., Wong, W.H., and McMahon, A.P. (2012). Six2 and Wnt regulate self-renewal and commitment of nephron progenitors through shared gene regulatory networks. *Dev. Cell* 23, 637–651. <https://doi.org/10.1016/j.devcel.2012.07.008>.
155. Melsted, P., Boeshaghi, A.S., Liu, L., Gao, F., Lu, L., Min, K.H. (Joseph), da Veiga Beltrame, E., Hjörleifsson, K.E., Gehring, J., and Pachter, L. (2021). Modular, efficient and constant-memory single-cell RNA-seq preprocessing. *Nat. Biotechnol.* 39, 813–818. <https://doi.org/10.1038/s41587-021-00870-2>.

156. Gegenhuber, B., Wu, M.V., Bronstein, R., and Tollkuhn, J. (2022). Gene regulation by gonadal hormone receptors underlies brain sex differences. *Nature* *606*, 153–159. <https://doi.org/10.1038/s41586-022-04686-1>.
157. Kang, H.J., Kawasawa, Y.I., Cheng, F., Zhu, Y., Xu, X., Li, M., Sousa, A.M.M., Pletikos, M., Meyer, K.A., Sedmak, G., et al. (2011). Spatio-temporal transcriptome of the human brain. *Nature* *478*, 483–489. <https://doi.org/10.1038/nature10523>.
158. Andrzejewski, D., Zhu, X., and Craven, M. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors.
159. Chen, J., Zhu, J., Lu, J., and Liu, S. (2018). Scalable training of hierarchical topic models. *Proc. VLDB Endow.* *11*, 826–839. <https://doi.org/10.14778/3192965.3192972>.
160. Dai, F.T., Mounia Lalmas and Zhenwen Scalable Dynamic Topic Modeling. Spotify Res. <https://research.atspotify.com/2022/11/scalable-dynamic-topic-modeling>.
161. Tomasi, F., Lalmas, M., and Dai, Z. (2022). Efficient inference for dynamic topic modeling with large vocabularies. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence (PMLR)*, pp. 1950–1959.
162. Tomasi, F., Chandar, P., Levy-Fix, G., Lalmas-Roelleke, M., and Dai, Z. (2020). Stochastic Variational Inference for Dynamic Correlated Topic Models. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI) (PMLR)*, pp. 859–868.
163. Seita, J., and Weissman, I.L. (2010). Hematopoietic stem cell: self-renewal versus differentiation. *Wiley Interdiscip. Rev. Syst. Biol. Med.* *2*, 640–653. <https://doi.org/10.1002/wsbm.86>.
164. Cheng, H., Zheng, Z., and Cheng, T. (2020). New paradigms on hematopoietic stem cell differentiation. *Protein Cell* *11*, 34–44. <https://doi.org/10.1007/s13238-019-0633-0>.
165. Palis, J. (2014). Primitive and definitive erythropoiesis in mammals. *Front. Physiol.* *5*, 3. <https://doi.org/10.3389/fphys.2014.00003>.
166. Mohandas, N., and Gallagher, P.G. (2008). Red cell membrane: past, present, and future. *Blood* *112*, 3939–3948. <https://doi.org/10.1182/blood-2008-07-161166>.
167. Machlus, K.R., and Italiano, J.E., Jr. (2013). The incredible journey: from megakaryocyte development to platelet formation. *J. Cell Biol.* *201*, 785–796. <https://doi.org/10.1083/jcb.201304054>.
168. Wang, S., Song, R., Wang, Z., Jing, Z., Wang, S., and Ma, J. (2018). S100A8/A9 in Inflammation. *Front. Immunol.* *9*, 1298. <https://doi.org/10.3389/fimmu.2018.01298>.
169. Borregaard, N. (2010). Neutrophils, from marrow to microbes. *Immunity* *33*, 657–670. <https://doi.org/10.1016/j.immuni.2010.11.011>.
170. Igarashi, K., Kurosaki, T., and Roychoudhuri, R. (2017). BACH transcription factors in innate and adaptive immunity. *Nat. Rev. Immunol.* *17*, 437–450. <https://doi.org/10.1038/nri.2017.26>.
171. Ginhoux, F., and Guilliams, M. (2016). Tissue-resident macrophage ontogeny and homeostasis. *Immunity* *44*, 439–449. <https://doi.org/10.1016/j.immuni.2016.02.024>.
172. Lemke, G. (2013). Biology of the TAM receptors. *Cold Spring Harb. Perspect. Biol.* *5*, a009076. <https://doi.org/10.1101/cshperspect.a009076>.
173. Scott, C.L., T’Jonck, W., Martens, L., Guilliams, M., and al, et (2018). The transcription factor ZEB2 is required to maintain the tissue-specific identities of macrophages. *Immunity* *49*, 312–325. <https://doi.org/10.1016/j.immuni.2018.07.004>.
174. Buckingham, M., Bajard, L., Chang, T., Montarras, D., and Rocancourt, D. (2003). The formation of skeletal muscle: from somite to limb. *J. Anat.* *202*, 59–68. <https://doi.org/10.1046/j.1469-7580.2003.00139.x>.
175. Chal, J., and Pourquié, O. (2017). Making muscle: skeletal myogenesis in vivo and in vitro. *Development* *144*, 2104–2122. <https://doi.org/10.1242/dev.151035>.

176. Relaix, F., and Zammit, P.S. (2012). Satellite cells are essential for skeletal muscle regeneration: the cell on the edge returns centre stage. *Development* 139, 2845–2856. <https://doi.org/10.1242/dev.069088>.
177. Olson, E.N. (2006). Gene regulatory networks in the evolution and development of the heart. *Science* 313, 1922–1927. <https://doi.org/10.1126/science.1132292>.
178. Pittenger, M.F., Mackay, A.M., Beck, S.C., Jaiswal, R.K., and Douglas, R. (1999). Multilineage potential of adult human mesenchymal stem cells. *Science* 284, 143–147. <https://doi.org/10.1126/science.284.5411.143>.
179. Long, F. (2012). Building strong bones: molecular regulation of the osteoblast lineage. *Nat. Rev. Mol. Cell Biol.* 13, 27–38. <https://doi.org/10.1038/nrm3254>.
180. Lefebvre, V., and Bhattaram, P. (2010). Vertebrate skeletogenesis. *Curr. Top. Dev. Biol.* 90, 291–317. [https://doi.org/10.1016/S0070-2153\(10\)90008-2](https://doi.org/10.1016/S0070-2153(10)90008-2).
181. Rodriguez-Boulan, E., and Macara, I.G. (2014). Organization and execution of the epithelial polarity programme. *Nat. Rev. Mol. Cell Biol.* 15, 225–242. <https://doi.org/10.1038/nrm3775>.
182. Whitsett, J.A., Wert, S.E., and Weaver, T.E. (2010). Alveolar surfactant homeostasis and the pathogenesis of pulmonary disease. *Annu. Rev. Med.* 61, 105–119. <https://doi.org/10.1146/annurev.med.60.041807.123500>.
183. Mou, H. Induced Pluripotent Stem Cell-derived Alveolar Type II Heterogeneity: Revealed by SFTPC Expression. *Am. J. Respir. Cell Mol. Biol.* 65, 345–346. <https://doi.org/10.1165/rcmb.2021-0242ED>.
184. McMahon, A.P. (2016). Development of the mammalian kidney. *Curr. Top. Dev. Biol.* 117, 31–64. <https://doi.org/10.1016/bs.ctdb.2015.10.010>.
185. Le Douarin, N.M., and Kalcheim, C. (1999). *The Neural Crest* 2nd ed. (Cambridge University Press) <https://doi.org/10.1017/CBO9780511897948>.
186. Simões-Costa, M., and Bronner, M.E. (2015). Establishing neural crest identity: a gene regulatory recipe. *Development* 142, 242–257. <https://doi.org/10.1242/dev.105445>.
187. Schwob, J.E. (2002). Neural regeneration and the peripheral olfactory system. *Anat. Rec.* 269, 33–49. <https://doi.org/10.1002/ar.10047>.
188. Zorn, A.M., and Wells, J.M. (2009). Vertebrate endoderm development and organ formation. *Annu. Rev. Cell Dev. Biol.* 25, 221–251. <https://doi.org/10.1146/annurev.cellbio.042308.113344>.
189. Gordillo, M., Evans, T., and Bhattacharya, S. (2015). Orchestrating liver development. *Development* 142, 2094–2108. <https://doi.org/10.1242/dev.114215>.
190. Si-Tayeb, K., Lemaigre, F.P., and Duncan, S.A. (2010). Organogenesis and development of the liver. *Dev. Cell* 18, 175–189. <https://doi.org/10.1016/j.devcel.2010.01.011>.
191. Shaker, A., and Rubin, D.C. (2010). Intestinal stem cells and epithelial-mesenchymal interactions in the crypt and stem cell niche. *Transl. Res. J. Lab. Clin. Med.* 156, 180–187. <https://doi.org/10.1016/j.trsl.2010.06.003>.
192. Barker, N. (2014). Adult intestinal stem cells: critical drivers of epithelial homeostasis and regeneration. *Nat. Rev. Mol. Cell Biol.* 15, 19–33. <https://doi.org/10.1038/nrm3721>.
193. Clevers, H. (2013). The Intestinal Crypt, A Prototype Stem Cell Compartment. *Cell* 154, 274–284. <https://doi.org/10.1016/j.cell.2013.07.004>.
194. Rosen, E.D., and MacDougald, O.A. (2006). Adipocyte differentiation from the inside out. *Nat. Rev. Mol. Cell Biol.* 7, 885–896. <https://doi.org/10.1038/nrm2066>.
195. Potente, M., Gerhardt, H., and Carmeliet, P. (2011). Basic and therapeutic aspects of angiogenesis. *Cell* 146, 873–887. <https://doi.org/10.1016/j.cell.2011.08.039>.